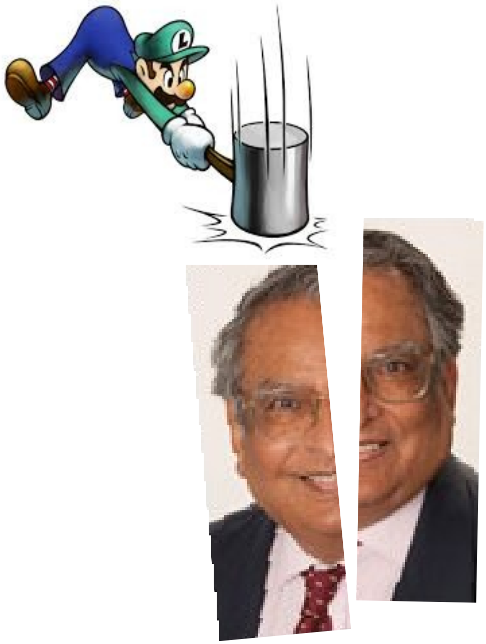
Advanced Statistical Methods

Lecture 4

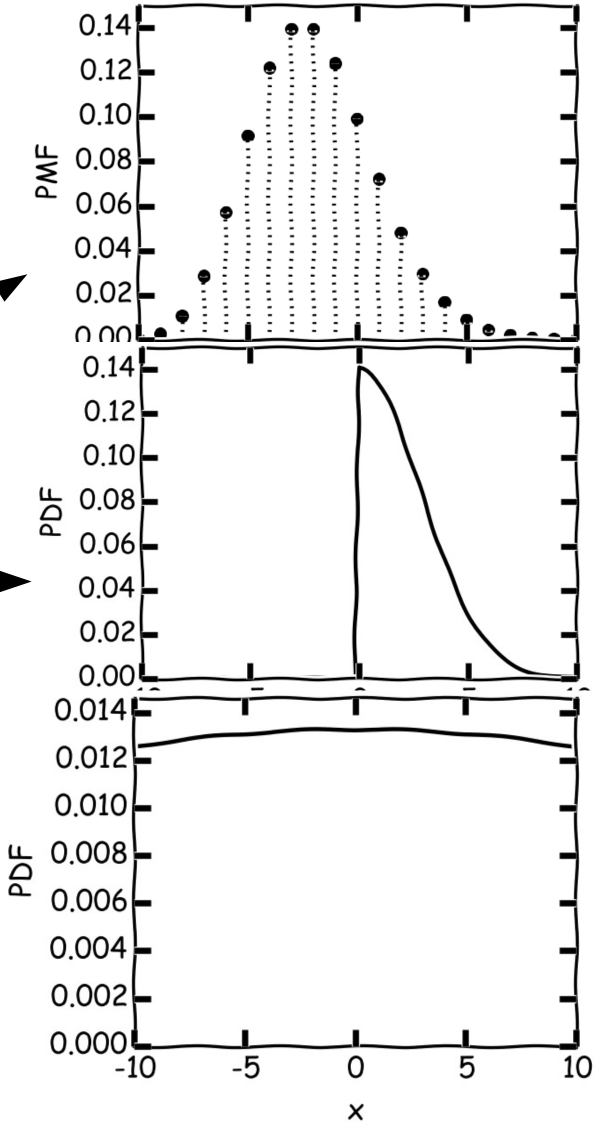
Breaking Wilks' theorem

Situation:
Maximum likelihood ratios can be approximated by a chi-squared distribution **if the MLEs of the model parameters are normal distributed.** Normal distribution of the data is often helpful, but not necessary.

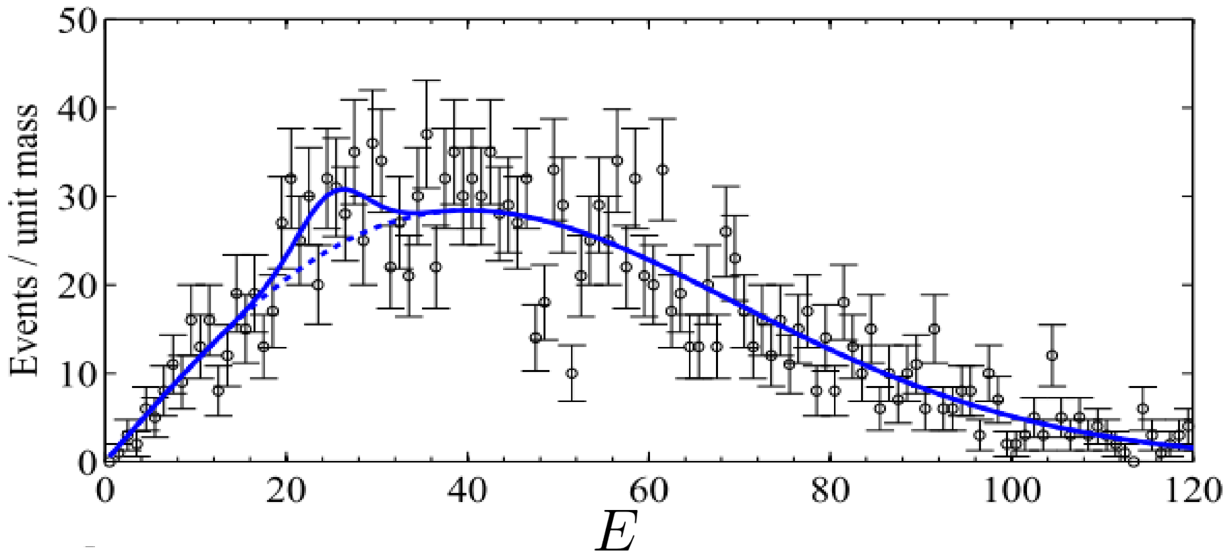


Ways to break Wilks' theorem

- Low number of events, Poisson noise
- Bounds on the model parameters (e.g. fluxes or masses can only be positive)
- Parameters that are irrelevant in the null hypothesis and have an (approximately) infinite variance



Example: Line searches



A very common scenario (X-ray and gamma-ray lines, Higgs bosons etc)

- Signal is "line" (narrow normal distribution with fixed width) with unknown strength and position
- Model for the BG depends on parameters

$$\frac{dN}{dE} = A_s \cdot N(E|\bar{E}, \Delta E) + bg(E|\zeta)$$

- The correct likelihood function is usually build up from Poisson distributions

$$\mathcal{L}(A_s, \bar{E}, \zeta|\vec{c}) = \prod_{i=1}^{n_{\rm bins}} P(c_i|\mu_i(A_s, \bar{E}, \zeta))$$

with expectation values $\mu_i(A_s, \bar{E}, \zeta) = \int_{E_i^-}^{E_i^+} dE' \frac{dN}{dE'}$ in energy range $[E_i^-, E_i^+]$

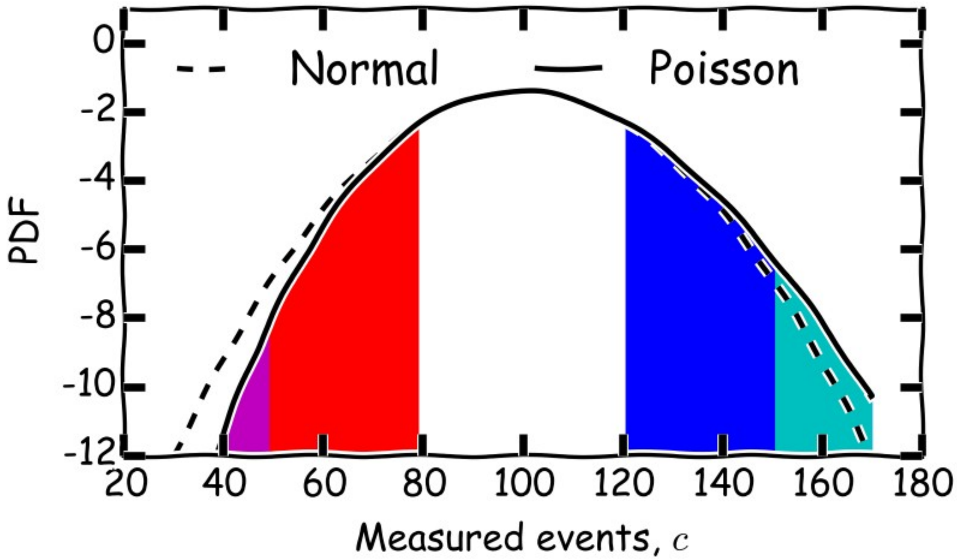
Low number of events

If the signal region contains only a very low number of events, the MLE for the signal will be Poisson distributed instead of normal distributed, and Wilks' theorem breaks down.

$$\mathcal{L}(\mu = b + \theta | c) = \frac{e^{-\theta - b}(\theta + b)^c}{c!}$$

$$\hat{\theta} = \max(0, c - b)$$

$\langle c \rangle = var(c) = 100$

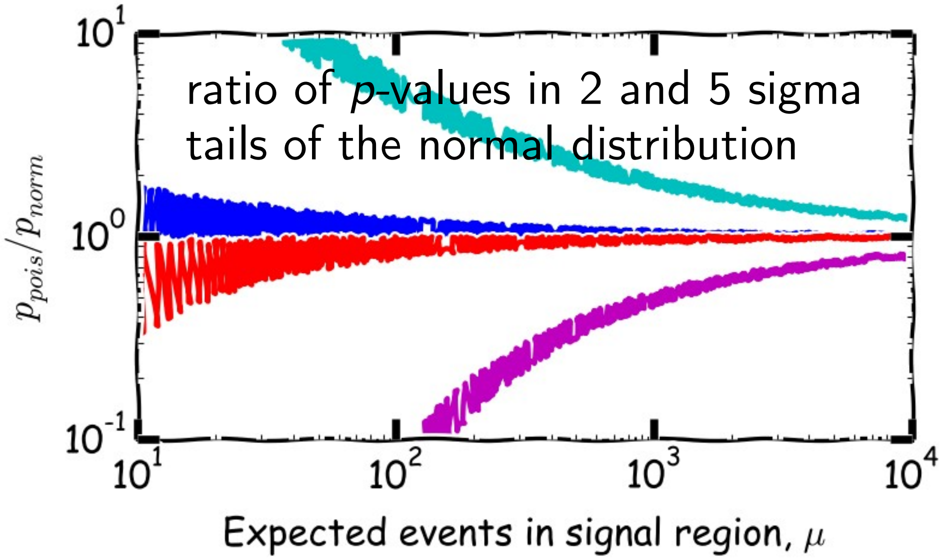


Notes:

- If the number of expected signal and background events is large enough, this can be reasonably well approximated by a normal distribution.

- What "reasonable large" means depends on the context:
  - For 2sigma confidence regions: around 100
  - For 5sigma discoveries: around 1000

- *If in doubt, double-check your results using Poisson statistics!*
- The problem is here not the discreteness of the events, but the non-gaussian tails.
  → Anscombe transform.


ratio of *p*-values in 2 and 5 sigma tails of the normal distribution

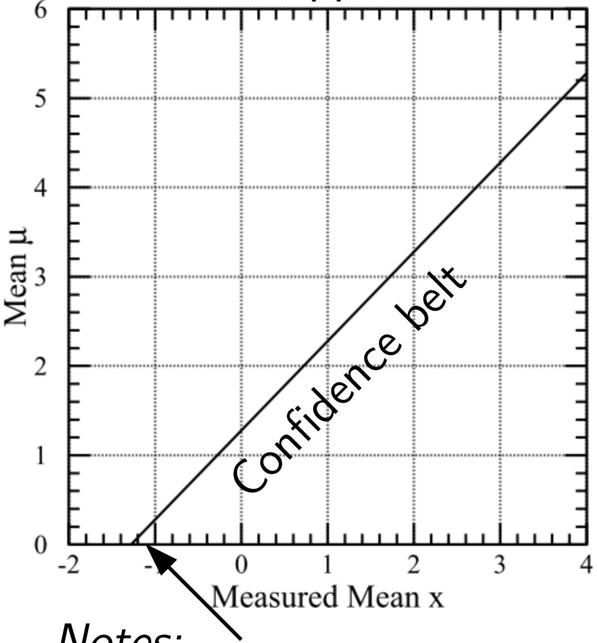
Flip-flopping: A problem close to thresholds

Let us suppose that a **Physicist *X*** takes the following attitude:

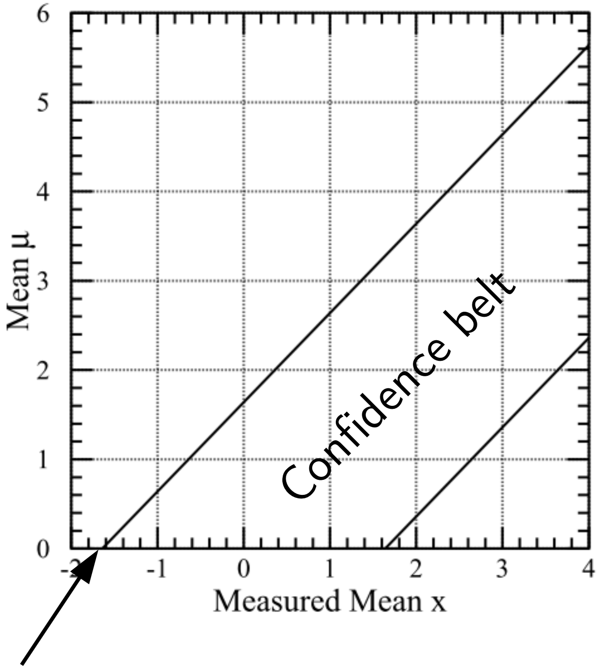
- If the result *x* is less than 3sigma, I will state an upper limit following standard procedures.
- If the result *x* is greater than 3sigma, I will state a central confidence interval.
- If the measured quantity (e.g. a flux) is below zero, I will pretend it is zero and quote the corresponding limit.

This policy is called ***flip-flopping*** and leads to wrong confidence intervals.

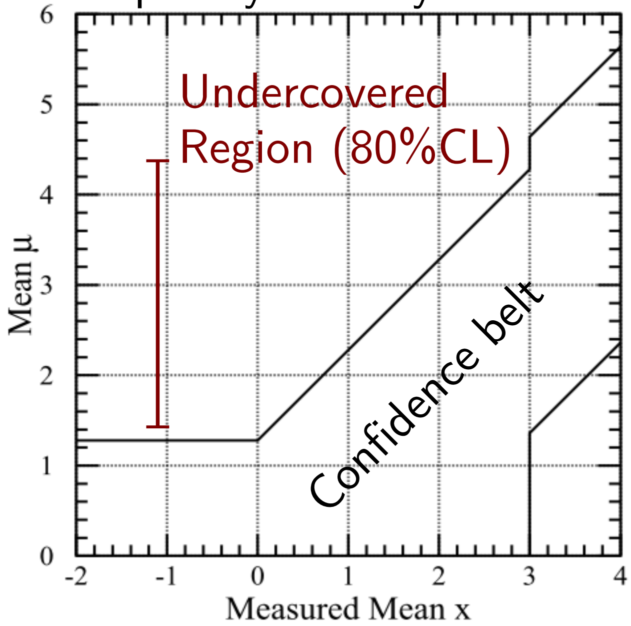
Proper
90% CL upper limits



Proper
90% CL central interval



Confidence interval
implicitly used by Mr. X



Notes:

- Both the upper limit and the standard central interval return an *empty* confidence interval for a sufficiently negative value of *x*. This is an example for a technically correct, but not useful construction of confidence intervals.

Feldman Cousins approach

One important application of the above discussion are **Poisson processes** with known background:

- Let us consider a process with known background *b*, and unknown signal mean μ. The number of measured events is here *n* and follows a Poisson distribution
$$P(n|\mu + b)$$
- Again, if *n* is much lower than *b*, the confidence interval will be empty for standard upper limits or central intervals.

Feldman-Cousins construction

- Feldman-Cousins propose to define confidence belt neither as upper limits nor as central intervals, but instead as the region with the largest *maximum likelihood ratio*

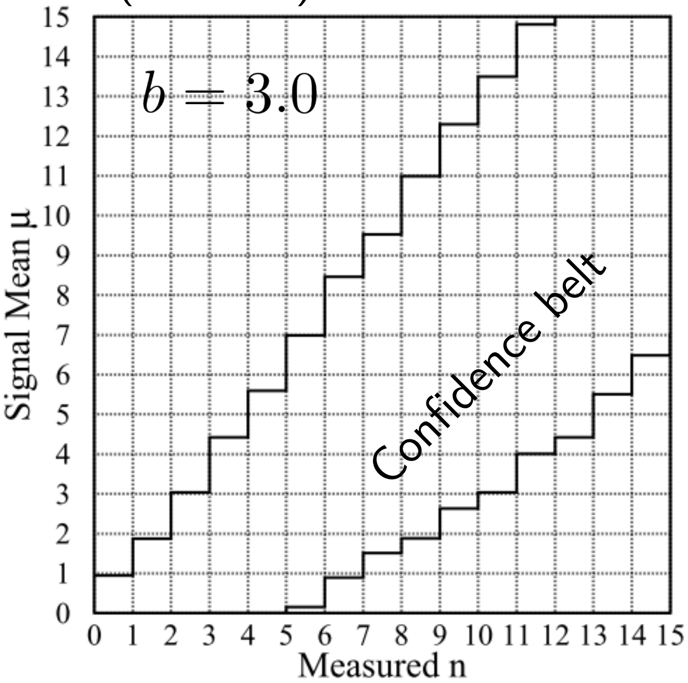
$$R(n) = \frac{P(n|\mu + b)}{P(n|\hat{\mu} + b)} \quad \hat{\mu} = \max(0, n - b) \text{ (MLE)}$$

- For each value of μ, the belt is then constructed by adding the *n* values that correspond to the largest *R(n)*, until the band covers at least 90% of the Poisson PDF.
- Example for *b=3.0* and μ *= 0.0*:

$$R(0) = \cdots = R(3) = 1$$
$$R(4) = 0.86$$
$$R(5) = 0.57$$

$$\sum_{n=0}^{5} P(n|3) = 0.92$$

Feldman-Cousins intervals (90% CL)



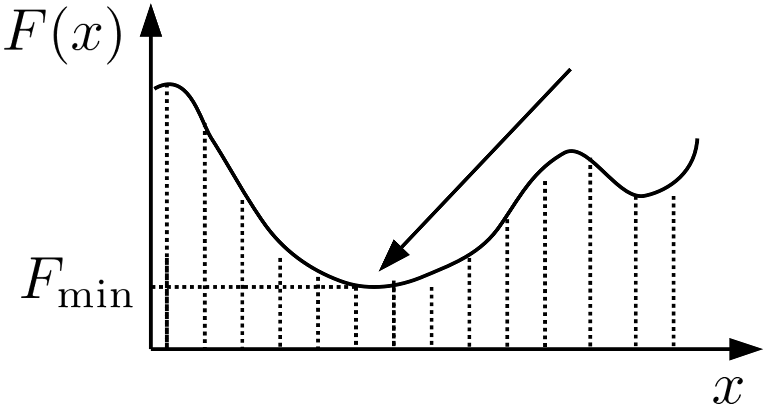
Numerical minimizers

General statements

- Minimizing (or, equivalently, maximizing) is one or the numerical *core challenges* of Frequentist statistics

- General goals
  - Evaluating the likelihood function for a single set of model parameters ("point") can be very time consuming. The number of likelihood evaluations should be reduced to a minimum.
  - In most cases, we are interested in the *global* minimum, not the *local*

- Most straightforward way is to evaluate function on a grid, and take minimum or maximum

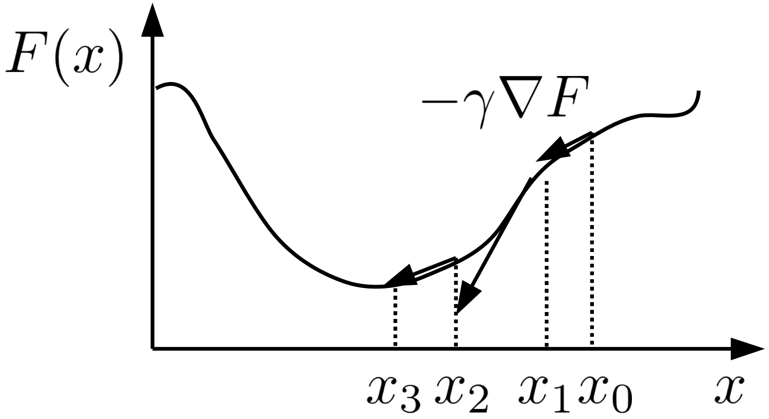


This works well in one or two dimensions, but becomes impossible if the number of model parameters is large (say, above three), since $\propto n_{\rm grid}^{n_{\rm dim}}$

Simplest method

Gradient decent

- Iterative procedure to find minimum of multi-dimensional function
- The direction and size of the is given by the *gradient* of the function
- Step size can be regulated with one free parameter, *gamma.*



Algorithm

- Select starting point close to minimum of interest
- Iterate following the rule

$$1 \geq \gamma > 0$$

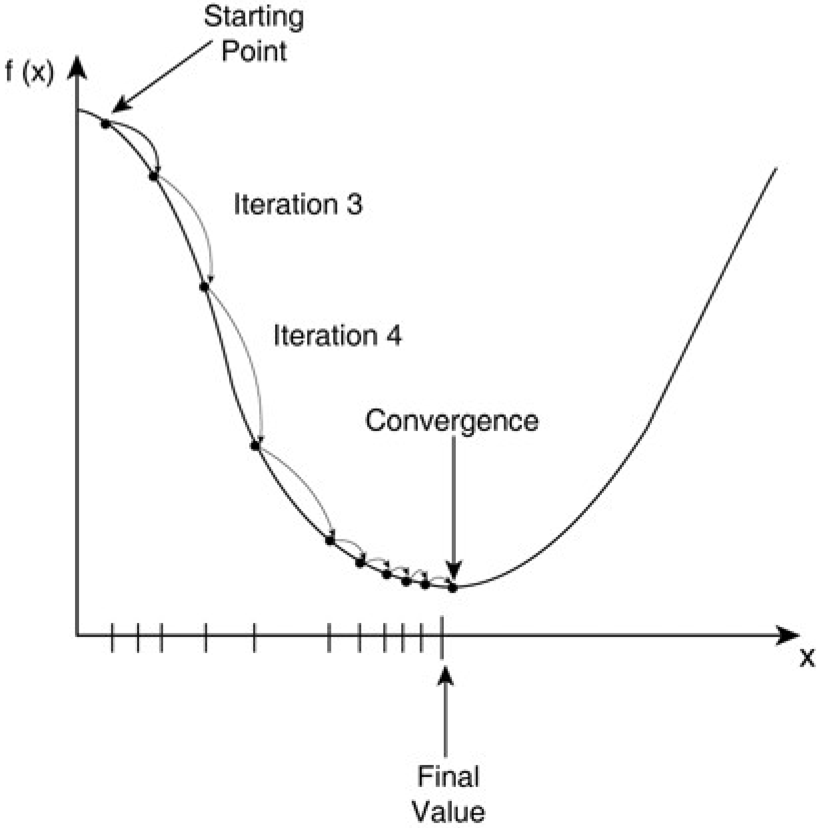
$$\vec{x}_{i+1} = \vec{x}_i - \gamma\nabla F(\vec{x}_i)$$

- Gamma should be selected small enough to not *overshoot* the minimum

$$F(x_{i+1}) \leq F(x_i)$$

- Stop minimization if gradient close enough to zero

$$|\nabla F(\vec{x}_i)| < \epsilon$$

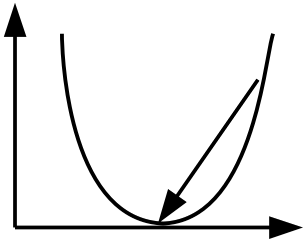


Newton method

Motivation

- *Problem with gradient decent*: The gradient only provides information about direction towards minimum, not the distance.
- **Newton methods** try to estimate the *distance to the minimum* by using the <u>first</u> and the <u>second</u> derivative of the function.

Example in one dimensions



- In one dimensions, Taylor expanding the function up to second order yields

$$f(x+\Delta x) = f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2$$

- Using this Taylor expansion, we estimate the distance to the minimum by

$$\frac{\partial f(x+\Delta x)}{\partial \Delta x} = f'(x) + f''(x)\Delta x \overset{!}{=} 0 \quad \Rightarrow \quad \Delta x = -\gamma \frac{f'(x)}{f''(x)}$$

- This gives an estimate for the step size (regulated by $1 \geq \gamma > 0$)

Newton method

- In multiple dimensions, the iteration step of the Newton method is given by

$$\vec{x}_{i+1} = \vec{x}_i - \gamma [H(x_i)]^{-1} \nabla F(x_i)$$

where *H* denotes the Hesse matrix $(H)_{kl} = \dfrac{\partial^2 F}{\partial x_k \partial x_l}$.

Gauss-Newton Method

Motivation

- *Problem with general Newton method*: Calculation of second derivative (Hesse matrix) is computationally very expensive.
- The score function has in many applications a specific form, that can be exploited when calculating derivative.

Gauss-Newton method

- This method requires the score function to be of quadratic form

$$F(\vec{\beta}) = \sum_{i=1}^{m} r_i(\vec{\beta})^2 \qquad r_i(\vec{\beta}) = c_i - \mu_i^1 \beta_1 - \mu_i^2 \beta_2$$

- In that case, the Hesse matrix gets contributions from the first and second derivate of *ri* w.r.t. the model parameters. If ri is approximately linear in beta, one can approximate the Hesse

$$(H)_{kl} = 2 \sum_{i=1}^{m} \left(\frac{\partial r_i}{\partial \beta_k} \frac{\partial r_i}{\partial \beta_l} + r_i \cancelto{0}{\frac{\partial^2 r_i}{\partial \beta_l \partial \beta_k}} \right)$$

- In some cases, it is useful to reduce the step size by using the **Levenberg-Marquardt damping factor**

$$H \to H + \lambda \mathrm{diag}(H)$$

Quasi-Newton Methods

Motivation

- *Problem with general Newton method*: Calculation of second derivative (Hesse matrix) is computationally very expensive.
- One can use *estimates* for the Hesse matrix that are updated iteratively

General idea

- The *i^th^* step of the iteration is given by $\vec{x}_{i+1} = \vec{x}_i + \Delta\vec{x}_i$. The step is calculated from an *approximated* Hesse matrix,

$$\Delta\vec{x}_i = \gamma[\hat{H}(x_i)]^{-1}\nabla F(x_i)$$

- Taylor expanding the *gradient* of the function at the old and new point yields

$$\nabla F(\vec{x}_i + \Delta\vec{x}_i) = \nabla F(\vec{x}_i) + H_{i+1}\Delta\vec{x}_i + \mathcal{O}(\Delta\vec{x}^2)$$

  If we truncate that expansion at first order in *x*, this gives a *constraint* on the i+1th Hesse matrix.

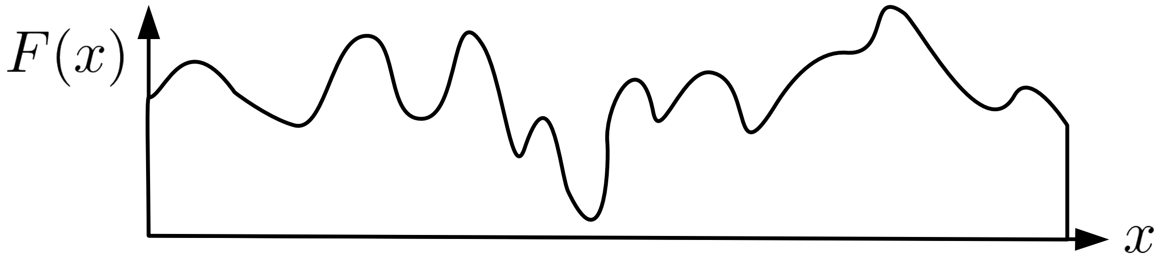
- Using this constraint and the ith Hesse matrix, one can define an *updated* i+1th approximated Hesse matrix $\hat{H}_{i+1}$. This new approximated Hesse matrix is then used in the next iteration.

- There are many different update rules available: DFP, BFGS, Broyden, SR1

Simulated Annealing

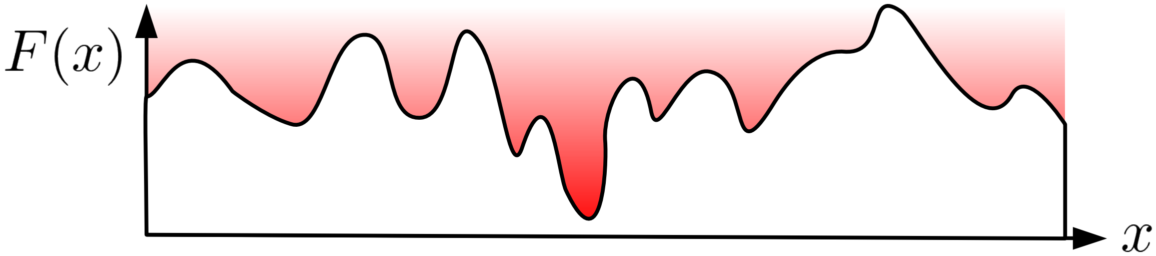
Motivation

- *Problem of all above minimizers*: They are optimized to find *local* minima.

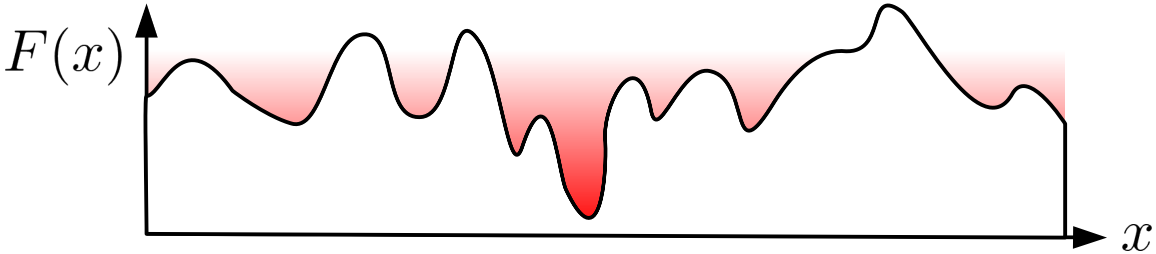


- This can be solved by introducing the concept of thermodynamic noise.

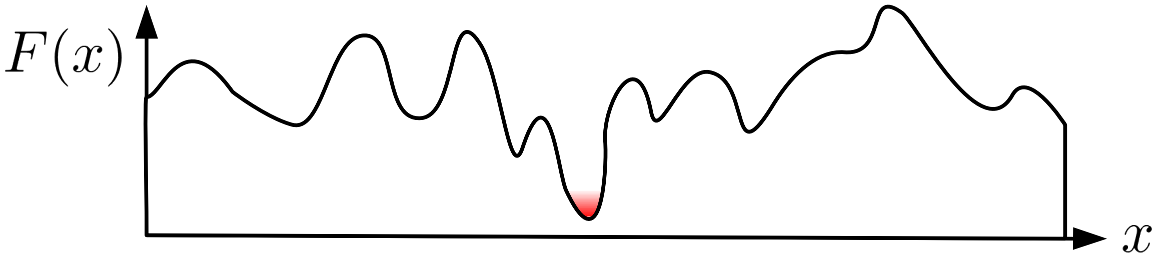
Imagine that function *F(x)* describes energy states in a system. At high temperature *T*, most of the states can be accessed.



If the temperature drops, only lower states in the system are occupied.

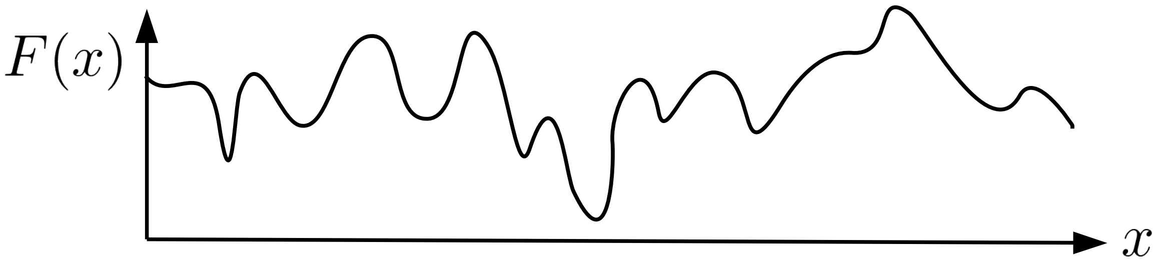


At minimum temperature, the system is in its minimum energy state.



Goal: Generate distribution of points $P(x) \propto e^{-F(x)/T}$ that follow excited states.

Simulated Annealing



Algorithm

1) Generate initial state x

2) Randomly pick new state according to proposal distribution $g(x \to x')$.

3) Accept state as new state with acceptance probability $A(x \to x')$, given by

$$A(x \to x') = \begin{cases} 1 & \text{if } F(x') \leq F(x) \\ e^{-F(x')/T} / e^{-F(x)/T} & \text{if } F(x') > F(x) \end{cases}$$

4) If the step is accepted, set $x = x'$ and save the new state in a list; else nothing happens.

5) Go back to step 2.

Convergence distribution

- After some time, the distribution of accepted points usually becomes stationary and follows the *detailed balance* criterion

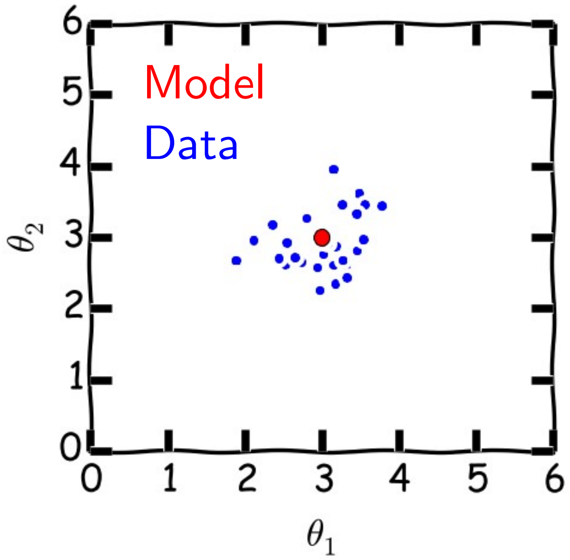
$$\pi(x \to x')P(x) = \pi(x' \to x)P(x'), \text{ where}$$
$$\pi(x \to x') = g(x \to x')A(x \to x')$$

- The convergence distribution is hence given by $$\frac{P(x)}{P(x')} = \frac{\exp(-F(x)/T)}{\exp(-F(x')/T)}$$

Frequentist vs Bayesian

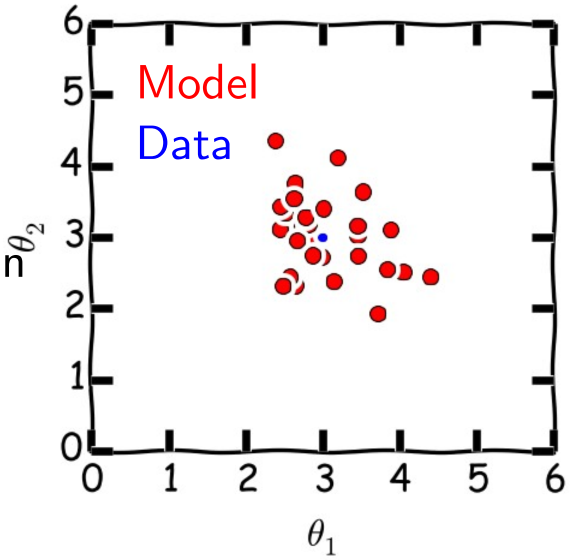
Frequentist

- Model is considered as fixed
- Measured values jump around true value



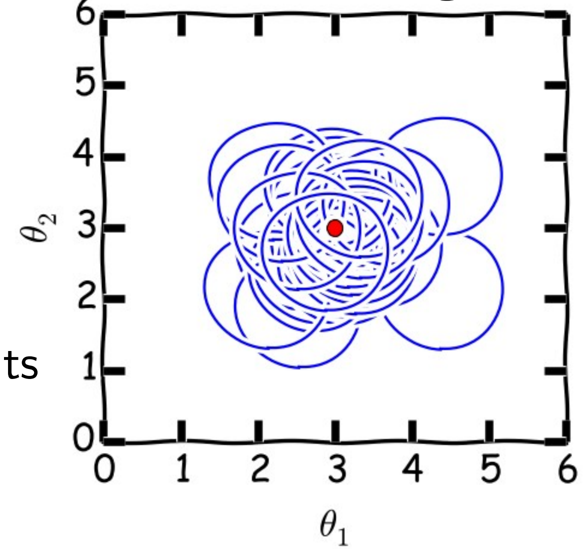
Bayesian

- Focus on consequences of single observation
- Focus on distribution or plausibility of models that could lead to that observation



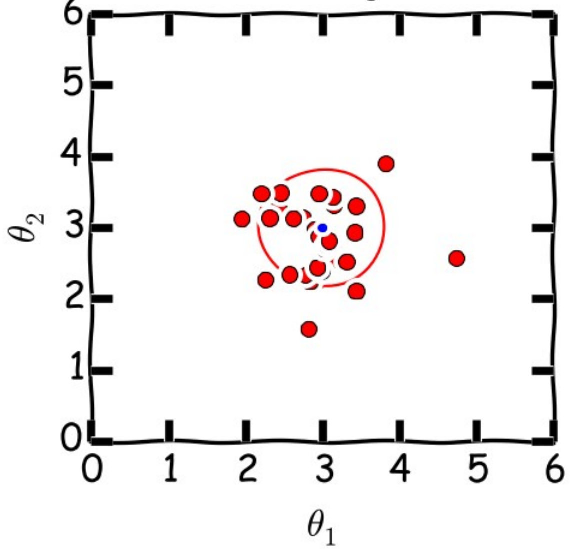
Confidence region

- Confidence regions are designed such that they *cover* the true value in a certain number of repeated experiments



Credible region

- A credible region contains a certain fraction of the models that most likely lead to the actual observation



Discussion up to now.

Discussion from now on.

Bayes' Theorem

Likelihood function

Posterior

Prior

$$P(H|D,I) = \frac{P(D|H,I) \cdot P(H,I)}{P(D|I)}$$

Global likelihood

It is a simple consequence of the rule for conditional probabilities:

$$P(X|Y,I) \cdot P(Y,I) = P(X,Y|I) = P(Y|X,I) \cdot P(X,I)$$

Notes:

- Bayes' theorem provides a rule for how to *update* the probability or plausibility of a certain hypothesis *H* to be true in light of data *D*. This always depends on additional background information *I*, which is often not made explicit.
- Frequentists are interested in likelihood functions *only*

$$\mathcal{L}(H|D,I) \propto P(D|H,I)$$

It is in general *not* equal to the posterior, which is most obvious looking at the normalization of the functions (with *x* and *θ* being data and model parameters, respectively).

$$\int d\theta P(\theta|x) = 1$$

Parameter estimation

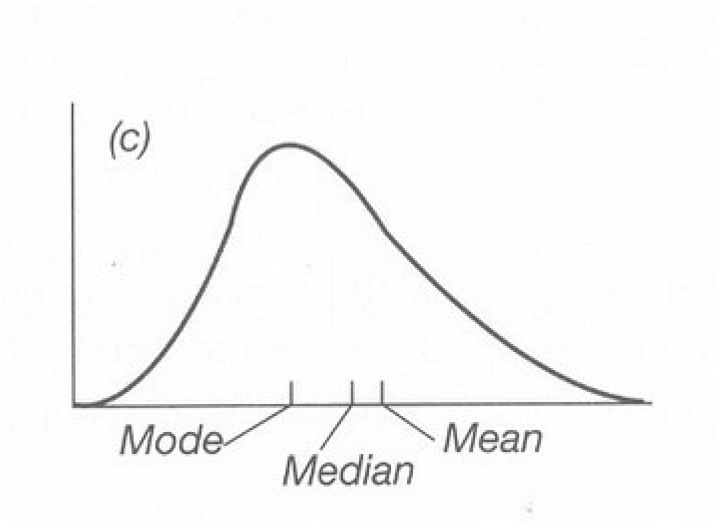
Bayesian Estimators

- The "posterior mean"

$$\langle\theta\rangle_\theta = \int d\theta\ \theta\ P(\theta|\vec{x}, M)$$

- The "posterior mode"

$$\theta_{\text{mode}} = \arg\max_\theta P(\theta|\vec{x}, M)$$



The posterior distribution:

- The posterior is obtained from Bayes' theorem (here for discrete hypotheses)

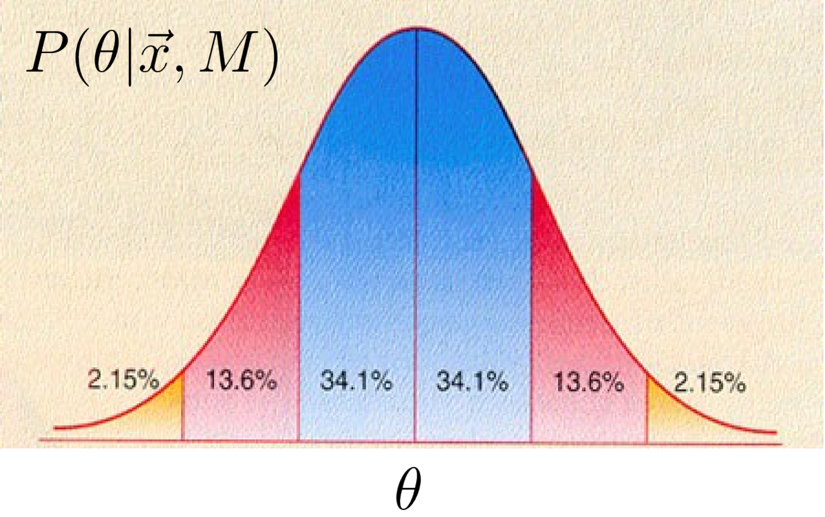
$$\underbrace{P(M_i|\vec{x}, I)}_{\textit{Posterior}} = \frac{\overbrace{P(\vec{x}|M_i, I)}^{\textit{Likelihood}} \cdot \overbrace{P(M_i|I)}^{\textit{Prior}}}{\underbrace{P(\vec{x}|I)}_{\textit{Global likelihood}}} \quad \text{with} \quad \sum_{i=1}^{n_{\text{models}}} P(M_i|\vec{x}, I) = 1$$

- In the case of composite hypotheses, the model likelihood is obtained from integrating (or "marginalizing") over the model parameters

$$\underbrace{P(\vec{x}|M_i, I)}_{\textit{(Model) Likelihood}} = \int d\theta\ \underbrace{P(\theta|M_i, I)}_{\textit{Prior on model parameter}}\ \underbrace{P(\vec{x}|\theta, M_i, I)}_{\textit{(Full) Likelihood}}$$

Credible intervals & marginalization

"Credible intervals" are regions in the parameter space that contain the true model parameters with certain probability (plausibility/degree of believe). Their definition *differs completely* from the definition of confidence intervals.



In one dimensions, a credible interval *R* with *probability content C* is given by:

$$\int_R d\theta \, P(\theta|\vec{x}, M) = C$$

This can be readily extended to two or more dimensions.

If several *nuisance parameters* exist in the model, they can be marginalized over to obtain the ***marginal posterior***.

$$P(\theta|\vec{x}, M) = \int d\phi \, P(\theta, \phi|\vec{x}, M)$$

This can again be used to define credible intervals.

Treatment of nuisance parameters

Profiling over unconstrained parameters

Using the above approach, irrelevant parameters (*nuisance parameters*) are profiled over when performing the statistical analysis. Effectively one only considers likelihood functions where the dependence on nuisance parameters is removed like

$$\mathcal{L}(\theta|x)_{\mathrm{prof}} = \max_{\xi} \mathcal{L}(\theta, \xi|x)$$

Marginalized likelihood

However, often additional information is available that can be used to calculate the *marginalized likelihood function*

$$\mathcal{L}(\theta|x)_{\mathrm{marg}} = \int d\xi \, \mathcal{L}(\theta, \xi|x) P(\xi)$$

Here, one effectively treats a Bayesian prior in a Frequentist interpretation, which can be adequate in many cases.

Model comparison & Jeffrey's scale

In Bayesian statistical inference, we perform ***Model comparison*** (this replaces *hypothesis testing* in the Frequentist approach). The central quantity is here the **odds ratio**

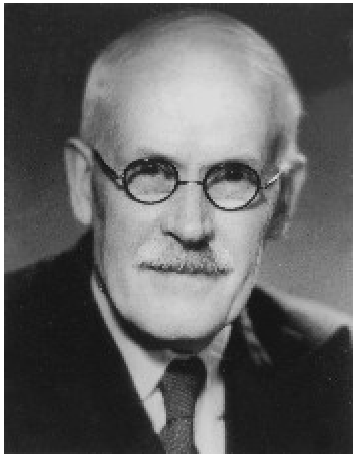
$$O_{ij} = \frac{P(M_i|\vec{x}, I)}{P(M_j|\vec{x}, I)}$$

which just depends on the model posteriors of a model *i* and *j*. Using Bayes' theorem, this can be written as

$$O_{ij} = \overset{\text{Priors}}{\frac{P(M_i|I)}{P(M_j|I)}} \underbrace{\overset{\text{Likelihoods}}{\frac{P(\vec{x}|M_i, I)}{P(\vec{x}|M_j, I)}}}_{= B_{ij}}$$

Bayes factor

- The Bayes factor describes how much *additional* credibility a model obtains due to the available data.
- It is independent of priors on the model (but can depend on priors on the model parameters!).
- A useful interpretation of the Bayes factor is given by the "Jeffrey's scale"



H. Jeffreys
1891-1989

Jeffrey's scale

B	"Strength of evidence"
< 1	Negative
1 – 3	Barely worth mentioning
3 – 10	Substantial
10 – 30	Strong
30 – 100	Very strong
> 100	Decisive