Advanced Statistical Methods

Lecture 6

Convergence distribution of M.-H. MCMC

We denote the PDF estimated by the MCMC as $\tilde{P}(x)$. It has the property

$$\int_V dx^n \tilde{P}(x) = \frac{\text{number of points inside volume } V (x_i \in V)}{\text{total number of points } n}$$

Convergence distribution

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

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

- The PDF estimated by the MCMC is hence proportional to the PDF in the acceptance ratio:

$$\frac{\tilde{P}(x)}{\tilde{P}(x')} = \frac{g(x' \to x)}{g(x \to x')} \frac{A(x' \to x)}{A(x \to x')} = \frac{P(x)}{P(x')}$$

$$\tilde{P}(x) \propto P(x)$$

(and equal if the sampled volume is the full volume).

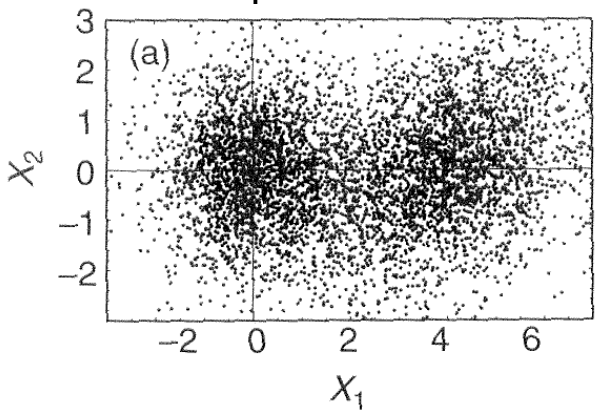
Exemplary run of M.-H. MCMC

Example

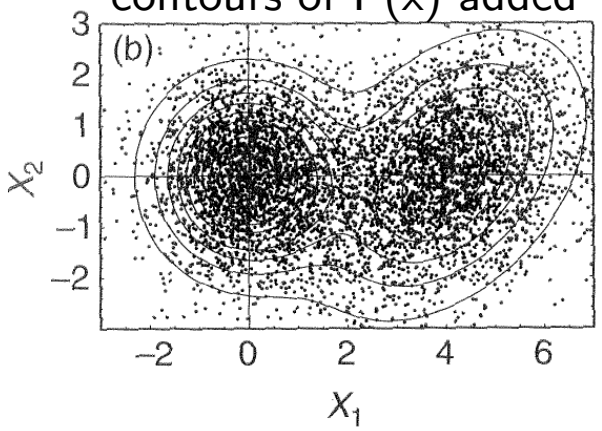
- Result from Metropolis Hastings scan of bi-modal distribution
- Proposal distribution has roughly the same extend as PDF:

$$g(x \to x') \propto e^{-\frac{1}{2\sigma^2}(x_1 - x_1')^2} e^{-\frac{1}{2\sigma^2}(x_2 - x_2')^2} \quad \text{with} \quad \sigma = 1.0$$

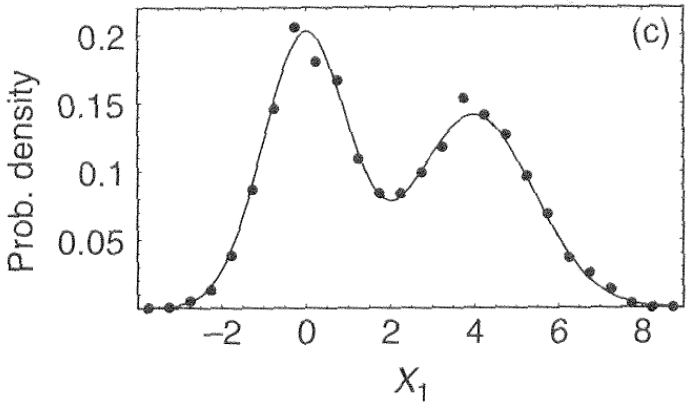
Chain points:

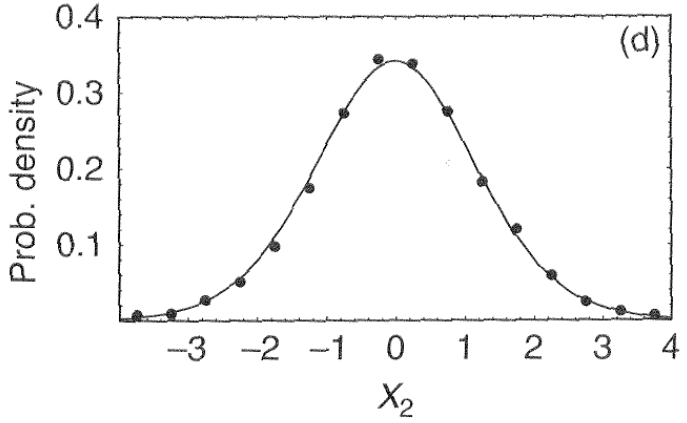


contours of P(x) added



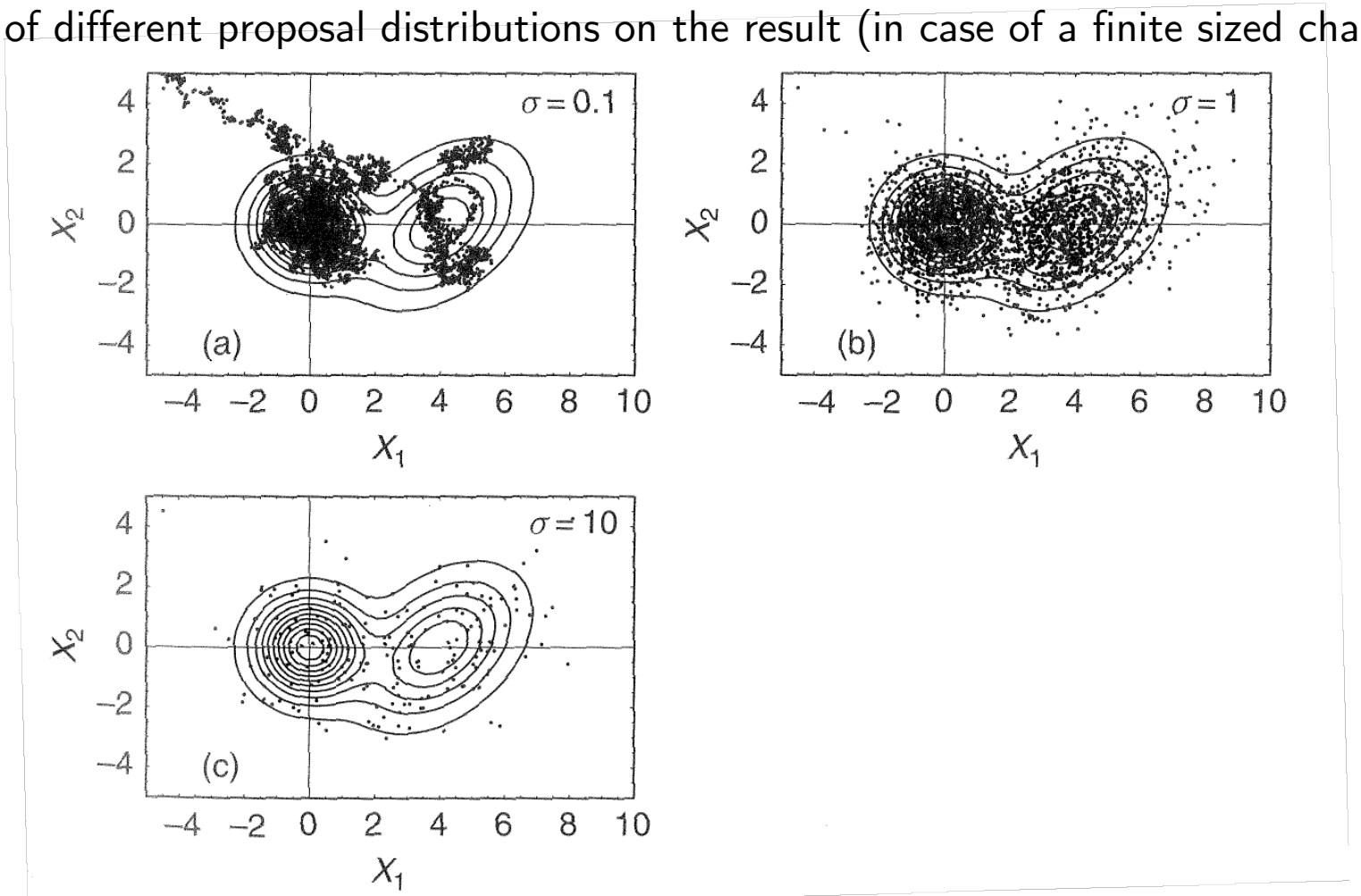
Marginalized posteriors:





Exemplary run of M.-H. MCMC

Effect of different proposal distributions on the result (in case of a finite sized chain)



Notes:

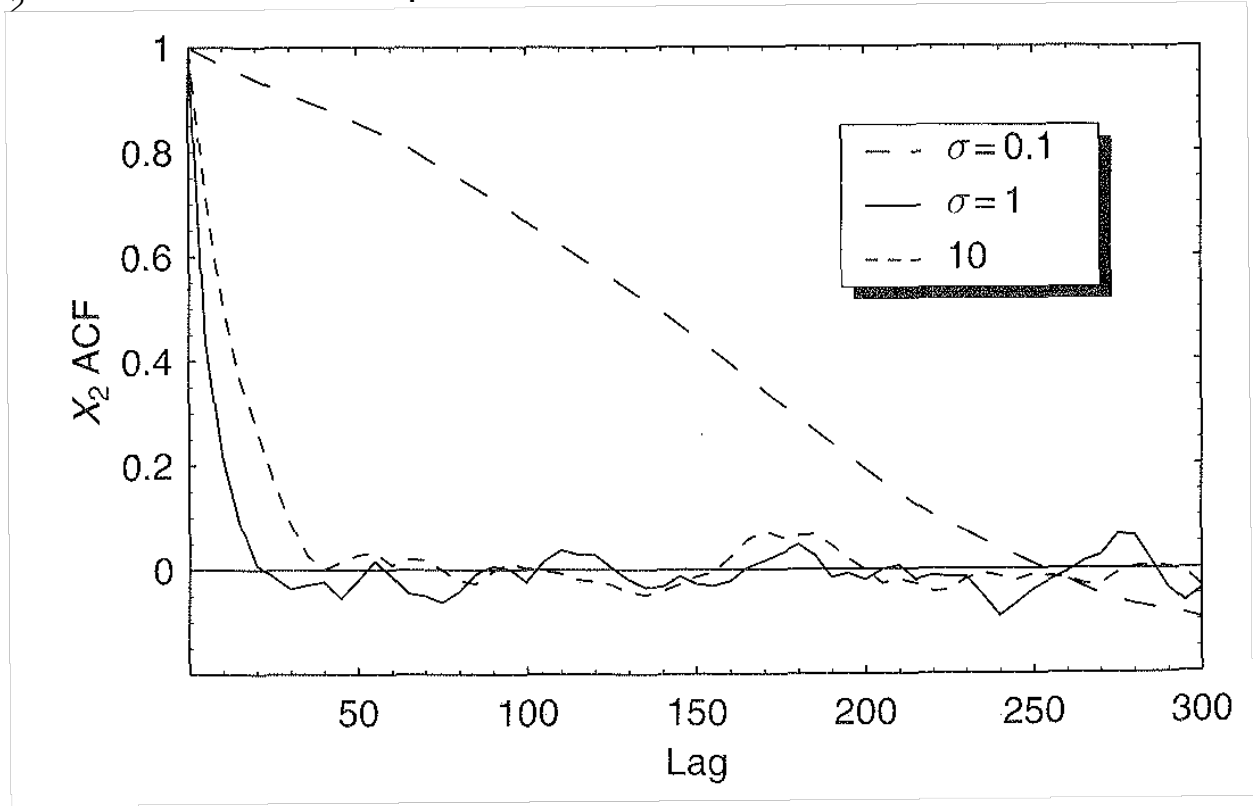
- Finding a reasonable proposal distribution that leads to a high acceptance rate is an art-form. It typically should roughly resemble the PDF of interest itself.
- Acceptance rates around 0.3 are high and to some degree optimal.
- The chain always takes some time to stabilize. The length of this "burn-in" period depends on the setup, and the corresponding points have to be removed before analyzing the chain.

Autocorrelation of MCMC chains

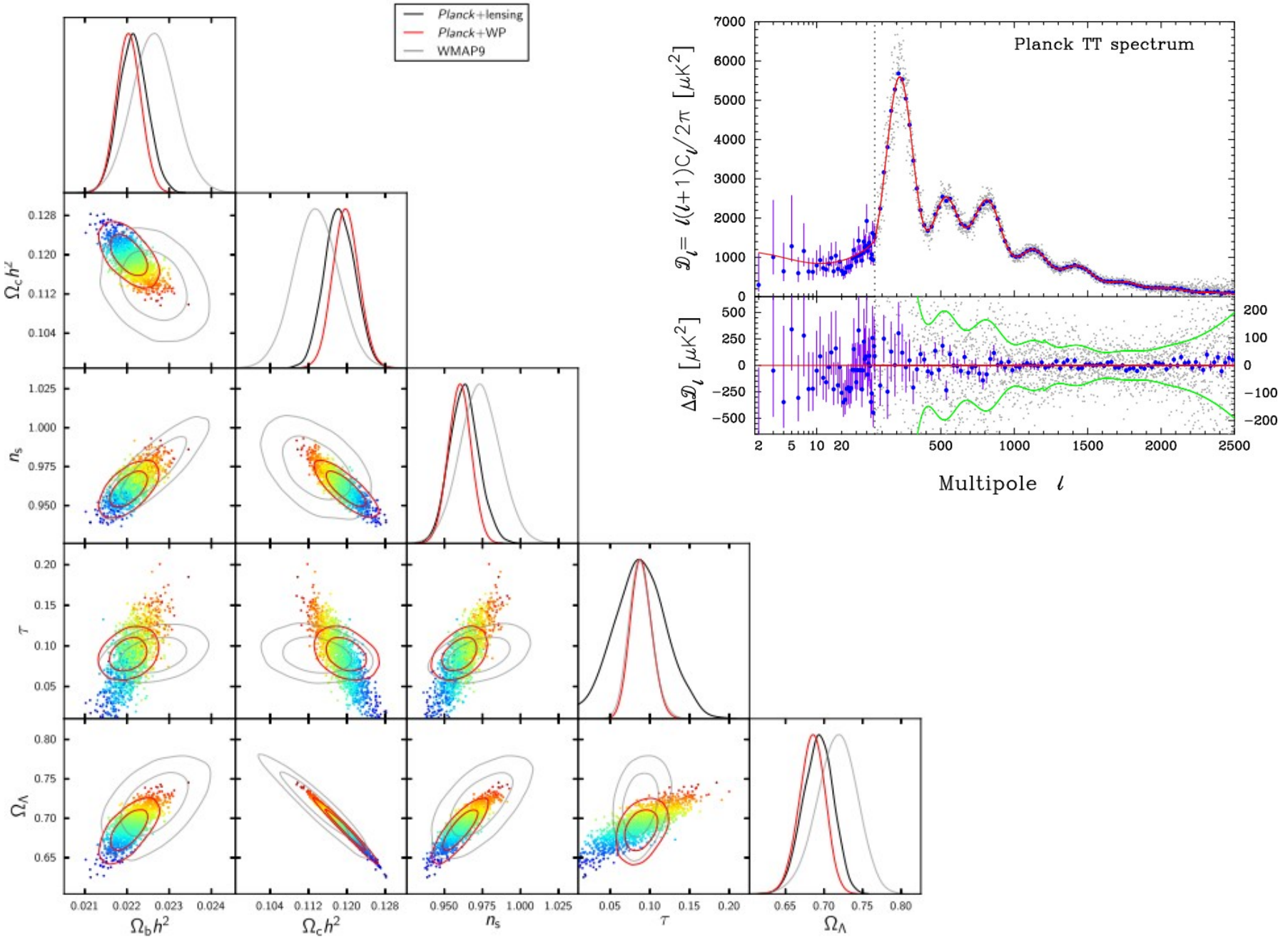
Depending on the proposal distribution and the acceptance rate, the Markov Chain can have a large degree of auto-correlation. This can be quantified by

$$\rho(h) = \frac{\sum_{i=1}^{n_{\text{points}}-h}(x_i - \bar{x})(x_{i+h} - \bar{x})}{\sqrt{\sum_{i=1}^{n_{\text{points}}-h}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n_{\text{points}}-h}(x_{i+h} - \bar{x})^2}} \qquad \bar{x} = \sum_{i=1}^{n_{\text{points}}} x_i$$

where the average is taken over all chain points. The variable x refers either to X_1 or X_2 in the above example.



Example: Cosmological fits



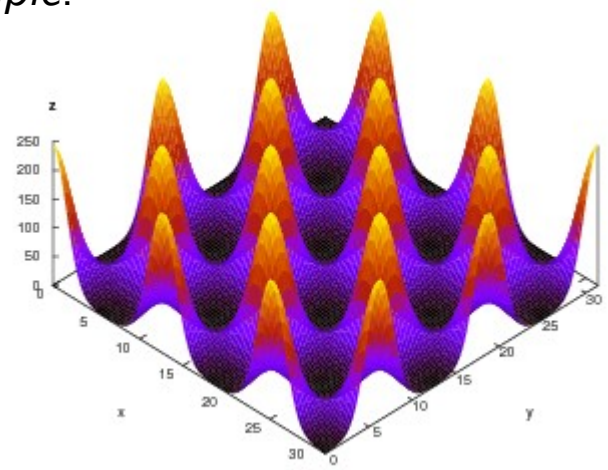
Advanced algorithms: Multinest

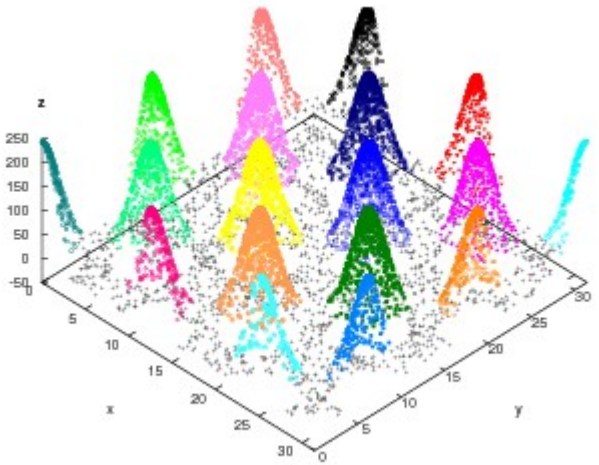
Problems:

- A traditional MCMC hinges on finding a good proposal distribution. Often, finding a good proposal distribution is equivalent of finding the PDF itself, which is kind of circular. This can be overcome with some adaptive methods.
- Multimodal distributions are very problematic for traditional MCMCs, since jumping between modes is *very* suppressed for standard proposal distributions (multivariate normal).

Advanced algorithms overcome these limitations. One well-known example is the *Multinest* algorithm. The basic idea is to start with a random distribution of points in the entire parameter region, and add points where the sampled PDF is large, while continuously updating the multimodal proposal distribution.

Example:





Generating random numbers

Problem: How to generate random numbers in the range [0, 1]? (or integers in the range [0, m-1])

Linear Congruent Generators (LCG)

- Select integers

$$a, c \gg m$$

and "seed"

$$I_0$$

- Iterate over

$$I_{i+1} = aI_i + c \ \ (\mathrm{Mod}\ m)$$

- Now, random numbers in the range [0, 1] are given by

$$u_i = \frac{I_i}{m-1}$$

Problems with this simple algorithm (overcome with more modern ones):

- The sequence repeats itself with some period that is less or equal *m*
- For certain choices of parameters, some generators might skip values and give an incomplete coverage of the interval [0, 1]
- Subtle correlations

Systematic uncertainties

One day before
publication.

Systematic uncertainties

Some definitions:

- Simplistic definition: "All uncertainties that are not a statistical error."

- Longer version of the same: We can consider all those uncertainties *statistical* for which we have a good model for their distribution in repeated experiments. *Systematic uncertainties* are all uncertainties where we lack such a model, for whatever reason.

- Scaling with amount of data: Statistical errors usually decrease like 1/sqrt(n) [more on that later], whereas systematic errors usually don't (actually, that is not always true, as also systematic errors can fluctuate and average out). In this sense statistical errors are *random* whereas systematic errors can be seen as a *bias*.

- Different context: Systematic uncertainties are usually related to uncertainties in the measurement apparatus, the modeling of the backgrounds, uncertainties in the parameters that enter the signal and background modeling.

Since systematic and statistical errors are usually independent, they can be often written like this:

$$x = 1.0 \pm 0.4(\text{stat.}) \pm 0.3(\text{syst.})$$

Examples for systematic uncertainties

- The **length of a ruler** that you use to calibrate length scales in your instrument could depend on humidity, temperature, age. This would affect all subsequent measurements.

- **Dust absorption** could affect the color and magnitude of a star that you are observing.

- The **theory predictions for g-2** (the muon magnetic moment) are hard, since they involve non-perturbative hadronic effects. Theoretical uncertainties can be estimated by theorists.

Notes:

- One can distinguish at least two types of uncertainties:
  - **Offset uncertainties** (calibration of zero point etc)

    $$x_{\mathrm{measured}} = x_{\mathrm{true}} + b$$

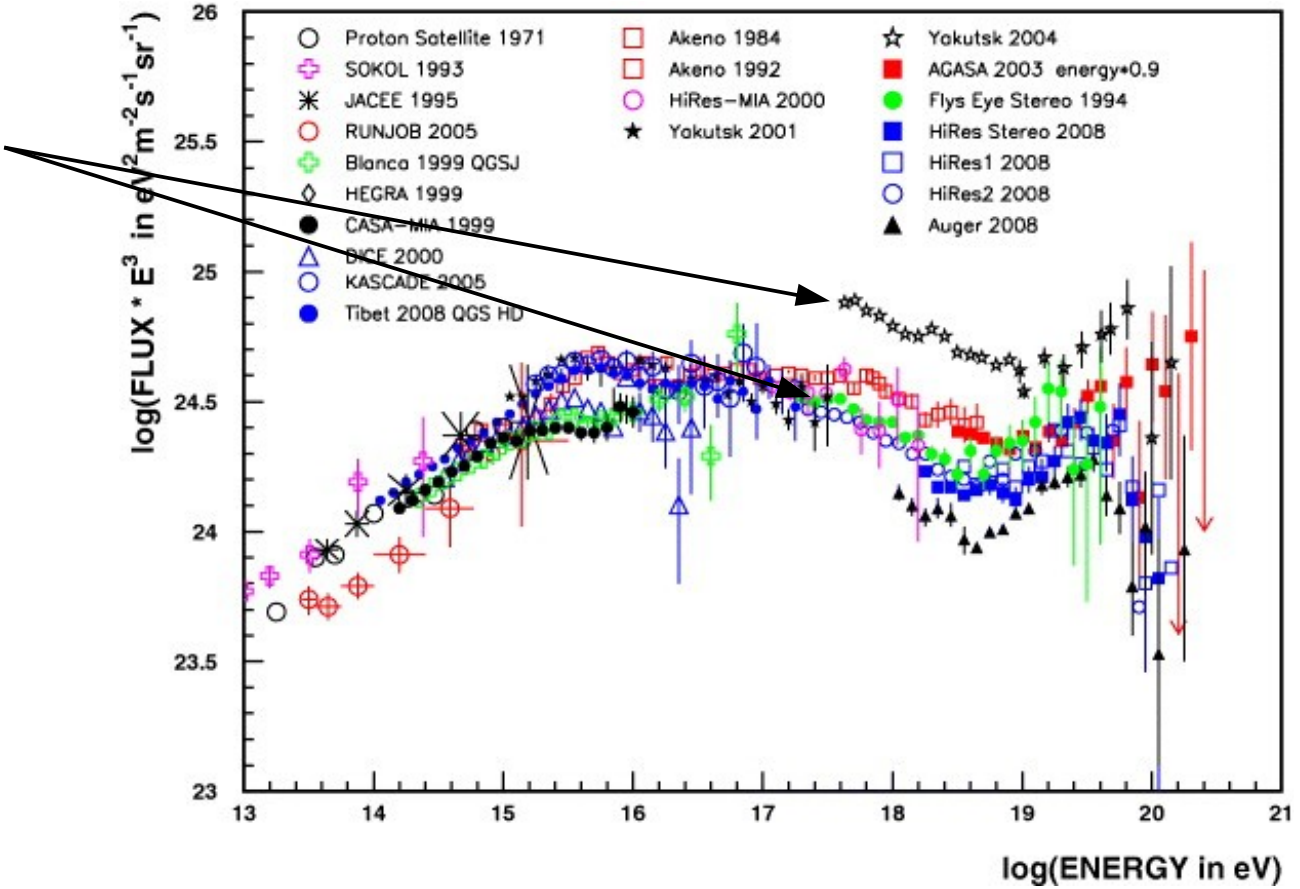
  - **Scaling errors** (contracted ruler etc)

    $$x_{\mathrm{measured}} = (1 + b) x_{\mathrm{true}}$$

Systematic uncertainties leave often statistical tests like the chi-squared goodness-of-fit unaffected. *The results can look very good, but might be complete rubbish.*

Correlations

Systematic errors typically look like this: small variations between data points, but still large deviations from the truth. How to parametrize that?



Systematic errors are in general *correlated.* This means that, when on looks at e.g. a flux measurement at different energies, the covariance matrix *is not diagonal.*

$$\Sigma_{ij} \equiv \langle (F_i - \langle F_i \rangle)(F_j - \langle F_j \rangle) \rangle \qquad \text{with} \quad \Sigma_{ij} \neq 0 \quad \text{for} \quad i \neq j$$

For example, in the case of an overall offset, given by

$$F_i = \langle F_i \rangle + \lambda F_i^{\text{offset}} \qquad \text{with} \qquad \lambda \sim N(0, \Delta\lambda)$$

the covariance matrix is $\Sigma_{ij} = \Delta\lambda^2 F_i^{\text{offset}} F_j^{\text{offset}}$.

How to propagate systematic errors

Often (like e.g. in your lab courses), systematic errors are treated on a similar footing as statistical errors. Making the strong assumption that they are normal distributed, their effect on the final result can be estimated by using Gaussian error propagation.

Consider a function that depends on various variables that can have both systematic and statistical errors

$$f = f(x, y, z)$$

The total uncertainty is then given by:

$$\Delta f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \Delta x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \Delta y^2 + \left(\frac{\partial f}{\partial z}\right)^2 \Delta z^2$$

with $\Delta x^2 = \Delta x^2_{\text{stat.}} + \Delta x^2_{\text{syst.}}$ etc

This gives the *total* uncertainty on the final result. The *systematic* part is the often obtained by subtracting the statistical error in square.

$$\Delta f^2_{\text{syst}} = \Delta f^2 - \Delta f^2_{\text{stat}}$$

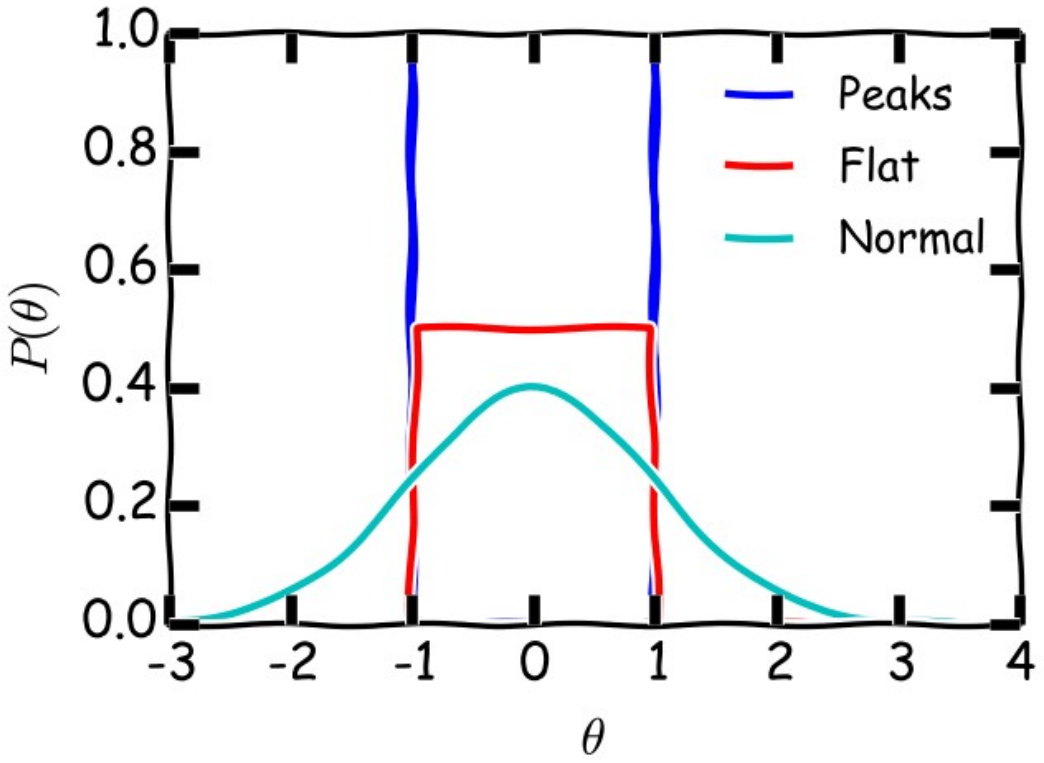
This is however a very simplistic approach. In many cases, there is no reason to believe that systematic errors should be normal distributed.

Bayesian approach

From the Bayesian perspective, systematic uncertainties can be incorporated in a simple way: as priors on the respective parameters. These are then marginalized over, with all its consequences (see Occam's razor).

$$\mathcal{L}(\theta|\vec{x}) = \int d\zeta \, \mathcal{L}(\theta, \zeta|\vec{x}) P(\zeta)$$

The question is here of course: What prior to use?



Example: *"The length of the ruler in known within 2%."*

- Typical assumption is a normal distribution, but this has long tails (is there *really* a chance that we are 6% off?).
- We can take a flat prior within the boundaries, but then the variance is too smaller than 2%.
- Only prior with correct variance and mean, and within the constraints, are two delta functions at the boundaries.

Frequentist

Frequentists in general have trouble to incorporate systematic uncertainties in their estimates. In the end, the ruler *is* too short, this is not a random process. Still, uncertainties can be incorporated by *profiling* or by *marginalization.* The latter is the same as in the Bayesian case. Profiling means:

$$\mathcal{L}(\theta|\vec{x}) = \max_{\zeta} \mathcal{L}(\theta, \zeta|\vec{x})\mathcal{L}(\zeta)$$

Note that we here need *a likelihood function* for the uncertain parameter. This is almost never available. But it can be approximated by pretending a prior is a likelihood, and just write

$$\mathcal{L}(\zeta) \propto P(\zeta)$$

Note that this introduces some dependence on the parametrization of the uncertain parameter: Let us suppose that there are two parameterization for the same quantity, which gives rise to two different likelihood functions

$$\zeta = \zeta(\xi) \quad P(\zeta) = \frac{d\xi}{d\zeta} P'(\xi) \qquad \begin{array}{c} \mathcal{L}(\zeta) \propto P(\zeta) \\ \mathcal{L}'(\xi) \propto P'(\xi) \end{array}$$

Then the likelihood functions are related like, $\mathcal{L}(\zeta) \propto \frac{d\xi}{d\zeta} \mathcal{L}'(\xi(\zeta))$

which means that in general the MLE is different in both cases $\hat{\zeta} \neq \zeta(\hat{\xi})$

Constraint term in Frequentist

Starting with the profile likelihood, one can motivate the use of *constraint terms* in the chi squared fits. Starting point is

$$\mathcal{L}(\theta|\vec{x}) \propto \max_{\zeta} \mathcal{L}(\theta, \zeta|\vec{x})\mathcal{P}(\zeta)$$

This can be rewritten as

$$-2\ln\mathcal{L}(\theta|\vec{x}) = \min_{\zeta}\left(-2\ln\mathcal{L}(\theta, \zeta|\vec{x}) - 2\ln P(\zeta)\right)$$

In the case of a Gaussian prior on zeta, the prior term becomes a constraint term in the chi square function:

$$\chi^2 = \min_{\zeta} \sum_i \frac{(\mu_i(\theta, \zeta) - x_i)^2}{\Delta\mu_i^2} + \frac{(\zeta - \zeta_0)^2}{\Delta\zeta^2}$$

This is a approximate, but very common treatment of systematics in the Frequentist context.

Best advices

- Question every input of your analysis. Do not take anything for granted or on trust.
- Different people tend to use different conventions while talking about the same thing. Double check all input parameters and equations.
- Think carefully about your instrument, and about every step in the analysis.
- Think of smart sanity checks, and if possible confirm your results with published results where they overlap. If they do not overlap with anything and you are the first, good luck.
- Do the same thing in different ways.
- Never trust the output of a computer program without validating its results.
- Mild paranoia helps.

$$\frac{1}{\sqrt{n}}$$

Sources of 1/sqrt(n)

When accumulating data, the statistical error on estimated parameters usually scale like 1/sqrt(*n*). We will discuss the reason for that in various ways.

Variance of the mean estimator:
Let's suppose we perform *n* measurements of variable *x*, which is following some distribution *p(x)*. A simple estimator for the mean of the distribution is given by

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The variance of *x* is given by $\sigma_x^2 \equiv \langle (x - \langle x \rangle)^2 \rangle$.

Question: What is the *variance* of the estimator?

$$\sigma_{\hat{\mu}_x}^2 \equiv \langle (\hat{\mu}_x - \langle \hat{\mu}_x \rangle)^2 \rangle = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \langle x \rangle)(x_j - \langle x \rangle)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \langle x \rangle)^2 = \frac{1}{n}\sigma_x^2$$

Hence, performing *n* measurements, we can estimate the mean of the distribution like

$$\mu_x \simeq \hat{\mu}_x \pm \frac{1}{\sqrt{n}}\sigma_x$$

Sources of 1/sqrt(n)

Imagine you perform a chi-squared fit to some data. The error bars on fitted model parameters theta can be as usual estimated using the Delta chi squared method.

$$\chi^2 = \sum_{i=1}^{n} \frac{(\mu_i(\vec{\theta}) - d_i)^2}{\Delta\mu_i(\vec{\theta})^2}$$

No consider the effect of increasing the amount of available data *d* by a factor of *N*. This means that you add additional terms to the chi-squared function. If your new data is measured in the same bins as your old data (say, you just measured a second time), and for the sake of notational simplicity has similar values, we simply obtain

$$\chi^2 = N \sum_{i=1}^{n} \frac{(\mu_i(\vec{\theta}) - d_i)^2}{\Delta\mu_i(\vec{\theta})^2} = \sum_{i=1}^{n} \frac{(\mu_i(\vec{\theta}) - d_i)^2}{(\Delta\mu_i(\vec{\theta})/\sqrt{N})^2}$$

In the much more realistic case that new and old data are not exactly the same, there will be additional terms, which only depend on the difference between the different data sets, and do not depend on the model parameters. Hence, they do not affect the fitting results.

We see that the errors that enter the chi-squared function effectively decrease by a factor of 1/sqrt(N). This will translate into an improvement of the constraints on model parameters by a factor of

$$\Delta\theta \propto \frac{1}{\sqrt{N}}$$

Sources of 1/sqrt(n)

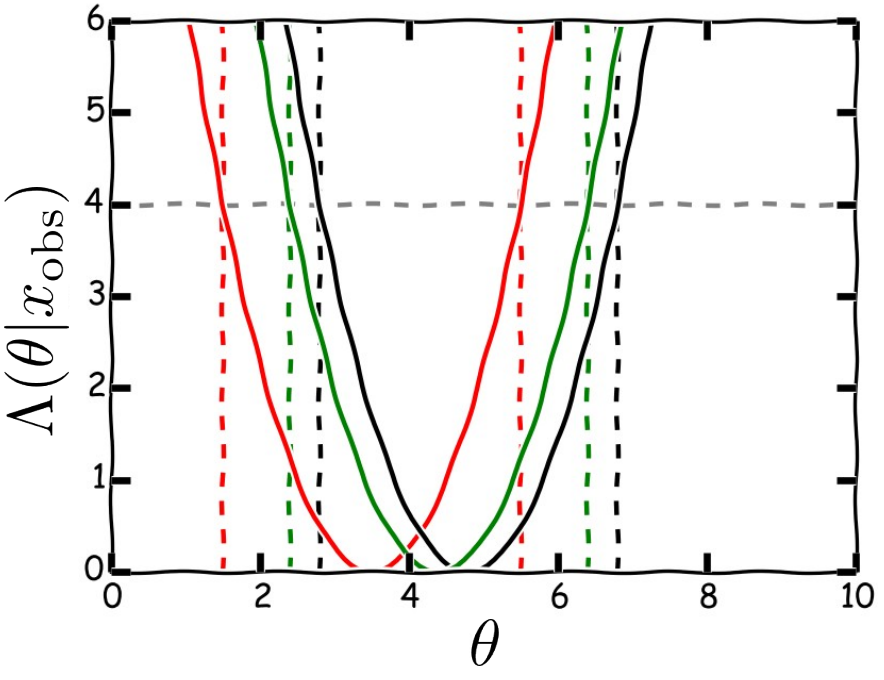
The same argument as on the previous slide holds also in general when looking at likelihood functions. Increasing the amount of data by a factor of *N* means approximately (again, to make the point, we assume here that old and new measurements are the same):

$$\mathcal{L}(\theta|\vec{x}) \to \mathcal{L}(\theta|\vec{x})^N$$

Usually, statistical inference is based on the log-likelihood function. Hence the quantity that is relevant for calculating statistical errors etc scales like *N*.

$$\Lambda(\theta|\vec{x}) = -2\ln\frac{\mathcal{L}_{\rm null}(\theta|\vec{x})^N}{\mathcal{L}_{\rm alt}(\theta|\vec{x})^N} = -2N\ln\frac{\mathcal{L}_{\rm null}(\theta|\vec{x})}{\mathcal{L}_{\rm alt}(\theta|\vec{x})}$$

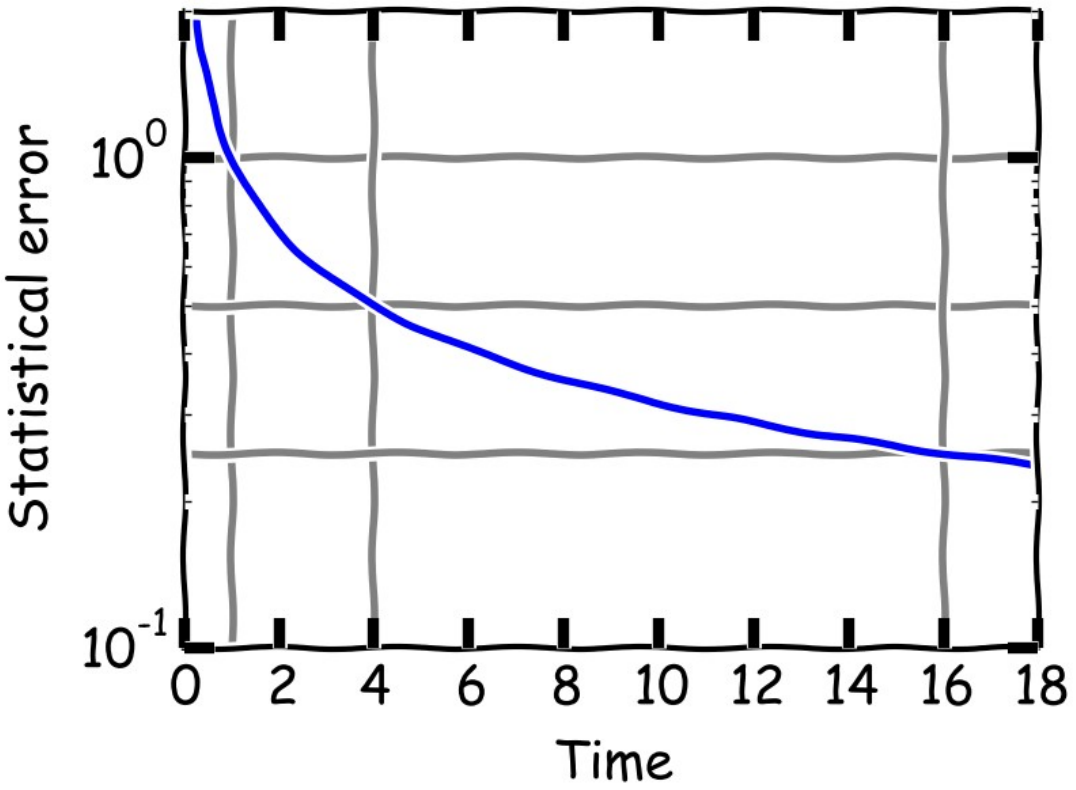
If null and alternative differ only in one degree of freedom, the Lambda is chi-squared distributed with *k=1*. One, two and three sigma confidence regions correspond to values where Lambda is smaller than 1, 4 or 9. Hence, the *size* of the confidence regions or intervals scales like 1/sqrt(N).



Statistical error vs instrument lifetime

This generic scaling with the amount of data (usually the number of measured events) is extremely important when considering how much worth it is to extend the lifetime of an experiment.

$$\Delta\theta \propto \frac{1}{\sqrt{t}}$$

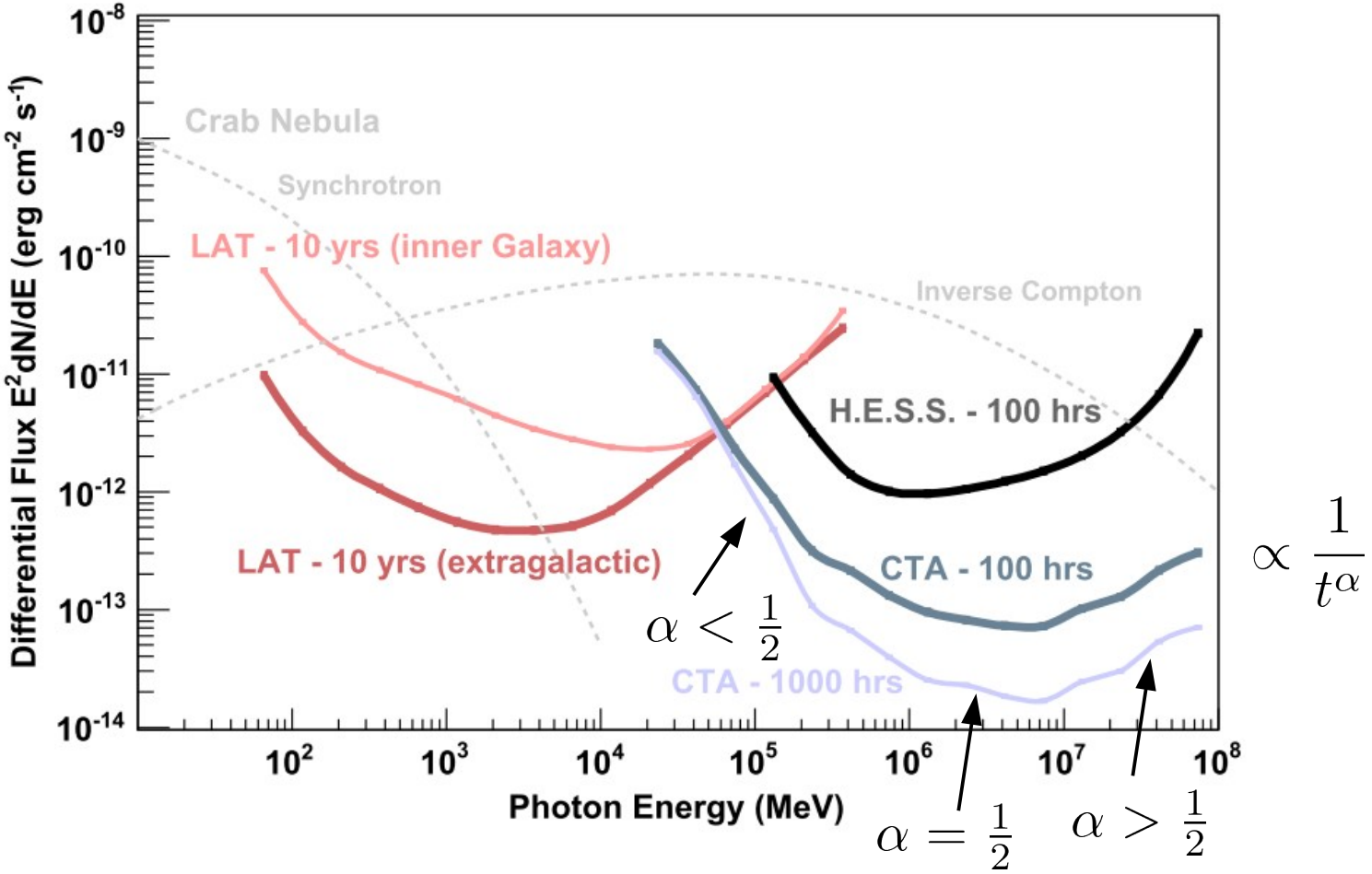


Improving the sensitivity of an instrument by a factor of two requires a *four times* longer runtime.

Why can it still be useful to extend e.g. the mission time of a satellite from 10 to 12 years?

Example: Projected sensitivity of CTA

Let's look at an example where this is *not* completely true. Below is shown the differential flux sensitivity of different gamma ray experiments, looking for point sources.



$$\propto \frac{1}{t^{\alpha}}$$

In the case of CTA, increasing the measurement time by a factor 10 increases the sensitivity at higher energies by a factor of roughly 3. This should not be surprising. However, at lower energies, it does not increase at all. At higher energies it actually increases *faster* (hard to see in the plot). *Why is this the case?*

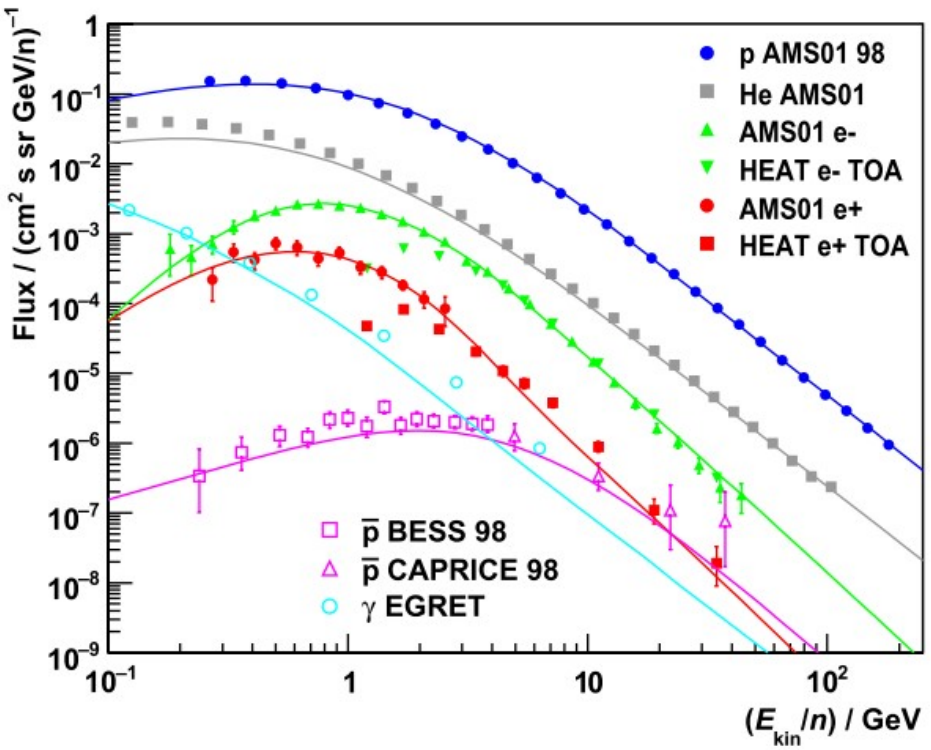
CTA PSC sensitivity

The sensitivity of CTA (and many other experiments) can be estimated by looking at the predicted number of signal and background events. In a given energy interval of interest, these can be calculated as

$$N_s = TA_{\mathrm{eff}} \cdot \int dE' \frac{dF_{\mathrm{sig}}}{dE'} \qquad N_b = TA_{\mathrm{eff}} \cdot \int dE' \frac{dF_{\mathrm{bg}}}{dE'}$$

where T is the measurement time, A the effective area, and the signal and background fluxes in the region of interest (usually of the size of the point spread function of the instrument) are denoted by dF/dE. Usually, the background is steeply falling with energy. Something like

$$\frac{dF_{\mathrm{bg}}}{dE} \propto E^{-2.5}$$



CTA PSC sensitivity

Often, the requirement for a signal detection in a certain energy range is three-fold, giving rise to sensitivity curves with three different regions.

$N_s > 0.05 N_b$

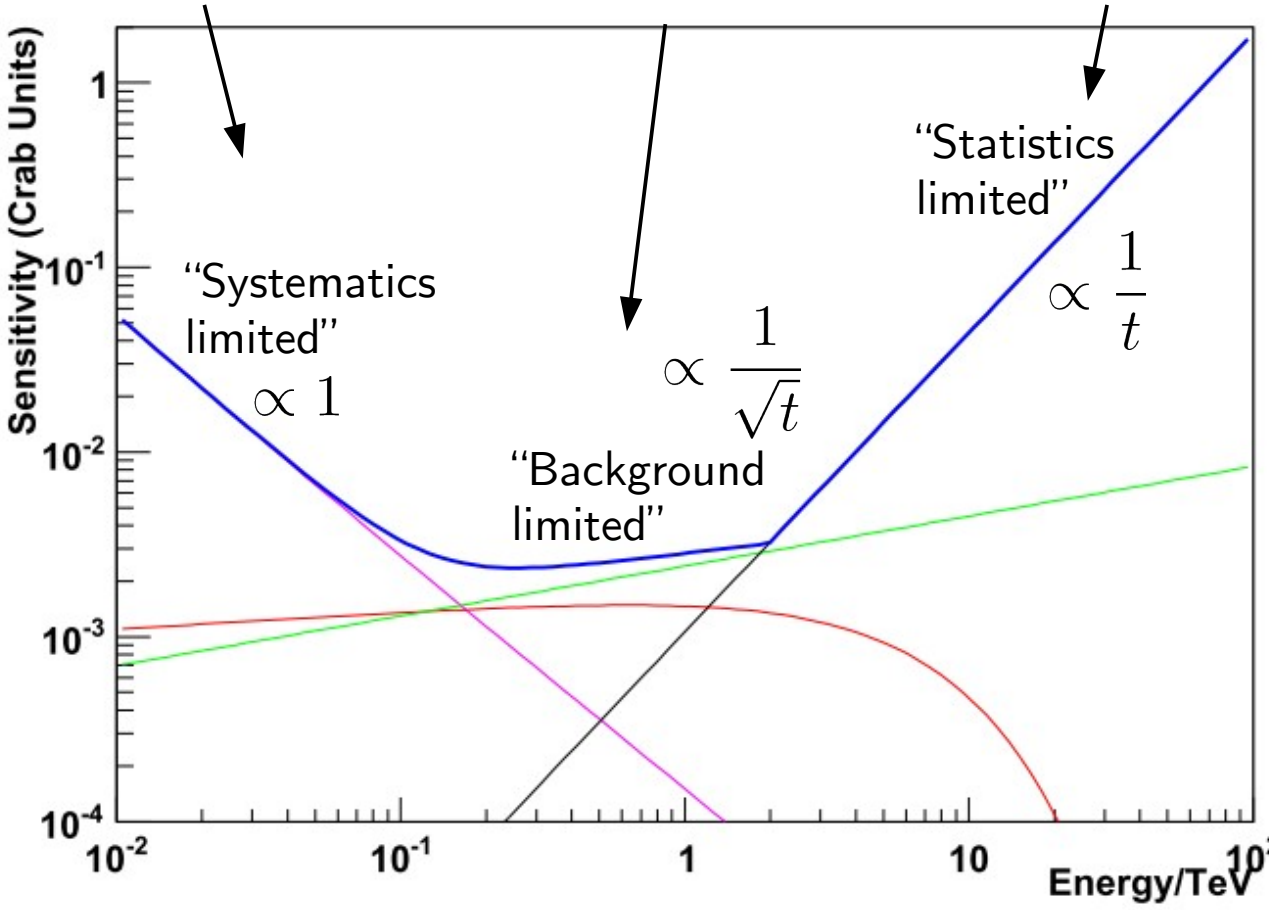
More than 5% of background.

$N_s > 5\sqrt{N_b}$

More than 5 sigma stat. significance.

$N_s \geq 10$

More than ten events.



In each of the three regions, the dependence on the measurement time is different.

Example Gravitational Waves