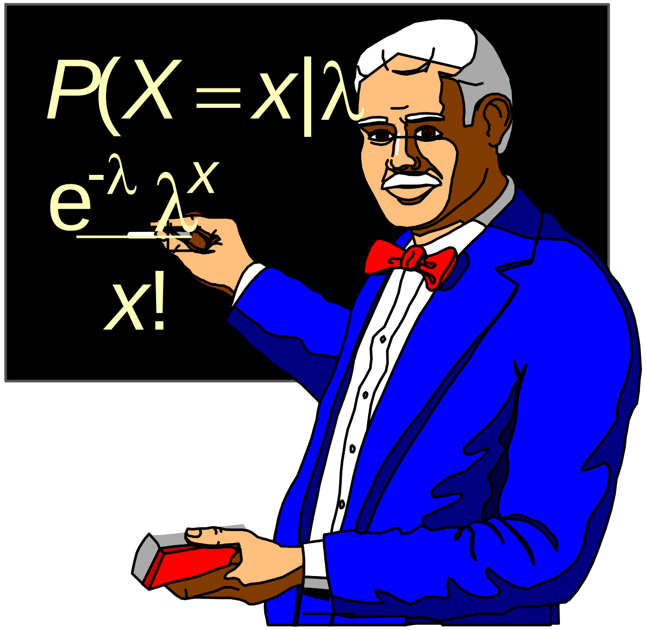
Phys 443
Computational Physics

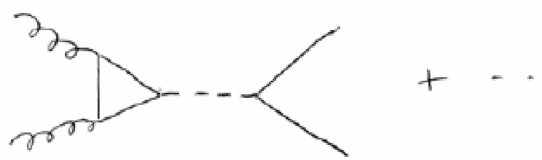
Maximum Likelihood Method



Theory-Statistics-Experiment

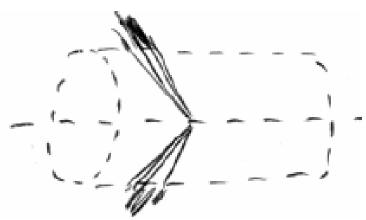
Theory (model, hypothesis):

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\not{D}\psi + \cdots$$

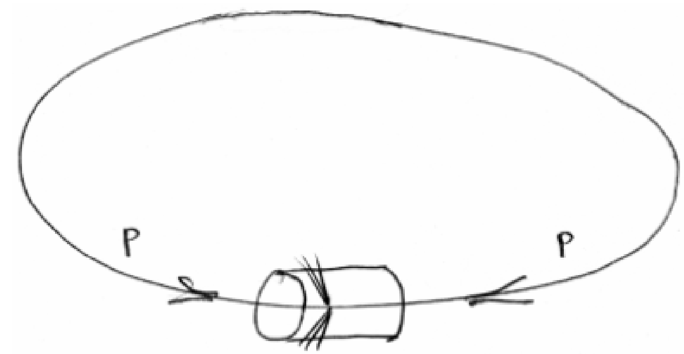


$$\sigma = \frac{G_F\,\alpha_s^2\,m_H^2}{288\sqrt{2\pi}} \times \sim$$

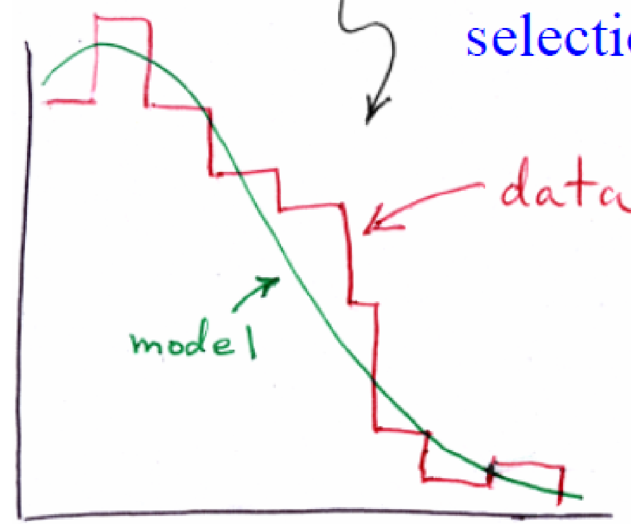
+ simulation of detector and cuts



Experiment:



+ data selection



Maximum Likelihood Method

- So far we have been finding estimators using our intuition and evaluated their properties (bias, variance, MSE) on a case-by-case basis. In (almost all of) our simple examples, the parameters were population averages (expectations) of some sort, which we estimated using sample means.

- What should we do if we work with complicated models for which we can't come up with intuitive estimators? A possible answer to this question is using Maximum Likelihood Estimation, which is a systematic method for finding estimators (given a model).

Maximum Likelihood Method

- **It is very general and poweful method for parameter estimation.**
- It provides estimators with desirable properties, and estimators are easy to find.
- The ML theory has a fundemantal position in all problems of parameter estimation where the functional form **of pdf is given.**
- For large samples the ML estimates are normally distributed. This makes the determination of variances on ML estimates very simple.

Maximum Likelihood Method

- Suppose we would like to measure the **true value of some quantity (x_T).** We make repeated measurements of this quantity $\{x_1, x_2, \ldots, x_n\}$. The standard way to estimate x_T from our measurements is to calculate the mean value of the measurements:

$$\boxed{\mu_x = \frac{\sum_{i=1}^{N} x_i}{N}}$$

and set $x_T = \mu_x$

Does this procedure make sense?

The MLM answers this question and provides **a method for estimating parameters from existing data**.

Maximum Likelihood Method

MLM: a general method for estimating parameters of interest from data.
Assume we have made *N measurements of* x $\{x_1, x_2, \ldots x_n\}$.
Assume we know the probability distribution function that describes x: $f(x, \alpha)$.
Assume we want to determine the parameter α.

MLM: pick α to maximize the probability of getting the measurements (the x_i's) that we did!
How do we use the MLM?

The probability of measuring x_1 is $f(x_1, \alpha)dx$
The probability of measuring x_2 is $f(x_2, \alpha)dx$
The probability of measuring x_n is $f(x_n, \alpha)dx$

If the measurements are independent, the probability of getting the measurements we did is:

$$L = f(x_1,\alpha)dx \,.\, f(x_2, \alpha)dx \,\ldots.\, f(x_n,\alpha)dx = f(x_1, \alpha)\,.f(x_2, \alpha) \ldots f(x_n,\alpha)dx^n$$

$$L = \prod_{i=1}^{N} f(x_i, \alpha)$$ Likelihood Function **We drop the dx^n since it is just proportionality constant**

Here ***L(α)*** **may** be considered proportional to the probability density associated to the random event "the true value of the parameter is ***α***"

Maximum Likelihood Method

We expect that ***L(α)*** will be higher for ***α*** values which are close to the true one, so we look for the value which makes ***L(α)* maximum**

determine the α that maximizes L:

$$\left.\frac{\partial L}{\partial \alpha}\right|_{\alpha=\alpha^*} = 0 \qquad \left.\left(\frac{\partial^2 \ln L}{\partial \alpha^2}\right)\right|_{\alpha=\alpha^*} < 0$$

Both L and lnL have maximum at the same location

Maximize lnL rather than L itself because *lnL* converts the product into a summation

$$\ln L = \sum_{i=1}^{N} \ln f(x_i, \alpha)$$

New maximization condition

$$\left.\frac{\partial \ln L}{\partial \alpha}\right|_{\alpha=\alpha^*} = \left.\sum_{i=1}^{N} \frac{\partial}{\partial \alpha} \ln f(x_i, \alpha)\right|_{\alpha=\alpha^*} = 0$$

α could be an array of parameters (i.e. slope and intercept) or just a single variable

Equations to determine α range from simple linear equations to coupled non linear equations.

Maximum Likelihood Method

- Numeric methods are often needed to find the maximum of *ln(L).* Especially difficult if there is more then one parameter. Standard tool in HEP: MINUIT

  http://inspirehep.net/record/1258345/files/mnusersguide.pdf

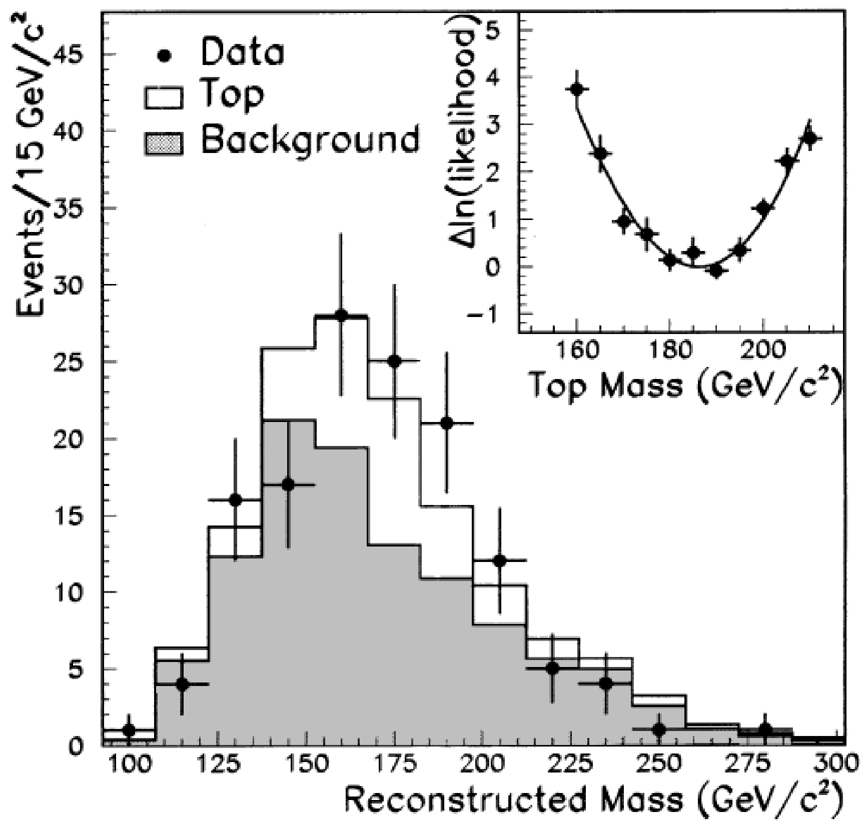
- It does not give you the 'most likely value of α' – it gives you the value of α for which this data is most likely.

Example

- In their paper* on all-hadronic decays of $t\bar{t}$ pairs CDF retained 136 events with at least one b-tagged jet and plotted the 3-jet invariant mass ($W^+b \longrightarrow q\bar{q}b \longrightarrow$ 3 jets).

The ML method was applied to extract the top quark mass: in the 11 HERWIG MC samples m_{top} was varied from 160 to 210 GeV, ln(likelihood) values were plotted to extract m_{top} = 186 GeV and a $\pm$10 GeV statistical error

The background is calculated by normalizing the spectrum of the untagged sample of 1121 events to 108 $\pm$ 9 events, estimated from the tag probability. A maximum likelihood method is applied to extract the top quark mass. The experimental data are compared to HERWIG Monte Carlo samples of $t\bar{t}$ events, in a top quark mass range from 160 to 210 GeV/c^2, and a background sample from the untagged events. The same method was applied to Refs. [1] and [2]. The difference in $-$ ln(likelihood) with respect to the minimum is shown in the inset to Fig. 1. The minimum is at 186 GeV/c^2, with a $\pm$10 GeV/c^2 statistical uncertainty. Systematic uncertainties in this fit arise from



* F. Abe et al., Phys. Rev. Lett. 79 (1997) 1992

Example

- Suppose that X is a discrete random variable with the following probability distribution function: where 0< θ<1 is a parameter. The following 10 independent observation

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

Where taken from such a distribution : (3,0,2,1,3,2,1,0,2,1). What is the maximumum likelihood estimate of θ.

The likelihood is

$$\begin{aligned} L(\theta) = \quad & P(X=3)P(X=0)P(X=2)P(X=1)P(X=3) \\ \times \quad & P(X=2)P(X=1)P(X=0)P(X=2)P(X=1) \end{aligned}$$

Example

$$L(\theta) = \quad P(X=3)P(X=0)P(X=2)P(X=1)P(X=3)$$
$$\times \quad P(X=2)P(X=1)P(X=0)P(X=2)P(X=1)$$

Exercise!

$$\begin{aligned} l(\theta) &= \log L(\theta) = \sum_{i=1}^{10} \log P(X_i|\theta) \\ &= 2\left(\log\frac{2}{3} + \log\theta\right) + 3\left(\log\frac{1}{3} + \log\theta\right) + 3\left(\log\frac{2}{3} + \log(1-\theta)\right) + 2\left(\log\frac{1}{3} + \log(1-\theta)\right) \\ &= C + 5\log\theta + 5\log(1-\theta) \end{aligned}$$

$$\frac{dl(\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} = 0$$

$$\hat{\theta} = 0.5.$$

Maximum Likelihood Method: Binomial

Example

- **Let *f(x, θ)* be given by a Binomial distribution**
- Let $\theta=np$ be the mean of the Binomial
- We want the best estimate of θ from our set of n measurements.

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f\left(x_i|\theta\right) \\
&= \theta^r (1-\theta)^{n-r}
\end{aligned}
$$

Here r is the number of "heads" observed and n−r is the number of tails. Note that we did not include the usual combinatorial (binomial) term in front of the expression above, to count the number of different ways that r heads could occur in n trials. Since his term does not involve θ, we can ignore this term.

Maximum Likelihood Method: Binomial

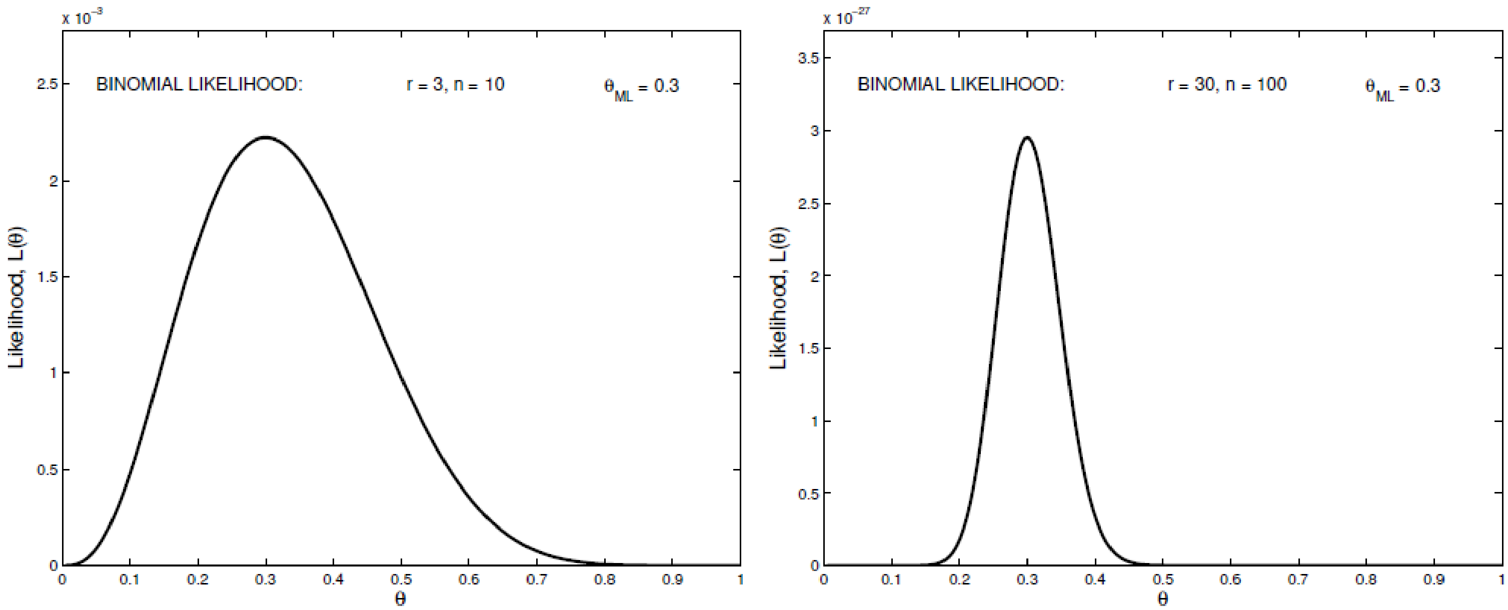


Figure 1: Binomial likelihood for (a) $r = 3, n = 10$, and (b) $r = 30, n = 100$.

Maximum Likelihood Method: Binomial

We can easily find the maximum likelihood estimate of θ

$$\log L(\theta) \;=\; l(\theta) \;=\; r\log\theta + (n-r)\log(1-\theta).$$

$$\frac{d}{d\theta}l(\theta) \;=\; \frac{r}{\theta} - \frac{n-r}{1-\theta} = 0, \qquad \text{at } \theta = \hat{\theta}_{ML}$$

$$\hat{\theta}_{ML} \;=\; \frac{r}{n}$$

Maximum Likelihood Method:Gaussian

Example

- **Let f(x,α) be given by a Gaussian distribution**
- Let α=μ be the mean of the Gaussian
- We want the best estimate of α from our set of n measurements.
- Let's assume that σ is the same for each measurement

$$f(x_i,\alpha) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\alpha)^2}{2\sigma^2}}$$

Likelihood function for this problem is:

$$L = \prod_{i=1}^{n} f(x_i,\alpha) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\alpha)^2}{2\sigma^2}} = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^n e^{-\frac{(x_1-\alpha)^2}{2\sigma^2}} e^{-\frac{(x_2-\alpha)^2}{2\sigma^2}} \cdots e^{-\frac{(x_n-\alpha)^2}{2\sigma^2}} = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^n e^{-\sum_{i=1}^{n}\frac{(x_i-\alpha)^2}{2\sigma^2}}$$

Find α maximizes the log likelihood function

Maximum Likelihood Method:Gaussian

$$\frac{\partial \ln L}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left[n \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(x_i - \alpha)^2}{2\sigma^2} \right] = 0$$

$$\frac{\partial}{\partial \alpha} \sum_{i=1}^{n} (x_i - \alpha)^2 = 0$$

$$\sum_{i=1}^{n} 2(x_i - \alpha)(-1) = 0$$

$$\sum_{i=1}^{n} x_i = n\alpha$$

$$\alpha = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Average !

- If σ are different for each data point
  - α is just the weighted average

$$\alpha = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}$$

Maximum Likelihood Method:Poisson

Example

- Let $f(x,\alpha)$ be given by a Poisson distribution
- Let $\alpha=\mu$ be the mean of the Poisson
- We want the best estimate of α from our set of n measurements $\{x_1,x_2,..x_n\}$
- Likelihood function in this case

$$L=\prod_{i=1}^{n} f(x_i,\alpha)=\prod_{i=1}^{n}\frac{e^{-\alpha}\alpha^{x_i}}{x_i!}=\frac{e^{-\alpha}\alpha^{x_1}}{x_1!}\frac{e^{-\alpha}\alpha^{x_2}}{x_2!}\cdots\frac{e^{-\alpha}\alpha^{x_n}}{x_n!}=\frac{e^{-n\alpha}\alpha^{\sum_{i=1}^{n}x_i}}{x_1!x_2!..x_n!}$$

Find α maximizes the log likelihood function

$$\frac{d\ln L}{d\alpha}=\frac{d}{d\alpha}\left(-n\alpha+\ln\alpha\cdot\sum_{i=1}^{n}x_i-\ln(x_1!x_2!..x_n!)\right)=-n+\frac{1}{\alpha}\sum_{i=1}^{n}x_i=0$$

$$\alpha=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$\boxed{\text{Average !}}$

Example: Maximum Likelihood Method

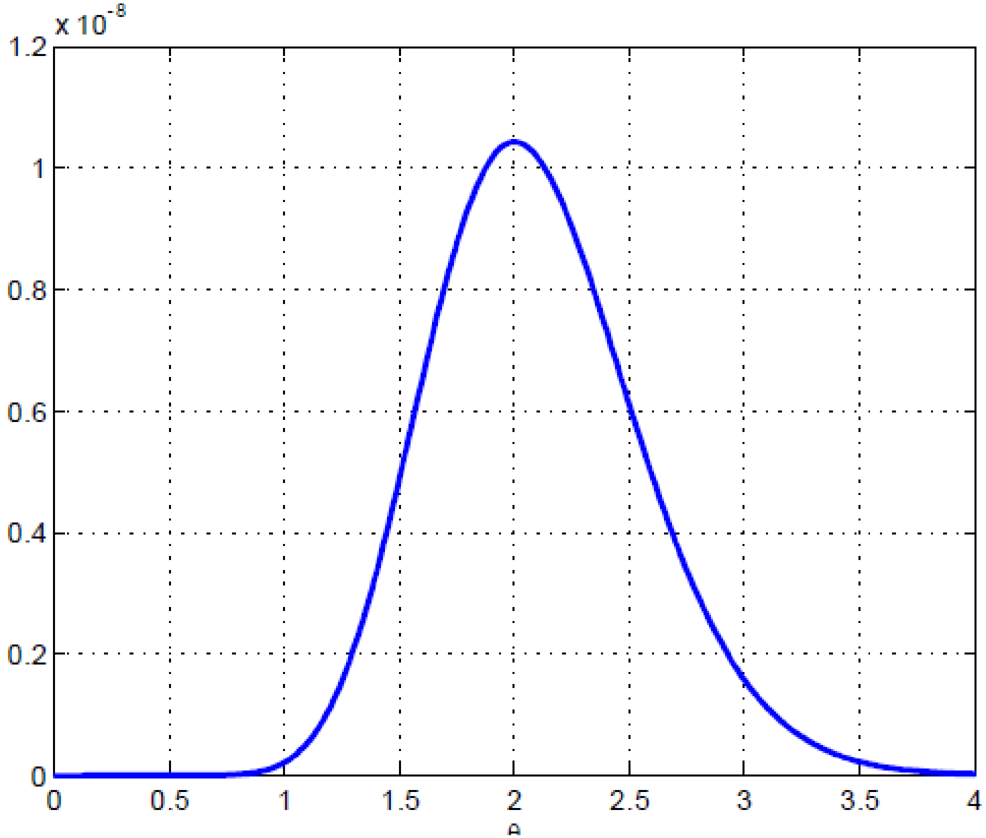
Example

- Let us asume that for n =10, we have a set measurements {5,0,1,1,0,3,2,3,4,1}

Then $$L_n(\alpha; x_1, x_2, x_3, \dots, x_{10}) = \frac{e^{-10\alpha}\alpha^{20}}{207{,}360}$$

What value of α would make this sample most probable?

Example: Maximum Likelihood Method



This Figure plots the function L_N (α;x) for various values of α. It has a single mode at α= 2, which would be the maximum likelihood estimate, or MLE, of α.

Example: Exponential Decay

Consider exponential pdf

$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$$

Independent measurements drawn this distribution: $t_1, t_2, \ldots t_n$

Likelihood function

$$L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$$

L(τ) is maximum when ln L(τ) is maximum

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

Example: Exponential Decay

Find maximum:

$$\frac{\partial \ln L(\tau)}{\partial \tau} = 0 \quad \rightsquigarrow \quad \sum_{i=1}^{n}\left(-\frac{1}{\tau} + \frac{t_i}{\tau^2}\right) = 0 \quad \rightsquigarrow \quad \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

Variance of the estimated decay time

$$\frac{\partial^2 \ln L(\tau)}{\partial^2 \tau} = \sum_{i=1}^{n}\left(\frac{1}{\tau^2} - 2\frac{t_i}{\tau^3}\right) = \frac{n}{\tau^2} - \frac{2}{\tau^3}\sum_{i=1}^{n} t_i = \frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)$$

$$V[\hat{\tau}] = -\left(\frac{\partial^2 \ln L}{\partial^2 \theta}\right)^{-1}_{\tau=\hat{\tau}} = \frac{\hat{\tau}^2}{n} \qquad \rightsquigarrow \qquad \hat{\sigma} = \frac{\hat{\tau}}{\sqrt{n}}$$

Example: Exponential Function

Suppose that the lifetime of light bulbs is modeled by an exponential distribution with(unknown) parameter λ We test 5 bulbs and find they have lifetimes of 2,3,1,3,and 4 years, respectively. What is the MLE for λ (= $1/\tau$) ?

Let X_i be the life time of the i^{th} bulb and let x_i be the value X_i takes.Then each X_i has pdf $F_{Xi}(x_i) = \lambda e^{-\lambda xi}$ We assume the lifetimes of the bulbs are independent, so the joint pdf is the product of then dividual densities:

$$f(x_1, x_2, x_3, x_4, x_5 \,|\, \lambda) = (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})(\lambda e^{-\lambda x_3})(\lambda e^{-\lambda x_4})(\lambda e^{-\lambda x_5}) = \lambda^5 e^{-\lambda(x_1+x_2+x_3+x_4+x_5)}$$

and our data has values

$$x_1 = 2,\ x_2 = 3,\ x_3 = 1,\ x_4 = 3,\ x_5 = 4.$$

So the likelihood and *log* likelihood functions with this data are

$$f(2,3,1,3,4 \,|\, \lambda) = \lambda^5 e^{-13\lambda}, \quad \ln(f(2,3,1,3,4 \,|\, \lambda) = 5\ln(\lambda) - 13\lambda$$

Example: Exponential Function

Finally we find the MLE

$$\frac{d}{d\lambda}(\text{log likelihood}) = \frac{5}{\lambda} - 13 = 0 \ \Rightarrow \ \boxed{\hat{\lambda} = \frac{5}{13}}.$$

Extended MLM

- Often we want to do a MLM fit to determine the number of a signal & background events. Let's assume we know the pdfs that describe the signal (p_s) and background (p_b) and the pdfs depend on some measured quantity x (e.g. energy, momentum, angle..) We can write the Likelihood for a single event (i) as:

$$L = \prod_{i=1}^{N} (f_s p_s(x_i) + (1 - f_s) p_b(x_i))$$

There are several drawbacks to this solution:

1) The number of signal and background are 100% correlated.

2) the (poisson) fluctuations in the number of events (N) is not taken into account

Extended MLM

Another solution which explicitly takes into account 2) is the EXTENDED MLM:

$$L = \frac{e^{-v} v^N}{N!} \prod_{i=1}^{N} (f_s p_s(x_i) + (1 - f_s) p_b(x_i)) = \frac{e^{-v}}{N!} \prod_{i=1}^{N} v(f_s p_s(x_i) + (1 - f_s) p_b(x_i))$$

$$\ln L = -v - \ln N! + \sum_{i=1}^{N} \ln[v(f_s p_s(x_i) + (1 - f_s) p_b(x_i))]$$

Here $v=N_s+N_b$ so we can re-write the likelihood function as:

$$\ln L = -(N_s + N_b) - \ln N! + \sum_{i=1}^{N} (\ln[N_s p_s(x_i) + N_b p_b(x_i)])$$

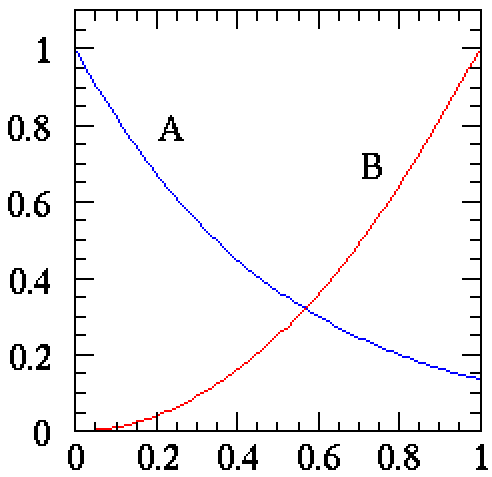
If N_s & N_b are poisson then so is their product for fixed N

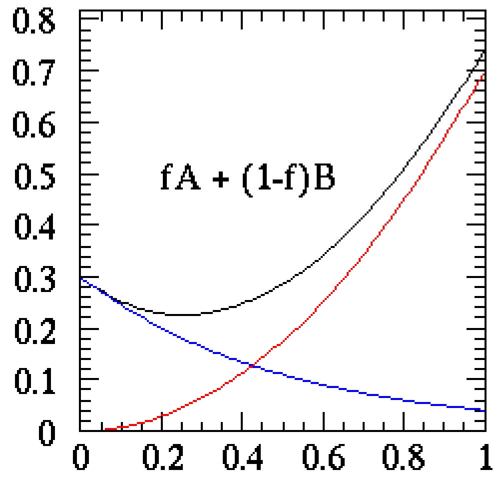
We maximize L in terms of N_s and N_b.

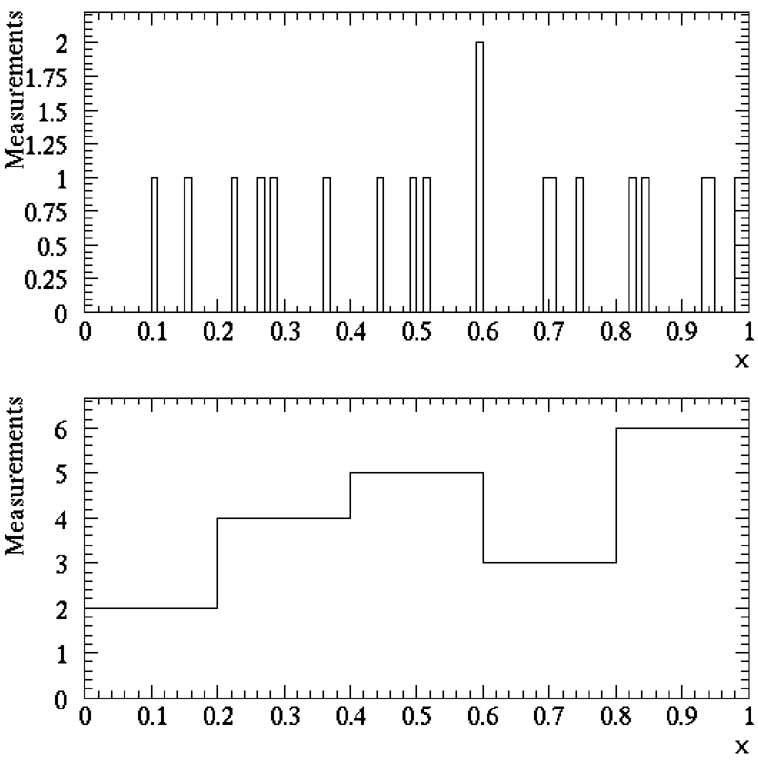
The N! term drops out when we take derivatives to max L.

Simple example of an ML estimator

- Suppose that our data sample is drawn from two different distributions. We know the shapes of the two distributions, but not what fraction of our population comes from distribution A vs. B. We have 20 random measurements of X from the population.







$$P_A(x) = \frac{2}{1 - e^{-2}} e^{-2x} \qquad P_B(x) = 3x^2$$

$$P_{tot}(x) = f\,P_A(x) + (1 - f)\,P_B(x)$$

Form for the log likelihood and the ML estimator

- Suppose that our data sample is drawn from two different distributions. We know the shapes of the two distributions, but not what fraction of our population comes from distribution A vs. B. We have 20 random measurements of X from the population.

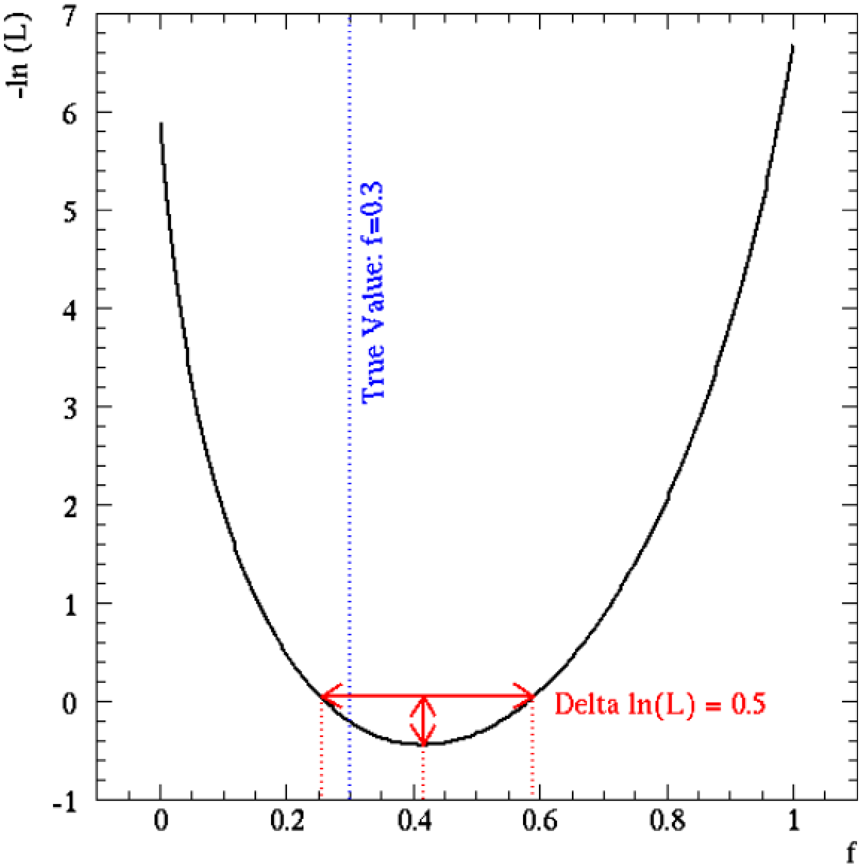
$$P_{tot}(x) = f\,P_A(x) + (1-f)\,P_B(x)$$

Form the negative log likelihood:

$$-\ln L(f) = -\sum_{i=1}^{N} \ln\left(P_{tot}(x_i|f)\right)$$

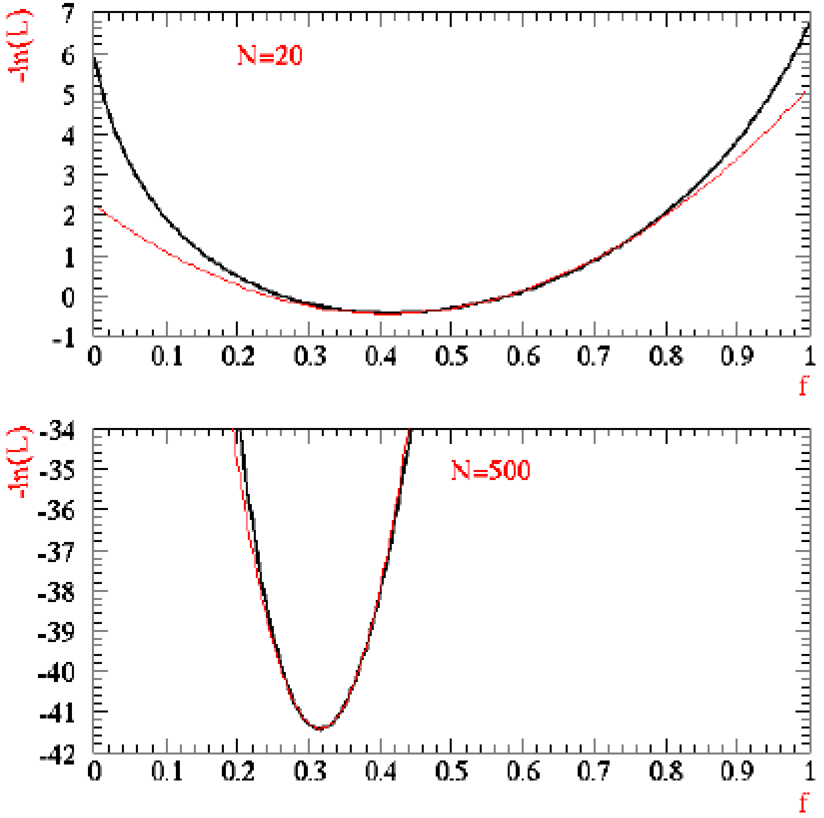
Minimize(Maximize) -ln(L) (ln(L)) with respect to f. Sometimes you can solve this analytically by setting the derivative equal to zero. More often you have to do it numerically.

log likelihood



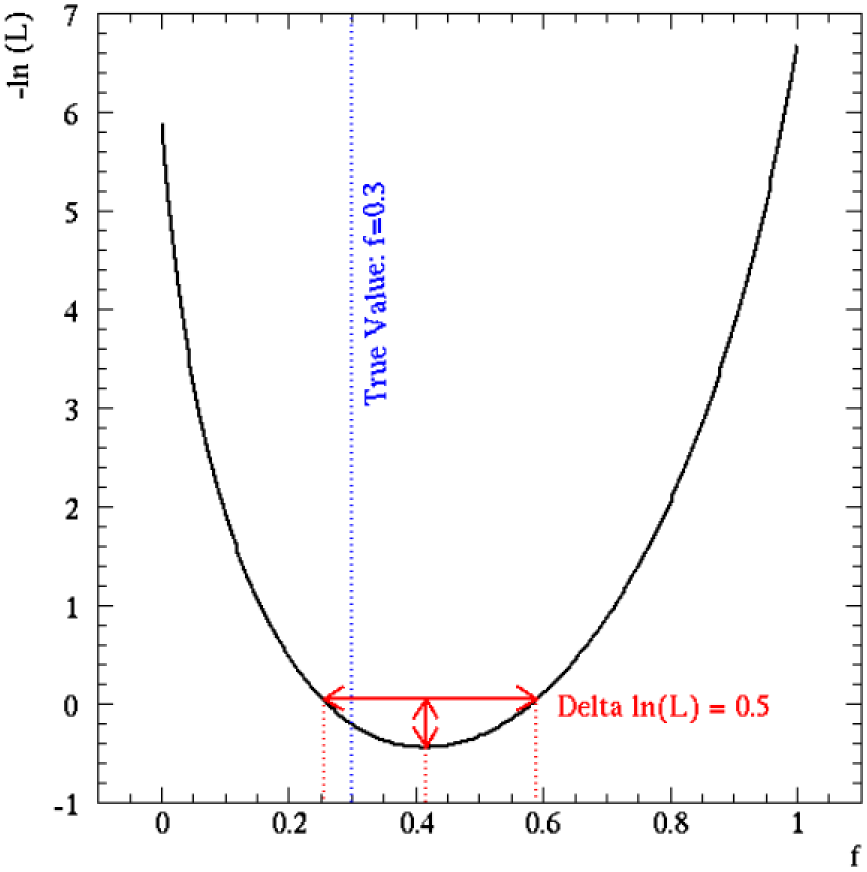
- The minimum is at f = 0.415 which is the ML estimate
- 1σ error range is defined by Δln(L) = ½ above the minimum.
- The set was actually drawn from a distribution with a true value of f = 0.3

log likelihood



- In general the log likelihood becomes more parabolic as N gets larger. The graphs at the right show the negative log likehoods for our exampleproblem for N=20 and N =500. The red curves are parabolic fits around minimum.

log likelihood



- Even when the log likelihood is not Gaussian, its nearly universal to define 1σ range by by Δln(L) = ½ This can result in asymmetric error bars such as

$$0.41^{+0.17}_{-0.15}$$

- Remember 1σ range would mean that the true value has 68% chance of being within that range.

In Class Exercise

The following data are the observed frequenices of occurrence of domestic accidents: we have n= 647 data as follows

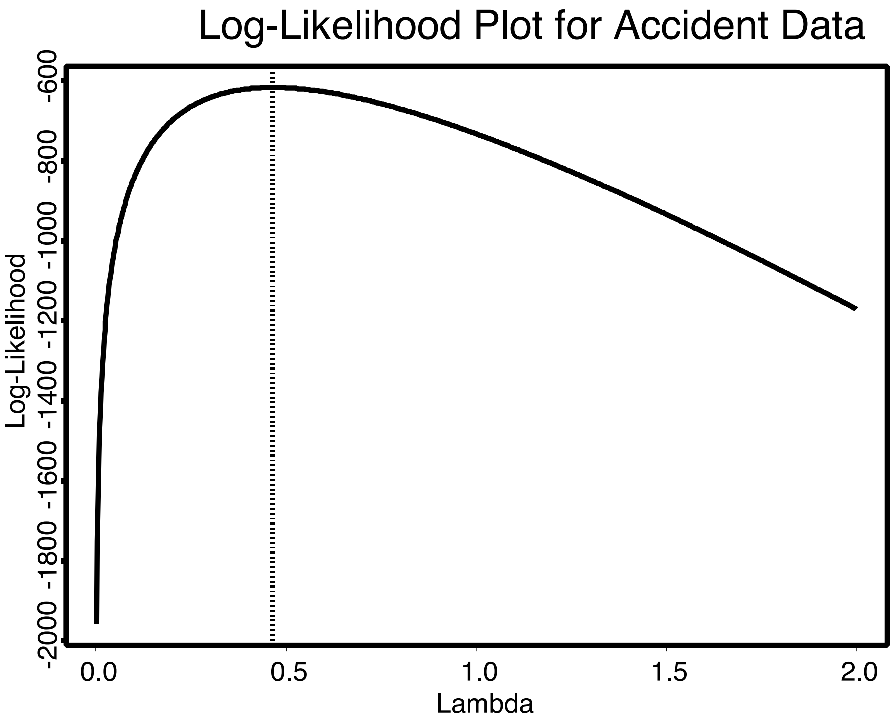
Number of accidents	Frequency
0	447
1	132
2	42
3	21
4	3
5	2

A Poisson model is assumed.

Write a python script to plot LogL versus λ and find estimate of λ_{ML} with error.
You may use only matplotlib and numpy

In class exercise

Write a python script to plot LogL versus λ and find estimate of λ_{ML} with error



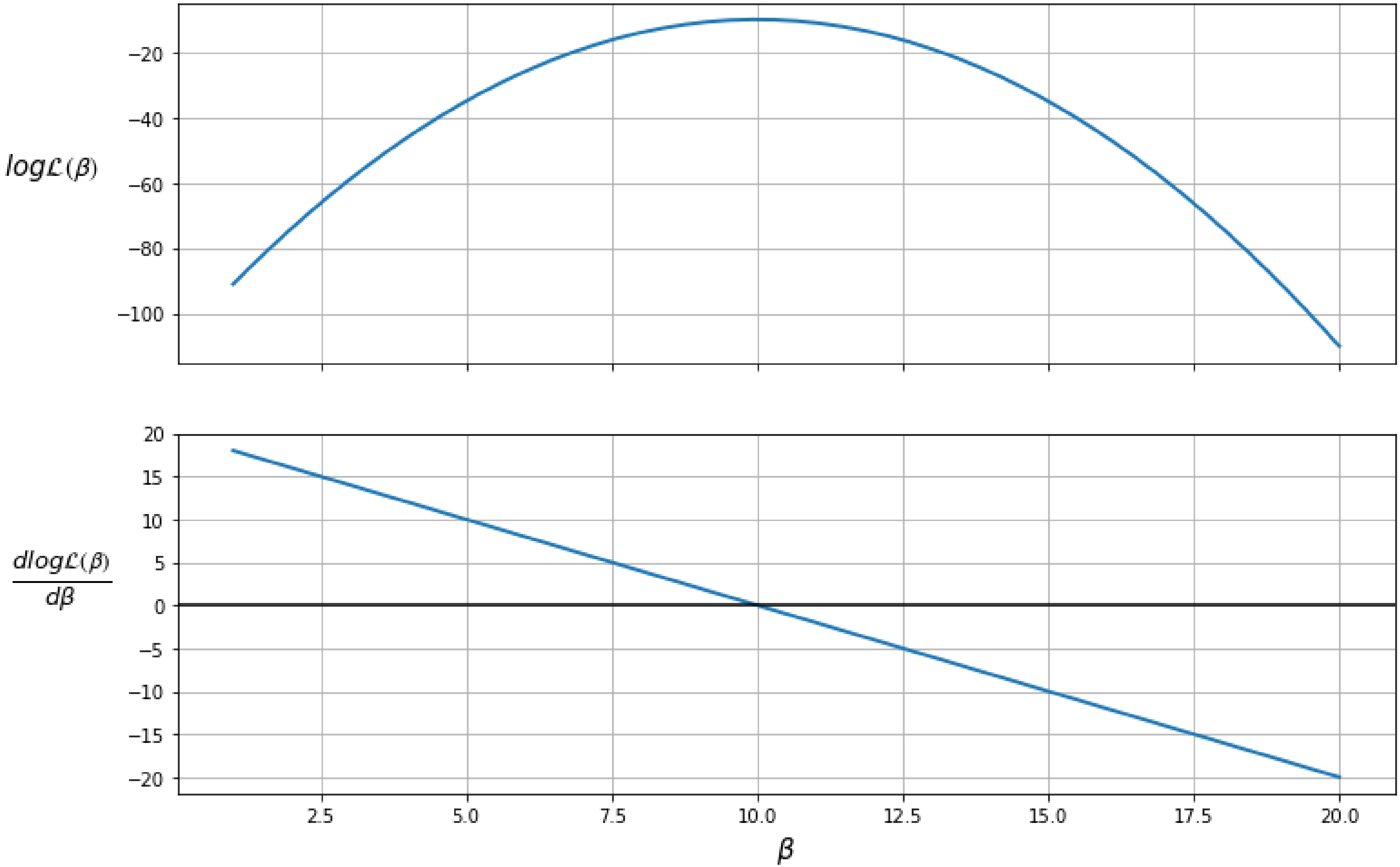
$\ln L = \ln L_{max} - 1/2$ ("1σ points")

Back up

Example: $\log L(\beta) = -(\beta - 10)^2 - 10$

```
import matplotlib.pyplot as plt
import numpy as np

p = np.linspace(1, 20)
logL = -(p - 10) ** 2 - 10
dlogL = -2 * p + 20

fig, (ax1, ax2) = plt.subplots(2, sharex=True, figsize=(12, 8))

ax1.plot(p, logL, lw=2)
ax2.plot(p, dlogL, lw=2)

ax1.set_ylabel(r'$log \mathcal{L(\beta)}$', rotation=0, labelpad=35, fontsize=15)
ax2.set_ylabel(r'$\frac{dlog \mathcal{L(\beta)}}{d \beta}$ ', rotation=0, labelpad=35,
           fontsize=19)
ax2.set_xlabel(r'$\beta$', fontsize=15)
ax1.grid(), ax2.grid()
plt.axhline(c='black')
plt.show()
```

Example



−20
−40
−60
−80
−100
$log\mathcal{L}(\beta)$
20
15
10
5
0
−5
−10
−15
−20
$\frac{dlog\mathcal{L}(\beta)}{d\beta}$
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
β

MLM signal/background

- Often we want to do a MLM fit to determine the number of a signal & background events. Let's assume we know the pdfs that describe the signal (p_s) and background (p_b) and the pdfs depend on some measured quantity x (e.g. energy, momentum, angle..) We can write the Likelihood for a single event (i) as:

$$L = f_s p_s(x_i) + (1 - f_s) p_b(x_i)$$

with f_s the fraction of signal events in the sample, and the number of signal events: $N_s = f_s N$

The likelihood function to maximize (with respect to f_s) is:

$$L = \prod_{i=1}^{N} (f_s p_s(x_i) + (1 - f_s) p_b(x_i))$$