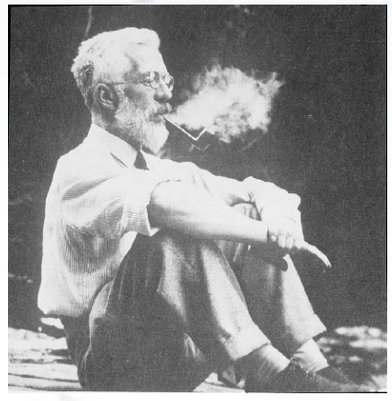
Applied Statistics

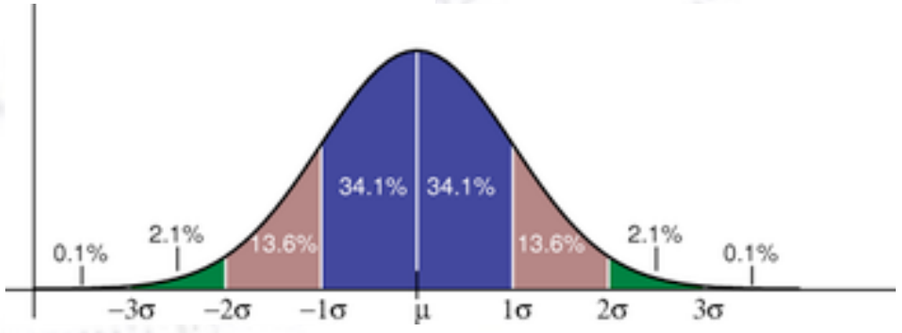
Bayesian statics and Markov Chains








Mathias S. Heltberg (NBI)



"Statistics is merely a quantisation of common sense"

Learning Objectives

<u>Fundamental insight into Bayesian statistics</u>

- What is a prior?
- How can you update the prior?

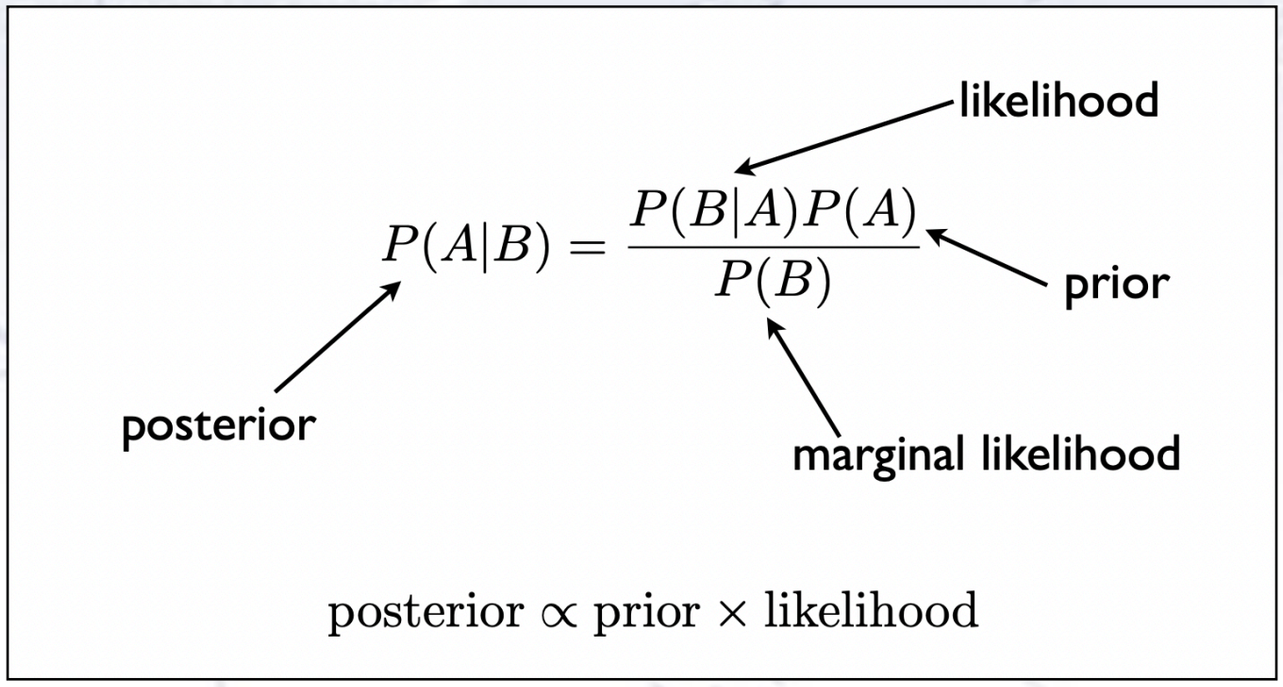
<u>Markov Chains</u>

- Calculating probabilities with Markov Chains
- Importance of detailed balance

How Markov Chains and Bayesian statistics work together in parameter estimations as Metropolis-Hastings

Bayesian statistics

Structure of the terms in the bayesian setup:



We calculate the likelihood based on our statistical methods.

The prior can cause concern - how to quantify our knowledge?

For population samples the prior is well known and can be used directly.

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.

Updating the prior

A classical example:

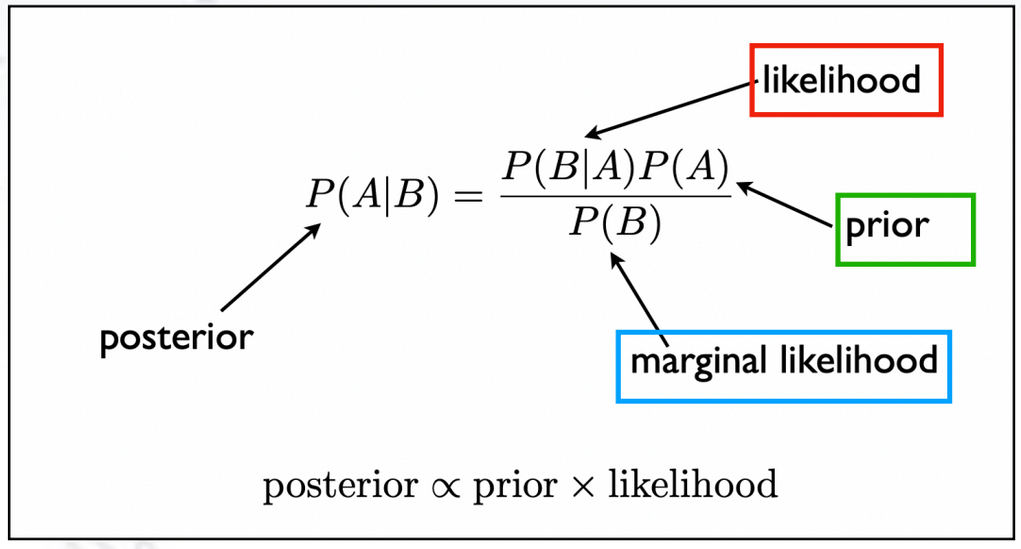
We take a test of some disease.

P(positive | disease) = 0.93.

P(negative | healthy) = 0.99.

Also the fraction of people having the disease in the population is:

p(disease) = 0.148%.



Result of test: Positive

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*(1-0.00148) = 0.01136

Resulting probability: P(disease | positive) = 0.12.

OK - so we only have a 12 percent chance of having the disease…

Updating the prior

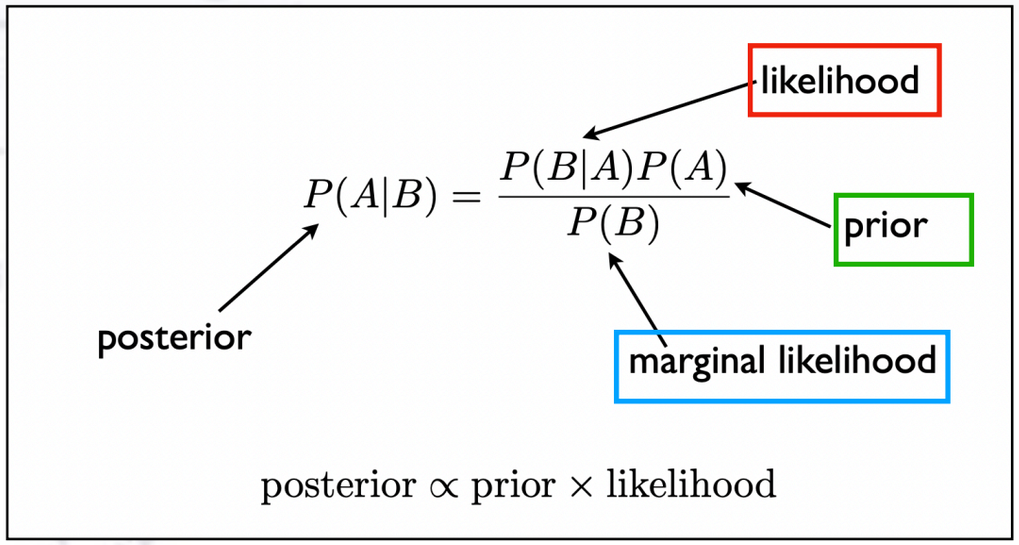
We take a new test. Result: positive!

Now the test statistics are naturally the same so we have:

P(positive | disease) = 0.93.

P(negative | healthy) = 0.99.

However *the prior is no longer the small 0.00148* but instead our posterior from the previous calculation: p(disease) = 0.12.



Calculations:

This means we can setup the following:

Likelihood: 0.93

Prior: 0.12

Marginal likelihood:

P(positive | disease)*P(disease) + P(positive | healthy)*P(healthy)

= 0.93*0.12 + 0.01*(1-0.12) = 0.1161

Resulting probability: P(disease | positive) = 0.9238.

OK - so now we have a 92 percent chance of having the disease because the prior is updated!

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 obtain the most probable value of some parameter given all probabilities.

The MAP estimates a parameter, that equals maximum posterior distribution. The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data.

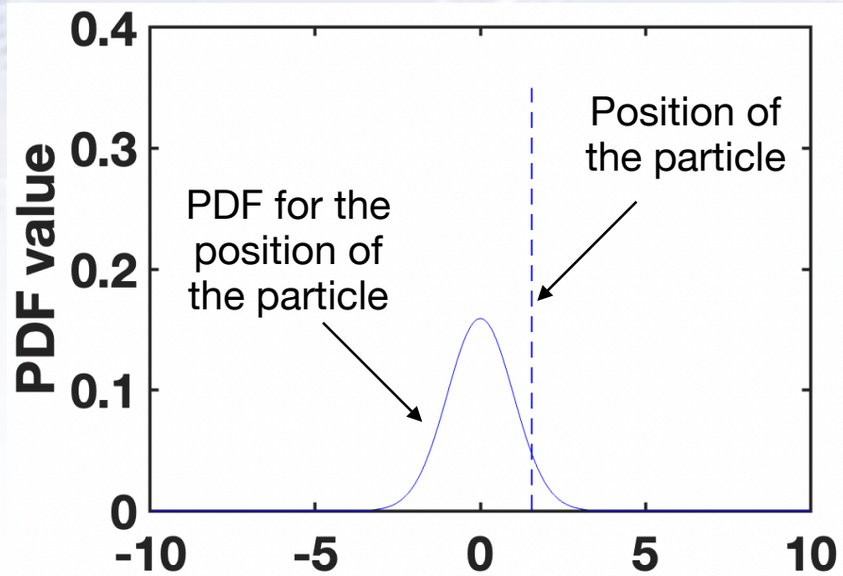
$$\begin{aligned}\hat{\theta}_{\mathrm{MAP}}(x) &= \arg\max_{\theta} f(\theta \mid x) \\ &= \arg\max_{\theta} \frac{f(x \mid \theta)\, g(\theta)}{\int_{\Theta} f(x \mid \vartheta)\, g(\vartheta)\, d\vartheta} \\ &= \arg\max_{\theta} f(x \mid \theta)\, g(\theta).\end{aligned}$$

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

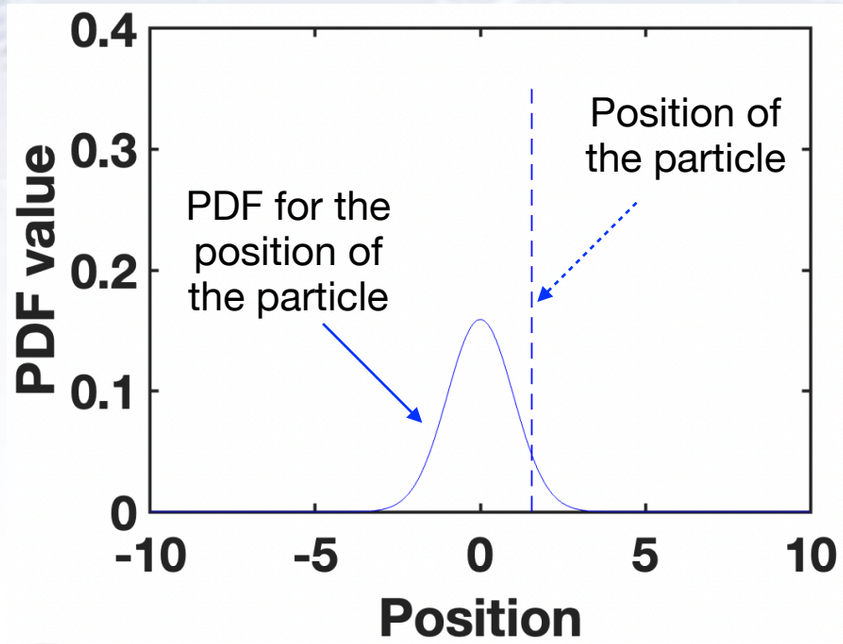
OK - lets look at an example.
Suppose I measure a diffusing particle. It diffuses like brownian motion and it takes a gaussianly distributed step:

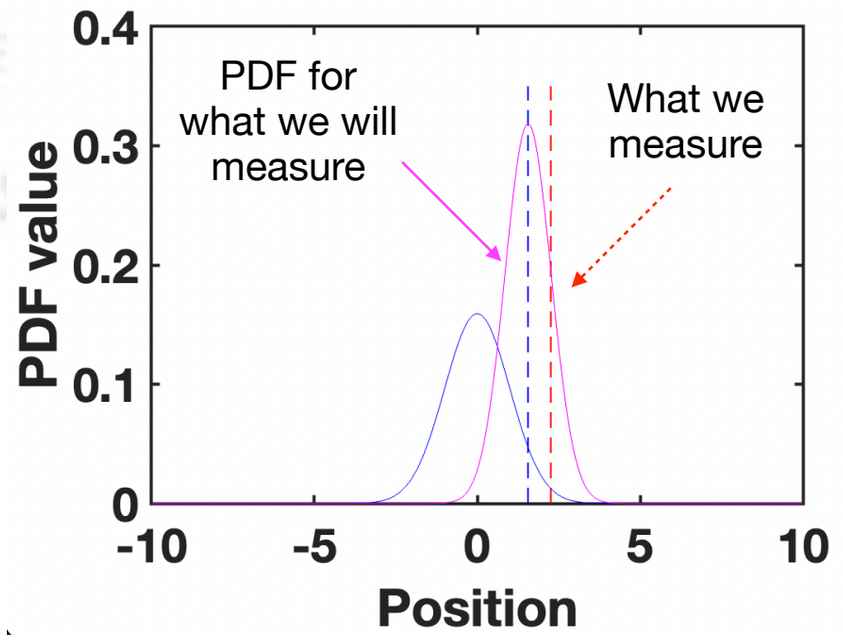
$$p_X(x) = \mathcal{N}(0, \sigma_x) = \frac{1}{2\pi\sigma_x} e^{-\frac{1}{2}\left(\frac{x}{\sigma_x}\right)^2}$$

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:

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





Maximum A Posteriori (MAP) Estimation

So let's say we have the standard deviation for both diffusion and experimental noise to be 1 m.

We now measure the value Y = 2 m.

What is our best estimate for the position of the particle?

Maximum A Posteriori (MAP) Estimation

Maximum Likelihood gives:

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

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 the exponent. This means I should minimise the function:

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

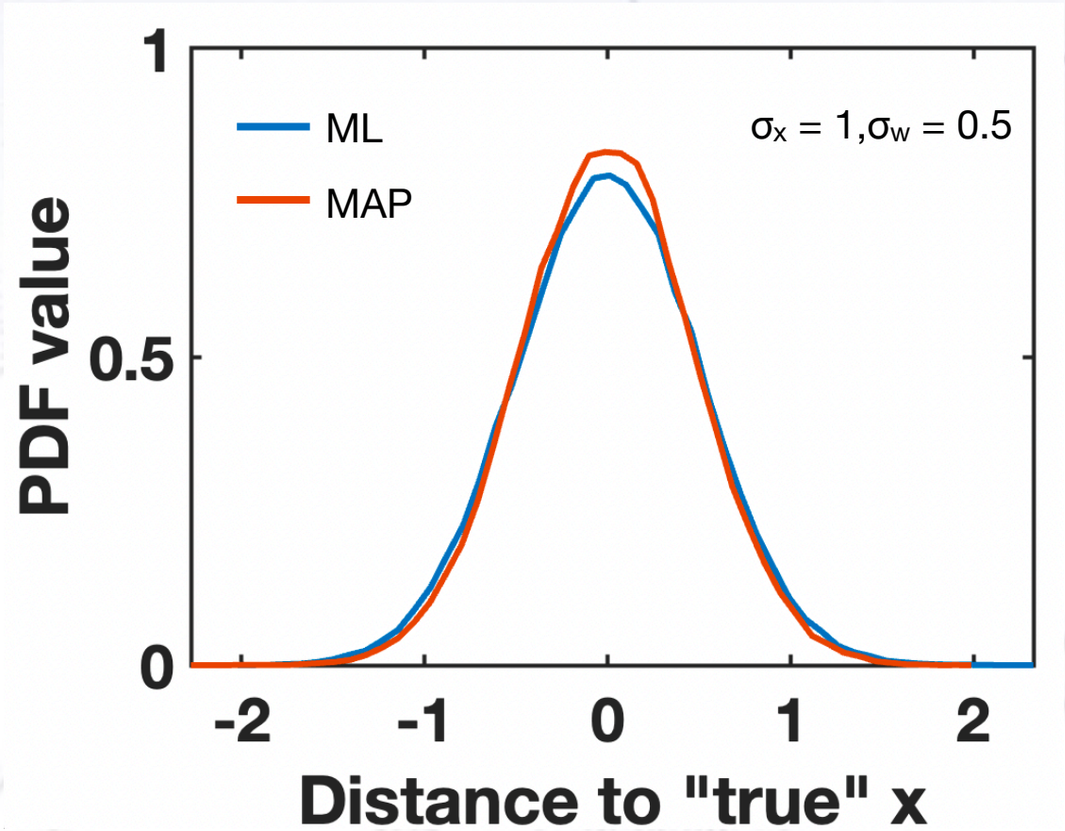
Simply differentiating this and setting to zero gives:

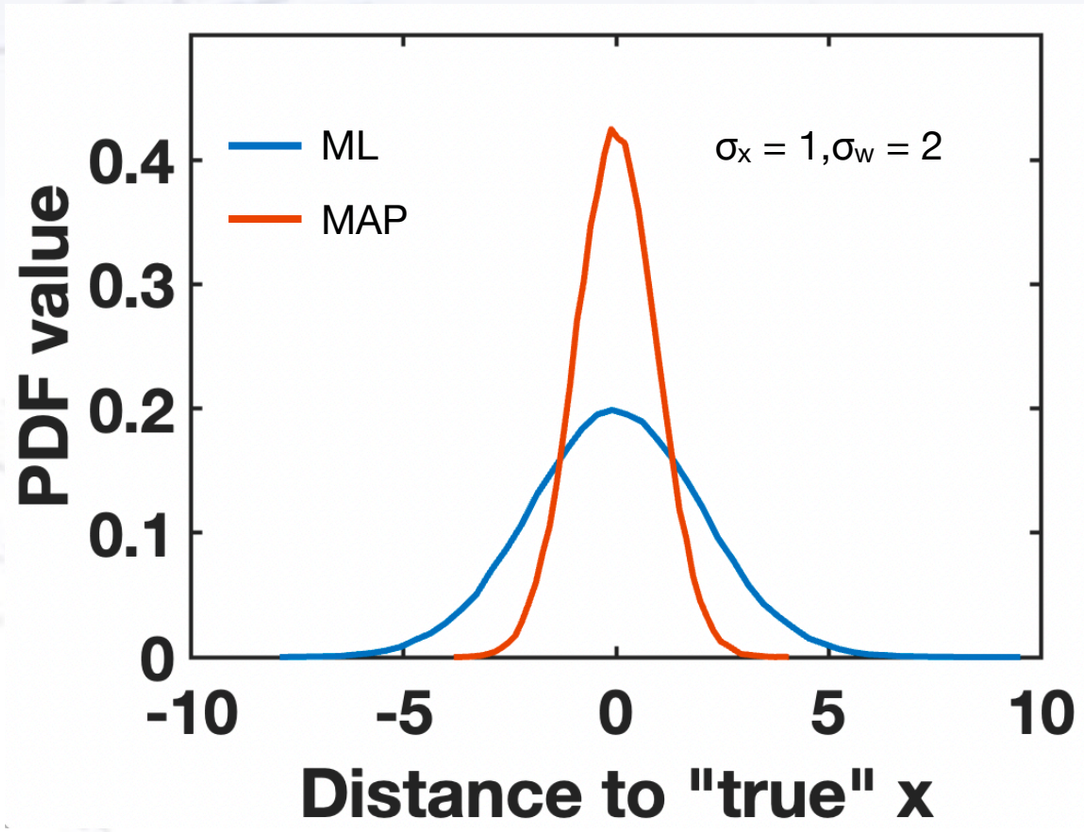
$$\frac{\partial f}{\partial x} = 0$$

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

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.

Is Bayesian statistics superior?

NO - that depends on the problem. But some problems are better suited for a bayesian approach.

If we have much data, frequentist and bayesian analysis yields the same *answer.* Lets assume we have a large set of data and we want to estimate the **mean**.

Frequentist maximum likelihood:

$$\widehat{\mu} = \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Maximum a Posteriori

$$p(\mu \mid \overrightarrow{X}) \propto p(\mu)\, p(\overrightarrow{X} \mid \mu) = \frac{1}{\sqrt{2\pi}\sigma_m} \exp\left(-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_m}\right)^2\right) \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(-\frac{1}{2}\left(\frac{x_j - \mu}{\sigma_v}\right)^2\right)$$

$$\hat{\mu}_{\text{MAP}} = \frac{\sigma_m^2\, n}{\sigma_m^2\, n + \sigma_v^2} \left(\frac{1}{n} \sum_{j=1}^{n} x_j\right) + \frac{\sigma_v^2}{\sigma_m^2\, n + \sigma_v^2} \mu_0$$

A maximum likelihood estimator coincides with the most probable Bayesian estimator given a uniform prior distribution on the parameters!!

The Kalman Filter

Dealing with sequential data is important. If you already have calculated a mean based on 1000 numbers - and you get a new number - it is nice not to calculate everuthing from scratch.

We all know how to calculate the weighted average of two numbers:

$$\hat{x} = \frac{x_1\sigma_2^2 + x_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

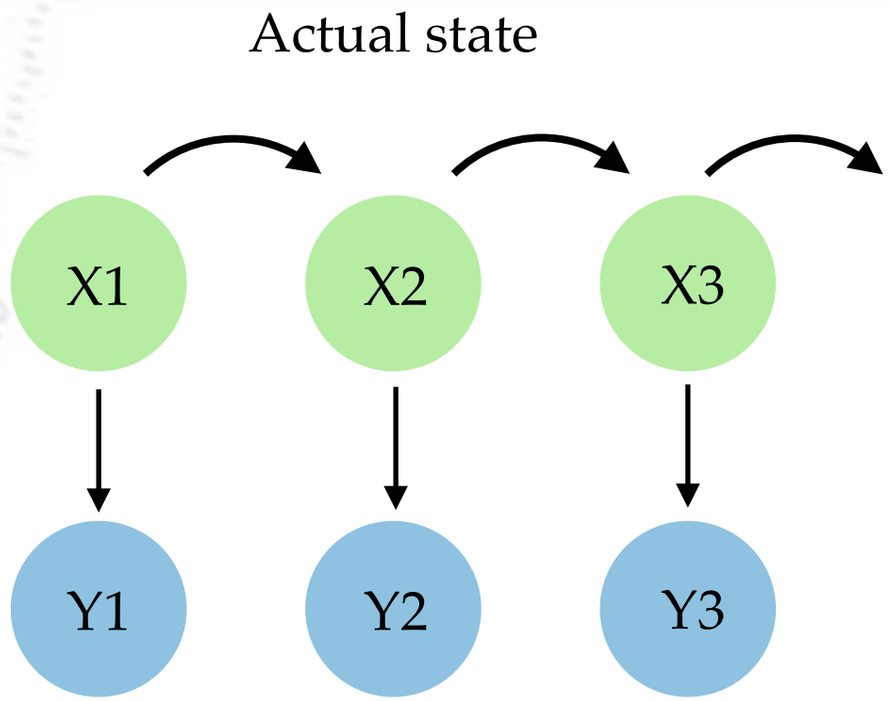
But a nice thing is that this can actually be expressed as:

$$\hat{x} = x_1 + K(x_2 - x_1) \quad K = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$$

This means that when we have sequential data, we can update our estimates as we include new points.

In this way we can for instance formulate the probability of the third datapoint as:

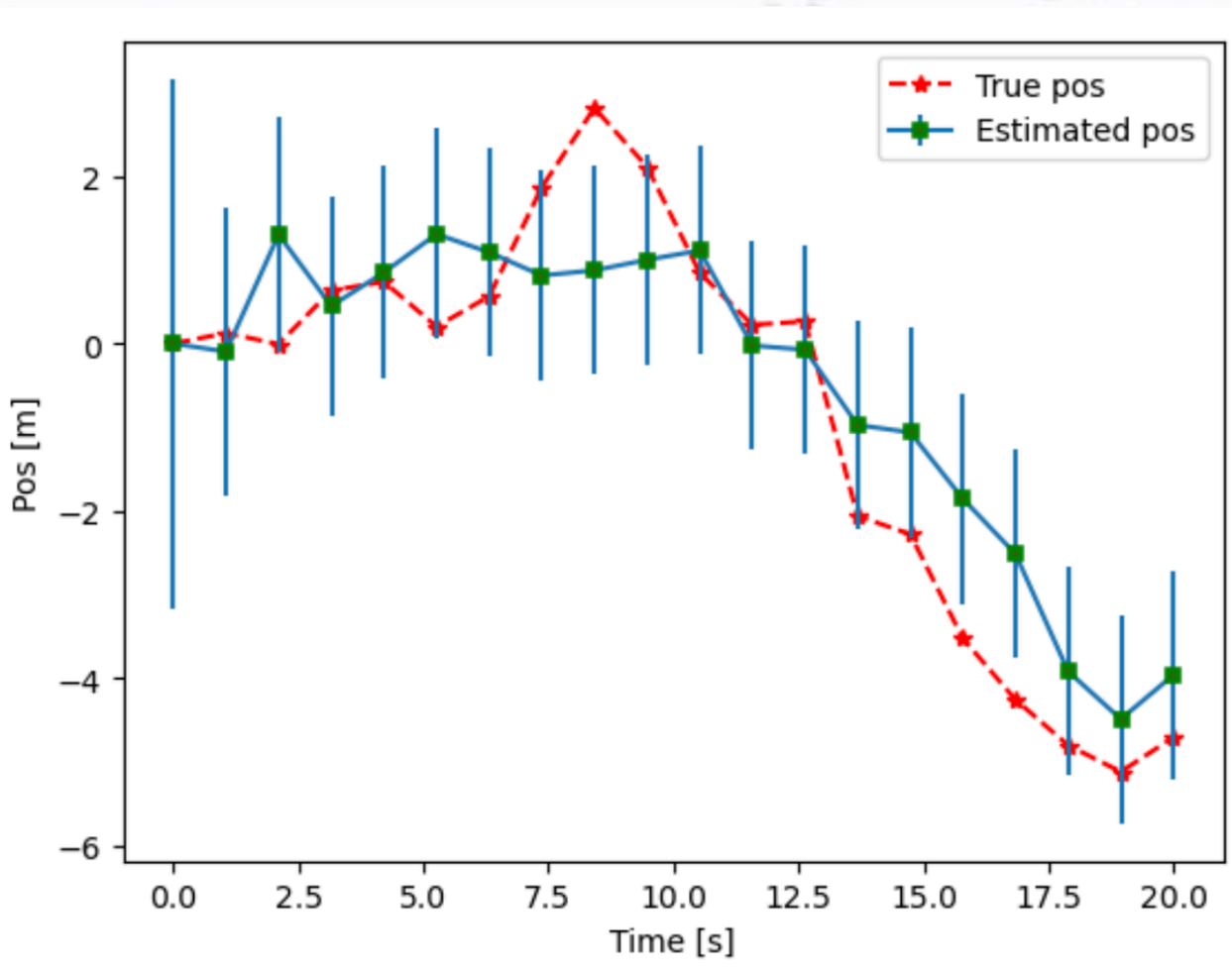
$$p(x_3 \mid y_1, y_2, y_3) \propto p(y_3 \mid x_3)p(x_3 \mid y_1, y_2)$$



The Kalman Filter

Lets see how this unfolds. Assume I have 20 positions of a particle - however this is the measured with experimental noise!

We know $\sigma_w = 2$ and $\sigma_x = 1$. We start by assuming $P(x_0) = \mathcal{N}(0,\sigma_{\hat{x}_1})$. Now updating the estiamte on means and variance allow us to estimate the position of the particle:



Bayesian Inference

Suppose we measure a diffusing particle at 5 positions with measurement error.

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

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:

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

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.

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

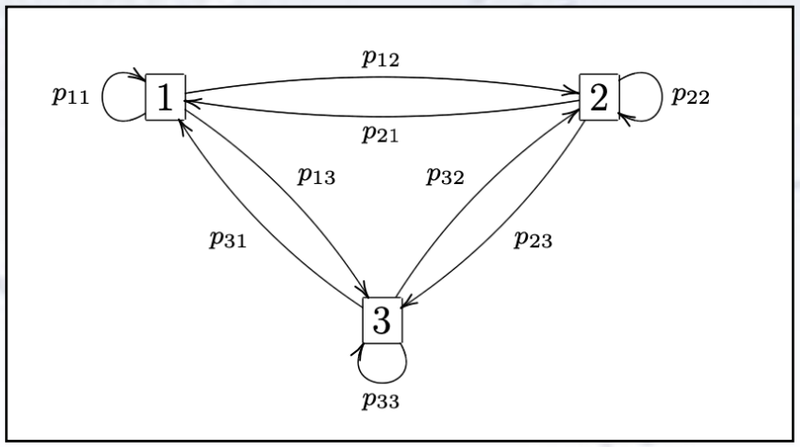
Mathematically we define Markov chains:

> 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 have 3 states and assign probabilities to move between states.



Based on these 9 probabilities, the matrix takes the form:

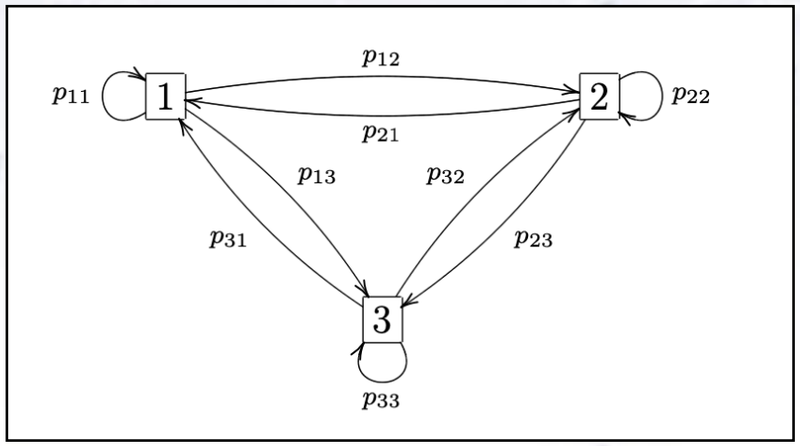
The general three-state Markov chain has transition matrix

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

Discrete Markov chains

So far so good - how can this be used?

Question: Given 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:

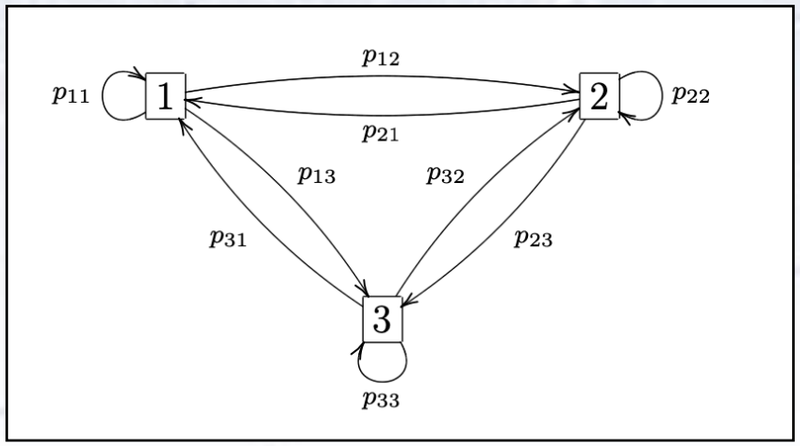
1 -> 2 -> 2 -> 2
1 -> 1 -> 2 -> 2
1 -> 1 -> 1 -> 2
1 -> 2 -> 1 -> 2
1 -> 2 -> 3 -> 2
1 -> 3 -> 2 -> 2

← Very tedious!!

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:

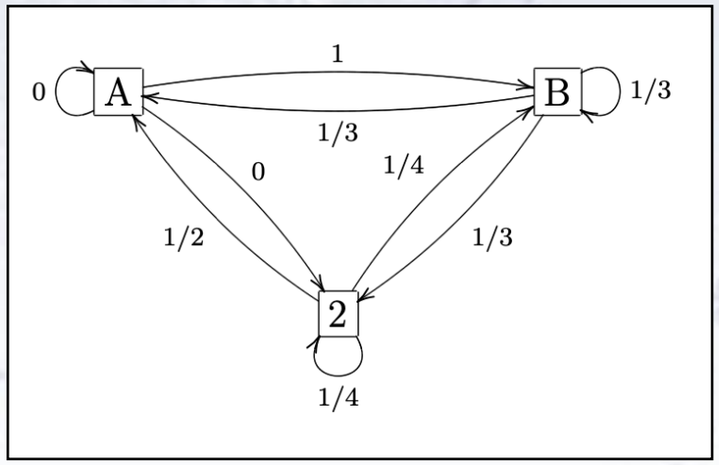
$$(P(X(n) = 1), \ldots, P(X(n) = N)) = \phi P^n.$$

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. We have a 3-state matrix:



Question: 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.

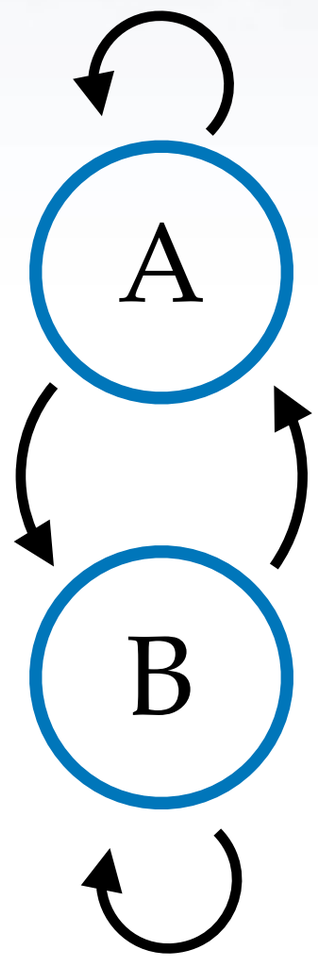
Markov Chains and detailed balance

For a Markov Chain we calculate the stationary distribution:
We get in our example:

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

$$\pi_A = \pi_A \cdot 0.7 + \pi_B \cdot 0.4$$

$$\pi_B = \pi_A \cdot 0.3 + \pi_B \cdot 0.6$$



And with the probability constraint:

$$\pi_A + \pi_B = 1$$

We get that: $\pi_A = \frac{4}{7}$ and $\pi_B = \frac{3}{7}$

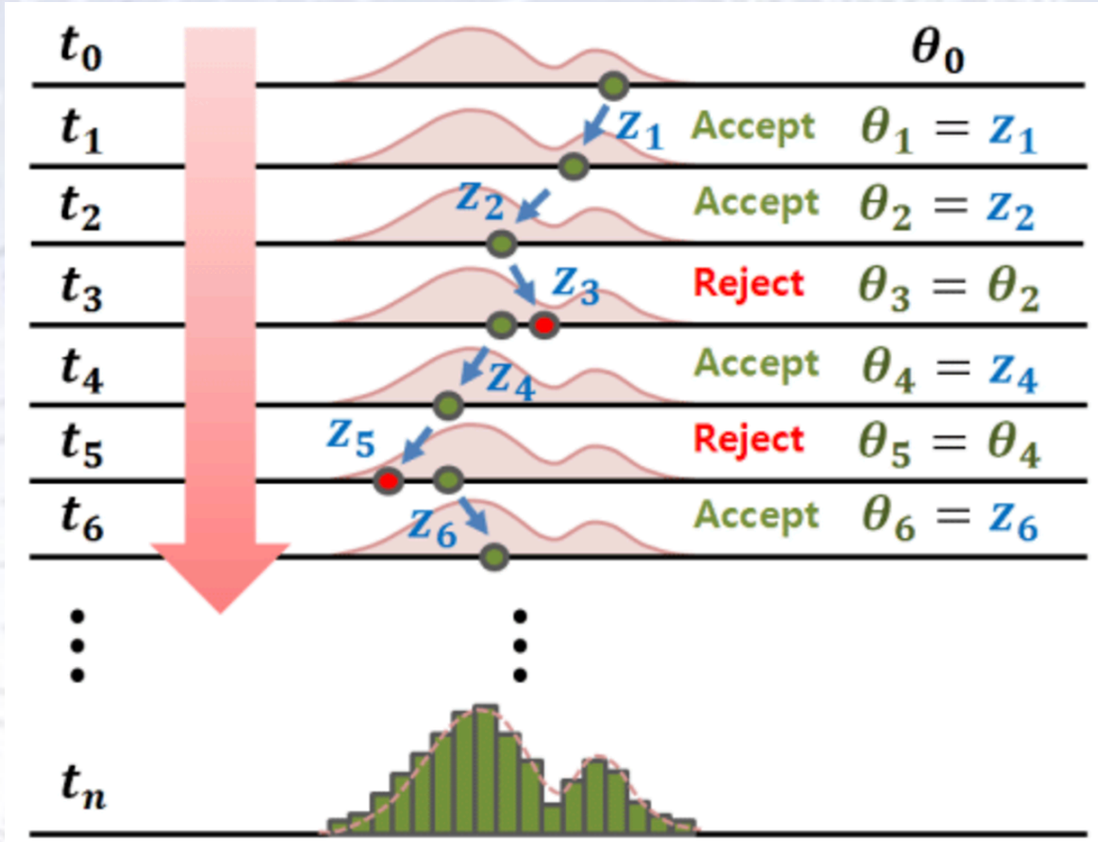
If it obeys detailed balance then: $\pi_i P_{ij} = \pi_j P_{ji},$

In our case this means: $\pi_A \frac{3}{10} = \pi_B \frac{4}{10}$

If a markov chain obey detailed balance it will converge to the stationary distribution from any starting state!

Markov Chain Monte Carlo

Markov chain Monte Carlo is a describes various techniques, that allow us to reach a distribution - which could be the uncertainty on a parameter.



They share the fundamental idea:

1) Propose a new parameter value as a state
2) Move to this state with probability that obeys detailed balance.
3) As we keep the sample, we will be sure to obtain the correct distribution of our parameters.

Metropolis Hastings

The best known MCMC technique is the Metropolis-Hastings algorithm.

The acceptance probability is now:

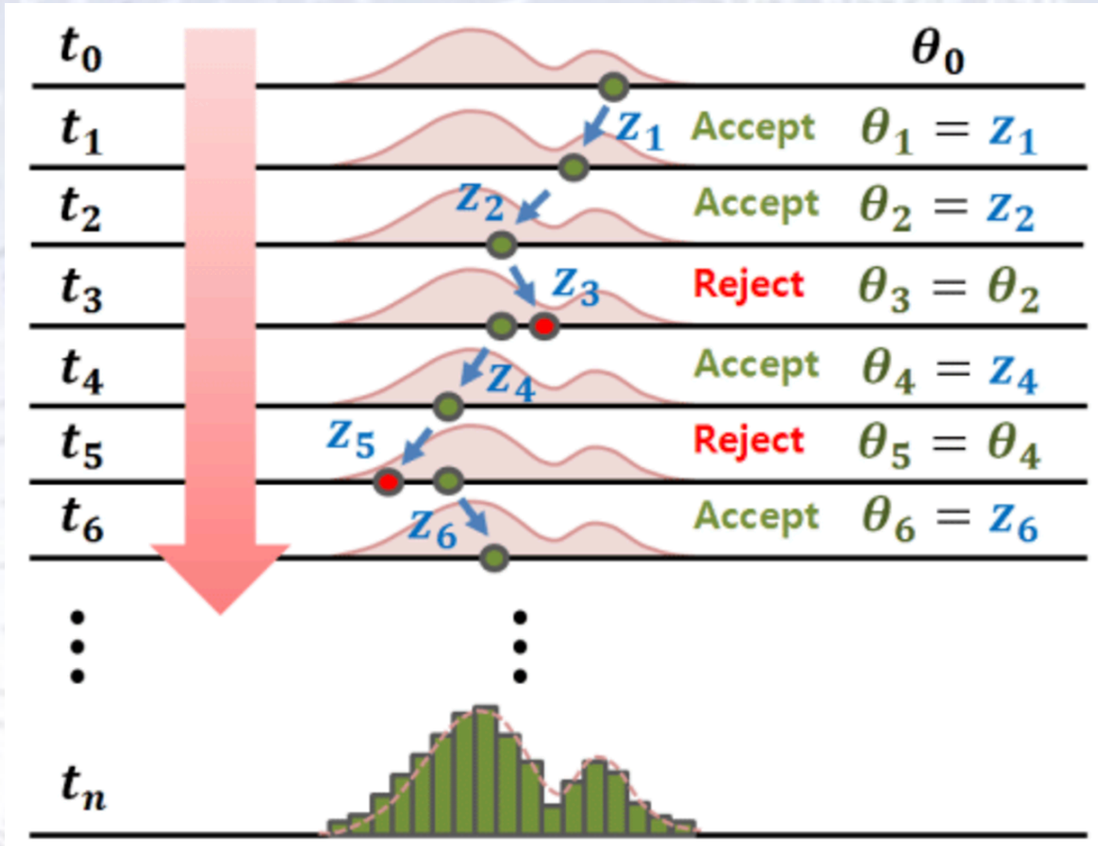
$$\alpha(\theta_t, \theta') = \min\left(1, \frac{\pi(\theta'|\text{data})q(\theta_t|\theta')}{\pi(\theta_t|\text{data})q(\theta'|\theta_t)}\right)$$

Here we have used bayesian statistics as:

$$\pi(\theta|\text{data}) \propto L(\theta|\text{data}) \cdot \pi(\theta)$$

The value q, is the probability to choose a new value. Typically this is symmetric, so it goes out.
If we have no prior, we end up with:

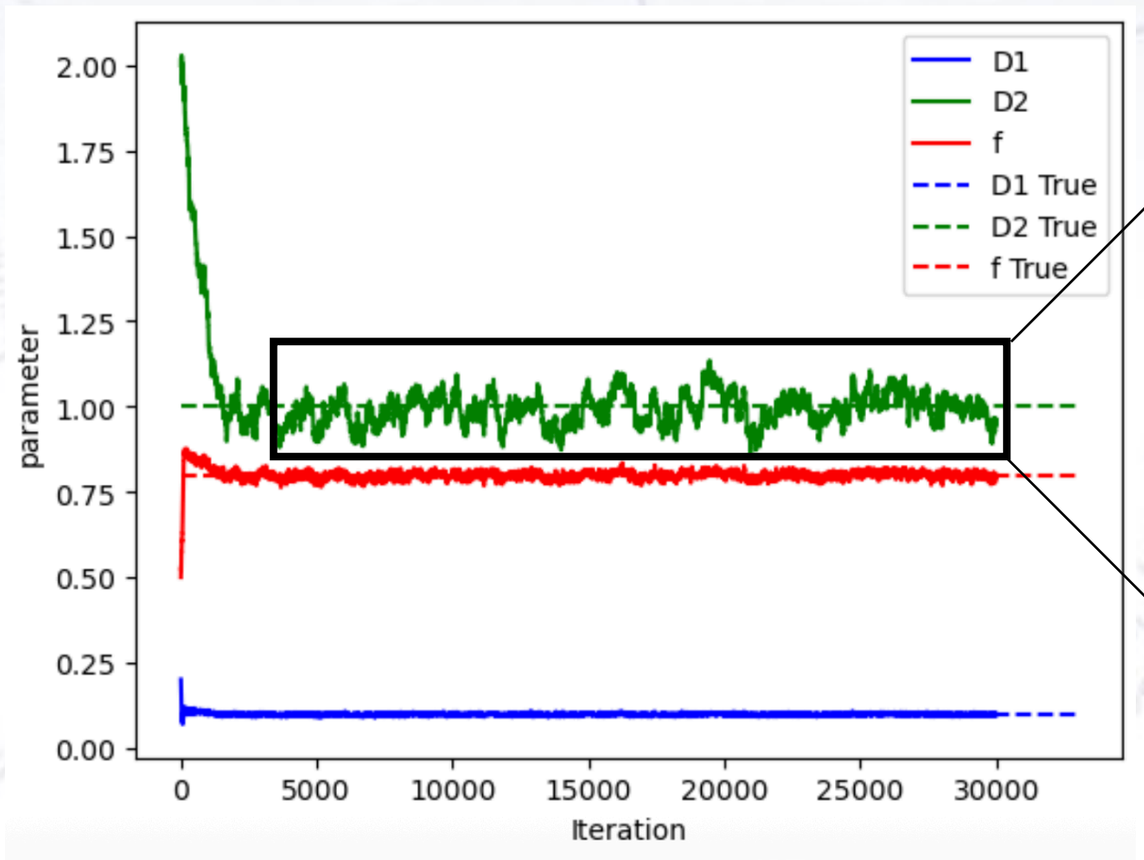
$$\alpha(\theta_t, \theta') = \min\left(1, \frac{L(\theta'|\text{data})}{L(\theta_t|\text{data})}\right)$$

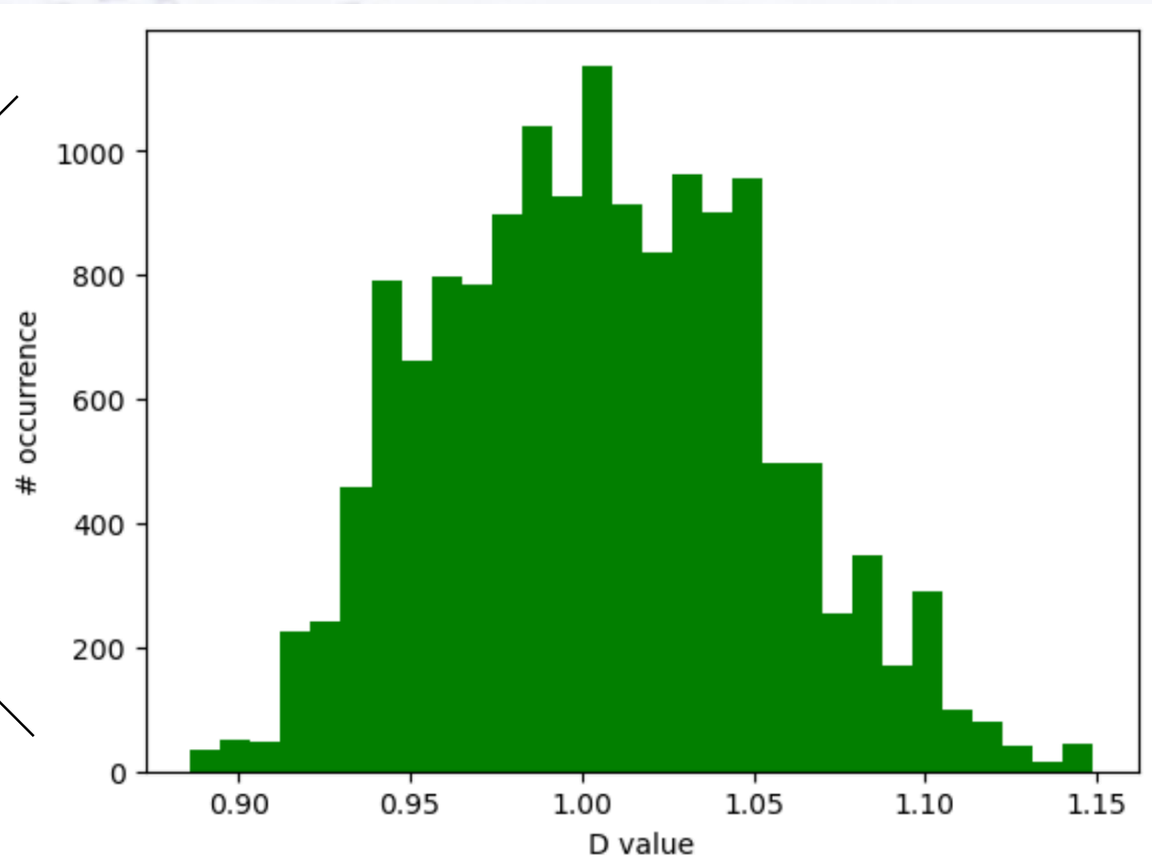


Metropolis Hastings

Imagine we measure diffusion positions from two different particles (1 and 2). We will estimate the diffusion foeccient of each and the fraction.

With this we can sample all parameters:

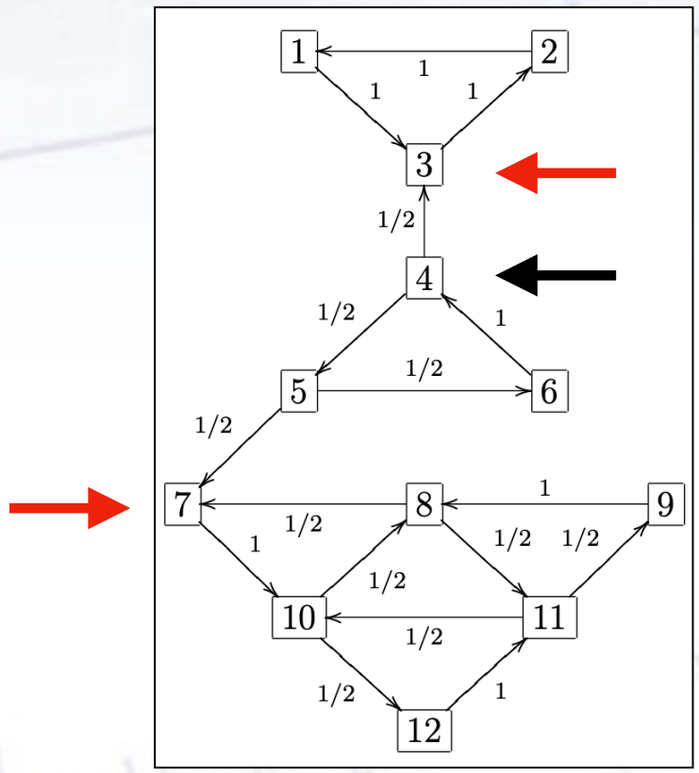





```
Diffusion 1 coefficient is 0.100 +/- 0.003  (D1_true = 0.1)
Diffusion 2 coefficient is 1.01 +/- 0.05    (D2_true = 1.0)
Fraction is 0.80 +/- 0.01                   (f_true = 0.8)
```

Irreducibility and communication classes

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



Once you reach state 3 or state 7 you can never return.

States 1-3 = recurrent class.
States 7-12 = recurrent class.
States 4-6 = transient class

Absorption probabilities

We are interested in the probabilities of reaching specifically state 3 or state 7.

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?

We can construct the states in the following way:

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

P is probability between recurring classes

Q is matrix between transient classes

S is the transition matrix with probabilites to go into the absorbing states

And we define:

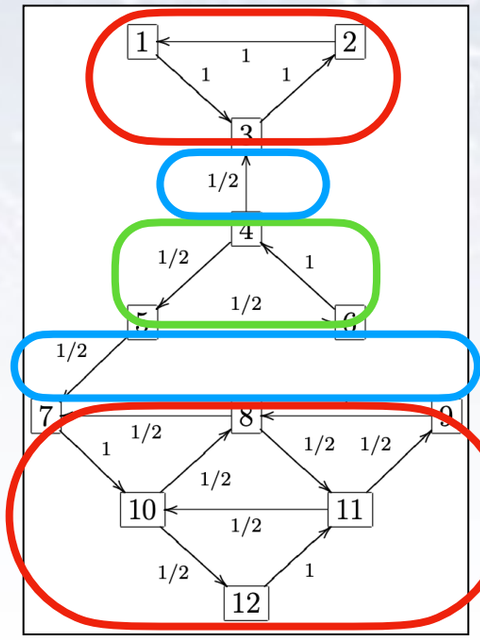
$M = (I - Q)^{-1}$ Expected number of visits to each state before absorption

$A = (I - Q)^{-1}S$ Probabilities to for each absorbing states being the firsts to visit

Absorption probabilities

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

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



I have restructured the matrix.

The two states 3 and 7 are absorbing - we don't care what is going on in state 12.

The expected number of visits to state 5 before absorption, given we start in state 4:

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

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

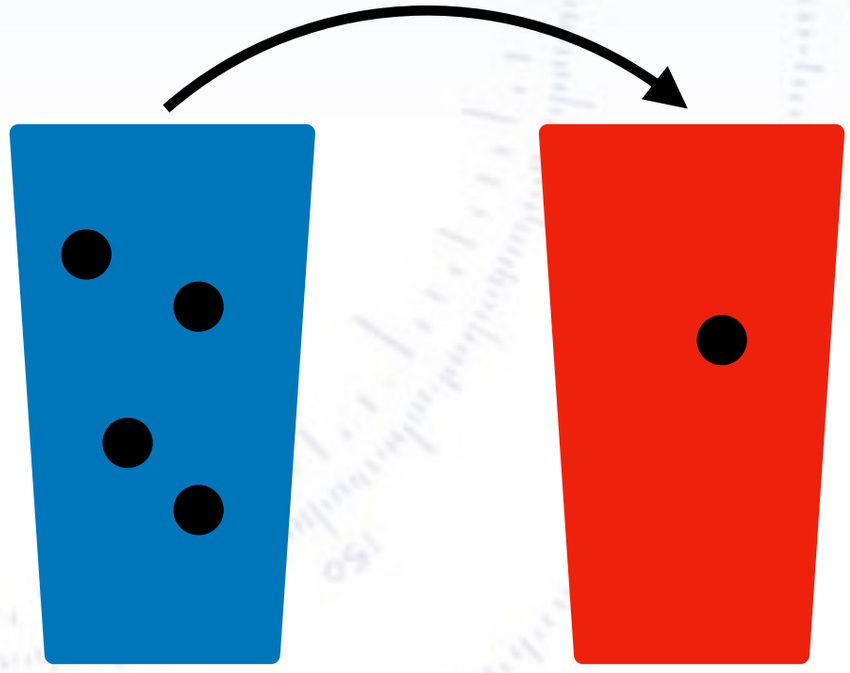
$$A_7 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} * \left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix} \right)^{-1} * \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \\ 0 & 0 \end{pmatrix} *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

Examples and hints for todays problems

Discrete Markov chains

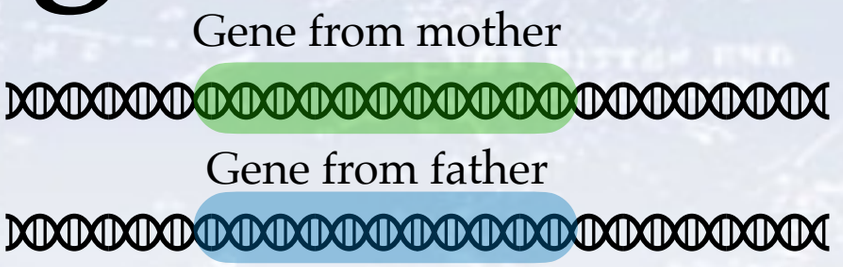
It happens quite often that some statistical problem can be formulated by a Markov Chain. We look at one example that is also covered in the excersizes - the Ehrenfest urn problem.

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.



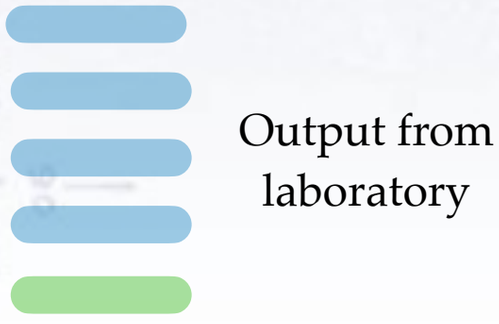
What is the probability to have 3 balls in the blue container after 10 iterations? This can be formulated as a Markov chain. Can you see why?

Estimating one gene



For one gene, each individual can have the same from mother and father (AA) or (BB) or one from each (AB)

Problem: Given you measure N random copies of that gene, what is the probability for each genotype.



We measure one gene (Say A_0) and we want to estimate the probability that this comes from AA. This is the probability

Note: The probability that it is an A is switched to a B0: 0.15 is one matrix (XX).
The transposed of this is used to calculate the probability of B: $P(B) = P(B \mid B0) + P(B \mid A0)$