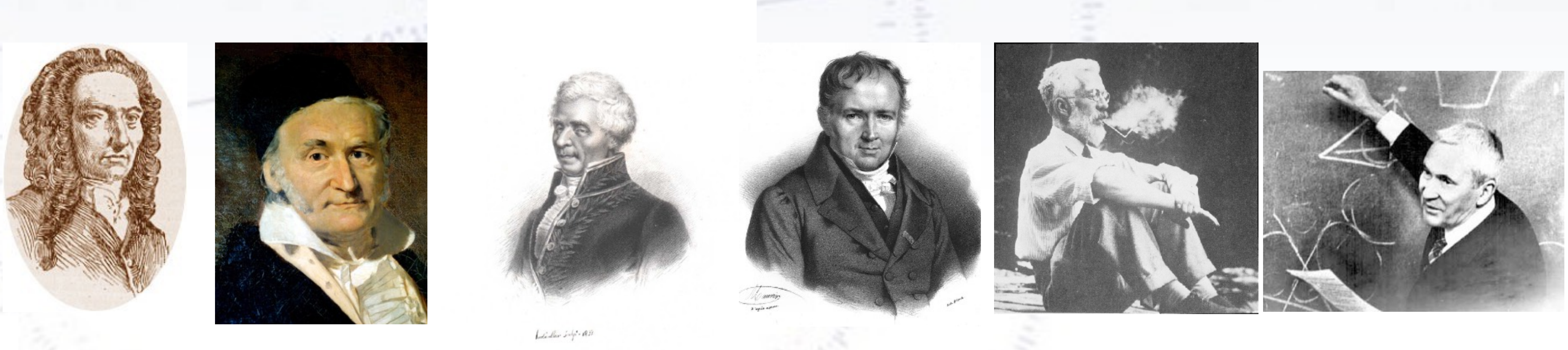
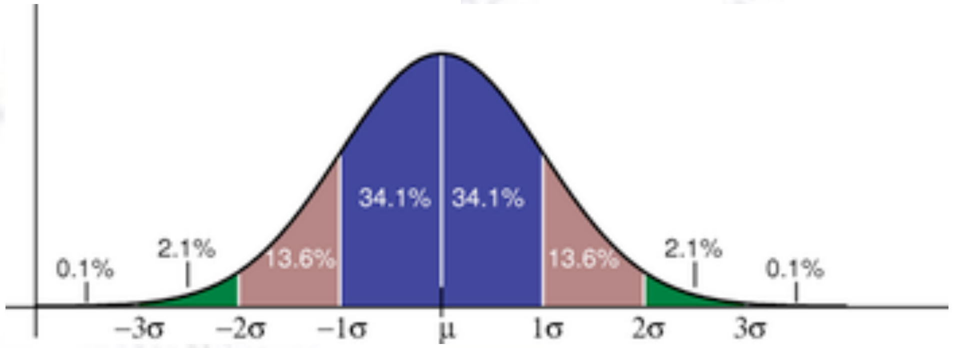
Applied Statistics

Bayesian statics and Markov Chains



Mathias Luidor Heltberg (NBI)



"Statistics is merely a quantisation of common sense"

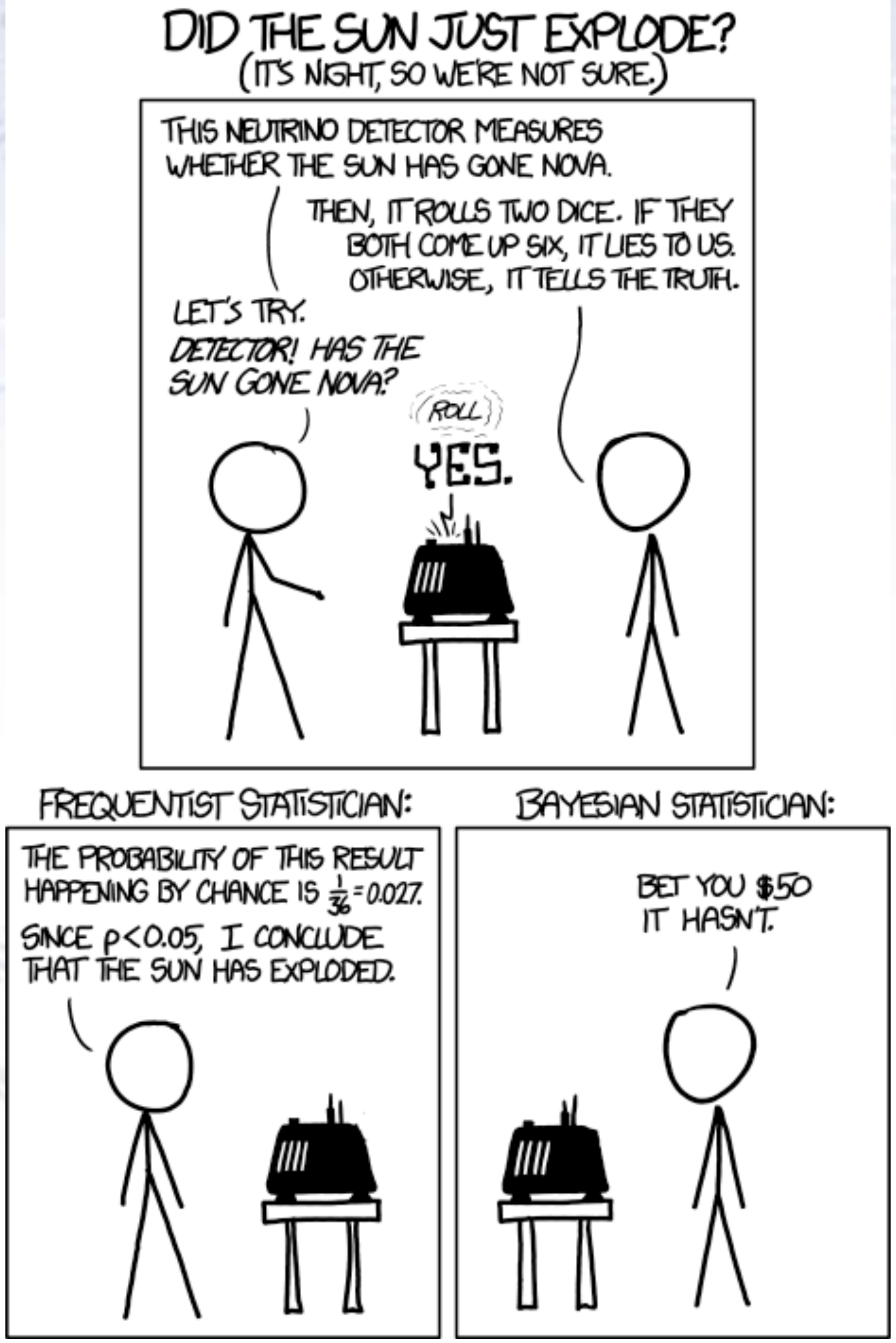
Bayesian statistics

Bayesian statistics is something that often first really appreciated after one has worked on a problem needs bayesian statistics.

In many situations, we have much information that can be used, before using the available data to draw a conclusion.

The criticism of Bayesian statistics is often the choice of a prior, where we as scientists should try to quantify our belief in something.

However often the choice of this prior can really be quantified - or it is something that does not alter the results (too much)…


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$=0.027. SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Bayesian statistics

You know by now that the Bayes theorem takes the form:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

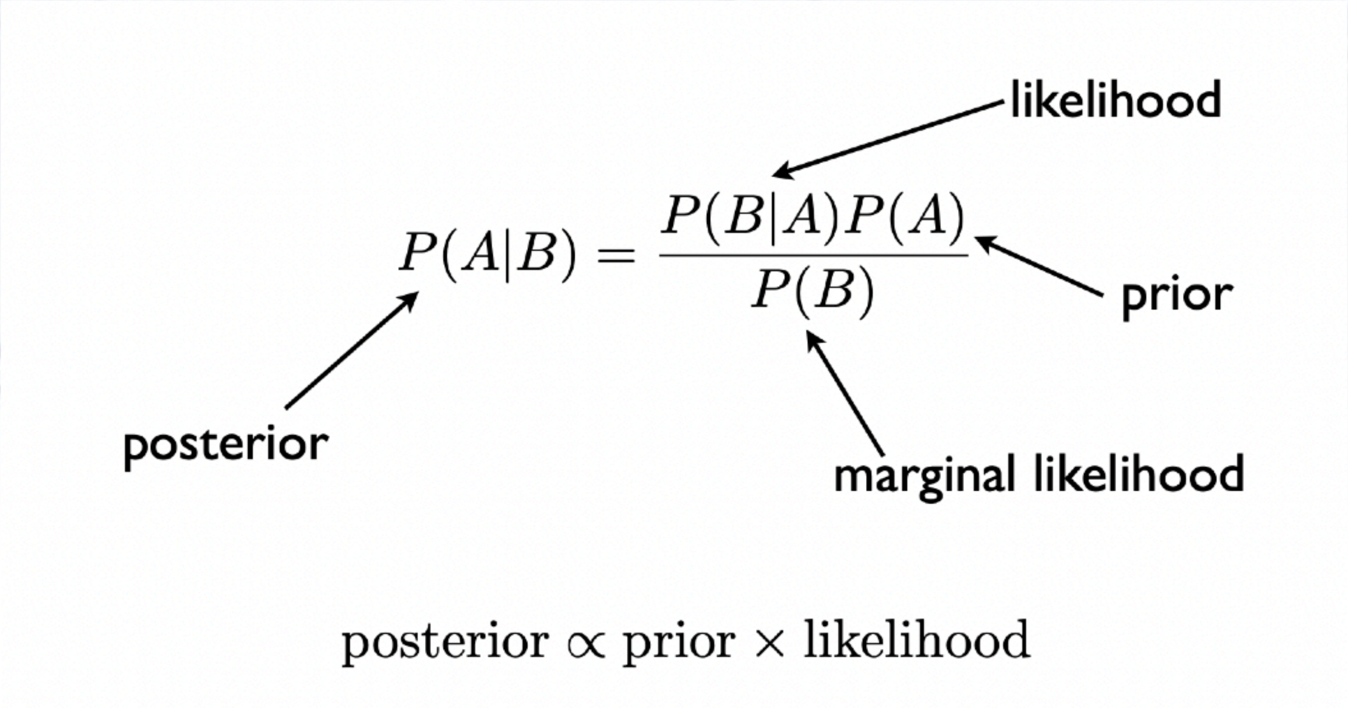
To write this out, there is a discrete version and a continuous version:

$$P(A|B) = \boxed{\frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}} \quad \boxed{\frac{P(B|A)P(A)}{\int P(B|A)P(A)dA}}$$

The point is that we need to integrate out the dependency of A in the denominator. That is to say: what is the probability of getting B, given I try all values of A.

Bayesian statistics

Structure of the terms in the bayesian setup:



likelihood

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

prior

posterior

marginal likelihood

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Typically we calculate the likelihood based on our statistical methods. Then the prior typically causes a lot of concern - because how to quantify our knowledge?

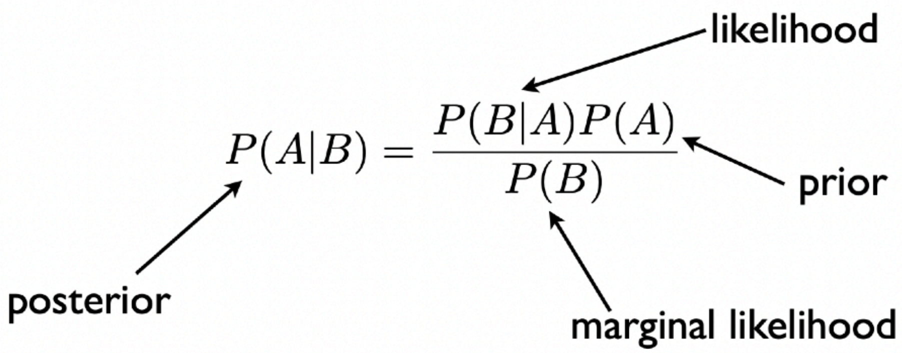
For many population samples the prior is well known and can be used directly. But for many cases, we can start with a flat prior and then update it as we move along.

Updating the prior

One important element in Bayesian statistics is the update of the prior probability. Lets start with a classical example:
We take a test of some disease.
P(positive | disease) = 0.93.
P(negative | healthy) = 0.99.
Also the fraction of people having the disease in the population is 0.148%.

Lets say we get a positive test result.
What is the probability that we have the disease:
Likelihood: 0.93
Prior: 0.00148
Marginal likelihood:
P(positive | disease)*P(disease) + P(positive | healthy)*P(healthy) =
0.93*0.00148 + 0.01*0.9985 = 0.01136

Combining these we have P(disease | positive) = 0.12.
OK - so we only have a 12 percent chance of having the disease…



$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Updating the prior

But lets now say we take a new test - and this is also positive!
Now the test statistics are naturally the same so we have:
P(positive | disease) = 0.93.
P(negative | healthy) = 0.99.
However now the prior is no longer the small 0.00148 but instead our posterior from the previous calculation: p(disease) = 0.12.

This means we can setup the following:
Likelihood: 0.93
Prior: 0.00148
Marginal likelihood:
P(positive | disease)*P(disease) + P(positive | healthy)*P(healthy) =
0.93*0.00148 + 0.01*0.9985 = 0.01136

Combining these we have P(disease | positive) = 0.12.
OK - so we only have a 12 percent chance of having the disease…

Maximum A Posteriori (MAP) Estimation

A concept of specific interest in the framework of Bayesian statistics is the concept of Maximum A Posteriori Estimation.

Note this sounds a lot like Maximum likelihood. Remember in maximum likelihood we use probabilities to obtain the most probable value given all probabilities.

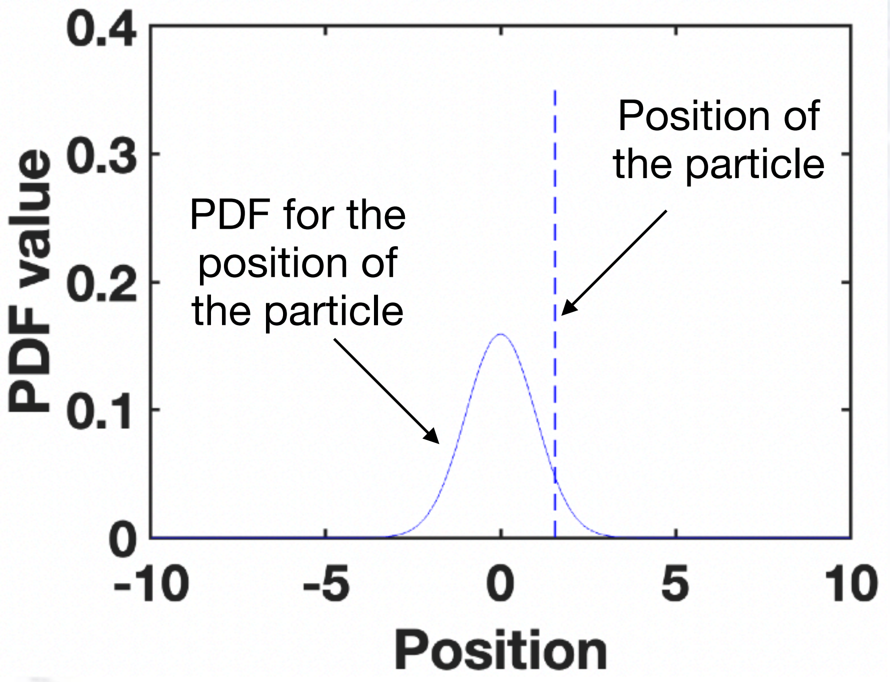
The MAP tries to measure some known quantity, that equals the mode of the posterior distribution. The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data.

$$\begin{aligned}\hat{\theta}_{\mathrm{MAP}}(x) &= \underset{\theta}{\arg\max}\ f(\theta \mid x)\\ &= \underset{\theta}{\arg\max}\ \frac{f(x \mid \theta)\, g(\theta)}{\displaystyle\int_{\Theta} f(x \mid \vartheta)\, g(\vartheta)\, d\vartheta}\\ &= \underset{\theta}{\arg\max}\ f(x \mid \theta)\, g(\theta).\end{aligned}$$

Maximum A Posteriori (MAP) Estimation

OK - lets look at an example.
Suppose I measure a diffusing particle. It diffuses like brownian motion and it takes a gaussianly distributed step:

$$p_X(x) = \mathcal{N}(0, \sigma_x) = \frac{1}{2\pi\sigma_x} e^{-\frac{1}{2}\left(\frac{x}{\sigma_x}\right)^2}$$



So after this step the particle has a true position X.
However there is noise in our measurements. This means that we measure a parameter Y = X+W.

Maximum A Posteriori (MAP) Estimation

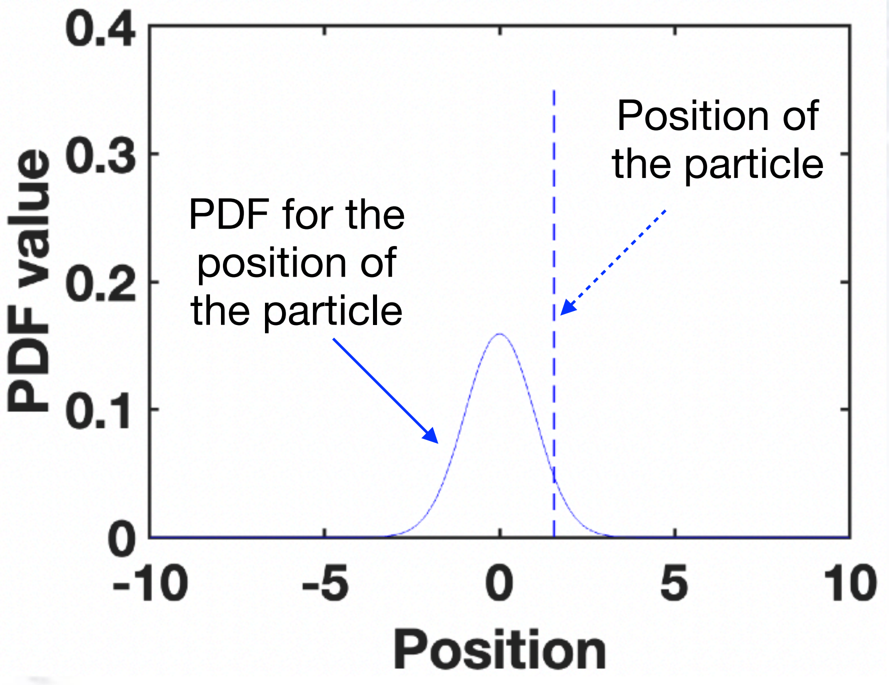
OK - lets look at an example.
Suppose I measure a diffusing particle. It diffuses like brownian motion and it takes a gaussianly distributed step:

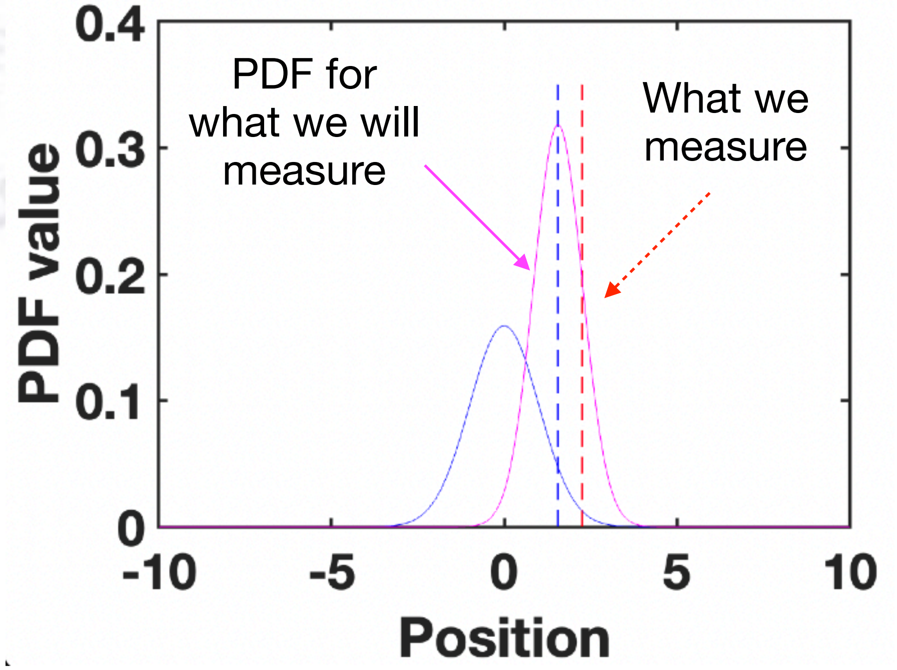
$$\boxed{p_X(x) = \mathcal{N}(0, \sigma_x) = \frac{1}{2\pi\sigma_x} e^{-\frac{1}{2}\left(\frac{x}{\sigma_x}\right)^2}}$$

So after this step the particle has a true position X. However there is noise in our measurements. This means that we measure a parameter Y = X+W.

However as is typically the case, the noise is also gaussian so:

$$\boxed{p_W(w) = \mathcal{N}(0, \sigma_w)}$$





Maximum A Posteriori (MAP) Estimation

So let’s say we measure the value Y = 2.25 m. What is our best estimate for the position of the particle?

Maximum A Posteriori (MAP) Estimation

Lets see what we happened if we directly attacked the problem using Maximal likelihood. This would simply be:

$$p_{Y|X}(y|x) = \frac{1}{2\pi\sigma_w} e^{-\frac{1}{2}\left(\frac{y-x}{\sigma_w}\right)^2}$$

From this is it clear that the most likely value is X = Y.

But what happens if we use the bayesian Maximum A Posteriori Estimation?

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} = \mathcal{C}\frac{1}{2\pi\sigma_w\sigma_x} e^{-\frac{1}{2}\left[\left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y-x}{\sigma_w}\right)^2\right]}$$

If I want to find the most probable value of the x-value, I should find the minimum of this function. This means I should minimise the function:

$$f = \frac{(y-x)^2}{2\sigma_w^2} + \frac{x^2}{2\sigma_x^2}$$

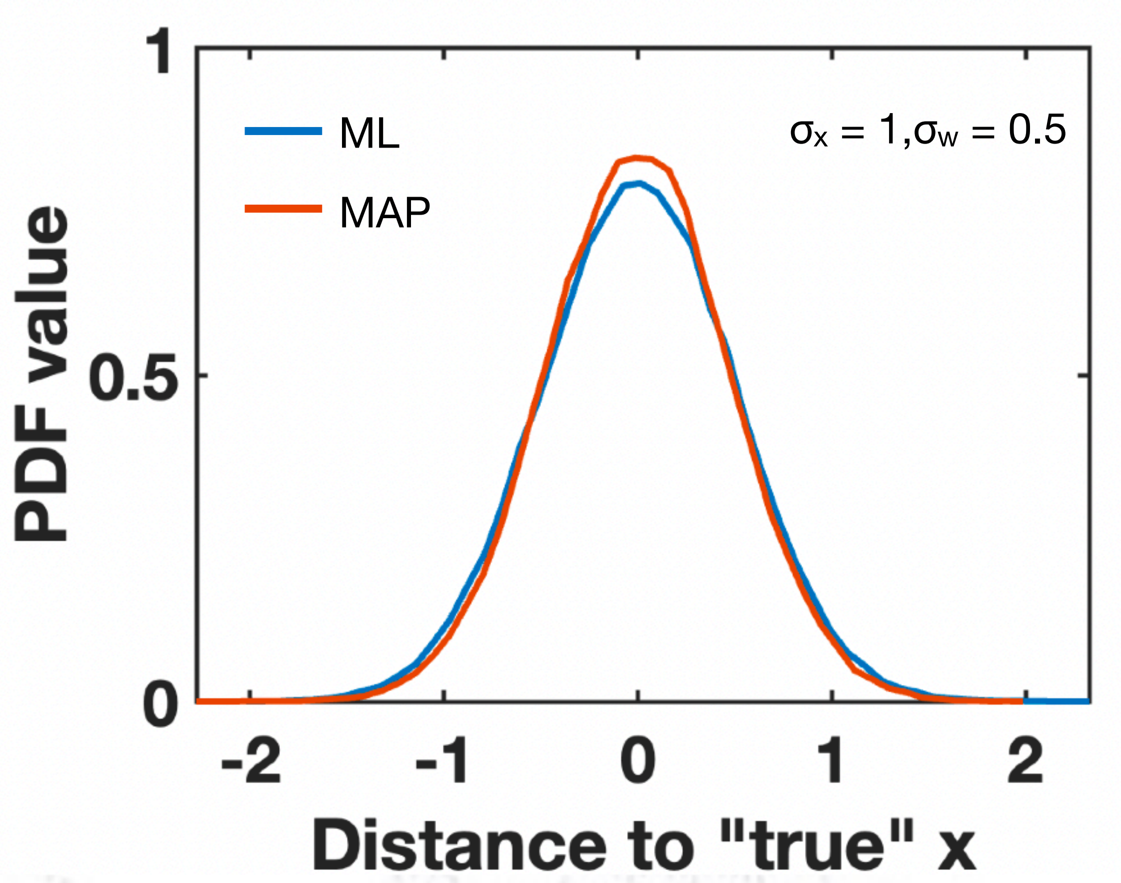
Simply differentiating this and setting to zero gives:

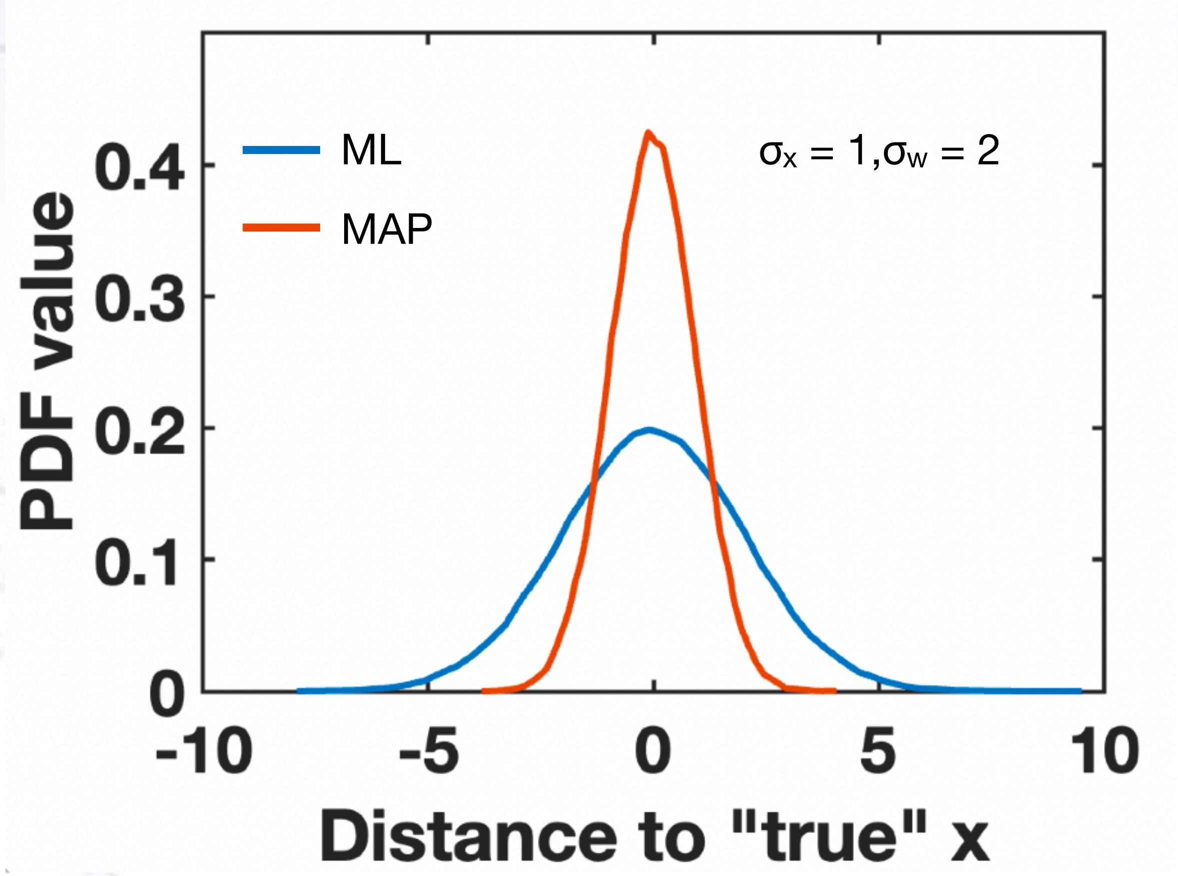
$$\frac{\partial f}{\partial x} = 0$$

$$\Rightarrow \hat{x} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2} y$$

Maximum A Posteriori (MAP) Estimation

This shows that the MAP gives a different result than the ML method. Does this matter?



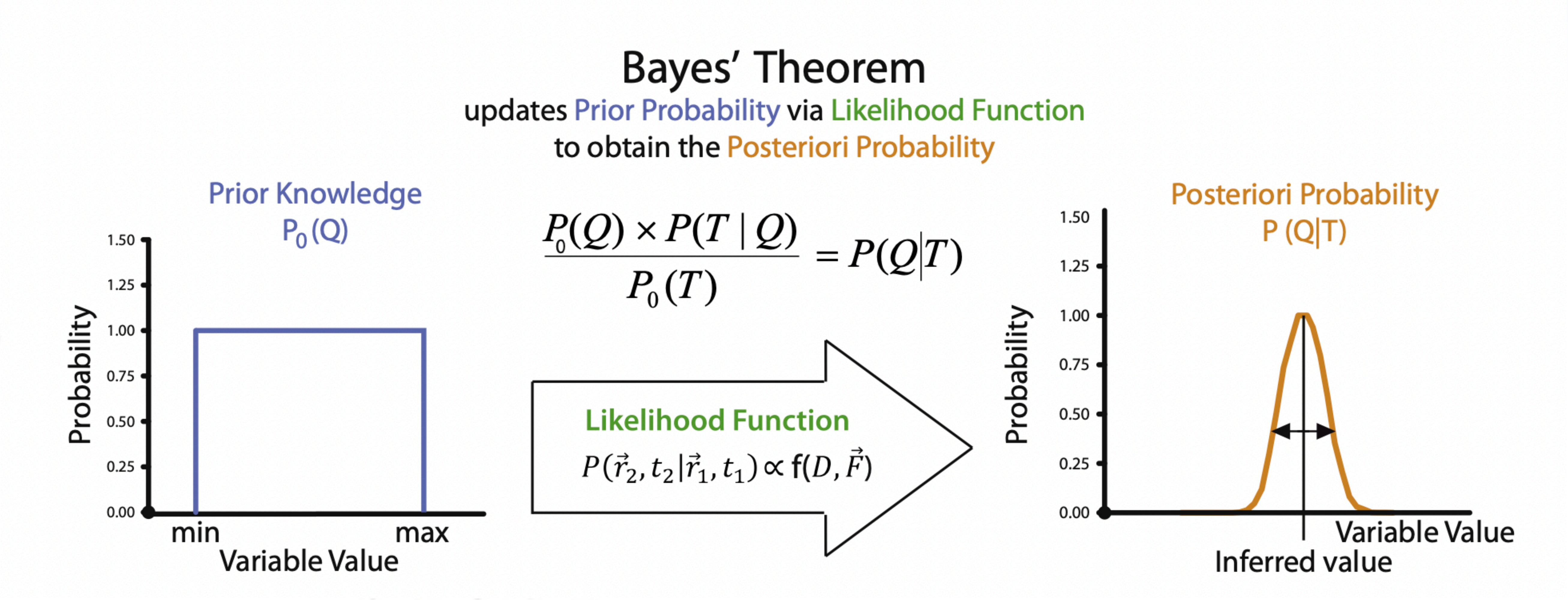


Conclusion: If we measure a point and we know the measurement error, the best estimate is not just point itself.

Bayesian Inference

The Bayesian approach is heavily used in parameter estimation called Bayesian inference. This is an enormous field, that we just want to touch upon in this lecture.

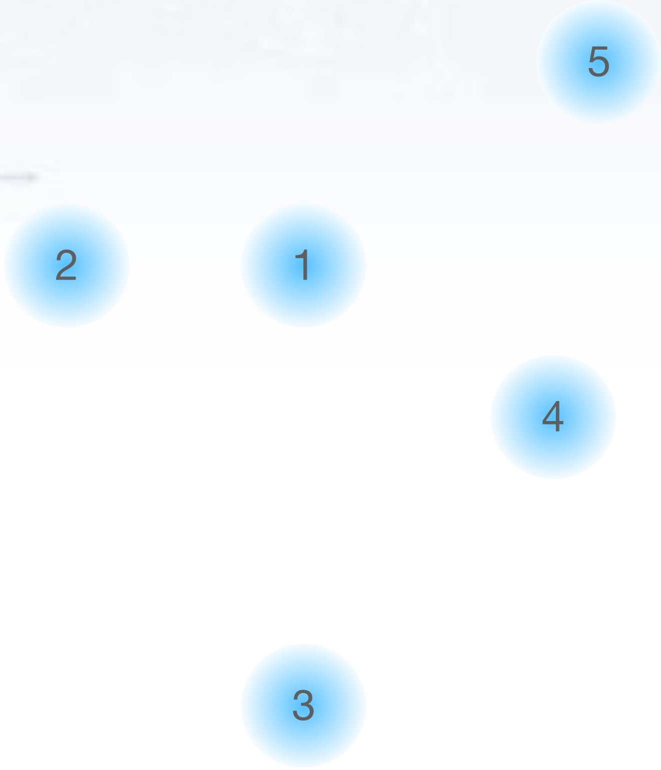
The idea is to start with some prior knowledge of the parameters, and then use the likelihood based on a series of events to extract optimal parameter values.



Bayesian Inference

Suppose we measure a diffusing particle at 5 positions. There is an experimental error to the measurements so we do not even know the precise position.

For this set of datapoints, we want to infer the underlying diffusion coefficient. We do know that there is structure in the data-points. That is 3 comes after 2 that comes after 1 etc.



But our certainty in the position of 3 is definitely affected by our certainty in the position of 2 and so on.

This means that we can calculate the probability of each position based on the previous position - and update these probabilities accordingly.

In the end we can use Bayes theorem to find infer the optimal parameters:

$$p(D,\sigma|r_1,...,r_5) = \frac{\prod p(r_i|r_{i-1}|D,\sigma)p(D,\sigma)}{p_0(r_1,...,r_5)}$$

Entering: Markov chains

It is now hopefully clear, that Bayesian statistics has a strong ability to estimate the structure of sequential data.

However we need a mathematical framework that can connect probabilities as we take new steps: entering Markov chains.

Andrey Markov was a Russian mathematician that developed the concept in the beginning of the 20th century and has further been studied by many mathematicians, but most notably our hero Kolmogorov.

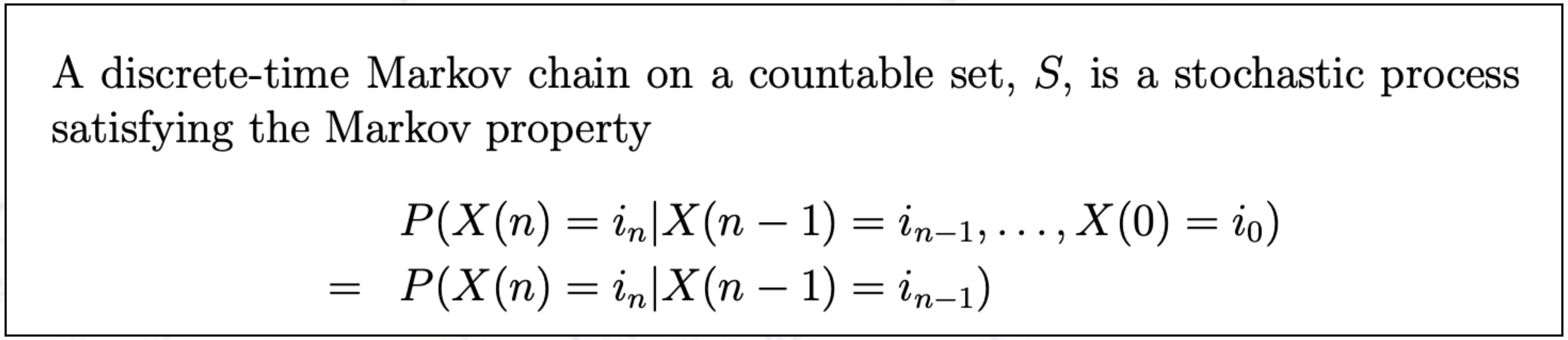
А. А. Марков (1886).

Markov chains

A very useful mathematical approach to statistics of a series of events, is the construction of Markov Chains.

Disclaimer: The use of Markov Chains is combination with Bayesian statistics and Monte Carlo methods is covered in depth in the course “Advanced methods in Applied Statistics” - so here we will have a small taste of it.

Mathematically we define Markov chains:



A discrete-time Markov chain on a countable set, S, is a stochastic process satisfying the Markov property

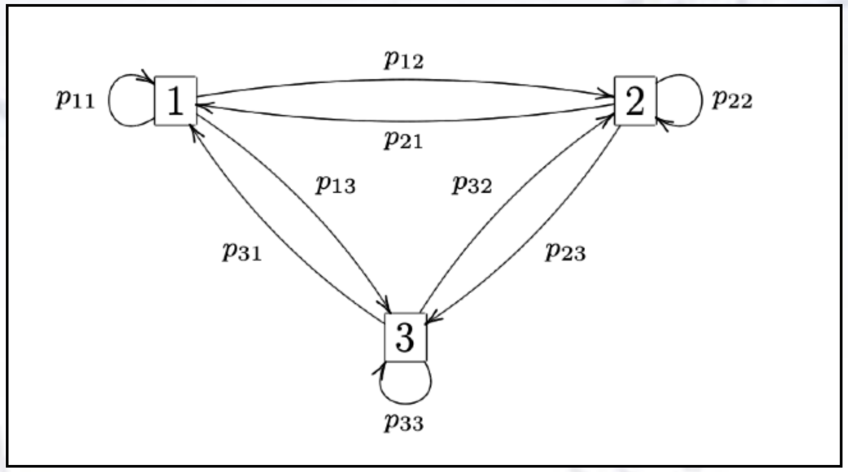
$$\begin{aligned} & P(X(n) = i_n | X(n-1) = i_{n-1}, \ldots, X(0) = i_0) \\ = \; & P(X(n) = i_n | X(n-1) = i_{n-1}) \end{aligned}$$

Translated this means that the probability to move to a specific state is completely determined by the state we are currently in - and not where we have been previously. It is therefore *memoryless.*

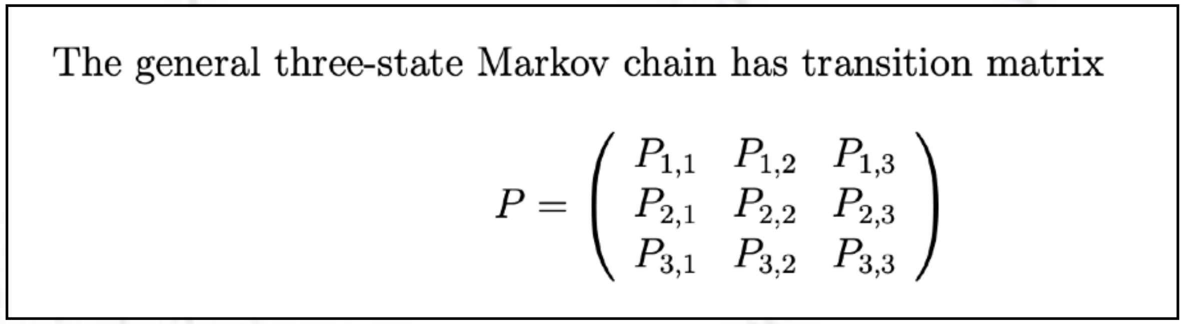
Discrete Markov chains

We can visualise this to make it much more easy to understand.

Lets assume I had a system for which I could quantify 3 states and assign probabilities to move between states. This could be visualised in the following way:



Based on these 9 probabilities, I can construct a matrix, taking the form:



The general three-state Markov chain has transition matrix

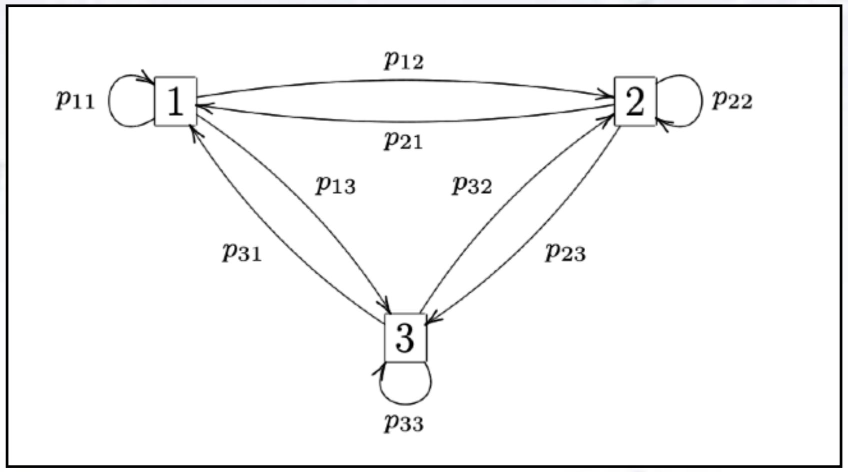
$$P = \begin{pmatrix} P_{1,1} & P_{1,2} & P_{1,3} \\ P_{2,1} & P_{2,2} & P_{2,3} \\ P_{3,1} & P_{3,2} & P_{3,3} \end{pmatrix}$$

Translated this means that the probability to move to a specific state is completely determined by the state we are currently in - and not where we have been previously. It is therefore *memoryless.*

Discrete Markov chains

So far so good - how can this be used?

Assume I as the question: we start in state 1. What is the probability to be in state 2 after 3 iterations?



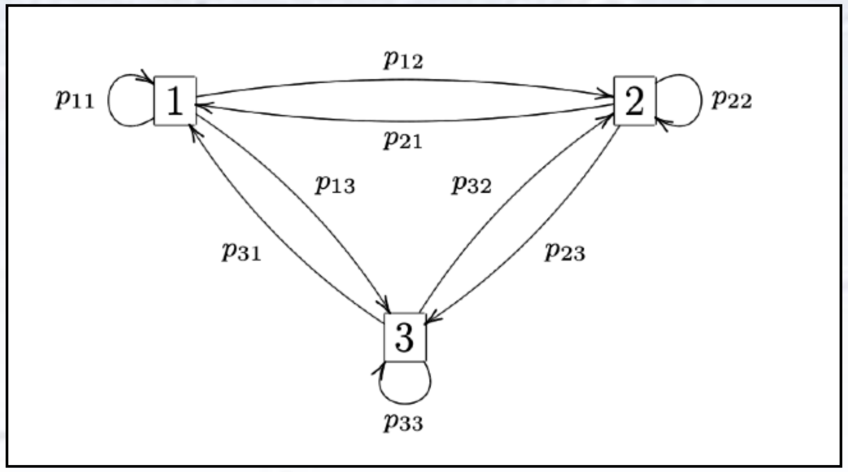
Try to count the number of ways we can end up in state 2. It could take the form:

1 -> 2 -> 2 -> 2
1 -> 1 -> 2 -> 2
1 -> 1 -> 1 -> 2
1 -> 2 -> 1 -> 2
1 -> 2 -> 3 -> 2
1 -> 3 -> 2 -> 2

Luckily there is a much nicer calculation that makes sure we do not have to count all possibilities every time. Imagine if it was after 20 iterations….

Discrete Markov chains

It turns out that the way we add the probabilities is exactly given by the structure of the transition matrix P.



$$P = \begin{pmatrix} P_{1,1} & P_{1,2} & P_{1,3} \\ P_{2,1} & P_{2,2} & P_{2,3} \\ P_{3,1} & P_{3,2} & P_{3,3} \end{pmatrix}$$

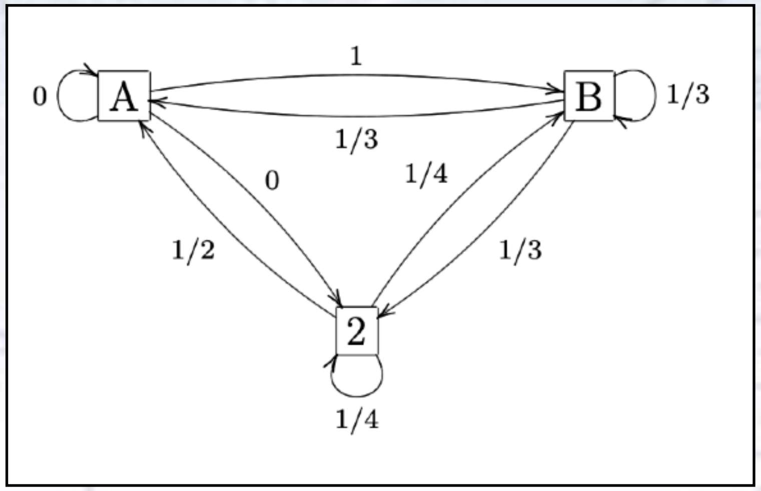
The probability to be in either of N states (here we have 3), after n iterations is given by the equation:

$$(P(X(n) = 1), \ldots, P(X(n) = N)) = \phi P^n.$$

Note that here ϕ is the row vector of initial probabilities. If we know it starts in state 1 in the above example it will take the form ϕ=(1,0,0).

Discrete Markov chains

So lets just see this in action. I now assign probabilities to the general 3-state matrix:



What is the probability to in the states after 4 iterations - given I start in state 2?

Our matrix takes the form: $P = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$

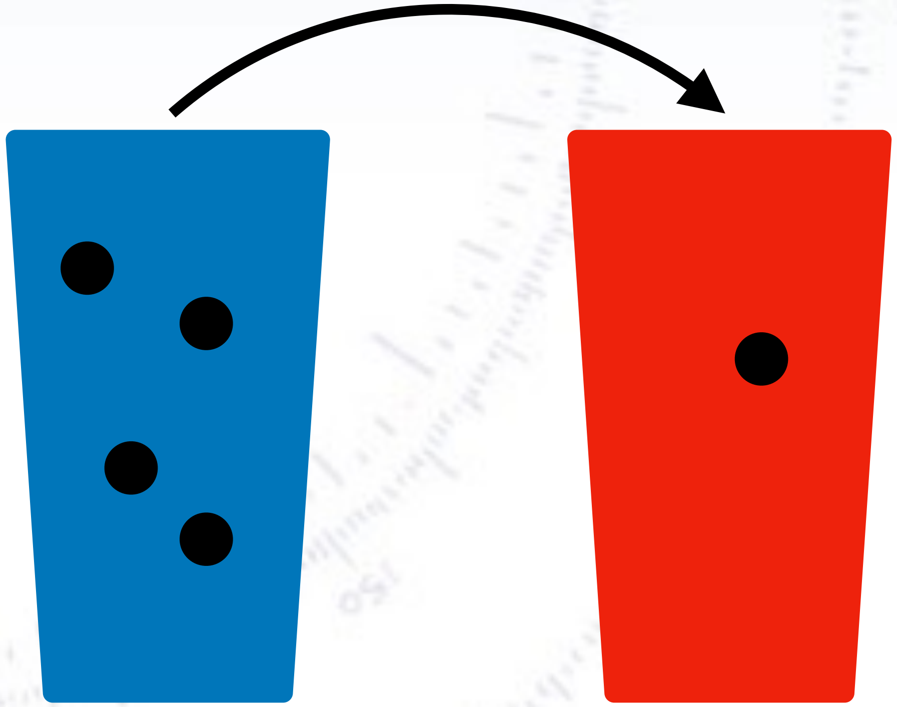
And making the calculations: $(0,1,0)*P^4 = (0.28 , 0.5 , 0.22)$

This is the vector of all probabilities after 4 iterations.

Discrete Markov chains

It happens quite often that some statistical problem can be formulated by a Markov Chain. We look at one example that is also covered in the excersizes - the Ehrenfest urn problem.

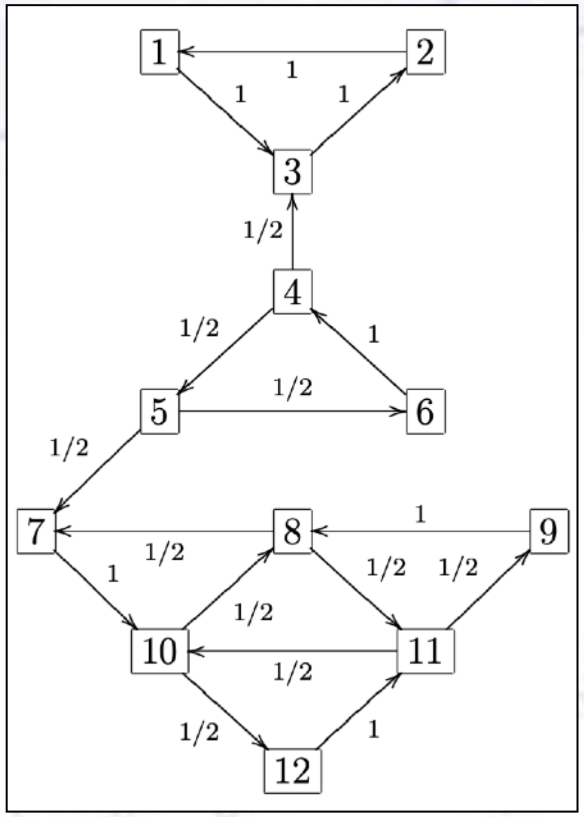
Suppose we have two urns (containers) and N balls (say 5). Now at each time step we pick a random ball and move it from one urn to the other.



What is the probability to have 3 balls in the blue container after 10 iterations? This can be formulated as a Markov chain. Can you see why?

Irreducibility and communication classes

Based on the structure of Markov chains, there can be separate communications classes. Suppose you start in state 4.

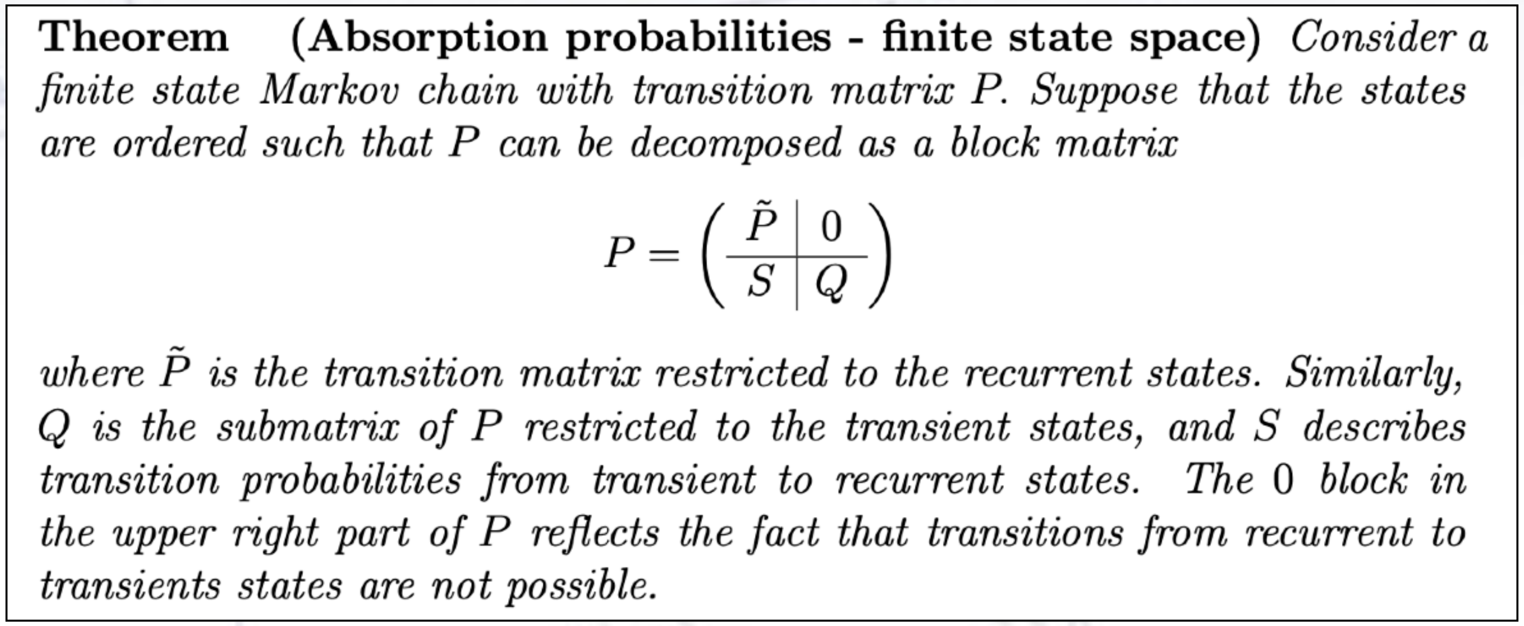


Once you reach state 3 or state 7 you can never return. And you can never go between state 3 and state 7. Therefore we say that 1-3 is a recurrent communication class, 4-6 is a transient class, and 7-12 is also a recurrent class.

If there is only one communication class the Chain is called irreducible.

Absorption probabilities

We can for many purposes say that state 3 and state 7 are absorbing states. Once you reach either of these there is no coming back. When we have absorbing states we can organise the matrix based on recurrent and transient states:



Theorem (Absorption probabilities - finite state space) *Consider a finite state Markov chain with transition matrix P. Suppose that the states are ordered such that P can be decomposed as a block matrix*

$$P = \left(\begin{array}{c|c} \tilde{P} & 0 \\ \hline S & Q \end{array} \right)$$

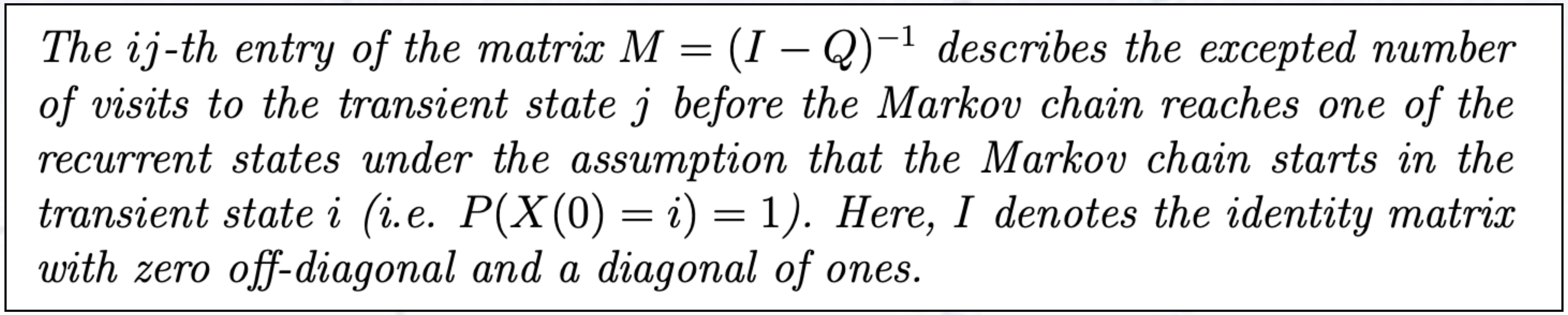
where $\tilde{P}$ is the transition matrix restricted to the recurrent states. Similarly, Q is the submatrix of P restricted to the transient states, and S describes transition probabilities from transient to recurrent states. The 0 block in the upper right part of P reflects the fact that transitions from recurrent to transients states are not possible.

This comes in handy, because often our question would be: Given we start in state 4, what is the expected number of iterations before absorption in to either state?

Or is the probability to be absorbed by 7 instead of 3, given I start in state 4?

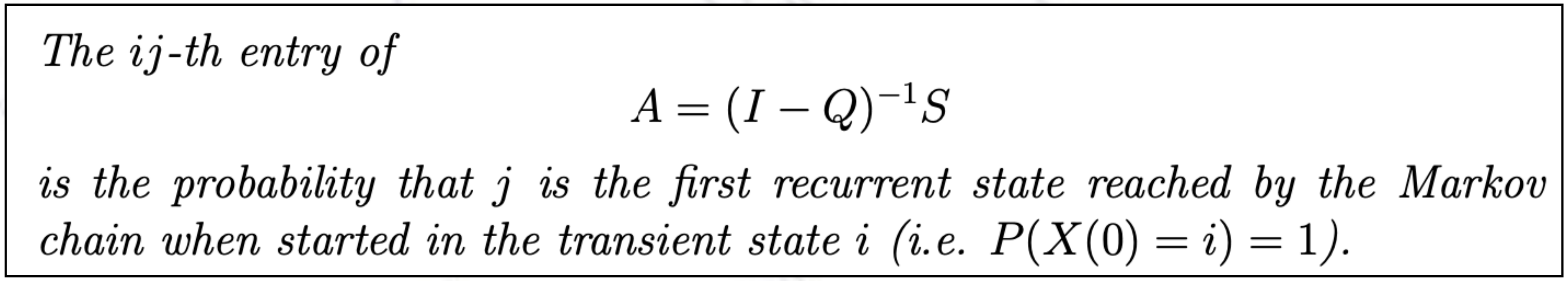
Absorption probabilities

Exactly this we can calculate by manipulation by the organisation mentioned above. We can calculate the expected number of visits in a specific transient state, by constructing the matrix M:



The ij-th entry of the matrix $M = (I - Q)^{-1}$ *describes the excepted number of visits to the transient state* j *before the Markov chain reaches one of the recurrent states under the assumption that the Markov chain starts in the transient state* i *(i.e.* $P(X(0) = i) = 1$*). Here,* I *denotes the identity matrix with zero off-diagonal and a diagonal of ones.*

Also, with this we can calculate the probability that a specific absorbing state will be the first we reach, by constructing the matrix:



The ij-th entry of

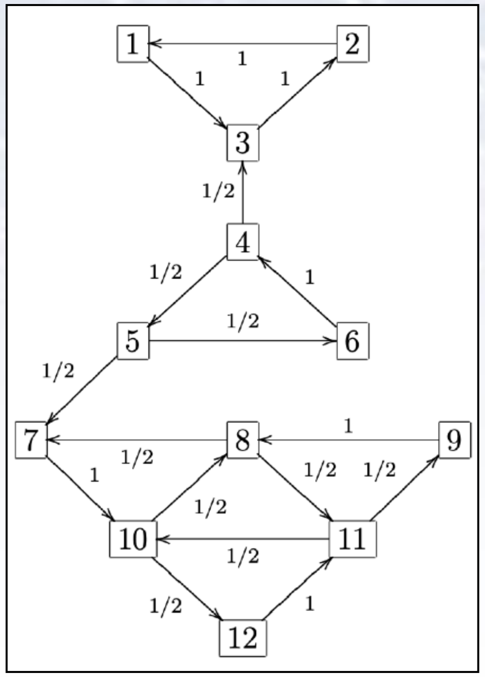
$$A = (I - Q)^{-1}S$$

is the probability that j *is the first recurrent state reached by the Markov chain when started in the transient state* i *(i.e.* $P(X(0) = i) = 1$*).*

Absorption probabilities

So far so good - let’s see it in action.

$$P = \left(\begin{array}{c|c} \tilde{P} & 0 \\ \hline S & Q \end{array}\right) \longrightarrow \quad P = \left(\begin{array}{cc|ccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & 0 \end{array}\right)$$



I have restructured the matrix, and made the two states 3 and 7 absorbing - so I don’t care what is going on in state 12.

Now the expected number of visits to state 5 before absorption, given we start in state 4, is now:

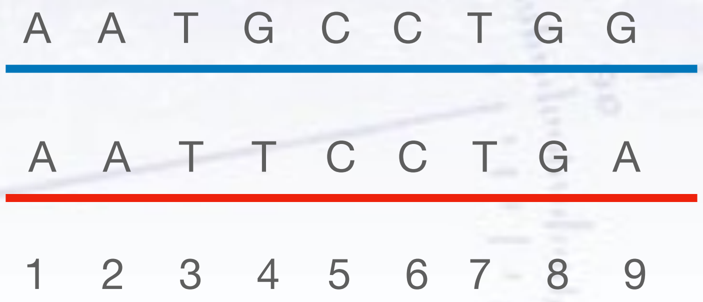
$$M_5 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} * \left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix} \right)^{-1} * \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = 2/3$$

And the probability that 7 is the first absorbing state we reach is:

$$A_7 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} * \left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix} \right)^{-1} * \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \\ 0 & 0 \end{pmatrix} * \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 1/3$$

Example from DNA estimation

Assume we are measuring DNA from ancient samples. At all places (alleles), we have two bases - one from the father and one from the mother. It could for instance look like this:



A A T G C C T G G

A A T T C C T G A

1 2 3 4 5 6 7 8 9

Now at position 1, both from the father and the mother we have an A. Lets say we have extracted some DNA for position 38 in an egyptic prince. We don't have much useful DNA so we get the sequence out:

[A,A,C,A,A]

Given that transitions can occur, and thus even though we measure a C, it could be an A, that has just changed due to errors in equipment and mutations in the DNA.

What is the probability for each of the possible genotypes?

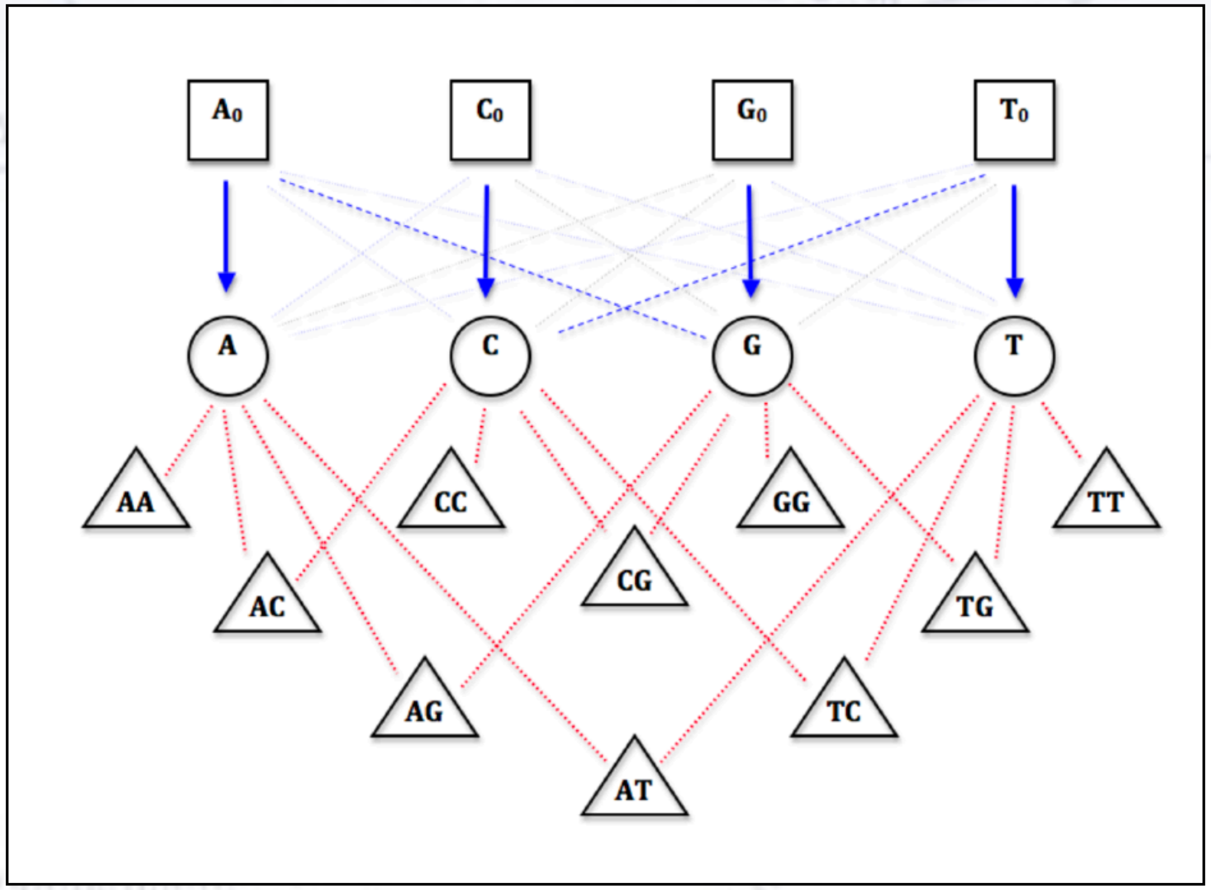
Example from DNA estimation

So here we measure some base, and we want to determine the probability of the true base and in the end that it comes from a specific genotype. To visualise this, we can set up a network:

What we measure

What it is

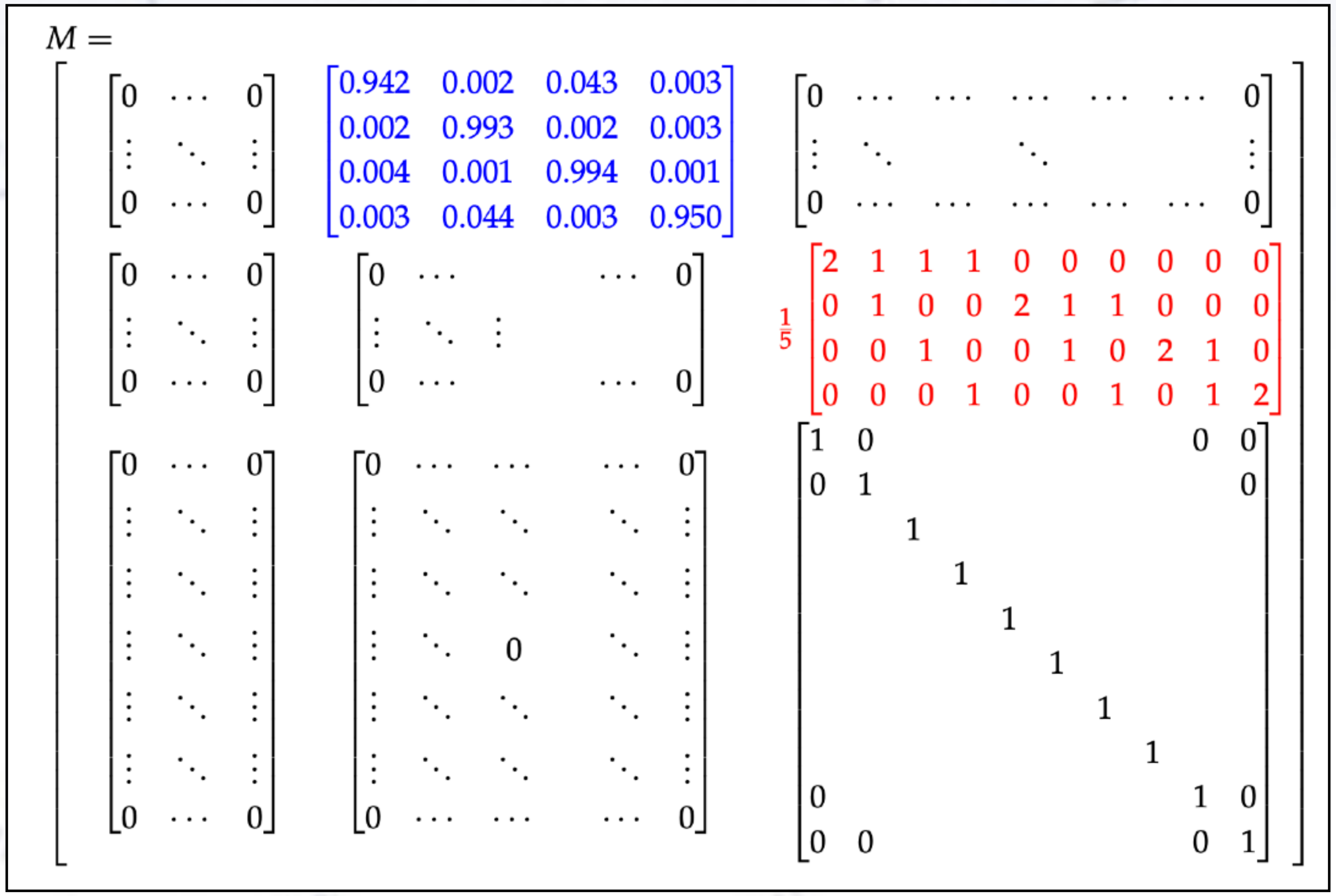
Where it comes from



But with this we could imagine it is a Markov chain where all genotypes in the bottom are absorbing states.

Example from DNA estimation

If I write up a total matrix it will be super large and ugly. But since I only need the transient state probabilities, I can take out the coloured parts and use these for matrix multiplication.



$$M = \begin{bmatrix}
\begin{bmatrix}0&\cdots&0\\ \vdots&\ddots&\vdots\\ 0&\cdots&0\end{bmatrix} &
\color{blue}{\begin{bmatrix}0.942&0.002&0.043&0.003\\ 0.002&0.993&0.002&0.003\\ 0.004&0.001&0.994&0.001\\ 0.003&0.044&0.003&0.950\end{bmatrix}} &
\begin{bmatrix}0&\cdots&\cdots&\cdots&\cdots&\cdots&0\\ \vdots&\ddots&&\ddots&&&\vdots\\ 0&\cdots&\cdots&\cdots&\cdots&\cdots&0\end{bmatrix} \\
\begin{bmatrix}0&\cdots&0\\ \vdots&\ddots&\vdots\\ 0&\cdots&0\end{bmatrix} &
\begin{bmatrix}0&\cdots&&\cdots&0\\ \vdots&\ddots&\vdots&&\\ 0&\cdots&&\cdots&0\end{bmatrix} &
\color{red}{\frac{1}{5}\begin{bmatrix}2&1&1&1&0&0&0&0&0&0\\ 0&1&0&0&2&1&1&0&0&0\\ 0&0&1&0&0&1&0&2&1&0\\ 0&0&0&1&0&0&1&0&1&2\end{bmatrix}} \\
\begin{bmatrix}0&\cdots&0\\ \vdots&\ddots&\vdots\\ \vdots&\ddots&\vdots\\ \vdots&\ddots&\vdots\\ \vdots&\ddots&\vdots\\ \vdots&\ddots&\vdots\\ 0&\cdots&0\end{bmatrix} &
\begin{bmatrix}0&\cdots&\cdots&\cdots&0\\ \vdots&\ddots&\ddots&\ddots&\vdots\\ \vdots&\ddots&\ddots&\ddots&\vdots\\ \vdots&\ddots&0&\ddots&\vdots\\ \vdots&\ddots&\ddots&\ddots&\vdots\\ \vdots&\ddots&\ddots&\ddots&\vdots\\ 0&\cdots&\cdots&\cdots&0\end{bmatrix} &
\begin{bmatrix}1&0&&&&&&&0&0\\ 0&1&&&&&&&&0\\ &&1&&&&&&&\\ &&&1&&&&&&\\ &&&&1&&&&&\\ &&&&&1&&&&\\ &&&&&&1&&&\\ &&&&&&&1&&\\ 0&&&&&&&&1&0\\ 0&0&&&&&&&0&1\end{bmatrix}
\end{bmatrix}$$

Example from DNA estimation

Now I get the information that some some genotypes are much more abundant than others. Should I use this in my analysis?

This is an example of a Bayesian prior that is easily quantifiable. If I do not include this, I directly ignore a lot of information and therefore I make the analysis more ignorant that is should be.

Therefore I can include a Bayesian prior on all genotype probabilities. Does this affect the results? Yes indeed.

But it quantifies how much the data should be overwhelming, before trusting a genotype that is otherwise rarely found.