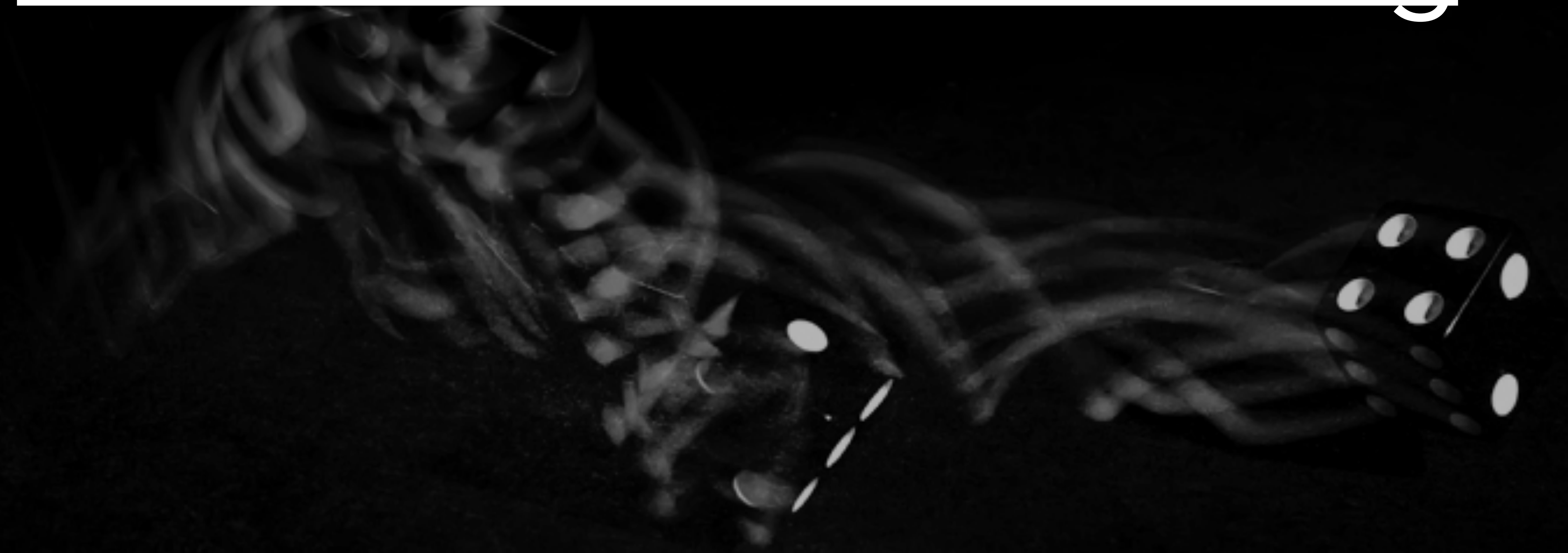
Lecture 3: Likelihoods and Numerical Minimizer Fitting



D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2022

Photo by Howard Jackman

Likelihoods and General Likelihood

- In today's lecture:
  - Maximum Likelihood
  - Extended Maximum Likelihood
  - Maximum Likelihood with binned (classified) data
- We will not be able to cover everything today

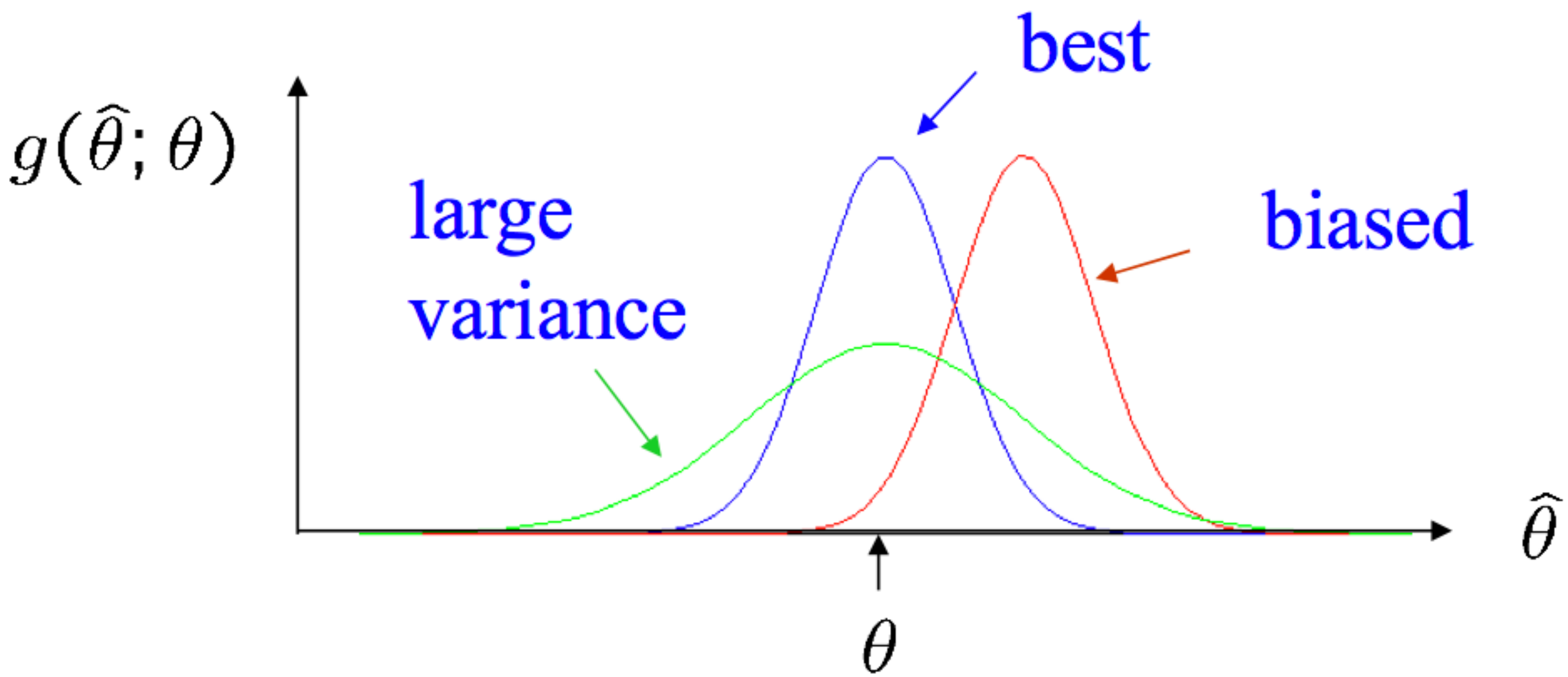
Material partially from D. Grant derived from lectures by A. Bellerive, G. Cowan, and D. Karlen

Estimating Parameters

- Given **n** observations one would like to describe the underlying (parent) probability distribution. The form of the parent distribution might be known, but there may be a number of unknown parameters
- The **n** observations may be used to determine the parameters as accurately as possible
- Define:
    - estimator - a function (f) of the observations used to determine the unknown parameter (θ)
    - estimate - the resulting value of the estimator, $\hat{\theta}$ or $\tilde{\theta}$
- A good estimator:
    - should not deviate from the true parameter value, in the limit of large **n**
    - the accuracy should improve as **n** increases

Estimating Parameters

- A good estimator cont.:
  - should be centered around the true parameter value for all n
  - should exhaust all the information in the measured data
  - should have a minimum variance (the best possible accuracy)
  - should be robust so as not to be sensitive to background or outliers



Estimating Parameters

- Test Statistics
  - The test statistic can be a function of one or more random variables (x, y, z, …) that are not dependent on any unknown parameters
  - One possibility for a test statistic is chi-squared, and today we're covering likelihoods
- Likelihood
  - For a random variable (x) distributed according to the PDF $f(x;\ \overrightarrow{\lambda})$, the functional form is known but the value of at least one parameter is unknown:

$$\vec{\lambda} = (\lambda_1, ..., \lambda_m)$$

  - The likelihood of observations in x for a specific λ is given by

$$L(x_1, x_2, ..., x_n; \lambda) = \prod_{i=1}^{n} f(x_i; \vec{\lambda})$$

Mini-Exercise

- We need data to test our likelihood code, so we will use Monte Carlo generated data where we **precisely know** the probability distribution function.

- Sample 4 data points from a Gaussian distribution. Do this for two different gaussian distributions (so 2 sets of 4 points):
  - Gaussian centered at 1.25 and a σ^2=0.11
  - Gaussian centered at 1.30 and a σ^2=0.50

- Plot the 4 points for each gaussian Monte Carlo sampling on the same plot, but in different colors.
  - It will look terrible and very non-gaussian, but the goal is to get code for generating MORE data, and also we can create a small enough Monte Carlo sample that we can check quickly

Likelihoods

$f()$ is commonly the probability distribution function

- The likelihood is the product of the individual 'probability' (or 'probabilities' for multiple parameters) of parameters (θ) which produce the observed outcomes (x_i)

$$\mathcal{L}(\vec{\theta}) = \prod_{i=1}^{N} f(x_i;\vec{\theta})$$

- The likelihood ($\mathscr{L}$ or L) given the observed data (x_i) for the parameters (θ) is equal to the probability (P) given the parameters (θ) of getting the observed data (x_i)

$$\mathcal{L}(\vec{\theta}|x) = P(x|\vec{\theta})$$

*changed from "λ" to "θ" to represent the parameter(s)

Simple Example

- We have 4 data points x = (1.01, 1.3, 1.35, 1.44) from a gaussian probability distribution function. What is the **likelihood** value for the data and the following values of the gaussian PDF:
  - Gaussian centered at 1.25 and a σ^2=0.11
  - Gaussian centered at 1.30 and a σ^2=0.50

$$\prod_{i=1}^{N} f(x_i;\vec{\theta}) \Rightarrow \frac{1}{\sqrt{2\pi\cdot 0.11}} e^{-\frac{(1.01-1.25)^2}{2\cdot 0.11}}$$

$$\times \frac{1}{\sqrt{2\pi\cdot 0.11}} e^{-\frac{(1.3-1.25)^2}{2\cdot 0.11}}$$

$$\times \frac{1}{\sqrt{2\pi\cdot 0.11}} e^{-\frac{(1.35-1.25)^2}{2\cdot 0.11}}$$

$$\times \frac{1}{\sqrt{2\pi\cdot 0.11}} e^{-\frac{(1.44-1.25)^2}{2\cdot 0.11}}$$

Simple Example Comments

- The probability distribution function is **normalized**. If it isn't normalized then you are very likely to get wrong and misleading values.
- Probability distribution (PDF) is the density of probability, and not the probability itself. Thus, the PDF calculated for individual data points can be above 1.
  - Gaussian A (centered at 1.25 and a σ^2=0.11) has a likelihood value of 1.29
  - Gaussian B (centered at 1.30 and a σ^2=0.50) has a likelihood value of 0.091
  - If we want to maximize the likelihood, Gaussian A is a better 'fit' to our data

log-Likelihoods

- Often "log" means "natural log" or better yet "$\ln$"
- Similar to using SI and non-SI units, **explicitly** use "$\ln$" or "$\log_{10}$" for the written form on plots or in write-ups.
- Why move from $\mathscr{L}$ to $\ln(\mathscr{L})$?
    - If you maximize/minimize $\ln(\mathscr{L})$ you also maximize/minimize in $\mathscr{L}$
    - Products ($\prod$) are converted to sums ($\sum$)
    - Exponentials and derivatives are easier to deal with in natural log space than straight likelihood space
    - It is common to use "LLH" to mean the "ln-likelihood"

Maximum Likelihood Method

- A very powerful and general method of parameter estimation when the functional form of the parent distribution, i.e. the probability distribution function, is known

- For large samples the estimates are often Gaussian distributed and hence the variances of the estimates are simple to determine

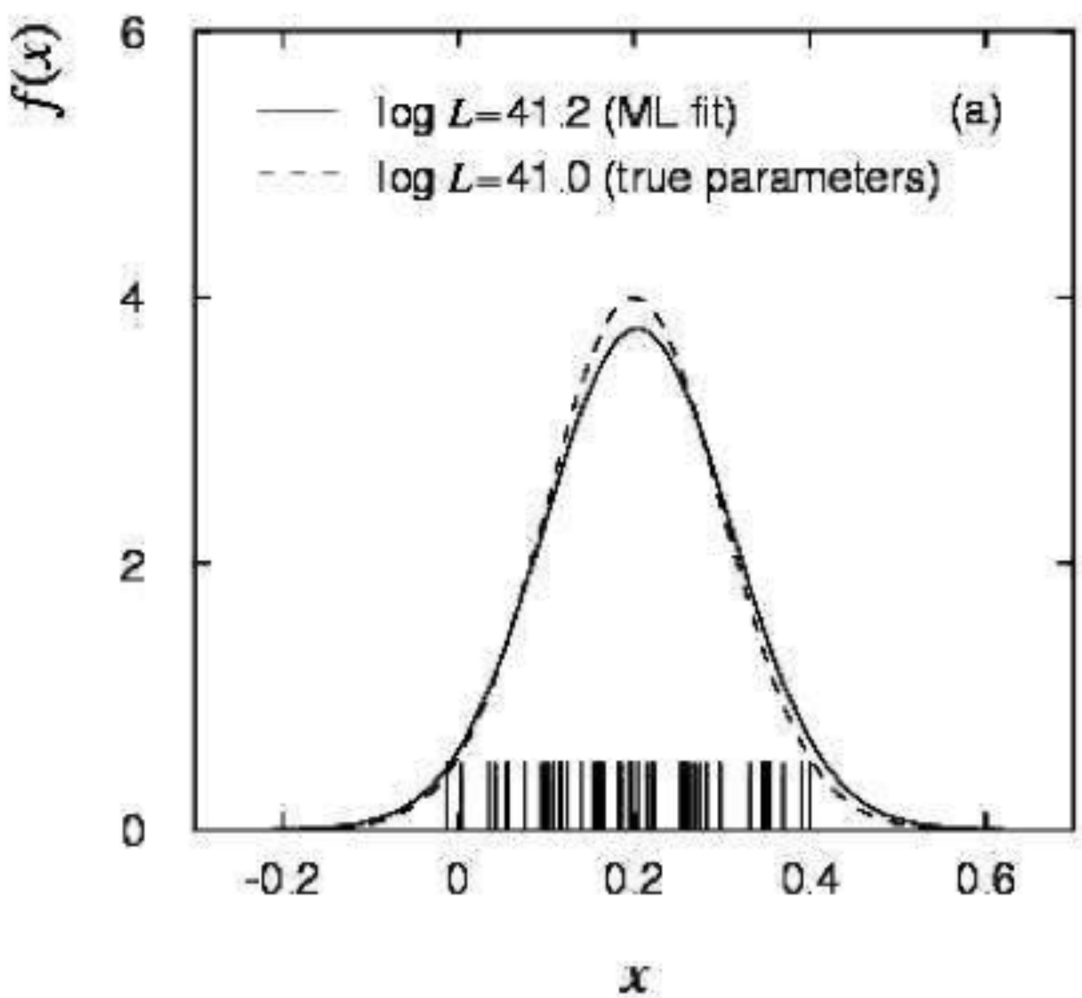
- Define: The estimate ($\hat{\lambda}$) is the value that maximizes the likelihood function

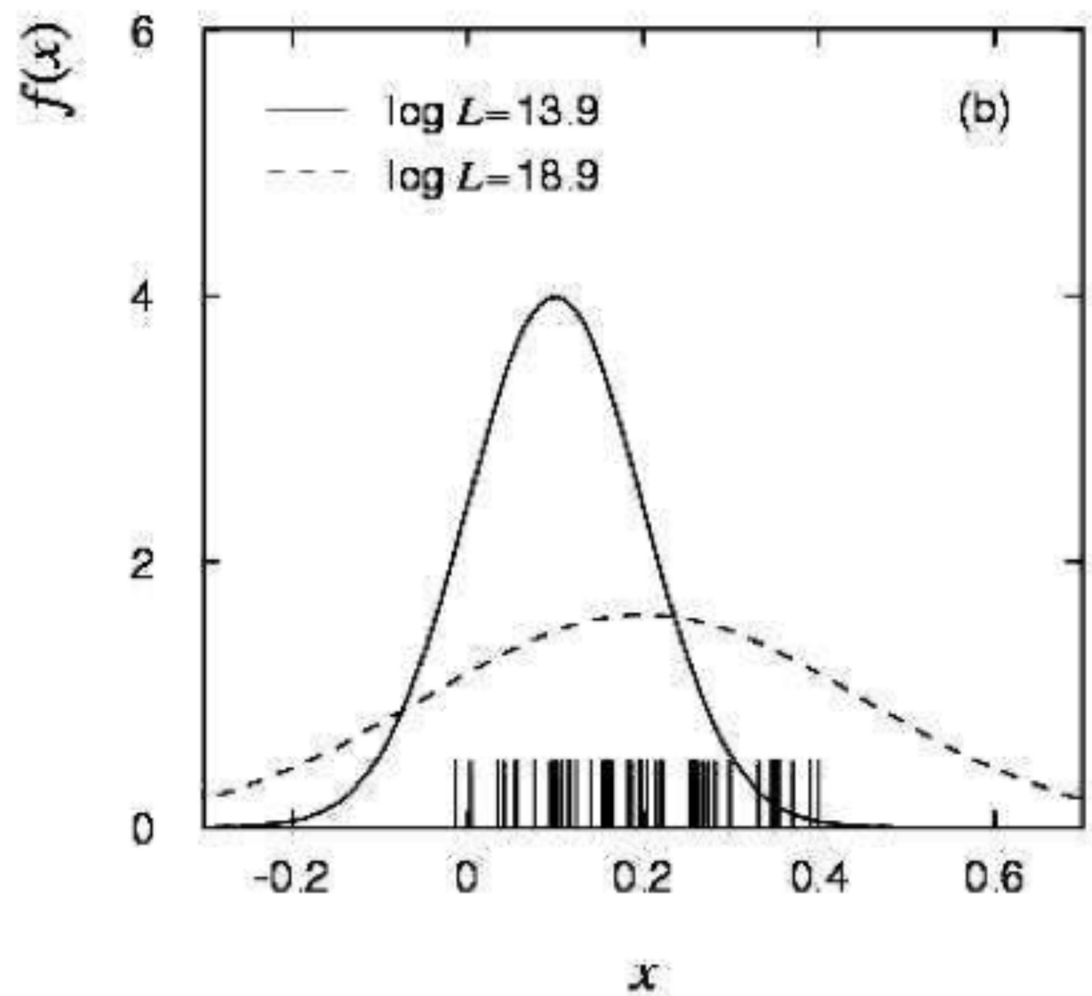
- Since the likelihood function and the natural logarithm ($\ln$) of the function have the same point for maximum values, analyzers will typically use the $\ln$(L)

$$\ln L = \sum_{i=1}^{n} \ln(f(x_i; \vec{\lambda}))$$

Maximum Likelihood Method

- Example of estimators
  - If the estimator is close to the true value then an expected high probability of obtaining data that matches exists

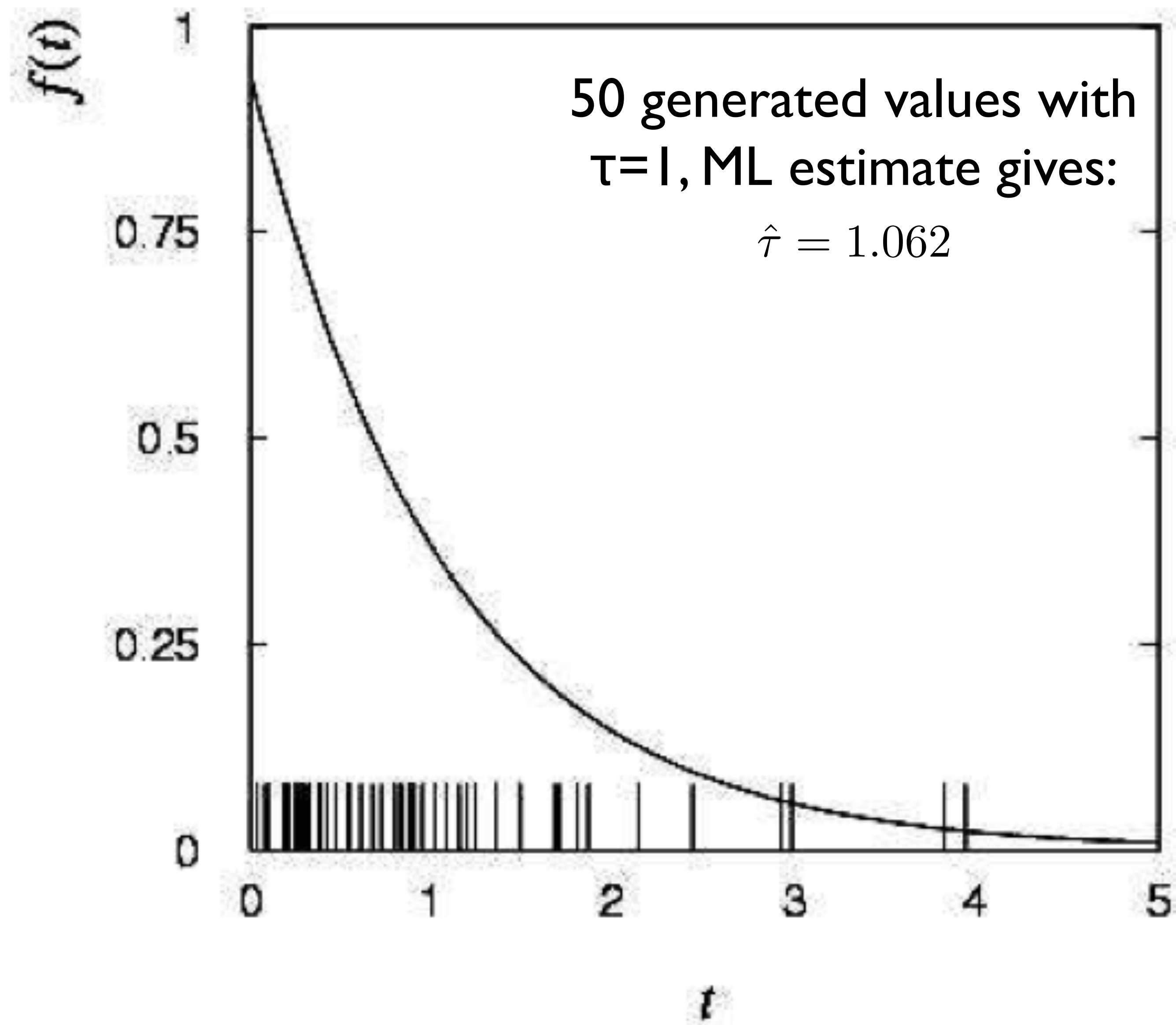




*taken as an example from “Statistical Data Analysis” - G. Cowan

Maximum Likelihood Method

- Example: Parameter of exponential PDF
  - Given an exponential PDF:
    $$f(t;\tau) = \frac{1}{\tau}e^{-t/\tau}$$
  - one can write a likelihood function for independent data (t_i):
    $$L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau}e^{-t_i/\tau}$$
  - The value where τ maximizes the likelihood function also gives the maximum value for the ln-likelihood function:
    $$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n}(\ln\frac{1}{\tau} - \frac{t_i}{\tau})$$
  - Maximum will be:
    $$\frac{\partial \ln L(\tau)}{\partial \tau} = 0 \rightarrow \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$



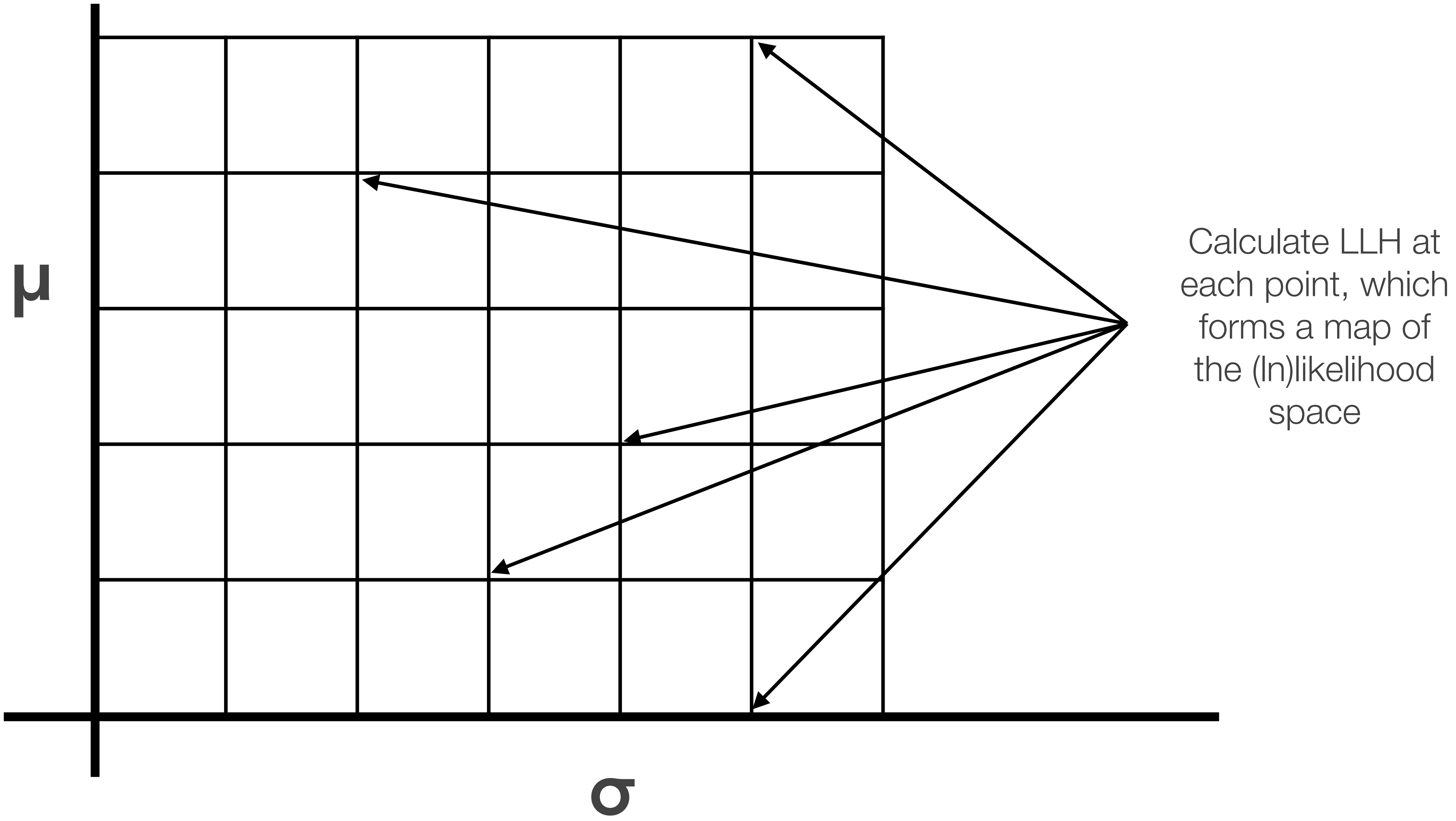
Exercise 1

- Once again making use of a gaussian PDF and the random number generator to calculate the LLH and the estimators
  - Be able to sample from a Gaussian PDF w/ the properties: $\mu=0.2$, $\sigma=0.1$, and 50 throws
  - You can establish the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$ analytically for such an easy example, but as a first option do a scan, i.e. 2D Raster scan. Note that technically an MLE has an analytic representation which is not always the same as what will be found by a scan, but the two are often close.

- What is meant by Raster Scan? Let's explore in the next slides

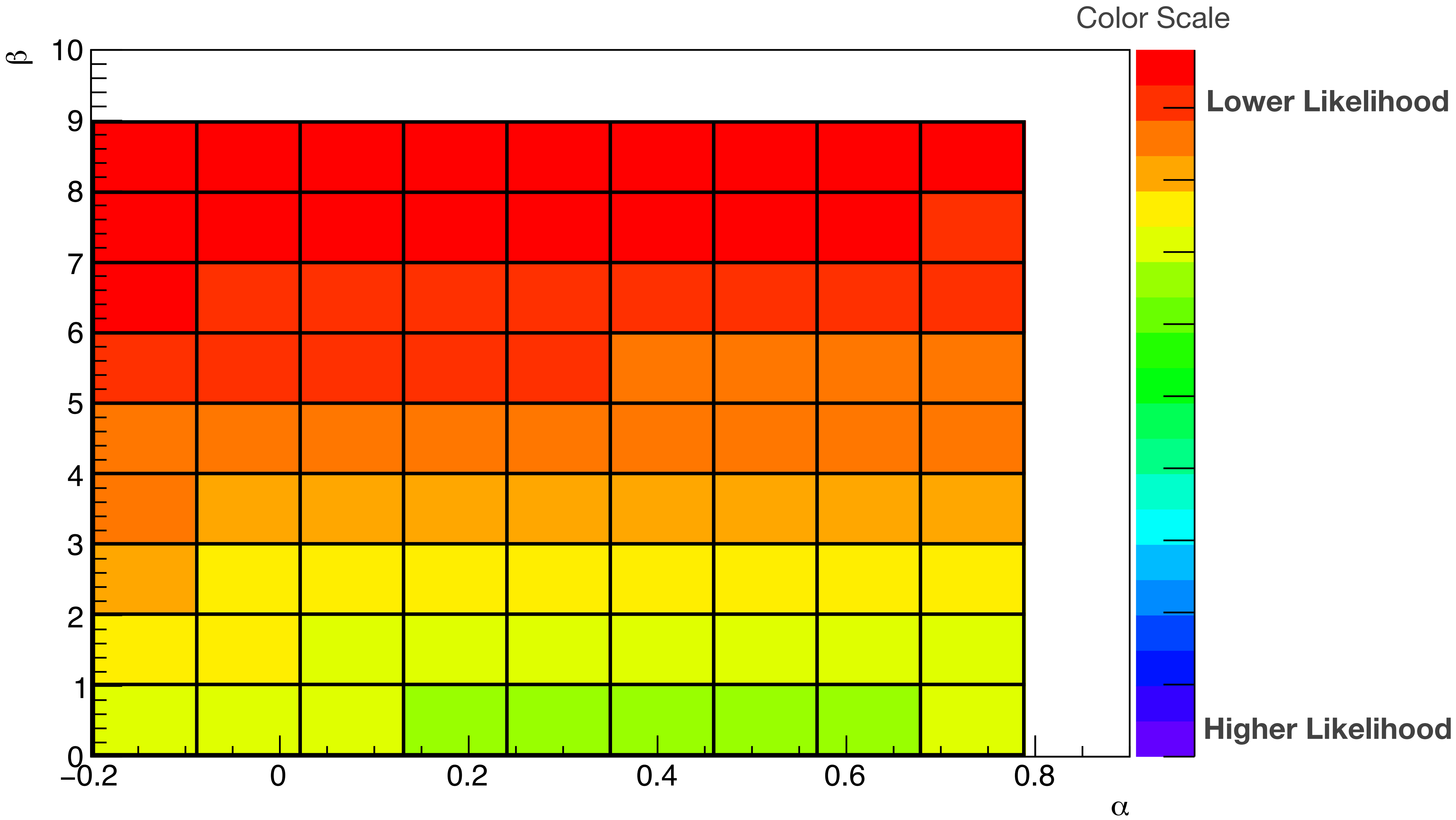
Raster Scan

- Fancy name for course/fine scanning the (ln)likelihood values σ and μ



Raster Scan

- From an exercise we will do later in this lecture



Scanning Comments

- The spacing and number of scanned points is determined by the analyzer
  - Too granular and the scan takes a long time
  - Too sparse can result in points that are not close to the actual MLE
- Dimensionality kills:
  - For the same number of points in each scanning dimension, the full likelihood calculation takes exponentially longer with each new dimension
  - In higher dimensions — and especially with semi-correlated parameters — the LLH space can be difficult to visualize

Exercise 1

MLE Values of 10^16 or 10^20 are okay.

- Once again making use of a gaussian PDF and the random number generator to calculate the LLH and the estimators
  - Gaussian w/ the properties: μ=0.2, σ=0.1, and 50 throws
  - You can establish the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$ analytically for such an easy example, but as a first option do a scan, i.e. 2D Raster scan. Note that technically an MLE has an analytic representation which is not always the same as what will be found by a scan, but the two are often nearby each other
- Compare the LLH for scanned MLE to the LLH using the input values. Do this for multiple iterations (50 new throws)
  - The analytic MLE should **always** have a better LLH than the LLH using the actual 'true' input values. The scanned MLE — for an appropriate precision in the scanning — is also likely have a 'better' LLH value than the LLH value from the known true inputs.

Exercise 2

- Multi-parameter likelihood
- Given a theoretical prediction with two independent parameters (**α**, **β**) which is:

$$f(x; \alpha, \beta) = 1 + \alpha x + \beta x^2$$

  - For **α**=0.5 and **β**=0.5, generate 2000 Monte Carlo data points using the above function transformed into a PDF over the range -1 ≤ x ≤ 1
  - Write your own likelihood function to 'fit' the estimators $\hat{\alpha}$ and $\hat{\beta}$ using the generated MC sample and a numerical minimizer/maximizer routine on either the -LLH or LLH to produce the estimator and if possible the parameter error

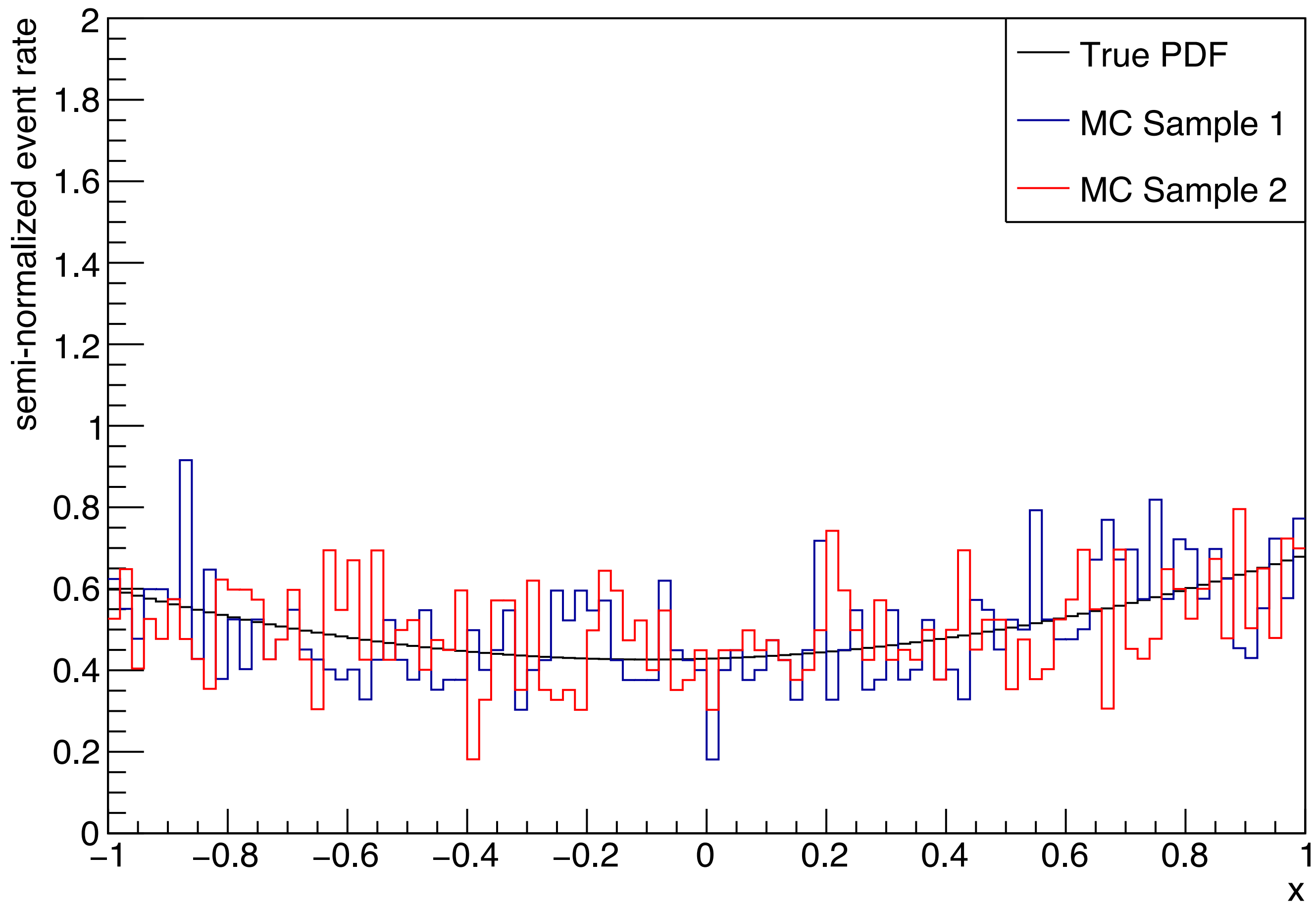
	Name	Value	Para Err	Err-	Err+	Limit-	Limit+
0	alpha =	0.4657	0.07151				
1	beta =	0.5227	0.1465				

Tips and Reminders

- Probability Distribution Functions are **normalized**
- ‘Fitters’ and ‘minimizers’ are often synonymous. It is customary to work in negative LLH space; hence the usage of ‘minimizers’ to find the ‘maximum’ likelihood estimator.
- MLE values on the previous slide will be different than your values. They should also differ for different ‘random’ samples of 2000 Monte Carlo data points.
- We are testing and developing the fitting method to get a parameter estimate, but it’s good to cross-check that the Monte Carlo data is reasonable.

Exercise 2 cont. w/ cross-check

- Write your own likelihood function to ‘fit’ the estimators $\hat{\alpha}$ and $\hat{\beta}$
- Fit using a numerical minimizer/maximizer routine on either the -LLH or LLH to produce the estimator and if possible the parameter error
- Using the new values **α**=0.1 and **β**=0.5, repeat the fitting procedure and plot the true distribution for the PDF and at least 1 of the samples w/ 2000 MC data points, and check that the returned values are good



*note, this doesn’t have to be a histogram, because the actual minimization should be unbinned

Exercise 2 cont.

- For those who want more…
- There are lots of different minimizers, so…
  - Figure out which minimizer you are using and benchmark the fitting time (or CPU resources)
  - Compare to a different minimizer. This can be either by name (MIGRAD versus BFGS) or by type (no-, first-, or second-derivative based algorithms)
  - Because the actual PDF is nicely analytic, smooth, and multi-parameter but not multi-dimensional, the derivative methods should be relatively quick

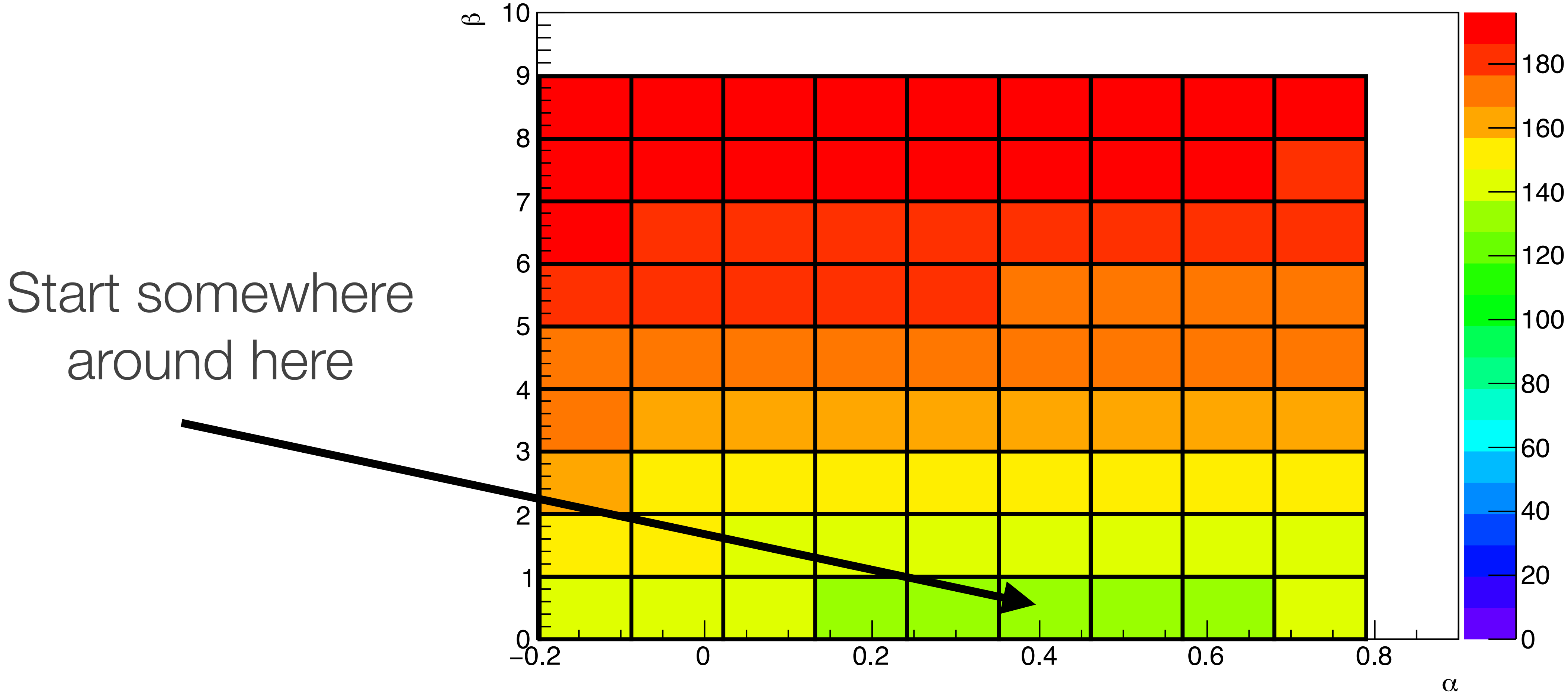
Numerical Minimization Notes

- The vast majority of numerical minimizers are dependent on initial settings and conditions to provide good fits in a reasonable time for real world PDFs
  - The higher the number of parameters, dimensionality, and the more complicated the LLH landscape the more important the initial settings and conditions
  - In the last example the LLH landscape is smooth, so the initial conditions shouldn't actually matter that much
    - You could speed it up by setting the starting point of the minimization closer to the true value, but that requires some initial guess of the true value
    - You can change the bounds on x, **α**, or **β**, which addresses problems of the fitting routine wandering away **and** keeps the fit in an appropriate range
    - You can do a coarse raster scan and start the minimizer in the cell/voxel with the best likelihood

Raster Scan

- This is a semi-coarse sampling of the LLH space. Establish which values of **α** and **β** have the best LLH and start your fit there, or at multiple points near the best LLH.

Start somewhere around here

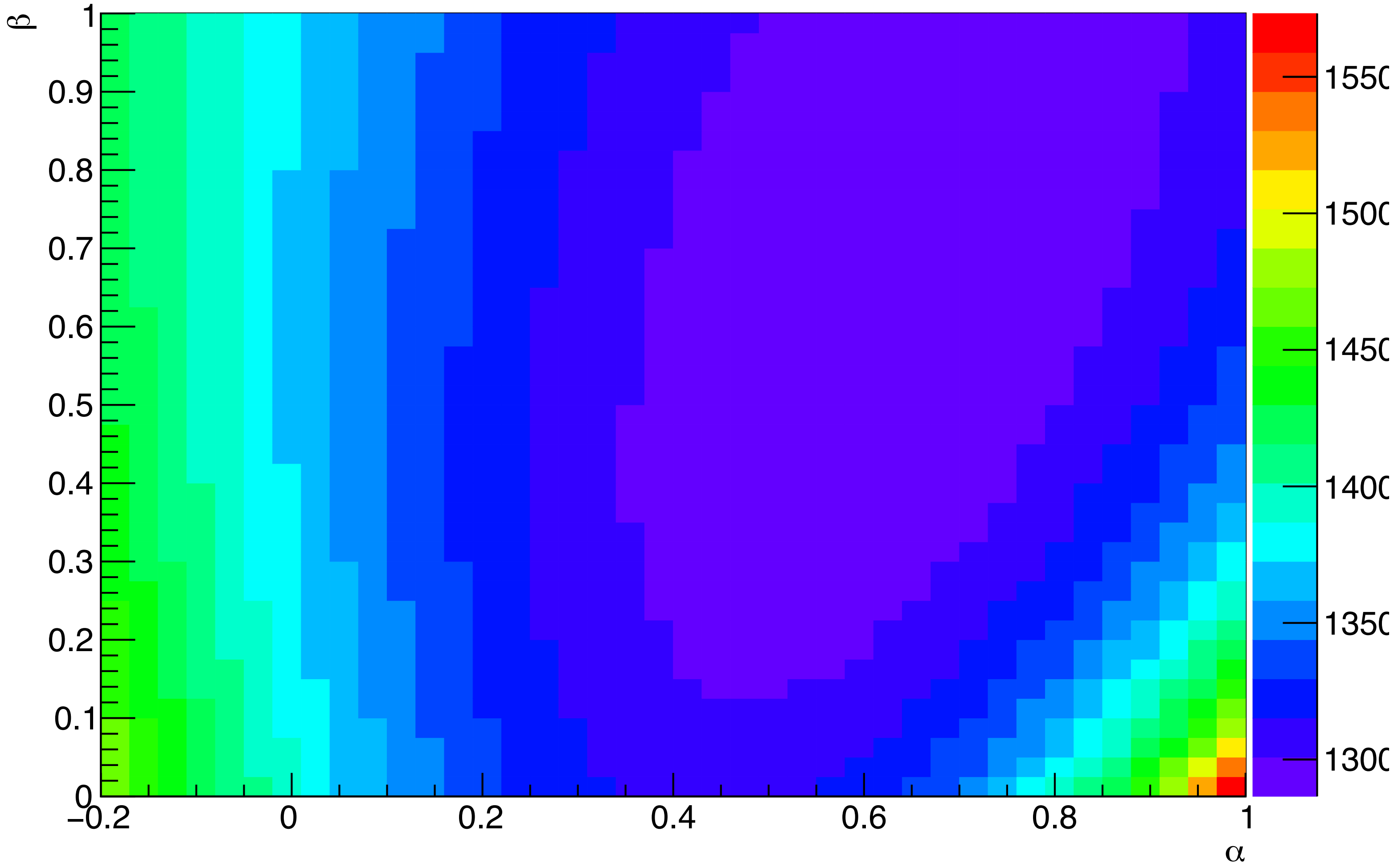


Numerical Minimization Ends

- Numerical minimizers require some criteria which terminates the minimization. Two common methods are:
  - **Number of steps**. This keeps the minimizer from ‘running away’, i.e. minimizing over infinite iterations.
  - **Estimated distance to minima** (EDM) or equivalent term for your minimizer. At some point near the true minima (at least at the precision of your data and minimizer) every infinitesimally small nearby point will have the same likelihood value. You can set the ΔLLH or ΔLH value criteria whereby when the minimizer encounters multiple steps below this threshold the minimization stops, and the MLE at the best LLH is considered the best-fit.

Exercise 3

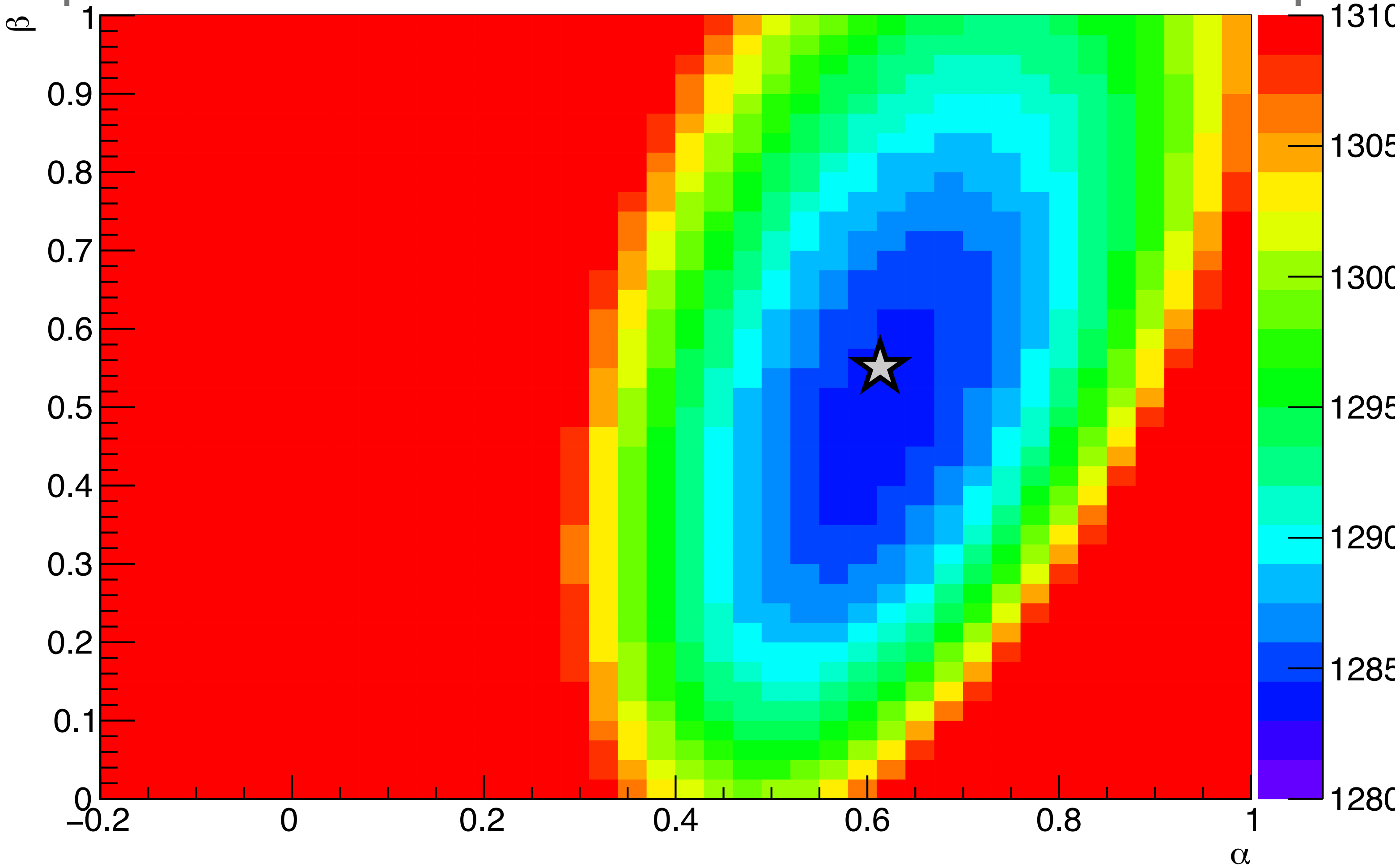
- Likelihood landscapes are important to visualize and understand… super important. Plot them whenever possible to understand the topology that your minimizer encounters
- For values of **α**=0.6 and **β**=0.5 for the previous formula/PDF make a 2D plot of the likelihood or LLH landscape



Exercise 3 cont.

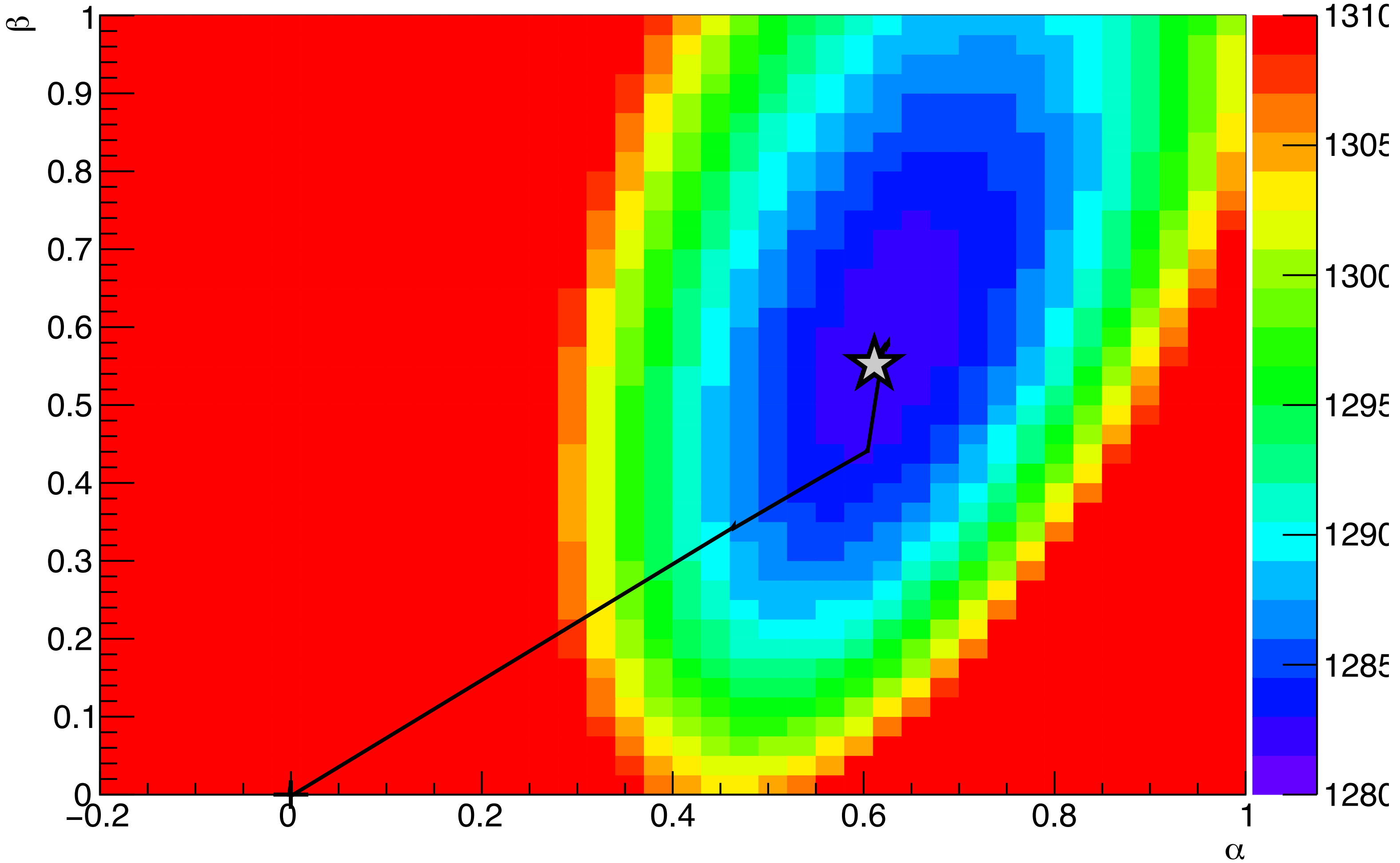
- Likelihood landscapes are important to visualize and understand… super important. Plot them whenever possible to understand the topology that your minimizer encounters
- For values of **α**=0.6 and **β**=0.5 for the previous formula/PDF make a 2D plot of the likelihood or LLH landscape

Zoomed in



Exercise 3 cont.

- For values of **α**=0.6 and **β**=0.5 for the previous formula/PDF make a 2D plot of the likelihood or LLH landscape and now plot the path of your minimizer as it 'steps' through the landscape



Zoomed in

Exercise 3 cont.

- For those who want more…
- Increase the number of Monte Carlo data points to 20000
  - Before you run the test, do you expect the value of the LLH at the MLE best-fit point to change versus 2000 points?
  - After you run the test, did the LLH change in a statistical meaningful way? Show empirically that it does or does not.
  - The 2D confidence interval can be assumed to be along iso-contour lines (contours) of constant ΔLLH from the best-fit. For the contour related to a ΔLLH=4, does it change between the sample with 2000 and 20000 MC data points?
  - Many statistical tests require that the ΔLLH be calculated versus the best-fit point. If your hypothesis test includes a fixed 'known' value, i.e. $\boldsymbol{\alpha}$=0.6 and $\boldsymbol{\beta}$=0.5, how much does the contour ΔLLH=4 change when calculated against the fixed values versus best-fit?

Exercise 3 cont.

- For those whom still want more…
  - Produce a ΔLLH landscape in reference to the best-fit point, i.e. the MLE
  - Produce Δ(ΔLLH) landscape for comparisons between using the best-fit as a LLH reference point and the 'true' values as the LLH reference point. This goal is a comparison of two hypothesis tests to see if it matters much, or at not at all, if the you use the best-fit, i.e. MLE, or true parameters. Commonly the 'true' would be replaced by the H_0 (null hypothesis), but here you can use the 'true' for testing.

Extended (General) Maximum Likelihood

- General idea
  - Given an assumed functional dependence f(x;θ) between the observable (x) and unknown parameter (θ) there are n events or observations. The likelihood function may be written as:

$$L = L(\vec{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

  - Sometimes n is not fixed but may instead be regarded as a Poisson random variable, with mean ν, which is the expected number of events. Written as a function of the parameters (θ) the information ν=ν(θ) may be used by generalizing the likelihood function:

$$L(n,\vec{x};\theta) = L(\nu,\theta) = \frac{\nu^n}{n!}e^{-\nu}L(\vec{x};\theta)$$

$$= \frac{\nu^n}{n!}e^{-\nu}\prod_{i=1}^{n} f(x_i;\theta)$$

Extended (General) Maximum Likelihood

- General idea
  - The ln-likelihood becomes:

$$\ln L(\vec{\theta}) = -\nu(\vec{\theta}) + \sum_{i=1}^{n} \ln(\nu(\vec{\theta}) f(x_i; \vec{\theta})) + C$$

  - The expression describes the joint probability for observing just n events and that those events provide the observations $x_1,\ldots,x_n$ when the number of observed events is assumed to be a Poisson variable with mean value **ν**.
  - The advantage of introducing the extended likelihood is the number of observed events (n) adds an additional constraint in determining the parameter(s), $\overrightarrow{\theta}$.
  - In problems where the **shape of the function** (f) is of primary interest we **gain little** by using the extended likelihood over the standard likelihood.
  - The extended likelihood should be applied in cases where the expected number of events can be calculated with considerable accuracy.

Extended (General) Maximum Likelihood

- Example
  - Consider two types of events, signal and background, each of which predicts a given PDF for the variable x: $f_s(x)$ and $f_b(x)$. What is observed is a mixture of the two event types. The signal fraction is given by θ, the expected total number ν, and observed total number n of events.
  - Let $\mu_s = \theta\nu$, $\mu_b = (1-\theta)\nu$. The goal is to estimate μ_s and μ_b which are poisson means of the signal and background.

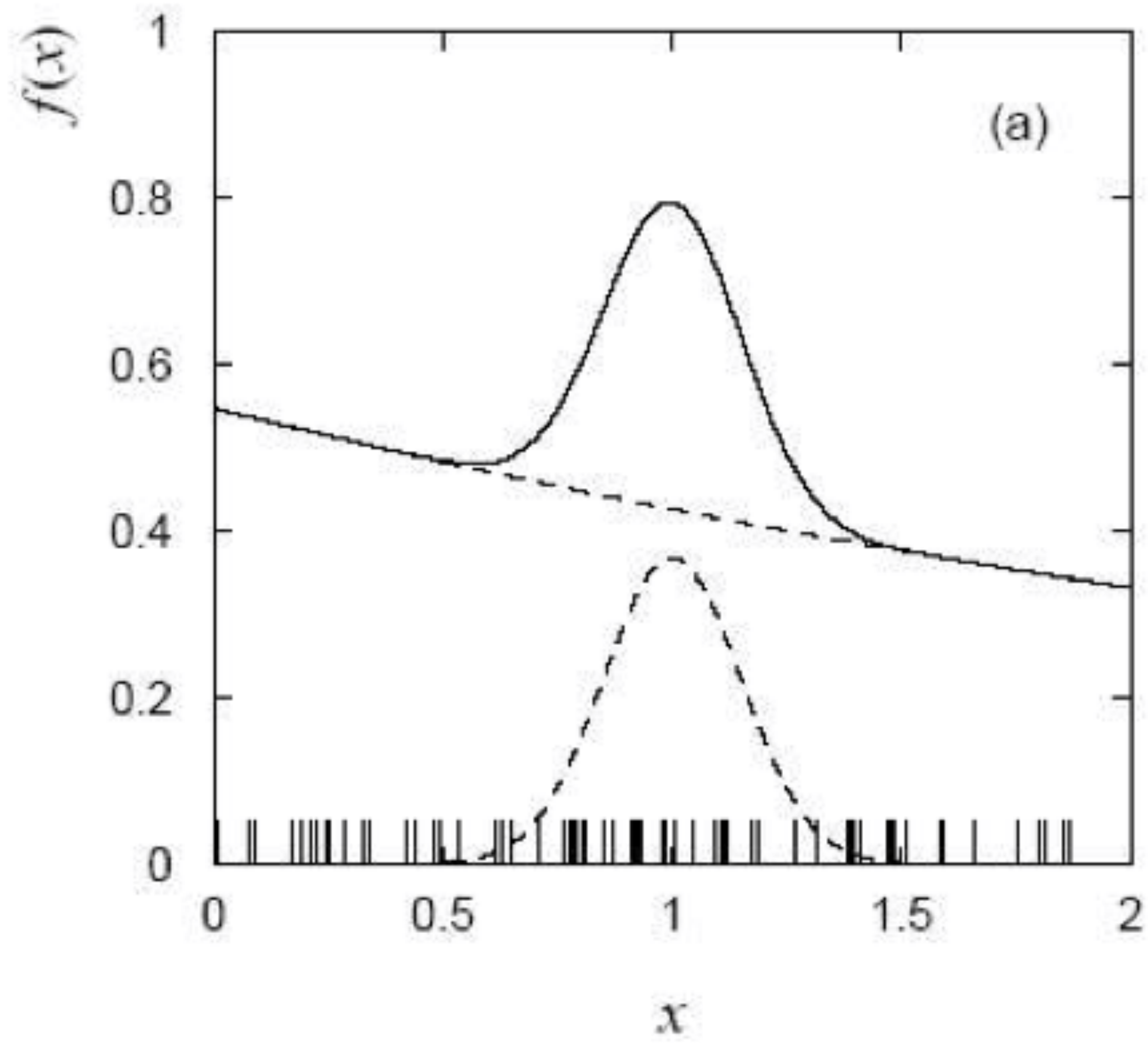
$$f(x;\mu_s,\mu_b) = \frac{\mu_s}{\mu_s+\mu_b} f_s(x) + \frac{\mu_b}{\mu_s+\mu_b} f_b(x)$$

$$P(n;\mu_s,\mu_b) = \frac{(\mu_s+\mu_b)^n}{n!} e^{-(\mu_s+\mu_b)}$$

$$\ln L(\mu_s,\mu_b) = -(\mu_s+\mu_b) + \sum_{i=1}^{n} \ln[(\mu_s+\mu_b) f(x_i;\mu_s,\mu_b)]$$

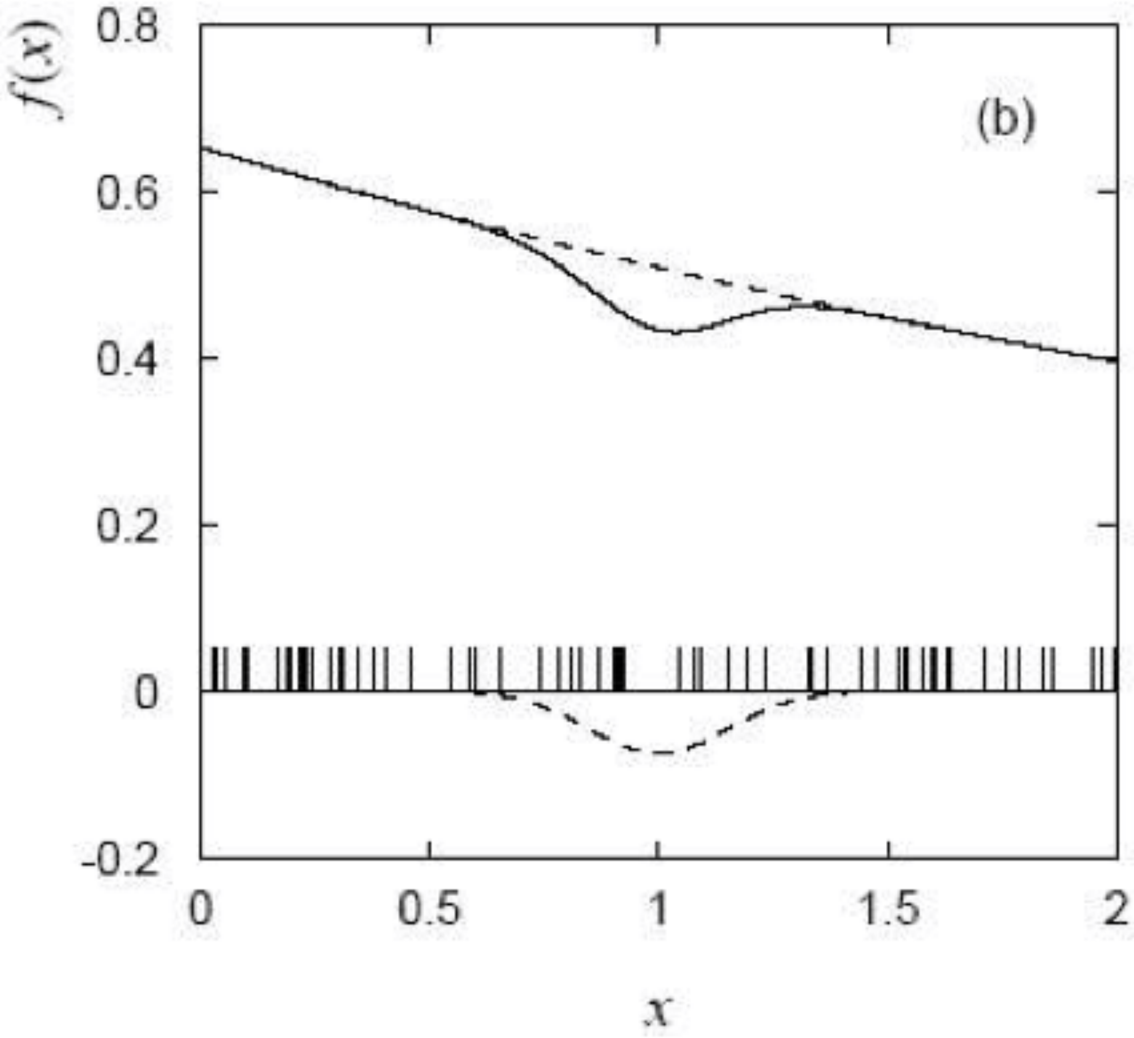
Extended (General) Maximum Likelihood

- Example
  - The signal distribution $f_s(x)$ is gaussian and background $f_b(x)$ is exponential, where the Monte Carlo sample (right plot) has true poisson means of: $\mu_s = 6;\ \mu_b = 60$
  - The ln-likelihood is maximized in terms of μ_s and μ_b : $\hat{\mu}_s = 8.7 \pm 5.5$
    $\hat{\mu}_b = 54.3 \pm 8.8$
  - In this case the errors reflect the total Poisson fluctuation as well as that in proportion to signal/background.



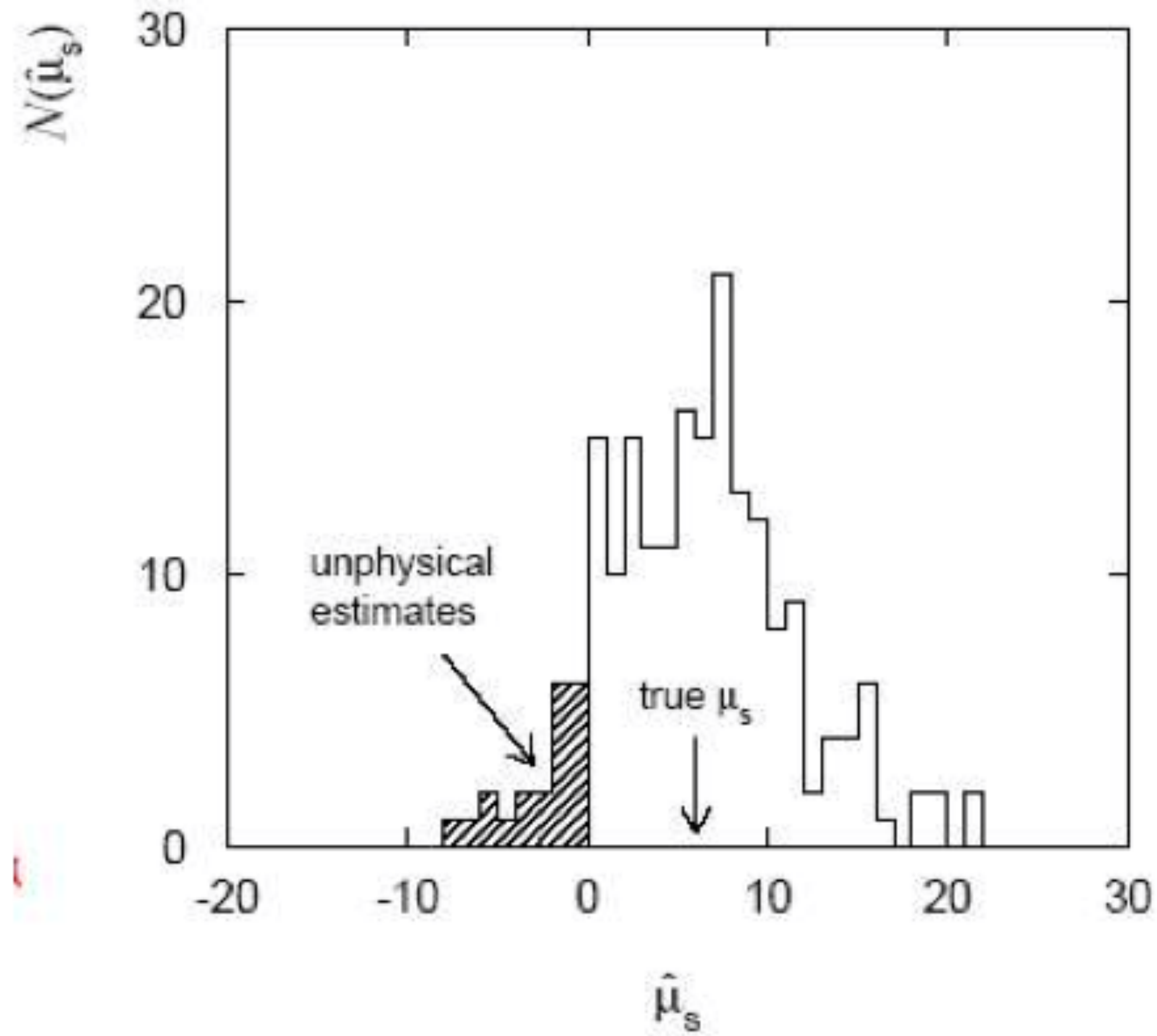
Extended (General) Maximum Likelihood

- Example
    - What if we now consider an unphysical estimate, e.g. a downward fluctuation of data in the peak region which can lead to fewer events that what would otherwise be obtained from background alone.
    - The estimate for μ_s is now pushed negative into an unphysical regime.
    - This is OK as long as the total PDF remains positive everywhere.



Extended (General) Maximum Likelihood

- Example
  - The unphysical estimator is unbiased and should ultimately be reported since the average of a large number of unbiased estimates will converge to the true value.
  - If you repeat the entire Monte Carlo many times then one may allow unphysical estimates.
  - In order to provide unbiased confidence limits and coverage



Additional

- A very useful minimization algorithm is “SIMPLEX”, which is also known by the authors names “Nelder & Mead”
  - https://academic.oup.com/comjnl/article/7/4/308/354237
- Review of the limited memory BFGS method
  - “On the limited memory BFGS method for large scale optimization” https://link.springer.com/article/10.1007/BF01589116
- The Powell algorithm does not require gradients
  - https://academic.oup.com/comjnl/article/7/2/155/335330
- Besides these references, wikipedia is quite good too as a place to start getting some backstory on the algorithms before going to the academic journal papers

Additional Material

(will not be covered in lecture, but should be reviewed)

Maximum Likelihood with Binned (Classified) Data

- In the case where the number of observations is very large, numerical evaluation of the likelihood function may become intensive, in particular if the PDF has a complex form.
- In such cases it is possible to reduce the amount of computation by grouping the data into subsets or classes and write the likelihood function as a product of a smaller number of averaged PDFs.
- In doing this there is clearly some loss of information. This loss will be modest if the variation of the distribution is small over each interval.
- Let the total number of events (n) be grouped into N classes for different intervals of the variable x. The joint probability to have n_1 events in class 1, n_2 events in class 2, etc. is given by a multinomial distribution.

Maximum Likelihood with Binned (Classified) Data

- Data will often be placed into a histogram:

$$\vec{n} = (n_1, ..., n_N),\ n_{tot} = \sum_{i=1}^{N} n_i$$

- The hypothesis is that:

$$\vec{\nu} = (\nu_1, ..., \nu_N),\ \nu_{tot} = \sum_{i=1}^{N} \nu_i \qquad \nu_i(\vec{\theta}) = \nu_{tot} \int_i f(x;\vec{\theta})dx$$

- If the data is modeled as a multinomial, then

$$f(\vec{n};\vec{\nu}) = \frac{n_{tot}!}{n_1!...n_N!}(\frac{\nu_1}{n_{tot}})^{n_1}...(\frac{\nu_N}{n_{tot}})^{n_N}$$

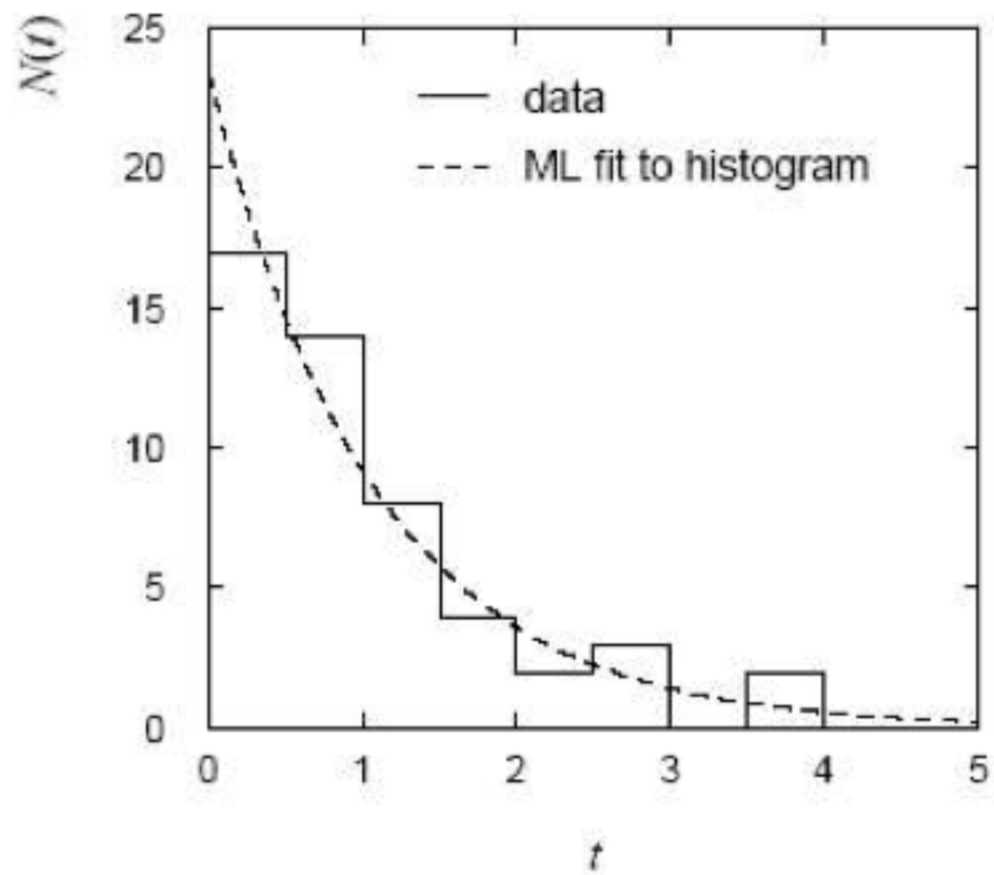
- and the ln-likelihood function becomes:

$$\ln L(\vec{\theta}) = \sum_{i=1}^{N} n_i \ln \nu_i(\vec{\theta}) + C$$

Maximum Likelihood with Binned (Classified) Data

- Take our historical example using the exponential, placing that data into a histogram.
- In the limit of zero bin width then one achieves the usual unbinned maximum likelihood.
- If each n is treated as a Poisson random variable, then we obtain the extended ln-likelihood:

$$\ln L(\nu_{tot}, \vec{\theta}) = -\nu_{tot} + \sum_{i=1}^{N} n_i \ln \nu_i(\nu_{tot}, \vec{\theta}) + C$$



$\hat{\tau} = 1.07 \pm 0.17$

(1.06 ± 0.15 for unbinned ML with same sample)

Maximum Likelihood with Binned (Classified) Data

- In the above problem it is equivalent to treat the number of events in each bin as an independent Poisson random variable, n_i, with mean value ν_i .

- The relationship that considers the dependence between this ν_{tot} and the other parameters, θ, is such that if there is no functional relation between ν_{tot} and the θ then one obtains $\hat{\nu}_{tot} = n_{tot}$ and the estimate for the parameters, $\hat{\theta}$, are the same as when the Poisson term is not included.

- If ν_{tot} is given as a function of θ, then the variance of the estimated parameters are in general reduced by including the Poisson term information.

- NOTE: the determination of parameters from histograms by quadratic sum minimization (chi-square) gives less precise results than those obtained by likelihood maximization. This is due to the assumption of the normal distribution for the values n_i requires large bin widths and therefore loss of information.

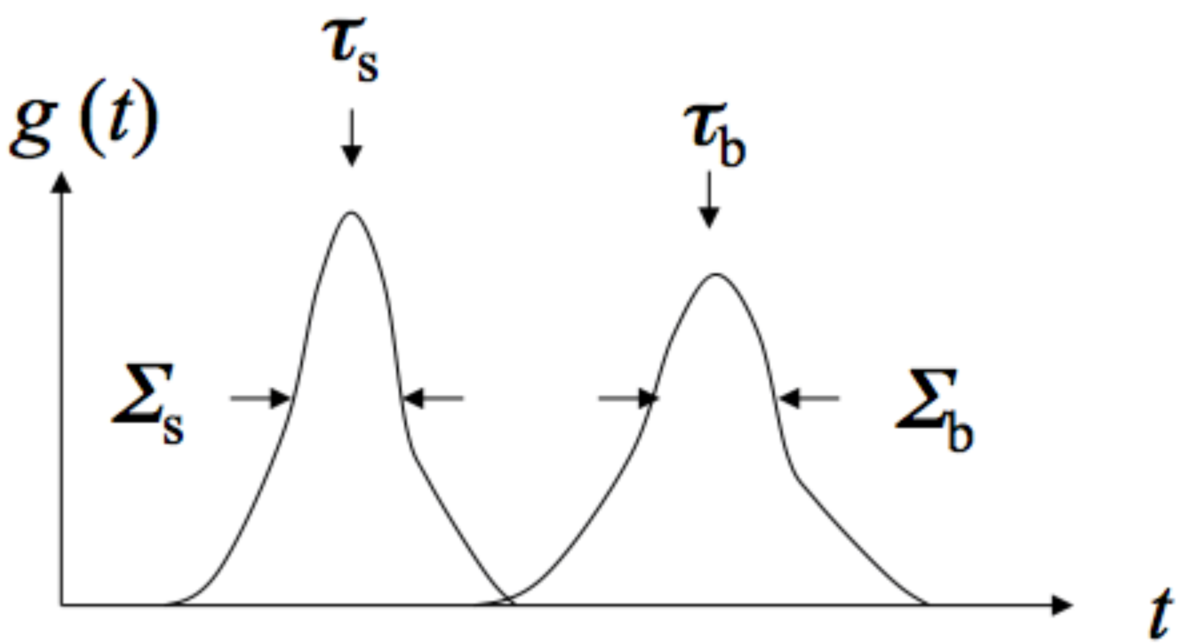
Multivariate methods for test statistics

- Linear Test Statistic
  - Try:

$$t(\vec{x}) = \sum_{i=1}^{n} a_i x_i$$

  - We choose the parameters, a, such that the PDFs will have maximum separation.
  - Construct the Fisher variable, which we maximize:

$$J(\vec{a}) = \frac{(\tau_s - \tau_b)^2}{\Sigma_x^2 + \Sigma_b^2}$$



large distance between mean values and small widths

Multivariate methods for test statistics

- Coefficients of maximum separation
  - We have

$$(\mu_k)_i = \int x_i f(\vec{x}|H_k) d\vec{x}$$

$$(V_k)_{ij} = \int (x-\mu_k)_i (x-\mu_k)_j f(\vec{x}|H_k) d\vec{x}$$

$k = 0, 1$ (hypothesis) $\quad\quad$ $i, j = 1, ..., n$ (component of $\vec{x}$)

  - In terms of mean and variance for the test statistic, t, then:

$$\tau_k = \int t(\vec{x}) f(\vec{x}|H_k) d\vec{x} = \vec{a}^T \vec{\mu}_k$$

$$\Sigma_k^2 = \int (t(\vec{x}) - \tau_k)^2 f(\vec{x}|H_k) d\vec{x} = \vec{a}^T V_k \vec{a}$$

Multivariate methods for test statistics

- Coefficients of maximum separation
  - The numerator of J(a) is:

$$(\tau_0 - \tau_1)^2 = \sum_{i,j=1}^{n} a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j$$

$$= \sum_{i,j=1}^{n} a_i a_j B_{ij} = \vec{a}^T B \vec{a}$$ Between classes

  - the denominator is:

$$\Sigma_0^2 + \Sigma_1^2 = \sum_{i,j=1}^{n} a_i a_j (V_0 + V_1)_{ij} = \vec{a}^T W \vec{a}$$ Within classes

  - Therefore we maximize:

$$J(\vec{a}) = \frac{\vec{a}^T B \vec{a}}{\vec{a}^T V \vec{a}} = \frac{\text{separation between classes}}{\text{separation within classes}}$$

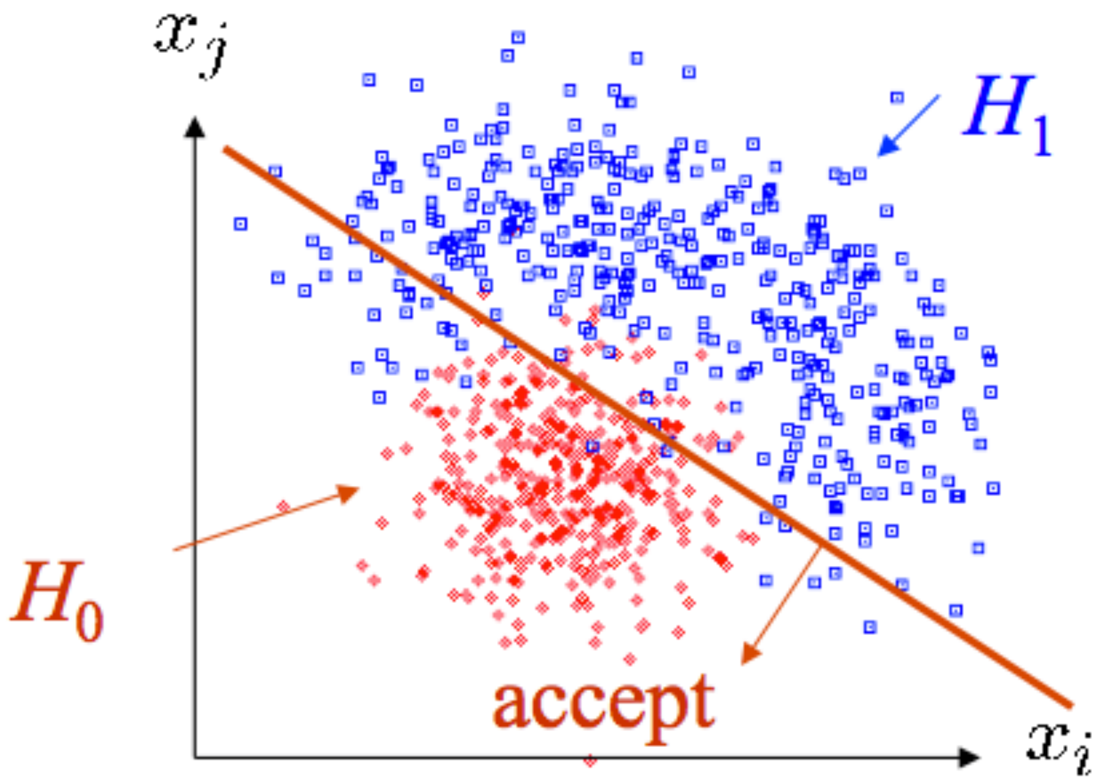
Multivariate methods for test statistics

- Fisher discriminant
  - Setting the first derivative of J equal to zero:

$$\frac{\partial J}{\partial a_i} = 0$$

  - gives us Fisher's linear discriminant

$$t(\vec{x}) = \vec{a}^T \vec{x} \qquad \vec{a} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1)$$



Multivariate methods for test statistics

- Fisher discriminant with Gaussian data
  - What if your PDF is a multivariate Gaussian with mean values given by:

$$E_0[\vec{x}] = \vec{\mu}_0 \text{ for } H_0 \qquad E_1[\vec{x}] = \vec{\mu}_1 \text{ for } H_1$$

  - In this case the covariance matrices are $V_0=V_1=V$ for both. The Fisher discriminant, with an offset, can be written as:

$$t(\vec{x}) = a_0 + (\vec{\mu}_0 - \vec{\mu}_1)^T V^{-1} \vec{x}$$

  - The likelihood ratio then becomes:

$$r = \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)}$$

$$= \exp[-\frac{1}{2}(\vec{x} - \vec{\mu})_0^T V^{-1} (\vec{x} - \vec{\mu}_0) + \frac{1}{2}(\vec{x} - \vec{\mu})_1^T V^{-1} (\vec{x} - \vec{\mu}_1)]$$

$$\propto e^t$$

Multivariate methods for test statistics

- Fisher discriminant with Gaussian data
    - Therefore, for this case

$$t \propto \ln r + C$$

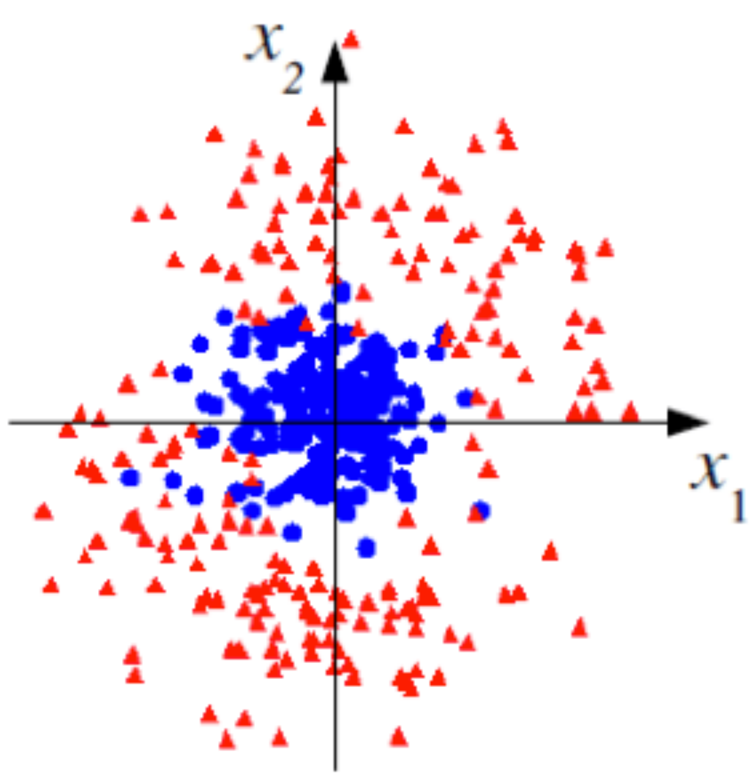
    - The Fisher discriminant is equivalent to the likelihood ratio and therefore gives maximum purity for a given efficiency.
    - When data is non-Gaussian this no longer holds, but the linear discriminant function may still be the simplest practical solution.
    - One often tries to transform data so that it better approximates a Gaussian before constructing the Fisher discriminant.

Multivariate methods for test statistics

- Fisher discriminant with Gaussian data
  - eg. non-linear transformation of inputs

$$x_1, ..., x_n \to \phi_1(\vec{x}), ..., \phi_m(\vec{x})$$

  - We have transformed the "feature space" var separated by a linear boundary.



$$\phi_1 = \tan^{-1}(x_2/x_1)$$

$$\phi_2 = \sqrt{x_1^2 + x_2^2}$$