

Lecture 1: Data, Lists and Models

Information is described by **data**. While data entries are often not numerical values initially, they can always be encoded in a numerical manner so that we can look at **numerical data**. Here are some examples: stock data, weather data, networks, vocabularies, address books, encyclopedias, surfaces, books, the human genome, pictures, movies or music pieces. Is there any information which is not described by data?

While data describe complicated objects, they can be organized in **lists**, lists of lists, or lists of lists of lists etc. Stock market or weather data can be represented as finite sequences of numbers, pictures are arrays of numbers, a movie is an array of pictures. A book is a list of letters. A graph is given by a list of nodes and connections between these nodes. A movie is a list of pictures and each picture is a list of list of pixels and each pixel is a list of red, green and blue values. If you think about these examples, you realized that **vectors**, **matrices** ore higher dimensional arrays are helpful to organize the data. We will define these concepts later but a vector is just a finite list of things and a matrix is a list of lists, a **spreadsheet**. In the case of pictures, or music pieces the translation into an array is obvious. Is it always possible to encode data as sequences of numbers or arrays of numbers or lists of arrays of numbers etc? Even for complicated objects like networks, one can use lists. A network can be encoded with an array of numbers, where we put a 1 at the node (i,j) if node i and j are connected and 0 otherwise. This example makes clear that we need a **mathematical language** to describe data. It looks like a difficult problem at first because data can appear in such different shapes and forms. It turns out that we can use vectors to describe data and use matrices to describe relations between these data. The fact that most popular databases are **relational databases** organized with tables vindicates this point of view. In this first lecture, we also want to see that linear algebra is a tool to organize and work with data. Even data manipulation can be described using linear algebra. Data of the same type can be added, scaled. We can mix for example two pictures to get a new picture. Already on a fundamental level, nature takes linear algebra seriously: while classical mechanics deals with differential equations which are in general nonlinear and complicated, quantum mechanics replaces this with a linear evolution of functions. Both in the classical and quantum world, we can describe the evolution of observables with linear laws.

Here are four important uses of linear algebra to describe data:

| Tool | Goal | Using | Example |
|-------------|---------------------|-------------|---------------------------------------|
| Databases | Describe the data | Lists | Relational database, Adjacency matrix |
| Modeling | Model the data | Probability | Markov process, Filtering, Smoothing |
| Fitting | Reduce the data | Projections | Linear regression. |
| Computation | Manipulate the data | Algebra | Fourier theory. |

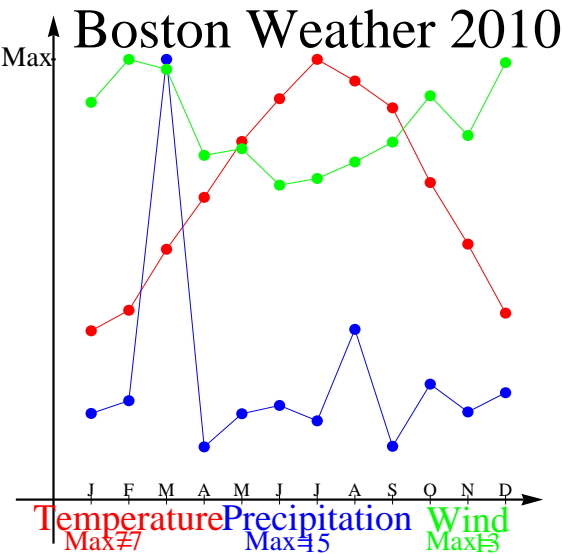
We have seen that a fundamental tool to organize data is the concept of a **list**. Mathematicians call this a **vector**. Since data can be added and scaled, data can be treated as vectors. We can also look at lists of lists. These are called **matrices**. Matrices are important because they allow to describe **relations** and **operations**. Given a matrix, we can access the data using coordinates. The entry $(3,4)$ for example is the forth element in the third row. Having data organized in lists, one can manipulate them more easily. One can use **arithmetic** on entire lists or arrays. This is what spreadsheets do.

Observed quantities are functions defined on lists. One calls them also **random variables**. Because observables can be added or subtracted, they can be treated as vectors. If the data themselves are not numbers like the strings of a DNA, we can add and multiply numerical functions on these data. The function $X(x)$ for example could count the number of A terms in a genome sequence x with letters A,G,C,T which abbreviate Adenin, Guanin, Cytosin and Tymin. It is a fundamental and pretty modern insight that all mathematics can be described using algebras and operators. We have mentioned that data are often related and organized in relational form and that an array of data achieves this. Lets look at weather data accessed from <http://www.nws.noaa.gov> on January 4'th 2011, where one of the row coordinates is "time". The different data vectors are listed side by side and listed in form of a **matrix**.

| Month | Year | Temperature | Precipitation | Wind |
|-------|------|-------------|---------------|------|
| 01 | 2010 | 29.6 | 2.91 | 12.0 |
| 02 | 2010 | 33.2 | 3.34 | 13.3 |
| 03 | 2010 | 43.9 | 14.87 | 13.0 |
| 04 | 2010 | 53.0 | 1.78 | 10.4 |
| 05 | 2010 | 62.8 | 2.90 | 10.6 |
| 06 | 2010 | 70.3 | 3.18 | 9.5 |
| 07 | 2010 | 77.2 | 2.66 | 9.7 |
| 08 | 2010 | 73.4 | 5.75 | 10.2 |
| 09 | 2010 | 68.7 | 1.80 | 10.8 |
| 10 | 2010 | 55.6 | 3.90 | 12.2 |
| 11 | 2010 | 44.8 | 2.96 | 11.0 |
| 12 | 2010 | 32.7 | 3.61 | 13.2 |

To illustrate how linear algebra enters, lets add up all the rows and divide by the number of rows. This is called the **average**. We can also look at the average squre distance to the mean, which is the variance. Its square root is called the **standard deviation**.

| Month | Year | Temperature | Precipitation | Wind |
|-------|------|-------------|---------------|--------|
| 6.5 | 2010 | 53.7667 | 4.13833 | 11.325 |



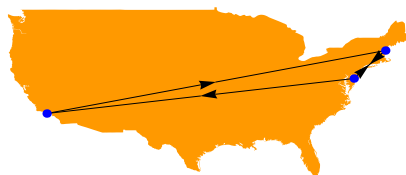
In the example data set, the average day was June 15th, 2010, the average temperature was 54 degrees Fahrenheit = 12 degrees Celsius, with an average of 4 inches of precipitation per month and an average wind speed of 11 miles per hour. The following figure visualizes the data. You see the unusually rainy March which had produced quite a bit of flooding. The rest is not so surprising. The temperatures are of course higher in summer than in winter and there is more wind in the cooler months. Since data often come with noise, one can simplify their model and reduce the amount of information to describe them. When looking at a particular precipitation data in Boston over a year, we have a lot of information which is not interesting. More interesting is the global trend, the deviation from this global trend and correlations within neighboring days. It is for example more likely to be rainy near a rainy day than near a sunny day. The data are not given by a random process like a dice. If we can identify the part of the data which are randomly chosen, how do we find the rule? This is a basic problem and can often be approached using linear algebra. The expectation will tell the most likely value of the data and the standard deviation tells us how noisy the data are. Already these notions have relations with geometry.

How do we model from a situation given only one data set? For example, given the DJI data, we would like to have a good model which predicts the nature of the process in the future. We can do this by looking for trends. Mathematically, this is done by data fitting and will be discussed extensively in this course. An other task is to model the process using a linear algebra model. Maybe it is the case that near a red pixel in a picture it is more likely to have a red pixel again or that after a gain in the DJI, we are more likely to gain more.

Lets illustrate how we can use lists of data to encode a traffic situation. Assume an airline services the towns of Boston, New York and Los Angeles. It flies from New York to Los Angeles, from Los Angeles to Boston and from Boston to New York as well as from New York to Boston. How can we compute the number of round trips of any length n in this network? Linea algebra helps: define the 3×3 **connection matrix** A given below and compute the n 'th power of the matrix. We will learn how to do that next week. In our case, there are 670976837021 different round trips of length 100 starting from Boston.

The matrix which encodes the situation is the following:

$$A = \begin{bmatrix} & BO & NY & LA \\ BO & 0 & 1 & 0 \\ NY & 1 & 0 & 1 \\ LA & 1 & 0 & 0 \end{bmatrix}$$



To summarize, linear algebra enters in many different ways into data analysis. Lists and lists of lists are fundamental ways to represent data. They will be called vectors and matrices. Linear algebra is needed to find good models or to reduce data. Finally, even if we have a model, we want to do computations efficiently.

Homework due February 2, 2011

1 In the movie "Fermat's room", some mathematicians are trapped in a giant press and have to solve mathematical problems to stop the press from crushing them to death. In one of the math problems, they get the data stream of $169 = 13^2$ digits:
 00000000000000011111111100011111111100111111111111111001110
 001000110011000100011001111101111100111100011110001111111
 1100000101010100000011010110000001111111000000000000000. Can you solve the riddle? Try to solve it yourself at first. If you need a hint, watch the clip <http://www.math.harvard.edu/~knill/mathmovies/text/fermat/riddle3.html>

2 This is problem 2 in Section 1.3 of the script, where 200 is replaced by 100: Flip a coin 100 times. Record the number of times, heads appeared in the first 10 experiments and call this n_1 . Then call the number of times, heads appears in the next $N = 10$ experiments and call it n_2 . This produces 10 inters n_1, \dots, n_{10} . Find the **mean**

$$m = (n_1 + n_2 + \dots + n_N)/N$$

of your data, then the **sample variance**

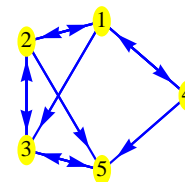
$$v = ((n_1 - m)^2 + (n_2 - m)^2 + \dots + (n_N - m)^2)/(N - 1)$$

and finally the **sample standard deviation** $\sigma = \sqrt{v}$. Remark: there are statistical reasons that $(N - 1)$ is chosen and not N . It is called **Bessel's correction**.

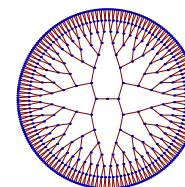
If you use a computer, make a million experiments at least.

```
R:=Random[Integer,1]; M=100; data=Table[R,{M}];
m=Sum[data[[k]],{k,M}]/M;
sigma=Sqrt[Sum[(data[[k]]-m)^2,{k,M}]/(M-1)];
```

3 a) Encode the following directed graph in a list of 5×5 integers 0 or 1. We call this a matrix. Use the rule that if there is an arrow from node i to node j , then the matrix A has an entry 1 in the i 'th row and j 'th column.



b) We often organize the data in a so called **binary tree**. Each vertex in the following tree can be described as a list of numbers $\{0, 1\}$. How is each point on the most outer circle determined by a list with 8 numbers? Assume you start at the origin and make steps to the right if 1 occurs in the list and left if 0 occurs. Assume you get the list $\{1, 1, 1, 1, 1, 1, 1, 1\}$. At which point do you end up? Which part of the outer circle corresponds to sequences which start with two 0's?



Lecture 2: Probability notions

A **probability space** consists of a set Ω called **laboratory** a set of subsets of Ω called **events** and a function P from events to the interval $[0, 1]$ so that probabilities of disjoint sets add up and such that the entire laboratory has probability 1. Every point in Ω is an **experiment**. Think of an event A as a collection of experiments and of $P[A]$ as the likelihood that A occurs.

Examples:

- 1 We turn a wheel of fortune and assume it is fair in the sense that every angle range $[a, b]$ appears with probability $(b - a)/2\pi$. What is the chance that the wheel stops with an angle between 30 and 90 degrees?

Answer: The laboratory here is the circle $[0, 2\pi)$. Every point in this circle is a possible experiment. The event that the wheel stops between 30 and 90 degrees is the interval $[\pi/6, \pi/2]$. Assuming that all angles are equally probable, the answer is $1/6$.

- 2 This example is called **Bertrand's paradox**. Assume we throw randomly a line into the unit disc. What is the probability that its length is larger than the length of the inscribed triangle?

Answer: Interestingly, the answer depends as we will see in the lecture.

- 3 Lets look at the DowJonesIndustrial average DJI from the start. What is the probability that the index will double in the next 50 years?

Answer: This is a strange question because we have **only one** data set. How can we talk about probability in this situation? One way is to see this graph as a sample of a larger probability space. A simple model would be to fit the data with some polynomial, then add random noise to it. The real DJI graph now looks very similar to a typical graph of those.

- 4 Lets look at the digits of π . What is the probability that the digit 5 appears? **Answer:** Also this is a strange example since the digits are not randomly generated. They are given by nature. There is no randomness involved. Still, one observes that the digits behave like a random number and that the number is "normal": every digit appears with the same frequency. This is independent of the base.

Here is a more precise list of conditions which need to be satisfied for events.

1. The entire laboratory Ω and the empty set \emptyset are events.
2. If A_j is a sequence of events, then $\bigcup_j A_j$ and $\bigcap_j A_j$ are events.

It follows that also the complement of an event is an event.

Here are the conditions which need to be satisfied for the **probability function** P :

1. $0 \leq P[A] \leq 1$ and $P[\Omega] = 1$.
2. A_j are disjoint events, then $P[\bigcup_{j=1}^{\infty} A_j] = \sum_{j=1}^{\infty} P[A_j]$.

$$\begin{aligned} P[\Omega \setminus A] &= 1 - P[A]. \\ P[A \cup B] &= P[A] + P[B] - P[A \cap B]. \end{aligned}$$

An important class of probability spaces are **finite probability spaces**, where every subset can be an event. The most natural choice is to assign them the probability $P[A] = |A|/|\Omega|$ where $|A|$ is the number of elements in A . This reads the "number of good cases" divided by the "number of all cases".

$$P[A] = \frac{|A|}{|\Omega|}$$

It is important that in any situation, we first find out what the laboratory is. This is often the hardest task. Once the setup is fixed, one has a combinatorics or counting problem.

- 5 We throw a dice twice. What is the probability that the sum is larger than 8? **Answer:** We can enumerate all possible cases in a matrix and get Let

$$\Omega = \begin{bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{bmatrix}.$$

be the possible cases, then there are only 8 cases where the sum is smaller or equal to 8.

- 6 Lets look at all 2×2 matrices for which the entries are either 0 or 1. What is the probability that such a matrix has a nonzero determinant $\det(A) = ad - bc$?

Answer: We have 16 different matrices. Our probability space is finite:

$$\Omega = \left\{ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \right. \\ \left. \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right\}.$$

Now lets look at the event that the determinant is nonzero. It contains the following matrices:

$$A = \left\{ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \right\}.$$

The probability is $P[A] = |A|/|\Omega| = 6/16 = 3/8 = 0.375$.

- 7 Lets pick 2 cards from a deck of 52 cards. What is the probability that we have 2 kings? **Answer:** Our laboratory Ω has $52 * 51$ possible experiments. To count the number of good cases, note that there are $4 * 3 = 12$ possible ordered pairs of two kings. Therefore $12/(52 * 51) = 1/221$ is the probability.

Some notation

Set theory in Ω :

The **intersection** $A \cap B$ contains the elements which are in A and B .

The **union** $A \cup B$ contains the elements which are in A or B .

The **complement** A^c contains the elements in Ω which are **not** in A .

The **difference** $A \setminus B$ contains the elements which are in A but not in B .

The **symmetric difference** $A \Delta B$ contains what is in A or B but not in both.

The **empty set** \emptyset is the set which does not contain any elements.

The algebra \mathcal{A} of events:

If Ω is the laboratory, the set \mathcal{A} of events is σ -**algebra**. It is a set of subsets of Ω in which one can perform countably many set theoretical operations and which contains Ω and \emptyset . In this set one can perform **countable unions** $\bigcup_j A_j$ for the union of a sequence of sets A_1, A_2, \dots or **countable intersections** $\bigcap_j A_j$.

The probability measure P :

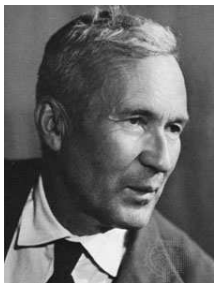
The probability function P from \mathcal{A} to $[0, 1]$ is assumed to be normalized that $P[\Omega] = 1$ and that $P[\bigcup_i A_i] = \sum_i P[A_i]$ if A_i are all disjoint events. The later property is called **σ -additivity**. One gets immediately that $P[A^c] = 1 - P[A]$, $P[\emptyset] = 0$ and that if $A \subset B$ then $P[A] \leq P[B]$.

The Kolmogorov axioms:

A **probability space** (Ω, \mathcal{A}, P) consists of a laboratory set Ω , a σ -algebra \mathcal{A} on Ω and a probability measure P . The number $P[A]$ is the **probability** of an event A . The elements in Ω are called **experiments**. The elements in \mathcal{A} are called **events**.

Some remarks:

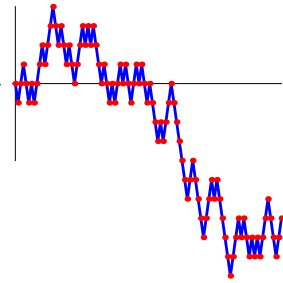
1) In a σ -algebra the operation $A \Delta B$ behaves like **addition** and $A \cap B$ like **multiplication**. We can "compute" with sets like $A \cap (B \Delta C) = (A \cap B) \Delta (A \cap C)$. It is therefore an algebra. One calls it a **Boolean algebra**. Beside the just mentioned **distributivity**, one has **commutativity**, and **associativity**. The "**zero**" is played by \emptyset because $\emptyset \Delta A = A$. The "**one**" is the set Ω because $\Omega \cap A = A$. The algebra is rather special because $A \cap A = A$ and $A \Delta A = \emptyset$. The "square" of each set is the set itself and adding a set to itself gives the zero set.



2) The Kolmogorov axioms form a solid foundation of probability theory. This has only been achieved in the 20th century (1931). Before that probability theory was a bit hazy. For infinite probability spaces it is necessary to restrict the set of all possible events. One can not take all subsets of an interval for example. There is no probability measure P which would work with all sets. There are just too many.

Homework due February 2, 2011

- 1 You walk 100 steps and chose in each step randomly one step forward or backward. You flip a coin. What is the chance to be back at your starting point 0 at the end of your walk?
 - a) Set up the probability space Ω . How many elements does it have?
 - b) Which subset A of Ω is the event to be back at 0 at time 100?
 - c) Find the probability $P[A]$.
 - d) What formula do you get for n steps.



- 2 Do problem 5) in Chapter 2 of the text but with 100 instead of 1000. You choose a random number from $\{1, \dots, 100\}$, where each of the numbers have the same probability. Let A denote the event that the number is divisible by 3 and B the event that the number is divisible by 5. What is the probability $P[A]$ of A , the probability $P[B]$ of B and the probability of $P[A \cap B]$? Compute also the probability $P[A \cap B]/P[B]$ which we will call the **conditional probability** next time. It is the probability that the number is divisible by 3 under the condition that the number is divisible by 5.
- 3 You choose randomly 6 cards from 52 and do not put the cards back. What is the probability that you got all aces? Make sure you describe the probability space Ω and the event A that we have all aces.

Lecture 3: Conditional probability

The **conditional probability** of an event A under the condition that the event B takes place is denoted with $P[A|B]$ and defined to be $P[A \cap B]/P[B]$.

- 1 We throw a coin 3 times. The first 2 times, we have seen head. What is the chance that we get tail the 3th time?

Answer: The probability space Ω consists of all words in the alphabet H, T of length 3. These are $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. The event B is the event that the first 2 cases were head. The event A is the event that the third dice is head.

- 2 **Problem:** Dave has two kids, one of them is a girl. What is the chance that the other is a girl?

Intuitively one would here give the answer $1/2$ because the second event looks independent of the first. However, this initial intuition is misleading and the probability only $1/3$.

Solution. We need to introduce the probability space of all possible events

$$\Omega = \{BG, GB, BB, GG\}$$

with $P[\{BG\}] = P[\{GB\}] = P[\{BB\}] = P[\{GG\}] = 1/4$. Let $B = \{BG, GB, GG\}$ be the event that there is at least one girl and $A = \{GG\}$ the event that both kids are girls. We have

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{(1/4)}{(3/4)} = \frac{1}{3}.$$

- 3 You are in the Monty-Hall game show and need to choose from three doors. Behind one door is a car and behind the others are goats. The host knows what is behind the doors. After you open the first door, he opens an other door with a goat. He asks you whether you want to switch. Do you want to?

Answer: Yes, you definitely should switch. You double your chances to win a car:

No switching: The probability space is the set of all possible experiments $\Omega = \{1, 2, 3\}$. You choose a door and win with probability $\boxed{1/3}$. The opening of the host does not affect any more your choice.

Switching: We have the same probability space. If you pick the car, then you lose because the switch will turn this into a goat. If you choose a door with a goat, the host opens the other door with the goat and you win. Since you win in two cases, the probability is $\boxed{2/3}$.

Also here, intuition can lead to **conditional probability traps** and suggest to have a win probability $1/3$ in general. Lets use the notion of **conditional probability** to give an other correct argument: the intervention of the host has narrowed the laboratory to $\Omega = \{12, 13, 21, 23, 31, 32\}$ where 21 for example means choosing first door 2 then door 1. Assume the car is behind door 1 (the other cases are similar). The host, who we assume always picks door 2 if you pick 1 with the car (the other case is similar) gives us the **condition** $B = \{13, 21, 31\}$ because the cases 23 and 32 are not possible. The winning event is $A = \{21, 31\}$. The answer to the problem is the conditional probability $P[A|B] = P[A \cap B]/P[B] = \boxed{2/3}$.

If A, B be events in the probability space (Ω, P) , then **Bayes rule** holds:

$$P[A|B] = \frac{P[B|A] \cdot P[A]}{P[B|A] + P[B|A^c]}.$$

It is a formula for the **conditional probability** $P[A|B]$ when we know the **unconditional probability** of A and $P[B|A]$, the **likelihood**. The formula immediately follows from the fact that $P[B|A] + P[B|A^c] = P[B]$.

While the formula followed directly from the definition of conditional probability, it is very useful since it allows us to compute the conditional probability $P[A|B]$ from the likelihoods $P[B|A], P[B|A^c]$. Here is an example:

- 4 **Problem.** The probability to die in a car accident in a 24 hour period is one in a million. The probability to die in a car accident at night is one in two millions. At night there is 30% traffic. You hear that a relative of yours died in a car accident. What is the chance that the accident took place at night?

Solution. Let B be the event to die in a car accident and A the event to drive at night. We apply the Bayes formula. We know $P[A \cap B] = P[B|A] \cdot P[A] = (1/2000000) \cdot (3/10) = 3/20000000$.

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = (3/20000000)/(1/1000000) = 3/20.$$

The accident took place at night with a 15 % chance.

A more general version of Bayes rule deals with more than just two possibilities. These possibilities are called A_1, A_2, \dots, A_n .

Bayes rule: If the disjoint events A_1, A_2, \dots, A_n cover the entire laboratory Ω , then

$$P[A_i|B] = \frac{P[B|A_i] \cdot P[A_i]}{\sum_{j=1}^n P[B|A_j] \cdot P[A_j]}.$$

Proof: Because the denominator is $P[B] = \sum_{j=1}^n P[B|A_j]P[A_j]$, the Bayes rule just says $P[A_i|B]P[B] = P[B|A_i]P[A_i]$. But these are by definition both $P[A_i \cap B]$.

- 5 **Problem:** A fair dice is rolled first. It gives a random number k from $\{1, 2, 3, 4, 5, 6\}$. Next, a fair coin is tossed k times. Assume, we know that all coins show heads, what is the probability that the score of the dice was equal to 5?

Solution. Let B be the event that all coins are heads and let A_j be the event that the dice showed the number j . The problem is to find $P[A_5|B]$. We know $P[B|A_j] = 2^{-j}$. Because the events $A_j, j = 1, \dots, 6$ are disjoint sets in Ω which cover it, we have $P[B] = \sum_{j=1}^6 P[B \cap A_j] = \sum_{j=1}^6 P[B|A_j]P[A_j] = \sum_{j=1}^6 2^{-j}/6 = (1/2 + 1/4 + 1/8 + 1/16 + 1/32 + 1/64)(1/6) = 21/128$. By Bayes rule,

$$P[A_5|B] = \frac{P[B|A_5]P[A_5]}{(\sum_{j=1}^6 P[B|A_j]P[A_j])} = \frac{(1/32)(1/6)}{21/128} = \frac{2}{63},$$

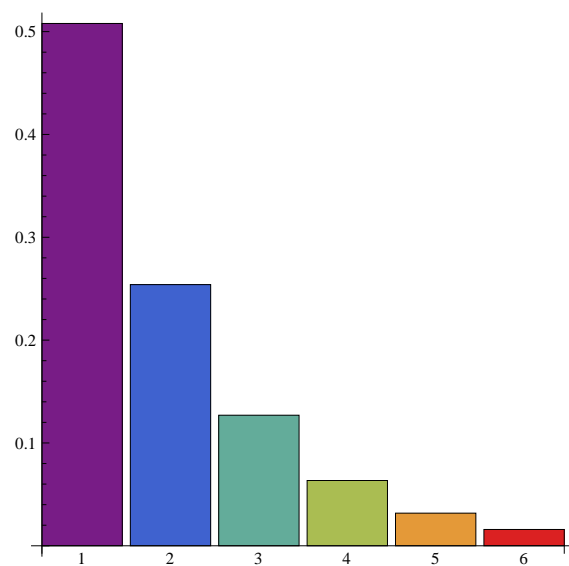


Figure: The conditional probabilities $P[A_j|B]$ in the previous problem. The knowledge that all coins show head makes it more likely to have thrown less coins.

Homework due February 2, 2011

- 1 Problem 2) in Chapter 3: if the probability that a student is sick at a given day is 1 percent and the probability that a student has an exam at a given day is 5 percent. Suppose that 6 percent of the students with exams go to the infirmary. What is the probability that a student in the infirmary has an exam on a given day?
- 2 Problem 5) in chapter 3: Suppose that A, B are subsets of a sample space with a probability function P . We know that $P[A] = 4/5$ and $P[B] = 3/5$. Explain why $P[B|A]$ is at least $1/2$.
- 3 Solve the Monty Hall problem with 4 doors. There are 4 doors with 3 goats and 1 car. What are the winning probabilities in the switching and no-switching cases? You can assume that the host always opens the still closed goat closest to the car.

Lecture 4: Linear equations from Probability

A **linear equation** for finitely many variables x_1, x_2, \dots, x_n is an equation of the form

$$a_1x_1 + a_2x_2 + \dots + a_nx_n = b.$$

Finitely many such equations form a **system of linear equations**. A system of linear equations can be written in matrix form $A\vec{x} = \vec{b}$. Here \vec{x} is the column vector containing the variables, A lists all the coefficients, and \vec{b} is the column vector which lists the numbers to the right.

- 1 Consider the system

$$\begin{array}{cccccccl} x & + & y & + & z & + & u & + & v & + & w & = & 3 \\ & & & & y & + & z & + & u & + & v & = & 2 \\ & & & & & & 2 & + & 2 & & & = & 4 \end{array}$$

There are 6 variables and 3 equations. Since we have less equations than unknowns, we expect infinitely many solutions. The system can be written as $A\vec{x} = \vec{b}$, where

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 & 0 & 0 \end{bmatrix}$$

and $\vec{b} = \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix}$.

- 2 Linear equations appear in probability theory. **Example:** Assume we have a probability space $\Omega = \{x, y, z, w\}$, with four elements. Assume we know the probabilities $P[\{x, y\}] = 4/10$, $P[\{z, w\}] = 6/10$, $P[\{x, z\}] = 7/10$, $P[\{b, d\}] = 3/10$. The question is to find the probabilities $x = P[\{a\}]$, $y = P[\{b\}]$, $z = P[\{c\}]$, $w = P[\{d\}]$.

Answer: This problem leads to a system of equations

$$\begin{array}{rcl} x & + & y & + & & & & & & & & & = & 4/10 \\ & & & & z & + & w & = & 6/10 \\ x & & & & z & + & & = & 7/10 \\ & & y & & & + & w & = & 3/10 \end{array}$$

The system can be solved by eliminating the variables. But the system has no unique solution.

- 3 **Example.** Assume we have two events B_1, B_2 which cover the probability space. We do not know their probabilities. We have two other events A_1, A_2 from which we know $P[A_i]$ and the conditional probabilities $P[A_i|B_j]$. We get the system of equations.

$$\begin{array}{rcl} P[A_1|B_1]P[B_1] & + & P[A_1|B_2]P[B_2] & = & P[A_1] \\ P[A_2|B_1]P[B_1] & + & P[A_2|B_2]P[B_2] & = & P[A_2] \end{array}$$

Here is a concrete example: Assume the chance that the first kid is a girl is 60% and that the probability to have a boy after a boy is $2/3$ and the probability to have a girl after a girl is $2/3$ too. What is the probability that the second kid is a girl?

Solution. Let B_1 be the event that the first kid is a boy and let B_2 the event that the first kid is a girl. Assume that for the first kid the probability to have a girl is 60%. But that $P[\text{Firstgirl}|\text{Secondgirl}] = 2/3$ and $P[\text{Firstboy}|\text{Secondboy}] = 2/3$. What are the probabilities that the first kid is a boy? This produces a system

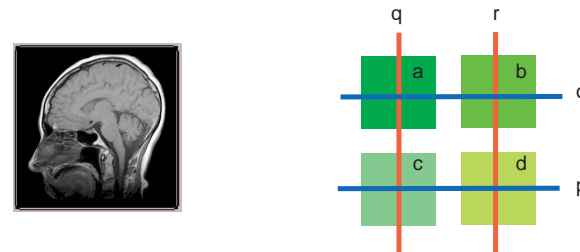
$$\begin{array}{rcl} 2/3P[B_1] & + & 1/3P[B_2] & = & 6/10 \\ 1/3P[B_1] & + & 2/3P[B_2] & = & 4/10 \end{array}$$

The probabilities are $8/15, 7/15$. There is still a slightly larger probability to have a girl. This example is also at the heart of Markov processes.

- 4 **Example** Here is a toy example of a problem one has to solve for magnetic resonance imaging (MRI). This technique makes use of the absorb and emission of energy in the radio frequency range of the electromagnetic spectrum.

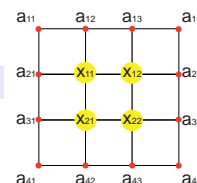
Assume we have 4 hydrogen atoms, whose nuclei are excited with energy intensity a, b, c, d . We measure the spin echo in 4 different directions. $3 = a+b, 7 = c+d, 5 = a+c$ and $5 = b+d$. What is a, b, c, d ? **Solution:** $a = 2, b = 1, c = 3, d = 4$. However, also $a = 0, b = 3, c = 5, d = 2$ solves the problem. This system has not a unique solution even so there are 4 equations

and 4 unknowns.



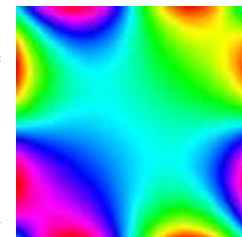
We model a drum by a fine net. The heights at each interior node needs the average the heights of the 4 neighboring nodes. The height at the boundary is fixed. With n^2 nodes in the interior, we have to solve a system of n^2 equations. For example, for $n = 2$ (see left), the $n^2 = 4$ equations are

5



$$\begin{array}{l} 4x_{11} = x_{21} + x_{12} + x_{21} + x_{12}, \\ 4x_{12} = x_{11} + x_{13} + x_{22} + x_{22}, \\ 4x_{21} = x_{31} + x_{11} + x_{22} + a_{43}, \\ 4x_{22} = x_{12} + x_{21} + a_{43} + a_{34}. \end{array}$$

To the right we see the solution to a problem with $n = 300$, where the computer had to solve a system with 90'000 variables. This problem is called a Dirichlet problem and has close ties to probability theory too.



- 6 The last example should show you that linear systems of equations also appear in data fitting even so we do not fit with linear functions. The task is to find a parabola

$$y = ax^2 + bx + c$$

through the points $(1, 3)$, $(2, 1)$ and $(4, 9)$. We have to solve the system

$$\begin{aligned} a + b + c &= 3 \\ 4a + 2b + c &= 1 \\ 16a + 4b + c &= 9 \end{aligned}$$

The solution is $(2, -8, 9)$. The parabola is $y = 2x^2 - 8x + 9$.

Homework due February 10, 2011

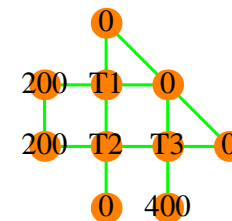
- 1 Problem 24 in 1.1 of Bretscher): On your next trip to Switzerland, you should take the scenic boat ride from Rheinfall to Rheinau and back. The trip downstream from Rheinfall to Rheinau takes 20 minutes, and the return trip takes 40 minutes; the distance between Rheinfall and Rheinau along the river is 8 kilometers. How fast does the boat travel (relative to the water), and how fast does the river Rhein flow in this area? You may assume both speeds to be constant throughout the journey.



- 2 (Problem 28 in 1.1 of Bretscher): In a grid of wires, the temperature at exterior mesh points is maintained at constant values as shown in the figure. When the grid is in thermal equilibrium, the temperature at each interior mesh point is the average of the temperatures at the four adjacent points. For example

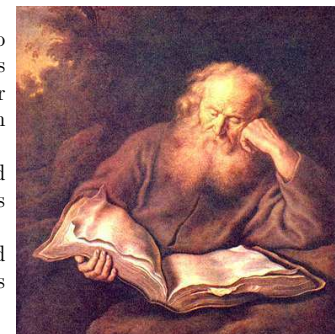
$$T_2 = (T_3 + T_1 + 200 + 0)/4.$$

Find the temperatures T_1, T_2, T_3 when the grid is in thermal equilibrium.



(Problem 46 in 1.1 of Bretscher): A hermit eats only two kinds of food: brown rice and yogurt. The rice contains 3 grams of protein and 30 grams of carbohydrates per serving, while the yogurt contains 12 grams of protein and 20 grams of carbohydrates.

- 3 a) If the hermit wants to take in 60 grams of protein and 300 grams of carbohydrates per day, how many servings of each item should he consume?
b) If the hermit wants to take in P grams of protein and C grams of carbohydrates per day, how many servings of each item should he consume?



Lecture 5: Gauss-Jordan elimination

We have seen in the last lecture that a system of linear equations like

$$\begin{cases} x + y + z = 3 \\ x - y - z = 5 \\ x + 2y - 5z = 9 \end{cases}$$

can be written in matrix form as $A\vec{x} = \vec{b}$, where A is a **matrix** called **coefficient matrix** and **column vectors** \vec{x} and \vec{b} .

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 2 & -5 \end{bmatrix}, \vec{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \vec{b} = \begin{bmatrix} 3 \\ 5 \\ 9 \end{bmatrix}.$$

The i 'th entry $(A\vec{x})_i$ is the dot product of the i 'th row of A with \vec{x} .

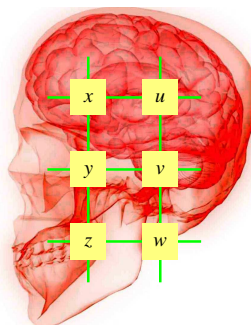
The **augmented matrix** is matrix, where other column has been added. This column contains the vector b . The last column is often separated with horizontal lines for clarity reasons.

$$B = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 1 & -1 & -1 & 5 \\ 1 & 2 & -5 & 9 \end{array} \right]. \quad \text{We will solve this equation using Gauss-Jordan}$$

elimination steps.

1 We aim is to find all the solutions to the system of linear equations

$$\begin{cases} x & & & + & u & & & = & 3 \\ & y & & & & + & v & & = & 5 \\ & & z & & & & + & w & = & 9 \\ x & + & y & + & z & & & & = & 8 \\ & & & u & + & v & + & w & = & 9 \end{cases}$$



This system appears in **tomography** like magnetic resonance

imaging. In this technology, a scanner can measure averages of tissue densities along lines.

The task is to compute the actual densities. We first write down the augmented matrix is

$$\left[\begin{array}{cccccc|c} 1 & 0 & 0 & 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 1 & 0 & 0 & 1 & 9 \\ 1 & 1 & 1 & 0 & 0 & 0 & 8 \\ 0 & 0 & 0 & 1 & 1 & 1 & 9 \end{array} \right].$$

Remove the sum of the first three rows from the 4th, then change sign of the 4'th row:

$$\left[\begin{array}{cccccc|c} 1 & 0 & 0 & 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 1 & 0 & 0 & 1 & 9 \\ 0 & 0 & 0 & 1 & 1 & 1 & 9 \\ 0 & 0 & 0 & 1 & 1 & 1 & 9 \end{array} \right].$$

Now subtract the 4th row from the last to get a row of zeros, then subtract the 4th row from the first. This is already the row reduced echelon form.

$$\left[\begin{array}{cccccc|c} 1 & 0 & 0 & 0 & -1 & -1 & -6 \\ 0 & 1 & 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 1 & 0 & 0 & 1 & 9 \\ 0 & 0 & 0 & 1 & 1 & 1 & 9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right].$$

The first 4 columns have leading 1. The other 2 variables are free variables r, s . We write the row reduced echelon form again as a system and get so the solution:

$$\begin{aligned} x &= -6 + r + s \\ y &= 5 - r \\ z &= 9 - s \\ u &= 9 - r - s \\ v &= r \\ w &= s \end{aligned}$$

There are infinitely many solutions. They are parametrized by 2 free variables.

Gauss-Jordan Elimination is a process, where successive subtraction of multiples of other rows or scaling or swapping operations brings the matrix into **reduced row echelon form**. The elimination process consists of three possible steps. They are called **elementary row operations**:

- Swap two rows.
- scale a row.
- subtract a multiple of a row from an other.

The process transfers a given matrix A into a new matrix $\text{rref}(A)$.

The first nonzero element in a row is called a **leading one**. The goal of the Gauss Jordan elimination process is to bring the matrix in a form for which the solution of the equations can be found. Such a matrix is called in **reduced row echelon form**.

- 1) if a row has nonzero entries, then the first nonzero entry is 1.
- 2) if a column contains a leading 1, then the other column entries are 0.
- 3) if a row has a leading 1, then every row above has a leading 1 to the left.

The number of leading 1 in $\text{rref}(A)$ is called the rank of A . It is an integer which we will use later.

A remark to the history: The process appeared already in the Chinese manuscript "Jiuzhang Suanshu" the 'Nine Chapters on the Mathematical art'. The manuscript or textbook appeared around 200 BC in the Han dynasty. The German geodesist **Wilhelm Jordan** (1842-1899) applied the Gauss-Jordan method to finding squared errors to work on surveying. (An other "Jordan", the French Mathematician Camille Jordan (1838-1922) worked on linear algebra topics also (Jordan form) and is often mistakenly credited with the Gauss-Jordan process.) **Gauss** developed Gaussian elimination around 1800 and used it to solve least squares problems in celestial mechanics and later in geodesic computations. In 1809, Gauss published the book "Theory of Motion of the Heavenly Bodies" in which he used the method for solving astronomical problems.

- 2 Find the rank of the following matrix:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \\ 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 \\ 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 \\ 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 \\ 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 \\ 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 \end{bmatrix}$$

- 3 More challenging is the question: what is the rank of the following matrix?

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 3 & 4 & 5 & 6 & 7 & 8 & 1 \\ 3 & 4 & 5 & 6 & 7 & 8 & 1 & 2 \\ 4 & 5 & 6 & 7 & 8 & 1 & 2 & 3 \\ 5 & 6 & 7 & 8 & 1 & 2 & 3 & 4 \\ 6 & 7 & 8 & 1 & 2 & 3 & 4 & 5 \\ 7 & 8 & 1 & 2 & 3 & 4 & 5 & 6 \\ 8 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{bmatrix}$$

- 4 What are the possible ranks for a 7×11 matrix?

Homework due February 10, 2011

- 1 Problem 10 In 1.2 of Bretscher: Find all solutions of the equations with paper and pencil using Gauss-Jordan elimination. Show all your work.

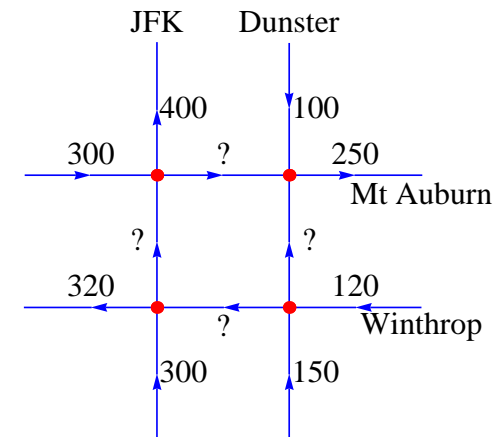
$$\begin{aligned} 4x_1 + 3x_2 + 2x_3 - x_4 &= 4 \\ 5x_1 + 4x_2 + 3x_3 - x_4 &= 4 \\ -2x_1 - 2x_2 - x_3 + 2x_4 &= -3 \\ 11x_1 + 6x_2 + 4x_3 + x_4 &= 11 \end{aligned}$$

- 2 Problem 20 in 1.2 of Bretscher: We say that two $n \times m$ matrices in reduced row-echelon form are of the same type if they contain the same number of leading 1's in the same positions. For example,

$$\begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

are of the same type. How many types of 2×2 matrices in reduced row-echelon form are there?

- 3 Problem 42 in 1.2 of Bretscher: The accompanying sketch represents a maze of one way streets in a city in the United States. The traffic volume through certain blocks during an hour has been measured. Suppose that the vehicles leaving the area during this hour were exactly the same as those entering it. What can you say about the traffic volume at the four locations indicated by a question mark? Can you figure out exactly how much traffic there was on each block? If not, describe one possible scenario. For each of the four locations, find the highest and the lowest possible traffic volume.



Lecture 6: The structure of solutions

Last time we have learned how to row reduce a **matrix**

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}.$$

And bring it to so called **row reduced echelon form**. We write $\text{rref}(A)$. In this form, we have in each row with nonzero entries a so called **leading one**.

The number of leading ones in $\text{rref}(A)$ is called the **rank** of the matrix.

- 1 If we have a unique solution to $A\vec{x} = \vec{b}$, then $\text{rref}(A)$ is the matrix which has a leading 1 in every column. This matrix is called the **identity matrix**.

A matrix with one column is also called a **column vector**. The entries of a matrix are denoted by a_{ij} , where i is the row number and j is the column number.

There are two ways how we can look a system of linear equation. It is called the "row picture" or "column picture":

Row picture: each b_i is the dot product of a row vector \vec{w}_i with \vec{x} .

$$A\vec{x} = \begin{bmatrix} -\vec{w}_1- \\ -\vec{w}_2- \\ \cdots \\ -\vec{w}_n- \end{bmatrix} \begin{bmatrix} | \\ \vec{x} \\ | \end{bmatrix} = \begin{bmatrix} \vec{w}_1 \cdot \vec{x} \\ \vec{w}_2 \cdot \vec{x} \\ \cdots \\ \vec{w}_n \cdot \vec{x} \end{bmatrix}$$

Column picture: \vec{b} is a sum of scaled column vectors \vec{v}_j .

$$A\vec{x} = \begin{bmatrix} | & | & \cdots & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_m \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} = x_1\vec{v}_1 + x_2\vec{v}_2 + \cdots + x_m\vec{v}_m = \vec{b}.$$

- 2 The system of linear equations

$$\begin{cases} 3x - 4y - 5z = 0 \\ -x + 2y - z = 0 \\ -x - y + 3z = 9 \end{cases}$$

is equivalent to $A\vec{x} = \vec{b}$, where A is a **coefficient matrix** and \vec{x} and \vec{b} are **vectors**.

$$A = \begin{bmatrix} 3 & -4 & -5 \\ -1 & 2 & -1 \\ -1 & -1 & 3 \end{bmatrix}, \vec{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \vec{b} = \begin{bmatrix} 0 \\ 0 \\ 9 \end{bmatrix}.$$

The **augmented matrix** is

$$B = \left[\begin{array}{ccc|c} 3 & -4 & -5 & 0 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & 9 \end{array} \right].$$

In this case, the row vectors of A are

$$\vec{w}_1 = \begin{bmatrix} 3 & -4 & -5 \\ -1 & 2 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

The column vectors are

$$\vec{v}_1 = \begin{bmatrix} 3 \\ -1 \\ -1 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} -4 \\ 2 \\ -1 \end{bmatrix}, \vec{v}_3 = \begin{bmatrix} -5 \\ -1 \\ 3 \end{bmatrix}$$

The **row picture** tells: $0 = b_1 = \begin{bmatrix} 3 & -4 & -5 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix}$

The **column picture** tells:

$$\begin{bmatrix} 0 \\ 0 \\ 9 \end{bmatrix} = x_1 \begin{bmatrix} 3 \\ -1 \\ -1 \end{bmatrix} + x_2 \begin{bmatrix} -4 \\ 2 \\ -1 \end{bmatrix} + x_3 \begin{bmatrix} -5 \\ -1 \\ -1 \end{bmatrix}.$$

A system of linear equations $A\vec{x} = \vec{b}$ with n equations and m unknowns is defined by the $n \times m$ matrix A and the vector \vec{b} . The row reduced matrix $\text{rref}(B)$ of the augmented matrix $B = [A|\vec{b}]$ determines the number of solutions of the system $Ax = b$. The **rank** $\text{rank}(A)$ of a matrix A is the number of leading ones in $\text{rref}(A)$.

Theorem. For any system of linear equations there are three possibilities:

- **Consistent with unique solution:** Exactly one solution. There is a leading 1 in each column of A but none in the last column of the augmented matrix B .
- **Inconsistent with no solutions.** There is a leading 1 in the last column of the augmented matrix B .
- **Consistent with infinitely many solutions.** There are columns of A without leading 1.

How do we determine in which case we are? It is the rank of A and the rank of the augmented matrix $B = [A|\vec{b}]$ as well as the number m of columns which determine everything:

If $\text{rank}(A) = \text{rank}(B) = m$: there is **exactly 1 solution**.

If $\text{rank}(A) < \text{rank}(B)$: there are **no solutions**.

If $\text{rank}(A) = \text{rank}(B) < m$: there are ∞ **many solutions**.

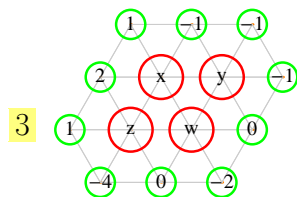
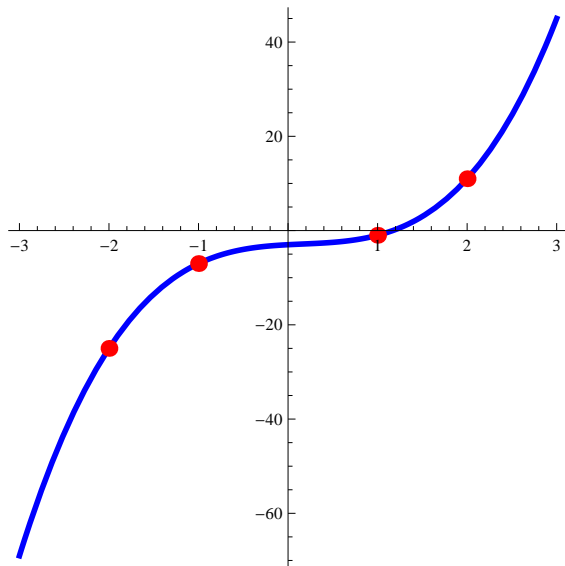
If we have n equations and n unknowns, it is most likely that we have exactly one solution. But remember Murphy's law "If anything can go wrong, it will go wrong". It happens!

- 3 What is the probability that we have exactly one solution if we look at all $n \times n$ matrices with entries 1 or 0? You explore this in the homework in the 2×2 case. During the lecture we look at the 3×3 case and higher, using a **Monte Carlo simulation**.

Homework due February 10, 2011

- 1 We look at the probability space of all 2×2 matrices with matrix entries 0 or 1.
- What is the probability that the rank of the matrix is 1?
 - What is the probability that the rank of the matrix is 0?
 - What is the probability that the rank of the matrix is 2?
- 2 Find a cubic equation $f(x) = ax^3 + bx^2 + cx + d = y$ such that the graph of f goes through the 4 points

$$A = (-1, -7), B = (1, -1), C = (2, 11), D = (-2, -25).$$



In a Herb garden, the humidity of its soil has the property that at any given point the humidity is the sum of the neighboring humidities. Samples are taken on a hexagonal grid on 14 spots. The humidity at the four locations x, y, z, w is unknown. Solve the equations

$$\begin{cases} x = y+z+w+2 \\ y = x+w-3 \\ z = x+w-1 \\ w = x+y+z-2 \end{cases} \text{ using row reduction.}$$

Lecture 7: Linear transformations

A **transformation** T from a set X to a set Y is a rule, which assigns to every x in X an element $y = T(x)$ in Y . One calls X the **domain** and Y the **codomain**. A transformation is also called a **map** from X to Y . A map T from \mathbf{R}^m to \mathbf{R}^n is called a **linear transformation** if there is a $n \times m$ matrix A such that $T(\vec{x}) = A\vec{x}$.

- To the linear transformation $T(x, y) = (3x + 4y, x + 5y)$ belongs the matrix $\begin{bmatrix} 3 & 4 \\ 1 & 5 \end{bmatrix}$. This transformation maps the two-dimensional plane onto itself.
- $T(x) = -33x$ is a linear transformation from the real line onto itself. The matrix is $A = [-33]$.
- To $T(\vec{x}) = \vec{y} \cdot \vec{x}$ from \mathbf{R}^3 to \mathbf{R} belongs the matrix $A = \vec{y} = \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}$. This 1×3 matrix is also called a **row vector**. If the codomain is the real axes, one calls the map also a **function**.
- $T(x) = x\vec{y}$ from \mathbf{R} to \mathbf{R}^3 . $A = \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$ is a 3×1 matrix which is also called a **column vector**. The map defines a line in space.
- $T(x, y, z) = (x, y)$ from \mathbf{R}^3 to \mathbf{R}^2 , A is the 2×3 matrix $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$. The map projects space onto a plane.
- To the map $T(x, y) = (x + y, x - y, 2x - 3y)$ belongs the 3×2 matrix $A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & -1 & -3 \end{bmatrix}$. The image of the map is a plane in three dimensional space.
- If $T(\vec{x}) = \vec{x}$, then T is called the **identity transformation**.

A transformation T is linear if and only if the following properties are satisfied:
 $T(\vec{0}) = \vec{0}$ $T(\vec{x} + \vec{y}) = T(\vec{x}) + T(\vec{y})$ $T(\lambda\vec{x}) = \lambda T(\vec{x})$

In other words, linear transformations are compatible with addition, scalar multiplication and also respect 0. It does not matter, whether we add two vectors before the transformation or add the transformed vectors.

Linear transformations are important in

- geometry (i.e. rotations, dilations, projections or reflections)
- art (i.e. perspective, coordinate transformations),
- computer aided design applications (i.e. projections),

- physics (i.e. Lorentz transformations),
- dynamics (linearizations of general maps are linear maps),
- compression (i.e. using Fourier transform or Wavelet transform),
- error correcting codes (many codes are linear codes),
- probability theory (i.e. Markov processes).
- linear equations (inversion is solving the equation)

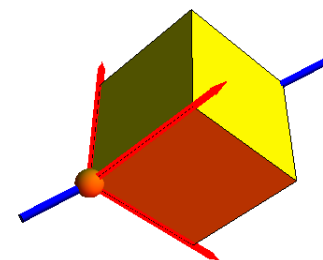
A linear transformation $T(x) = Ax$ with $A = \begin{bmatrix} | & | & \cdots & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & \cdots & | \end{bmatrix}$ has the property that the

column vector $\vec{v}_1, \vec{v}_2, \vec{v}_n$ are the images of the **standard vectors** $\vec{e}_1 = \begin{bmatrix} 1 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$, $\vec{e}_i = \begin{bmatrix} 0 \\ \cdot \\ 1 \\ \cdot \end{bmatrix}$, and

$$\vec{e}_n = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ 1 \end{bmatrix}.$$

In order to find the matrix of a linear transformation, look at the image of the standard vectors and use those to build the columns of the matrix.

- Find the matrix belonging to the linear transformation, which rotates a cube around the diagonal $(1, 1, 1)$ by 120 degrees ($2\pi/3$).



- Find the linear transformation, which reflects a vector at the line containing the vector $(1, 1, 1)$.

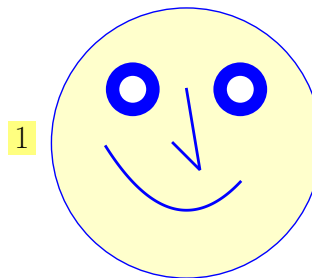
If there is a linear transformation S such that $S(T\vec{x}) = \vec{x}$ for every \vec{x} , then S is called the **inverse** of T . We will discuss inverse transformations later in more detail.

$A\vec{x} = \vec{b}$ means to invert the linear transformation $\vec{x} \mapsto A\vec{x}$. If the linear system has exactly one solution, then an inverse exists. We will write $\vec{x} = A^{-1}\vec{b}$ and see that the inverse of a linear transformation is again a linear transformation.

- 3 Otto Bretscher's book contains as a motivation a "code", where the encryption happens with the linear map $T(x, y) = (x + 3y, 2x + 5y)$. It is an variant of a Hill code. The map has the inverse $T^{-1}(x, y) = (-5x + 3y, 2x - y)$. Assume we know, the other party uses a Bretscher code and can find out that $T(1, 1) = (3, 5)$ and $T(2, 1) = (7, 5)$. Can we reconstruct the code? The problem is to find the matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. It is useful to decode the Hill code in general. If $ax + by = X$ and $cx + dy = Y$, then $x = (dX - bY)/(ad - bc)$, $y = (cX - aY)/(ad - bc)$. This is a linear transformation with matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and the corresponding matrix is $A^{-1} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} / (ad - bc)$.

"Switch diagonally, negate the wings and scale with a cross".

Homework due February 16, 2011



This is Problem 24-40 in Bretscher: Consider the circular face in the accompanying figure. For each of the matrices A_1, \dots, A_6 , draw a sketch showing the effect of the linear transformation $T(x) = Ax$ on this face.

$$A_1 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad A_5 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix},$$

$$A_6 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

- 2 This is problem 50 in Bretscher. A goldsmith uses a platinum alloy and a silver alloy to make jewelry; the densities of these alloys are exactly 20 and 10 grams per cubic centimeter, respectively.
- a) King Hiero of Syracuse orders a crown from this goldsmith, with a total mass of 5 kilograms (or 5,000 grams), with the stipulation that the platinum alloy must make up at least 90% of the mass. The goldsmith delivers a beautiful piece, but the king's friend Archimedes has doubts about its purity. While taking a bath, he comes up with a method to check the composition of the crown (famously shouting "Eureka!" in the process, and running to the king's palace naked). Submerging the crown in water, he finds its volume to be 370 cubic centimeters. How much of each alloy went into this piece (by mass)? Is this goldsmith a crook?

b) Find the matrix A that transforms the vector

$$\begin{bmatrix} \text{mass of platinum alloy} \\ \text{mass of silver alloy} \end{bmatrix}$$

into the vector

$$\begin{bmatrix} \text{total mass} \\ \text{total volume} \end{bmatrix}$$

for any piece of jewelry this goldsmith makes.

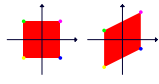
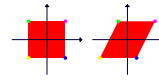
c) Is the matrix A in part (b) invertible? If so, find its inverse. Use the result to check your answer in part a)

- 3 In the first week we have seen how to compute the mean and standard deviation of data.
- a) Given some data $(x_1, x_2, x_3, \dots, x_6)$. Is the transformation from $\mathbf{R}^6 \rightarrow \mathbf{R}$ which maps the data to its mean m linear?
- b) Is the map which assigns to the data the standard deviation σ a linear map? c) Is the map which assigns to the data the difference (y_1, y_2, \dots, y_6) defined by $y_1 = x_1, y_2 = x_2 - x_1, \dots, y_6 = x_6 - x_5$ linear? Find its matrix. d) Is the map which assigns to the data the normalized data $(x_1 - m, x_2 - m, \dots, x_n - m)$ given by a linear transformation?

Lecture 8: Examples of linear transformations

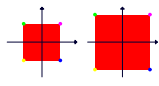
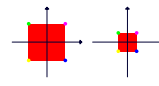
While the space of linear transformations is large, there are few types of transformations which are typical. We look here at dilations, shears, rotations, reflections and projections.

Shear transformations

1 $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$  $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ 

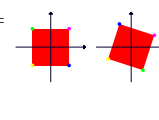
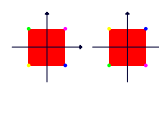
In general, shears are transformation in the plane with the property that there is a vector \vec{w} such that $T(\vec{w}) = \vec{w}$ and $T(\vec{x}) - \vec{x}$ is a multiple of \vec{w} for all \vec{x} . Shear transformations are invertible, and are important in general because they are examples which can not be diagonalized.

Scaling transformations

2 $A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$  $A = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$ 

One can also look at transformations which scale x differently than y and where A is a diagonal matrix. Scaling transformations can also be written as $A = \lambda I_2$ where I_2 is the identity matrix. They are also called **dilations**.

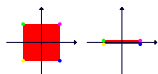
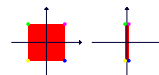
Reflection

3 $A = \begin{bmatrix} \cos(2\alpha) & \sin(2\alpha) \\ \sin(2\alpha) & -\cos(2\alpha) \end{bmatrix}$  $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ 

Any reflection at a line has the form of the matrix to the left. A reflection at a line containing a unit vector \vec{u} is $T(\vec{x}) = 2(\vec{x} \cdot \vec{u})\vec{u} - \vec{x}$ with matrix $A = \begin{bmatrix} 2u_1^2 - 1 & 2u_1u_2 \\ 2u_1u_2 & 2u_2^2 - 1 \end{bmatrix}$

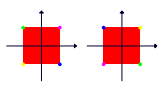
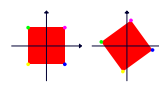
Reflections have the property that they are their own inverse. If we combine a reflection with a dilation, we get a **reflection-dilation**.

Projection

$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$  $A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ 

- 4 A projection onto a line containing unit vector \vec{u} is $T(\vec{x}) = (\vec{x} \cdot \vec{u})\vec{u}$ with matrix $A = \begin{bmatrix} u_1u_1 & u_2u_1 \\ u_1u_2 & u_2u_2 \end{bmatrix}$. Projections are also important in statistics. Projections are not invertible except if we project onto the entire space. Projections also have the property that $P^2 = P$. If we do it twice, it is the same transformation. If we combine a projection with a dilation, we get a **rotation dilation**.

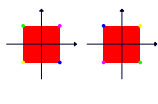
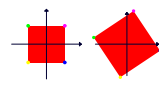
Rotation

5 $A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$  $A = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}$ 

Any rotation has the form of the matrix to the right.

Rotations are examples of orthogonal transformations. If we combine a rotation with a dilation, we get a **rotation-dilation**.

Rotation-Dilation

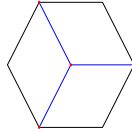
6 $A = \begin{bmatrix} 2 & -3 \\ 3 & 2 \end{bmatrix}$  $A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ 

A rotation dilation is a composition of a rotation by angle $\arctan(y/x)$ and a dilation by a factor $\sqrt{x^2 + y^2}$.

If $z = x + iy$ and $w = a + ib$ and $T(x, y) = (X, Y)$, then $X + iY = zw$. So a rotation dilation is tied to the process of the multiplication with a complex number.

Rotations in space

- 7 Rotations in space are determined by an axis of rotation and an angle. A rotation by 120° around a line containing $(0,0,0)$ and $(1,1,1)$ belongs to $A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ which permutes $\vec{e}_1 \rightarrow \vec{e}_2 \rightarrow \vec{e}_3$.



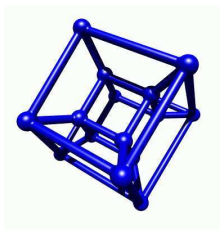
Reflection at xy-plane

- 8 To a reflection at the xy -plane belongs the matrix $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ as can be seen by looking at the images of \vec{e}_i . The picture to the right shows the linear algebra textbook reflected at two different mirrors.



Projection into space

- 9 To project a 4d-object into the three dimensional xyz-space, use for example the matrix $A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$. The picture shows the projection of the four dimensional cube (tesseract, hypercube) with 16 edges $(\pm 1, \pm 1, \pm 1, \pm 1)$. The tesseract is the theme of the horror movie "hypercube".



Homework due February 16, 2011

- 1 What transformation in space do you get if you reflect first at the xy -plane, then rotate around the z axes by 90 degrees (counterclockwise when watching in the direction of the z -axes), and finally reflect at the x axes?
- 2 a) One of the following matrices can be composed with a dilation to become an orthogonal projection onto a line. Which one?

$$A = \begin{bmatrix} 3 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 \end{bmatrix} \quad B = \begin{bmatrix} 3 & 1 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 1 & 3 \end{bmatrix} \quad C = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

$$D = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad E = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad F = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

- b) The **smiley face** visible to the right is transformed with various linear transformations represented by matrices $A - F$. Find out which matrix does which transformation:

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

$$D = \begin{bmatrix} 1 & -1 \\ 0 & -1 \end{bmatrix}, \quad E = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad F = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} / 2$$



| A-F | image | A-F | image | A-F | image |
|-----|-------|-----|-------|-----|-------|
| | | | | | |
| | | | | | |

- 3 This is homework 28 in Bretscher 2.2: Each of the linear transformations in parts (a) through (e) corresponds to one and only one of the matrices A) through J). Match them up.

a) Scaling b) Shear c) Rotation d) Orthogonal Projection e) Reflection

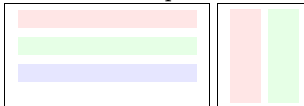
$$A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} \quad C = \begin{bmatrix} -0.6 & 0.8 \\ 0.8 & -0.6 \end{bmatrix} \quad D = \begin{bmatrix} 7 & 0 \\ 0 & 7 \end{bmatrix} \quad E = \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix}$$

$$F = \begin{bmatrix} 0.6 & 0.8 \\ 0.8 & -0.6 \end{bmatrix} \quad G = \begin{bmatrix} 0.6 & 0.6 \\ 0.8 & 0.8 \end{bmatrix} \quad H = \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} \quad I = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \quad J = \begin{bmatrix} 0.8 & -0.6 \\ 0.6 & -0.8 \end{bmatrix}$$

Lecture 9: Matrix algebra

If A is a $n \times m$ matrix and B is a $m \times p$ matrix, then the **matrix product** AB is defined as the

$n \times p$ matrix with entries $(BA)_{ij} = \sum_{k=1}^m B_{ik}A_{kj}$.



1 If B is a 3×4 matrix, and A is a 4×2 matrix then BA is a 3×2 matrix. For example:

$$B = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 3 & 1 & 8 & 1 \\ 1 & 0 & 9 & 2 \end{bmatrix}, A = \begin{bmatrix} 1 & 3 \\ 3 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, BA = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 3 & 1 & 8 & 1 \\ 1 & 0 & 9 & 2 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 3 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 15 & 13 \\ 14 & 11 \\ 10 & 5 \end{bmatrix}.$$

If A is a $n \times n$ matrix and $T : \vec{x} \mapsto A\vec{x}$ has an inverse S , then S is linear. The matrix A^{-1} belonging to $S = T^{-1}$ is called the **inverse matrix** of A .

Matrix multiplication generalizes the common multiplication of numbers. We can write the dot product between two vectors as a matrix product when writing the first vector as a $1 \times n$ matrix (= row vector) and the second as a $n \times 1$ matrix (=column vector) like in

$$[1 \ 2 \ 3] \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} = 20. \text{ Note that } AB \neq BA \text{ in general and for } n \times n \text{ matrices, the inverse } A^{-1}$$

does not always exist, otherwise, for $n \times n$ matrices the same rules apply as for numbers: $A(BC) = (AB)C$, $AA^{-1} = A^{-1}A = 1_n$, $(AB)^{-1} = B^{-1}A^{-1}$, $A(B+C) = AB+AC$, $(B+C)A = BA+CA$ etc.

2 The entries of matrices can themselves be matrices. If B is a $n \times p$ matrix and A is a $p \times m$ matrix, and assume the entries are $k \times k$ matrices, then BA is a $n \times m$ matrix, where each entry $(BA)_{ij} = \sum_{l=1}^p B_{il}A_{lj}$ is a $k \times k$ matrix. Partitioning matrices can be useful to improve the speed of matrix multiplication. If $A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$, where A_{ij} are $k \times k$ matrices with the property that A_{11} and A_{22} are invertible, then one can write the inverse as $B = \begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22}^{-1} \\ 0 & A_{22}^{-1} \end{bmatrix}$ is the inverse of A .

3 Let us associate to a small blogging network a matrix $\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$ and look at the spread

of some news. Assume the source of the news about some politician is the first entry (maybe the gossip news "gawker") $\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$. The vector $A\vec{x}$ has a 1 at the places, where the news could be in the next hour. The vector $(AA)(\vec{x})$ tells, in how many ways the news can go in

2 steps. In our case, it can go in three different ways back to the page itself.

Matrices help to solve combinatorial problems. One appears in the movie "Good will hunting". For example, what does $[A^{100}]$ tell about the news distribution on a large network. What does it mean if A^{100} has no zero entries?

If A is a $n \times n$ matrix and the system of linear equations $A\vec{x} = \vec{y}$ has a unique solution for all \vec{y} , we write $\vec{x} = A^{-1}\vec{y}$. The inverse matrix can be computed using Gauss-Jordan elimination. Lets see how this works.

Let 1_n be the $n \times n$ identity matrix. Start with $[A|1_n]$ and perform Gauss-Jordan elimination. Then

$$\text{rref}([A|1_n]) = [1_n|A^{-1}]$$

Proof. The elimination process solves $A\vec{x} = \vec{e}_i$ simultaneously. This leads to solutions \vec{v}_i which are the columns of the inverse matrix A^{-1} because $A^{-1}\vec{e}_i = \vec{v}_i$.

$$\begin{array}{cc} \left[\begin{array}{cc|cc} 2 & 6 & 1 & 0 \\ 1 & 4 & 0 & 1 \end{array} \right] & \left[\begin{array}{c|c} A & 1_2 \end{array} \right] \\ \left[\begin{array}{cc|cc} 1 & 3 & 1/2 & 0 \\ 1 & 4 & 0 & 1 \end{array} \right] & \left[\begin{array}{c|c} \dots & \dots \end{array} \right] \\ \left[\begin{array}{cc|cc} 1 & 3 & 1/2 & 0 \\ 0 & 1 & -1/2 & 1 \end{array} \right] & \left[\begin{array}{c|c} \dots & \dots \end{array} \right] \\ \left[\begin{array}{cc|cc} 1 & 0 & 2 & -3 \\ 0 & 1 & -1/2 & 1 \end{array} \right] & \left[\begin{array}{c|c} 1_2 & A^{-1} \end{array} \right] \end{array}$$

$$\text{The inverse is } A^{-1} = \begin{bmatrix} 2 & -3 \\ -1/2 & 1 \end{bmatrix}.$$

If $ad - bc \neq 0$, the inverse of a linear transformation $\vec{x} \mapsto A\vec{x}$ with $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is given by

$$\text{the matrix } A^{-1} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} / (ad - bc).$$

Shear:

$$A = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

Diagonal:

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \end{bmatrix}$$

Reflection:

$$A = \begin{bmatrix} \cos(2\alpha) & \sin(2\alpha) \\ \sin(2\alpha) & -\cos(2\alpha) \end{bmatrix}$$

$$A^{-1} = A = \begin{bmatrix} \cos(2\alpha) & \sin(2\alpha) \\ \sin(2\alpha) & -\cos(2\alpha) \end{bmatrix}$$

Rotation:

$$A = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}$$

Rotation=Dilation:

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} a/r^2 & b/r^2 \\ -b/r^2 & a/r^2 \end{bmatrix}, r^2 = a^2 + b^2$$

Homework due February 16, 2011

- 1 Find the inverse of the following matrix

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

- 2 The probability density of a multivariate normal distribution centered at the origin is a multiple of

$$f(x) = \exp(-x \cdot A^{-1}x)$$

We will see the covariance matrix later. It encodes how the different coordinates of a random vector are correlated. Assume

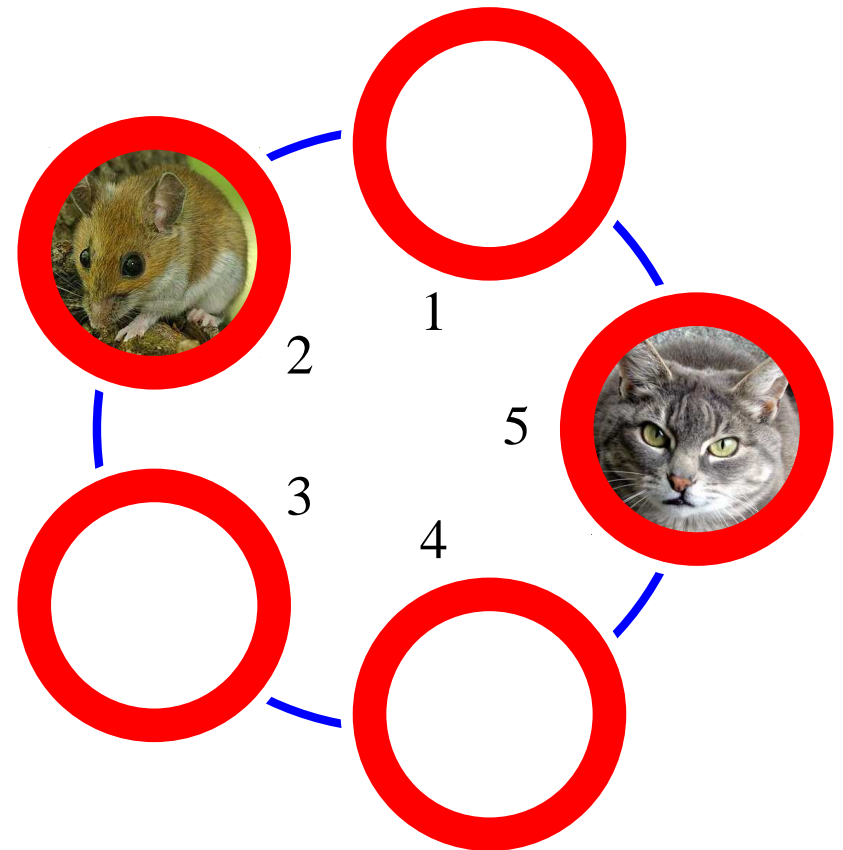
$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Find $f([1, 2, -3])$.

- 3 This is a system we will analyze more later. For now we are only interested in the algebra. **Tom** the cat moves each minute randomly from on spots 1,5,4 jumping to neighboring sites only. At the same time **Jerry**, the mouse, moves on spots 1,2,3, also jumping to neighboring sites. The possible position combinations (2, 5), (3, 4), (3, 1), (1, 4), (1, 1) and transitions are encoded in a matrix

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 1 \end{bmatrix}$$

This means for example that we can go from state (2,5) with equal probability to all the other states. In state (3,4) or (3,1) the pair (Tom,Jerry) moves back to (2,5). If state (1,1) is reached, then Jerry's life ends and Tom goes to sleep there. We can read off that the probability that Jerry gets eaten in one step as $1/4$. Compute A^4 . The first column of this matrix gives the probability distribution after four steps. The last entry of the first column gives the probability that Jerry got swallowed after 4 steps.



Lecture 10: Random variables

In this lecture, we define random variables, the expectation, mean and standard deviation.

A **random variable** is a function X from the probability space to the real line with the property that for every interval the set $\{X \in [a, b]\}$ is an event.

There is nothing complicated about random variables. They are just functions on the laboratory Ω . The reason for the difficulty in understanding random variables is solely due to the name "variable". It is not a variable we solve for. It is just a function. It quantifies properties of experiments. In any applications, the sets $X \in [a, b]$ are automatically events. The last condition in the definition is something we **do not have to worry about in general**.

If our probability space is finite, all subsets are events. In that case, **any** function on Ω is a random variable. In the case of continuous probability spaces like intervals, any piecewise continuous function is a random variable. In general, any function which can be constructed with a sequence of operations is a random variable.

- 1 We throw two dice and assign to each experiment the sum of the eyes when rolling two dice. For example $X[(1,2)] = 3$ or $X[(4,5)] = 9$. This random variable takes values in the set $\{2, 3, 4, \dots, 12\}$.

Given a random variable X , we can look at probabilities like $P[\{X = 3\}]$. We usually leave out the brackets and abbreviate this as $P[X = 3]$. It is read as "the probability that $X = 3$."

- 2 Assume Ω is the set of all 10 letter sequences made of the four nucleotides G, C, A, T in a string of DNA. An example is $\omega = (G, C, A, T, T, A, G, G, C, T)$. Define $X(\omega)$ as the number of Guanine basis elements. In the particular sample ω just given, we have $X(\omega) = 3$.

Problem Assume $X(\omega)$ is the number of Guanine basis elements in a sequence. What is the probability of the event $\{X(\omega) = 2\}$? **Answer** Our probability space has $4^{10} = 1048576$ elements. There are 3^8 cases, where the first two elements are G . There are 3^8 elements where the first and third element is G , etc. For any pair, there are 3^8 sequences. We have $(10 \cdot 9/2) = 45$ possible ways to choose a pair from the 10. There are therefore $3^8 \cdot 45$ sequences with exactly 2 amino acids G . This is the cardinality of the event $A = \{X(\omega) = 2\}$. The probability is $|A|/|\Omega| = 45 \cdot 3^8/4^{10}$ which is about 0.28.

For random variables taking finitely many values we can look at the probabilities $p_j = P[X = c_j]$. This collection of numbers is called a **discrete probability distribution** of the random variable.

- 3 We throw a dice 10 times and call $X(\omega)$ the number of times that "heads" shows up. We have $P[X = k] = \binom{10}{k} (1/2)^{10}$, because we chose k elements from $n = 10$. This distribution is called the **Binominal distribution** on the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

- 4 In the 10 nucleotide example, where X counts the number of G nucleotides, we have

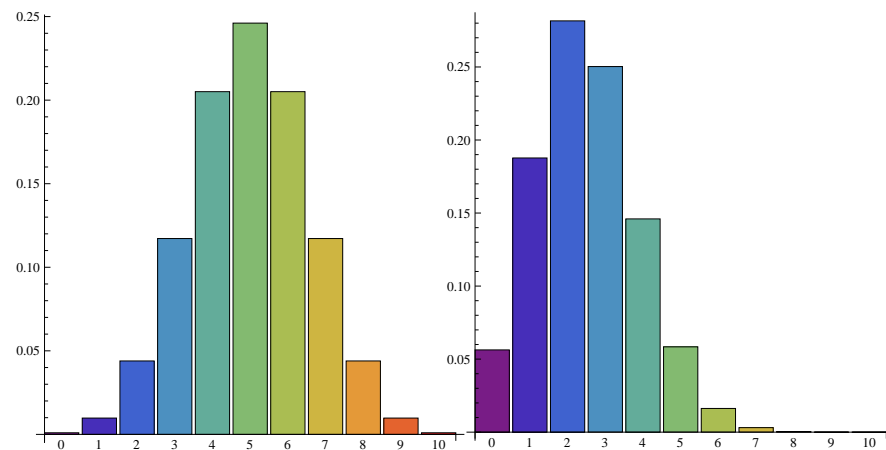
$$P[X = k] = \binom{10}{k} \frac{3^{n-k}}{4^n}.$$

We can write this as $\binom{10}{k} p^k (1-p)^{n-k}$ with $p = 1/4$ and interpret it as having "heads" turn up k times if it appears with probability p and "tails" with probability $1-p$.

If $X(k)$ counts the number of 1 in a sequence of length n and each 1 occurs with a probability p , then

$$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

This **Binomial distribution** is an extremely important example of a probability distribution.



The Binominal distribution with $p = 1/2$. The Binominal distribution with $p = 1/4$.

For a random variable X taking finitely many values, we define the **expectation** as $m = E[X] = \sum_x x P[X = x]$. Define the **variance** as $\text{Var}[X] = E[(X - m)^2] = E[X^2] - E[X]^2$ and the **standard deviation** as $\sigma[X] = \sqrt{\text{Var}[X]}$.

- 5 In the case of throwing a coin 10 times and head appears with probability $p = 1/2$ we have

$$E[X] = 0 \cdot P[X = 0] + 1 \cdot P[X = 1] + 2 \cdot P[X = 2] + 3 \cdot P[X = 3] + \dots + 10 \cdot P[X = 10].$$

The average adds up to $10 \times p = 5$, which is what we expect. We will see next time when we discuss independence, how we can get this immediately. The variance is

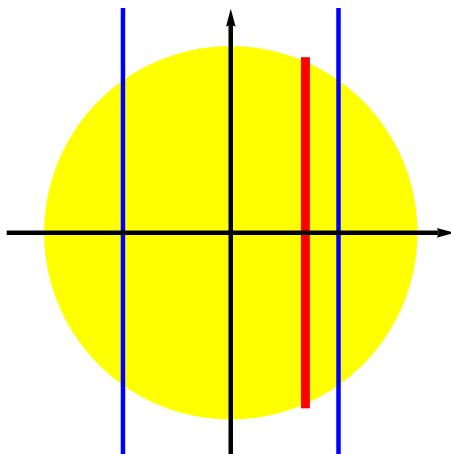
$$\text{Var}[X] = (0 - 5)^2 \cdot P[X = 0] + (1 - 5)^2 \cdot P[X = 1] + \dots + (10 - 5)^2 \cdot P[X = 10].$$

It is $10p(1-p) = 10/4$. Again, we will have to wait until next lecture to see how we can get this without counting.

All these examples so far, the random variable has taken only a discrete set of values. Here is an example, where the random variable can take values in an interval. It is called a variable with a **continuous distribution**.

- 6 Throw a vertical line randomly into the unit disc. Let $X[\omega]$ be the length of the segment cut out from the circle. What is $P[X > 1]$?

Solution: we need to hit the x axes in $|x| < 1/\sqrt{3}$. Comparing lengths gives the probability is $1/\sqrt{3}$. We have assumed here that every interval $[c, d]$ in the interval $[-1, 1]$ appears with probability $(d - c)/2$.

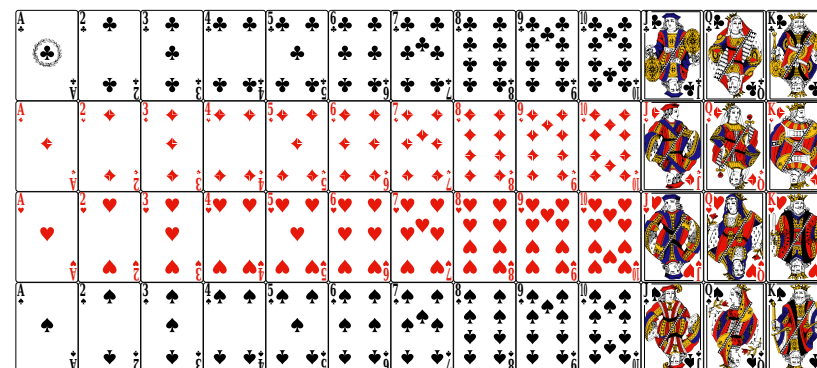


If a random variable has the property that $P[X \in [a, b]] = \int_a^b f(x) dx$ where f is a nonnegative function satisfying $\int_{-\infty}^{\infty} f(x) dx = 1$. Then the expectation of X is defined as $E[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$. The function f is called the **probability density function** and we will talk about it later in the course.

In the previous example, we have seen again the Bertrand example, but because we insisted on vertical sticks, the probability density was determined. The other two cases we have seen produced different probability densities. A probability model always needs a probability function P .

Homework due February 23, 2011

- 1 In the card game **blackjack**, each of the 52 cards is assigned a value. You see the French card deck below in the picture. **Numbered cards 2-10** have their natural value, the **picture cards jack, queen, and king** count as 10, and aces are valued as either 1 or 11. Draw the probability distribution of the random variable X which gives the value of the card assuming that we assign to the hearts ace and diamond aces the value 1 and to the club ace and spades ace the value 11. Find the mean the variance and the standard deviation of the random variable X .

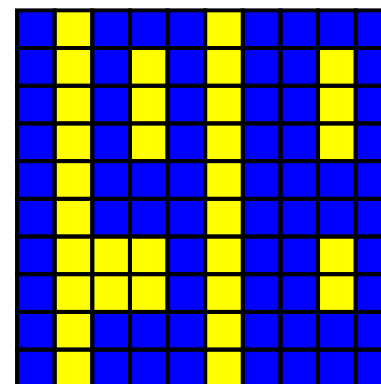


- 2 We look at the probability space of all 2×2 matrices, where the entries are either 1 or -1 . Define the random variable $X(\omega) = \det(\omega)$, where ω is one of the matrices. The determinant is

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc.$$

Draw the probability distribution of this random variable and find the expectation as well as the Variance and standard deviation.

- 3 A LCD display with 100 pixels is described by a 10×10 matrix with entries 0 and 1. Assume, that each of the pixels fails independently with probability $p = 1/20$ during the year. Define the random variable $X(\omega)$ as the number of dead pixels after a year.
- What is the probability of the event $P[X > 3]$, the probability that more than 3 pixels have died during the year?
 - What is the expected number of pixels failing during the year?



Lecture 11: Independence

If Ω is a finite probability space where each experiment has the same probability, then

$$E[X] = \frac{1}{|\Omega|} \sum_{\omega} X(\omega) \quad (1)$$

is the expectation of the random variable. Last time, we wrote this as

$$E[X] = \sum_{x_j} x_j P[X = x_j], \quad (2)$$

where x_j are the possible values of X . The later expression is the same but involves less terms.

In real life we often do not know the probability space. Or, the probability space is so large that we have no way to enumerate it. The only thing we can access is the distribution, the frequency with which data occur. **Statistics** helps to build a model. Formula (1) is often not computable, but (2) is since we can build a model with that distribution.

1 Lets illustrate this with data

$$X = (1, 2, 3, 3, 4, 1, 1, 1, 2, 6)$$

To compute the expectation of X , write it as the result of a random variable $X(1) = 1, X(2) = 2, X(3) = 3, \dots, X(10) = 6$ on a probability space of 10 elements. In this case, $E[X] = (1 + 2 + 3 + 3 + 4 + 1 + 1 + 1 + 2 + 6)/10 = 24/10$. But we can look at these data also differently and say $P[X = 1] = 4/10, P[X = 2] = P[X = 3] = 2/10, P[X = 4] = P[X = 6] = 1/6$. Now,

$$\begin{aligned} E[X] &= 1 P[X = 1] + 2 P[X = 2] + 3 P[X = 3] + 4 P[X = 4] + 6 P[X = 6] \\ &= 1 \frac{4}{10} + 2 \frac{2}{10} + 3 \frac{2}{10} + 4 \frac{1}{10} + 6 \frac{1}{10} = \frac{12}{5}. \end{aligned}$$

The first expression has 10 terms, the second 5. Not an impressive gain, but look at the next example.

2 We throw 100 coins and let X denote the number of "heads". Formula (1) involves 2^{100} terms. This is too many to sum over. The expression (2) however

$$\sum_{k=1}^{100} k P[X = k] = \sum_{k=1}^{100} k \binom{100}{k} \frac{1}{2^{100}}$$

has only 100 terms and sums up to $100 * (1/2) = 50$ because in general

$$\frac{1}{2^n} \sum_{k=0}^n k \binom{n}{k} = \frac{n}{2}.$$

By the way, one can see this by writing out the factorials $k \binom{n}{k} = n \binom{n-1}{k}$. Summing over the probability space is unmanageable. Even if we would have looked at 10 trillion cases every millisecond since 14 billion years, we would not be through. But this is not an obstacle. Despite the huge probability space, we have a simple model which tells us what the probability is to have k heads.

Two events A, B are called **independent**, if $P[A \cap B] = P[A] \cdot P[B]$.

3 Let Ω be the probability space obtained by throwing two dice. It has 36 elements. Let A be the event that the first dice shows an odd number and let B be the event that the second dice shows less than 3 eyes. The probability of A is $18/36 = 1/2$ the probability of B is $12/36 = 1/3$. The event $A \cap B$ consists of the cases $\{(1, 1), (1, 2), (3, 1), (3, 2), (5, 1), (5, 2)\}$ and has probability $1/6$. The two events are independent.

4 If Ω is the probability space of throwing 5 coins. It has $2^5 = 32$ elements. The event A that the first 4 coins are head and the event B that the last coin is head are uncorrelated: $P[A] = 1/2^4$ and $P[B] = 1/2$. And $P[A \cap B] = 1/2^5$. We might think that if 4 heads have come, "justice" or "fairness" should tilt the chance towards "tails" since in average the same number of heads and tails occur. But this is not the case. The two events are independent. The coin flying the 5 times does not know about the previous cases.

| | | | | | |
|----|----|----|----|----|----|
| 11 | 12 | 13 | 14 | 15 | 16 |
| 21 | 22 | 23 | 24 | 25 | 26 |
| 31 | 32 | 33 | 34 | 35 | 36 |
| 41 | 42 | 43 | 44 | 45 | 46 |
| 51 | 52 | 53 | 54 | 55 | 56 |
| 61 | 62 | 63 | 64 | 65 | 66 |

We can rephrase correlation using conditional probability

If A, B are independent then $P[A|B] = P[A]$. Knowing about B does not change the probability of A .

This follows from the definition $P[A|B] = P[A \cap B]/P[B]$ and $P[A \cap B] = P[A] \cdot P[B]$.

Two **random variables** X, Y are called **independent** if for every x, y , the events $\{X = x\}$ and $\{Y = y\}$ are independent.

5 If Ω is the probability space of throwing two dice. Let X be the random variable which gives the value of the first dice and Y the random variable which gives the value of the second dice. Then $X((a, b)) = a$ and $Y((a, b)) = b$. The events $X = x$ and $Y = y$ are independent because each has probability $1/6$ and event $\{X = x, Y = y\}$ has probability $1/36$.

Two **random variables** X, Y are called **uncorrelated**, if $E[XY] = E[X] \cdot E[Y]$.

6 Let X be the random variable which is 1 on the event A and zero everywhere else. Let Y be the random variable which is 1 on the event B and zero everywhere else. Now $E[X] = 0P[X = 0] + 1P[X = 1] = P[A]$. Similarly $P[Y] = P[B]$. and $P[XY] = P[A \cap B]$ because $XY(\omega) = 1$ only if ω is in A and in B .

7 Let X be the random variable on the probability space of two dice which gives the dice value of the first dice. Let Y be the value of the second dice. These two random variables are uncorrelated.

$$E[XY] = \frac{1}{36} \sum_{i=1}^6 \sum_{j=1}^6 ij = [(1+2+3+4+5+6) \cdot (1+2+3+4+5+6)]/36 = \frac{21^2}{36} = \frac{49}{4}.$$

We also have $E[X] = (1+2+3+4+5+6)/6 = \frac{7}{2}$.

Define the **covariance** of two random variables X, Y as

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

Two random variables are uncorrelated if and only if their correlation is zero.

To see this, just multiply out $E[(X - E[X]) \cdot (Y - E[Y])] = E[XY] - 2E[X]E[Y] + E[X]E[Y] = E[XY] - E[X]E[Y]$.

If two random variables are independent, then they are uncorrelated.

Proof. Let $\{a_1, \dots, a_n\}$ be the values of the variable X and $\{b_1, \dots, b_n\}$ be the value of the variable Y . For an event A we define the random variable $1_A(\omega) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$ Let $A_i = \{X = a_i\}$ and $B_j = \{Y = b_j\}$. We can write $X = \sum_{i=1}^n a_i 1_{A_i}$, $Y = \sum_{j=1}^m b_j 1_{B_j}$, where the events A_i and B_j are independent. Because $E[1_{A_i}] = P[A_i]$ and $E[1_{B_j}] = P[B_j]$ we have $E[1_{A_i} \cdot 1_{B_j}] = P[A_i] \cdot P[B_j]$. This implies $E[XY] = E[X]E[Y]$.

For uncorrelated random variables, we have $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$.

To see this, subtract first the mean from X and Y . This does not change the variance but now the random variables have mean 0. We have $\text{Var}[X+Y] = E[(X+Y)^2] = E[X^2 + 2XY + Y^2] = E[X^2] + 2E[XY] + E[Y^2]$.

8 Let X be the random variable of one single Bernoulli trial with $P[X = 1] = p$ and $P[X = 0] = 1 - p$. This implies $E[X] = 0P[X = 0] + pP[X = 1]$ and

$$\text{Var}[X] = (0 - p)^2 P[X = 0] + (1 - p)^2 P[X = 1] = p^2(1 - p) + (1 - p)^2 p = p(1 - p).$$

If we add n independent random variables of this type, then $E[X_1 + \dots + X_n] = np$ and $\text{Var}[X_1 + \dots + X_n] = \text{Var}[X_1 + \dots + X_n] = np(1 - p)$.

Homework due February 23, 2011

1 Look at the first n digits of $\pi = 3.1415926535897932385$ after the comma. For $n = 20$ this is 1415926535897932385. If you have no access to a computer, do this with $n = 20$. Find the mean and standard deviation of these data. Then draw the discrete distribution of the random variable which gives $X(k) = k$ 'th digit after the comma.

If you should have access to Mathematica. Here is the line which produces a histogram of the first n digits of π with respect to base $b = 10$:

```
b=10; n=100; Histogram[IntegerDigits[Floor[Pi b^n], b], b]
```

And here is the code which produces the mean and standard deviation of the first n digits:

```
b=10; n=100;
s=IntegerDigits[Floor[Pi b^n], b]; m=N[Sum[s[[k]], {k, n}]/n];
sigma=Sqrt[N[Sum[(s[[k]]-m)^2, {k, n}]/n]]; {m, sigma}
```

- 2 a) Verify that the empty set is independent to any other set.
b) Verify that the full laboratory Ω is independent to any other set.
c) Two disjoint sets of positive probability are not independent.
d) Find subsets A, B, C of $\{1, 2, 3, 4\}$ with probability $P[A] = |A|/4$ such that A is independent of B and B is independent of C but A is not independent of C .
- 3 Let Ω be the probability space of throwing two dice. Let X denote the difference of the two dice values and let Y be the sum. Find the correlation between these two random variables.

Lecture 12: Correlation

Independence and correlation

What is the difference between "uncorrelated" and "independent"? We have already mentioned the important fact:

If two random variables are independent, then they are uncorrelated.

The proof uses the notation $1_A(\omega) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$. We can write $X = \sum_{i=1}^n a_i 1_{A_i}$, $Y = \sum_{j=1}^m b_j 1_{B_j}$, where $A_i = \{X = a_i\}$ and $B_j = \{Y = b_j\}$ are independent. Because $E[1_{A_i}] = P[A_i]$ and $E[1_{B_j}] = P[B_j]$ we have $E[1_{A_i} \cdot 1_{B_j}] = P[A_i] \cdot P[B_j]$. Compare

$$E[XY] = E[(\sum_i a_i 1_{A_i})(\sum_j b_j 1_{B_j})] = \sum_{i,j} a_i b_j E[1_{A_i} 1_{B_j}] = \sum_{i,j} a_i b_j E[1_{A_i}] E[1_{B_j}].$$

with

$$E[X]E[Y] = E[(\sum_i a_i 1_{A_i})]E[(\sum_j b_j 1_{B_j})] = (\sum_i a_i E[1_{A_i}])(\sum_j b_j E[1_{B_j}]) = \sum_{i,j} a_i b_j E[1_{A_i}] E[1_{B_j}].$$

to see that the random variables are uncorrelated.

Remember the covariance

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

with which one has

$$\text{Var}[X] = \text{Cov}[X, X] = E[X \cdot X] - E[X]^2.$$

One defines also the correlation

$$\text{Corr}[XY] = \frac{\text{Cov}[XY]}{\sigma[X]\sigma[Y]}.$$

Here is a key connection between linear algebra and probability theory:

If X, Y are two random variables of zero mean, then the covariance $\text{Cov}[XY] = E[X \cdot Y]$ is the **dot product** of X and Y . The standard deviation of X is the length of X . The correlation is the cosine of the angle between the two vectors. Positive correlation means an acute angle, negative correlation means an obtuse angle. Uncorrelated means orthogonal.

If correlation can be seen geometrically, what is the geometric significance of independence?

Two random variables X, Y are independent if and only if for any functions f, g the random variables $f(X)$ and $f(Y)$ are uncorrelated.

You can check the above proof using $E[f(X)] = \sum_j f(a_j)E[A_j]$ and $E[g(X)] = \sum_j g(b_j)E[B_j]$. It still remains true. The only thing which changes are the numbers $f(a_i)$ and $g(b_j)$. By choosing suitable functions we can assure that all events $A_i = X = x_i$ and $B_j = Y = y_j$ are independent.

Lets explain this in a very small example, where the probability space has only three elements. In that case, random variables are vectors. We look at centered random variables, random variables of zero mean so that the covariance is the dot product. We refer here as vectors as random

variables, meaning that $X = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ is the function on the probability space $\{1, 2, 3\}$ given by

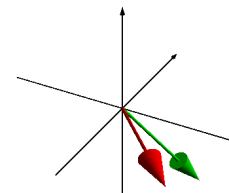
$f(1) = a, f(2) = b, f(3) = c$. As you know from linear algebra books, it is more common to write X_k instead of $X(k)$. Lets state an almost too obvious relation between linear algebra and probability theory because it is at the heart of the matter:

Vectors in R^n can be seen as random variables on the probability space $\{1, 2, \dots, n\}$.

It is because of this relation that it makes sense to combine the two subjects of linear algebra and probability theory. It is the reason why methods of linear algebra are immediately applicable to probability theory. It also reinforces the picture given in the first lecture that data are vectors. The expectation of data can be seen as the expectation of a random variable.

1 Here are two random variables of zero mean: $X = \begin{bmatrix} 3 \\ -3 \\ 0 \end{bmatrix}$ and $Y = \begin{bmatrix} 4 \\ 4 \\ -8 \end{bmatrix}$. They are

uncorrelated because their dot product $E[XY] = 3 \cdot 4 + (-3) \cdot 4 + 0 \cdot 8$ is zero. Are they independent? No, the event $A = \{X = 3\} = \{1\}$ and the event $B = \{Y = 4\} = \{1, 2\}$ are not independent. We have $P[A] = 1/3$, $P[B] = 2/3$ and $P[A \cap B] = 1/3$. We can also see it as follows: the random variables $X^2 = [9, 9, 0]$ and $Y^2 = [16, 16, 64]$ are no more uncorrelated: $E[X^2 \cdot Y^2] - E[X^2]E[Y^2] = 31040 - 746496$ is no more zero.



2 Lets take the case of throwing two coins. The probability space is $\{HH, HT, TH, TT\}$.

The random variable that the first dice is 1 is $X = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$. The random variable that

the second dice is 1 is $Y = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$. These random variables are independent. We can

center them to get centered random variables which are independent. [Alert: the random variables $Y = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}$ written down earlier are not independent, because the sets $A = \{X = 1\}$ and $\{Y = 1\}$ are disjoint and $P[A \cap B] = P[A] \cdot P[B]$ does not hold.]

- 3 The random variables $X = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $Y = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ are not uncorrelated because $E[(X - E[X])(Y - E[Y])]$ is the dot product $\begin{bmatrix} 2/3 \\ -1/3 \\ -1/3 \end{bmatrix} \begin{bmatrix} -1/3 \\ 2/3 \\ -1/3 \end{bmatrix}$ is not zero. Interestingly enough there are no nonconstant random variables on a probability space with three elements which are independent.¹

Finally lets mention again the important relation

Pythagoras theorem: $\text{Var}[X] + \text{Var}[Y] = \text{Var}[X + Y]$

if X, Y are uncorrelated random variables. It shows that not only the expectation but also the variance adds up, if we have independence. It is **Pythagoras theorem** because the notion "uncorrelated" means geometrically that the centered random variables are perpendicular and the variance is the length of the vector squared.

Parameter estimation

Parameter estimation is a central subject in statistics. We will look at it in the case of the Binomial distribution. As you know, if we have a coin which shows "heads" with probability p then the probability to have $X = k$ heads in n coin tosses is

The **Binomial distribution**

$$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

Keep this distribution in mind. It is one of the most important distributions in probability theory. Since this is the distribution of a sum of k random variables which are independent $X_k = 1$ if k 'th coin is head and $X_k = 0$ if it is tail, we know the mean and standard deviation of these variables $E[X] = np$ and $\text{Var}[X] = np(1-p)$.

- 4 Look at the data $X = (1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1)$.² Assume the data have a binomial distribution. What is our best bet for p ? We can compute its expectation $E[X] = 12/21 = 2/3$. We can also compute the variance $\text{Var}[X] = 2/9$. We know that the average of binomial distributed random variables has mean p and standard deviation $p(1-p)$. We want both to fit of course. Which one do we trust more, the mean or the standard deviation? In our example we got $p = 2/3$ and $p(1-p) = 2/9$. We were so lucky, were't we?

It turns out we were not lucky at all. There is a small miracle going on which is true for all $0-1$ data. For $0-1$ data **the mean determines the variance!**

Given any $0-1$ data of length n . Let k be the number of ones. If $p = k/n$ is the mean, then the variance of the data is $p(1-p)$.

Proof. Here is the statisticians proof: $\frac{1}{n} \sum_{i=1}^n (x_i - p)^2 = \frac{1}{n} (k(1-p)^2 + (n-k)(0-p)^2) = (k - 2kp + np^2)/n = p - 2p + p^2 = p^2 - p = p(1-p)$. And here is the probabilists proof: since $E[X^2] = E[X]$ we have $\text{Var}[X] = E[X^2] - E[X]^2 = E[X](1 - E[X]) = p(1-p)$.

Homework due February 23, 2011

- 1 Find the correlation coefficient $\text{Corr}[X, Y] = \text{Cov}[X, Y] / (\sigma[X]\sigma[Y])$ of the following π and e data

$$X = (31415926535897932385)$$

$$Y = (27182818284590452354)$$

- 2 Independence depends on the coordinate system. Find two random variables X, Y such that X, Y are independent but $X - 2Y, X + 2Y$ are not independent.
- 3 Assume you have a string X of $n = 1000$ numbers which takes the two values 0 and $a = 4$. You compute the mean of these data $p = (1/n) \sum_k X(k)$ and find $p = 1/5$. Can you figure out the standard deviation of these data?

¹This is true for finite probability spaces with prime $|\Omega|$ and uniform measure on it.

²These data were obtained with *IntegerDigits[Prime[100000], 2]* which writes the 100'000th prime $p = 1299709$ in binary form.

Lecture 13: Image and Kernel

The image of a matrix

If $T : \mathbf{R}^m \rightarrow \mathbf{R}^n$ is a linear transformation, then $\{T(\vec{x}) \mid \vec{x} \in \mathbf{R}^m\}$ is called the **image** of T . If $T(\vec{x}) = A\vec{x}$, where A is a matrix then the image of T is also called the image of A . We write $\text{im}(A)$ or $\text{im}(T)$.

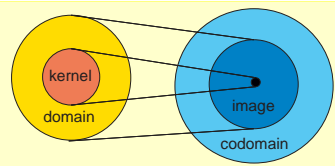
- 1 The map $T(x, y, z) = (x, y, 0)$ maps the three dimensional space into itself. It is linear because we can find a matrix A for which $T(\vec{x}) = A \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$. The image of T is the xy -plane.
- 2 If $T(x, y) = (\cos(\phi)x - \sin(\phi)y, \sin(\phi)x + \cos(\phi)y)$ is a rotation in the plane, then the image of T is the whole plane.
- 3 The averaging map $T(x, y, z) = (x + y + z)/3$ from \mathbf{R}^3 to \mathbf{R} has as image the entire real axes \mathbf{R} .

The **span** of vectors $\vec{v}_1, \dots, \vec{v}_k$ in \mathbf{R}^n is the set of all linear combinations $c_1\vec{v}_1 + \dots + c_k\vec{v}_k$.

- 4 The span of the standard basis vectors e_1, e_2 is the xy -plane.

A subset V of \mathbf{R}^n is called a **linear space** if it is closed under addition scalar multiplication and contains 0.

The image of a linear transformation $\vec{x} \mapsto A\vec{x}$ is the span of the column vectors of A . The image is a linear space.



How do we compute the image? If we are given a matrix for the transformation, then the image is the span of the column vectors. But we do not need all of them in general.

A column vector of A is called a **pivot column** if it contains a leading one after row reduction. The other columns are called **redundant columns**.

The pivot columns of A span the image of A .

Proof. You can see this by deleting the other columns. The new matrix B still allows to solve $Bx = b$ if $Ax = b$ could be solved.

- 5 Find the image of

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 & 4 \\ 5 & 5 & 5 & 5 & 5 \end{bmatrix}.$$

The kernel of a matrix

If $T : \mathbf{R}^m \rightarrow \mathbf{R}^n$ is a linear transformation, then the set $\{x \mid T(x) = 0\}$ is called the **kernel** of T . These are all vectors which are annihilated by the transformation. If $T(\vec{x}) = A\vec{x}$, then the kernel of T is also called the **kernel of A** . The kernel of A are all solutions to the linear system $Ax = 0$. We write $\ker(A)$ or $\ker(T)$.

- 6 The kernel of $T(x, y, z) = (x, y, 0)$ is the z -axes. Every vector $(0, 0, z)$ is mapped to 0.
- 7 The kernel of a rotation in the plane consists only of the zero point.
- 8 The kernel of the averaging map consists of all vector (x, y, z) for which $x + y + z = 0$. The kernel is a plane. In the language of random variables, the kernel of T consists of the centered random variables.

Also the kernel of a matrix A is a linear space.

How do we compute the kernel? Just solve the linear system of equations $A\vec{x} = \vec{0}$. Form $\text{rref}(A)$. For every column without leading 1 we can introduce a **free variable** s_i . If \vec{x} is the solution to $A\vec{x}_i = 0$, where all s_j are zero except $s_i = 1$, then $\vec{x} = \sum_j s_j \vec{x}_j$ is a general vector in the kernel.

- 9 Find the kernel of $A = \begin{bmatrix} 1 & 3 & 0 \\ 2 & 6 & 5 \\ 3 & 9 & 1 \\ -2 & -6 & 0 \end{bmatrix}$. Gauss-Jordan elimination gives: $B = \text{rref}(A) =$

$\begin{bmatrix} 1 & 3 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. There are two pivot columns and one redundant column. The equation $B\vec{x} = 0$ is equivalent to the system $x + 3y = 0, z = 0$. After fixing $z = 0$, can chose $y = t$ freely and obtain from the first equation $x = -3t$. Therefore, the kernel consists of vectors $t \begin{bmatrix} -3 \\ 1 \\ 0 \end{bmatrix}$.

Homework due March 2, 2011

- 1 Find the image and kernel of the **chess matrix**:

$$A = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

- 2 Find the image and kernel of the following **Pascal triangle matrix**:

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 3 & 0 & 3 & 0 & 1 & 0 \\ 1 & 0 & 4 & 0 & 6 & 0 & 4 & 0 & 1 \end{bmatrix}.$$

- 3 We work on the error correcting code as in the book (problem 53-54 in 3.1). Your task is to do the encoding and decoding using the initials from your name and write in one sentence using the terminology of "image" and "kernel", what the essence of this error correcting code is.

Step I) Encoding. To do so, we encode the letters of the alphabet by pairs of three vectors containing zeros and ones:

$$\begin{array}{lll} A = (0, 0, 0, 1), (0, 0, 0, 1) & B = (0, 0, 0, 1), (0, 0, 1, 0) & C = (0, 0, 0, 1), (0, 0, 1, 1) \\ D = (0, 0, 0, 1), (0, 1, 0, 1) & E = (0, 0, 0, 1), (0, 1, 1, 0) & F = (0, 0, 0, 1), (0, 1, 1, 1) \\ G = (0, 0, 0, 1), (1, 0, 0, 1) & H = (0, 0, 0, 1), (1, 0, 1, 0) & I = (0, 0, 0, 1), (1, 0, 1, 1) \\ J = (0, 0, 0, 1), (1, 1, 0, 1) & K = (0, 0, 0, 1), (1, 1, 1, 0) & L = (0, 0, 0, 1), (1, 1, 1, 1) \\ M = (0, 0, 1, 0), (0, 0, 0, 1) & N = (0, 0, 1, 0), (0, 0, 1, 0) & O = (0, 0, 1, 0), (0, 0, 1, 1) \\ P = (0, 0, 1, 0), (0, 1, 0, 1) & Q = (0, 0, 1, 0), (0, 1, 1, 0) & R = (0, 0, 1, 0), (0, 1, 1, 1) \\ S = (0, 0, 1, 0), (1, 0, 0, 1) & T = (0, 0, 1, 0), (1, 0, 1, 0) & U = (0, 0, 1, 0), (1, 0, 1, 1) \\ V = (0, 0, 1, 0), (1, 1, 0, 1) & W = (0, 0, 1, 0), (1, 1, 1, 0) & X = (0, 0, 1, 0), (1, 1, 1, 1) \\ Y = (0, 0, 1, 1), (1, 0, 0, 1) & Z = (0, 0, 1, 1), (1, 0, 1, 0) & ? = (0, 0, 1, 1), (1, 0, 1, 1) \\ ! = (0, 0, 1, 1), (1, 0, 0, 1) & . = (0, 0, 1, 1), (1, 0, 1, 0) & , = (0, 0, 1, 1), (1, 0, 1, 1) \end{array}$$

Choose a letter to get a pair of vectors (x, y) . $x = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}, y = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$. Use $1 + 1 = 0$ in

the matrix multiplications to build

$$Mx = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \quad My = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

Step II) Transmission.

Now add an error by switching one entry in each vector:

$$u = Mx + e = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \quad v = My + f = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

Step III) Detect the error e and f.

Form

$$Hu = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix}, Hv = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix}.$$

Now look in which column Hu or Hv is. Assume this column is k . Place 0's everywhere in the vectors e except at the k 'th entry where you put 1. For example if Hu is the second $e = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$, $f = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$. column, then put a 1 at the second place. We obtain vectors e and f :

Step IV) Decode the message.

$$\text{Use } P \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{bmatrix} = \begin{bmatrix} x_4 \\ x_5 \\ x_6 \\ x_7 \end{bmatrix} \text{ to determine } Pe = P \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}, Pf = P \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}. \text{ In}$$

an error-free transmission (Pu, Pv) would give the right result back. Now

$$Pu = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}, \quad Pv = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

satisfy $Pu = x + Pe, Pv = y + Pf$. We recover the original message (x, y) and so the letter from

$$x = Pu - Pe = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}, \quad y = Pv - Pf = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

Assignment: perform the encoding with your initials and check that you got the right letter back. Tell then in one sentence what the essence of this error correction is. In particular, how does the image of the "encoding matrix" M fit with the kernel of the "healing matrix" H ?

Lecture 14: Basis

Recall that subset X of \mathbf{R}^n which is closed under addition and scalar multiplication is called a **linear subspace** of \mathbf{R}^n . We have to check three conditions: (a) $0 \in V$, (b) $\vec{v} + \vec{w} \in V$ if $\vec{v}, \vec{w} \in V$. (c) $\lambda \vec{v} \in V$ if \vec{v} and λ is a real number.

- 1 The image and kernel of a transformation are linear spaces. This is an important example since this is how we describe linear spaces, either as the image of a linear transformation or the kernel of a linear transformations. Both are useful and they are somehow dual to each other. The kernel is associated to row vectors because we are perpendicular to all row vectors, the image is associated to column vectors because we are perpendicular to all column vectors.

A set \mathcal{B} of vectors $\vec{v}_1, \dots, \vec{v}_m$ is called **basis** of a linear subspace X of \mathbf{R}^n if they are **linear independent** and if they **span** the space X . Linear independent means that there are no nontrivial **linear relations** $a_1 \vec{v}_1 + \dots + a_m \vec{v}_m = 0$. Spanning the space means that every vector \vec{v} can be written as a linear combination $\vec{v} = a_1 \vec{v}_1 + \dots + a_m \vec{v}_m$ of basis vectors.

- 2 Two nonzero vectors in the plane form a basis if they are not parallel.

- 3 The standard basis vectors $e_1 = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}$, $e_2 = \begin{bmatrix} 0 \\ 1 \\ \dots \\ 0 \end{bmatrix}$, \dots , $e_n = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 1 \end{bmatrix}$ form a basis in \mathbf{R}^n .

Given a basis \mathcal{B} in V . Every vector in V can be written in a **unique manner** as a linear combination of vectors in \mathcal{B} .

To see this, assume \vec{v} is written in two different ways

$$\vec{v} = a_1 v_1 + a_2 v_2 + \dots + a_n v_n = b_1 v_1 + b_2 v_2 + \dots + b_n v_n.$$

a Then $(a_1 - b_1)v_1 + (a_2 - b_2)v_2 + \dots + (a_n - b_n)v_n = 0$. But this shows that the vectors are not linearly independent.

- 4 Given a probability space with 4 elements. Any random variable can be written in a unique way as a linear combination of the 4 random variables

$$\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

In general, one a finite probability space. A basis defines n random variables such that every random variable X can be written as a linear combination of X_1, \dots, X_n .

A set of random variables X_1, \dots, X_n which form a basis on a finite probability space $\Omega = \{1, \dots, n\}$ describe everything. Every random variable X which we want to compute can be expressed using these random variables.

Given a bunch of vectors in a linear space V , we can construct a basis of V by sticking the vectors as columns into a matrix A , then pick the pivot columns of A .

- 5 Find a basis of the space V spanned by the three vectors

$$A = \begin{bmatrix} 1 \\ 2 \\ 3 \\ -2 \end{bmatrix}, \begin{bmatrix} 3 \\ 6 \\ 9 \\ -6 \end{bmatrix}, \begin{bmatrix} 0 \\ 5 \\ 1 \\ 0 \end{bmatrix}.$$

Solution: Form the matrix

$$A = \begin{bmatrix} 1 & 3 & 0 \\ 2 & 6 & 5 \\ 3 & 9 & 1 \\ -2 & -6 & 0 \end{bmatrix}.$$

Row reduction shows that the first and third vector span the space V . This is a basis for V .

A $n \times n$ matrix $A = \begin{bmatrix} | & | & \dots & | \\ \vec{v}_1 & \vec{v}_2 & \dots & \vec{v}_n \\ | & | & \dots & | \end{bmatrix}$ is invertible if and only if $\vec{v}_1, \dots, \vec{v}_n$ is a basis in \mathbf{R}^n .

- 6 Find a basis for the image and the kernel of $A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$. **Solution:** In reduced row

echelon form is $B = \text{rref}(A) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$. To determine a basis of the kernel we write

$Bx = 0$ as a system of linear equations: $x + y = 0, z = 0$. The variable y is the free variable. With $y = t$, $x = -t$ is fixed. The linear system $\text{rref}(A)x = 0$ is solved by

$\vec{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = t \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$. So, $\vec{v} = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$ is a basis of the kernel. Because the first and third

vectors in $\text{rref}(A)$ are pivot columns, the vectors $\vec{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$, $\vec{v}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ form a basis of the

image of A .

Why do we not just always stick to the standard basis vectors $\vec{e}_1, \dots, \vec{e}_n$? The reason for the need of more general basis vectors is that they allow a **more flexible adaptation** to the situation. In geometry, the reflection of a ray at a plane or at a curve is better done in a basis adapted to the situation. For differential equations, the system can be solved in a suitable basis. Basis also matters in statistics. Given a set of random variables, we often can find a basis for them which consists of uncorrelated vectors.

How do we check that a set of vectors form a basis in \mathbf{R}^n ?

A set of n vectors $\vec{v}_1, \dots, \vec{v}_n$ in \mathbf{R}^n form a basis in \mathbf{R}^n if and only if the matrix A containing the vectors as column vectors is invertible.

7 The vectors $\vec{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$, $\vec{v}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ and $\vec{v}_3 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$ form a basis of \mathbf{R}^3 because the matrix

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 1 & 1 \end{bmatrix}$$

is invertible.

More generally, the pivot columns of an arbitrary matrix A form a basis for the image of A . Since we represent linear spaces always as the kernel or image of a linear map, the problem of finding a basis to a linear space is always the problem of finding a basis for the image or finding a basis for the kernel of a matrix.

Homework due March 2, 2011

1 Find a basis for the image and kernel of the Chess matrix:

$$A = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

and verify the rank-nullity theorem in this case.

2 Find a basis for the set of vectors perpendicular to the image of A , where A is the Pascal matrix.

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 3 & 0 & 3 & 0 & 1 & 0 \\ 1 & 0 & 4 & 0 & 6 & 0 & 4 & 0 & 1 \end{bmatrix}.$$

3 Verify that a vector is in the kernel of a matrix A^T if and only if it is perpendicular to the image of A .

Verify it in the following concrete example

$$A = \begin{bmatrix} 1 & 3 & 2 \\ 2 & 6 & 0 \\ 3 & 9 & 1 \\ 4 & 1 & 1 \end{bmatrix}.$$

The matrix A^T denotes the matrix where the columns and rows of A are switched. It is called the transpose of the matrix A .

Lecture 15: Dimension

Remember that $X \subset \mathbb{R}^n$ is called a **linear space** if $\vec{0} \in X$ and if X is closed under addition and scalar multiplication. Examples are \mathbb{R}^n , $X = \ker(A)$, $X = \text{im}(A)$, or the row space of a matrix. In order to describe linear spaces, we had the notion of a basis:

$\mathcal{B} = \{\vec{v}_1, \dots, \vec{v}_n\} \subset X$ is a basis if two conditions are satisfied: \mathcal{B} is **linear independent** meaning that $c_1\vec{v}_1 + \dots + c_n\vec{v}_n = \vec{0}$ implies $c_1 = \dots = c_n = 0$. Then \mathcal{B} **span** X : $\vec{v} \in X$ then $\vec{v} = a_1\vec{v}_1 + \dots + a_n\vec{v}_n$. The spanning condition for a basis assures that there are **enough** vectors to represent any other vector, the linear independence condition assures that there are **not too many** vectors. Every $\vec{v} \in X$ can be written uniquely as a sum $\vec{v} = a_1\vec{v}_1 + \dots + a_n\vec{v}_n$ of basis vectors.

The **dimension** of a linear space V is the number of basis vectors in V .

The dimension of three dimensional space is 3. The dimension is independent on where the space is embedded in. For example: a line in the plane and a line embedded in space have both the dimension 1.

- 1 The dimension of \mathbb{R}^n is n . The standard basis is

$$\begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \dots \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ \dots \\ 1 \end{bmatrix}.$$

- 2 The dimension of $\{0\}$ is zero. The dimension of any line is 1. The dimension of a plane is 2.
- 3 The dimension of the image of a matrix is the number of pivot columns. We can construct a basis of the kernel and image of a linear transformation $T(x) = Ax$ by forming $B = \text{rref}(A)$. The set of Pivot columns in A form a basis of the image of T .
- 4 The dimension of the kernel of a matrix is the number of free variables. It is also called **nulity**. A basis for the kernel is obtained by solving $Bx = 0$ and introducing free variables for the redundant columns.

Given a basis $\mathcal{A} = \{v_1, \dots, v_n\}$ and a basis $\mathcal{B} = \{w_1, \dots, w_m\}$ of X , then $m = n$.

Lemma: if q vectors $\vec{w}_1, \dots, \vec{w}_q$ span X and $\vec{v}_1, \dots, \vec{v}_p$ are linearly independent in X , then $q \geq p$.

Assume $q < p$. Because \vec{w}_i span, each vector \vec{v}_i can be written as $\sum_{j=1}^q a_{ij}\vec{w}_j = \vec{v}_i$. Now do Gauss-

Jordan elimination of the augmented $(p \times (q+n))$ -matrix to this system: $\left[\begin{array}{ccc|c} a_{11} & \dots & a_{1q} & \vec{v}_1^T \\ \dots & \dots & \dots & \dots \\ a_{p1} & \dots & a_{pq} & \vec{v}_p^T \end{array} \right]$,

where \vec{v}_i^T is the vector \vec{v}_i written as a row vector. Each row of A of this $[A|b]$ contains some nonzero entry. We end up with a matrix, which contains a last row $\left[0 \dots 0 \mid b_1\vec{w}_1^T + \dots + b_q\vec{w}_q^T \right]$ showing that $b_1\vec{w}_1^T + \dots + b_q\vec{w}_q^T = 0$. Not all b_j are zero because we had to eliminate some nonzero

entries in the last row of A . This nontrivial relation of \vec{w}_i^T (and the same relation for column vectors \vec{w}) is a contradiction to the linear independence of the \vec{w}_j . The assumption $q < p$ can not be true.

To prove the proposition, use the lemma in two ways. Because \mathcal{A} spans and \mathcal{B} is linearly independent, we have $m \leq n$. Because \mathcal{B} spans and \mathcal{A} is linearly independent, we have $n \leq m$.

The following theorem is also called the **rank-nulity** theorem because $\dim(\text{im}(A))$ is the rank and $\dim(\ker(A))$ is the nulity.

Fundamental theorem of linear algebra: Let $A : \mathbb{R}^m \rightarrow \mathbb{R}^n$ be a linear map.

$$\dim(\ker(A)) + \dim(\text{im}(A)) = m$$

There are n columns. $\dim(\ker(A))$ is the number of columns without leading 1, $\dim(\text{im}(A))$ is the number of columns with leading 1.

- 5 If A is an invertible $n \times n$ matrix, then the dimension of the image is n and that the $\dim(\ker(A)) = 0$.
- 6 The first grade **multiplication table** matrix

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 4 & 6 & 8 & 10 & 12 & 14 & 16 & 18 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 & 27 \\ 4 & 8 & 12 & 16 & 20 & 24 & 28 & 32 & 36 \\ 5 & 10 & 15 & 20 & 25 & 30 & 35 & 40 & 45 \\ 6 & 12 & 18 & 24 & 30 & 36 & 42 & 48 & 54 \\ 7 & 14 & 21 & 28 & 35 & 42 & 49 & 56 & 63 \\ 8 & 16 & 24 & 32 & 40 & 48 & 56 & 64 & 72 \\ 9 & 18 & 27 & 36 & 45 & 54 & 63 & 72 & 81 \end{bmatrix}.$$

has rank 1. The nulity is therefore 8.

- 7 Are there a 4×4 matrices A, B of ranks 3 and 1 such that $\text{ran}(AB) = 0$? **Solution.** Yes, we can even find examples which are diagonal.
- 8 Is there 4×4 matrices A, B of rank 3 and 1 such that $\text{ran}(AB) = 2$? **Solution.** No, the kernel of B is three dimensional by the rank-nulity theorem. But this also means the kernel of AB is three dimensional (the same vectors are annihilated). But this implies that the rank of AB can maximally be 1.

The rank of AB can not be larger than the rank of A or the rank of B .
The nulity of AB can not be smaller than the nulity of B .

We end this lecture with an informal remark about fractal dimension: Mathematicians study objects with non-integer dimension since the early 20'th century. The topic became fashion in the 1980'ies, when mathematicians started to generate fractals on computers. To define fractals, the notion of dimension is extended. Here is an informal definition which can be remembered easily and allows to compute the dimension of most "star fractals" you find on

the internet when searching for fractals. It assumes that X is a bounded set. You can pick up this definition also in the Startreck movie (2009) when little Spock gets some math and ethics lectures in school. It is the simplest definition and also called box counting dimension in the math literature on earth.

Assume we can cover X with $n = n(r)$ cubes of size r and not less. The **fractal dimension** is defined as the limit

$$\dim(X) = \frac{-\log(n)}{\log(r)}$$

as $r \rightarrow 0$.

For linear spaces X , the fractal dimension of X intersected with the unit cube agrees with the usual dimension in linear algebra.

Proof. Take a basis $\mathcal{B} = \{v_1, \dots, v_m\}$ in X . We can assume that this basis vectors are all orthogonal and each vector has length 1. For given $r > 0$, place cubes at the lattice points $\sum_{j=1}^m k_j r v_j$ with integer k_j . This covers the intersection X with the unit cube with (C/r^m) cubes where $1/\sqrt{m} \leq C \leq \sqrt{m}$. The dimension of X is

$$\dim(X) = \log(C/r^m)/\log(r) = \log(C)/\log(r) + m$$

which converges to m as $r \rightarrow 0$.

9 We cover the **unit interval** $[0, 1]$ with $n = 1/r$ intervals of length r . Now,

$$\dim(X) = \frac{-\log(1/r)}{\log(r)} = 1.$$

10 We cover the **unit square** with $n = 1/r^2$ squares of length r . Now,

$$\dim(X) = \frac{-\log(1/r^2)}{\log(r)} = 2.$$

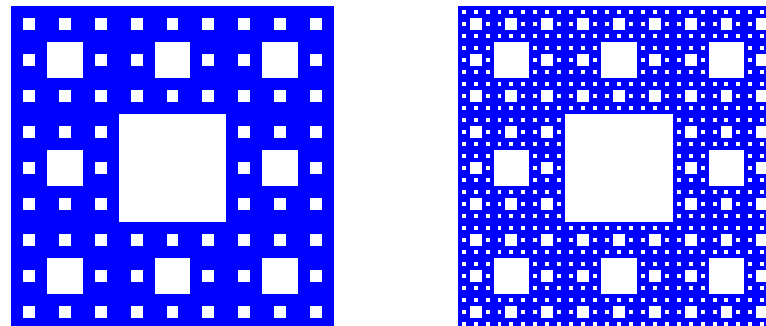
11 The **Cantor set** is obtained recursively by dividing intervals into 3 pieces and throwing away the middle one. We can cover the Cantor set with $n = 2^k$ intervals of length $r = 1/3^k$ so that

$$\dim(X) = \frac{-\log(2^k)}{\log(1/3^k)} = \log(2)/\log(3).$$

12 The **Shirpinski carpet** is constructed recursively by dividing a square in 9 equal squares and throwing away the middle one, repeating this procedure with each of the squares etc. At the k 'th step, we need $n = 8^k$ squares of length $r = 1/3^k$ to cover X . The dimension is

$$\dim(X) = \frac{-\log(8^k)}{\log(1/3^k)} = \log(8)/\log(3).$$

This is smaller than $2 = \log(9)/\log(3)$ but larger than $1 = \log(3)/\log(3)$.



Homework due March 9, 2011

- 1 a) Give an example of a 5×6 matrix with $\dim(\ker(A)) = 3$ or argue why it does not exist.
b) Give an example 5×8 matrix with $\dim(\ker(A)) = 2$ or argue why it does not exist.

- 2 a) Find a basis for the subspace of all vectors in R^5 satisfying

$$x_1 + 2x_2 + 3x_3 - x_4 + x_5 = 0.$$

- b) Find a basis for the space spanned by the rows of the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 4 & 6 & 7 \end{bmatrix}.$$

- 3 a) Assume two linear subspaces V, W of R^m have the property that $V \cap W = \{0\}$ and such that every vector in R^m can be written as $x + y$ with $x \in V, y \in W$. Find a formula which relates $\dim(V), \dim(W)$ and m .
b) Assume now that $V \cap W$ is 1 dimensional. What is the relation between $\dim(V), \dim(W)$ and m .

Lecture 16: Coordinates

A basis $\mathcal{B} = \{\vec{v}_1, \dots, \vec{v}_n\}$ of \mathbf{R}^n defines the matrix $S = \begin{bmatrix} | & \dots & | \\ \vec{v}_1 & \dots & \vec{v}_n \\ | & \dots & | \end{bmatrix}$. It is called the **coordinate transformation matrix** of the basis.

By definition, the matrix S is invertible: the linear independence of the column vectors implies S has no kernel. By the rank-nullity theorem, the image is the entire space \mathbf{R}^n .

If \vec{x} is a vector in \mathbf{R}^n and $\vec{x} = c_1\vec{v}_1 + \dots + c_n\vec{v}_n$, then c_i are called the **\mathcal{B} -coordinates** of \vec{v} .

We have seen that such a representation is unique if the basis is fixed.

We write $[\vec{x}]_{\mathcal{B}} = \begin{bmatrix} c_1 \\ \dots \\ c_n \end{bmatrix}$. If $\vec{x} = \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix}$, we have $\vec{x} = S([\vec{x}]_{\mathcal{B}})$.

The \mathcal{B} -coordinates of \vec{x} are obtained by applying S^{-1} to the coordinates of the standard basis:

$$[\vec{x}]_{\mathcal{B}} = S^{-1}(\vec{x})$$

This just rephrases that $S([\vec{x}]_{\mathcal{B}}) = \vec{x}$. Remember the column picture. The left hand side is just $c_1\vec{v}_1 + \dots + c_n\vec{v}_n$ where the v_j are the column vectors of S .

- 1 If $\vec{x} = \begin{bmatrix} 4 \\ -2 \\ 3 \end{bmatrix}$, then 4, -2, 3 are the standard coordinates. With the standard basis $\mathcal{B} = \{e_1, e_2, e_3\}$ we have $\vec{x} = 4e_1 - 2e_2 + 3e_3$.

- 2 If $\vec{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\vec{v}_2 = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$, then $S = \begin{bmatrix} 1 & 3 \\ 2 & 5 \end{bmatrix}$. A vector $\vec{v} = \begin{bmatrix} 6 \\ 9 \end{bmatrix}$ has the coordinates

$$S^{-1}\vec{v} = \begin{bmatrix} -5 & 3 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 6 \\ 9 \end{bmatrix} = \begin{bmatrix} -3 \\ 3 \end{bmatrix}.$$

Indeed, as we can check, $-3\vec{v}_1 + 3\vec{v}_2 = \vec{v}$.

- 3 Find the coordinates of $\vec{v} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ with respect to the basis $\mathcal{B} = \{\vec{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}\}$. We have $S = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $S^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$. Therefore $[\vec{v}]_{\mathcal{B}} = S^{-1}\vec{v} = \begin{bmatrix} -1 \\ 3 \end{bmatrix}$. Indeed $-1\vec{v}_1 + 3\vec{v}_2 = \vec{v}$.

If $\mathcal{B} = \{v_1, \dots, v_n\}$ is a basis in \mathbf{R}^n and T is a linear transformation on \mathbf{R}^n , then the \mathcal{B} -matrix of T is

$$B = \begin{bmatrix} | & \dots & | \\ [T(\vec{v}_1)]_{\mathcal{B}} & \dots & [T(\vec{v}_n)]_{\mathcal{B}} \\ | & \dots & | \end{bmatrix}.$$

- 4 Find a clever basis for the reflection of a light ray at the line $x + 2y = 0$. **Solution:** Use one vector in the line and an other one perpendicular to it: $\vec{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\vec{v}_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$. We achieved so $B = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = S^{-1}AA$ with $S = \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}$.

If A is the matrix of a transformation in the standard coordinates then

$$B = S^{-1}AS$$

is the matrix in the new coordinates.

The transformation S^{-1} maps the coordinates from the standard basis into the coordinates of the new basis. In order to see what a transformation A does in the new coordinates, we first map it back to the old coordinates, apply A and then map it back again to the new coordinates.

$$\begin{array}{ccc} \vec{v} & \xleftarrow{S} & \vec{w} = [\vec{v}]_{\mathcal{B}} \\ A \downarrow & & \downarrow B \\ A\vec{v} & \xrightarrow{S^{-1}} & B\vec{w} \end{array}$$

- 5 Let T be the reflection at the plane $x + 2y + 3z = 0$. Find the transformation matrix B in the basis $\vec{v}_1 = \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}$, $\vec{v}_2 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, $\vec{v}_3 = \begin{bmatrix} 0 \\ 3 \\ -2 \end{bmatrix}$. Because $T(\vec{v}_1) = \vec{v}_1 = [\vec{e}_1]_{\mathcal{B}}$, $T(\vec{v}_2) = \vec{v}_2 = [\vec{e}_2]_{\mathcal{B}}$, $T(\vec{v}_3) = -\vec{v}_3 = -[\vec{e}_3]_{\mathcal{B}}$, the solution is $B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Two matrices A, B which are related by $B = S^{-1}AS$ are called **similar**.

- 6 If A is similar to B , then $A^2 + A + 1$ is similar to $B^2 + B + 1$. $B = S^{-1}AS$, $B^2 = S^{-1}B^2S$, $S^{-1}S = \mathbf{1}$, $S^{-1}(A^2 + A + 1)S = B^2 + B + \mathbf{1}$.
- 7 If A is a general 2×2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and let $S = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, What is $S^{-1}AS$? **Solution:** $\begin{bmatrix} d & c \\ b & a \end{bmatrix}$. Both the rows and columns have switched. This example shows that the matrices $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $\begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ are similar.

Homework due March 9, 2011

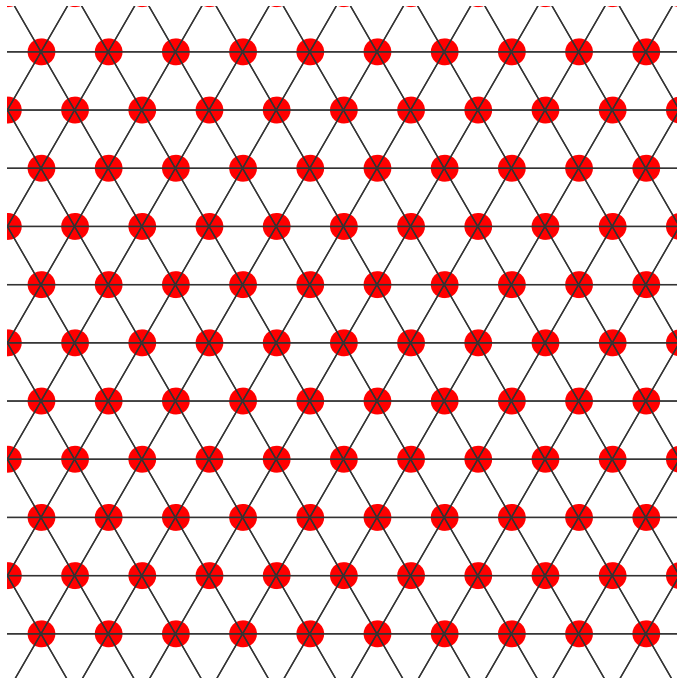
- 1 Find the \mathcal{B} -matrix B of the linear transformation which is given in standard coordinates as

$$T \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

if $\mathcal{B} = \left\{ \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}$.

- 2 Let V be the plane spanned by $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$. Find the matrix A of reflection at the plane V by using a suitable coordinate system.

- 3 a) Find a basis which describes best the points in the following lattice: We aim to describe the lattice points with integer coordinates (k, l) .
b) Once you find the basis, draw all the points which have (x, y) coordinates in the disc $x^2 + y^2 \leq 10$



Lecture 17: Orthogonality

Two vectors \vec{v} and \vec{w} are called **orthogonal** if their dot product is zero $\vec{v} \cdot \vec{w} = 0$.

1 $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 6 \\ -3 \end{bmatrix}$ are orthogonal in \mathbf{R}^2 .

2 \vec{v} and \vec{w} are both orthogonal to the cross product $\vec{v} \times \vec{w}$ in \mathbf{R}^3 . The dot product between \vec{v} and $\vec{v} \times \vec{w}$ is the determinant

$$\det \begin{pmatrix} v_1 & v_2 & v_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix}.$$

\vec{v} is called a **unit vector** if its length is one: $||\vec{v}|| = \sqrt{\vec{v} \cdot \vec{v}} = 1$.

A set of vectors $\mathcal{B} = \{\vec{v}_1, \dots, \vec{v}_n\}$ is called **orthogonal** if they are pairwise orthogonal. They are called **orthonormal** if they are also unit vectors. A basis is called an **orthonormal basis** if it is a basis which is orthonormal. For an orthonormal basis, the matrix with entries $A_{ij} = \vec{v}_i \cdot \vec{v}_j$ is the unit matrix.

Orthogonal vectors are linearly independent. A set of n orthogonal vectors in \mathbf{R}^n automatically form a basis.

Proof: The dot product of a **linear relation** $a_1\vec{v}_1 + \dots + a_n\vec{v}_n = 0$ with \vec{v}_k gives $a_k\vec{v}_k \cdot \vec{v}_k = a_k||\vec{v}_k||^2 = 0$ so that $a_k = 0$. If we have n linear independent vectors in \mathbf{R}^n , they automatically span the space because the fundamental theorem of linear algebra shows that the image has then dimension n .

A vector $\vec{w} \in \mathbf{R}^n$ is called **orthogonal** to a linear space V , if \vec{w} is orthogonal to every vector $\vec{v} \in V$. The **orthogonal complement** of a linear space V is the set W of all vectors which are orthogonal to V .

The orthogonal complement of a linear space V is a linear space. It is the kernel of A^T , if the image of A is V .

To check this, take two vectors in the orthogonal complement. They satisfy $\vec{v} \cdot \vec{w}_1 = 0, \vec{v} \cdot \vec{w}_2 = 0$. Therefore, also $\vec{v} \cdot (\vec{w}_1 + \vec{w}_2) = 0$.

Pythagoras theorem: If \vec{x} and \vec{y} are orthogonal, then $||\vec{x} + \vec{y}||^2 = ||\vec{x}||^2 + ||\vec{y}||^2$.

Proof. Expand $(\vec{x} + \vec{y}) \cdot (\vec{x} + \vec{y})$.

Cauchy-Schwarz: $|\vec{x} \cdot \vec{y}| \leq ||\vec{x}|| ||\vec{y}||$.

Proof: $\vec{x} \cdot \vec{y} = ||\vec{x}|| ||\vec{y}|| \cos(\alpha)$. If $|\vec{x} \cdot \vec{y}| = ||\vec{x}|| ||\vec{y}||$, then \vec{x} and \vec{y} are parallel.

Triangle inequality: $||\vec{x} + \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}||$.

Proof: $(\vec{x} + \vec{y}) \cdot (\vec{x} + \vec{y}) = ||\vec{x}||^2 + ||\vec{y}||^2 + 2\vec{x} \cdot \vec{y} \leq ||\vec{x}||^2 + ||\vec{y}||^2 + 2||\vec{x}|| ||\vec{y}|| = (||\vec{x}|| + ||\vec{y}||)^2$.

Angle: The **angle** between two vectors \vec{x}, \vec{y} is $\alpha = \arccos\left(\frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| ||\vec{y}||}\right)$.

$\cos(\alpha) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| ||\vec{y}||} \in [-1, 1]$ is the **statistical correlation** of \vec{x} and \vec{y} if the vectors \vec{x}, \vec{y} represent data of zero mean.

3 Express the fact that \vec{x} is in the kernel of a matrix A using orthogonality. **Answer** $A\vec{x} = 0$ means that $\vec{w}_k \cdot \vec{x} = 0$ for every row vector \vec{w}_k of \mathbf{R}^n . Therefore, the orthogonal complement of the row space is the kernel of a matrix.

The **transpose** of a matrix A is the matrix $(A^T)_{ij} = A_{ji}$. If A is a $n \times m$ matrix, then A^T is a $m \times n$ matrix. Its rows are the columns of A . For square matrices, the transposed matrix is obtained by reflecting the matrix at the diagonal.

4 The transpose of a vector $A = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ is the row vector $A^T = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$.
The transpose of the matrix $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ is the matrix $\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$.

$(AB)^T = B^T A^T$, $(A^T)^T = A$, $(A^{-1})^T = (A^T)^{-1}$.
 $\vec{v}^T \vec{w}$ is the dot product $\vec{v} \cdot \vec{w}$.
 $\vec{x} \cdot A\vec{y} = A^T \vec{x} \cdot \vec{y}$.

The proofs are direct computations. Here is the first identity:

$$(AB)_{kl}^T = (AB)_{lk} = \sum_i A_{li} B_{ik} = \sum_i B_{ki}^T A_{li}^T = (B^T A^T)_{kl}.$$

A linear transformation is called **orthogonal** if $A^T A = I_n$.

We see that a matrix is orthogonal if and only if the column vectors form an orthonormal basis.

Orthogonal matrices preserve length and angles. They satisfy $A^{-1} = A^T$.

5 A rotation is orthogonal.

Orthogonality over time

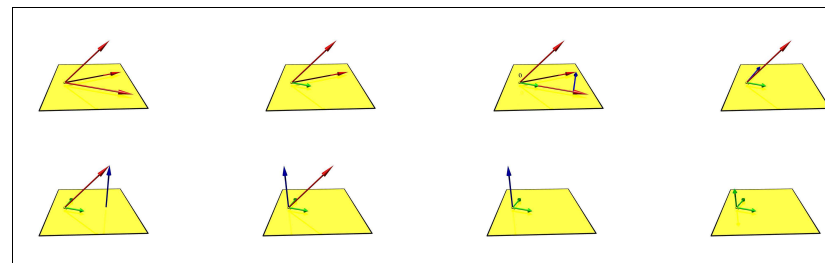
- From -2800 BC until -2300 BC, Egyptians used ropes divided into length ratios like 3 : 4 : 5 to build triangles. This allowed them to triangulate areas quite precisely: for example to build irrigation needed because the Nile was reshaping the land constantly or to build the pyramids: for the **great pyramid at Giza** with a base length of 230 meters, the average error on each side is less than 20cm, an error of less than 1/1000. A key to achieve this was **orthogonality**.
- During one of Thales (-624 BC to (-548 BC)) journeys to Egypt, he used a geometrical trick to **measure the height** of the great pyramid. He measured the size of the shadow of the pyramid. Using a stick, he found the relation between the length of the stick and the length of its shadow. The same length ratio applies to the pyramid (**orthogonal** triangles). Thales found also that triangles inscribed into a circle and having as the base as the diameter must have a right angle.
- The Pythagoreans (-572 until -507) were interested in the discovery that the squares of a lengths of a triangle with two **orthogonal** sides would add up as $a^2 + b^2 = c^2$. They were puzzled in assigning a length to the diagonal of the unit square, which is $\sqrt{2}$. This number is irrational because $\sqrt{2} = p/q$ would imply that $q^2 = 2p^2$. While the prime factorization of q^2 contains an even power of 2, the prime factorization of $2p^2$ contains an odd power of 2.
- Eratosthenes (-274 until 194) realized that while the sun rays were **orthogonal** to the ground in the town of Scene, this did no more do so at the town of Alexandria, where they would hit the ground at 7.2 degrees). Because the distance was about 500 miles and 7.2 is 1/50 of 360 degrees, he measured the circumference of the earth as 25'000 miles - pretty close to the actual value 24'874 miles.
- Closely related to **orthogonality** is **parallelism**. Mathematicians tried for ages to prove Euclid's parallel axiom using other postulates of Euclid (-325 until -265). These attempts had to fail because there are geometries in which parallel lines always meet (like on the sphere) or geometries, where parallel lines never meet (the Poincaré half plane). Also these geometries can be studied using linear algebra. The geometry on the sphere with **rotations**, the geometry on the half plane uses Möbius transformations, 2×2 matrices with determinant one.
- The question whether the angles of a right triangle do always add up to 180 degrees became an issue when geometries were discovered, in which the measurement depends on the position in space. Riemannian geometry, founded 150 years ago, is the foundation of **general relativity**, a theory which describes gravity geometrically: the presence of mass bends space-time, where the dot product can depend on space. **Orthogonality** becomes relative. On a sphere for example, the three angles of a triangle are bigger than 180° . Space is curved.
- In **probability theory**, the notion of **independence** or **decorrelation** is used. For example, when throwing a dice, the number shown by the first dice is independent and decorrelated from the number shown by the second dice. Decorrelation is identical to **orthogonality**, when vectors are associated to the random variables. The **correlation coefficient** between two vectors \vec{v}, \vec{w} is defined as $\vec{v} \cdot \vec{w} / (|\vec{v}| |\vec{w}|)$. It is the cosine of the angle between these vectors.

- In **quantum mechanics**, states of atoms are described by functions in a linear space of functions. The states with energy $-E_B/n^2$ (where $E_B = 13.6\text{eV}$ is the Bohr energy) in a hydrogen atom. States in an atom are **orthogonal**. Two states of two different atoms which don't interact are **orthogonal**. One of the challenges in quantum computation, where the computation deals with qubits (=vectors) is that orthogonality is not preserved during the computation (because we don't know all the information). Different states can interact.

Homework due March 9, 2011

- Assume X, Y, Z, U, V are independent random variables which have all standard deviation 1. Find the standard deviation of $X + Y + 2Z + 3U - V$.
 - We have two random variables X and Y of standard deviation 1 and 2 and correlation -0.5 . Can you find a combination $Z = aX + bY$ such that X, Z are uncorrelated?
 - Verify that the standard deviation of $X + Y$ is smaller or equal than the sum of the standard deviations of X and Y .
- Verify that if A, B are orthogonal matrices then their product $A.B$ and $B.A$ are orthogonal matrices.
 - Verify that if A, B are orthogonal matrices, then their inverse is an orthogonal matrix.
 - Verify that 1_n is an orthogonal matrix.

These properties show that the space of $n \times n$ orthogonal matrices form a "group". It is called $O(n)$.
- Given a basis \mathcal{B} , we describe a process called **Gram-Schmidt orthogonalization** which produces an orthonormal basis. If $\vec{v}_1, \dots, \vec{v}_n$ are the basis vectors let $\vec{w}_1 = \vec{v}_1$ and $\vec{u}_1 = \vec{w}_1 / |\vec{w}_1|$. The Gram-Schmidt process recursively constructs from the already constructed orthonormal set $\vec{u}_1, \dots, \vec{u}_{i-1}$ which spans a linear space V_{i-1} the new vector $\vec{w}_i = (\vec{v}_i - \text{proj}_{V_{i-1}}(\vec{v}_i))$ which is orthogonal to V_{i-1} , and then normalizing \vec{w}_i to get $\vec{u}_i = \vec{w}_i / |\vec{w}_i|$. Each vector \vec{w}_i is orthonormal to the linear space V_{i-1} . The vectors $\{\vec{u}_1, \dots, \vec{u}_n\}$ form an orthonormal basis in V .



Find an orthonormal basis for $\vec{v}_1 = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$, $\vec{v}_2 = \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix}$ and $\vec{v}_3 = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$.

Lecture 18: Projections

A linear transformation P is called an **orthogonal projection** if the image of P is V and the kernel is perpendicular to V and $P^2 = P$.

Orthogonal projections are useful for many reasons. First of all however:

In an orthonormal basis $P = P^T$. The point Px is the point on V which is closest to x .

Proof. $Px - x$ is perpendicular to Px because

$$(Px - x) \cdot Px = Px \cdot Px - x \cdot Px = P^2x \cdot x - x \cdot Px = Px \cdot x - x \cdot Px = 0.$$

We have used that $P^2 = P$ and $Av \cdot w = v \cdot A^T w$.

For an orthogonal projection P there is a basis in which the matrix is diagonal and contains only 0 and 1.

Proof. Chose a basis \mathcal{B}_∞ of the kernel of P and a basis \mathcal{B}_ϵ of V , the image of P . Since for every $\vec{v} \in \mathcal{B}_1$, we have $Pv = 0$ and for every $\vec{v} \in \mathcal{B}_2$, we have $Pv = v$, the matrix of P in the basis $\mathcal{B}_1 \cup \mathcal{B}_2$ is diagonal.

1 The matrix

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} / 3$$

is a projection onto the one dimensional space spanned by $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

2 The matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

is a projection onto the xy -plane.

3 If V is a line containing the unit vector \vec{v} then $Px = v(v \cdot x)$, where \cdot is the dot product. Writing this as a matrix product shows $Px = AA^T x$ where A is the $n \times 1$ matrix which contains \vec{v} as the column. If v is not a unit vector, we know from multivariable calculus that $Px = v(v \cdot x)/|v|^2$. Since $|v|^2 = A^T A$ we have $Px = A(A^T A)^{-1} A^T x$.

How do we construct the matrix of an orthogonal projection? Lets look at an other example

4 Let v, w be two vectors in three dimensional space which both have length 1 and are perpendicular to each other. Now

$$Px = (v \cdot x)\vec{v} + (w \cdot x)\vec{w}.$$

We can write this as AA^T , where A is the matrix which contains the two vectors as column vectors. For example, if $v = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} / \sqrt{3}$ and $w = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} / \sqrt{6}$, then

$$P = AA^T = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{6} \\ -1/\sqrt{3} & 2/\sqrt{6} \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & -1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{6} & 2/\sqrt{6} \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

For any matrix, we have $(\text{im}(A))^\perp = \ker(A^T)$.

Remember that a vector is in the kernel of A^T if and only if it is orthogonal to the rows of A^T and so to the columns of A . The kernel of A^T is therefore the orthogonal complement of $\text{im}(A)$ for any matrix A :

For any matrix, we have $\ker(A) = \ker(A^T A)$.

Proof. \subset is clear. On the other hand $A^T Av = 0$ means that Av is in the kernel of A^T . But since the image of A is orthogonal to the kernel of A^T , we have $Av = 0$, which means v is in the kernel of A .

If V is the image of a matrix A with trivial kernel, then the projection P onto V is

$$Px = A(A^T A)^{-1} A^T x.$$

Proof. Let y be the vector on V which is closest to Ax . Since $y - Ax$ is perpendicular to the image of A , it must be in the kernel of A^T . This means $A^T(y - Ax) = 0$. Now solve for x to get the **least square solution**

$$x = (A^T A)^{-1} A^T y.$$

The projection is $Ax = A(A^T A)^{-1} A^T y$.

5 Let $A = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$. The orthogonal projection onto $V = \text{im}(A)$ is $\vec{b} \mapsto A(A^T A)^{-1} A^T \vec{b}$. We

$$\text{have } A^T A = \begin{bmatrix} 5 & 0 \\ 2 & 1 \end{bmatrix} \text{ and } A(A^T A)^{-1} A^T = \begin{bmatrix} 1/5 & 2/5 & 0 \\ 2/5 & 4/5 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

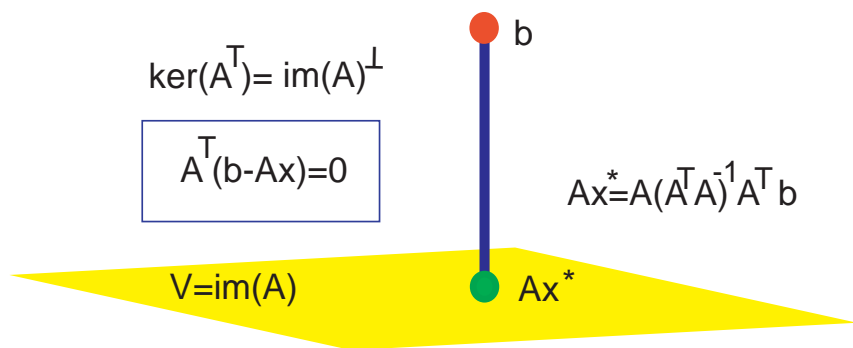
For example, the projection of $\vec{b} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ is $\vec{x}^* = \begin{bmatrix} 2/5 \\ 4/5 \\ 0 \end{bmatrix}$ and the distance to \vec{b} is $1/\sqrt{5}$.

The point \vec{x}^* is the point on V which is closest to \vec{b} .

6 Let $A = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \end{bmatrix}$. Problem: find the matrix of the orthogonal projection onto the image of A .

The image of A is a one-dimensional line spanned by the vector $\vec{v} = (1, 2, 0, 1)$. We calculate $A^T A = 6$. Then

$$A(A^T A)^{-1} A^T = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 & 1 \end{bmatrix} / 6 = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 2 & 4 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix} / 6.$$



Homework due March 23, 2011

1 a) Find the orthogonal projection of $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ onto the subspace of \mathbb{R}^4 spanned by

$$\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

b) Find the orthogonal projection of $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ onto the subspace of \mathbb{R}^5 which has the basis

$$\mathcal{B} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

2 Let A be a matrix with trivial kernel. Define the matrix $P = A(A^T A)^{-1} A^T$.

a) Verify that we have $P^T = P$.

b) Verify that we have $P^2 = P$.

For this problem, just use the basis properties of matrix algebra like $(AB)^T = B^T A^T$.

3 a) Verify that the identity matrix is a projection.

b) Verify that the zero matrix is a projection.

c) Find two orthogonal projections P, Q such that $P + Q$ is not a projection.

d) Find two orthogonal projections P, Q such that PQ is not a projection.

Lecture 19: Data fitting

Last time we have derived the important formula

$$P = A(A^T A)^{-1} A^T.$$

which gives the projection matrix when projecting onto the image of a matrix A .

Given a system of linear equations $Ax = b$, the point $x = (A^T A)^{-1} A^T b$ is called the **least square solution** of the system.

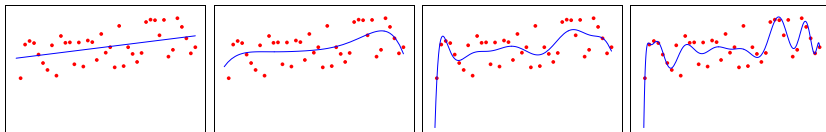
If A has no kernel, then the least square solution exists.

Proof. We know that if A has no kernel then the square matrix $A^T A$ has no kernel and is therefore invertible.

In applications we do not have to worry about this. In general, A is a $n \times m$ matrix where n is much larger than m meaning that we have **lots** of equations but few variables. Such matrices in general have a trivial kernel. For linear regression for example, it only appears if all data points are on a vertical axes like $(0, 2), (0, 6), (0, 0), (0, 4)$ and where any line $y = mx + 3$ is a least square solution. If in real life you should get into a situation where A has a kernel, you use the wrong model or have not enough few data.

If x is the least square solution of $Ax = b$ then Ax is the closest point on the image of A to b . The least square solution is the best solution of $Ax = b$ we can find. Since $Px = Ax$, it is the closest point to b on V . Our knowledge about kernel and the image of linear transformations helped us to derive this.

- 1 Finding the best polynomial which passes through a set of points is a **data fitting** problem. If we wanted to accommodate **all data**, the degree of the polynomial would become too large. The fit would look too wiggly. Taking a smaller degree polynomial will not only be more convenient but also give a better picture. Especially important is **regression**, the fitting of data with linear polynomials.

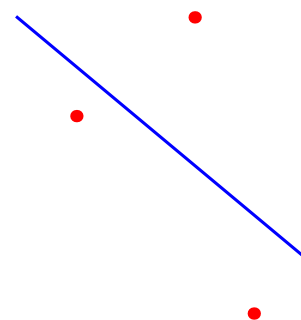


The above pictures show 30 data points which are fitted best with polynomials of degree 1, 6, 11 and 16. The first linear fit maybe tells most about the trend of the data.

- 2 The simplest fitting problem is fitting by lines. This is called linear regression. Find the best line $y = ax + b$ which fits the data

| x | y |
|----|----|
| -1 | 1 |
| 1 | 2 |
| 2 | -1 |

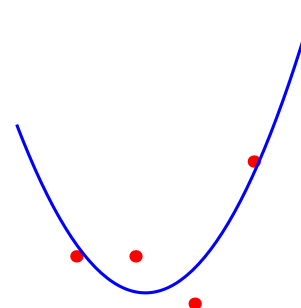
Solution. We will do this in class. The best solution is $y = -x/2 + 1$.



- 3 Find the best parabola $y = ax^2 + bx + c$ which fits the points

| x | y |
|----|----|
| -1 | 8 |
| 0 | 8 |
| 1 | 4 |
| 2 | 16 |

We do this in class. The best solution is $f(x) = 3x^2 - x + 5$.



- 4 Find the function $y = f(x) = a \cos(\pi x) + b \sin(\pi x)$, which best fits the data

| x | y |
|-----|---|
| 0 | 1 |
| 1/2 | 3 |
| 1 | 7 |

Solution: We have to find the least square solution to the system of equations

$$\begin{aligned} 1a + 0b &= 1 \\ 0a + 1b &= 3 \\ -1a + 0b &= 7 \end{aligned}$$

which is in matrix form written as $A\vec{x} = \vec{b}$ with

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}, \vec{b} = \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix}.$$

Now $A^T\vec{b} = \begin{bmatrix} -6 \\ 3 \end{bmatrix}$ and $A^T A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ and $(A^T A)^{-1} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix}$ and $(A^T A)^{-1} A^T \vec{b}$ is $\begin{bmatrix} -3 \\ 3 \end{bmatrix}$. The best fit is the function $f(x) = -3\cos(\pi x) + 3\sin(\pi x)$.

5 Find the circle $a(x^2 + y^2) + b(x + y) = 1$ which best fits the data

| x | y |
|-----|-----|
| 0 | 1 |
| -1 | 0 |
| 1 | -1 |
| 1 | 1 |

In other words, find the least square solution for the system of equations for the unknowns a, b which aims to have all 4 data points (x_i, y_i) on the circle. To get system of linear equations $Ax = b$, plug in the data

$$\begin{aligned} 11a + b &= 1 \\ a - b &= 1 \\ 2a &= 1 \\ 2a + 2b &= 1. \end{aligned}$$

This can be written as $Ax = b$, where

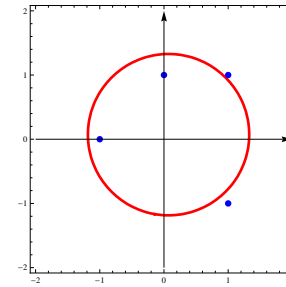
$$A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 2 & 0 \\ 2 & 2 \end{bmatrix}, b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

We get the least square solution with the usual formula. First compute

$$(A^T A)^{-1} = \begin{bmatrix} 3 & -2 \\ -2 & 5 \end{bmatrix} / 22$$

and then

$$A^T b = \begin{bmatrix} 6 \\ 2 \end{bmatrix},$$



Homework due March 23, 2011

1 Find the function $y = f(x) = ax^2 + bx^3$, which best fits the data

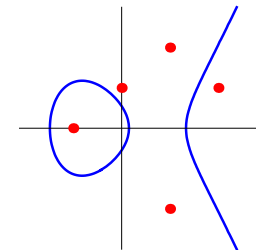
| x | y |
|-----|-----|
| -1 | 1 |
| 1 | 3 |
| 0 | 10 |

2 A curve of the form

$$y^2 = x^3 + ax + b$$

is called an **elliptic curve** in Weierstrass form. Elliptic curves are important in cryptography. Use data fitting to find the best parameters (a, b) for an elliptic curve given the following points:

$$\begin{aligned} (x_1, y_1) &= (1, 2) \\ (x_2, y_2) &= (-1, 0) \\ (x_3, y_3) &= (2, 1) \\ (x_4, y_4) &= (0, 1) \end{aligned}$$



3 Find the function of the form

$$f(t) = a \sin(t) + b \cos(t) + c$$

which best fits the data points $(0, 0), (\pi, 1), (\pi/2, 2), (-\pi, 3)$.

Lecture 20: More data fitting

Last time, we saw how the geometric formula $P = A(A^T A)^{-1} A^T$ for the projection on the image of a matrix A allows us to fit data. Given a fitting problem, we write it as a system of linear equations

$$Ax = b.$$

While this system is not solvable in general, we can look for the point on the image of A which is closest to b . This is the "best possible choice" of a solution and called the **least square solution**:

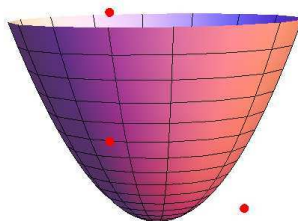
The vector $x = (A^T A)^{-1} A^T b$ is the **least square solution** of the system $Ax = b$.

The most popular example of a data fitting problem is **linear regression**. Here we have data points (x_i, y_i) and want to find the best line $y = ax + b$ which fits these data. But data fitting can be done with any finite set of functions. Data fitting can be done in higher dimensions too. We can for example look for the best surface fit through a given set of points (x_i, y_i, z_i) in space. Also here, we find the least square solution of the corresponding system $Ax = b$ which is obtained by assuming all points to be on the surface.

1 Which paraboloid $ax^2 + by^2 = z$ best fits the data

| x | y | z |
|----|----|---|
| 0 | 1 | 2 |
| -1 | 0 | 4 |
| 1 | -1 | 3 |

In other words, find the least square solution for the system of equations for the unknowns a, b which aims to have all data points on the paraboloid.



Solution: We have to find the least square solution to the system of equations

$$\begin{aligned} a \cdot 0 + b \cdot 1 &= 2 \\ a \cdot 1 + b \cdot 0 &= 4 \\ a \cdot 1 + b \cdot 1 &= 3. \end{aligned}$$

In matrix form this can be written as $A\vec{x} = \vec{b}$ with

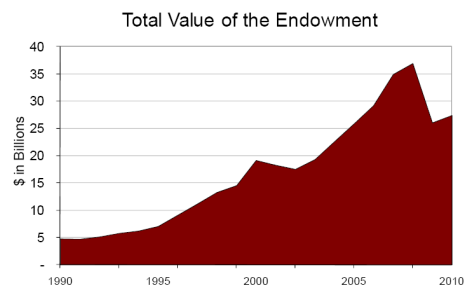
$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}, \vec{b} = \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix}.$$

We have $A^T A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ and $A^T b = \begin{bmatrix} 7 \\ 5 \end{bmatrix}$. We get the least square solution with the formula

$$x = (A^T A)^{-1} A^T b = \begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

The best fit is the function $f(x, y) = 3x^2 + y^2$ which produces an elliptic paraboloid.

2



A graphic from the Harvard Management Company Endowment Report of October 2010 is shown to the left. Assume we want to fit the growth using functions $1, x, x^2$ and assume the years are numbered starting with 1990. What is the best parabola $a + bx + cx^2 = y$ which fits these data?

| quintennium | endowment in billions |
|-------------|-----------------------|
| 1 | 5 |
| 2 | 7 |
| 3 | 18 |
| 4 | 25 |
| 5 | 27 |

We solved this example in class with linear regression. We saw that the best fit. With a quadratic fit, we get the system $A\vec{x} = \vec{b}$ with

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix}, \quad \vec{b} = \begin{bmatrix} 5 \\ 7 \\ 18 \\ 25 \\ 27 \end{bmatrix}.$$

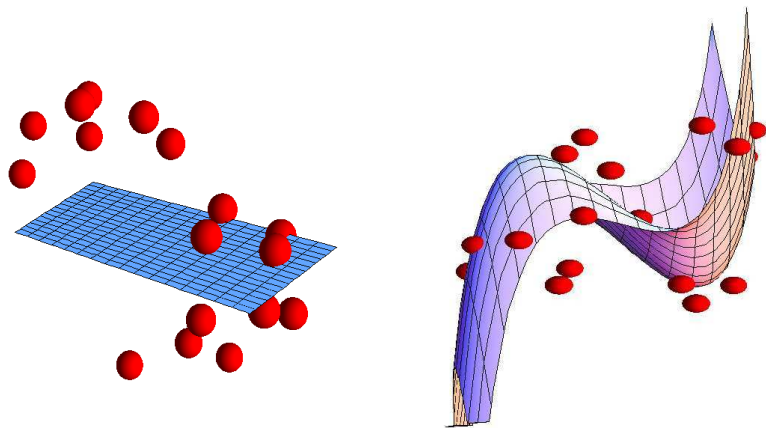
The solution vector $\vec{x} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} -21/5 \\ 277/35 \\ -2/7 \end{bmatrix}$ which indicates strong linear growth but some slow down.

- 3 Here is a problem on data analysis from a website. We collect some data from users but not everybody fills in all the data

| | | | | | | | | | | |
|----------|---|---|---|---|---|---|---|---|---|---|
| Person 1 | 3 | 5 | - | 3 | 9 | - | - | - | 2 | 9 |
| Person 2 | 4 | - | - | 8 | - | 5 | 6 | 2 | - | 9 |
| Person 3 | - | 4 | 2 | 5 | 7 | - | 1 | 9 | 8 | - |
| Person 4 | 1 | - | - | - | - | - | - | - | - | - |

It is difficult to do statistic with this. One possibility is to filter out all data from people who do not fulfill a minimal requirement. Person 4 for example did not do the survey seriously enough. We would throw this data away. Now, one could sort the data according to some important row. After that one could fit the data with a function $f(x, y)$ of two variables. This function could be used to fill in the missing data. After that, we would go and seek correlations between different rows.

Whenever doing data reduction like this, one must always compare different scenarios and investigate how much the outcome changes when changing the data.



The left picture shows a linear fit of the above data. The second picture shows a fit with cubic functions.

Homework due March 23, 2011

- 1 Here is an example of a fitting problem, where the solution is not unique:

| x | y |
|---|---|
| 0 | 1 |
| 0 | 2 |
| 0 | 3 |

Write down the corresponding fitting problem for linear functions $f(x) = ax + b = y$. What is going wrong?

- 2 If we fit data with a polynomial of the form $y = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots + a_nx^n$. How many data points $(x_1, y_1), \dots, (x_m, y_m)$ do you expect to fit exactly if the points x_1, x_2, \dots, x_m are all different?
- 3 The first 6 prime numbers 2, 3, 5, 7, 11 define the data points $(1, 2), (2, 3), (3, 5), (5, 7), (6, 11)$ in the plane. Find the best parabola of the form $y = ax^2 + c$ which fits these data.

Lecture 21: Midterm checklist

Probability theory

- ☐ **Probability space** $(\Omega, \mathcal{A}, P) = (\text{laboratory}, \text{events}, \text{probability measure})$
- ☐ **Random variable** A function from Ω to the reals.
- ☐ **Data** $(x_1, x_2, \dots, x_n) = \text{Vector} \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix} = \text{random variables over } \Omega = \{1, \dots, n\}$
- ☐ **Event A or B** the intersection $A \cap B$.
- ☐ **Event A and B** the union $A \cup B$.
- ☐ **Not the event A** the complement $\Omega \setminus A$.
- ☐ **Event A under the condition B.** $P[A|B] = P[A \cap B]/P[B]$
- ☐ **Independent events** $P[A \cap B] = P[A] \cdot P[B]$.
- ☐ **Independent random variables** $\{X \in [a, b]\}, \{Y \in [c, d]\}$ are independent events.
- ☐ **Independence and correlation** Independent random variables are uncorrelated.
- ☐ **Expectation** $E[X] = \sum_{\omega} X(\omega) = \sum_{x_i} x_i P[X = x_i]$
- ☐ **Variance** $\text{Var}[X] = E[X^2] - E[X]^2$.
- ☐ **Standard deviation** $\sigma[X] = \sqrt{\text{Var}[X]}$.
- ☐ **Covariance** $\text{Cov}[X, Y] = E[XY] - E[X] \cdot E[Y]$.
- ☐ **Correlation** $\text{Corr}[X, Y] = \text{Cov}[X, Y]/(\sigma[X]\sigma[Y])$.
- ☐ **Uncorrelated** $\text{Corr}[X, Y] = 0$.
- ☐ **Variance formula** $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$.
- ☐ **Pythagoras** $\text{Var}[X] + \text{Var}[Y] = \text{Var}[X + Y]$ for uncorrelated random variables.
- ☐ **Bayes formula** $P[A|B] = \frac{P[B|A] \cdot P[A]}{P[B|A] + P[B|A^c]}$.
- ☐ **Bayes rule** $P[A_i|B] = \frac{P[B|A_i] \cdot P[A_i]}{\sum_{j=1}^n P[B|A_j] \cdot P[A_j]}$.
- ☐ **Permutations** $n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$ possibilities.
- ☐ **Combinations** n^k possibilities to select from n and put back.
- ☐ **Ordered selection** Choose k from n with order $n!/(n-k)!$
- ☐ **Unordered selection** Choose k from n gives $\binom{n}{k} = n!/(k!(n-k)!)$
- ☐ **Binomial distribution** $P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$.
- ☐ **Expectation of binomial distribution** pn .
- ☐ **Variance of binomial distribution** $p(1-p)n$.
- ☐ **For 0-1 data.** The expectation determines the variance.

Linear algebra

- ☐ **Matrix A** is a $n \times m$ matrix, it has m columns and n rows, maps \mathbf{R}^m to \mathbf{R}^n .
- ☐ **Square matrix** $n \times n$ matrix, maps \mathbf{R}^n to \mathbf{R}^n .
- ☐ **Identity matrix** the diagonal matrix I_n satisfies $I_n v = v$ for all vectors v .
- ☐ **Column Vector** $n \times 1$ matrix = column vector
- ☐ **Row Vector** $1 \times n$ matrix = row vector.
- ☐ **Linear transformation** $\vec{x} \mapsto A\vec{x}$, $T(\vec{x} + \vec{y}) = T(\vec{x}) + T(\vec{y})$, $T(\lambda\vec{x}) = \lambda T(\vec{x})$.
- ☐ **Column vectors** of A are images of standard basis vectors $\vec{e}_1, \dots, \vec{e}_n$.
- ☐ **Linear system of equations** $A\vec{x} = \vec{b}$, have n equations, m unknowns.
- ☐ **Consistent system** $A\vec{x} = \vec{b}$: there is at least one solution \vec{x} .
- ☐ **Vector form of linear equation** $x_1 \vec{v}_1 + \dots + x_n \vec{v}_n = \vec{b}$, \vec{v}_i columns of A .
- ☐ **Matrix form of linear equation** $\vec{w}_i \cdot \vec{x} = b_i$, \vec{w}_i rows of A .
- ☐ **Augmented matrix** of $A\vec{x} = \vec{b}$ is the matrix $[A|b]$ which has one column more as A .
- ☐ **Coefficient matrix** of $A\vec{x} = \vec{b}$ is the matrix A .
- ☐ **Matrix multiplication** $[AB]_{ij} = \sum_k A_{ik} B_{kj}$, dot i -th row with j 'th column.
- ☐ **Gauss-Jordan elimination** $A \rightarrow \text{rref}(A)$ in row reduced echelon form.
- ☐ **Gauss-Jordan elimination steps** SSS: Swapping, Scaling, Subtracting rows.
- ☐ **Leading one** First nonzero entry in a row is equal to 1. Write $\boxed{1}$.
- ☐ **Row reduced echelon form** (1) nonzero row has $\boxed{1}$, (2) columns with $\boxed{1}$ are zero except at $\boxed{1}$, (3) every row above row with $\boxed{1}$ has $\boxed{1}$ to the left.
- ☐ **Pivot column** column with $\boxed{1}$ in $\text{rref}(A)$.
- ☐ **Redundant column** column with no $\boxed{1}$ in $\text{rref}(A)$.
- ☐ **Rank of matrix A** number of $\boxed{1}$ in $\text{rref}(A)$. It is equal to $\dim(\text{im}(A))$.
- ☐ **Nulley of matrix A:** is defined as $\dim(\ker(A))$.
- ☐ **Kernel of matrix** $\{\vec{x} \in \mathbf{R}^n, A\vec{x} = \vec{0}\}$.
- ☐ **Image of matrix** $\{A\vec{x}, \vec{x} \in \mathbf{R}^n\}$.
- ☐ **Inverse transformation of T** A transformation satisfying $S(T(x)) = x = T(S(x))$.
- ☐ **Inverse matrix of A** Matrix $B = A^{-1}$ satisfies $AB = BA = I_n$
- ☐ **Rotation in plane** $A = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}$, rotate counter-clock by α .
- ☐ **Dilation in plane** $\vec{x} \mapsto \lambda\vec{x}$, also called scaling. Given by diagonal $A = I_2$
- ☐ **Rotation-Dilation** $A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$. Scale by $\sqrt{a^2 + b^2}$, rotate by $\arctan(b/a)$.
- ☐ **Reflection-Dilation** $A = \begin{bmatrix} a & b \\ b & -a \end{bmatrix}$. Scale by $\sqrt{a^2 + b^2}$, reflect at line w, slope b/a .
- ☐ **Horizontal and vertical shear** $\vec{x} \mapsto A\vec{x}$, $A = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}$, $\vec{x} \mapsto A\vec{x}$, $A = \begin{bmatrix} 1 & 0 \\ b & 1 \end{bmatrix}$.
- ☐ **Reflection about line** $\vec{x} \mapsto A\vec{x}$, $A = \begin{bmatrix} \cos(2\alpha) & \sin(2\alpha) \\ \sin(2\alpha) & -\cos(2\alpha) \end{bmatrix}$.
- ☐ **Projection onto line** containing unit vector u : $A = \begin{bmatrix} u_1 u_1 & u_1 u_2 \\ u_2 u_1 & u_2 u_2 \end{bmatrix}$.

Linear subspace check $\vec{0} \in X$, $\vec{x}, \vec{y} \in X, \lambda \in \mathbf{R} \Rightarrow \vec{x} + \vec{y} \in X, \lambda \vec{x} \in X$.

$\mathcal{B} = \{\vec{v}_1, \dots, \vec{v}_n\}$ **span** X : Every $\vec{x} \in X$ can be written as $\vec{x} = a_1 \vec{v}_1 + \dots + a_n \vec{v}_n$.

$\mathcal{B} = \{\vec{v}_1, \dots, \vec{v}_n\}$ **linear independent** X : $\sum_i a_i \vec{v}_i = \vec{0}$ implies $a_1 = \dots = a_n = 0$.

$\mathcal{B} = \{\vec{v}_1, \dots, \vec{v}_n\}$ **basis in** X : linear independent in X and span X .

Dimension of linear space X : number of basis elements of a basis in X .

S-matrix Coordinate transformation matrix containing basis vectors as columns.

\mathcal{B} **coordinates** $[\vec{v}]_{\mathcal{B}} = S^{-1} \vec{v}$, where $S = [\vec{v}_1, \dots, \vec{v}_n]$ contains basis vectors \vec{v}_i as columns.

\mathcal{B} **matrix** of T in basis \mathcal{B} . The matrix is $B = S^{-1}AS$.

A **similar to** B : defined as $B = S^{-1}AS$. We write $A \sim B$.

Row reduction SSS: scale rows, swap rows and subtract row from other row.

Row reduced echelon form is a matrix in row reduced echelon form?

Matrix-Transformation The columns of A are the images of the basis vectors.

Kernel-Image Compute the kernel and the image by row reduction.

System of linear equations Solve a system of linear equation by row reduction.

How many solutions Are there 0, 1, ∞ solutions? $\text{rank}(A), \text{rank}[A, b]$ matter.

Similar? Check whether B^n, A^n are similar. Both invertible or not. Possibly find S .

Linear space 0 is in V , $v + w$ is in V and λv is in V .

Linear transformation is a given transformation linear or not?

Space orthogonal to given space write as row space of a matrix and find kernel.

Number of solutions. A linear system of equations has either exactly 0, 1 or ∞ many solutions.

Solve system Row reduce $[A|b]$ to get $[I_n|x]$ with solution x .

Vectors perpendicular to a set of vectors, get kernel of matrix which contains vectors as rows.

Rank-nullity theorem $\dim(\ker(A)) + \dim(\text{im}(A)) = m$, where A is $n \times m$ matrix.

Number of basis elements is independent of basis. Is equal to dimension.

Basis of image of A pivot columns of A form a basis of the image of A .

Basis of kernel of A introduce free variables for each redundant column of A .

Inverse of 2×2 **matrix** switch diagonal, change sign of wings and divide by det.

Inverse of $n \times n$ **matrix** Row reduce $[A|I_n]$ to get $[I_n|A^{-1}]$.

Matrix algebra $(AB)^{-1} = B^{-1}A^{-1}$, $A(B+C) = AB+AC$, etc. $AB \neq BA$ i.g.

Invertible $\Leftrightarrow \text{rref}(A) = I_n \Leftrightarrow$ columns form basis $\Leftrightarrow \text{rank}(A) = n, \Leftrightarrow \text{nullity}(A) = 0$.

Similarity properties: $A \sim B$ implies $A^n \sim B^n$. If A is invertible, B is invertible.

Orthogonal vectors $\vec{v} \cdot \vec{w} = 0$.

length $||\vec{v}|| = \sqrt{\vec{v} \cdot \vec{v}}$, **unit vector** \vec{v} with $||\vec{v}|| = \sqrt{\vec{v} \cdot \vec{v}} = 1$.

Orthogonal basis basis such that v_1, \dots, v_n are pairwise orthogonal, and length 1.

Orthogonal complement of \mathbf{V} $V^\perp = \{v|v \text{ perpendicular to } V\}$.

Projection onto \mathbf{V} orth. basis $P = QQ^T$ if Q has orthonormal columns.

Orthogonal projection onto V is $A(A^T A)^{-1}A^T$.

Least square solution of $A\vec{x} = \vec{b}$ is $\vec{x}_* = (A^T A)^{-1}A^T \vec{b}$.

Data fitting Find least square solution of equations when data are fitted exactly.

Lecture 22: Distributions

A random variable is a function from a probability space Ω to the real line R . There are two important classes of random variables:

1) For **discrete random variables**, the random variable X takes a discrete set of values. This means that the random variable takes values x_k and the probabilities $P[X = x_k] = p_k$ add up to 1.

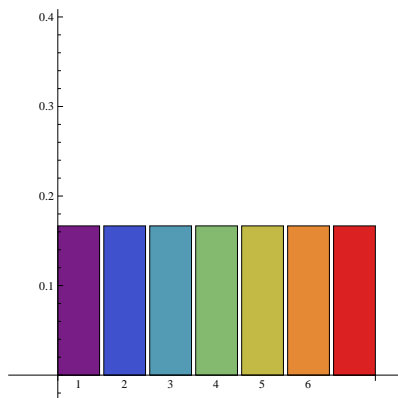
2) For **continuous random variables**, there is a probability density function $f(x)$ such that $P[X \in [a, b]] = \int_a^b f(x) dx$ and $\int_{-\infty}^{\infty} f(x) dx = 1$.

Discrete distributions

- 1 We throw a dice and assume that each side appears with the same probability. The random variable X which gives the number of eyes satisfies

$$P[X = k] = \frac{1}{6}.$$

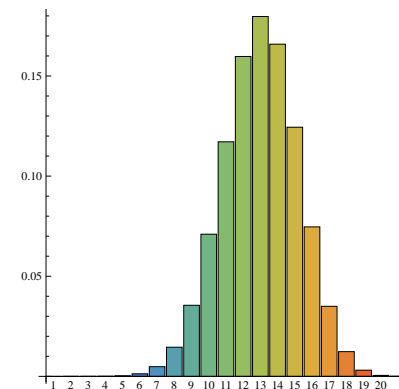
This is a discrete distribution called the **uniform distribution** on the set $\{1, 2, 3, 4, 5, 6\}$.



- 2 Throw n coins for which head appears with probability p . Let X denote the number of heads. This random variable takes values $0, 1, 2, \dots, n$ and

$$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

This is the Binomial distribution we know.



- 3 The probability distribution on $N = \{0, 1, 2, \dots\}$

$$P[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}$$

is called the **Poisson distribution**. It is used to describe the number of radioactive decays in a sample, the number of newborns with a certain defect. You show in the homework that the mean and standard deviation is λ . Poisson distribution is the most important distribution on $\{0, 1, 2, 3, \dots\}$. It is a limiting case of Binomial distributions

- 4 An epidemiology example from Cliffs notes: the UHS sees $X=10$ pneumonia cases each winter. Assuming independence and unchanged conditions, what is the probability of there being 20 cases of pneumonia this winter? We use the Poisson distribution with $\lambda = 10$ to see $P[X = 20] = 10^{20} e^{-10} / 20! = 0.0018$.

The Poisson distribution is the $n \rightarrow \infty$ limit of the binomial distribution if we chose for each n the probability p such that $\lambda = np$ is fixed.

Proof. Setting $p = \lambda/n$ gives

$$\begin{aligned} P[X = k] &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \left(\frac{n!}{n^k (n-k)!}\right) \left(\frac{\lambda^k}{k!}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^k. \end{aligned}$$

This is a product of four factors. The first factor $n(n-1)\dots(n-k+1)/n^k$ converges to 1. Leave the second factor $\lambda^k/k!$ as it is. The third factor converges to $e^{-\lambda}$ by the definition of the exponential. The last factor $(1 - \frac{\lambda}{n})^k$ converges to 1 for $n \rightarrow \infty$ since k is kept fixed. We see that $P[X = k] \rightarrow \lambda^k e^{-\lambda} / k!$.

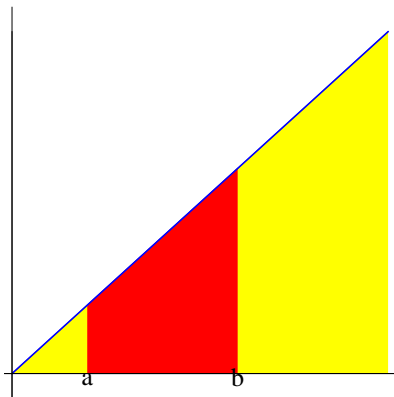
Continuous distributions

5 A random variable is called a **uniform distribution** if $P[X \in [a, b]] = (b - a)$ for all $0 \leq a < b \leq 1$. We can realize this random variable on the probability space $\Omega = [a, b]$ with the function $X(x) = x$, where $P[I]$ is the length of an interval I . The uniform distribution is the most natural distribution on a finite interval.

6 The random variable X on $[0, 1]$ where $P[[a, b]] = b - a$ is given by $X(x) = \sqrt{x}$. We have

$$P[X \in [a, b]] = P[\sqrt{x} \in [a, b]] = P[x \in [a^2, b^2]] = b^2 - a^2.$$

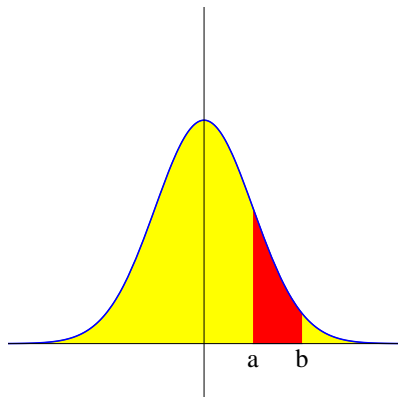
We have $f(x) = 2x$ because $\int_a^b f(x) dx = x^2|_a^b = b^2 - a^2$. The function $f(x)$ is the probability density function of the random variable.



7 A random variable with normal distribution with mean 1 and standard deviation 1 has the probability density

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

This is an example of a continuous distribution. The normal distribution is the most natural distribution on the real line.



8 The probability density function

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

is called the **exponential distribution**. The exponential distribution is the most natural distribution on the positive real line.

Remark. The statements "most natural" can be made more precise. Given a subset $X \subset \mathbb{R}$ of the real line and the mean and standard deviation we can look at the distribution f on X for which the **entropy** $-\int_X f \log(f) dx$ is maximal. The uniform, exponential and normal distributions extremize entropy on an interval, half line or real line. The reason why they appear so often is that adding independent random variables increases entropy. If different processes influence an experiment then the entropy becomes large. Nature tries to maximize entropy. That's why these distributions are "natural".

9 The distribution on the positive real axis with the density function

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

is called the **log normal distribution** with mean m . Examples of quantities which have log normal distribution is the size of a living tissue like length or height of a population or the size of cities. An other example is the **blood pressure** of adult humans. A quantity which has a log normal distribution is a quantity which has a logarithm which is normally distributed.

Homework due March 30, 2011

- 1 a) Find the mean of the exponential distribution.
b) Find the variance and standard deviation of the exponential distribution.
c) Find the **entropy** $-\int_0^\infty f(x) \log(f(x)) dx$ in the case $\lambda = 1$.

2 Here is a special case of the **Students t distribution**

$$f(x) = \frac{2}{\pi} (1 + x^2)^{-2}.$$

- a) Verify that it is a probability distribution.
b) Find the mean. (No computation needed, just look at the symmetry).
c) Find the standard deviation.

To compute the integrals in a),c), you can of course use a computer algebra system if needed.

- 3 a) Verify that the Poisson distribution is a probability distribution: $\sum_{k=0}^\infty P[X = k] = 1$.
b) Find the mean $m = \sum_{k=0}^\infty k P[X = k]$.
c) Find the standard deviation $\sum_{k=0}^\infty (k - m)^2 P[X = k]$.

Lecture 23: Chebychev theorem

In this lecture we look at more probability distributions and prove the fantastically useful Chebychev's theorem.

Remember that a **continuous probability density** is a nonnegative function f such that $\int_{\mathbb{R}} f(x) dx = 1$. A random variable X has this probability density if

$$P[X \in [a, b]] = \int_a^b f(x) dx$$

for all intervals $[a, b]$.

If we know the probability density of a random variable, we can compute all the important quantities like the expectation or the variance.

If X has the probability density f , then $m = E[X] = \int x f(x) dx$ and $\text{Var}[X] = \int (x - m)^2 f(x) dx$.

The **distribution function** of a random variable with probability density f is defined as

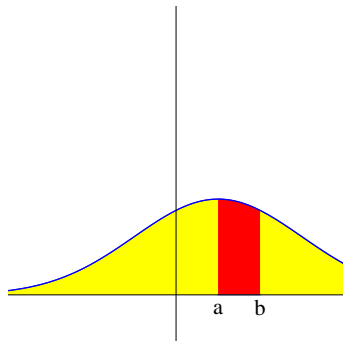
$$F(s) = \int_{-\infty}^s f(x) dx = P[X \leq s] .$$

By definition F is a monotone function: $F(b) \geq F(a)$ for $b \geq a$. One abbreviates the probability density function with *PDF* and the distribution function with *CDF* which abbreviates cumulative distribution function.

- 1 The most important distribution on the real line is the **normal distribution**

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

It has mean m and standard deviation σ . This is a probability measure because after a change of variables $y = (x - m)/(\sqrt{2}\sigma)$, the integral $\int_{-\infty}^{\infty} f(x) dx$ becomes $\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} dy = 1$.



- 2 The most important distribution on the positive real line is the **exponential distribution**

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

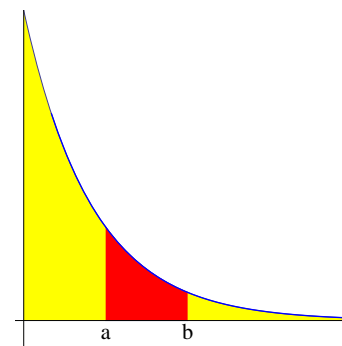
Lets compute its mean:

$$m = \int_0^{\infty} x f(x) dx = \frac{1}{\lambda} .$$

From $\lambda \int_0^{\infty} x^2 \exp(-\lambda x) dx = 2/\lambda^2$, we get the variance

$$2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$$

and the standard deviation $1/\lambda$.



- 3 The most important distribution on a finite interval $[a, b]$ is the **uniform distribution**

$$f(x) = 1_{[a,b]} \frac{1}{b-a} ,$$

where 1_I is the characteristic function

$$1_I(x) = \begin{cases} 1 & x \in I \\ 0 & x \notin I \end{cases} .$$

The following theorem is very important for estimation purposes. Despite the simplicity of its proof it has a lot of applications:

Chebychev theorem If X is a random variable with finite variance, then

$$P[|X - E[X]| \geq c] \leq \frac{\text{Var}[X]}{c^2} .$$

Proof. The random variable $Y = X - E[X]$ has zero mean and the same variance. We need only to show $P[|Y| \geq c] \leq \frac{\text{Var}[Y]}{c^2}$. Taking the expectation of the inequality

$$c^2 1_{\{|Y| \geq c\}} \leq Y^2$$

gives

$$c^2 P[|Y| \geq c] \leq E[Y^2] = \text{Var}[Y]$$

finishing the proof.

The theorem also gives more meaning to the notion "Variance" as a measure for the deviation from the mean. The following example is similar to the one section 11.6 of Cliff's notes:

- 4 A die is rolled 144 times. What is the probability to see 50 or more times the number 6 shows up? Let X be the random variable which counts the number of times, the number 6 appears. This random variable has a binomial distribution with $p = 1/6$ and $n = 144$. It has the expectation $E[X] = np = 144/6 = 24$ and the variance $\text{Var}[X] = np(1-p) = 20$. Setting $c = (50 - 24) = 26$ in Chebychev, we get $P[|X - 24| \geq 26] \leq 20/26^2 \sim 0.0296\dots$. The chance is smaller than 3 percent. The actual value $\sum_{k=50}^{144} \binom{144}{k} p^k (1-p)^{144-k} \sim 1.17 \cdot 10^{-7}$ is much smaller. Chebychev does not necessarily give good estimates, but it is a handy and universal "rule of thumb".

Finally, let's look at a practical application of the use of the cumulative distribution function. It is the task to **generate random variables with a given distribution**:

- 5 Assume we want to generate random variables X with a given distribution function F . Then $Y = F(X)$ has the uniform distribution on $[0, 1]$. We can reverse this. If we want to produce random variables with a distribution function F , just take a random variable Y with uniform distribution on $[0, 1]$ and define $X = F^{-1}(Y)$. This random variable has the distribution function F because $\{X \in [a, b]\} = \{F^{-1}(Y) \in [a, b]\} = \{Y \in F([a, b])\} = \{Y \in [F(a), F(b)]\} = F(b) - F(a)$. We see that we need only to have a random number generator which produces uniformly distributed random variables in $[0, 1]$ to get a random number generator for a given continuous distribution. A computer scientist implementing random processes on the computer only needs to have access to a random number generator producing uniformly distributed random numbers. The later are provided in **any** programming language which deserves this name.

To generate random variables with cumulative distribution function F , we produce random variables X with uniform distribution in $[0, 1]$ and form $Y = F^{-1}(X)$.

With computer algebra systems

- 1) In Mathematica, you can generate random variables with a certain distribution with a command like in the following example:

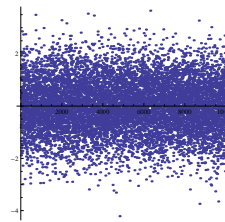
```
X:=Random[NormalDistribution[0,1]]
ListPlot[Table[X,{10000}],PlotRange->All]
```

- 2) Here is how to access the **probability density function** (PDF)

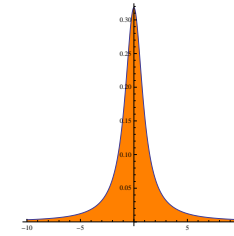
```
f=PDF[CauchyDistribution[0,1]];
S=Plot[f[x],{x,-10,10},PlotRange->All,Filling->Axis]
```

- 3) And the **cumulative probability distribution** (CDF)

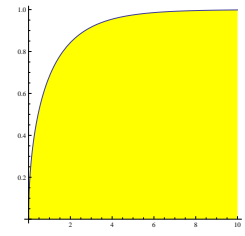
```
f=CDF[ChiSquareDistribution[1]];
S=Plot[f[x],{x,0,10},PlotRange->All,Filling->Bottom]
```



Random numbers



The PDF



The CDF

Homework due March 30, 2011

- The random variable X has a normal distribution with standard deviation 2 and mean 5. Estimate the probability that $|X - 5| > 3$.
- Estimate the probability of the event $X > 10$ for a Poisson distributed random variable X with mean 4.
- Verify that $\phi(x) = \tan(x\pi)$ maps the interval $[0, 1]$ onto the real line so that its inverse $F(y) = \arctan(y)/\pi$ is a map from \mathbb{R} to $[0, 1]$.
 - Show that $f = F'(y) = \frac{1}{\pi} \frac{1}{1+y^2}$.
 - Assume we have random numbers in $[0, 1]$ handy and want to random variables which have the probability density f . How do we achieve this?
 - The mean $\int_{-\infty}^{\infty} xf(x) dx$ does not exist as an indefinite integral but can be assigned the value 0 by taking the limit $\int_{-R}^R xf(x) dx = 0$ for $R \rightarrow \infty$. Is it possible to assign a value to the variance $\int_{-\infty}^{\infty} x^2 f(x) dx$?

The probability distribution with density

$$\frac{1}{\pi} \frac{1}{1+y^2}$$

which appeared in this homework problem is called the **Cauchy distribution**. Physicists call it the **Cauchy-Lorentz distribution**.

Why is the Cauchy distribution natural? As one can deduce from the homework, if you chose a random point P on the unit circle, then the slope of the line OP has a Cauchy distribution. Instead of the circle, we can take a rotationally symmetric probability distribution like the Gaussian with probability measure $P[A] = \int_A e^{-x^2-y^2}/\pi dx dy$ on the plane. Random points can be written as (X, Y) where both X, Y have the normal distribution with density $e^{-x^2}/\sqrt{\pi}$. We have just shown

If we take independent Gaussian random variables X, Y of zero mean and with the same variance and form $Z = X/Y$, then the random variable Z has the Cauchy distribution.

Now, it becomes clear why the distribution appears so often. Comparing quantities is often done by looking at their ratio X/Y . Since the normal distribution is so prevalent, there is no surprise, that the Cauchy distribution also appears so often in applications.

Lecture 24: Determinants

In this lecture, we define the determinant for a general $n \times n$ matrix and look at the Laplace expansion method to compute them. A determinant attaches a number to a square matrix, which determines a lot about the matrix, like whether the matrix is invertible.

The 2×2 case

The determinant of a 2×2 matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is defined as $\det(A) = ad - bc$.

We have seen that this is useful for the inverse:

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

This formula shows:

A 2×2 matrix A is invertible if and only if $\det(A) \neq 0$. The determinant determines the invertibility of A .

$$1 \quad \det \begin{pmatrix} 5 & 4 \\ 2 & 1 \end{pmatrix} = 5 \cdot 1 - 4 \cdot 2 = -3.$$

We also see already that the determinant changes sign if we flip rows, that the determinant is linear in each of the rows.

We can write the formula as a sum over all permutations of $1, 2$. The first permutation $\pi = (1, 2)$ gives the sum $A_{1,\pi(1)}A_{2,\pi(2)} = A_{1,1}A_{2,2} = ad$ and the second permutation $\pi = (2, 1)$ gives the sum $A_{1,\pi(1)}A_{2,\pi(2)} = A_{1,2}A_{2,1} = bc$. The second permutation has $|\pi|$ upcrossing and the sign $(-1)^{|\pi|} = -1$. We can write the above formula as $\sum_{\pi} (-1)^{|\pi|} A_{1\pi(1)}A_{2\pi(2)}$.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} - \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

The 3×3 case

The determinant of a 3×3 matrix

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

is defined as $aei + bfg + cdh - ceg - bdi - afh$.

We can write this as a sum over all permutations of $\{1, 2, 3\}$. Each permutation produces a "pattern" along we multiply the matrix entries. The patterns π with an even number $|\pi|$ of upcrossings are taken with a positive sign the other with a negative sign.

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} + \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} + \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} - \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} - \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} - \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

$$2 \quad \det \begin{pmatrix} 2 & 0 & 4 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} = 2 - 4 = -2.$$

The general definition

A **permutation** is an invertible transformation $\{1, 2, 3, \dots, n\}$ onto itself. We can visualize it using the **permutation matrix** which is everywhere 0 except $A_{i,\pi(i)} = 1$.

There are $n! = n(n-1)(n-2) \cdots 2 \cdot 1$ permutations.

3 For $\pi = (6, 4, 2, 5, 3, 1)$ if $\pi(1) = 6, \pi(2) = 4, \pi(3) = 2, \pi(4) = 5, \pi(5) = 3, \pi(6) = 1$ we have the permutation matrix

$$P_{\pi} = \begin{bmatrix} & & & & & 1 \\ & & & 1 & & \\ & 1 & & & & \\ & & & & 1 & \\ & & 1 & & & \\ 1 & & & & & \end{bmatrix}.$$

It has $|\pi| = 5 + 2 + 3 + 1 + 1 = 12$ up-crossings. The determinant of a matrix which has everywhere zeros except $A_{i\pi(j)} = 1$ is the number $(-1)^{|\pi|}$ which is called the **sign** of the permutation.

The **determinant** of a $n \times n$ matrix A is defined as the sum

$$\sum_{\pi} (-1)^{|\pi|} A_{1\pi(1)} A_{2\pi(2)} \cdots A_{n\pi(n)},$$

where π is a permutation of $\{1, 2, \dots, n\}$ and $|\pi|$ is the number of up-crossings.

4

$$\det(A) = \det \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 7 & 0 \\ 0 & 0 & 11 & 0 & 0 & 0 \\ 13 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} = 2 * 3 * 5 * 7 * 11 * 13.$$

The determinant of an upper triangular matrix or lower triangular matrix is the product of the diagonal entries.

Laplace expansion

This Laplace expansion is a convenient way to sum over all permutations. We group the permutations by taking first all the ones where the first entry is 1, then the one where the first entry is 2 etc. In that case we have a permutation of $(n-1)$ elements. the sum over these entries produces a determinant of a smaller matrix.

For each entry a_{j1} in the first column form the $(n-1) \times (n-1)$ matrix B_{j1} which does not contain the first and j 'th row. The determinant of B_{j1} is called a **minor**.

Laplace expansion $\det(A) = (-1)^{1+1}A_{11}\det(B_{11}) + \dots + (-1)^{1+n}A_{n1}\det(B_{n1})$

5 Find the determinant of

$$\begin{bmatrix} 0 & 0 & 7 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

We have two nonzero entries in the first column.

$$\begin{aligned} \det(A) &= (-1)^{2+1}8\det \begin{bmatrix} 0 & 7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 2 & 0 & 0 & 0 & 0 \end{bmatrix} + (-1)^{4+1}3\det \begin{bmatrix} 0 & 7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 5 & 0 \\ 2 & 0 & 0 & 0 & 0 \end{bmatrix} \\ &= -8(2*7*1*5*1) + -3(2*7*0*5*1) = -560 \end{aligned}$$

6 Find the determinant of

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

The answer is -1 .

7 Find the determinant of

$$A = \begin{bmatrix} 3 & 2 & 3 & 0 & 0 & 0 \\ 0 & 4 & 2 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Answer: 60.

Homework due April 6, 2011

1 Find the determinant of the following matrix

$$\begin{bmatrix} 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 \\ 1 & 0 & 3 & 4 & 7 & 1 \\ 2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 4 & 1 & 0 \\ 0 & 0 & 6 & 0 & 0 & 0 \end{bmatrix}.$$

2 Give the reason in terms of permutations why the determinant of a **partitioned matrix**

$\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ is the product $\det(A)\det(B)$.

$$\text{Example } \det \begin{pmatrix} 3 & 4 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 4 & -2 \\ 0 & 0 & 2 & 2 \end{pmatrix} = 2 \cdot 12 = 24.$$

3 Find the determinant of the diamond matrix:

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 8 & 8 & 8 & 0 & 0 & 0 \\ 0 & 0 & 8 & 8 & 8 & 8 & 8 & 0 & 0 \\ 0 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 0 \\ 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 \\ 0 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 0 \\ 0 & 0 & 8 & 8 & 8 & 8 & 8 & 0 & 0 \\ 0 & 0 & 0 & 8 & 8 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Hint. Do not compute too much. Investigate what happens with the determinant if you switch two rows.

Lecture 25: More Determinants

In this lecture, we learn about a faster method to compute the determinant of a $n \times n$ matrix. Summing over all possible permutations is often not efficient. For a 20×20 matrix, we would already have to sum over $20! = 2432902008176640000 \sim 2.4 \cdot 10^{18}$ entries. As a comparison, there are $4.3 \cdot 10^{17}$ seconds (≈ 13.7 billion years) since the big bang.

Linearity of the determinant

Lets take a general $n \times n$ matrix A . The following linearity property of determinants follows pretty much from the definition because for every pattern, the sum is right. The determinant is a sum over all patterns.

The determinant is linear in every row and every column.

Lets see what this means for rows. For columns it is similar.

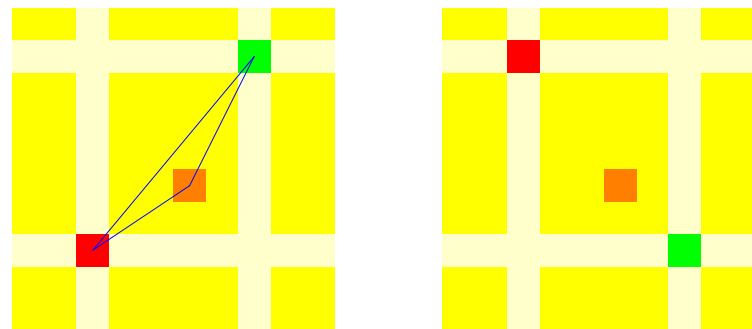
$$\begin{aligned} & \det \begin{pmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ v_1 & v_2 & v_3 & \dots & v_n \\ \dots & \dots & \dots & \dots & \dots \\ A_{n1} & A_{n2} & A_{n3} & \dots & A_{nn} \end{pmatrix} + \det \begin{pmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ w_1 & w_2 & w_3 & \dots & w_n \\ \dots & \dots & \dots & \dots & \dots \\ A_{n1} & A_{n2} & A_{n3} & \dots & A_{nn} \end{pmatrix} \\ &= \det \begin{pmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ (v+w)_1 & (v+w)_2 & (v+w)_3 & \dots & (v+w)_n \\ \dots & \dots & \dots & \dots & \dots \\ A_{n1} & A_{n2} & A_{n3} & \dots & A_{nn} \end{pmatrix} . \\ & \lambda \det \begin{pmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ v_1 & v_2 & v_3 & \dots & v_n \\ \dots & \dots & \dots & \dots & \dots \\ A_{n1} & A_{n2} & A_{n3} & \dots & A_{nn} \end{pmatrix} = \det \begin{pmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ \lambda v_1 & \lambda v_2 & \lambda v_3 & \dots & \lambda v_n \\ \dots & \dots & \dots & \dots & \dots \\ A_{n1} & A_{n2} & A_{n3} & \dots & A_{nn} \end{pmatrix} . \end{aligned}$$

Swapping two rows changes the sign of the determinant.

$$\det \begin{pmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ v_1 & v_2 & v_3 & \dots & v_n \\ \dots & \dots & \dots & \dots & \dots \\ w_1 & w_2 & w_3 & \dots & w_n \\ \dots & \dots & \dots & \dots & \dots \\ A_{n1} & A_{n2} & A_{n3} & \dots & A_{nn} \end{pmatrix} = -\det \begin{pmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ w_1 & w_2 & w_3 & \dots & w_n \\ \dots & \dots & \dots & \dots & \dots \\ v_1 & v_2 & v_3 & \dots & v_n \\ \dots & \dots & \dots & \dots & \dots \\ A_{n1} & A_{n2} & A_{n3} & \dots & A_{nn} \end{pmatrix} .$$

Proof. We have to show that the number of upcrossing changes by an odd number. Lets count the number of upcrossings before and after the switch. Assume row a and c are switched. We look at one pattern and assume that (a,b) be an entry on row a and (c,d) is an entry on row b . The entry (a,b) changes the number of upcrossings to (c,d) by 1 (there is one upcrossing from (a,b) to (c,d) before which is absent after).

For each entry (x,y) inside the rectangle $(a,c) \times (b,d)$, the number of upcrossings from and to (x,y) changes by two. (there are two upcrossings to and from the orange squares before which are absent after). For each entry outside the rectangle and different from $(a,b),(c,d)$, the number of upcrossings does not change.



It follows that if two rows are the same, then the determinant is zero.

Row reduction

We immediately get from the above properties what happens if we do row reduction. Subtracting a row from an other row does not change the determinant since by linearity we subtract the determinant of a matrix with two equal rows. Swapping two rows changes the sign and scaling a row scales the determinant.

If c_1, \dots, c_k are the row reduction scale factors and m is the number of row swaps during row reduction, then

$$\det(A) = \frac{(-1)^m}{c_1 \cdots c_k} \det(rref(A)) .$$

Since row reduction is fast, we can compute the determinant of a 20×20 matrix in a jiffy. It takes about 400 operations and thats nothing for a computer.

1

$$\begin{bmatrix} 4 & 1 & 1 & 1 \\ 4 & 2 & 2 & 2 \\ 4 & 1 & 6 & 3 \\ 4 & 1 & 1 & 7 \end{bmatrix} .$$

Row reduce.

- 2 Compute the following determinant.

$$\det \begin{bmatrix} 0 & 2 & 5 & 6 \\ 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 4 \\ 0 & 0 & 3 & 2 \end{bmatrix}.$$

We could use the Laplace expansion or see that there is only one pattern. The simplest way however is to swap two rows to get an upper triangular matrix

$$\det \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 5 & 6 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 4 \end{bmatrix} = 24.$$

- 3 The determinant of

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

is -1 because swapping the last row to the first row gives the identity matrix. Alternatively we could see that this permutation matrix has 5 upcrossings and that the determinant is -1 .

A matrix is invertible if and only if $\det(A) \neq 0$.

Product of matrices

One of the main reasons why determinants are interesting is because of the following property

$$\det(A \cdot B) = \det(A)\det(B)$$

Proof. One can bring the $n \times n$ matrix $[A|AB]$ into row reduced echelon form. Similar than the augmented matrix $[A|b]$ was brought into the form $[1|A^{-1}b]$, we end up with $[1|A^{-1}AB] = [1|B]$. By looking at the $n \times n$ matrix to the left during the Gauss-Jordan elimination process, the determinant has changed by a factor $\det(A)$. We end up with a matrix B which has determinant $\det(B)$. Therefore, $\det(AB) = \det(A)\det(B)$.

$$\det(A^T) = \det(A)$$

Proof. Every upcrossing is a pair of entries A_{ij}, A_{kl} where $k > i, l > j$. If we look at the transpose, this pair of entries appears again as an upcrossing. So, every summand in the permutation definition of the determinant appears with the same sign also in the determinant of the transpose.

What are determinants useful for?

As the name tells, determinants determine a lot about matrices. We can see from the determinant whether the matrix is invertible.

An other reason is that determinants allow explicit formulas for the inverse of a matrix. We might look at this next time. Next week we will see that determinants allow to define the characteristic polynomial of a matrix whose roots are the important eigenvalues. In analysis, the determinant appears in change of variable formulas:

$$\int_S f(x) dx = \int_{u(S)} f(y) |\det(Du^{-1}(y))| dy.$$

Physicists are excited about determinants because summation over all possible "paths" is used as a quantization method. The Feynmann path integral is a "summation" over a suitable class of paths and leads to quantum mechanics. The relation with determinants comes because each summand in a determinant can be interpreted as a contribution of a path in a finite graph with n nodes.

Homework due April 6, 2011

- 1 Find the determinant of

$$\begin{bmatrix} 3 & 2 & 0 & 0 & 0 & 0 \\ 3 & 3 & 2 & 0 & 0 & 0 \\ 3 & 3 & 3 & 2 & 0 & 0 \\ 3 & 3 & 3 & 3 & 2 & 0 \\ 3 & 3 & 3 & 3 & 3 & 2 \\ 3 & 3 & 3 & 3 & 3 & 3 \end{bmatrix}$$

- 2 Find the determinant of

$$\begin{bmatrix} 3 & 1 & 1 & 2 & 2 & 2 \\ 1 & 0 & 1 & 2 & 2 & 2 \\ 1 & 0 & 3 & 2 & 2 & 2 \\ 0 & 0 & 0 & 4 & 1 & 0 \\ 0 & 0 & 0 & 1 & 4 & 0 \\ 0 & 0 & 0 & 1 & 1 & 4 \end{bmatrix}$$

- 3 a) Find an example showing that $\det(A+B) \neq \det(A) + \det(B)$.
b) How do you modify $\det(\lambda A) = \lambda \det(A)$ to make it correct if A is a $n \times n$ matrix?

Lecture 26: Determinants part three

Geometry

For a $n \times n$ matrix, the determinant defines a volume of the parallelepiped spanned by the column vectors.

Since we have not defined volume in higher dimensions, we can take the absolute value of the determinant as the definition of the volume. The sign of the determinant gives additional information, it defines the **orientation**. Determinants are useful to **define** these concepts in higher dimensions. The linearity result shown last time illustrates why this makes sense: scaling one of the vector by a factor 2 for example changes the determinant by a factor 2. This also increases the volume by a factor of 2 which can be interpreted as stacking two such solids on top of each other.

- 1 The area of the parallelogram spanned by $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ is the determinant of

$$A = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}$$

which is 4.

- 2 Find the volume of the parallelepiped spanned by the column vectors of $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$.

Solution: The determinant is -1 . The volume is 1. The fact that the determinant is negative reflects the fact that these three vectors form a "left handed" coordinate system.

The volume of a k dimensional parallelepiped defined by the vectors v_1, \dots, v_k is $\sqrt{\det(A^T A)}$.

We can take also this as the definition of the volume. Note that $A^T A$ is a square matrix.

- 3 The area of the parallelogram in space spanned by the vectors $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ is

$$\det(A^T A) = \det\left(\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}\right) = \det\left(\begin{bmatrix} 3 & 2 \\ 2 & 2 \end{bmatrix}\right) = 2.$$

The area is therefore $\sqrt{2}$. If you have seen multivariable calculus, you could also have computed the area using the cross product. The area is the length of the cross product $\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$ of the two vectors which is $\sqrt{2}$ too.

Cramer's rule

Solution to the linear system equations $A\vec{x} = \vec{b}$ can be given explicitly using determinants:

Cramers rule: If A_i is the matrix, where the column \vec{v}_i of A is replaced by \vec{b} , then

$$x_i = \frac{\det(A_i)}{\det(A)}$$

Proof. $\det(A_i) = \det([v_1, \dots, b, \dots, v_n]) = \det([v_1, \dots, (Ax), \dots, v_n]) = \det([v_1, \dots, \sum_i x_i v_i, \dots, v_n]) = x_i \det([v_1, \dots, v_i, \dots, v_n]) = x_i \det(A)$.

- 4 Solve the system $5x + 3y = 8, 8x + 5y = 2$ using Cramer's rule. **Solution.** This linear system with $A = \begin{bmatrix} 5 & 3 \\ 8 & 5 \end{bmatrix}$ and $b = \begin{bmatrix} 8 \\ 2 \end{bmatrix}$. We get $x = \det\left[\begin{smallmatrix} 8 & 3 \\ 2 & 5 \end{smallmatrix}\right] = 34y = \det\left[\begin{smallmatrix} 5 & 8 \\ 8 & 2 \end{smallmatrix}\right] = -54$.

Gabriel Cramer was born in 1704 in Geneva. He worked on geometry and analysis until his death at in 1752 during a trip to France. Cramer used the rule named after him in a book "Introduction à l'analyse des lignes courbes algébrique", where he used the method to solve systems of equations with 5 unknowns. According to a short biography of Cramer by J.J O'Connor and E F Robertson, the rule had been used already before by other mathematicians. Solving systems with Cramer's formulas is slower than by Gaussian elimination. But it is useful for example if the matrix A or the vector b depends on a parameter t , and we want to see how x depends on the parameter t . One can find explicit formulas for $(d/dt)x_i(t)$ for example.

Cramer's rule leads to an explicit formula for the inverse of a matrix inverse of a matrix:

Let A_{ij} be the matrix where the i 'th row and the j 'th column is deleted. $B_{ij} = (-1)^{i+j} \det(A_{ji})$ is called the **classical adjoint** or **adjugate** of A . The determinant of the classical adjugate is called **minor**.

$$[A^{-1}]_{ij} = (-1)^{i+j} \frac{\det(A_{ji})}{\det(A)}$$

Proof. The columns of A^{-1} are the solutions of

$$A\vec{x} = \vec{e}_j$$

where \vec{e}_j are basis vectors.

Don't confuse the classical adjoint with the **transpose** A^T . The classical adjoint is the transpose of the matrix where the i 'th row and j 'th column is deleted. The mix up is easy to do since the transpose is often also called the **adjoint**.

5 $A = \begin{bmatrix} 2 & 3 & 1 \\ 5 & 2 & 4 \\ 6 & 0 & 7 \end{bmatrix}$ has $\det(A) = -17$ and we get $A^{-1} = \begin{bmatrix} 14 & -21 & 10 \\ -11 & 8 & -3 \\ -12 & 18 & -11 \end{bmatrix} / (-17)$:

$B_{11} = (-1)^2 \det \begin{bmatrix} 2 & 4 \\ 0 & 7 \end{bmatrix} = 14$. $B_{12} = (-1)^3 \det \begin{bmatrix} 3 & 1 \\ 0 & 7 \end{bmatrix} = -21$. $B_{13} = (-1)^4 \det \begin{bmatrix} 3 & 1 \\ 2 & 4 \end{bmatrix} = 10$.

$B_{21} = (-1)^3 \det \begin{bmatrix} 5 & 4 \\ 6 & 7 \end{bmatrix} = -11$. $B_{22} = (-1)^4 \det \begin{bmatrix} 2 & 1 \\ 6 & 7 \end{bmatrix} = 8$. $B_{23} = (-1)^5 \det \begin{bmatrix} 2 & 1 \\ 5 & 4 \end{bmatrix} = -3$.

$B_{31} = (-1)^4 \det \begin{bmatrix} 5 & 2 \\ 6 & 0 \end{bmatrix} = -12$. $B_{32} = (-1)^5 \det \begin{bmatrix} 2 & 3 \\ 6 & 0 \end{bmatrix} = 18$. $B_{33} = (-1)^6 \det \begin{bmatrix} 2 & 3 \\ 5 & 2 \end{bmatrix} = -11$.

Random matrices

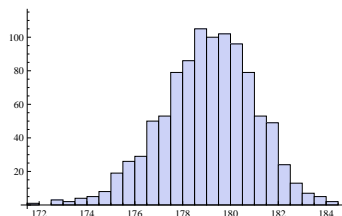
If the entries of a matrix are random variables with a continuous distribution, then the determinant is nonzero with probability one.

If the entries of a matrix are random variables which have the property that $P[X = x] = p > 0$ for some x , then there is a nonzero probability that the determinant is zero.

Proof. We have with probability p^{2n} that the first two rows have the same entry x .

What is the distribution of the determinant of a random matrix? These are questions which are hard to analyze theoretically. Here is an experiment: we take random 100×100 matrices and look at the distribution of the logarithm of the determinant.

```
M=100; T:=Log[Abs[Det[Table[Random[NormalDistribution[0,1]],{M},{M}]]]];
s=Table[T,{1000}]; S=Histogram[s]
```



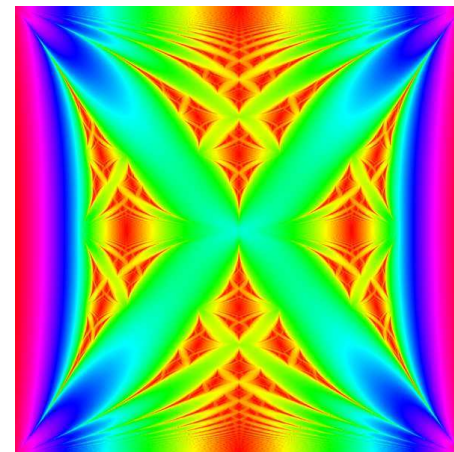
Applications of determinants

In solid state physics, one is interested in the function $f(E) = \det(L - EI_n)$, where

$$L = \begin{bmatrix} \lambda \cos(\alpha) & 1 & 0 & \cdot & 0 & 1 \\ 1 & \lambda \cos(2\alpha) & 1 & \cdot & \cdot & 0 \\ 0 & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & 0 \\ 0 & \cdot & \cdot & 1 & \lambda \cos((n-1)\alpha) & 1 \\ 1 & 0 & \cdot & 0 & 1 & \lambda \cos(n\alpha) \end{bmatrix}$$

describes an electron in a periodic crystal, E is the energy and $\alpha = 2\pi/n$. The electron can move as a Bloch wave whenever the determinant is negative. These intervals form the **spectrum** of the quantum mechanical system. A physicist is interested in the rate of change of $f(E)$ or its dependence on λ when E is fixed.

The graph to the left shows the function $E \mapsto \log(|\det(L - EI_n)|)$ in the case $\lambda = 2$ and $n = 5$. In the energy intervals, where this function is zero, the electron can move, otherwise the crystal is an insulator. The picture to the right shows the spectrum of the crystal depending on α . It is called the "Hofstadter butterfly" made popular in the book "Gödel, Escher, Bach" by Douglas Hofstadter.



Homework due April 6, 2011

1 Find the following determinants

a) $A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 6 & 8 & 10 \\ 5 & 5 & 5 & 5 & 4 \\ 1 & 3 & 2 & 7 & 4 \\ 3 & 2 & 8 & 4 & 9 \end{bmatrix}$

b) $A = \begin{bmatrix} 2 & 1 & 4 & 4 & 2 \\ 1 & 1 & 1 & 2 & 3 \\ 0 & 0 & 2 & 1 & 1 \\ 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix}$

2 Find the following determinants

a) $A = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 2 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 4 & 2 \end{bmatrix}$

b) $A = \begin{bmatrix} 1 & 6 & 10 & 1 & 15 \\ 2 & 8 & 17 & 1 & 29 \\ 0 & 0 & 3 & 8 & 12 \\ 0 & 0 & 0 & 4 & 9 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix}$

3 Find a 4×4 matrix A with entries 0, +1 and -1 for which the determinant is maximal. Hint. Think about the volume. How do we get a maximal volume?

Lecture 27: Discrete dynamical systems

Eigenvectors and Eigenvalues

Markets, population evolutions or ingredients in a chemical reaction are often nonlinear. A linear description often can give a good approximation and solve the system explicitly. Eigenvectors and eigenvalues provide us with the key to do so.

A nonzero vector v is called an **eigenvector** of a $n \times n$ matrix A if $[Av = \lambda v]$ for some number λ . The later is called an **eigenvalue** of A .

We first look at real eigenvalues but also consider complex eigenvalues.

- 1 A vector v is an eigenvector to the eigenvalue 0 if and only if \vec{v} is in the kernel of A because $A\vec{v} = 0\vec{v}$ means that \vec{v} is in the kernel.
- 2 A rotation A in three dimensional space has an eigenvalue 1, with eigenvector spanning the axes of rotation. This vector satisfies $A\vec{v} = \vec{v}$.
- 3 Every standard basis vector \vec{w}_i is an eigenvector if A is a diagonal matrix. For example, $\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 3 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.
- 4 For an orthogonal projection P onto a space V , every vector in V is an eigenvector to the eigenvalue 1 and every vector perpendicular to V is an eigenvector to the eigenvalue 0.
- 5 For a reflection R at a space V , every vector v in V is an eigenvector with eigenvalue 1. Every vector perpendicular to v is an eigenvector to the eigenvalue -1 .

Discrete dynamical systems

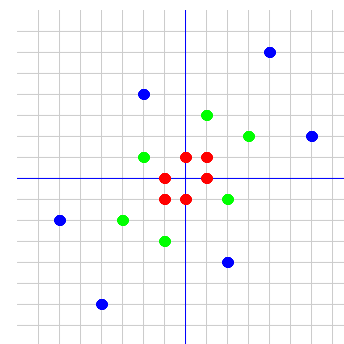
When applying a linear map $x \mapsto Ax$ again and again, we obtain a **discrete dynamical system**. We want to understand what happens with the **orbit** $x_1 = Ax, x_2 = AAx = A^2x, x_3 = AAAx = A^3x, \dots$ and find a closed formula for $A^n x$

- 6 The one-dimensional discrete dynamical system $x \mapsto ax$ or $x_{n+1} = ax_n$ has the solution $x_n = a^n x_0$. The value $1.03^{20} \cdot 1000 = 1806.11$ for example is the balance on a bank account which had 1000 dollars 20 years ago if the interest rate was a constant 3 percent.

- 7 Look at the recursion

$u_{n+1} = u_n - u_{n-1}$

with $u_0 = 0, u_1 = 1$. Because $\begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u_n \\ u_{n-1} \end{bmatrix} = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}$ we have a discrete dynamical system. Lets compute some orbits: $A = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}, A^2 = \begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}, A^3 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$. We see that A^6 is the identity. Every initial vector is mapped after 6 iterations back to its original starting point.



- 8 $A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \vec{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, A\vec{v} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, A^2\vec{v} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}, A^3\vec{v} = \begin{bmatrix} 7 \\ 1 \end{bmatrix}, A^4\vec{v} = \begin{bmatrix} 9 \\ 1 \end{bmatrix}$ etc.
Do you see a pattern?

The following example shows why eigenvalues and eigenvectors are so important:

- 9 If \vec{v} is an eigenvector with eigenvalue λ , then $A\vec{v} = \lambda\vec{v}, A^2\vec{v} = A(A\vec{v}) = A\lambda\vec{v} = \lambda A\vec{v} = \lambda^2\vec{v}$ and more generally $A^n\vec{v} = \lambda^n\vec{v}$.

For an eigenvector, we have a closed form solution for $A^n\vec{v}$. It is $\lambda^n\vec{v}$.

- 10 The recursion

$$x_{n+1} = x_n + x_{n-1}$$

with $x_0 = 0$ and $x_1 = 1$ produces the **Fibonacci sequence**

$$(1, 1, 2, 3, 5, 8, 13, 21, \dots)$$

This can be computed with a discrete dynamical system because

$$\begin{bmatrix} x_{n+1} \\ x_n \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_n \\ x_{n-1} \end{bmatrix}.$$

Can we find a formula for the n 'th term?



In the third section of **Liber abbaci**, published in 1202, the mathematician Fibonacci, with real name **Leonardo di Pisa** (1170-1250) writes: "A certain man put a pair of rabbits in a place surrounded on all sides by a wall. How many pairs of rabbits can be produced from that pair in a year if it is supposed that every month each pair begets a new pair which from the second month on becomes productive?"

Markov case

We will discuss the following situation a bit more in detail:

An $n \times n$ matrix is called a **Markov matrix** if all entries are nonnegative and each column adds up to 1.

- 11 Customers using **Apple IOS** and **Google Android** are represented by a vector $\begin{bmatrix} A \\ G \end{bmatrix}$. Each cycle 1/3 of IOS users switch to Android and 2/3 stays. Also lets assume that 1/2 of the Android OS users switch to IOS and 1/2 stay. The matrix $A = \begin{bmatrix} 2/3 & 1/2 \\ 1/3 & 1/2 \end{bmatrix}$ is a **Markov matrix**. What customer ratio do we have in the limit? The matrix A has an eigenvector $(3/5, 2/5)$ which belongs to the eigenvalue 1.

$$A\vec{v} = \vec{v}$$

means that 60 to 40 percent is the final stable distribution.

The following fact motivates to find good methods to compute eigenvalues and eigenvectors.

If $A\vec{v}_1 = \lambda_1\vec{v}_1$, $A\vec{v}_2 = \lambda_2\vec{v}_2$ and $\vec{v} = c_1\vec{v}_1 + c_2\vec{v}_2$, we have **closed form solution** $A^n\vec{v} = c_1\lambda_1^n\vec{v}_1 + c_2\lambda_2^n\vec{v}_2$.

Lets try this in the Fibonacci case. We will see next time how we find the eigenvalues and eigenvectors:

- 12 Lets try to find a number ϕ such that

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \phi \\ 1 \end{bmatrix} = \phi \begin{bmatrix} \phi \\ 1 \end{bmatrix}$$

This leads to the quadratic equation $\phi + 1 = \phi^2$ which has the solutions $\phi_+ = (1 + \sqrt{5})/2$ and $\phi_- = (1 - \sqrt{5})/2$. The number ϕ^+ is one of the most famous and symmetric numbers

in mathematics called the **golden ratio**. We have found our eigenvalues and eigenvectors. Now find c_1, c_2 such that

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} = c_1 \begin{bmatrix} \phi^+ \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} \phi^- \\ 1 \end{bmatrix}$$

We see $c_1 = -c_2 = 1/\sqrt{5}$. We can write

$$\begin{bmatrix} x_{n+1} \\ x_n \end{bmatrix} = A^n \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{5}}\phi_+^n \begin{bmatrix} \phi_+ \\ 1 \end{bmatrix} - \frac{1}{\sqrt{5}}\phi_-^n \begin{bmatrix} \phi_- \\ 1 \end{bmatrix}$$

and can read off $x_n = (\phi_+^n - \phi_-^n)/\sqrt{5}$.

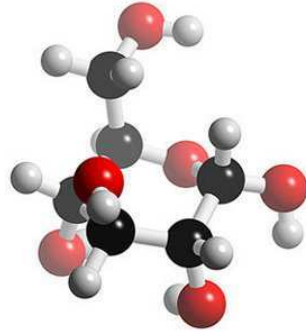
Homework due April 13, 2011

Compare Problem 52 in Chapter 7.1 of Bretscher's Book. The glucose and excess hormone concentration in your blood are modeled by a vector $\vec{v} = \begin{bmatrix} g \\ h \end{bmatrix}$. Between meals the concentration changes to $\vec{v} \rightarrow A\vec{v}$, where

1

$$A = \begin{bmatrix} 0.978 & -0.006 \\ 0.004 & 0.992 \end{bmatrix}.$$

Check that $\begin{bmatrix} -1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 3 \\ -1 \end{bmatrix}$ are eigenvectors of A . Find the eigenvalues.



Compare Problem 54 in Chapter 7.1 of Bretscher's Book. The dynamical system $v_{n+1} = Av_n$ with

$$A = \begin{bmatrix} 0 & 2 \\ 1 & 1 \end{bmatrix}$$

2

models the growth of a **lilac bush**. The vector $\vec{v} = \begin{bmatrix} n \\ a \end{bmatrix}$ models the number of new branches and the number of old branches. Verify that $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 2 \\ -1 \end{bmatrix}$ are eigenvectors of A . Find the eigenvalues and find the close form solution starting with $\vec{v} = [2, 3]^T$.



Compare problem 50 in Chapter 7.1 of Bretscher's Book. Two interacting populations of **hares and foxes** can be modeled by the discrete dynamical system $v_{n+1} = Av_n$ with

$$A = \begin{bmatrix} 4 & -2 \\ 1 & 1 \end{bmatrix}$$

3

Find a closed form solutions in the following three cases: a) $\vec{v}_0 = \begin{bmatrix} h_0 \\ f_0 \end{bmatrix} = \begin{bmatrix} 100 \\ 100 \end{bmatrix}$.

b) $\vec{v}_0 = \begin{bmatrix} h_0 \\ f_0 \end{bmatrix} = \begin{bmatrix} 200 \\ 100 \end{bmatrix}.$

c) $\vec{v}_0 = \begin{bmatrix} h_0 \\ f_0 \end{bmatrix} = \begin{bmatrix} 600 \\ 500 \end{bmatrix}.$



Lecture 28: Eigenvalues

We have seen that $\det(A) \neq 0$ if and only if A is invertible.

The polynomial $f_A(\lambda) = \det(A - \lambda I_n)$ is called the **characteristic polynomial** of A .

The eigenvalues of A are the roots of the characteristic polynomial.

Proof. If $Av = \lambda v$, then v is in the kernel of $A - \lambda I_n$. Consequently, $A - \lambda I_n$ is not invertible and

$$\det(A - \lambda I_n) = 0.$$

1 For the matrix $A = \begin{bmatrix} 2 & 1 \\ 4 & -1 \end{bmatrix}$, the characteristic polynomial is

$$\det(A - \lambda I_2) = \det\left(\begin{bmatrix} 2-\lambda & 1 \\ 4 & -1-\lambda \end{bmatrix}\right) = \lambda^2 - \lambda - 6.$$

This polynomial has the roots 3, -2.

Let $\text{tr}(A)$ denote the **trace** of a matrix, the sum of the diagonal elements of A .

For the matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, the characteristic polynomial is

$$\lambda^2 - \text{tr}(A)\lambda + \det(A).$$

We can see this directly by writing out the determinant of the matrix $A - \lambda I_2$. The trace is important because it always appears in the characteristic polynomial, also if the matrix is larger:

For any $n \times n$ matrix, the characteristic polynomial is of the form

$$f_A(\lambda) = (-\lambda)^n + \text{tr}(A)(-\lambda)^{n-1} + \dots + \det(A).$$

Proof. The pattern, where all the entries are in the diagonal leads to a term $(A_{11} - \lambda) \cdot (A_{22} - \lambda) \dots (A_{nn} - \lambda)$ which is $(-\lambda)^n + (A_{11} + \dots + A_{nn})(-\lambda)^{n-1} + \dots$. The rest of this as well as the other patterns only give us terms which are of order λ^{n-2} or smaller.

How many eigenvalues do we have? For real eigenvalues, it depends. A rotation in the plane with an angle different from 0 or π has no real eigenvector. The eigenvalues are complex in that case:

2 For a rotation $A = \begin{bmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{bmatrix}$ the characteristic polynomial is $\lambda^2 - 2\cos(\alpha) + 1$ which has the roots $\cos(\alpha) \pm i\sin(\alpha) = e^{i\alpha}$.

Allowing complex eigenvalues is really a blessing. The structure is very simple:

Fundamental theorem of algebra: For a $n \times n$ matrix A , the characteristic polynomial has exactly n roots. There are therefore exactly n eigenvalues of A if we count them with multiplicity.

Proof¹ One only has to show a polynomial $p(z) = z^n + a_{n-1}z^{n-1} + \dots + a_1z + a_0$ always has a root z_0 . We can then factor out $p(z) = (z - z_0)g(z)$ where $g(z)$ is a polynomial of degree $(n - 1)$ and use induction in n . Assume now that in contrary the polynomial p has no root. Cauchy's integral theorem then tells

$$\int_{|z|=r} \frac{dz}{zp(z)} = \frac{2\pi i}{p(0)} \neq 0. \quad (1)$$

On the other hand, for all r ,

$$\left| \int_{|z|=r} \frac{dz}{zp(z)} \right| \leq 2\pi r \max_{|z|=r} \frac{1}{|zp(z)|} = \frac{2\pi}{\min_{|z|=r} |p(z)|}. \quad (2)$$

The right hand side goes to 0 for $r \rightarrow \infty$ because

$$|p(z)| \geq |z|^n \left(1 - \frac{|a_{n-1}|}{|z|} - \dots - \frac{|a_0|}{|z|^n}\right)$$

which goes to infinity for $r \rightarrow \infty$. The two equations (1) and (2) form a contradiction. The assumption that p has no root was therefore not possible.

If $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A , then

$$f_A(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \dots (\lambda_n - \lambda).$$

Comparing coefficients, we know now the following important fact:

The determinant of A is the product of the eigenvalues. The trace is the sum of the eigenvalues.

We can therefore often compute the eigenvalues

3 Find the eigenvalues of the matrix

$$A = \begin{bmatrix} 3 & 7 \\ 5 & 5 \end{bmatrix}$$

Because each row adds up to 10, this is an eigenvalue: you can check that $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. We can also read off the trace 8. Because the eigenvalues add up to 8 the other eigenvalue is -2 . This example seems special but it often occurs in textbooks. Try it out: what are the eigenvalues of

$$A = \begin{bmatrix} 11 & 100 \\ 12 & 101 \end{bmatrix}?$$

¹A. R. Schep. A Simple Complex Analysis and an Advanced Calculus Proof of the Fundamental theorem of Algebra. Mathematical Monthly, 116, p 67-68, 2009

- 4 Find the eigenvalues of the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 2 & 3 & 4 & 5 \\ 0 & 0 & 3 & 4 & 5 \\ 0 & 0 & 0 & 4 & 5 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix}$$

We can immediately compute the characteristic polynomial in this case because $A - \lambda I_5$ is still upper triangular so that the determinant is the product of the diagonal entries. We see that the eigenvalues are 1, 2, 3, 4, 5.

The eigenvalues of an upper or lower triangular matrix are the diagonal entries of the matrix.

- 5 How do we construct 2×2 matrices which have integer eigenvectors and integer eigenvalues? Just take an integer matrix for which the row vectors have the same sum. Then this sum is an eigenvalue to the eigenvector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. The other eigenvalue can be obtained by noticing that the trace of the matrix is the sum of the eigenvalues. For example, the matrix $\begin{bmatrix} 6 & 7 \\ 2 & 11 \end{bmatrix}$ has the eigenvalue 13 and because the sum of the eigenvalues is 18 a second eigenvalue 5.

A matrix with nonnegative entries for which the sum of the columns entries add up to 1 is called a **Markov matrix**.

Markov Matrices have an eigenvalue 1.

Proof. The eigenvalues of A and A^T are the same because they have the same characteristic polynomial. The matrix A^T has an eigenvector $[1, 1, 1, 1, 1]^T$.

6

$$A = \begin{bmatrix} 1/2 & 1/3 & 1/4 \\ 1/4 & 1/3 & 1/3 \\ 1/4 & 1/3 & 5/12 \end{bmatrix}$$

This vector \vec{v} defines an equilibrium point of the Markov process.

- 7 If $A = \begin{bmatrix} 1/3 & 1/2 \\ 2/3 & 1/2 \end{bmatrix}$. Then $[3/7, 4/7]$ is the equilibrium eigenvector to the eigenvalue 1.

Homework due April 13, 2011

- 1 a) Find the characteristic polynomial and the eigenvalues of the matrix

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \\ 4 & -1 & 0 \end{bmatrix}.$$

b) Find the eigenvalues of $A = \begin{bmatrix} 100 & 1 & 1 & 1 & 1 \\ 1 & 100 & 1 & 1 & 1 \\ 1 & 1 & 100 & 1 & 1 \\ 1 & 1 & 1 & 100 & 1 \\ 1 & 1 & 1 & 1 & 100 \end{bmatrix}.$

- 2 a) Verify that $n \times n$ matrix has a at least one real eigenvalue if n is odd.
b) Find a 4×4 matrix, for which there is no real eigenvalue.
c) Verify that a symmetric 2×2 matrix has only real eigenvalues.

- 3 a) Verify that for a partitioned matrix

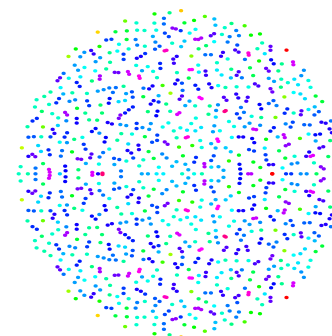
$$C = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix},$$

the union of the eigenvalues of A and B are the eigenvalues of C .

b) Assume we have an eigenvalue \vec{v} of A use this to find an eigenvector of C . Similarly, if \vec{w} is an eigenvector of B , build an eigenvector of C .

(*) Optional: Make some experiments with random matrices: The following Mathematica code computes Eigenvalues of random matrices. You will observe Girko's circular law.

```
M=1000;
A=Table[Random[]-1/2,{M},{M}];
e=Eigenvalues[A];
d=Table[Min[Table[If[i==j,10,Abs[e[[i]]-e[[j]]]],{j,M}]],{i,M}];
a=Max[d]; b=Min[d];
Graphics[Table[{Hue[(d[[j]]-a)/(b-a)],
Point[{Re[e[[j]]],Im[e[[j]]]}]}, {j,M}]]
```



Lecture 29: Eigenvectors

Eigenvectors

Assume we know an eigenvalue λ . How do we compute the corresponding eigenvector?

The **eigenspace** of an eigenvalue λ is defined to be the linear space of all eigenvectors of A to the eigenvalue λ .

The eigenspace is the kernel of $A - \lambda I_n$.

Since we have computed the kernel a lot already, we know how to do that.

The dimension of the eigenspace of λ is called the **geometric multiplicity** of λ .

Remember that the multiplicity with which an eigenvalue appears is called the algebraic multiplicity of λ :

The algebraic multiplicity is larger or equal than the geometric multiplicity.

Proof. Let λ be the eigenvalue. Assume it has geometric multiplicity m . If v_1, \dots, v_m is a basis of the eigenspace E_μ form the matrix S which contains these vectors in the first m columns. Fill the other columns arbitrarily. Now $B = S^{-1}AS$ has the property that the first m columns are $\mu e_1, \dots, \mu e_m$, where e_i are the standard vectors. Because A and B are similar, they have the same eigenvalues. Since B has m eigenvalues λ also A has this property and the algebraic multiplicity is $\geq m$.

You can remember this with an analogy: the **geometric mean** \sqrt{ab} of two numbers is smaller or equal to the **algebraic mean** $(a+b)/2$.

- 1 Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

This matrix has a large kernel. Row reduction indeed shows that the kernel is 4 dimensional. Because the algebraic multiplicity is larger or equal than the geometric multiplicity there are 4 eigenvalues 0. We can also immediately get the last eigenvalue from the trace 15. The eigenvalues of A are 0, 0, 0, 0, 15.

- 2 Find the eigenvalues of B .

$$B = \begin{bmatrix} 101 & 2 & 3 & 4 & 5 \\ 1 & 102 & 3 & 4 & 5 \\ 1 & 2 & 103 & 4 & 5 \\ 1 & 2 & 3 & 104 & 5 \\ 1 & 2 & 3 & 4 & 105 \end{bmatrix}$$

This matrix is $A + 100I_5$ where A is the matrix from the previous example. Note that if $Bv = \lambda v$ then $(A + 100I_5)v = \lambda + 100)v$ so that A, B have the same eigenvectors and the eigenvalues of B are 100, 100, 100, 100, 115.

- 3 Find the determinant of the previous matrix B . **Solution:** Since the determinant is the product of the eigenvalues, the determinant is $100^4 \cdot 115$.

- 4 The shear $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ has the eigenvalue 1 with algebraic multiplicity 2. The kernel of $A - I_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ is spanned by $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and the geometric multiplicity is 1.

- 5 The matrix $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ has eigenvalue 1 with algebraic multiplicity 2 and the eigenvalue 0 with multiplicity 1. Eigenvectors to the eigenvalue $\lambda = 1$ are in the kernel of $A - 1$ which is the kernel of $\begin{bmatrix} 0 & 1 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ and spanned by $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$. The geometric multiplicity is 1.

If all eigenvalues are different, then all eigenvectors are linearly independent and all geometric and algebraic multiplicities are 1. The eigenvectors form then an eigenbasis.

Proof. If all are different, there is one of them λ_i which is different from 0. We use induction with respect to n and assume the result is true for $n-1$. Assume that in contrary the eigenvectors are linearly dependent. We have $v_i = \sum_{j \neq i} a_j v_j$ and $\lambda_i v_i = A v_i = A(\sum_{j \neq i} a_j v_j) = \sum_{j \neq i} a_j \lambda_j v_j$ so that $v_i = \sum_{j \neq i} b_j v_j$ with $b_j = a_j \lambda_j / \lambda_i$. If the eigenvalues are different, then $a_j \neq b_j$ and by subtracting $v_i = \sum_{j \neq i} a_j v_j$ from $v_i = \sum_{j \neq i} b_j v_j$, we get $0 = \sum_{j \neq i} (b_j - a_j) v_j = 0$. Now $(n-1)$ eigenvectors of the n eigenvectors are linearly dependent. Now use the induction assumption.

Here is an other example of an eigenvector computation:

- 6 Find all the eigenvalues and eigenvectors of the matrix

$$B = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Solution. The characteristic polynomial is $\lambda^4 - 1$. It has the roots $1, -1, i, -i$. Instead of computing the eigenvectors for each eigenvalue, write

$$v = \begin{bmatrix} 1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix},$$

and look at $Bv = \lambda v$.

Where are eigenvectors used: in class we will look at some applications: **Hückel theory**, **orbitals** of the Hydrogen atom and **Page rank**. In all these cases, the eigenvectors have immediate interpretations. We will talk about page rank more when we deal with Markov processes.

The page rank vector is an eigenvector to the Google matrix.

These matrices can be huge. The google matrix is a $n \times n$ matrix where n is larger than 10 billion!¹

¹The book of Lanville and Meyher of 2006 gives 8 billion. This was 5 years ago.

Homework due April 13, 2011

- 1 Find the eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & -2 \\ 3 & 6 & -5 \end{bmatrix}.$$

- 2 a) Find the eigenvectors of A^{10} , where A is the previous matrix.
b) Find the eigenvectors of A^T , where A is the previous matrix.

- 3 This is homework problem 40 in section 7.3 of the Bretscher book.



Photos of the
Swiss lakes
in the text.
The pollution
story is fiction
fortunately.



The vector $A^n(x)b$ gives pollution levels in the Silvaplana, Sils and St Moritz lake n weeks after an oil spill. The matrix is $A = \begin{bmatrix} 0.7 & 0 & 0 \\ 0.1 & 0.6 & 0 \\ 0 & 0.2 & 0.8 \end{bmatrix}$ and $b = \begin{bmatrix} 100 \\ 0 \\ 0 \end{bmatrix}$ is the initial pollution level. Find a closed form solution for the pollution after n days.

Lecture 30: Diagonalization

Diagonalization

Two matrices are called **similar** if $S^{-1}AS$. A matrix is called **diagonalizable** if it is similar to a diagonal matrix.

A matrix is diagonalizable if and only if it has an eigenbasis, a basis consisting of eigenvectors.

Proof. If we have an eigenbasis, we have a coordinate transformation matrix S which contains the eigenvectors v_i as column vectors. To see that the matrix $S^{-1}AS$ is diagonal, we check

$$S^{-1}ASe_i - S^{-1}Av_i = S^{-1}\lambda_i v_i = \lambda_i S^{-1}v_i = \lambda_i e_i.$$

On the other hand if A is diagonalizable, then we have a matrix S for which $S^{-1}AS = B$ is diagonal. The column vectors of S are eigenvectors because the k 'th column of the equation $AS = BS$ shows $Av_i = \lambda_i v_i$.

Are all matrices diagonalizable? No! We need to have an eigenbasis and therefore that the geometric multiplicities all agree with the algebraic multiplicities. We have seen that the shear matrix

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

has the eigenvalues 1 for which the geometric multiplicity is smaller than the algebraic one. This matrix is not diagonalizable.

Simple spectrum

A matrix has simple spectrum, if all eigenvalues have algebraic multiplicity 1.

If a matrix has simple spectrum, then it is diagonalizable.

Proof. Because the algebraic multiplicity is 1 for each eigenvalue and the geometric multiplicity is always at least 1, we have an eigenvector for each eigenvalue and so n eigenvalues.

1 We have computed the eigenvalues of the rotation matrix

$$A = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}$$

We have seen that the eigenvalues are $e^{i\alpha} = \cos(\alpha) + i\sin(\alpha)$, the eigenvectors are $\begin{bmatrix} \pm i \\ 1 \end{bmatrix}$. The eigenvectors are the same for every rotation-dilation matrix. With

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}, S = \begin{bmatrix} i & -i \\ 1 & 1 \end{bmatrix}$$

we have

$$S^{-1}AS = \begin{bmatrix} a+ib & 0 \\ 0 & a-ib \end{bmatrix}.$$

Functional calculus

2 What is $A^{100} + A^{37} - 1$ if $A = \begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix}$? The matrix has the eigenvalues $\lambda_1 = 2 + \sqrt{3}$ with eigenvector $\vec{v}_1 = [\sqrt{3}, 1]$ and the eigenvalues $\lambda_2 = 2 - \sqrt{3}$ with eigenvector $\vec{v}_2 = [-\sqrt{3}, 1]$. Form $S = \begin{bmatrix} \sqrt{3} & -\sqrt{3} \\ 1 & 1 \end{bmatrix}$ and check $S^{-1}AS = D$ is diagonal. Because $B^k = S^{-1}A^kS$ can easily be computed, we know $A^{100} + A^{37} - 1 = S(B^{100} + B^{37} - 1)S^{-1}$.

Establishing similarity

3 Show that the matrices $A = \begin{bmatrix} 3 & 5 \\ 2 & 6 \end{bmatrix}$ $B = \begin{bmatrix} 4 & 4 \\ 3 & 5 \end{bmatrix}$ are similar. Proof. They have the same eigenvalues 8, 9 as you can see by inspecting the sum of rows and the trace. Both matrices are therefore diagonalizable and similar to the matrix

$$\begin{bmatrix} 8 & 0 \\ 0 & 9 \end{bmatrix}.$$

- If A and B have the same characteristic polynomial and diagonalizable, then they are similar.
- If A and B have a different determinant or trace, they are not similar.
- If A has an eigenvalue which is not an eigenvalue of B , then they are not similar.
- If A and B have the same eigenvalues but different geometric multiplicities, then they are not similar.

Without proof we mention the following result which gives an if and only if result for similarity:

If A and B have the same eigenvalues with geometric multiplicities which agree and the same holds for all powers A^k and B^k , then A is similar to B .

Cayley Hamilton theorem

For any polynomial p ,⁴ we can form the matrix $p(A)$. For example, for $p(x) = x^2 + 2x + 3$, we have $p(A) = A^2 + 2A + 3$.

If f_A is the characteristic polynomial, we can form $f_A(A)$

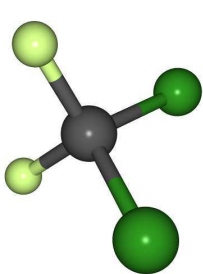
If A is diagonalizable, then $f_A(A) = 0$.

The matrix $B = S^{-1}AS$ has the eigenvalues in the diagonal. So $f_A(B)$, which contains $f_A(\lambda_i)$ in the diagonal is zero. From $f_A(B) = 0$ we get $Sf_A(B)S^{-1} = f_A(A) = 0$.

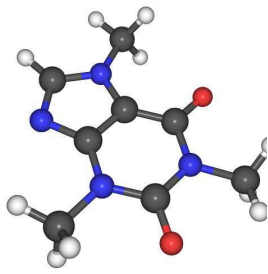
The theorem holds for all matrices: the coefficients of a general matrix can be changed a tiny bit so that all eigenvalues are different. For any such perturbations one has $f_A(A) = 0$. Because the coefficients of $f_A(A)$ depend continuously on A , they are zero in general.

An application in chemistry

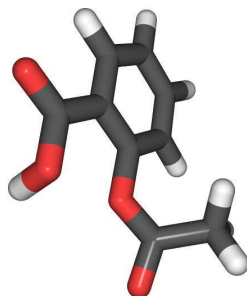
While quantum mechanics describes the motion of atoms in molecules, the vibrations can be described classically, when treating the atoms as "balls" connected with springs. Such approximations are necessary when dealing with large atoms, where quantum mechanical computations would be too costly. Examples of simple molecules are white phosphorus P_4 , which has tetrahedral shape or methane CH_4 the simplest organic compound or **freon**, CF_2Cl_2 which is used in refrigerants.



Freon CF_2Cl_2



Caffeine $C_8H_{10}N_4O_2$



Aspirin $C_9H_8O_4$

1

Let x_1, x_2, x_3, x_4 be the positions of the four phosphorus atoms (each of them is a 3-vector). The inter-atomic forces bonding the atoms is modeled by springs. The first atom feels a force $x_2 - x_1 + x_3 - x_1 + x_4 - x_1$ and is accelerated in the same amount. Let's just choose units so that the force is equal to the acceleration. Then

$$\begin{aligned}\ddot{x}_1 &= (x_2 - x_1) + (x_3 - x_1) + (x_4 - x_1) \\ \ddot{x}_2 &= (x_3 - x_2) + (x_4 - x_2) + (x_1 - x_2) \\ \ddot{x}_3 &= (x_4 - x_3) + (x_1 - x_3) + (x_2 - x_3) \\ \ddot{x}_4 &= (x_1 - x_4) + (x_2 - x_4) + (x_3 - x_4)\end{aligned}$$

which has the form
 $\ddot{x} = Ax$, where the
 4×4 matrix

$$A = \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix}, v_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, v_2 = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, v_3 = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, v_4 = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{are}$$

the eigenvectors to the eigenvalues $\lambda_1 = 0, \lambda_2 = -4, \lambda_3 = -4, \lambda_4 = -4$. With $S = [v_1 v_2 v_3 v_4]$, the matrix $B = S^{-1}BS$ is diagonal with entries $0, -4, -4, -4$. The coordinates $y_i = Sx_i$ satisfy $\ddot{y}_1 = 0, \ddot{y}_2 = -4y_2, \ddot{y}_3 = -4y_3, \ddot{y}_4 = -4y_4$ which we can solve y_0 which is the center of mass satisfies $y_0 = a + bt$ (move molecule with constant speed). The motions $y_i = a_i \cos(2t) + b_i \sin(2t)$ of the other eigenvectors are oscillations, called **normal modes**. The general motion of the molecule is a superposition of these modes.

Homework due April 20, 2011

- 1 What is the probability that an upper triangular 3×3 matrix with entries 0 and 1 is diagonalizable?
- 2 Which of the following matrices are similar?

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

- 3 Diagonalize the following matrix in the complex:

$$A = \begin{bmatrix} 2 & -3 & 0 & 0 \\ 3 & 2 & 0 & 0 \\ 0 & 0 & 5 & 6 \\ 0 & 0 & 6 & 5 \end{bmatrix}$$

⁴We grabbed the pdb Molecule files from <http://www.sci.ouc.bc.ca>, translated them with "povchem" from .pdb to .pov rendered them under Povray.

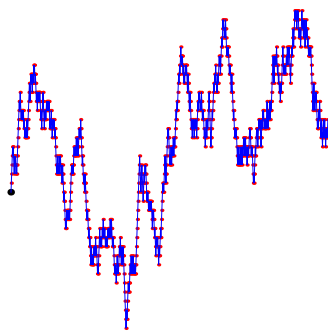
Lecture 31: The law of large numbers

A sequence of random variables is called IID abbreviating **independent, identically distributed** if they all have the same distribution and if they are independent.

We assume that all random variables have a finite variance $\text{Var}[X]$ and expectation $E[X]$.

A sequence of random variables defines a **random walk** $S_n = \sum_{k=1}^n X_k$. The interpretation is that X_k are the individual steps. If we take n steps, we reach S_n .

Here is a typical trajectory of a random walk. We throw a dice and if the dice shows head we go up, if the dice shows tail, we go down.



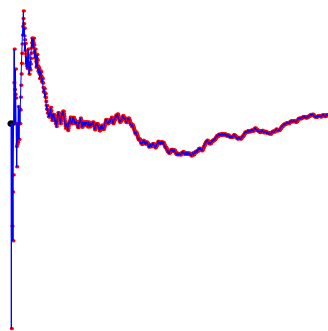
The following result is one of the three most important results in probability theory:

Law of large numbers. For almost all ω , we have $S_n/n \rightarrow E[X]$.

Proof. We only prove the weak law of large numbers which deals with a weaker convergence: We have $\text{Var}[S_n/n] = n\text{Var}[X]/n^2 = \text{Var}[X]/n$ so that by Chebyshev's theorem

$$P[|S_n/n - E[X]| < \epsilon] \leq \text{Var}[X]/n\epsilon^2$$

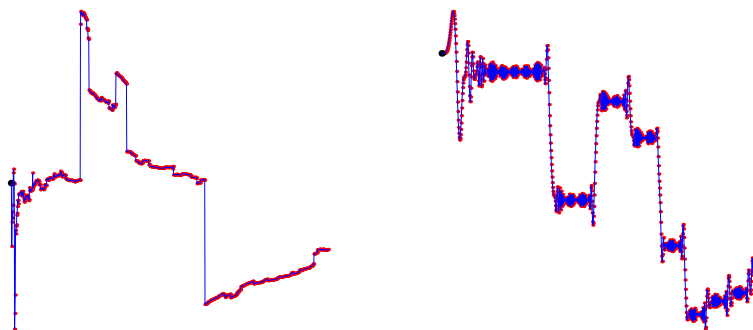
for $n \rightarrow \infty$. We see that the probability that S_n/n deviates by a certain amount from the mean goes to zero as $n \rightarrow \infty$. The strong law would need a half an hour for a careful proof.¹



- 1 If X_i are random variables which take the values 0, 1 and 1 is chosen with probability p , then S_n has the binomial distribution and $E[S_n] = np$. Since $E[X] = p$, the law of large numbers is satisfied.
- 2 If X_i are random variables which take the value 1, -1 with equal probability 1/2, then S_n is a symmetric random walk. In this case, $S_n/n \rightarrow 0$.
- 3 Here is a strange paradox called the **Martingale paradox**. We try it out in class. Go into a Casino and play the doubling strategy. Enter 1 dollar, if you lose, double to 2 dollars, if you lose, double to 4 dollars etc. The first time you win, stop and leave the Casino. You won 1 dollar because you lost maybe 4 times and $1 + 2 + 4 + 8 = 15$ dollars but won 16. The paradox is that the expected win is zero and in actual Casinos even negative. The usual solution to the paradox is that as longer you play as more you win but also increase the chance that you lose huge leading to a zero net win. It does not quite solve the paradox because in a Casino where you are allowed to borrow arbitrary amounts and where no bet limit exists, you can not lose.
How close is S_n/n to $E[X]$? Experiment:
- 4 Throw a dice n times and add up the total number S_n of eyes. Estimate $S_n/n - E[X]$ with experiments. Below is example code for Mathematica. How fast does the error decrease?

```
f[n_] := Sum[Random[Integer, 5] + 1, {n}]/n - 7/2;
data = Table[{k, f[k]}, {k, 1000}];
Fit[data, {1, Exp[-x]}, x]
```

- 5 Here is the situation where the random variables are Cauchy distributed. The expectation is not defined. The left picture below shows this situation.
- 6 What happens if we relax the assumption that the random variables are uncorrelated? The illustration to the right below shows an experiment, where we take a periodic function $f(x)$ and an irrational number α and where $X_k(x) = f(k\alpha)$.



It turns out that no randomness is necessary to establish the strong law of large numbers. It is enough to have "ergodicity"

A probability preserving transformation T on a probability space (Ω, P) is called **ergodic** if every event A which is left invariant has probability 0 or 1.

¹O.Knill, Probability and Stochastic Processes with applications, 2009

- 7 If Ω is the interval $[0, 1]$ with measure $P[[c, d]] = d - c$, then $T(x) = x + \alpha \bmod 1$ is ergodic if α is irrational.

Birkhoff's ergodic theorem. If $X_k = f(T^k x)$ is a sequence of random variables obtained from an ergodic process, then $S_n(\omega)/n \rightarrow E[X]$ for almost all ω .

This theorem is the reason that ideas from probability theory can be applied in much more general contexts, like in **number theory** or in **celestial mechanics**.

Application: normal numbers

A real number is called **normal** to base 10 if in its decimal expansion, every digit appears with the same frequency $1/10$.

Almost every real number is normal

The reason is that we can look at the k 'th digit of a number as the value of a random variable $X_k(\omega)$ where $\omega \in [0, 1]$. These random variables are all independent and have the same distribution. For the digit 7 for example, look at the random variables $Y_k(\omega) = \begin{cases} 1 & \omega_k = 7 \\ 0 & \text{else} \end{cases}$ which have expectation $1/10$. The average $S_n(\omega)/n =$ "number of digits 7 in the first k digits of the decimal expansion" of ω converges to $1/10$ by the law of large numbers. We can do that for any digit and therefore, almost all numbers are normal.

Application: Monte Carlo integration

The limit

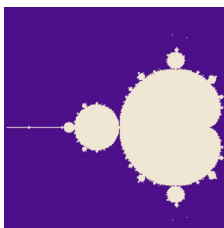
$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(x_k)$$

where x_k are IID random variables in $[a, b]$ is called the **Monte-Carlo integral**.

The Monte Carlo integral is the same than the Riemann integral for continuous functions.

We can use this to compute areas of complicated regions:

The following two lines evaluate the **area of the Mandelbrot fractal** using Monte Carlo integration. The function F is equal to 1, if the parameter value c of the quadratic map $z \rightarrow z^2 + c$ is in the Mandelbrot set and 0 else. It shoots 100'000 random points and counts what fraction of the square of area 9 is covered by the set. Numerical experiments give values close to the actual value around 1.51.... One could use more points to get more accurate estimates.



```
F[c_] := Block[{z=c, u=1}, Do[z=N[z^2+c]; If[Abs[z]>3, u=0; z=3], {99}]; u];
M=10^5; Sum[F[-2.5+3 Random[]+I(-1.5+3 Random[])], {M}]/(9.0/M)
```

Application: first significant digits

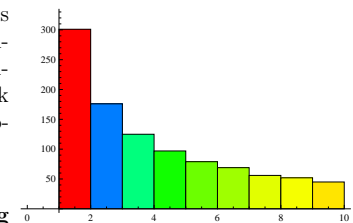
If you look at the distribution of the first digits of the numbers 2^n : 2, 4, 8, 1, 3, 6, 1, 2, Lets experiment:

```
data=Table[First[IntegerDigits[2^n]], {n, 1, 100}];
Histogram[data, 10]
```

Interestingly, the digits are not equally distributed. The smaller digits appear with larger probability. This is called **Benford's law** and it is abundant. Lists of numbers of real-life source data are distributed in a nonuniform way. Examples are bills, accounting books, stock prices. Benford's law states the digit k appears with probability

$$p_k = \log_{10}(k+1) - \log_{10}(k)$$

where $k = 1, \dots, 9$. It is useful for **forensic accounting** or investigating **election frauds**.



The probability distribution p_k on $\{1, \dots, 9\}$ is called the **Benford distribution**.

The reason for this distribution is that it is uniform on a logarithmic scale. Since numbers x for which the first digit is 1 satisfy $0 \leq \log(x) \bmod 1 < \log_{10}(2) = 0.301\dots$, the chance to have a digit 1 is about 30 percent. The numbers x for which the first digit is 6 satisfy $0.778\dots = \log_{10}(6) \leq \log(x) \bmod 1 < \log_{10}(7) = 0.845\dots$, the chance to see a 6 is about 6.7 percent.

Homework due April 20, 2011

- 1 Look the first significant digit X_n of the sequence 2^n . For example $X_5 = 3$ because $2^5 = 32$. To which number does S_n/n converge? We know that X_n has the Benford distribution $P[X_n = k] = p_k$. [The actual process which generates the random variables is the irrational rotation $x \rightarrow x + \log_{10}(2) \bmod 1$ which is ergodic since the $\log_{10}(2)$ is irrational. You can therefore assume that the law of large numbers (Birkhoff's generalization) applies.]
- 2 By going through the proof of the weak law of large numbers, does the proof also work if X_n are only uncorrelated?
- 3 Assume A_n is a sequence of $n \times n$ upper-triangular random matrices for which each entry is either 1 or 2 and 2 is chosen with probability $p = 1/4$.
 - a) What can you say about $\text{tr}(A_n)/n$ in the limit $n \rightarrow \infty$?
 - b) What can you say about $\log(\det(A_n))/n$ in the limit $n \rightarrow \infty$?

Lecture 32: Central limit theorem

The central limit theorem explains why the normal distribution

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

is prevalent. If we add independent random variables and normalize them so that the mean is zero and the standard deviation is 1, then the distribution of the sum converges to the normal distribution.

Given a random variable X with expectation m and standard deviation σ define the **normalized random variable** $X^* = (X - m)/\sigma$.

The normalized random variable has the mean 0 and the standard deviation 1. The standard normal distribution mentioned above is an example. We have seen that S_n/n converges to a definite number if the random variables are uncorrelated. We have also seen that the standard deviation of S_n/n goes to zero.

A sequence X_n of random variables **converges in distribution** to a random variable X if for every trigonometric polynomial f , we have $E[f(X_n)] \rightarrow E[f(X)]$.

This means that $E[\cos(tX_n)] \rightarrow E[\cos(tX)]$ or $E[\sin(tX_n)] \rightarrow E[\sin(tX)]$ converge for every t . We can combine the cos and sin to $\exp(itx) = \cos(tx) + i\sin(tx)$ and cover both at once by showing $E[e^{itX_n}] \rightarrow E[e^{itX}]$ for $n \rightarrow \infty$. So, checking the last statement for every t is equivalent to check the convergence in distribution.

The function $\phi_X(t) = E[e^{itX}]$ is called the **characteristic function** of a random variable X .

Convergence in distribution is equivalent to the statement that the cumulative distribution functions $F_n(c) = P[X_n \leq c]$ converge to $F(c) = P[X \leq c]$ at every point c at which F is continuous. An other statement which is intuitive is that if the distribution is such that all moments $E[X^m]$ exist, it is enough to check that the moments $E[X_n^m]$ converge to $E[X^m]$ for all m . Trigonometric polynomials are preferred because they do not require the boundedness of the moments. "Convergence in distribution" is also called "convergence in law" or "weak convergence".

The following result is one of the most important theorems in probability theory. It explains why the standard normal distribution is so important.

Central limit theorem: Given a sequence of IID random variables with finite mean and variance and finite $E[X^3]$. Then S_n^* converges in distribution to the standard normal distribution.

Proof. Let X be a $N(0, 1)$ distributed random variable. We show $E[e^{itS_n^*}] \rightarrow E[e^{itX}]$ for any fixed t . Since any of the two random variables X_k, X_l are independent,

$$E[\exp(i(X_k + X_l))] = E[\exp(iX_k)]E[\exp(iX_l)] .$$

More generally

$$E[\exp(itS_n)] = E[\exp(it(X_1 + \dots + X_n))] = E[\exp(itX_1)] \cdots E[\exp(itX_n)] .$$

Since we normalize the random variables, we can assume that each X_k has zero expectation and variance 1. If $E[X_k] = 0, E[X_k^2] = 1$, we have for each of the n factors

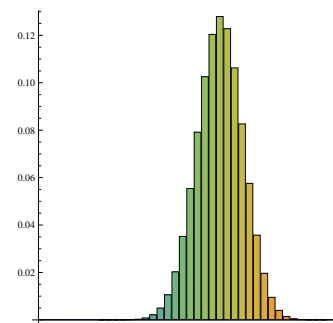
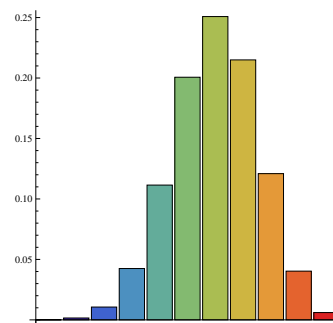
$$E[e^{itX_k/\sqrt{n}}] = (1 - \frac{t^2}{2n} - \frac{it^3 E[X^3]}{3n^{3/2}} + \dots) .$$

Using $e^{-t^2/2} = 1 - t^2/2 + \dots$, we get

$$E[e^{itS_n/\sqrt{n}}] = (1 - \frac{t^2}{2n} + R_n/(n^{3/2}))^n \rightarrow e^{-t^2/2} .$$

The last step uses a Taylor remainder term $R_n/n^{3/2}$ term. It is here that the $E[X^3] < \infty$ assumption has been used. The statement now follows from

$$E[e^{itX}] = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx = e^{-t^2/2} .$$



- 1 We throw a fair dice n times. The distribution of S_n is a binomial distribution. The mean is np and the standard deviation is $\sqrt{np(1-p)}$. The distribution of $(S_n - np)/\sqrt{np(1-p)}$ looks close to the normal distribution. This special case of the central limit theorem is called the **de Moivre-Laplace** theorem. It was proven by de Moivre in 1730 in the case $p = 1/2$ and in 1812 for general $0 < p < 1$ by Laplace.

Statistical inference

For the following see also Cliffs notes page 89-93 (section 14.6): what is the probability that the average S_n/n is within ϵ to the mean $E[X]$?

The probability that S_n/n deviates more than $R\sigma/\sqrt{n}$ from $E[X]$ can for large n be estimated by

$$\frac{1}{\sqrt{2\pi}} \int_R^\infty e^{-x^2/2} dx .$$

Proof. Let $p = E[X]$ denote the mean of X_k and σ the standard deviation. Denote by X a random variable which has the standard normal distribution $N(0,1)$. We use the notation $X \sim Y$ if X and Y are close in distribution. By the central limit theorem

$$\frac{S_n - np}{\sqrt{n}\sigma} \sim X .$$

Dividing nominator and denominator by n gives $\frac{\sqrt{n}}{\sigma}(\frac{S_n}{n} - p) \sim X$ so that

$$\frac{S_n}{n} - p \sim X \frac{\sigma}{\sqrt{n}} .$$

The term σ/\sqrt{n} is called the **standard error**. The central limit theorem gives some insight why the standard error is important.

In scientific publications, the standard error should be displayed rather than the standard deviation.¹

A squirrel accidentally drank from liquor leaking from a garbage bag. Tipsy, it walks forth and back on a telephone cable, randomly taking a step of one foot forward or backwards each second. How far do we expect him to be drifted off after 3 minutes

2 **Answer:** The mean is zero, the standard deviation is 1. By the central limit theorem we expect the drift off from $p = 0$ to be \sqrt{n} because

$$S_n \sim X\sigma\sqrt{n} .$$

That means we can expect the squirrel to be within $\sqrt{180} = 13$ feet. The chance to see it in this neighborhood is about 2/3 because the probability to be within the standard deviation interval is about 2/3 (see homework).



A probability space and a random variable X define a **null hypothesis**, a model for your experiment. Assume you measure $X = c$. Assuming c is larger than the expectation, the **P-value** of this experiment is defined as $P[X \geq c]$. If c is smaller than the expectation, we would define the P-value as $P[X \leq c]$.

The **P-value** is the probability that the test statistics is at least as extreme as the experiment.

3 The assumption in the previous problem that our squirrel is completely drunk the **null hypothesis**. Assume we observe the squirrel after 3 minutes at 20 feet from the original

place. What is the **P-value** of this observation? The **P-value** is $P[S_{180} \geq 20]$. Since $S_{180}/\sqrt{180}$ is close to normal and $c = 20/\sqrt{180} = 1.49..$ we can estimate the **P** value as

$$\int_c^\infty f_{normal}(x) dx = 0.068..$$

Since we know that the actual distribution is a Binomial distribution, we could have computed the P-value exactly as $\sum_{k=100}^{180} B(180, k)p^k(1-p)^{n-k} = 0.078$ with $p = 1/2$. A P-value smaller than 5 percent is called **significant**. We would have to reject the null hypothesis and the squirrel is not drunk. In our case, the experiment was not significant.

The central limit theorem is so important because it gives us a tool to estimate the **P-value**. It is much better in general than the estimate given by Chebyshev.

¹Geoff Cumming, Fiona Fidler, I and David L. Vaux: Error bars in experimental biology, Journal of Cell biology, 177, 1, 2007 7-11

Homework due April 20, 2011

- 1 Look at the standard normal probability distribution function

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

Verify that for this distribution, there is a probability of 68 percent to land within the interval $[-\sigma, \sigma]$ is slightly larger than $2/3$.

You have verified the rule of thumb:

For the normal distribution $N(m, \sigma)$, the chance to be in the interval $[m - \sigma, m + \sigma]$ is about two thirds.

- 2 In many popular children games in which one has to throw a dice and then move forward by the number of eyes seen. After n rounds, we are at position S_n . If you play such a game, find an interval of positions so that you expect to be with at least $2/3$ percent chance in that interval after n rounds.



Remark: An example is the game "Ladder" which kids play a lot in Switzerland: it is a random walk with drift 3.5. What makes the game exciting are occasional accelerations or setbacks. Mathematically it is a **Markov process**. If you hit certain fields, you get pushed ahead (sometimes significantly) but for other fields, you can almost lose everything. The game stays interesting because even if you are ahead you can still end up last or you trail behind all the game and win in the end.

- 3 We play in a Casino with the following version of the martingale strategy. We play all evening and bet one dollar on black until we reach our goal of winning 10 dollars. Assume each game lasts a minute. How long do we expect to wait until we can go home? (You can assume that the game is fair and that you win or lose with probability $1/2$ in each game.)

Remark: this strategy appears frequently in movies, usually when characters are desperate. Examples are "Run Lola Run", "Casino Royale" or "Hangover".



Lecture 33: Markov matrices

A $n \times n$ matrix is called a **Markov matrix** if all entries are nonnegative and the sum of each column vector is equal to 1.

1 The matrix

$$A = \begin{bmatrix} 1/2 & 1/3 \\ 1/2 & 2/3 \end{bmatrix}$$

is a Markov matrix.

Markov matrices are also called **stochastic matrices**. Many authors write the transpose of the matrix and apply the matrix to the right of a row vector. In linear algebra we write Ap . This is of course equivalent.

Lets call a vector with nonnegative entries p_k for which all the p_k add up to 1 a **stochastic vector**. For a stochastic matrix, every column is a stochastic vector.

If p is a stochastic vector and A is a stochastic matrix, then Ap is a stochastic vector.

Proof. Let v_1, \dots, v_n be the column vectors of A . Then

$$Ap = \begin{bmatrix} p_1 \\ p_2 \\ \dots \\ p_n \end{bmatrix} = p_1 v_1 + \dots + p_n v_n.$$

If we sum this up we get $p_1 + p_2 + \dots + p_n = 1$.

A Markov matrix A always has an eigenvalue 1. All other eigenvalues are in absolute value smaller or equal to 1.

Proof. For the transpose matrix A^T , the sum of the row vectors is equal to 1. The matrix A^T therefore has the eigenvector

$$\begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}.$$

Because A and A^T have the same determinant also $A - \lambda I_n$ and $A^T - \lambda I_n$ have the same determinant so that the eigenvalues of A and A^T are the same. With A^T having an eigenvalue 1 also A has an eigenvalue 1.

Assume now that v is an eigenvector with an eigenvalue $|\lambda| > 1$. Then $A^n v = |\lambda|^n v$ has exponentially growing length for $n \rightarrow \infty$. This implies that there is for large n one coefficient $[A^n]_{ij}$ which is larger than 1. But A^n is a stochastic matrix (see homework) and has all entries ≤ 1 . The assumption of an eigenvalue larger than 1 can not be valid.

2 The example

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$$

shows that a Markov matrix can have zero eigenvalues and determinant.

3 The example

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

shows that a Markov matrix can have negative eigenvalues. and determinant.

4 The example

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

shows that a Markov matrix can have several eigenvalues 1.

5 If all entries are positive and A is a 2×2 Markov matrix, then there is only one eigenvalue 1 and one eigenvalue smaller than 1.

$$A = \begin{bmatrix} a & b \\ 1-a & 1-b \end{bmatrix}$$

Proof: we have seen that there is one eigenvalue 1 because A^T has $[1, 1]^T$ as an eigenvector. The trace of A is $1 + a - b$ which is smaller than 2. Because the trace is the sum of the eigenvalues, the second eigenvalue is smaller than 1.

6 The example

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

shows that a Markov matrix can have complex eigenvalues and that Markov matrices can be orthogonal.

The following example shows that stochastic matrices do not need to be diagonalizable, not even in the complex:

7 The matrix

$$A = \begin{bmatrix} 5/12 & 1/4 & 1/3 \\ 5/12 & 1/4 & 1/3 \\ 1/6 & 1/2 & 1/3 \end{bmatrix}$$

is a stochastic matrix, even doubly stochastic. Its transpose is stochastic too. Its row reduced echelon form is

$$A = \begin{bmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 0 & 0 & 0 \end{bmatrix}$$

so that it has a one dimensional kernel. Its characteristic polynomial is $f_A(x) = x^2 - x^3$ which shows that the eigenvalues are 1, 0, 0. The algebraic multiplicity of 0 is 2. The geometric multiplicity of 0 is 1. The matrix is not diagonalizable. ¹

¹This example appeared in <http://mathoverflow.net/questions/51887/non-diagonalizable-doubly-stochastic-matrices>

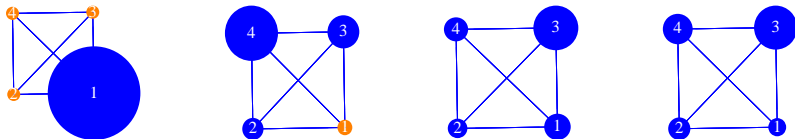
The eigenvector v to the eigenvalue 1 is called the **stable equilibrium distribution** of A . It is also called **Perron-Frobenius eigenvector**.

Typically, the discrete dynamical system converges to the stable equilibrium. But the above rotation matrix shows that we do not have to have convergence at all.

8 Assume

$$A = \begin{bmatrix} 0 & 0.1 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.2 & 0.1 \\ 0.3 & 0.2 & 0.5 & 0.4 \\ 0.5 & 0.4 & 0.1 & 0.2 \end{bmatrix}.$$

Lets visualize this. We start with the vector $\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$.



Many games are Markov games. Lets look at a simple example of a **mini monopoly**, where no property is bought:

9 Lets have a simple “monopoly” game with 6 fields. We start at field 1 and throw a coin. If the coin shows head, we move 2 fields forward. If the coin shows tail, we move back to the field number 2. If you reach the end, you win a dollar. If you overshoot you pay a fee of a dollar and move to the first field. **Question:** in the long term, do you win or lose if $p_6 - p_5$ measures this win? Here $p = (p_1, p_2, p_3, p_4, p_5, p_6)$ is the stable equilibrium solution with eigenvalue 1 of the game.

10 Take the same example but now throw also a dice and move with probability $1/6$. The matrix is now

[illegible]

In the homework, you will see that there is only one stable equilibrium now.

Homework due April 27, 2011

1 Find the stable equilibrium distribution of the matrix

$$A = \begin{bmatrix} 1/2 & 1/3 \\ 1/2 & 2/3 \end{bmatrix}.$$

2 a) Verify that the product of two Markov matrices is a Markov matrix.
b) Is the inverse of a Markov matrix always a Markov matrix? Hint for a): Let A, B be Markov matrices. You have to verify that BAe_k is a stochastic vector.

3 Find all the eigenvalues and eigenvectors of the doubly stochastic matrix in the modified game above

[illegible]

Lecture 34: Perron Frobenius theorem

This is a second lecture on Markov processes. We want to see why the following result is true:

If all entries of a Markov matrix A are positive then A has a unique equilibrium: there is only one eigenvalue 1. All other eigenvalues are smaller than 1.

To illustrate the importance of the result, we look how it is used in chaos theory and how it can be used for search engines to rank pages.

- 1 The matrix

$$A = \begin{bmatrix} 1/2 & 1/3 \\ 1/2 & 2/3 \end{bmatrix}$$

is a Markov matrix for which all entries are positive. The eigenvalue 1 is unique because the sum of the eigenvalues is $1/2 + 2/3 < 2$.

- 2 We have already proven Perron-Frobenius for 2×2 Markov matrices: such a matrix is of the form

$$A = \begin{bmatrix} a & b \\ 1-a & 1-b \end{bmatrix}$$

and has an eigenvalue 1 and a second eigenvalue smaller than 1 because $\text{tr}(A)$ the sum of the eigenvalues is smaller than 2.

- 3 Lets give a brute force proof of the Perron-Frobenius theorem in the case of 3×3 matrices: such a matrix is of the form

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ 1-a-d & 1-b-e & 1-c-f \end{bmatrix}.$$

and has an eigenvalue 1. The determinant is $D = c(d-e) + a(e-f) + b(-d+f)$ and is the product of the two remaining eigenvalues. The trace is $1 + (a-c) + (e-f)$ so that $T = (a-c) + (e-f)$ is the sum of the two remaining eigenvalues. An ugly verification shows that these eigenvalues are in absolute value smaller than 1.

The Markov assumption is actually not needed. Here is a more general statement which is useful in other parts mathematics. It is also one the theorems with the most applications like **Leontief's models** in economics, **chaos theory** in dynamical systems or **page rank** for search engines.

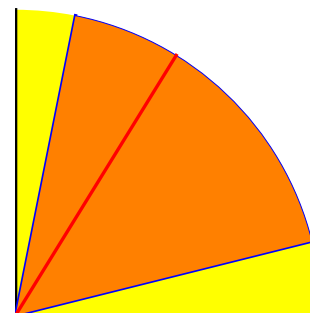
Perron Frobenius theorem: If all entries of a $n \times n$ matrix A are positive, then it has a unique maximal eigenvalue. Its eigenvector has positive entries.

Proof. The proof is quite geometric and intuitive. Look at the sphere $x_1^2 + \dots + x_n^2 = 1$ and intersect it with the space $\{x_1 \geq 0, \dots, x_n \geq 0\}$ which is a quadrant for $n = 2$ and octant for $n = 3$. This gives a closed, bounded set X . The matrix A defines a map $T(v) = Av/|Av|$ on X because the entries of the matrix are nonnegative. Because they are positive, TX is contained in the interior of X . This map is a **contraction**, there exists $0 < k < 1$ such that $d(Tx, Ty) \leq kd(x, y)$ where d

is the geodesic sphere distance. Such a map has a unique fixed point v by Banach's fixed point theorem. This is the eigenvector $Av = \lambda v$ we were looking for. We have seen now that on X , there is only one eigenvector. Every other eigenvector $Aw = \mu w$ must have a coordinate entry which is negative. Write $|w|$ for the vector with coordinates $|w_j|$. The computation

$$|\mu||w|_i = |\mu w_i| = \left| \sum_j A_{ij} w_j \right| \leq \sum_j |A_{ij}| |w_j| = \sum_j A_{ij} |w_j| = (A|w|)_i$$

shows that $|\mu|L \leq \lambda L$ because $(A|w|)$ is a vector with length smaller than λL , where L is the length of w . From $|\mu|L \leq \lambda L$ with nonzero L we get $|\mu| \leq \lambda$. The first " \leq " which appears in the displayed formula is however an inequality for some i if one of the coordinate entries is negative. Having established $|\mu| < \lambda$ the proof is finished.



Remark. The theorem generalizes to situations considered in **chaos theory**, where **products of random matrices** are considered which all have the same distribution but which do not need to be independent. Given such a sequence of random matrices A_k , define $S_n = A_n \cdot A_{n-1} \cdots A_1$. This is a non commutative analogue of the random walk $S_n = X_1 + \dots + X_n$ for usual random variables. But it is much more intricate because matrices do not commute. Laws of large numbers are now more subtle.

Application: Chaos

The **Lyapunov exponent** of a random sequence of matrices is defined as

$$\lim_{n \rightarrow \infty} \frac{1}{2n} \log \lambda(S_n^T S_n),$$

where $\lambda(B)$ is the maximal eigenvalue of the symmetric matrix $S_n^T S_n$.

Here is a prototype result in Chaos theory due to Anosov for which the proof of Perron-Frobenius can be modified using different contractions. It can be seen as an example of a noncommutative law of large numbers:

If A_k is a sequence of identically distributed random positive matrices of determinant 1, then the Lyapunov exponent is positive.

- 4 Let A_k be either $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ or $\begin{bmatrix} 3 & 2 \\ 1 & 1 \end{bmatrix}$ with probability $1/2$. Since the matrices do not commute, we can not determine the long term behavior of S_n so easily and laws of large numbers do not apply. The Perron-Frobenius generalization above however shows that still, S_n grows exponentially fast.

Positive Lyapunov exponent is also called **sensitive dependence on initial conditions** for the system or simply dubbed **"chaos"**. Nearby trajectories will deviate exponentially. Edward Lorentz, who studied about 50 years ago models of the complicated equations which govern our weather stated this in a poetic way in 1972:

The flap of a butterfly's wing in Brazil can set off a tornado in Texas.

Unfortunately, Mathematics is quite weak still to mathematically prove positive Lyapunov exponents if the system does not a priori feature positive matrices. There are cases which can be settled quite easily. For example, if the matrices A_k are IID random matrices of determinant 1 and eigenvalues 1 have not full probability, then the Lyapunov exponent is positive due to work of Fuerstenberg and others. In real systems, like for the motion of our solar system or particles in a box, positive Lyapunov exponents is measured but can not be proven yet. Even for simple toy systems like $S_n = dT^n$, where dT is the Jacobean of a map T like $T(x, y) = (2x - c \sin(x), y)$ and T^n is the n 'th iterate, things are unsettled. One measures $\lambda \geq \log(c/2)$ but is unable to prove it yet. For our real weather system, where the Navier stokes equations apply, one is even more helpless. One does not even know whether trajectories exist for all times. This existence problem looks like an esoteric ontological question if it were not for the fact that a one million dollar bounty is offered for its solution.

Application: Pagerank

A set of nodes with connections is a **graph**. Any network can be described by a graph. The link structure of the web forms a graph, where the individual websites are the nodes and if there is an arrow from site a_i to site a_j if a_i links to a_j . The adjacency matrix A of this graph is called the **web graph**. If there are n sites, then the adjacency matrix is a $n \times n$ matrix with entries $A_{ij} = 1$ if there exists a link from a_j to a_i . If we divide each column by the number of 1 in that column, we obtain a Markov matrix A which is called the **normalized web matrix**. Define the matrix E which satisfies $E_{ij} = 1/n$ for all i, j . The graduate students and later entrepreneurs **Sergey Brin** and **Lawrence Page** had in 1996 the following one billion dollar idea:

The **Google matrix** is the matrix $G = dA + (1 - d)E$, where $0 < d < 1$ is a parameter called **damping factor** and A is the Markov matrix obtained from the adjacency matrix by scaling the rows to become stochastic matrices. This is a $n \times n$ Markov matrix with eigenvalue 1.

Its Perron-Frobenius eigenvector v scaled so that the largest value is 10 is called **page rank** of the damping factor d .

The **page rank equation** is

$$[dA + (1 - d)E]v = v$$

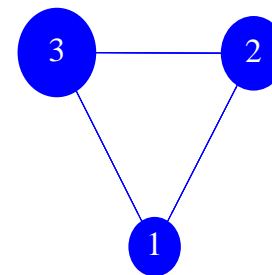
The damping factor can look a bit mysterious. Brin and Page write:

"PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. And, the d damping factor is the probability at each page the "random surfer" will get bored and request another random page. One important variation is to only add the damping factor d to a single page, or a group of pages. This allows for personalization and can make it nearly impossible to deliberately mislead the system in order to get a higher ranking. We have several other extensions to PageRank." ¹

- 5 Consider 3 sites A, B, C , where A is connected to B, C and B is connected to C and C is connected to A . Find the page rank to $d = 0.1$. **Solution.** The adjacency matrix of the

graph is $A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$. The Google matrix is

$$G = \begin{bmatrix} d & d & d \\ d & d & d \\ d & d & d \end{bmatrix} / 3 + (1 - d) \begin{bmatrix} 0 & 0 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{bmatrix}.$$



It is said now that page rank is the world's largest matrix computation. The $n \times n$ matrix is huge. It was 8.1 billion 5 years ago. ²

Homework due April 27, 2011

- 1 Verify that if a Markov matrix A has the property that A^2 has only positive entries, then A has a unique eigenvalue 1.
- 2 Take 4 sites A, B, C, D where A links to B, C, D , and B links to C, D and C links to D and D links to A . Find the Google matrix with the damping factor $1/2$.
- 3 Determine the Page rank of the previous system, possibly using technology like Mathematica.

¹<http://infolab.stanford.edu/backrub/google.html>

²Amy Langville and Carl Meyer, *Google's PageRank and Beyond*, Princeton University Press, 2006.

Lecture 35: Symmetric matrices

In this lecture, we look at the spectrum of symmetric matrices. Symmetric matrices appear in **geometry**, for example, when introducing **more general dot products** $v \cdot Av$ or in **statistics** as **correlation matrices** $\text{Cov}[X_k, X_l]$ or in quantum mechanics as **observables** or in **neural networks** as **learning maps** $x \mapsto \text{sign}(Wx)$ or in **graph theory** as **adjacency matrices**. Symmetric matrices play the same role as the **real numbers** do among the complex numbers. Their eigenvalues often have physical or geometrical interpretations. One can also calculate with symmetric matrices like with numbers: for example, we can solve $B^2 = A$ for B if A is symmetric matrix and B is square root of A .) This is not possible in general. There is no matrix B for example such that $B^2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Recall the following definition:

A real matrix is called **symmetric** if $A^T = A$. Symmetric matrices are also called selfadjoint. For complex matrices we would ask $A^* = \overline{A}^T = A$.

1 The matrix

$$A = \begin{bmatrix} 3 & 4 \\ 4 & 3 \end{bmatrix}$$

is symmetric.

A symmetric matrix has real eigenvalues.

Proof. Extend the dot product to complex vectors by $(v, w) = \sum_i \bar{v}_i w_i$, where \bar{v} is the complex conjugate. For real vectors it is the usual dot product $(v, w) = v \cdot w$. The new product has the property $(Av, w) = (v, A^T w)$ for real matrices A and $(\lambda v, w) = \bar{\lambda}(v, w)$ as well as $(v, \lambda w) = \lambda(v, w)$. Now $\bar{\lambda}(v, v) = (\lambda v, v) = (Av, v) = (v, A^T v) = (v, Av) = (v, \lambda v) = \lambda(v, v)$ shows that $\bar{\lambda} = \lambda$ because $(v, v) \neq 0$ for $v \neq 0$.

There is an orthogonal eigenbasis for a symmetric matrix A if all the eigenvalues of A are different.

Proof. Assume $Av = \lambda v$ and $Aw = \mu w$. The relation

$$\lambda(v, w) = (\lambda v, w) = (Av, w) = (v, A^T w) = (v, Aw) = (v, \mu w) = \mu(v, w)$$

is only possible if $(v, w) = 0$ if $\lambda \neq \mu$.

Spectral theorem A symmetric matrix can be diagonalized with an orthonormal matrix S .

The result is called spectral theorem. I present now an intuitive proof, which gives more insight why the result is true. The linear algebra book of Bretscher has an inductive proof.

Proof. We have seen already that if all eigenvalues are different, there is an eigenbasis and diagonalization is possible. The eigenvectors are all orthogonal and $B = S^{-1}AS$ is diagonal containing the eigenvalues. In general, we can change the matrix A to $A = A + (C - A)t$ where C is a matrix with pairwise different eigenvalues. Then the eigenvalues are different for all except finitely many t . The orthogonal matrices S_t converges for $t \rightarrow 0$ to an orthogonal matrix S and S diagonalizes A .¹

Why could we not perturb a general matrix A_t to have disjoint eigenvalues and A_t could be diagonalized: $S_t^{-1}A_t S_t = B_t$? The problem is that S_t might become singular for $t \rightarrow 0$.

2 The matrix $A = \begin{bmatrix} 1 & 1 \\ 0 & 1+t \end{bmatrix}$ has the eigenvalues $1, 1+t$ and is diagonalizable for $t > 0$ but not diagonalizable for $t = 0$. What happens with the diagonalization in the limit? **Solution:** Because the matrix is upper triangular, the eigenvalues are $1, 1+t$. The eigenvector to the eigenvalue 1 is $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$. The eigenvector to the eigenvalue $1+t$ is $\begin{bmatrix} 1 \\ t \end{bmatrix}$. We see that in the limit $t \rightarrow 0$, the second eigenvector collides with the first one. For symmetric matrices, where the eigenvectors are always perpendicular to each other, such a collision can not happen.

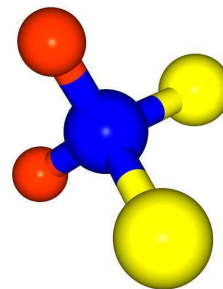
3 The **Freon molecule** (Dichlorodifluoromethane or shortly CFC-12) CCl_2F_2 has 5 atoms. It is a CFC was used in refrigerators, solvents and propellants but contributes to ozone depletion in the atmosphere. The adjacency matrix is

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

a) Verify that the matrix has the characteristic polynomial $x^5 - 4x^3$.

b) Find the eigenvalues of A .

c) Find the eigenvectors of A .

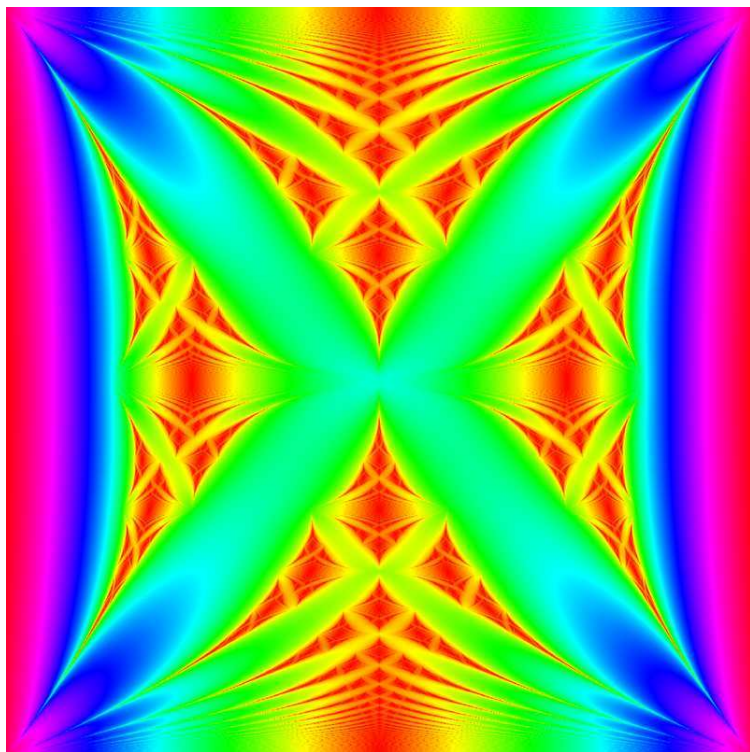


¹This is justified by a result of Neumann-Wigner who proved that the set of symmetric matrices with simple eigenvalues is path connected and dense in the linear space of all symmetric $n \times n$ matrices.

In solid state physics or quantum mechanics, one is interested in matrices like

$$L = \begin{bmatrix} \lambda \cos(\alpha) & 1 & 0 & \cdot & 0 & 1 \\ 1 & \lambda \cos(2\alpha) & 1 & \cdot & \cdot & 0 \\ 0 & 1 & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 1 & \lambda \cos((n-1)\alpha) & 1 \\ 1 & 0 & \cdot & 0 & 1 & \lambda \cos(n\alpha) \end{bmatrix}$$

It appears in models describing an electron in a periodic crystal. The eigenvalues form what one calls the **spectrum** of the matrix. A physicist is interested in it because it determines what conductivity properties the system has. This depends on α .



The picture shows the eigenvalues of L for $\lambda = 2$ with large n . The vertical axis is α which runs from $\alpha = 0$ at the bottom to $\alpha = 2\pi$ on the top. Due to its nature, the picture is called "Hofstadter butterfly". It has been popularized in the book "Gödel, Escher Bach" by Douglas Hofstadter.

Homework due April 27, 2011

- 1 For the following question, give a reason why it is true or give a counter example.
 - a) Is the sum of two symmetric matrices symmetric?
 - b) Is the product of a symmetric matrix symmetric?
 - c) Is the inverse of an invertible symmetric matrix symmetric?
 - d) If B is an arbitrary $n \times m$ matrix, is $A = B^T B$ symmetric? e) If A is similar to B and A is symmetric, then B is symmetric.
 - f) If A is similar to B with an orthogonal coordinate change S and A is symmetric, then B is symmetric.

- 2 Find all the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 10001 & 3 & 5 & 7 & 9 & 11 \\ 1 & 10003 & 5 & 7 & 9 & 11 \\ 1 & 3 & 10005 & 7 & 9 & 11 \\ 1 & 3 & 5 & 10007 & 9 & 11 \\ 1 & 3 & 5 & 7 & 10009 & 11 \\ 1 & 3 & 5 & 7 & 9 & 10011 \end{bmatrix}.$$

As usual, document all your reasoning.

- 3 Which of the following classes of linear transformations are described by symmetric?
 - a) Reflections in the plane.
 - b) Rotations in the plane.
 - c) Orthogonal projections.
 - d) Shears.

Lecture 36: Final checklist

Please see lecture 21 handout for the topics before the midterm.

Probability theory

- ☐ **Discrete random variable** $P[X = x_k] = p_k$ discrete set of values.
- ☐ **Continuous random variable** $P[X \in [a, b]] = \int_a^b f(x) dx$
- ☐ **Properties of random variables** Var, E, Cov, σ .
- ☐ **Discrete distributions** Uniform, Binomial, Poisson
- ☐ **Continuous distributions** Uniform, Normal, Exponential
- ☐ **Chebychev's theorem** $P[|X - E[X]| \geq c] \leq \frac{\text{Var}[X]}{c^2}$.
- ☐ **Random walk** $S_n = \sum_{k=1}^n X_k$.
- ☐ **Law of large numbers** $S_n/n \rightarrow E[X]$.
- ☐ **Normalized random variable** $X^* = (X - E[X])/\sigma[X]$.
- ☐ **Convergence in distribution** $E[\cos(tX_n)] \rightarrow E[\cos(tX)]$ and $E[\sin(tX_n)] \rightarrow E[\sin(tX)]$ for all t .
- ☐ **Characteristic function** $\phi_X(t) = E[e^{itX}]$.
- ☐ **Central limit theorem** $S_n^* \rightarrow N(0, 1)$ in distribution if X_k are IID random variables with $E[X^3] < \infty$.
- ☐ **Statistical inference** $\frac{S_n}{n} - E[X] \sim N(0, 1) \frac{\sigma}{\sqrt{n}}$.
- ☐ **Standard error** of X_1, \dots, X_n is σ/\sqrt{n} .
- ☐ **Stochastic vector** Nonnegative entries which add up to 1.
- ☐ **Markov Matrix = Stochastic Matrix** All columns are stochastic vectors.
- ☐ **Perron Frobenius theorem** Markov matrices have an eigenvalue 1. All other eigenvalue are smaller or equal than 1.
- ☐ **Positive matrix.** There is a unique largest eigenvalue 1.

Linear algebra

- ☐ **Determinant** $\sum_{\pi} (-1)^{|\pi|} A_{1\pi(1)} A_{2\pi(2)} \cdots A_{n\pi(n)}$
- ☐ **Laplace expansion** $\det(A) = (-1)^{1+1} A_{11} \det(B_{11}) + \cdots + (-1)^{1+n} A_{n1} \det(B_{n1})$
- ☐ **Linearity of determinants** $\det(A)$ is linear in each row.
- ☐ **Determinant after row reduction** $\det(A) = \frac{(-1)^m}{c_1 \cdots c_k} \det(\text{rref}(A))$.
- ☐ **Determinant and invertibility** $\det(A \cdot B) = \det(A) \det(B)$.
- ☐ **Determinant and volume** $|\det(A)|$ is the volume of the parallelepiped spanned by the columns.
- ☐ **Cramer's rule** A_i replaces column by b , then $x_i = \frac{\det(A_i)}{\det(A)}$ solves $Ax = b$.
- ☐ **Adjugate** A_{ij} delete row i and column j . Call $B_{ij} = (-1)^{i+j} \det(A_{ji})$ the classical adjoint
- ☐ **Inverse** $[A^{-1}]_{ij} = (-1)^{i+j} \frac{\det(A_{ji})}{\det(A)}$
- ☐ **Eigenvectors and eigenvalues** $Av = \lambda v$
- ☐ **Characteristic polynomial** $f_A(\lambda) = \det(A - \lambda I_n)$
- ☐ **Eigenvalues** Roots of characteristic polynomial.
- ☐ **Eigenspace** kernel of $A - \lambda I_n$.
- ☐ **Geometric multiplicity** of λ dimension of eigenspace of λ .
- ☐ **Algebraic multiplicity** of λ_k m if $f_A(\lambda) = (\lambda - \lambda_k)^m g(\lambda)$ and $g(\lambda_k) \neq 0$.
- ☐ **Trace of matrix** Sum $\lambda_1 + \cdots + \lambda_n$ of eigenvalues.
- ☐ **Determinant of matrix** Product $\lambda_1 \lambda_2 \cdots \lambda_n$ of eigenvalues.
- ☐ **Discrete dynamical system** orbit $A^m x$
- ☐ **Closed form solution** $x = c_1 v_1 + \cdots + c_n v_n$, $A^m x = c_1 \lambda_1^m v_1 + \cdots + c_n \lambda_n^m v_n$.
- ☐ **Similar matrices** $S^{-1}AS$.
- ☐ **Diagonalizable matrix** There is an orthogonal eigenbasis.
- ☐ **Simple spectrum** All eigenvalues are different. Matrices with simple spectrum are diagonalizable.
- ☐ **Symmetric matrices** $A^T = A$. Symmetric matrices are diagonalizable.
- ☐ **Google matrix** $G = dA + (1-d)E$, where $E_{ij} = 1/n$, A is the adjusted adjacency matrix and d is a damping factor.