Data analysis (in latice QCD)

Christian Hoelbling
Bergische Universität Wuppertal

New Horizons in Lattice Field Theory
Mar. 18-19, 2013, IIP Natal

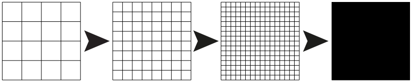
SFB
TRR55

BERGISCHE
UNIVERSITÄT
WUPPERTAL

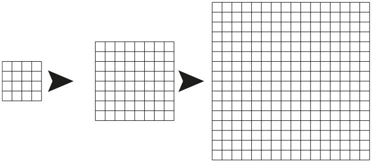
Lattice

Lattice QCD=QCD when

- Cutoff removed (continuum limit)



- Infinite volume limit taken



- At physical hadron masses (Especially π)
  - Numerically challenging to reach light quark masses

Statistical error from stochastic estimate of the path integral

Basic task

- Goal of phenomenological lattice QCD:
  - Compute expectation values of physical observables (masses, matrix elements,...)
  - Get reliable total errors of physical predictions
  - Use a minimum amount of computer time to obtain them
- Data analysis should:
  - provide results with reliable total erros
  - show how to efficiently improve the results

It's not about the final number, it's all about reliable errors

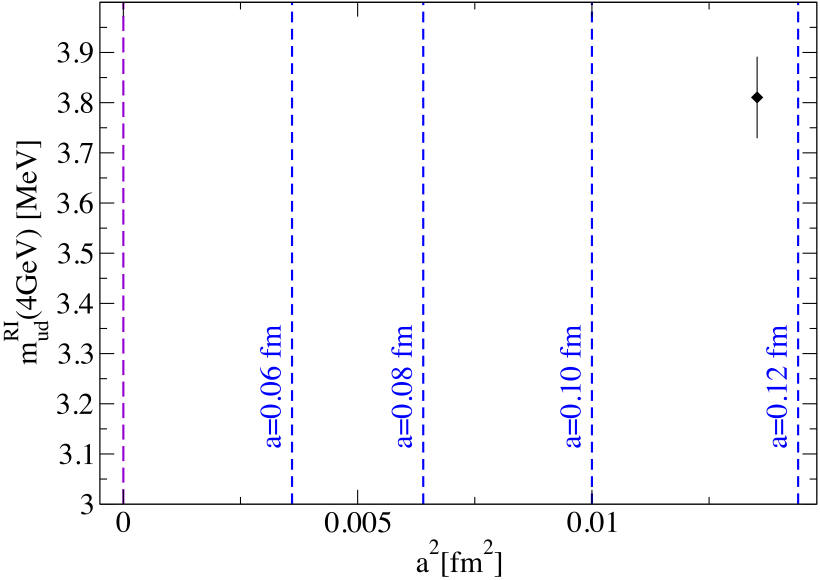
Errors

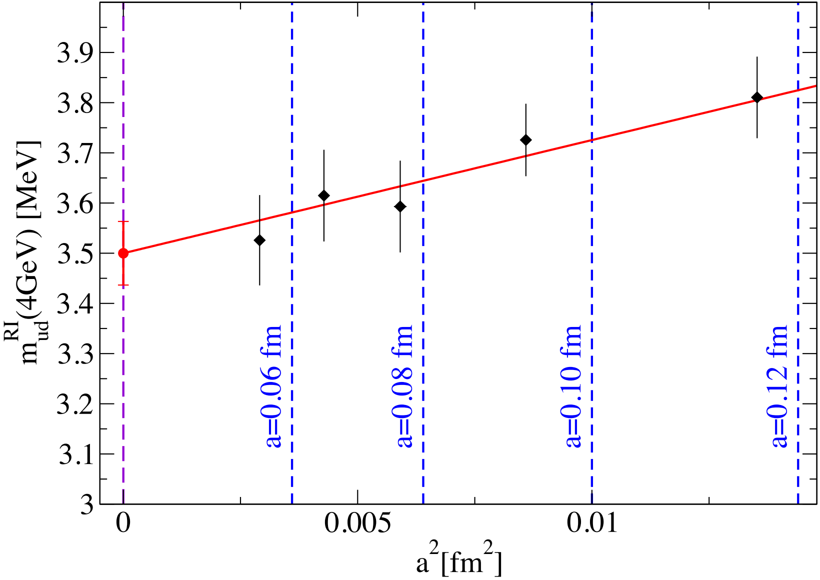
Errors fall into 2 broad categories:

- Statistical errors:
  - Origin: stochastic evaluation of the path integral
  - Can be treated by standard methods (e.g. bootstrap)
- Systematic errors:
  - Origin: our lack of knowledge
  - Can not be computed, only estimated

Keep good balance between the two!

- All systematics needs to be included for a correct result!


3.9
3.8
3.7
3.6
3.5
3.4
3.3
3.2
3.1
3
m_{ud}^{RI}(4GeV) [MeV]
0
0.005
0.01
$a^2[fm^2]$
a=0.06 fm
a=0.08 fm
a=0.10 fm
a=0.12 fm


3.9
3.8
3.7
3.6
3.5
3.4
3.3
3.2
3.1
3
0
0.005
0.01
$m_{ud}^{RI}(4GeV)$ [MeV]
$a^2 [fm^2]$
a=0.06 fm
a=0.08 fm
a=0.10 fm
a=0.12 fm

What we will practice

In this course, we will:

- Generate fake propagators
  - Everyone deals with a separate set
  - We know the solution
- Extract ground state mass (exercise 1)
- Extra/interpolate an observable to the “physical point” (exercise 2)
- Tutorial: focus on practical aspects

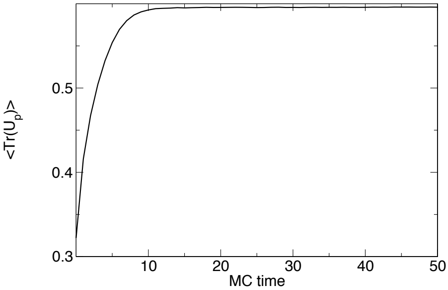
Autocorrelation

Lattice data are typically Markov chains:

- Each ensemble is based on the previous one
- Need independent ensembles in equilibrium distribution

Two problems:

- Thermalization
  - Affects only beginning
  - Cut initial configs
- Autocorrelation
  - Reduces number of independent configs
  - Different per observable



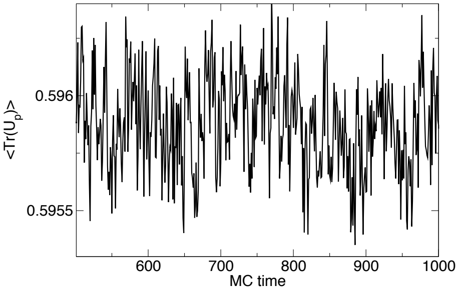
Autocorrelation

Lattice data are typically Markov chains:

- Each ensemble is based on the previous one
- Need independent ensembles in equilibrium distribution

Two problems:

- Thermalization
  - Affects only beginning
  - Cut initial configs
- Autocorrelation
  - Reduces number of independent configs
  - Different per observable



Autocorrelation - definitions

Given a time series a_t, the autocorrelation is the correlation of the time series with itself at a lag T

$$R(T) = \frac{\left\langle (a_t - \langle a_t \rangle)(a_{T+t} - \langle a_{T+t} \rangle) \right\rangle}{\langle a_t \rangle \langle a_{T+t} \rangle}$$

In a stationary random process

$$R(T) \sim e^{-T/\tau}$$

with the autocorrelation time τ

Autocorrelation - effects

We usually compute the integrated autocorrelation time

$$\tau_{\text{int}} = \sum_{T=1}^{N} R(T) \sim \int_0^\infty dT e^{-T/\tau} = \tau$$

Autocorrelation reduces the effective number of measurements

$$\sigma^2_{\langle a \rangle} \approx \frac{\sigma_a^2}{N}$$

Minimize autocorrelation: blocking the data

$$a_X = \frac{1}{B} \sum_{b=0}^{B-1} a_{BX+b}$$

Autocorrelation - effects

We usually compute the integrated autocorrelation time

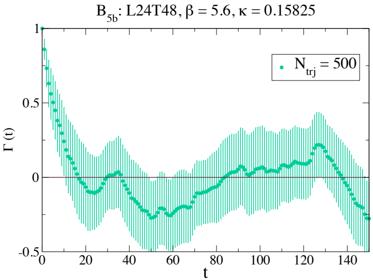
$$\tau_{\text{int}} = \sum_{T=1}^{N} R(T) \sim \int_0^\infty dT e^{-T/\tau} = \tau$$

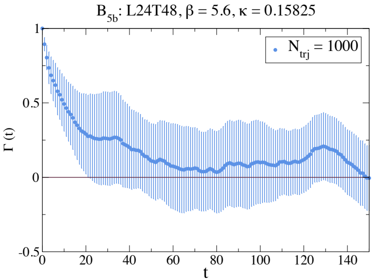
Autocorrelation reduces the effective number of measurements

$$\sigma^2_{\langle a\rangle} \approx \frac{\sigma_a^2}{N}(1 + 2\tau_{\text{int}})$$

Minimize autocorrelation: blocking the data

$$a_X = \frac{1}{B}\sum_{b=0}^{B-1} a_{BX+b}$$



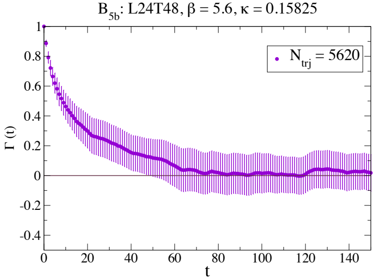


Difficult to compute τ_{int} accurately

- Time series long enough
- Observable dependent
- Global observables slower

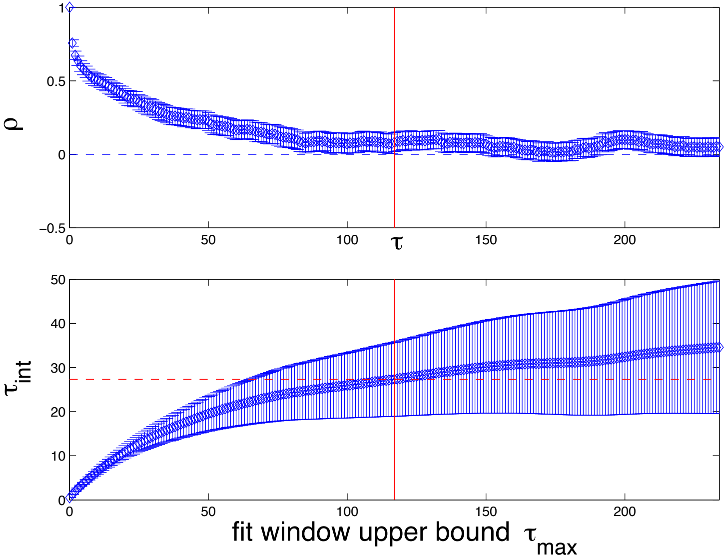
Example: plaquette in DDHMC

(Chowdhury et. al (2012))



Autocorrelation - packages

There is a standard package you can feed your time series to:
U. Wolff, Monte Carlo errors with less errors,
Comput.Phys.Commun. 156:143-153,2004;
Erratum-ibid.176:383,2007
hep-lat/0306017
MATLAB code can be found at:
http://www.physik.hu-berlin.de/com/ALPHAsoft/


1
0.5
0
−0.5
ρ
0
50
100
τ
150
200
50
40
30
20
10
0
τ_int
0
50
100
150
200
fit window upper bound τ_max

Ground state extraction

☞ Euclidean correlation function

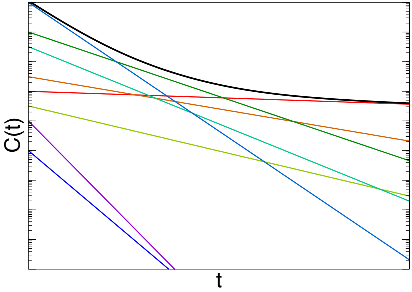
$$c_t = \langle 0|\mathcal{O}^\dagger(t)\mathcal{O}(0)|0\rangle$$

☞ Insert $1 = |i\rangle\langle i|$

$$\sum_i \langle 0|e^{Ht}\mathcal{O}^\dagger(0)e^{-Ht}|i\rangle\langle i|\mathcal{O}(0)|0\rangle$$

☞ Eigenbasis $|i\rangle$ of H

$$\sum_i |\langle 0|\mathcal{O}(0)|i\rangle|^2 e^{-(E_i-E_0)t}$$



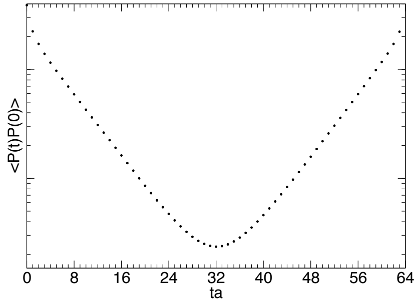
For $t \to \infty$:

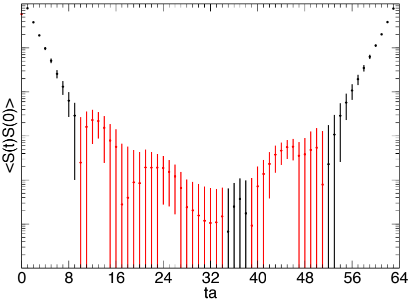
- ☞ Lightest state couling to $\mathcal{O}$ dominates: $c_t \propto e^{-M\cdot t}$
- ☞ $M_{t+\frac{1}{2}} = log[c_t/c_{t+1}]$, prefactor ➛matrix element

Signals from propagators

There are several complications

- Ground state coupling may be small
- Signal decays exponentially, noise not always
- There are backward (periodic BC) or border (open/fixed BC) contributions

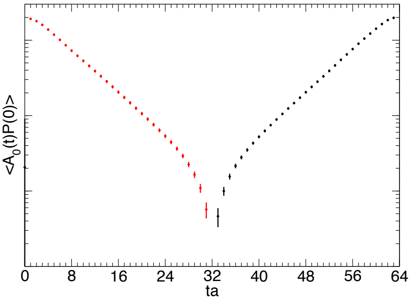


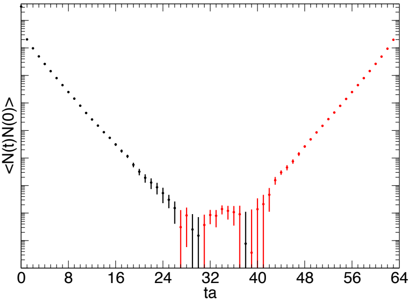


Signals from propagators

There are several complications

- Ground state coupling may be small
- Signal decays exponentially, noise not always
- There are backward (periodic BC) or border (open/fixed BC) contributions

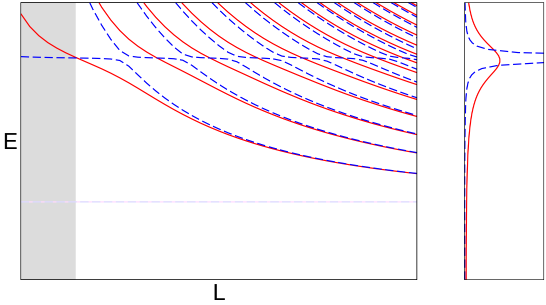




Excited state dominance

Small coupling of ground state is not an academic problem

- Occurs especially in resonant channels
- Ground state needs virtual $q\bar{q}$ production
- Different operators couple very differently



Propagator forms

Single state, propagating forward:

$$c_f(t) = c_f^0 e^{-mt}$$

The backward contribution:

$$c_b(t) = c_b^0 e^{-m(T-t)}$$

Include contributions warping around the lattice (tiny):

$$\begin{aligned} c_f(t) &= c_f^0 \left(e^{-mt} + e^{-m(T+t)} + \dots \right) \\ &= c_f^0 e^{-mt} \times \sum_{n=0}^{\infty} e^{-nmT} \\ &= c_f^0 e^{-mt} \frac{1}{1 - e^{-mT}} \end{aligned}$$

Propagator forms

For T (P) symmetric ($c^0 = c_f^0 = c_b^0$)
resp. antisymmetric ($c^0 = c_f^0 = -c_b^0$):

$$c_t = \frac{c^0}{1 - e^{-mT}} \left(e^{-mt} \pm e^{-m(T-t)} \right)$$

$$= \frac{c^0}{1 - e^{-mT}} e^{-m\frac{T}{2}} \times \begin{cases} \cosh\left(m\left(\frac{T}{2} - t\right)\right) \\ \sinh\left(m\left(\frac{T}{2} - t\right)\right) \end{cases}$$

Effective mass $M_{t+\frac{1}{2}}$ from numerical solution of:

$$\frac{c_{t+1}}{c_t} = \frac{\cosh\left(M_{t+\frac{1}{2}}\left(\frac{T}{2} - t - 1\right)\right)}{\cosh\left(M_{t+\frac{1}{2}}\left(\frac{T}{2} - t\right)\right)}$$

Propagator forms

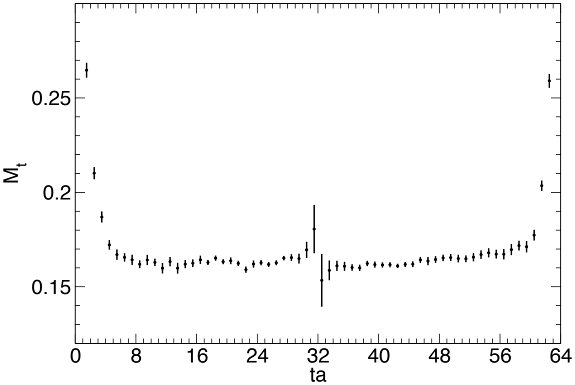
For T (P) symmetric ($c^0 = c^0_f = c^0_b$)
resp. antisymmetric ($c^0 = c^0_f = -c^0_b$):

$$
\begin{aligned}
c_t &= \frac{c^0}{1-e^{-mT}} \left(e^{-mt} \pm e^{-m(T-t)} \right) \\
&= \frac{c^0}{1-e^{-mT}} e^{-m\frac{T}{2}} \times \begin{cases} \cosh\left(m\left(\frac{T}{2}-t\right)\right) \\ \sinh\left(m\left(\frac{T}{2}-t\right)\right) \end{cases}
\end{aligned}
$$

Effective mass $M_{t+\frac{1}{2}}$ from numerical solution of:

$$
\frac{c_{t+1}}{c_t} = \frac{\sinh\left(M_{t+\frac{1}{2}}\left(\frac{T}{2}-t-1\right)\right)}{\sinh\left(M_{t+\frac{1}{2}}\left(\frac{T}{2}-t\right)\right)}
$$

Mass plateaus



Analytical 3-point expression (we will use this):

$$M_{t+\frac{1}{2}} = acosh\frac{c_{t+1} + c_{t-1}}{2c_t}$$

Mass fit

After identifying plateau range, we fit the propagators with

$$p_t = \frac{c^0}{1 - e^{-mT}} \left(e^{-mt} \pm e^{-m(T-t)} \right)$$

where m and c^0 are fit parameters
Maximum likelihood fit assuming normal error distribution:

$$\chi^2 = (c - p)_s (\Sigma^{-1})_{st} (c - p)_t \to \text{min}$$

Data points c, fit function p and covariance matrix Σ

$$\Sigma_{st} = \langle (c_s - \langle c_s \rangle)(c_t - \langle c_t \rangle) \rangle$$

Usual variance in diagonal elements $\Sigma_{tt} = \sigma(c_t)^2$

Fit results

From a fit we in principle get 3 things:

✓ The most likely value of the fit parameters

- Values of the parameters at $\chi^2 \to \text{min}$

✓ Standard errors of the parameters (more generally, confidence regions)

- Contours of constant $\Delta\chi^2 = \chi^2 - \chi^2_{\text{min}}$

✓ The quality of the fit

- From $Q = \frac{\Gamma(\frac{n}{2}, \frac{\chi^2}{2})}{\Gamma(\frac{n}{2})} = \frac{\int_{\frac{\chi^2}{2}}^{\infty} t^{\frac{n}{2}-1} e^{-t} dt}{\int_0^{\infty} t^{\frac{n}{2}-1} e^{-t} dt}$

- Q: probability that - given the model - the data are at least as far off the prediction as the real data

☞ Q should be a flat random value $\in [0, 1]$

Correlations

For uncorrelated data, Σ is diagonal

$$C_{st} = \frac{\Sigma_{st}}{\sigma(c_s)\sigma(c_t)}$$

Typical (estimated) normalized covariance C for a correlator:

1.0000	0.9963	0.9840	0.9746	0.9509
0.9963	1.0000	0.9912	0.9801	0.9595
0.9840	0.9912	1.0000	0.9934	0.9846
0.9746	0.9801	0.9934	1.0000	0.9912
0.9509	0.9595	0.9846	0.9912	1.0000

Eigenvalues:

4.9224	0.0661	0.0059	0.0041	0.0014

Problems with correlations

The structure of the covariance matrix can be problematic

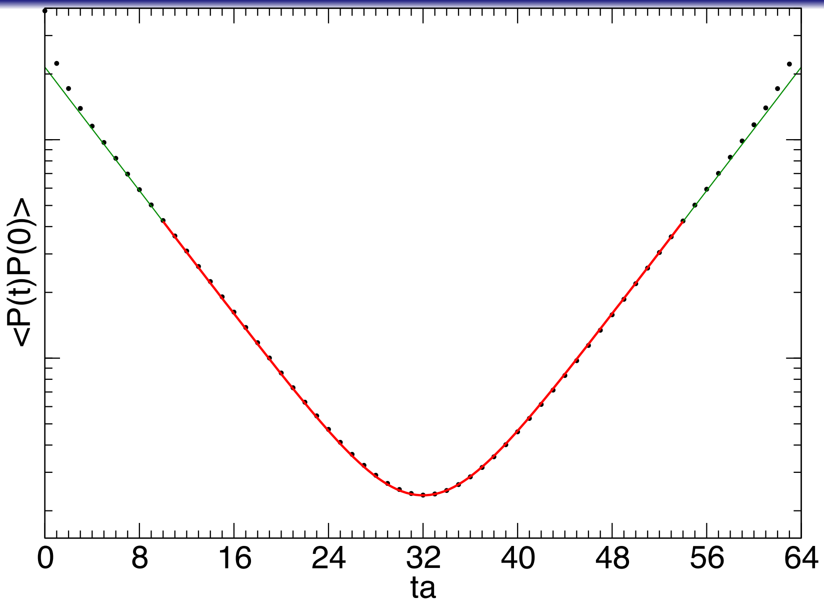
- Covariance matrix determined statistically
- In C^{-1}, small modes dominate
- Smallest modes have large errors

One can:

- Do an uncorrelated fit: Σ diagonal
- Truncate small eigenmodes
  - Truncate them (optionally correct diagonal)
  - Average them (Michael, Mc Kerrell, 1994)

Problem: Q and parameter errors useless

➢ Need to be determined in some other way



<P(t)P(0)>
0
8
16
24
32
40
48
56
64
ta

Computing errors

When you make N measurements a_i, you compute

- the estimate of the expectation value

$$\langle a \rangle = \frac{1}{N} \sum_{i=1}^{N} a_i$$

- the estimated error of the expectation value

$$\sigma^2_{\langle a \rangle} = \frac{1}{N} \frac{1}{N-1} \sum_{i=1}^{N} (a_i - \langle a \rangle)$$

From $\mathcal{O}(100)$ configs, we get one mass measurement!
Do we have to repeat this $\mathcal{O}(100)$ times to estimate σ^2?

Resampling

No! We can resample our ensemble:

- Given N configs c_i and the full ensemble $E = \{1, \ldots, N\}$
- Given an observable $O(A)$ on an arbitrary Ensemble A

➛ We can produce one resampled ensembles B_1 by drawing with repetition N configs from E

➛ We actually draw N_B resampled ensembles B_i

➛ We compute $\overline{O} = O(E)$ and $O_i = O(B_i)$

The distribution of O_i mimics independent measurements!

$$\sigma_O^2 \approx \sigma^2(O_i) \qquad \langle O \rangle \approx \overline{O} + \overline{O} - \langle O_i \rangle$$

Resampling

No! We can resample our ensemble:

- Given N configs c_i and the full ensemble $E = \{1, \ldots, N\}$
- Given an observable $O(A)$ on an arbitrary Ensemble A

➛ We can produce one resampled ensembles B_1 by drawing with repetition N configs from E

➛ We actually draw N_B resampled ensembles B_i

➛ We compute $\overline{O} = O(E)$ and $O_i = O(B_i)$

The distribution of O_i mimics independent measurements!

$$\sigma_O^2 \approx \sigma^2(O_i) \qquad \langle O\rangle \approx \overline{O} + \cancel{\overline{O}} \cancel{-} \langle \cancel{O_i} \rangle$$

Usually better not to correct (stability)

Jackknife

Jackknife is similar to bootstrap:

- Cut the ensemble E into N_J same size blocks
- Form N_J resampled ensembles J_i by leaving out one block from E at a time
- Compute $\overline{O} = O(E)$ and $O_i = O(J_i)$

$$\sigma_O^2 \approx (N_J - 1)\sigma^2(O_i) \qquad \langle O \rangle \approx \overline{O} + (N_J - 1)\left(\overline{O} - \langle O_i \rangle\right)$$

Jackknife

Jackknife is similar to bootstrap:

- Cut the ensemble E into N_J same size blocks
- Form N_J resampled ensembles J_i by leaving out one block from E at a time
- Compute $\overline{O} = O(E)$ and $O_i = O(J_i)$

$$\sigma_O^2 \approx (N_J - 1)\sigma^2(O_i) \qquad \langle O \rangle \approx \overline{O} + \cancel{(N_J - 1)(\overline{O} - \langle O_i \rangle)}$$

Usually better not to correct (stability)

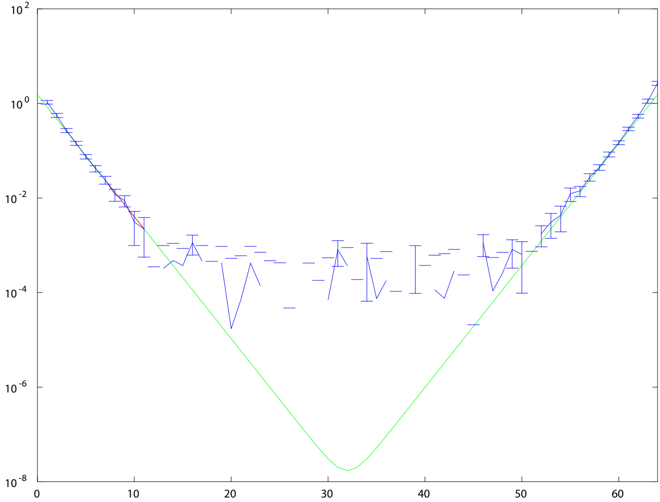
Using bootstrap

Some practical notes:

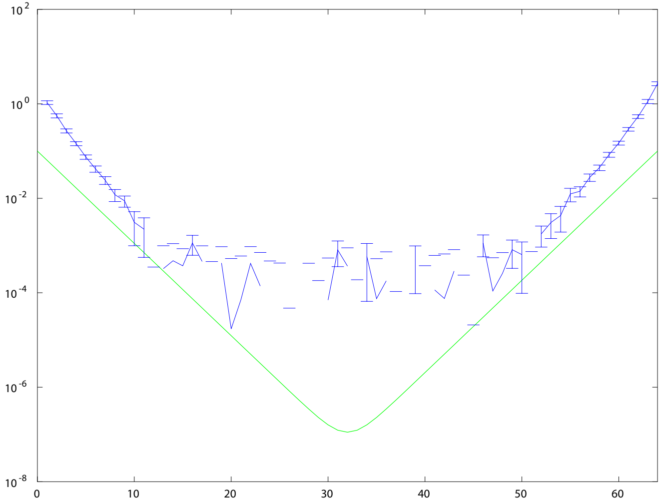
- Use bootstrap if you can (more expensive though)
- Choose N_B as large as you can
- Do the complete analysis within the bootstrap
  - This does even include averaging over different analysis procedures for systematics etc.
  - Only exception are estimates of global ensemble properties like e.g. (co-)variances needed for fits within the bootstrap.

  ➛ nesting bootstraps usually not necessary
- Not necessary if O is linear: $\sigma_{\mathrm{JN}} \equiv \sigma_{\mathrm{naive}}$
- You may extract more information from distribution of O_i
  - Confidence intervals, percentiles, etc.

Rho propagator



Rho propagator



Fits with x-errors

A typical analysis situation:

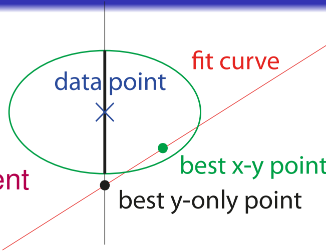
- We have collected data at different bare quark masses
- We want to make a prediction at the physical point (for simplicity we ignore continuum and infinite volume)

How do we proceed?

- Define the physical point (e.g. M_π)
- Extra/interpolate target observable there

M_π is not a parameter!

X-errors



Fitting data with errors in the x-axis:

- ➛ add each x-value as a fit parameter
- ➛ constrain each x-value with measurement

Uncorrelated case:

$$\chi^2 \to \chi^2 + \sum_i (x_i - p_i)^2 / \sigma_{x_i}^2$$

Generalization with full covariance matrix

✎ Big covariance matrices lead to uncontrolled fits

Mandatory to eliminate spurious correlations

Correlated errors

Special case: x_i, y_i correlated, but uncorrelated with x_j, y_j $i \neq j$

- ➢ Appears naturally in fit of independent ensembles
- ➢ Covariance matrix reduces to block diagonal form

Contribution to χ^2:

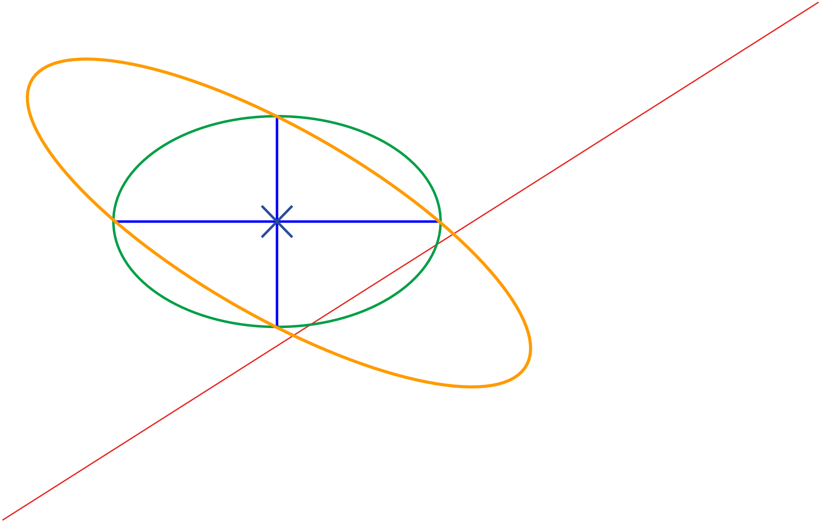
$$\chi^2 \supset \chi_i^2 \begin{pmatrix} \Delta x & \Delta y \end{pmatrix} \begin{pmatrix} \Sigma_{xx}^{-1} & \Sigma_{xy}^{-1} \\ \Sigma_{xy}^{-1} & \Sigma_{yy}^{-1} \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

- χ_i^2 constant along an ellipse
- Covariance Σ_{xy}^{-1} tilts the axis

✓ Including x-errors can never increase χ_i^2

✓ Including x-errors does not change n (d.o.f.)

Error ellipses



General strategy

Sometimes subsets of data points are correlated

- 3 independent ensembles at each of 3 lattice spacings
- A measurement of each of the 3 lattice spacings a_i

How do you extrapolate the observable M to the continuum?

- Form $M = M_{\text{lat}}/a_i$ for each ensemble
- Error on $M = M_{\text{lat}}/a_i$ is combination of error on M_{lat} and a_i

✗ Introduces correlations between independent ensembles

General strategy

Sometimes subsets of data points are correlated

- 3 independent ensembles at each of 3 lattice spacings
- A measurement of each of the 3 lattice spacings a_i

How do you extrapolate the observable M to the continuum?

- Form $M = M_{\text{lat}}/a_i$ for each ensemble
- Error on $M = M_{\text{lat}}/a_i$ error on M_{lat}, ignore a_i
- ✗ Lattice spacing error not accounted for

General strategy

Sometimes subsets of data points are correlated

- 3 independent ensembles at each of 3 lattice spacings
- A measurement of each of the 3 lattice spacings a_i

How do you extrapolate the observable M to the continuum?

- Introduce a fit parameter $\hat{a}_i$ for each lattice spacing
- Constrain $\hat{a}_i$ with measurement
- Fit $M_{\text{lat}} = M\hat{a}_i$ for each ensemble

Combined fit quality

When doing your continuum/chiral/infinite volume fit

- Data points are often results of fits themselves
- How do you compute the quality of cascaded fits?

Theoretical ideal (not feasible):

- Do one big fit

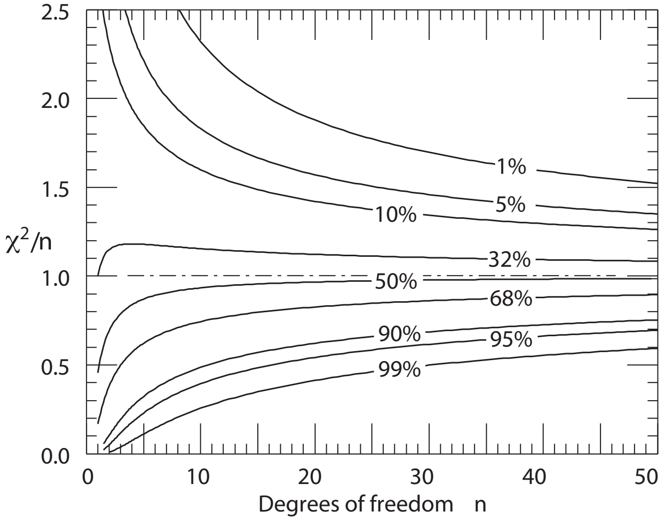
All original fits worked fully correlated:

- Sum χ^2 and d.o.f. of all fits $\rightarrow Q$

Original fits not fully correlated:

- Treat data points as input, just compute Q of final fit

Fit quality



(PDG, 2012)

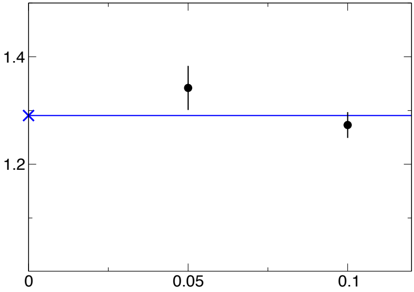
Which fit is better?

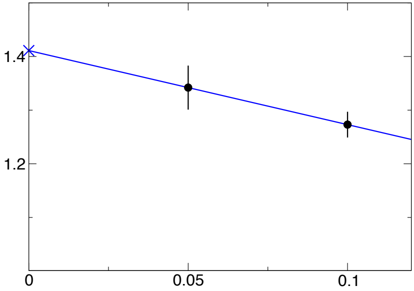
The following slides compare 2 fits each

All data are uncorrelated

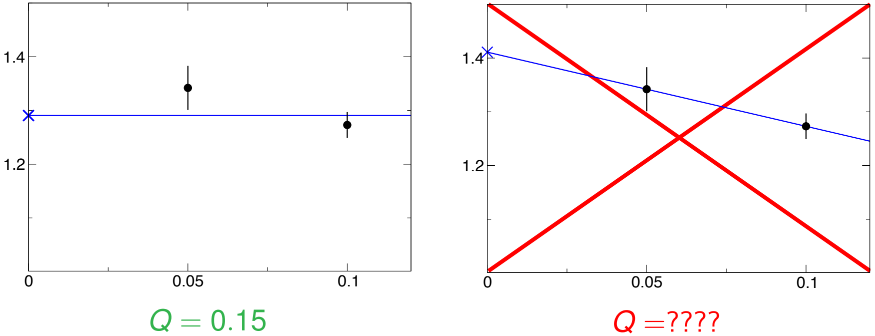
Which fit can be trusted more?

Which fit is better?



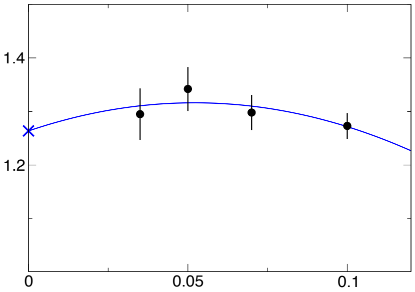


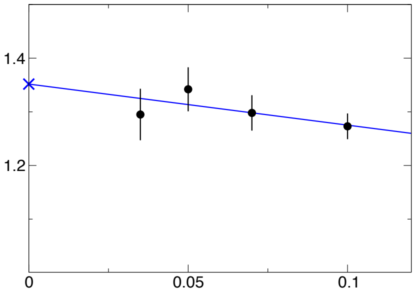
Which fit is better?



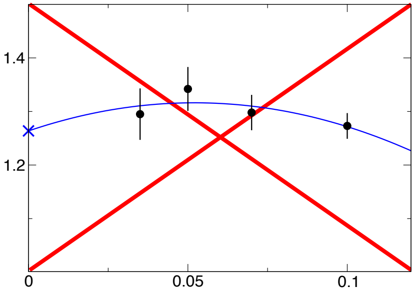
Never leave 0 d.o.f., you loose control over fit quality

Which fit is better?

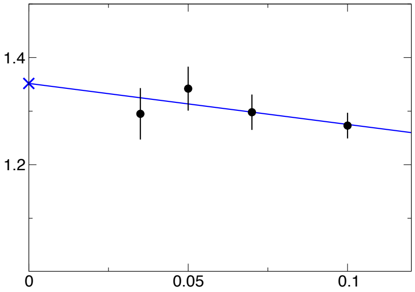




Which fit is better?



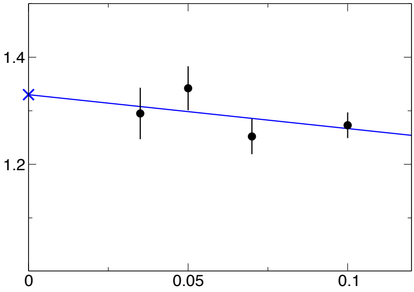
$Q = 0.42$

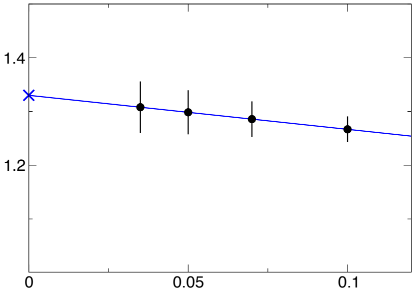


$Q = 0.64$

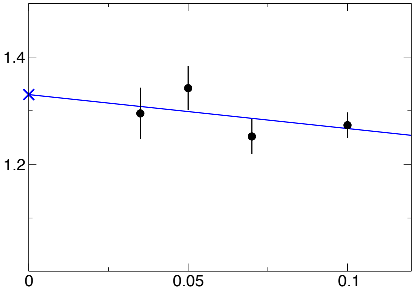
- Do not try to extract too much from the data
- The displayed data have no sensitivity towards a curvature term. It is compatible with 0.

Which fit is better?

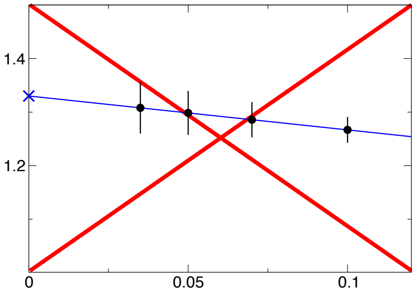




Which fit is better?



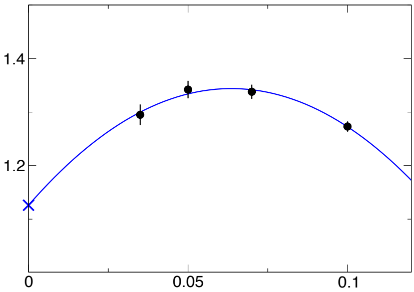
$Q = 0.31$

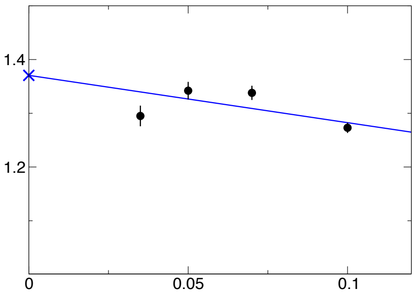


$Q = 1.00$

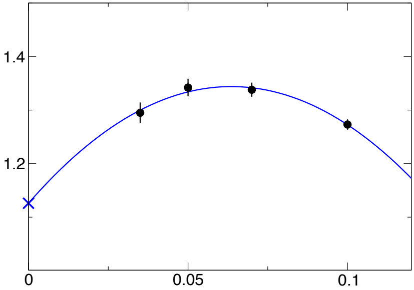
- $1 - Q = 8 \times 10^{-13}$ ➛winning the lottery is more probable than having a result this good by chance
- Data are suspicious (unrecognized correlation)

Which fit is better?

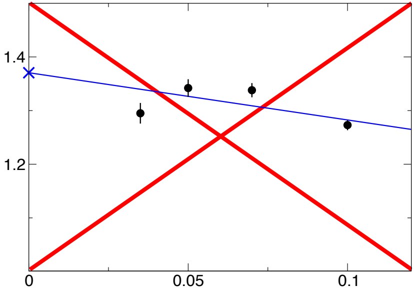




Which fit is better?



$Q = 0.54$



$Q = 0.002$

Linear modell is not sufficient for these data

Practical hints

Some hints for numerically minimizing a complex χ^2 function

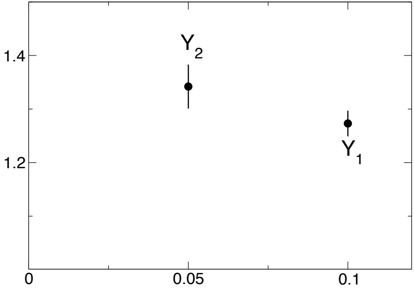
- Give reasonable starting values
  - Solver might find a wrong minimum or crash
- Build up your fit parameter by parameter
  - Start with all but the most relevant parameters constrained
  - Minimize the constrained fit first
  - When it has converged, free one more parameter
- Check pulls and bootstrap samples for outliers
  - A good fit can identify problematic input data
- Always look at the fit to check it does fit the data

Systematics

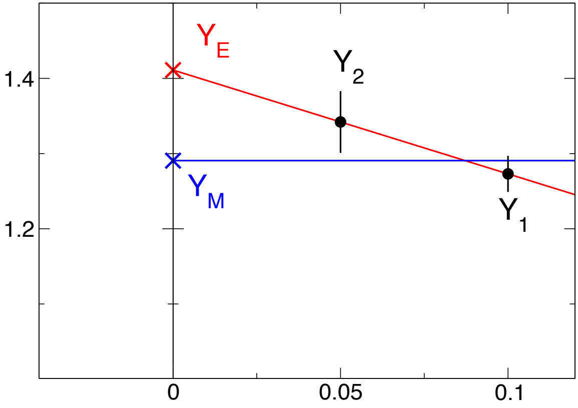
How do we compute the systematic error?

- We don't
- Systematics can only be estimated
- There is no single correct procedure

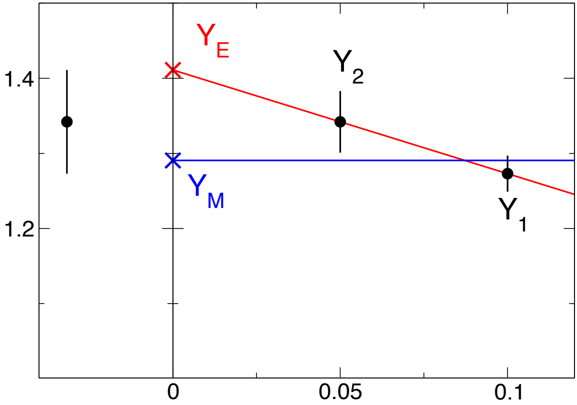
Example: systematic error of $x \to 0$



Simple estimates



Simple estimates

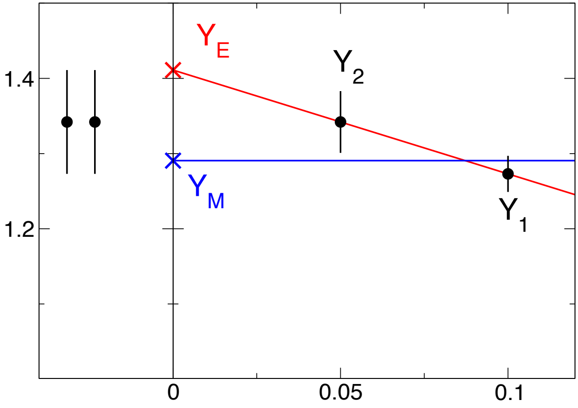


central value Y_2

error $|Y_2 - Y_E|$

no linear fit

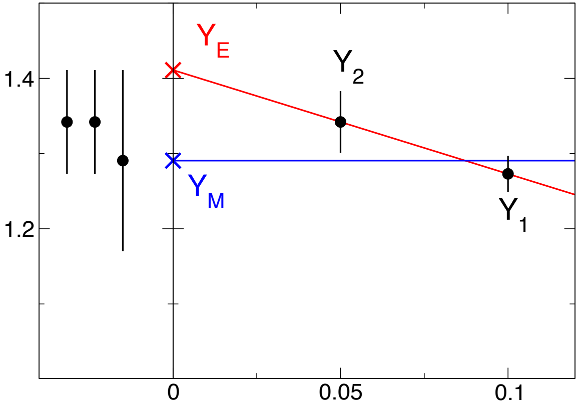
Simple estimates



central value Y_2

error $|Y_2 - Y_1|$

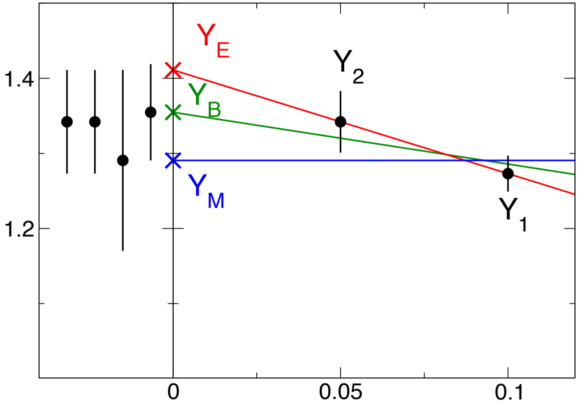
Simple estimates



central value Y_M

error $|Y_M - Y_E|$

Simple estimates



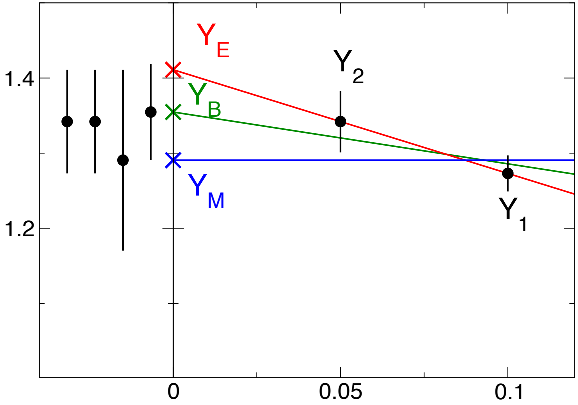
central value Y_B

error $|Y_M - Y_B|$

You can do a linear fit if you have prior knowledge on the slope

☞Constraint on slope is an additional data point

Simple estimates



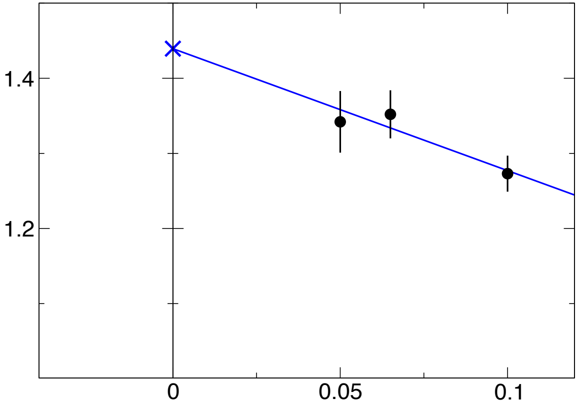
Constant fit reasonable

$Q = 0.15$

These are estimates for what systematics?

☞ Neglecting first order (linear) corrections to constant

Simple estimates



One more data point: error on linear term is now statistical

☞Now we need to estimate systematic due to higher ordes

Systematics

One conservative strategy for systematics:

- Identify all higher order effects you have to neglect
- For each of them:
  - Repeat the entire analysis treating this one effect differently
  - Add the spread of results to systematics
- Important:
  - Do not do suboptimal analyses
  - Do not double-count analyses

make sure there are no unknown unknowns

Let's practice!