Maximum Likelihood

Airlie McCoy

Friday 17th May 2013

UNIVERSITY OF CAMBRIDGE

1. Collect diffraction data

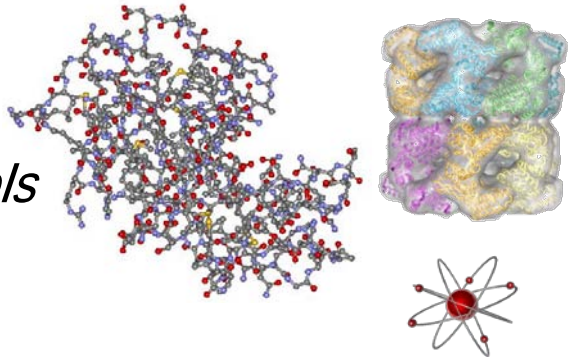
Need to find phases to calculate electron density

2. Propose models

For EP, heavy atoms

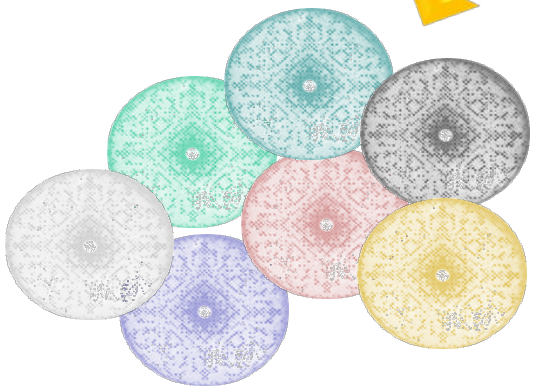
For MR, positioned models

For refinement, atoms



3. Calculate "diffraction" (F's) from proposed models

Phases <u>and</u> amplitudes

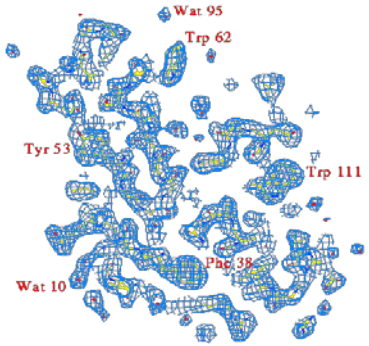


4. Compare and find best match

Compare observed and calculated structure factors (F's) with scoring function

6. Find differences

Solve structure

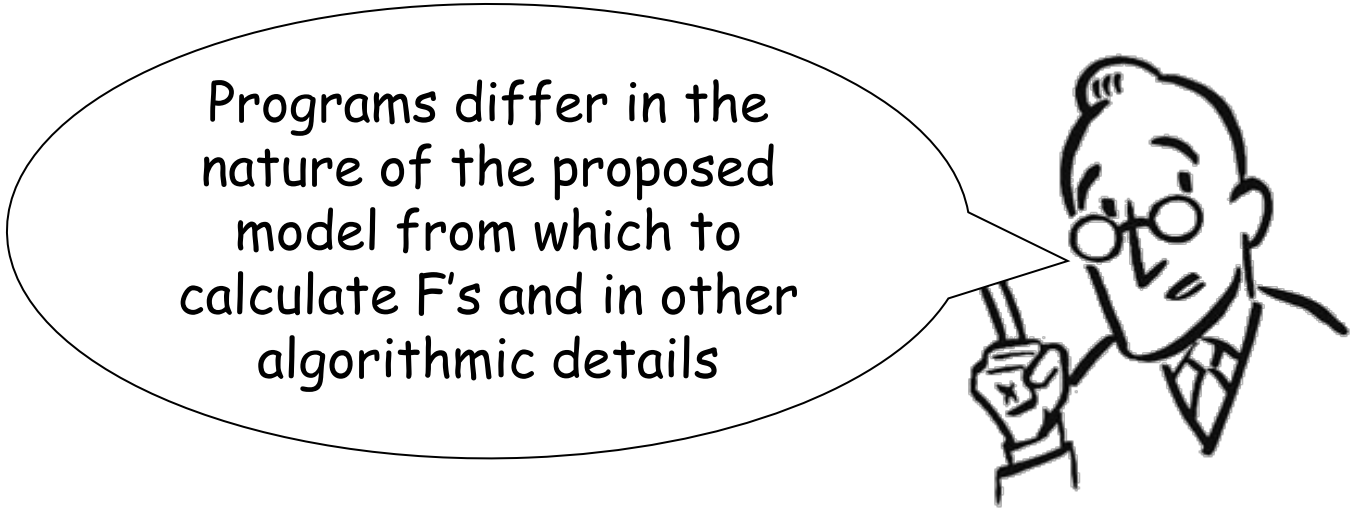


5. Calculate electron density

Using observed amplitudes and phases of selected model

Maximum Likelihood

- What is the “best match” between the observed and calculated structure factors?
  - Need a scoring function for comparison
- Modern programs use Maximum Likelihood
  - Phaser
  - SHARP
  - Refmac
  - Phenix.refine
  - Phenix.autosol
  - Solve/resolve



Likelihood Function

$$\min\left[-\sum_{hkl}\log\left(\frac{2F_{obs}}{\sigma_{\Delta}^{2}}e^{\frac{F_{obs}^{2}+D^{2}F_{calc}^{2}}{\sigma_{\Delta}^{2}}}I_{O}\left(\frac{2F_{obs}DF_{calc}}{\sigma_{\Delta}^{2}}\right)\right)\right]$$

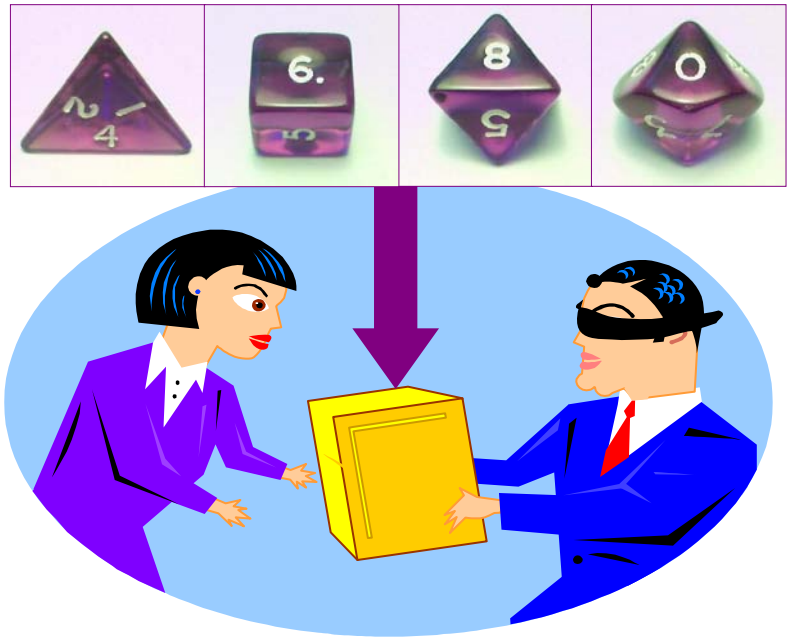
- This ML equation/function is the basis for ML molecular replacement and refinement software
- Equations very similar to this are used in ML experimental phasing and density modification
- The aim of this talk is to understand this fundamental equation

Concepts:

- Maximum Likelihood
- Independence
- Log(Likelihood)
- Bayes' Theorem
- Integrating out Variables
- Central limit theorem

Maximum Likelihood

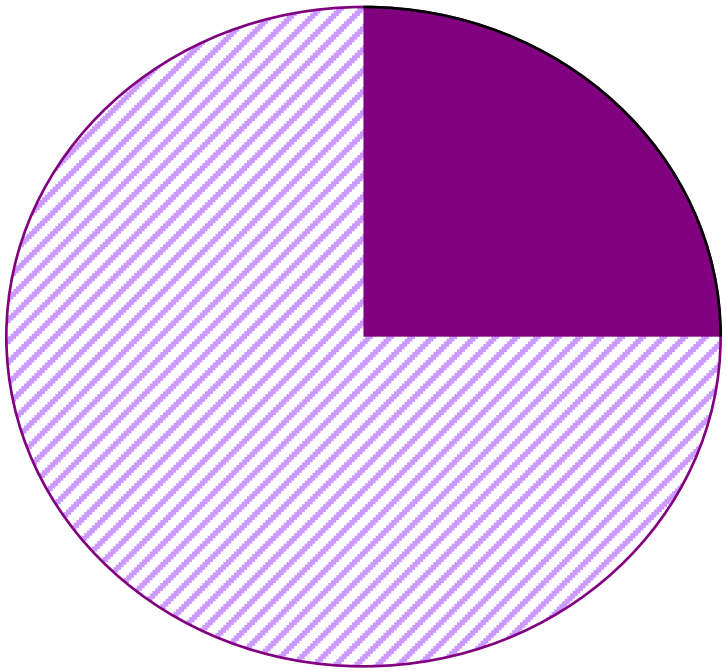
A game of dice



- Put four unbiased dice in a box
- I select a die at random
- How often will you guess correctly which die I selected?

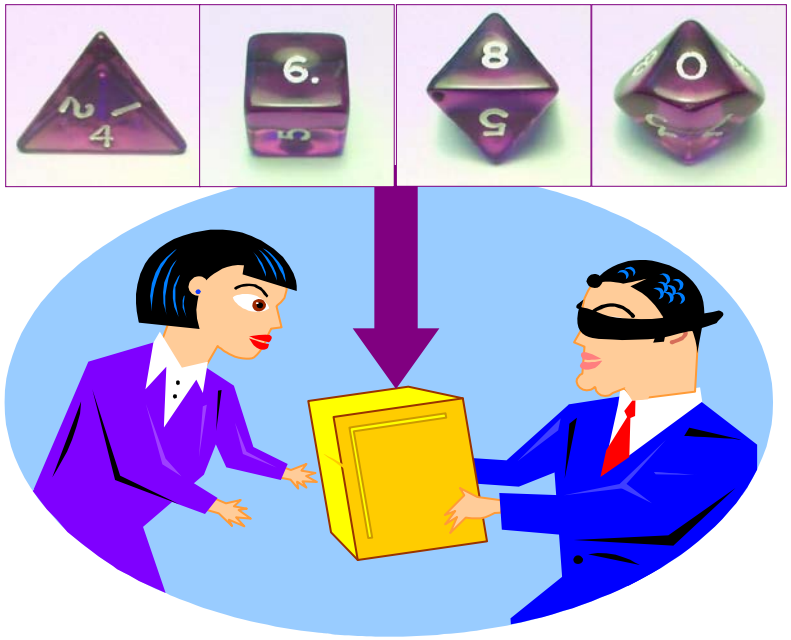
Probability

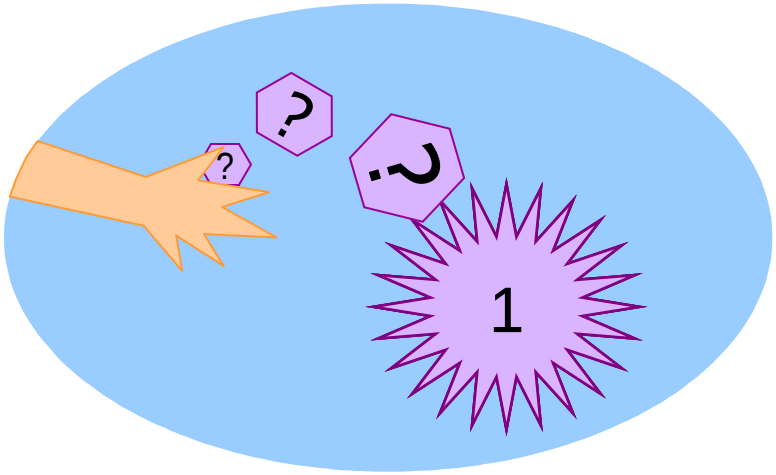
- In the game of dice you have a 1 in 4 chance of being right
- If a large number of people guessed, one quarter would be right each time
- If you play the game many times, you will be right a quarter of the time



Maximum likelihood

A game of dice with data



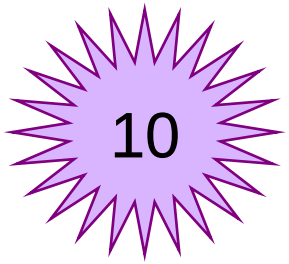


- Put four unbiased dice in a box
- I select a die at random
- I roll the die and tell you the result of the roll
- Which die did I most likely select?

Maximum likelihood

Roll a 10

- The die obviously must have been the 10 sided die
- What does “must” mean in probabilities?



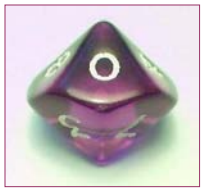
$$P(10;\boxed{10}) = \tfrac{1}{10}$$

$$P(10;\boxed{8}) = 0$$

$$P(10;\boxed{6}) = 0$$

$$P(10;\boxed{4}) = 0$$

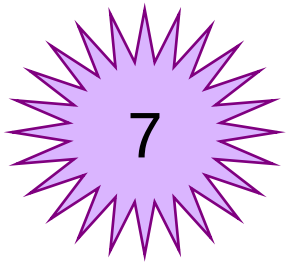
most likely



Maximum likelihood

Roll a 7

- The die could have been the 10 sided or the 8 sided die
- Which die is most likely?



$$P\left(7;\boxed{10}\right) = \tfrac{1}{10}$$

$$P\left(7;\boxed{8}\right) = \tfrac{1}{8}$$

$$P\left(7;\boxed{6}\right) = 0$$

$$P\left(7;\boxed{4}\right) = 0$$

most likely

Maximum likelihood

Roll a 1

- Could have been rolled by any of the dice
- The most likely die is the one with the highest probability of generating the data

1

$$P(1;\boxed{10}) = \tfrac{1}{10}$$

$$P(1;\boxed{8}) = \tfrac{1}{8}$$

$$P(1;\boxed{6}) = \tfrac{1}{6}$$

$$P(1;\boxed{4}) = \tfrac{1}{4}$$

most likely

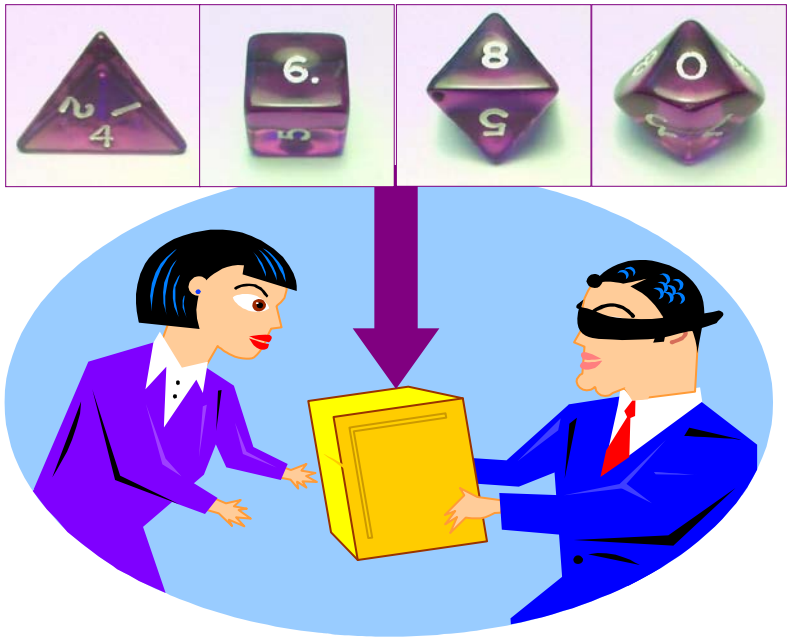
Maximum likelihood

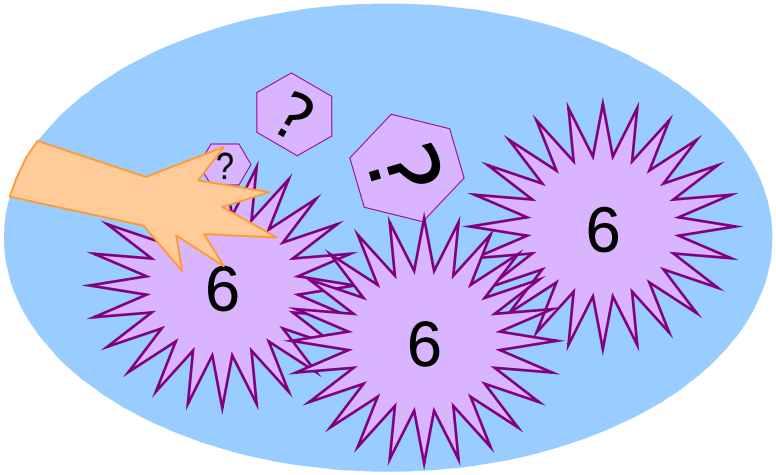
Crystallography

- Data are the F_O in <u>reciprocal space</u>
  - or merged F_O and ΔF_O
  - or merged F_O^+ and F_O^-
  - or merged I_O (or I_O and ΔI_O – or I_O^+ and I_O^-)
  - or unmerged I_O's
  - time of collection t_O
- Model is the structure in <u>real space</u>
- Need to calculate the structure factor $\mathbf{F}_C$ from the model in order to compare with data
- "Solution" is the model with the $\mathbf{F}_C$ with the highest likelihood of generating F_O

Independence
and log-likelihood

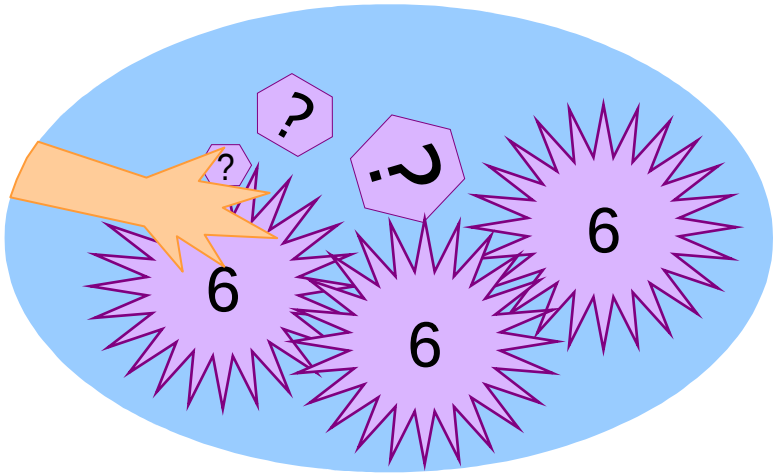
A game of dice with more data





- Put four unbiased dice in a box
- I select a die at random
- I roll that die <u>three</u> times and tell you the results
- Which die did I most likely select?

A game of dice with more data

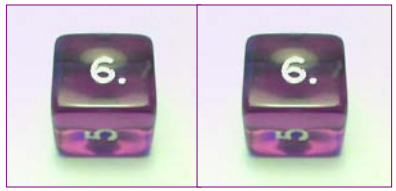


- What is the chance of throwing a 6 three times from a 6-sided die?
- The chance of throwing a 6, or any other number, the second, or third time is not influenced by the value of the first roll - they are <u>independent</u>

Multiplying probabilities

- When probabilities are independent they multiply

$$P\left(6;\boxed{6}\right) = \tfrac{1}{6} = 0.16666667$$



$$P\left(6,6;\boxed{6}\right) = \tfrac{1}{6} \times \tfrac{1}{6} = \tfrac{1}{36} = 0.0277778$$

$$P\left(6,6,6;\boxed{6}\right) = \tfrac{1}{6} \times \tfrac{1}{6} \times \tfrac{1}{6} = \tfrac{1}{216} = 0.0046296$$

100 times

$$P\left(6\ldots\times 100;\boxed{6}\right) = 6^{-100} = 1.53064\times 10^{-78}$$

Computers and small numbers

"Oh great one, what is the probability of throwing a 6 from a six sided die one billion times?"

```
> SYSTEM-F-FLTOVF_F, arithmetic fault,
floating overflow at PC=00006244,
PSL=03C0 0020 %TRACE-F-TRACEBACK,
symbolic stack dump follows
module name     routine name    line
OVERF           OVERF           104
DPARA$MAIN      DPARA$MAIN      276
```

Computers can not store numbers very close to zero

Computers and log(small numbers)

"Oh great one, what is the logarithm of the probability of throwing a 6 from a six sided die one billion times?"

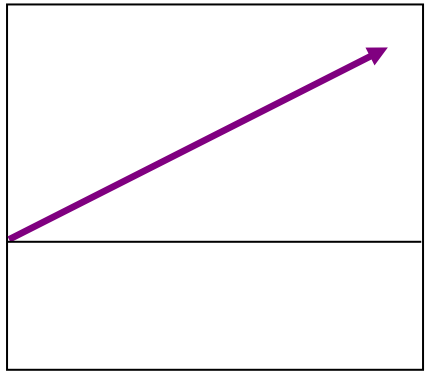
```
> -778151250.4
```

log(likelihood) is not close to zero

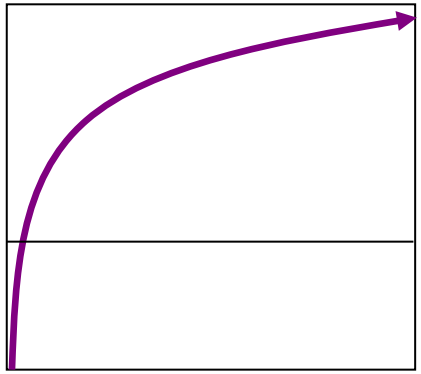
- So the log(likelihood) solves the small number problem
- But can we just switch to using the log(likelihood)?

Optimisation and logarithms

- Logarithmic functions are “monotonic” functions
  - *i.e.* they “preserve the given order”
  - If $y_1 < y_2$ for all $x_1 < x_2$ then $\log(x_1) < \log(x_2)$
- The parameter values obtained optimising log(likelihood) are the same as those obtained optimising likelihood
  - **Optimising log(likelihood) ≡ Optimising likelihood**



$y = x$



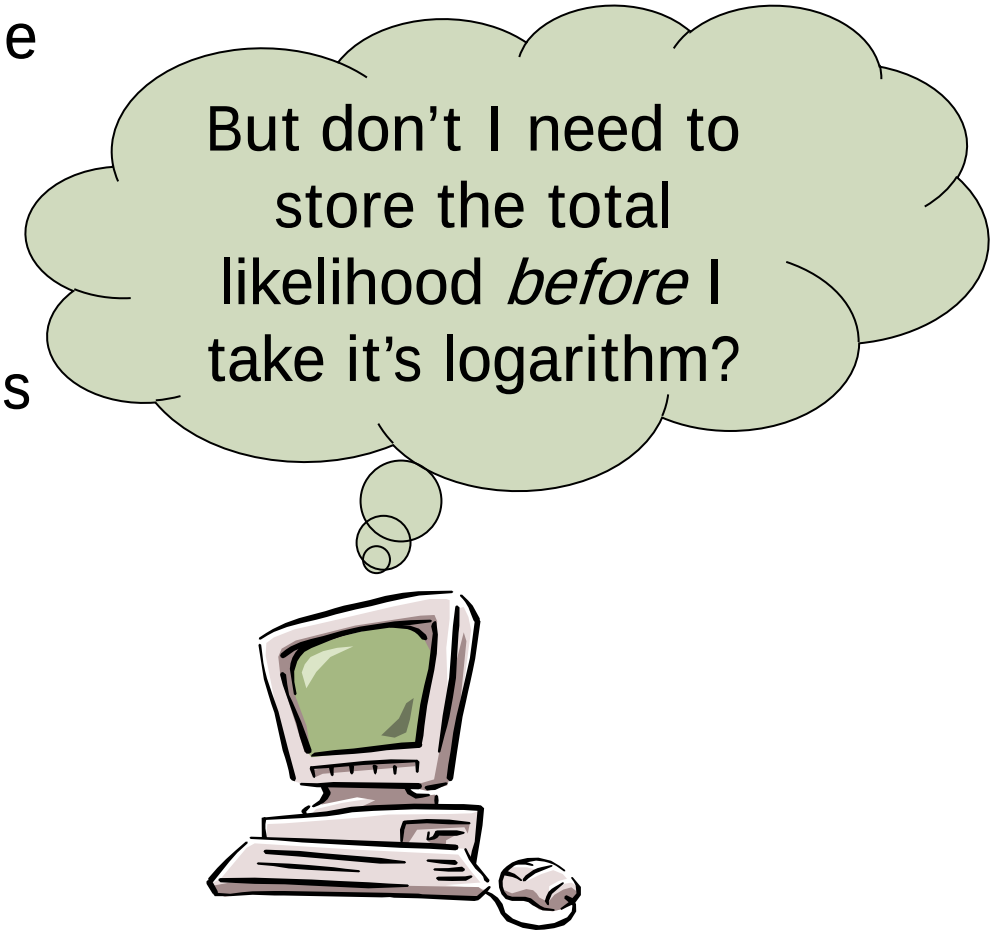
$y = \log(x)$

Logarithms, products and sums

- No, there is a shortcut to the log(total likelihood) when total likelihood is a product of likelihoods
- If log(total likelihood) equals log(Π likelihoods)

  product
- Then log(total likelihood) also equals Σlog(likelihoods)

  sum



Logarithms and independence

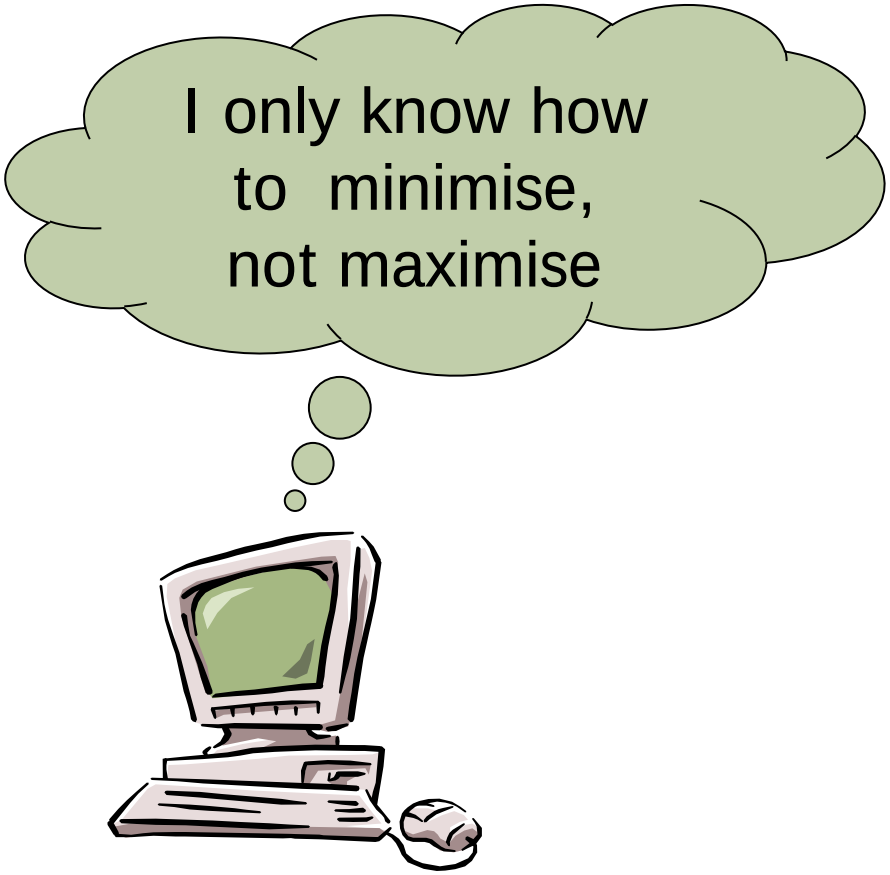
$$\log(\prod \text{likelihoods}) = \sum \log(\text{likelihoods})$$

$$\log\left(P\left(3,3;\boxed{6}\right)\right)$$
$$= \log\left(P\left(3;\boxed{6}\right)\times P\left(3;\boxed{6}\right)\right)$$
$$= \log\left(\tfrac{1}{6}\times\tfrac{1}{6}\right)$$
$$= \log(0.0277)$$
$$= -1.556$$

$$\log\left(P\left(3,3;\boxed{6}\right)\right)$$
$$= \log\left(P\left(3;\boxed{6}\right)\right)+\log\left(P\left(3;\boxed{6}\right)\right)$$
$$= \log\left(\tfrac{1}{6}\right)+\log\left(\tfrac{1}{6}\right)$$
$$= -0.778-0.778$$
$$= -1.556$$

Minimising

- Computer algorithms are designed to minimise
- Therefore we optimise our parameters by minimising the -log(likelihood)



Independence and log-likelihood

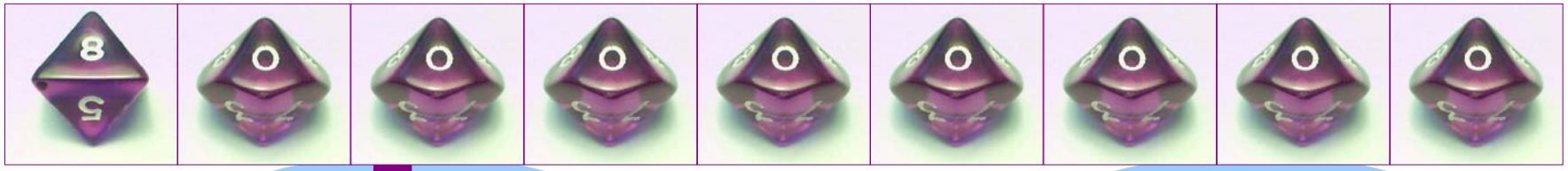
Crystallography

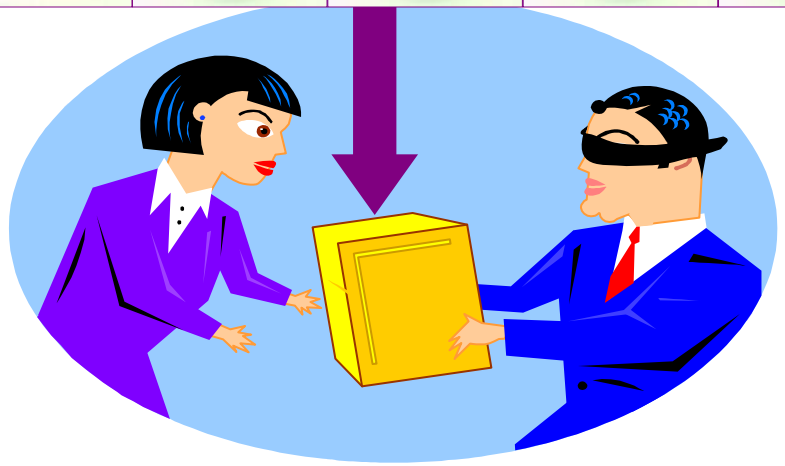
- ML algorithms assume reflections are independent
- This is an approximation: reflections are not independent, due to the presence of solvent and any non-crystallographic symmetry
- However, the approximation is very good
- Total likelihood is the product of the reflection likelihoods
- The algorithms actually calculate the log(likelihood)
- Total log(likelihood) is the sum of the reflection log(likelihoods)
- Maximum likelihood search and refinement algorithms minimise the –log(likelihood)

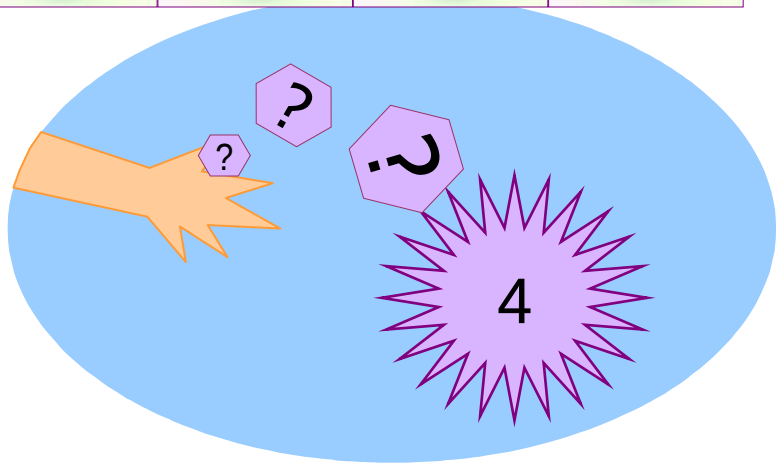
Bayes' Theorem

and prior probability

A game of dice with copies of a die







- Put one 8-sided die and eight 10-sided dice in a box
- I select a die at random
- I roll the die and tell you the result of the roll
- Which die did I most likely select?

Prior probability and Bayes' theorem

- In this case the prior probability of selecting the 10-sided die dominates the higher likelihood of throwing a 4 from the 8-sided die than from the 10-sided die

Bayes' Theorem

$$P(\text{model};\text{data}) = \frac{P(\text{model})}{P(\text{data})} \times P(\text{data};\text{model})$$

- In experimental situations, $P(\text{data})$ is constant, and when <u>comparing probabilities</u> can be ignored

$$P(\text{model};\text{data}) = P(\text{model}) \times P(\text{data};\text{model})$$

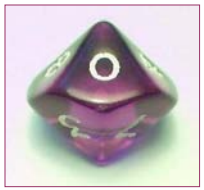
prior probability likelihood

Roll a 4

$$
\begin{aligned}
P\left(\boxed{10};4\right) &= P\left(\boxed{10}\right)P\left(4;\boxed{10}\right) \\
&= \tfrac{8}{9}\times\tfrac{1}{10} \\
&= \tfrac{8}{90} \\
&= 0.0888
\end{aligned}
$$

$$
\begin{aligned}
P\left(\boxed{8};4\right) &= P\left(\boxed{8}\right)P\left(4;\boxed{8}\right) \\
&= \tfrac{1}{9}\times\tfrac{1}{8} \\
&= \tfrac{1}{72} \\
&= 0.01388
\end{aligned}
$$

most likely



Crystallography

- Bayes' Theorem is used in refinement

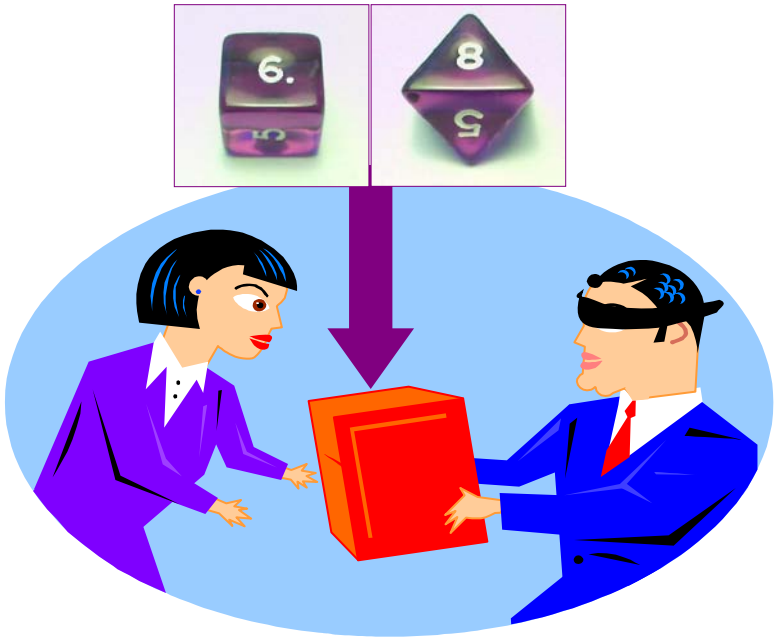
$$P(\text{model};\text{data}) = P(\text{model}) \times P(\text{data};\text{model})$$

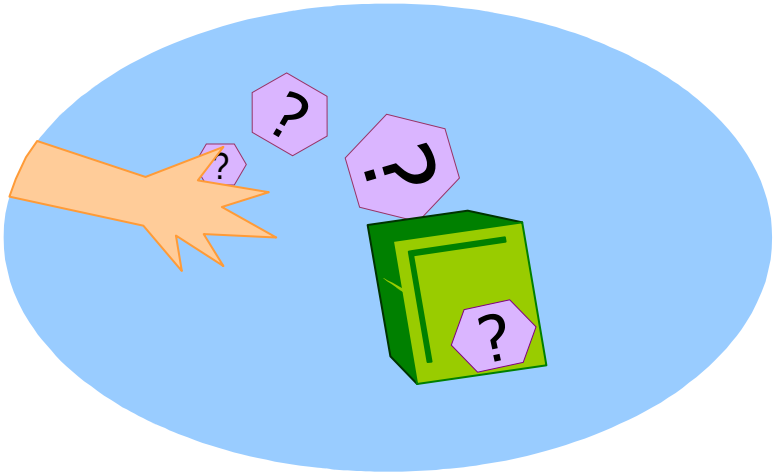
  - Prior probability from the chemistry i.e. knowledge of bond-lengths, bond-angles, planarity etc
  - Likelihood from the X-ray diffraction experiment
- Also used in density modification
- Fundamental principle in the method of "integrating out nuisance variables"...

Integrating out nuisance variables

Integrating out nuisance variables

A game of dice with unknown dice

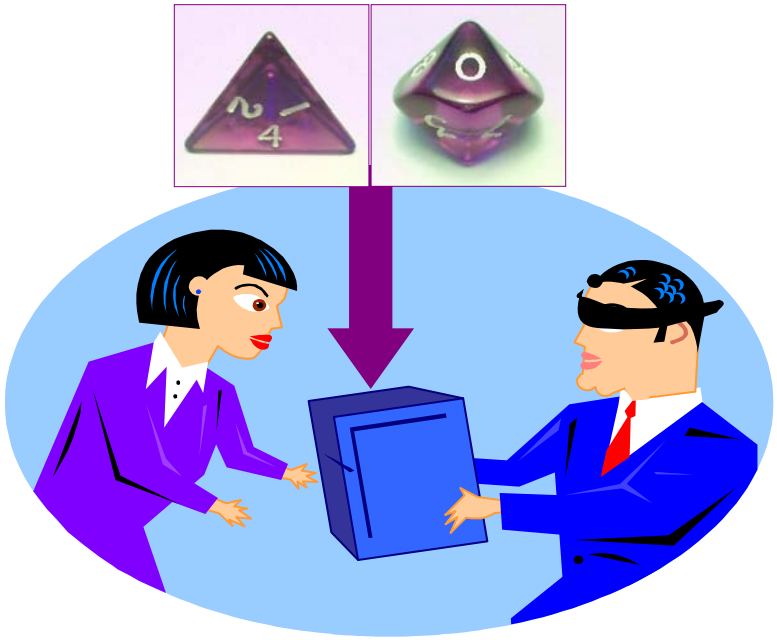


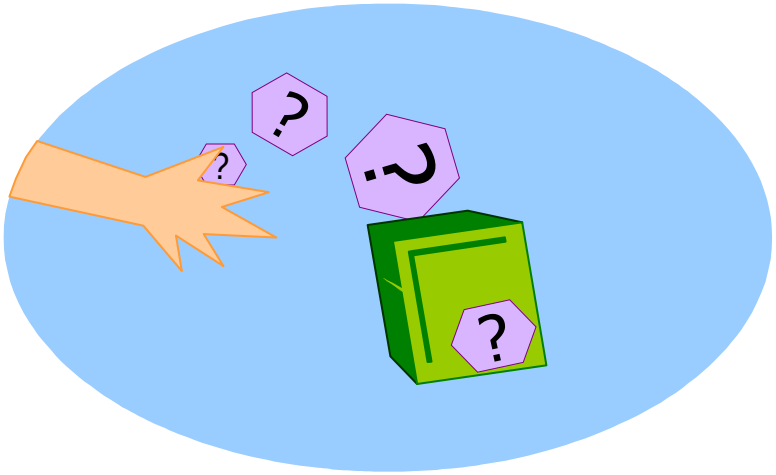


- Put a 6-sided and an 8-sided die in a red box
- I select a die at random and put it in a green box

Integrating out nuisance variables

A game of dice with unknown dice

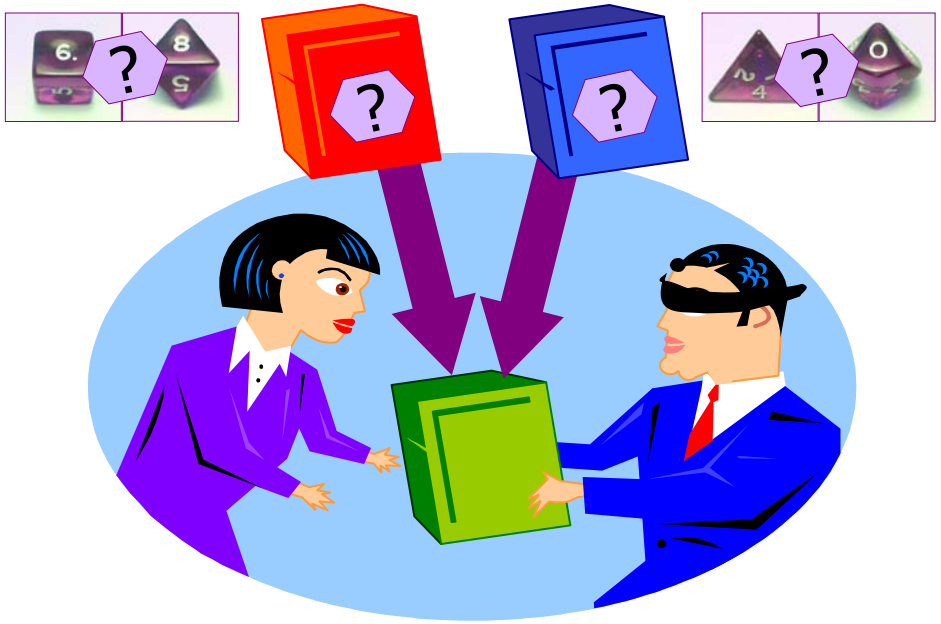


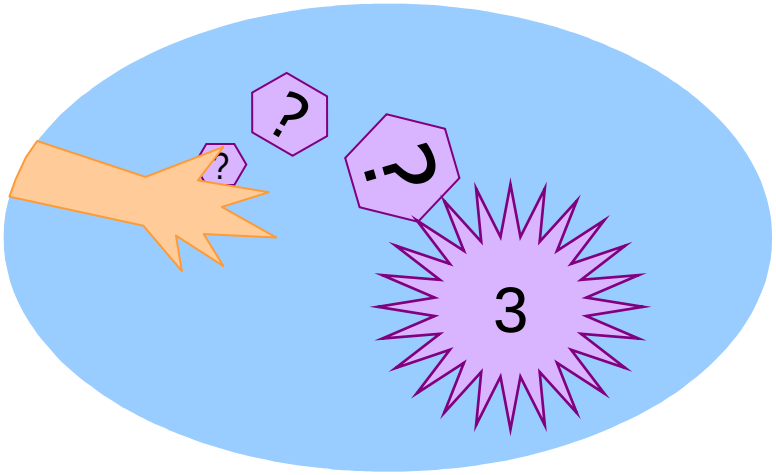


- Put a 4-sided and 10-sided die in a blue box
- I select a die at random and put it in the same green box as the first die (from the red box)

Integrating out nuisance variables

A game of dice with unknown dice

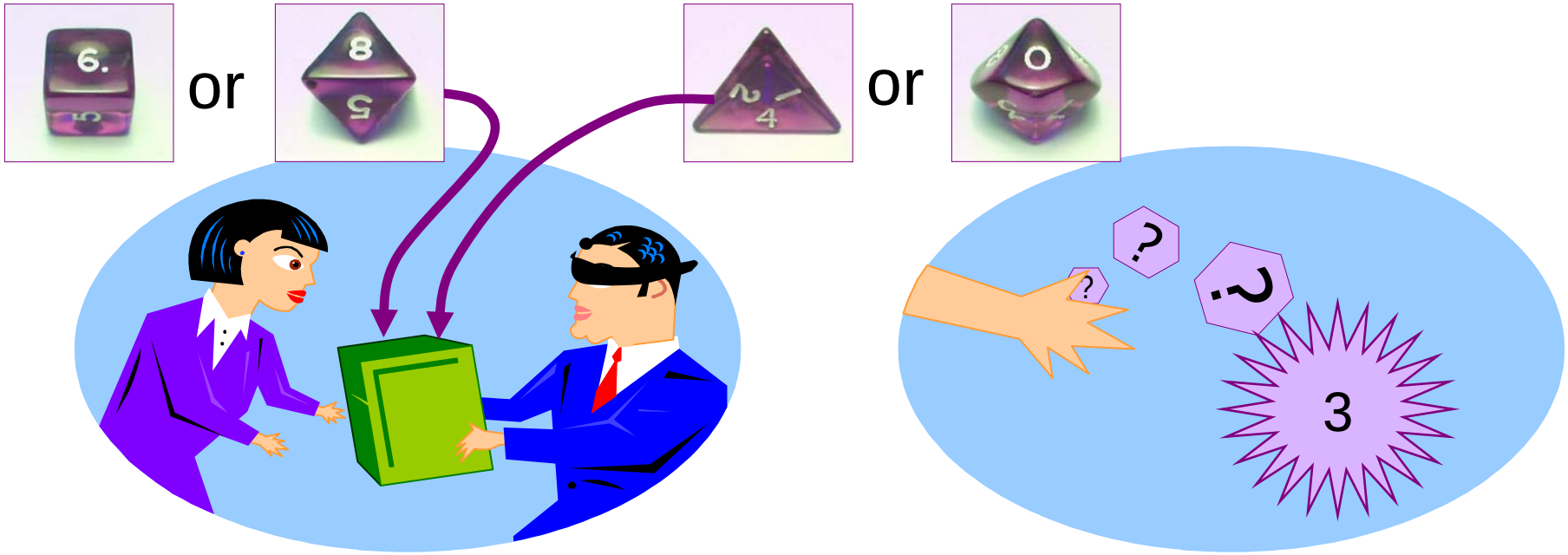




- I select a die at random from the green box, roll the die and tell you the result
- Did the die come from the red box or the blue box?

Integrating out nuisance variables

A game of dice with unknown dice



- There are two dice in the box. One is <u>either</u> a 6 <u>or</u> an 8 sided die and the other is <u>either</u> a 4 <u>or</u> a 10 sided die
- I select a die, roll, and tell you the result
- Which of the two dice possibilities did I select?

Integrating out nuisance variables

Roll a 3

$$P\left(\boxed{6 \text{ or } 8};3\right) = P\left(\boxed{6};3\right) + P\left(\boxed{8};3\right)$$

$$= P\left(3;\boxed{6}\right) \times P\left(\boxed{6}\right) + P\left(3;\boxed{8}\right) \times P\left(\boxed{8}\right)$$

$$= \left(\tfrac{1}{6} \times \tfrac{1}{2}\right) + \left(\tfrac{1}{8} \times \tfrac{1}{2}\right)$$

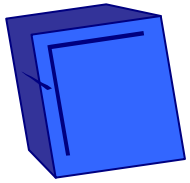
$$= 0.1458\overline{3}$$

$$P\left(\boxed{4 \text{ or } 10};3\right) = P\left(\boxed{4};3\right) + P\left(\boxed{10};3\right)$$

$$= P\left(3;\boxed{4}\right) \times P\left(\boxed{4}\right) + P\left(3;\boxed{10}\right) \times P\left(\boxed{10}\right)$$

$$= \left(\tfrac{1}{4} \times \tfrac{1}{2}\right) + \left(\tfrac{1}{10} \times \tfrac{1}{2}\right)$$

$$= 0.175 \quad \textit{most likely}$$



Integrating out nuisance variables

Discrete and continuous probabilities

- Probability for discrete probabilities

$$P(\text{data};\text{model}) = \sum_{i=1}^{n} P(\text{data},x_i;\text{model}), \quad \text{where } a \le x_i \le b$$

- For continuous probability, sum becomes an integral

$$P\left(\text{data};\text{model}\right) = \int_a^b P\left(\text{data},x;\text{model}\right) dx$$

- The unknown variable is called a nuisance variable
- The removal of a nuisance variable from a probability distribution by integration is called integrating out the nuisance variable
- Nuisance variables can be very useful!

Integrating out nuisance variables

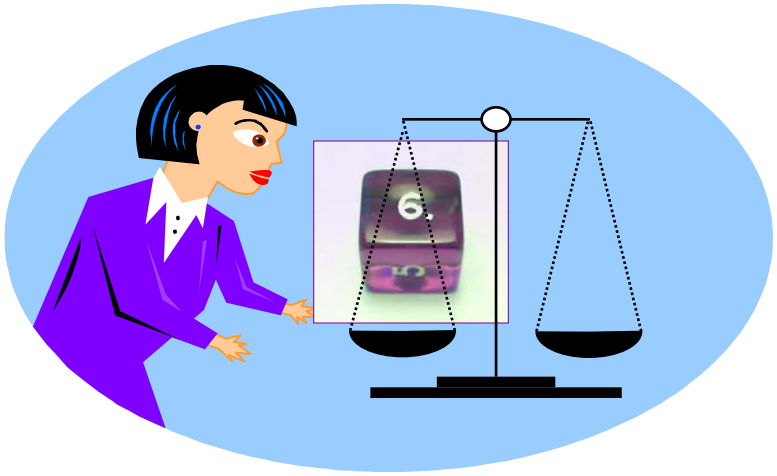
Crystallography

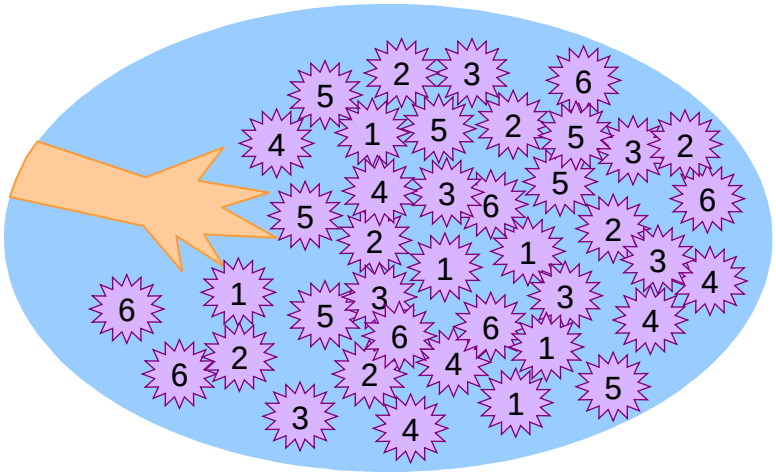
- Data (for each reflection) is the observed structure factor amplitude $|\mathbf{F}_O|$
- Model is the calculated structure factor $\mathbf{F}_C$
- Clever bit: Probabilities are calculated in terms of the phased observed structure factor $\mathbf{F}_O$ (and the calculated structure factor $\mathbf{F}_C$)
- The introduced *phase difference* is a nuisance variable
- Probability of $|\mathbf{F}_O|$ is then found by integrating out the "nuisance" (but very useful) phase

Central limit theorem

Central limit theorem

The average of many games of dice



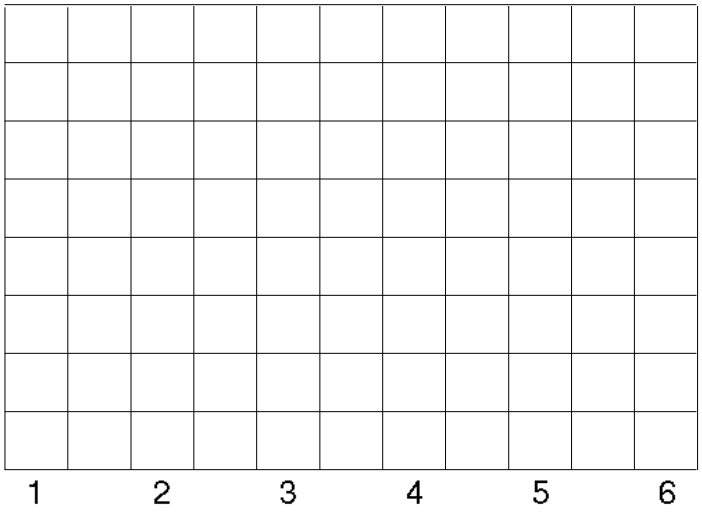


- I have an unbiased 6-sided die
- I roll the die 40 times and take the average of the values
- I do this 10000 times, plotting the average values from each game on a histogram

The average of many games of dice

Build up a histogram of the average 40 throws of the dice

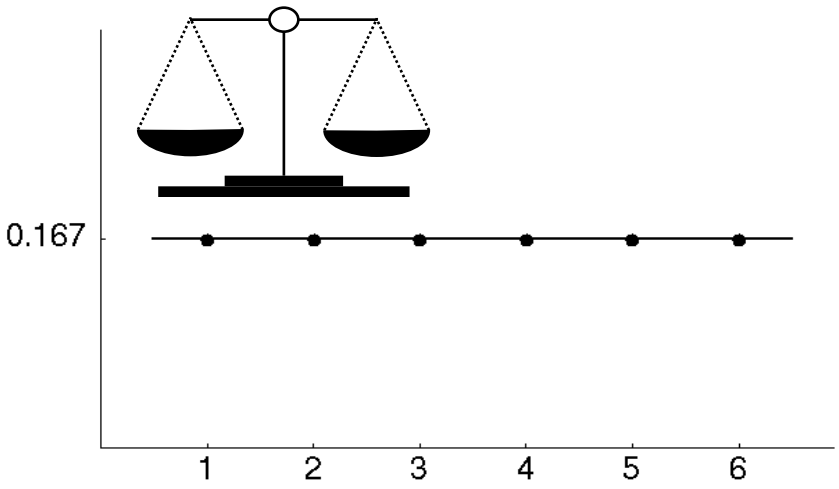
1. $40 \times \boxed{3}$
2. $20 \times \boxed{3} + 20 \times \boxed{4}$
3. $10 \times \boxed{1} + 10 \times \boxed{2} + 10 \times \boxed{5} + 10 \times \boxed{6}$
4. $40 \times \boxed{4}$
5. $10 \times \boxed{1} + 10 \times \boxed{2} + 10 \times \boxed{3} + 10 \times \boxed{4}$
6. $10 \times \boxed{2} + 10 \times \boxed{3} + 10 \times \boxed{4} + 10 \times \boxed{5}$
7. $20 \times \boxed{4} + 20 \times \boxed{5}$
8. $10 \times \boxed{3} + 10 \times \boxed{5} + 20 \times \boxed{4}$
9. $10 \times \boxed{1} + 10 \times \boxed{5} + 20 \times \boxed{3}$
10. $20 \times \boxed{2} + 20 \times \boxed{5}$



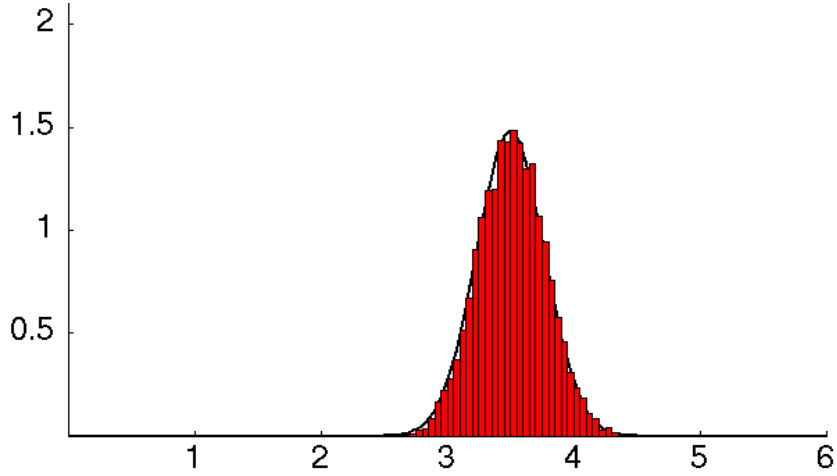
Histogram

The average of many games of dice

- The histogram is Gaussian (bell-shaped curve) with a mean at 3.5



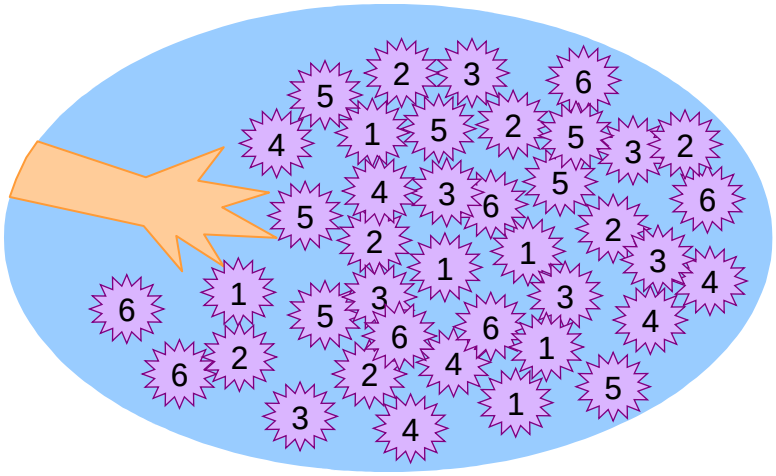
Bias of die



Histogram

Central limit theorem

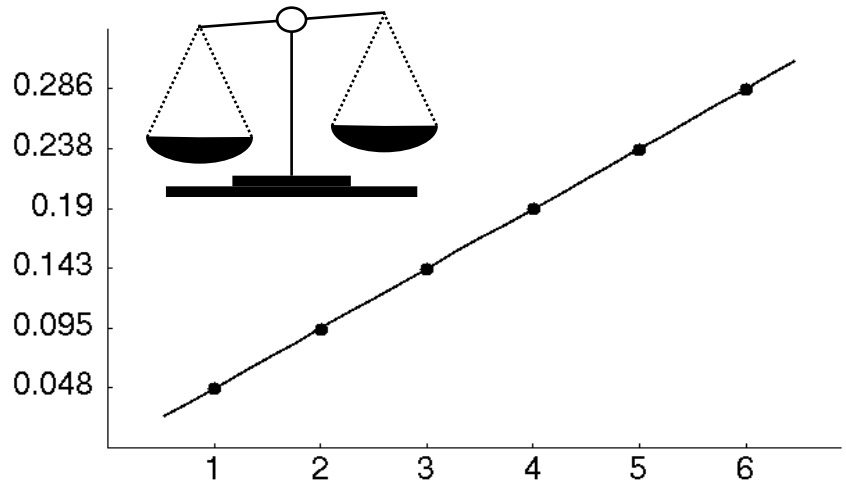
Linearly biased die



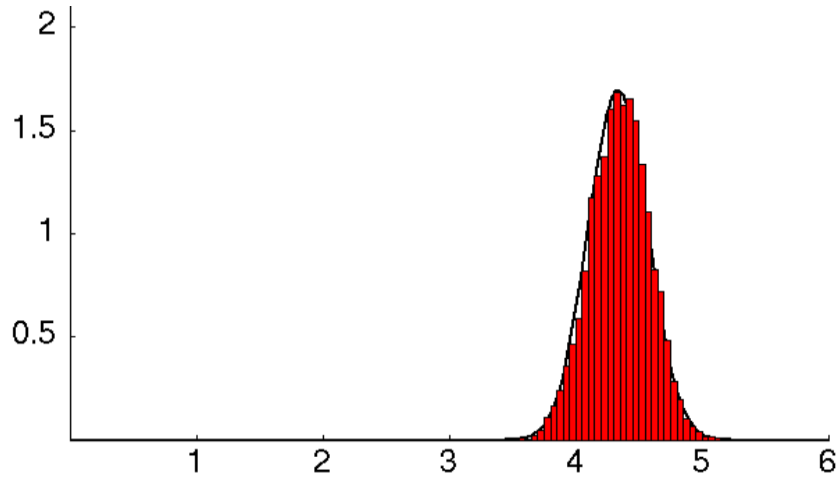
- I have an “linearly” biased 6-sided die
- I roll the die 40 times and take the average of the values
- I do this 10000 times, plotting the average values from each game on a histogram

Linearly biased die

- The histogram is Gaussian with a mean at 4.3, and the variance (width) of the distribution is smaller



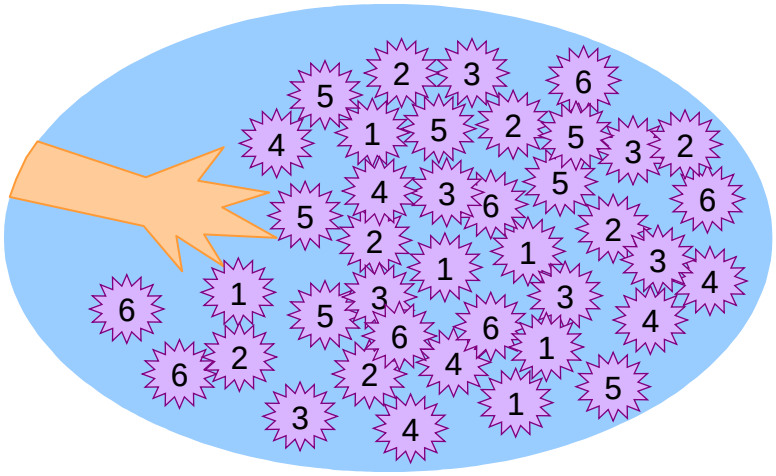
Bias of die



Histogram

Central limit theorem

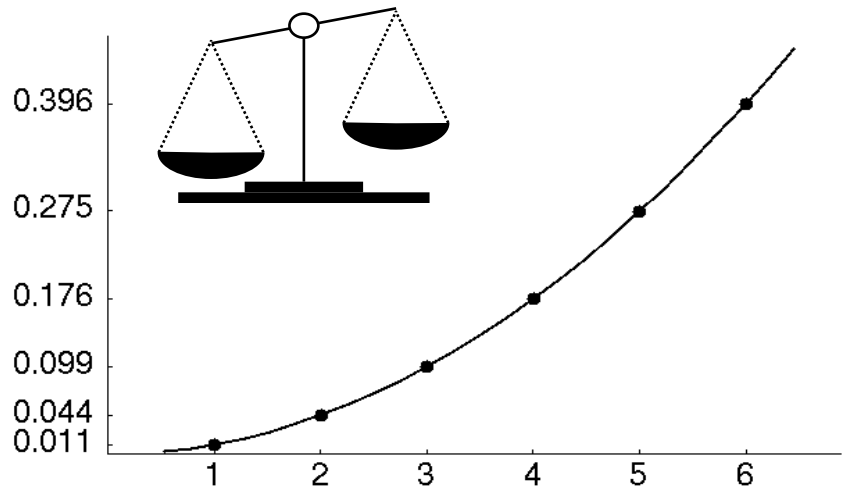
Quadratically biased die



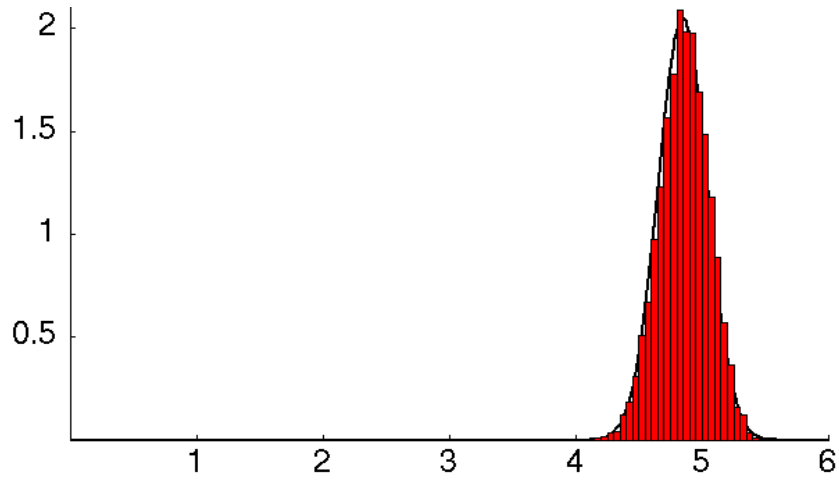
- I have an “quadratically” biased 6-sided die
- I roll the die 40 times and take the average of the values
- I do this 10000 times, plotting the average values from each game on a histogram

Quadratically biased die

- The histogram is Gaussian with an even higher mean and smaller variance (width)



Bias of die



Histogram

Central limit theorem

- No matter what the bias of the die, the histogram generated by the average of many rolls of the die is a Gaussian
- This is true even when the bias if the die (from which the average is computed) is decidedly <u>not</u> Gaussian
- This property is called the central limit theorem
- Historically, the central limit theorem was called the “law of errors”

Central limit theorem

Crystallography

- The atomic structure factor contributions to a given reflection $\mathbf{F}_C$ are very complicated
- However, the central limit theorem says that when you take the average of all these complicated structure factor contributions you get a Gaussian distribution
- This is lucky, because it is easy to integrate out the phase from a Gaussian distribution
- The result of the integration is the “Wilson” or “Rice” distributions, which are ubiquitous in maximum likelihood crystallography

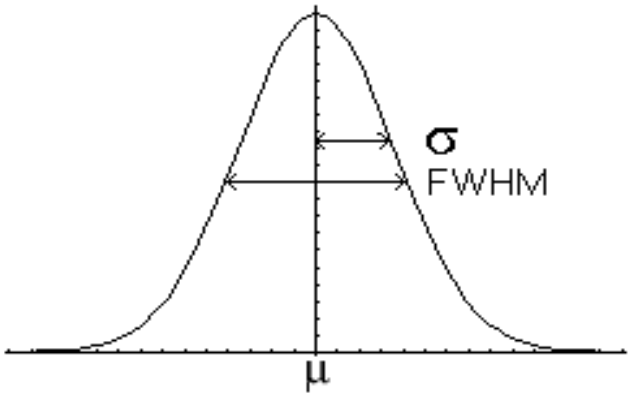
Central limit theorem

Gaussians and random walks

- 1D Gaussian

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

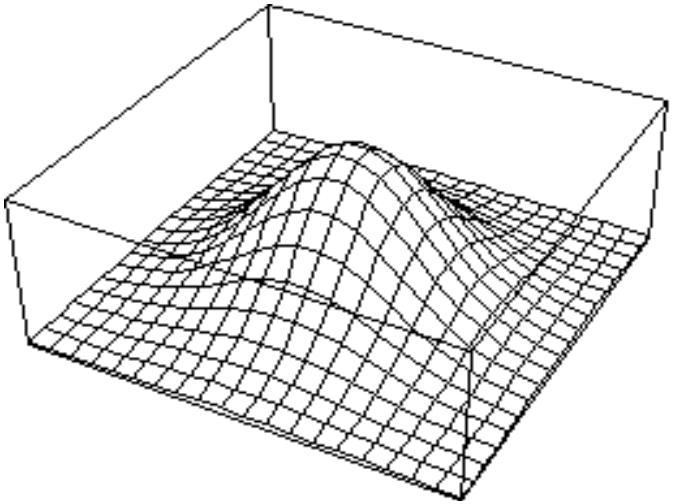
"1D random walk"

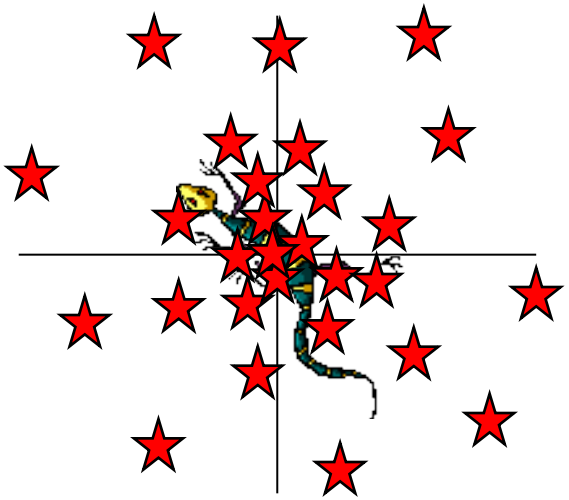


- 2D Gaussian

$$\frac{1}{2\pi\sigma^2} e^{-\frac{|\mathbf{x}-\boldsymbol{\mu}|^2}{2\sigma^2}}$$

"2D random walk"





Summary

- MAXIMUM LIKELIHOOD: the best model is the one that maximizes the probability of observing the data
- INDEPENDENCE: probabilities multiply when the experimental data points are independent
- LOG-LIKELIHOOD: used instead of the likelihood as it has a maximum at the same value as the likelihood but the numbers are not too small for computers to use
- BAYES' THEOREM: P(model;data) = prior $\times$ likelihood
- INTEGRATING OUT PARAMETERS: removes nuisance variables in a joint probability distribution
- CENTRAL LIMIT THEOREM: the distribution of the average is Gaussian, even when the distribution from which the average is drawn is not Gaussian

Maximum Likelihood

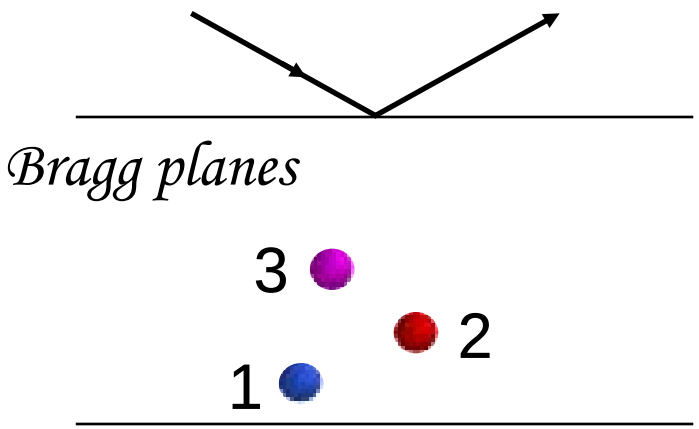
- Recap: What is the “best match” between the observed and calculated structure factors?
- Use **probability** as the scoring function
- Probabilities account for errors/uncertainties
  - Can model the errors/uncertainties in the positions of atoms and B-factors and occupancies
- Use the method of Maximum Likelihood to select the best model for the calculation of the phases

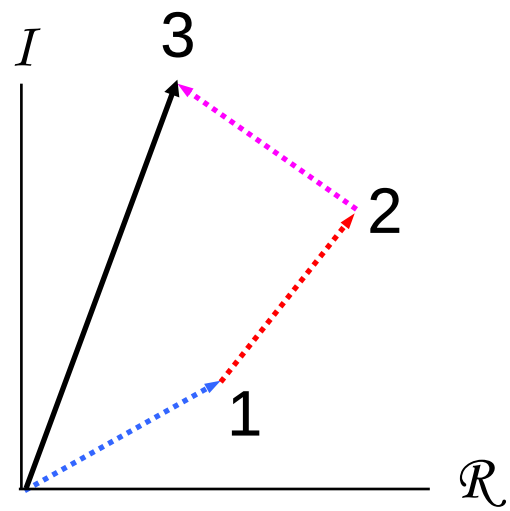
Likelihood Function

- The model consists of atoms with errors
  - The (phased) structure factors also have errors
- Total likelihood for a reflection is a 2D Gaussian
  - Because central limit theorem applies
- Integrate out observed phase from 2D Gaussian
  - Gives the likelihood for structure factor amplitude
  - “The phase problem”
- Assume reflections are independent
  - Total likelihood = Π_h likelihood
- Use log(likelihood)
- Total log(likelihood) = Σ_h log(likelihood)
- Minimise the -log(likelihood)

Three atoms

- The total structure factor of three atoms for a reflection is the sum of the structure factors
- The phase (0 to 360°) depends on the distance of the atoms between Bragg planes





Atoms have errors

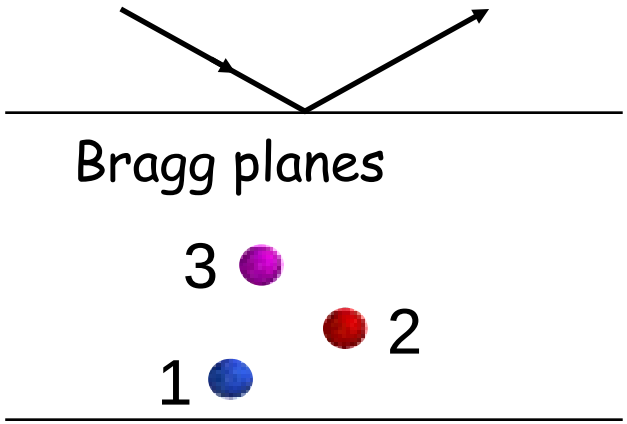
(or there are uncertainties)

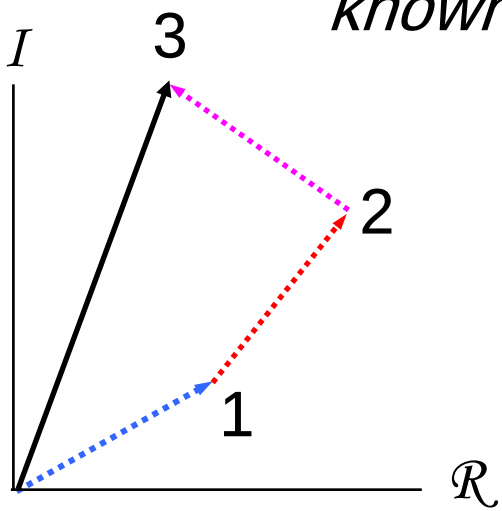
No errors

Structure factors known exactly

No errors

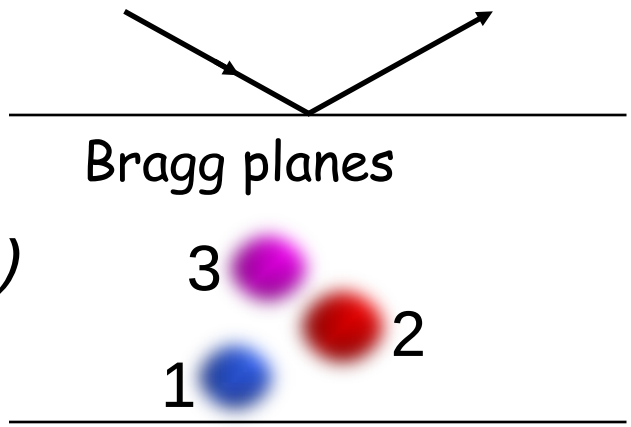
Atomic positions known exactly

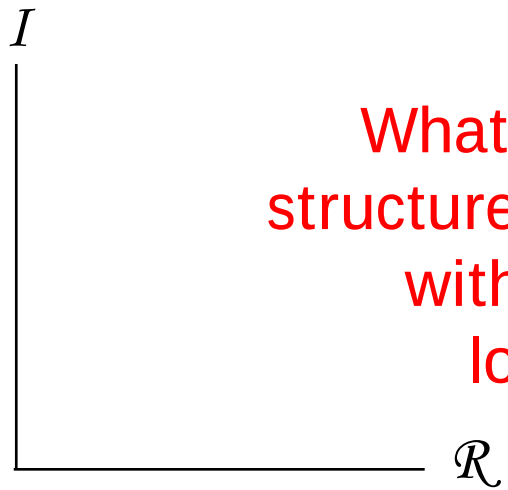




With errors

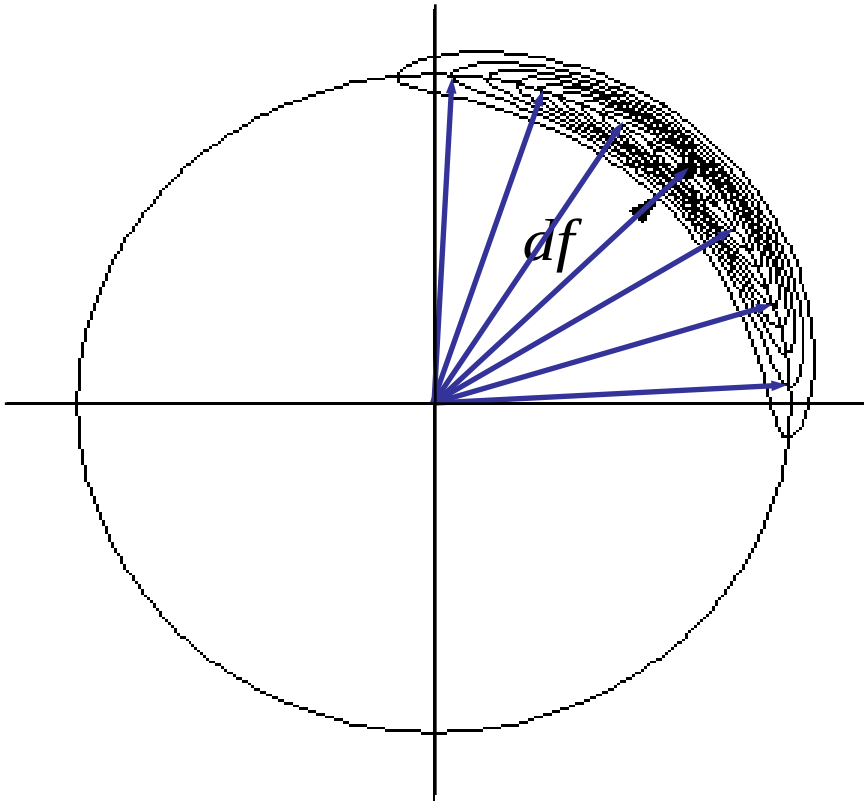
Atomic scattering (positions/B-factors) not known exactly

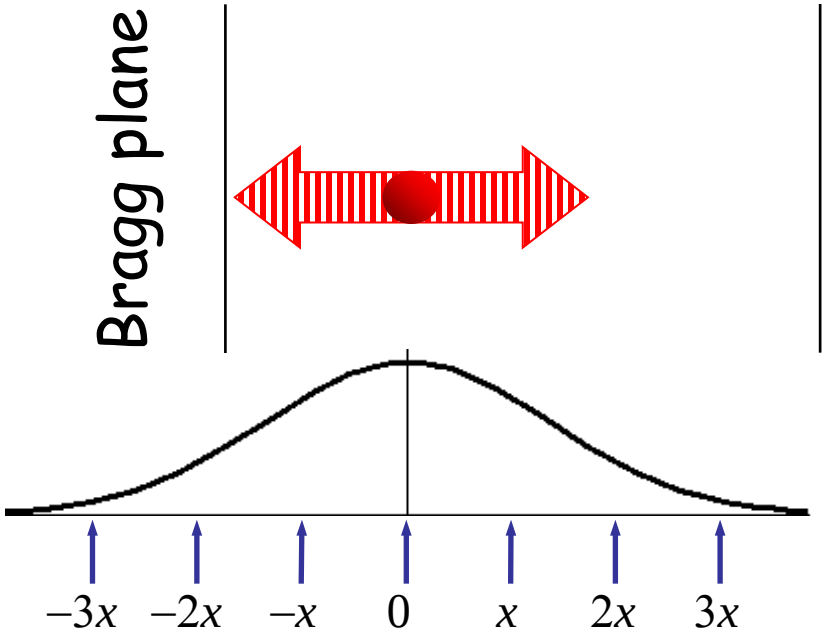




What do the structure factors with errors look like?

Structure factor for atom with errors





Only the error/uncertainty in the position perpendicular to the Bragg planes is relevant for any given structure factor

What do the structure factors with errors look like?

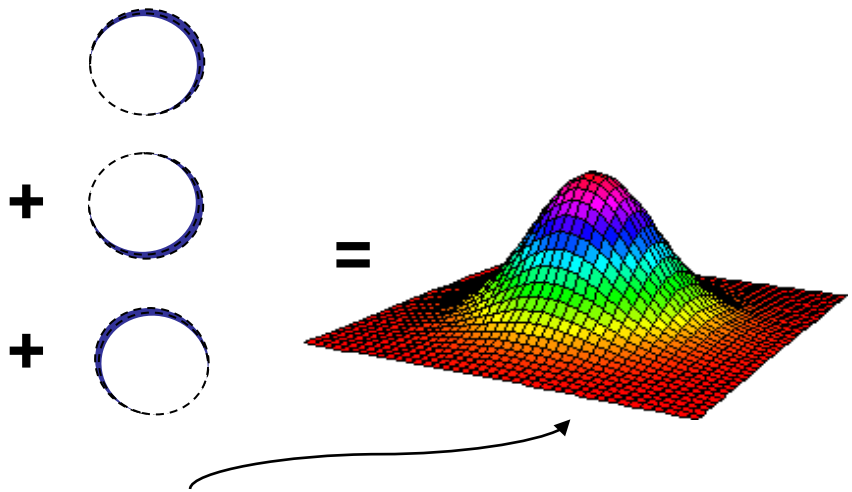
A boomerang!

Likelihood Function

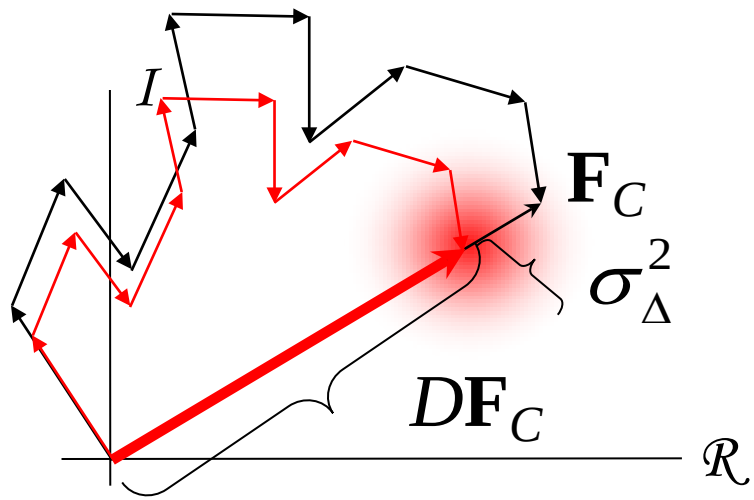
- The model consists of atoms with errors
  - The (phased) structure factors also have errors
- Total likelihood for a reflection is a 2D Gaussian
  - Because central limit theorem applies
- Integrate out observed phase from 2D Gaussian
  - Gives the likelihood for structure factor amplitude
  - “The phase problem”
- Assume reflections are independent
  - Total likelihood = Π_h likelihood
- Use log(likelihood)
- Total log(likelihood) = Σ_h log(likelihood)
- Minimise the -log(likelihood)

Structure factor for a model with errors

Boomerangs
independent errors
(approximation)



Sum of boomerangs
gives 2D Gaussian by
Central Limit Theorem



2D Gaussian centred on DFc
with width (variance) σ_Δ^2

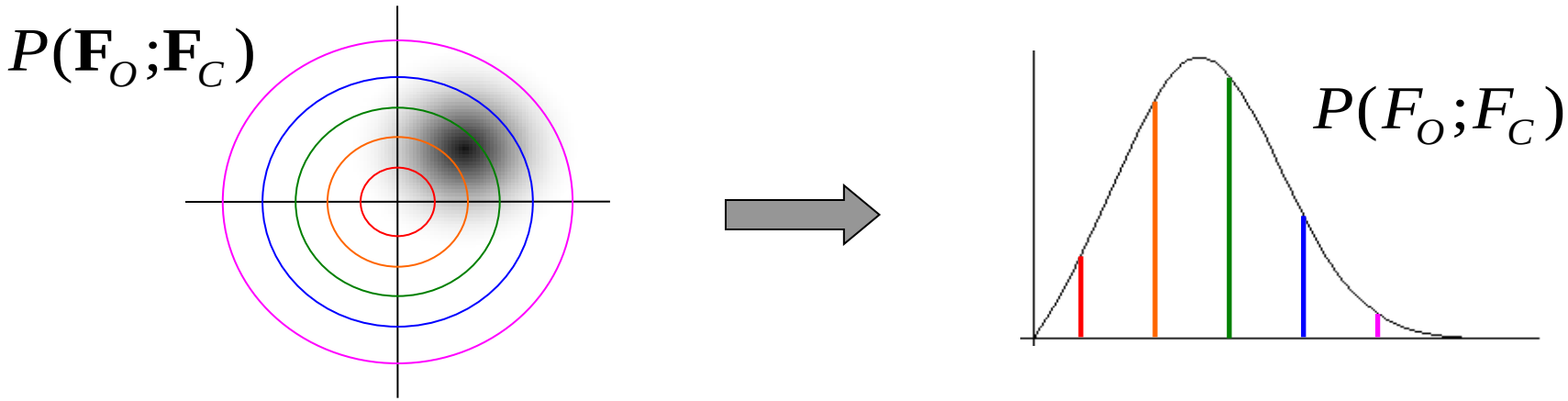
$$P(\mathbf{F}_O;\mathbf{F}_C) = \frac{1}{\pi\sigma_\Delta^2} e^{-\frac{|\mathbf{F}_O - D\mathbf{F}_C|^2}{\sigma_\Delta^2}}$$

Likelihood Function

- The model consists of atoms with errors
  - The (phased) structure factors also have errors
- Total likelihood for a reflection is a 2D Gaussian
  - Because central limit theorem applies
- Integrate out observed phase from 2D Gaussian
  - Gives the likelihood for structure factor amplitude
  - "The phase problem"
- Assume reflections are independent
  - Total likelihood = Π_h likelihood
- Use log(likelihood)
- Total log(likelihood) = Σ_h log(likelihood)
- Minimise the -log(likelihood)

Likelihood

- But we do not measure the phase of the observed structure factor!
- Integrate out the phase to get the likelihood for the unphased structure factors



$$P(F_O;F_C) = \frac{2F_O}{\sigma_\Delta^2} e^{-\frac{F_O{}^2 + D^2 F_C{}^2}{\sigma_\Delta^2}} I_0\left(\frac{2F_O D F_C}{\sigma_\Delta^2}\right)$$

Likelihood Function

- The model consists of atoms with errors
  - The (phased) structure factors also have errors
- Total likelihood for a reflection is a 2D Gaussian
  - Because central limit theorem applies
- Integrate out observed phase from 2D Gaussian
  - Gives the likelihood for structure factor amplitude
  - “The phase problem”
- Assume reflections are independent
  - Total likelihood = Π_h likelihood
- Use log(likelihood)
- Total log(likelihood) = Σ_h log(likelihood)
- Minimise the -log(likelihood)

Likelihood Function

Optimize over models

Log-likelihood

Likelihood per reflection
P(data;model)

$$\min\left[-\sum_{hkl}\log\left(\frac{2F_{obs}}{\sigma_\Delta^2}e^{\frac{F_{obs}^2+D^2F_{calc}^2}{\sigma_\Delta^2}}I_O\left(\frac{2F_{obs}DF_{calc}}{\sigma_\Delta^2}\right)\right)\right]$$

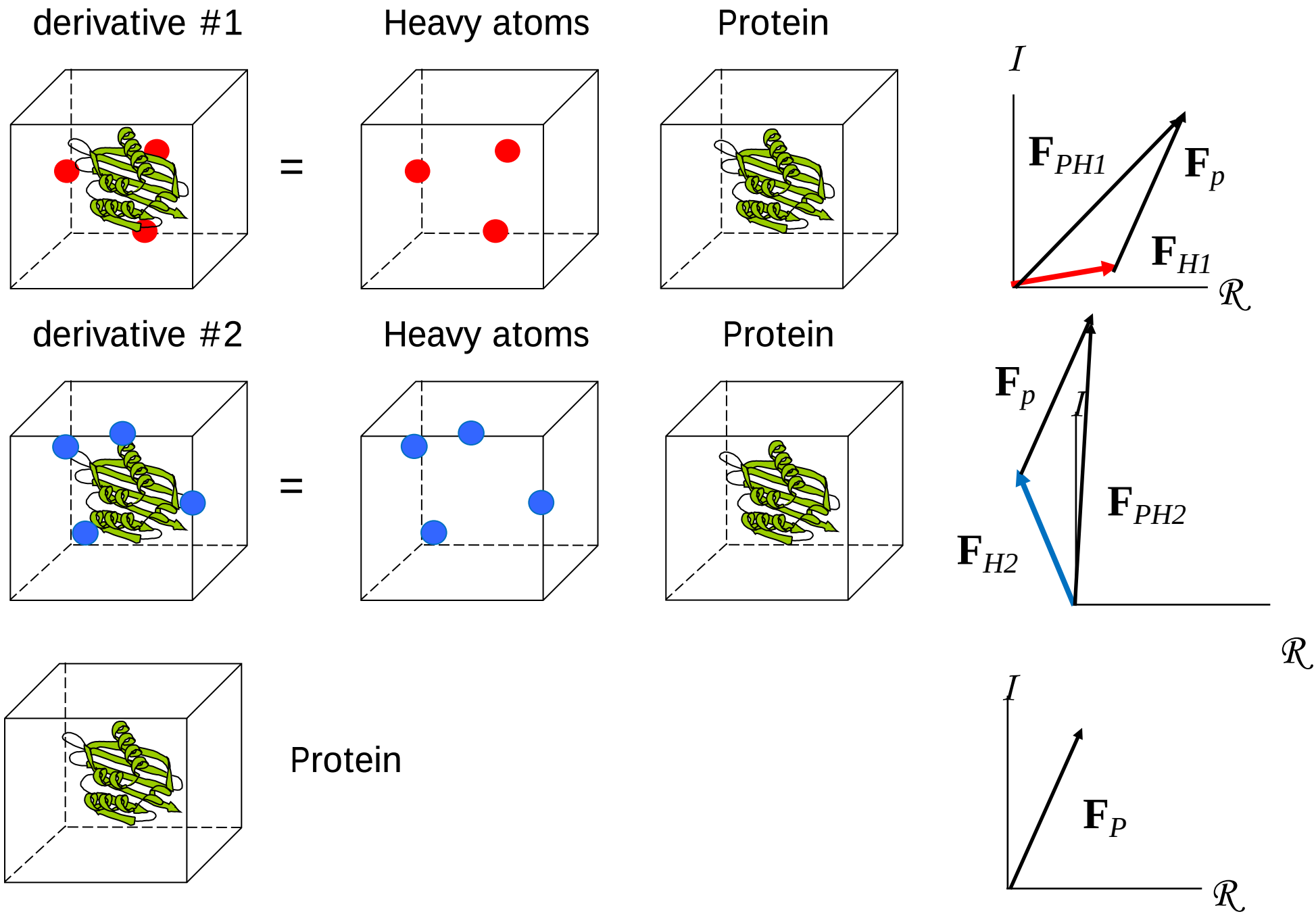
Independence of the reflections

Data

Model

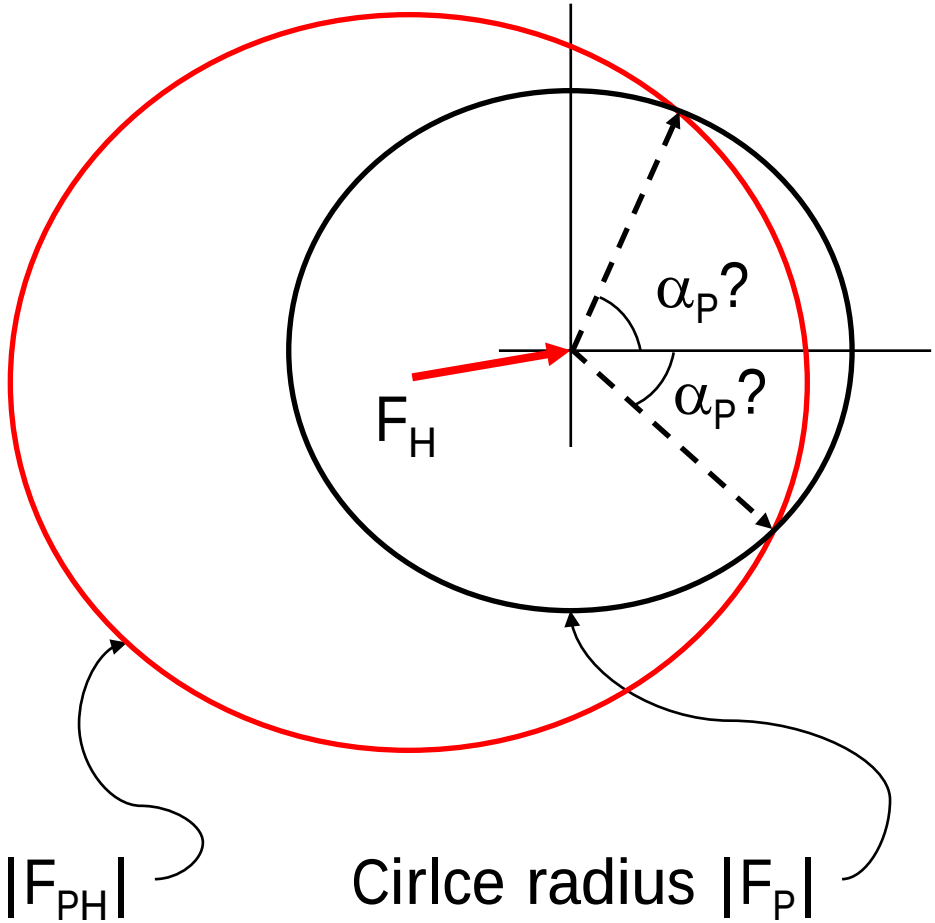
Q.E.D.

Experimental Phasing *with* Maximum Likelihood


derivative #1
Heavy atoms
Protein
=
I
$\mathbf{F}_{PH1}$
$\mathbf{F}_p$
$\mathbf{F}_{H1}$
$\mathcal{R}$
derivative #2
Heavy atoms
Protein
=
$\mathbf{F}_p$
I
$\mathbf{F}_{PH2}$
$\mathbf{F}_{H2}$
$\mathcal{R}$
Protein
I
$\mathbf{F}_P$
$\mathcal{R}$

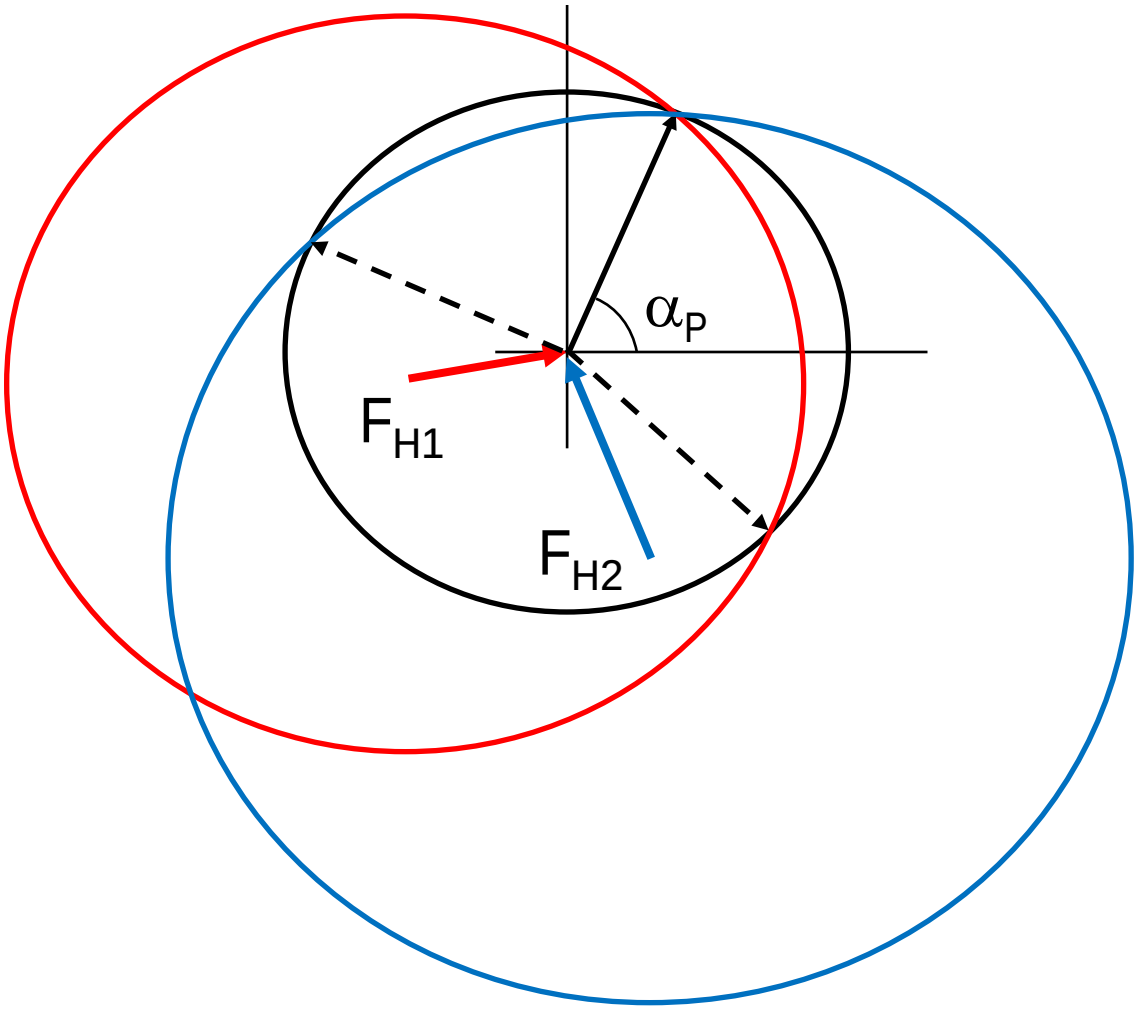
Harker Diagram

- The interference experiment can be represented on a Harker diagram
- The phase ambiguity for the native protein structure factor is shown as the intersection of two circles with centres displaced by F_H



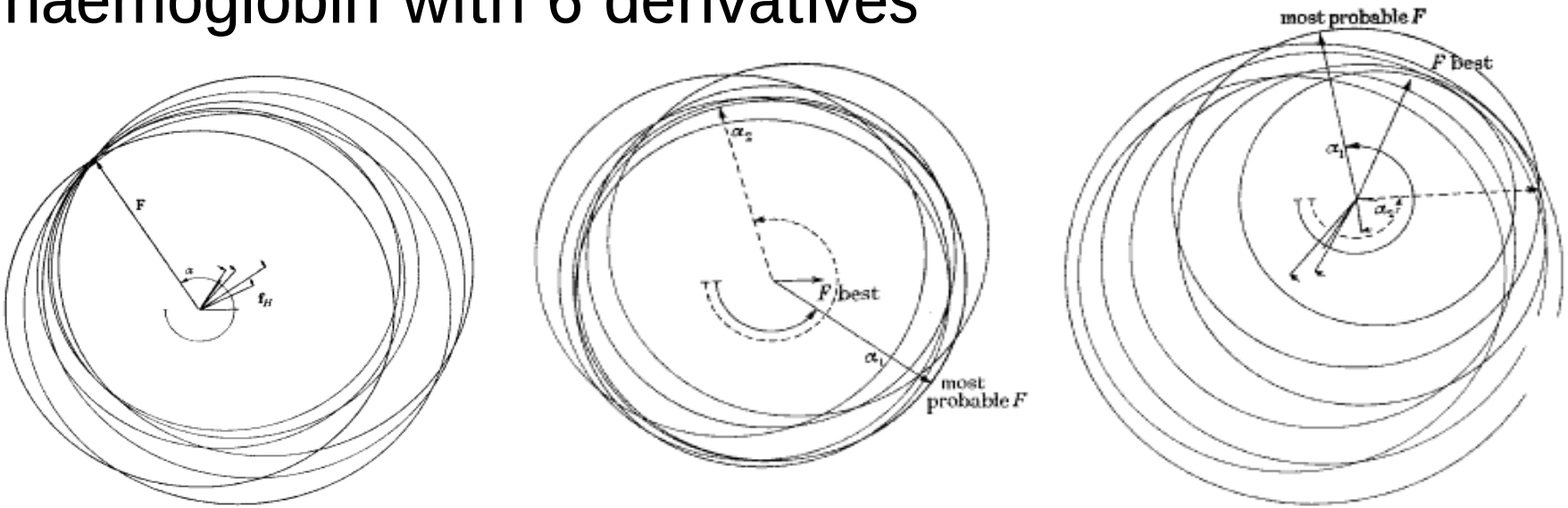
Harker Diagram

- The interference experiment for a second derivative can be shown on the same Harker diagram
- A second derivative breaks the phase ambiguity



Reality...

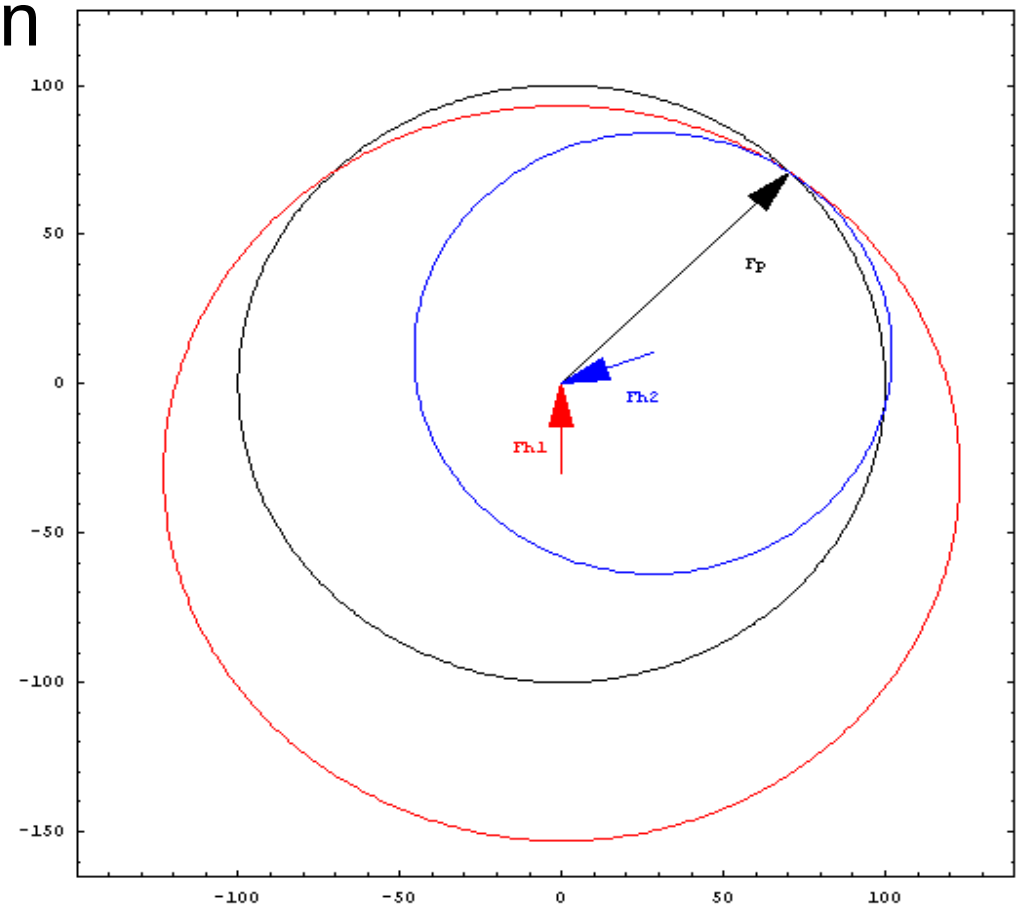
- Some real Harker diagrams from the phasing of haemoglobin with 6 derivatives



- Phase circles <u>rarely</u> cross exactly
- Need a **probabilistic** approach to determining the phase

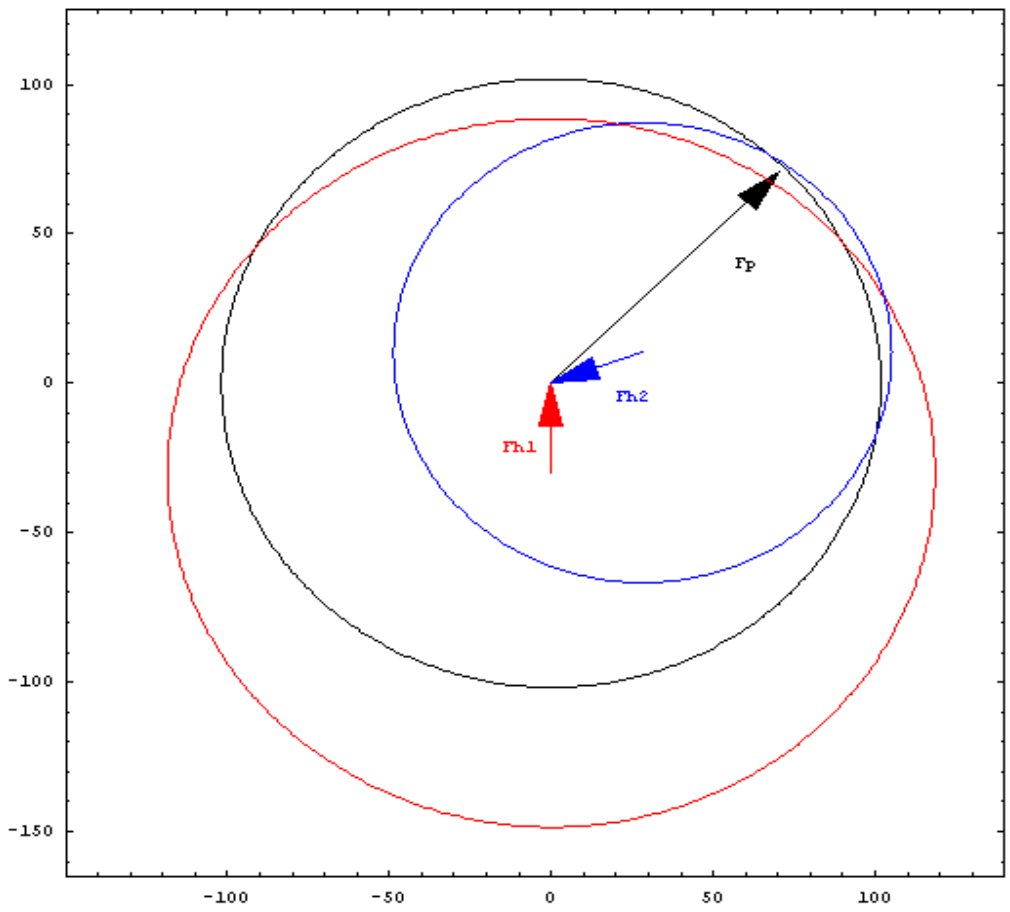
Harker construction

- Phasing of one refection using two derivatives with no errors
- Phase determined with very high probability



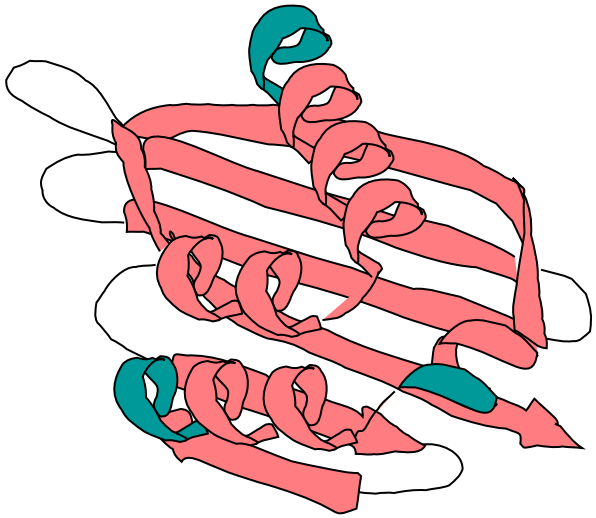
Harker construction

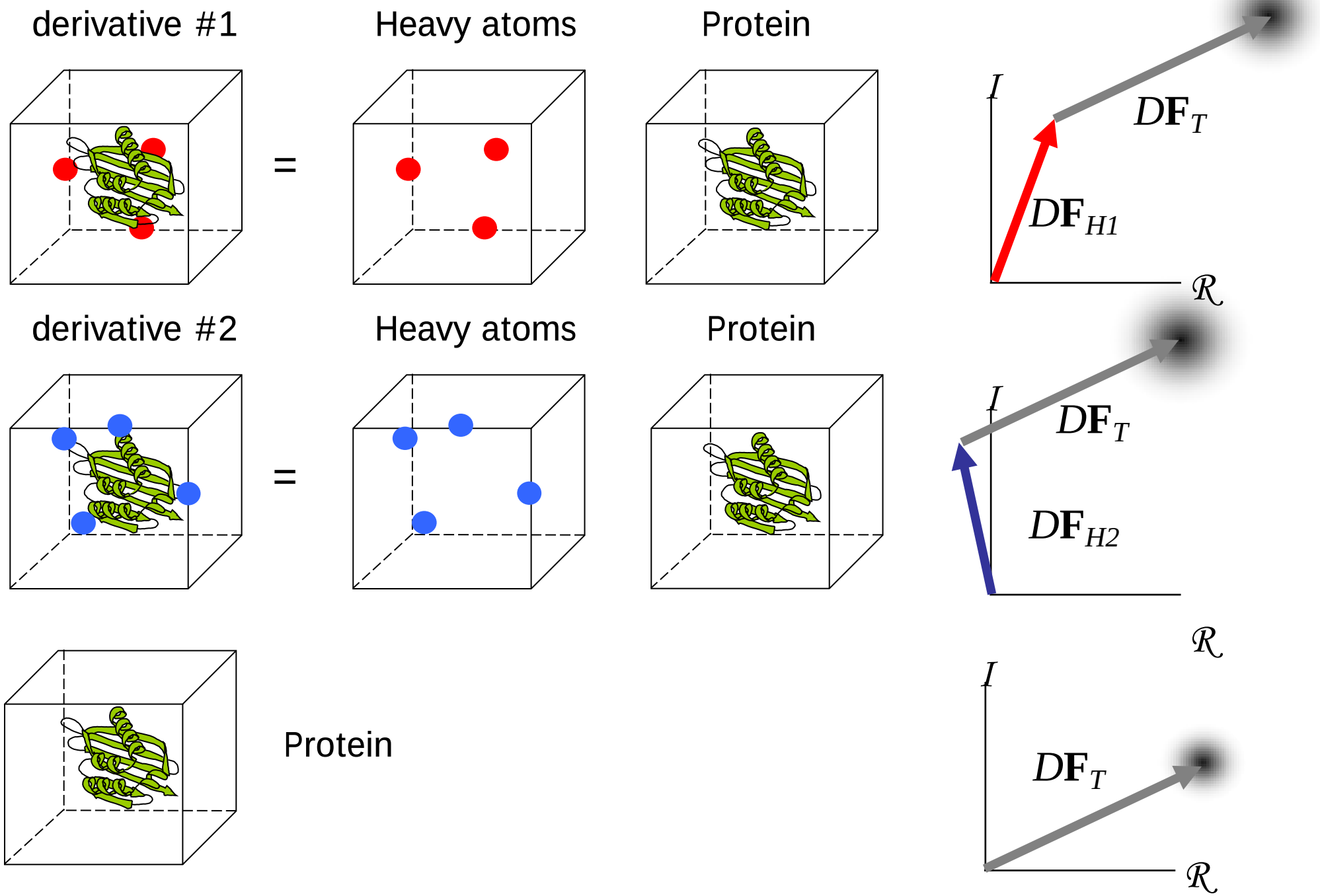
- There are many sources of error
  - Data errors
  - Model errors
- The errors are large
- We are looking for the <u>best</u> phase
- We therefore need a probability function



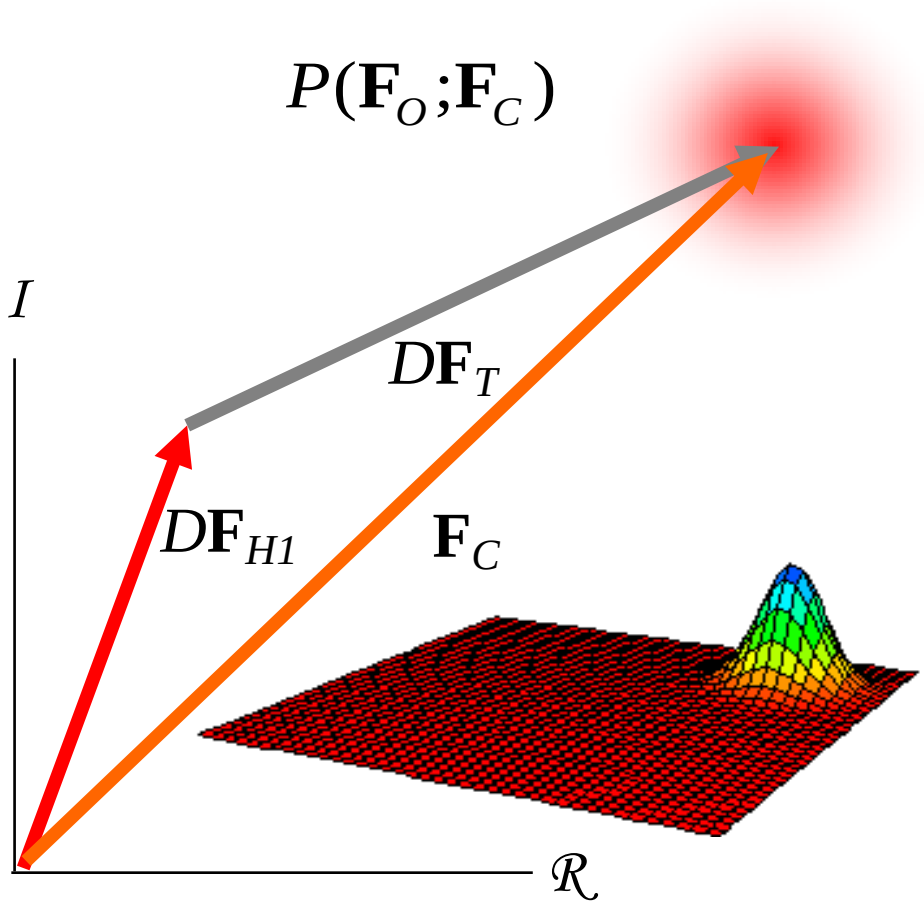
Introducing.... the True F

- Introduce a nuisance parameter, the “true F”
- The “true F” is the component of scattering shared by the native and derivatives
  - The “left over” parts of the structure factors are independent
- The “true F” is integrated out at the end of the analysis
  - Numerical integration




derivative #1
Heavy atoms
Protein
=
I
DF_T
DF_H1
ℛ
derivative #2
Heavy atoms
Protein
=
I
DF_T
DF_H2
ℛ
Protein
I
DF_T
ℛ

2D Gaussian probability function



- The 2D Gaussian is the probability of measuring an $\mathbf{F}_O$ given $\mathbf{F}_C$ = D$\mathbf{F}_H$ + D$\mathbf{F}_T$
- Probability accounts for
  - Errors in $\mathbf{F}_{H,}$
  - Non-isomorphism
  - Measurement errors in $\mathbf{F}_O$
- However, we measure $|\mathbf{F}_O|$ not $\mathbf{F}_O$
- Integrate out the phase to get a Rice function

Likelihood Function

Optimize over models

Log-likelihood

Integrate out F_{true}

Likelihood per derivative
P(data;model)

$$\min\left[\sum_{hkl} -\log\left[\iint_{\mathbf{F}_T}\prod_{j}\left(\frac{2F_{Oj}}{\sigma_{\Delta j}^2}e^{\frac{F_{Oj}^2+D^2F_{Cj}^2}{\sigma_{\Delta j}^2}}I_O\left(\frac{2F_{Oj}DF_{Cj}}{\sigma_{\Delta j}^2}\right)\right)d\mathbf{F}_T\right]\right]$$

Data

Model

Independence of the reflections

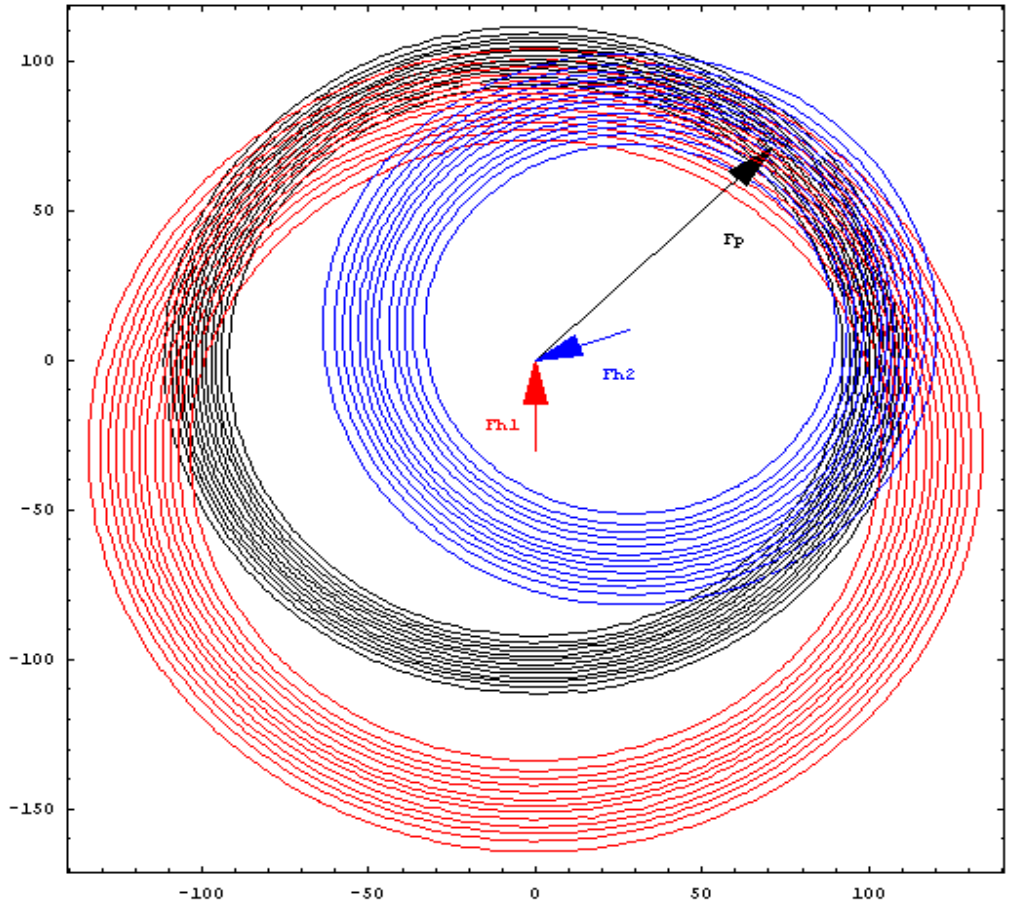
Independence of the derivatives

Integrate out F_{true}

Q.E.D.

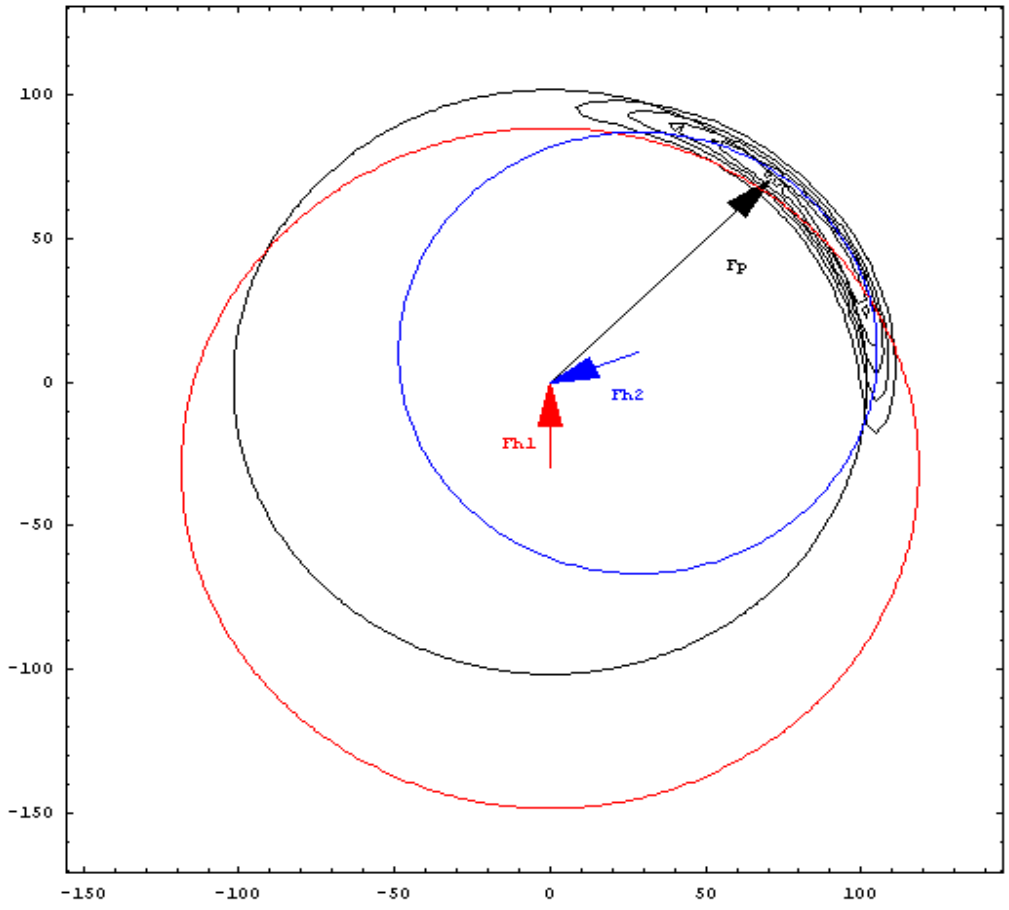
Probabilistic Harker Diagram

- Each circle has an error associated with it to give a distribution
- The total likelihood is the volume under the curve of the product of the distributions



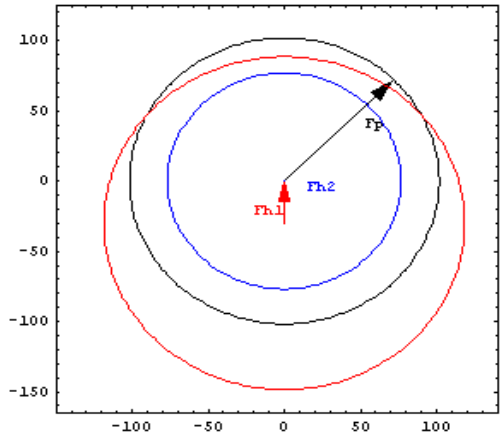
Probabilistic Harker Diagram

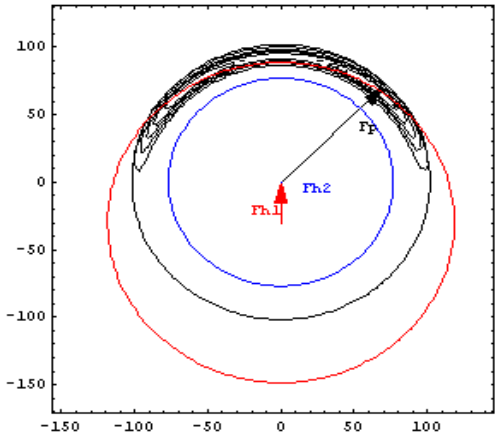
- The final distribution is high only where all three circles overlap



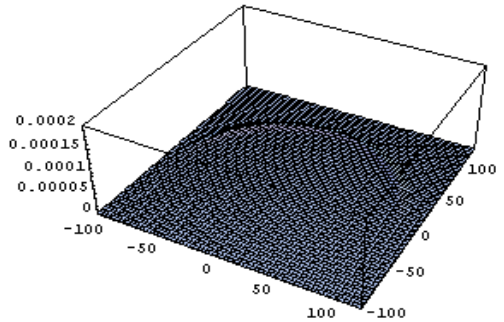
Refining Occupancy

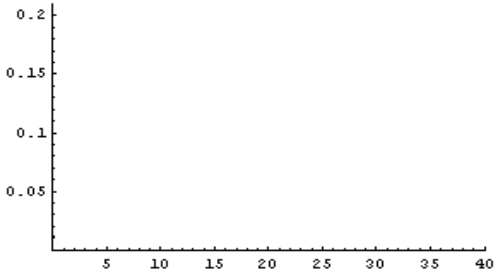
To refine the occupancy of a heavy atom, maximise the likelihood (area under the curve)





Final refined value is the optimum for ALL reflections (movie shows ONE reflection)

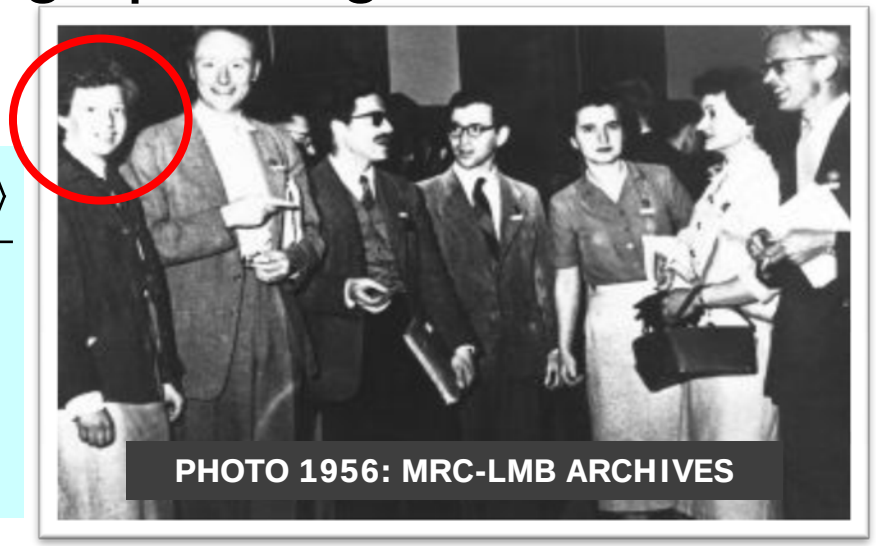




Other Phasing Statistics

- They are non-likelihood measures of phasing
- Heuristic formula that help judge phasing

Ann Cullis



$$R_{Cullis} = \frac{\langle \text{prob. weighted lack of closure} \rangle}{\langle \text{isomorphous difference} \rangle}$$

< 0.6 excellent

< 0.9 usable

$$\text{Isomorphous Phasing Power} = \frac{\langle \text{heavy atom amplitude} \rangle}{\langle \text{prob. weighted lack of closure} \rangle}$$

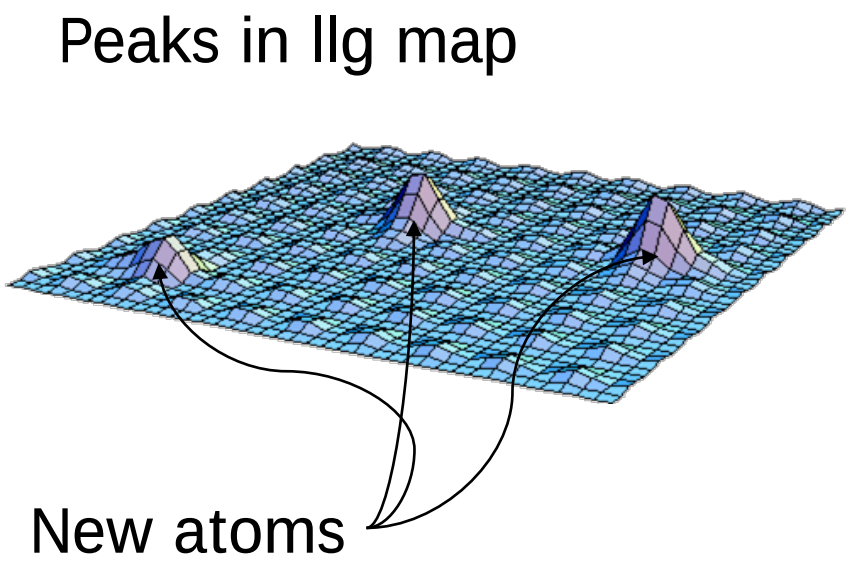
$$\text{Anomalous Phasing Power} = \frac{\langle \text{anomalous amplitude} \rangle}{\langle \text{prob. weighted lack of closure} \rangle}$$

> 1.5 excellent
> 1.0 good
> 0.5 usable

Completion of sub-structure

- Inclusion of minor sites improves the phases
  - Want the biggest “substructure” that can be found
- Minor sites often not detectable from Patterson
  - For example, anomalously scattering intrinsic sulphurs
- Compute derivative of log-likelihood with respect to heavy atom structure factor
  - FT gives a map – the “log-likelihood gradient map”
  - shows where likelihood function would **like** to see changes in anomalous scatterer model
  - But can’t do anything about it because there are no atoms there to change occupancy/B-factor of…

Log-likelihood gradient maps

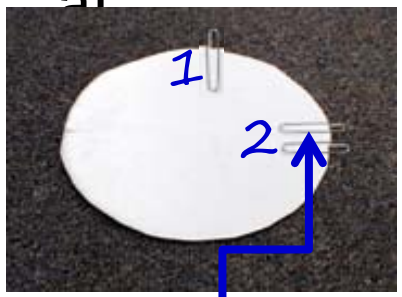


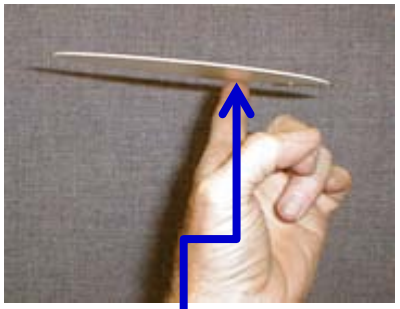
- Gradient of likelihood function w.r.t. coordinates xyz indicates where the function wants new atoms added
- Substructure completion is done by adding new atoms at these locations and then re-refining
  - **Repeat until convergence**
- Can also be used to confirm/identify anomalously scattering atoms

Calculating Electron Density

- ML function is good for refining the parameters, but what phase should be used in the electron density equation?
  - Have to pick <u>one</u> phase
- **We want the phase that gives the electron density with the lowest rms error**
  - Parseval's theorem relates the rms error in real space to the rms error in reciprocal space and vice versa
- This phase (the "**best phase**") is the probability-weighted average of all the phases
  - It is <u>not</u> the "most probable phase"

- Cut the centre out of a polystyrene foam plate
- Balance the disk on your finger
  - The centre of mass is the centre
- Now put 3 paperclips on edge of the disc
  - 2 together
  - 1 a distance away
- The balancing point is **between** the 3 paperclips
  - Not on the 2 paperclips

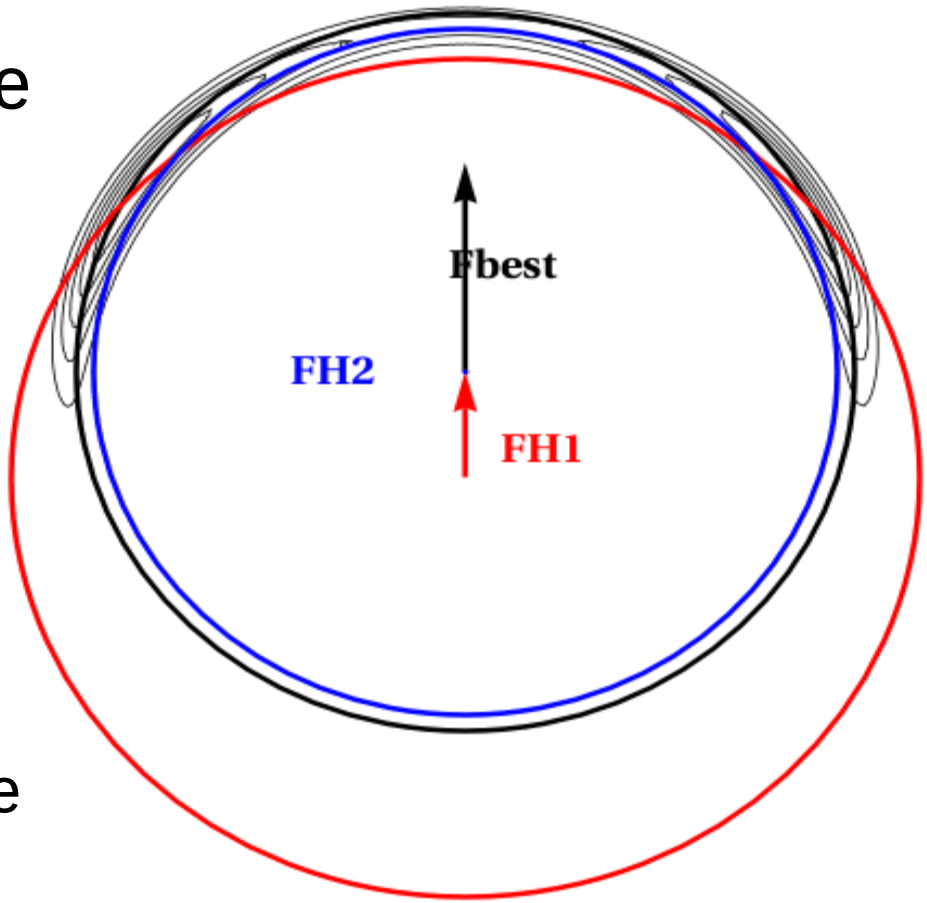




Most Probable Structure Factor

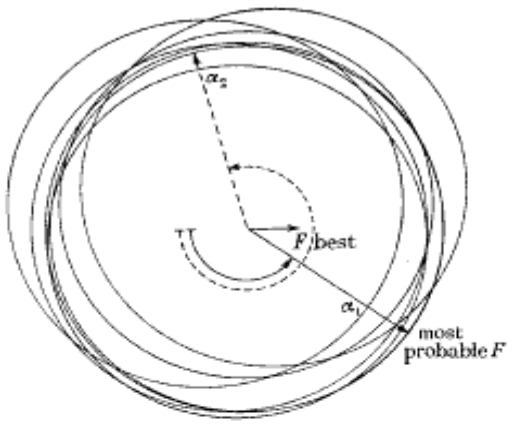
Best Structure Factor

- F_{best} has a lower |F| amplitude than F_{obs}
- The reduction in F_{obs} to give F_{best} is expressed as the "figure of merit" (m)
  - **$0 < m < 1$:** F_{best} lies inside the F_{obs} circle
  - **$m = 1$:** Perfect phase information
  - **$m = 0$:** No phase information
  - The higher the average value of the figure of merit, the better

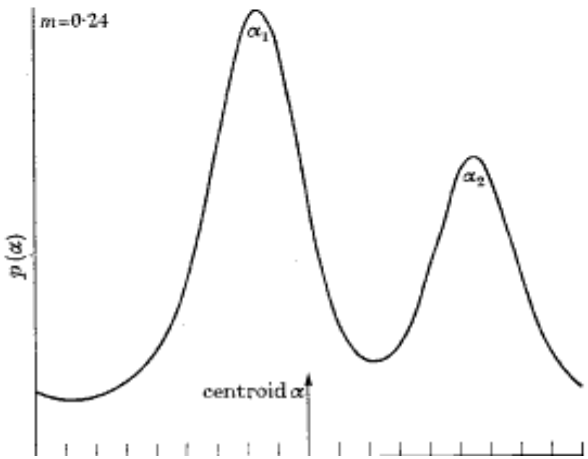


Phase probability

- Each reflection really has a phase probability density function (PDF) rather than a single phase
- This is a complicated mathematical function
  - Requires lots of memory
- Four Hendrickson-Lattman coefficients (A,B,C,D) are used to store this PDF in a compact form

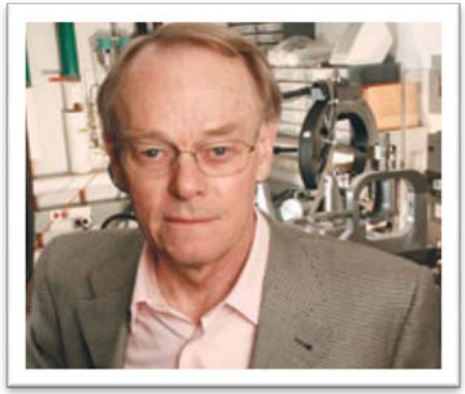


(b)



Hendrickson-Lattman Coefficients

- Each reflection has a (different) structure factor PDF
- There is one set of (four) HL coefficients for each reflection
  - They are different for each reflection
- These four parameters generate a curve that **approximates** the PDF for each reflection



Wayne Hendrickson

Eaton Lattman

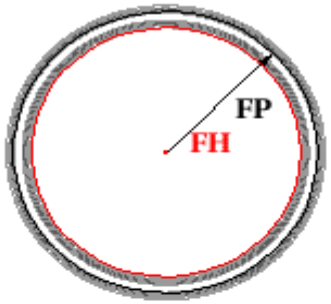
Hendrickson-Lattman Coefficients

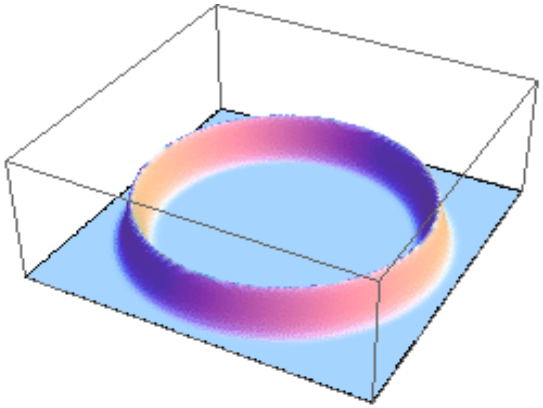
$$P(\alpha) \propto \exp\left[A\cos(\alpha) + B\sin(\alpha) + C\cos(2\alpha) + D\sin(2\alpha)\right]$$

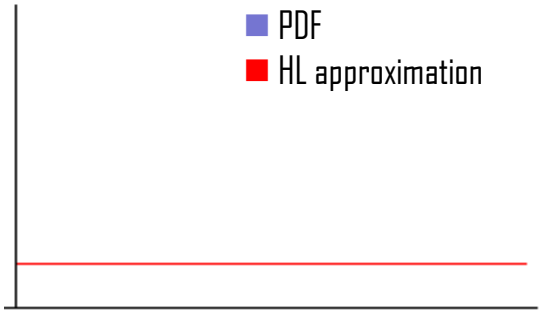
- Hendrickson Lattman (HL) coefficients can only (completely) describe a bi-modal distribution
  - Since the highest frequency is 2α
  - Most PDFs do not have more than two peaks
- HL coefficients allow for easy combination of phase information from multiple sources
  - the combined PDF is formed simply by adding the A,B,C, and D from the two distributions

Hendrickson-Lattman Coefficients

HL coefficients as a function of FH occupancy



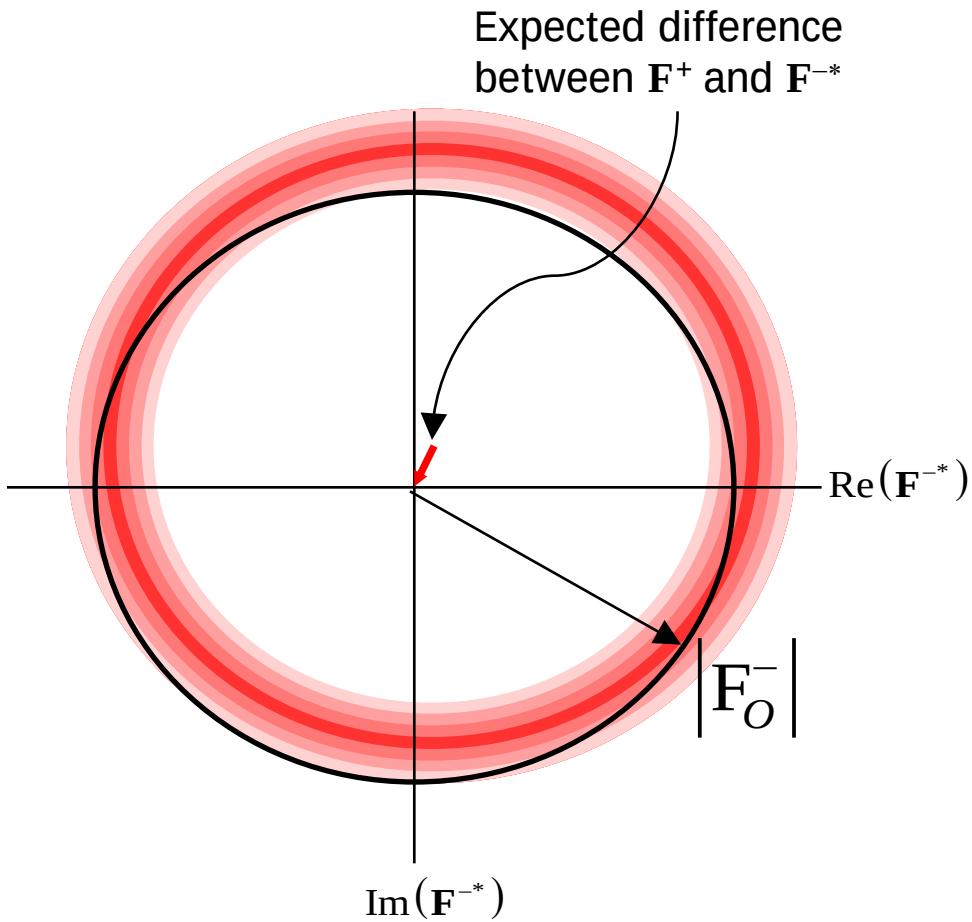




SAD Phasing

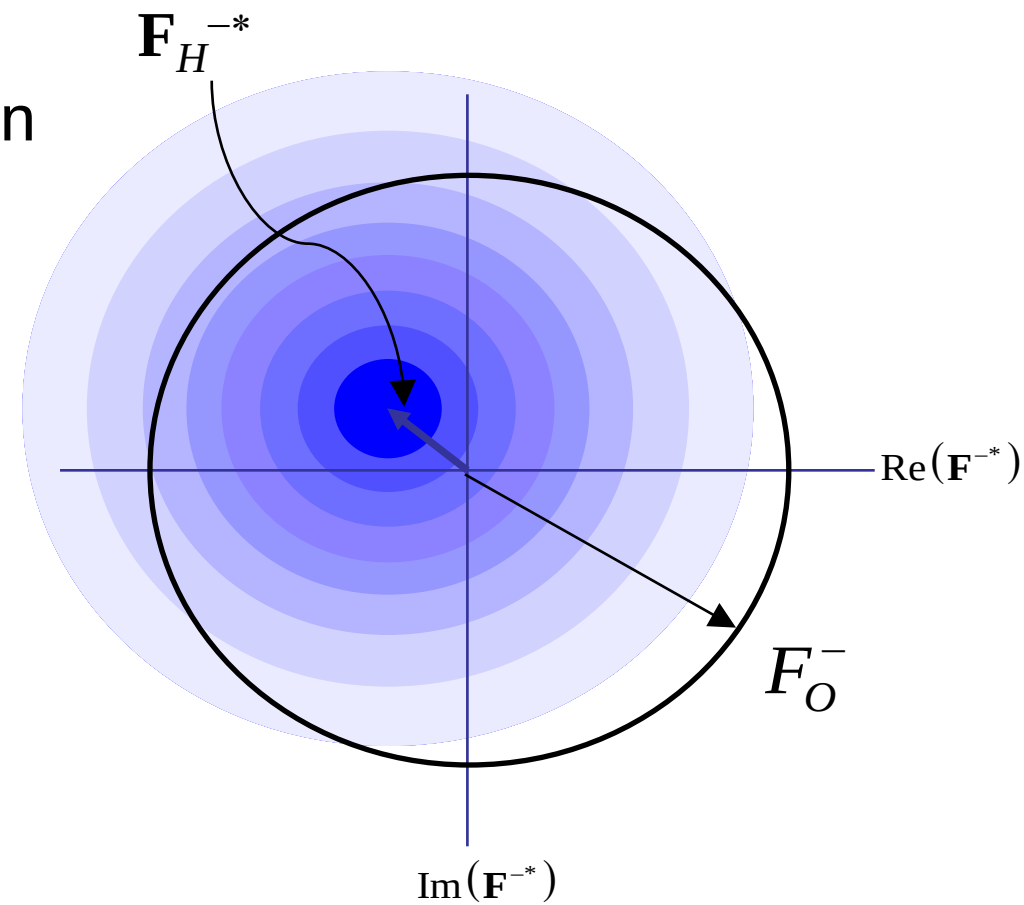
Rice term

- Primarily anomalous scattering
- “tight” probability distribution



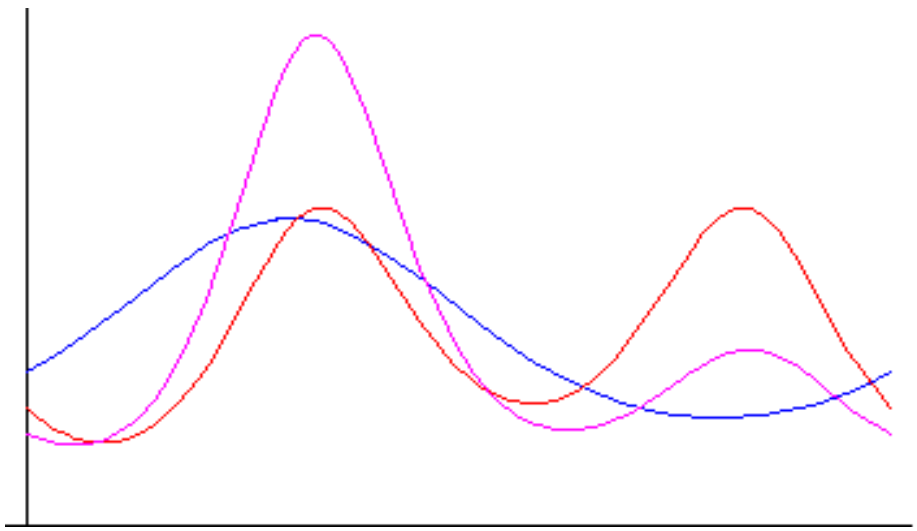
2D-Gaussian Term

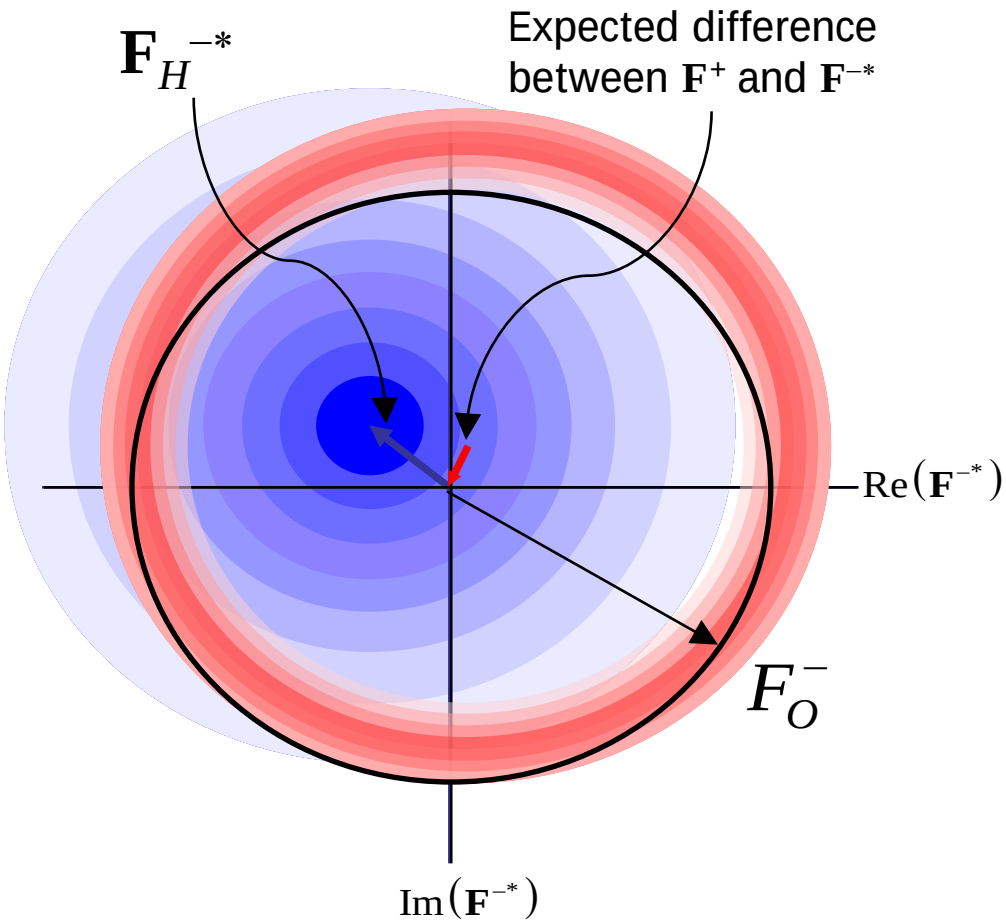
- Primarily Normal scattering
- “diffuse” probability distribution



Rice & 2D-Gaussian Product

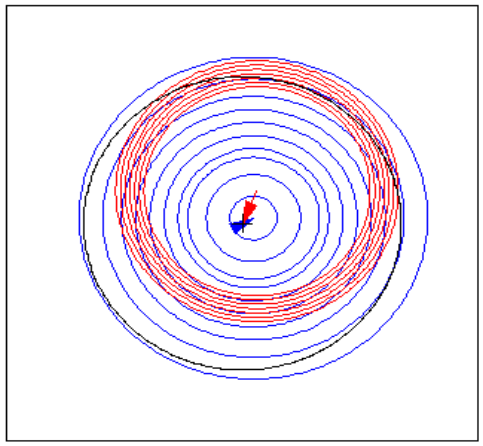
Likelihood is proportional to the product of the two distributions (magenta) under the black circle

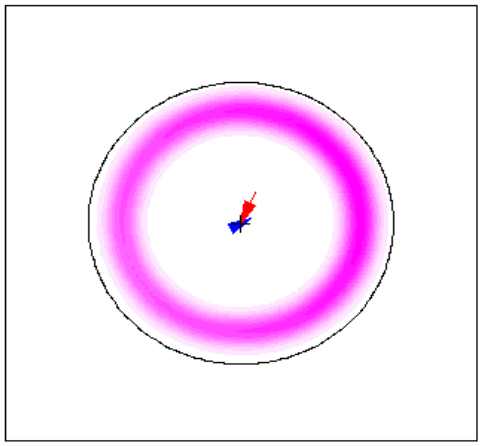




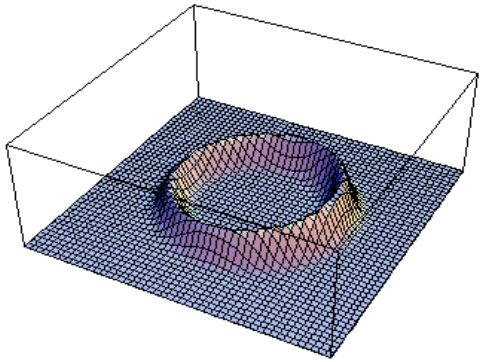
Refining heavy atom occupancy

To refine the occupancy of a heavy atom, maximise the SAD likelihood





Final refined value is the optimum for ALL reflections (movie shows ONE reflection)



SAD Phasing

DNA Hexamer

Software for Experimental Phasing

- SHARP
  - Maximum likelihood phasing
  - SAD/SIR/MIR/MAD/MIRAS
- Solve
  - Maximum likelihood phasing – algorithms different from SHARP
  - SAD/SIR/MIR/MAD/MIRAS
  - Phenix
- Phaser
  - SAD - Correlated maximum likelihood phasing
  - Easily used with partial MR models
  - CCP4/Phenix
- Mlphare
  - Pseudo-maximum likelihood phasing
  - Not under active development
  - CCP4

Further Reading

- Liking Likelihood
  - Acta Cryst D60 2169 2004
- Likelihood-Based Experimental Phasing
  - In “Evolving Methods for Macromolecular Crystallography”, proceedings of the 2005 Erice Crystallography School
- Simple algorithm for a maximum likelihood SAD function
  - Acta Cryst D60 1220 2004