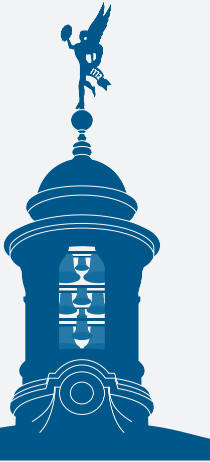
WWU MÜNSTER

Graduiertenkolleg 2149 Research Training Group

nCTEQ nuclear parton distribution functions

Towards a Markov chain Monte Carlo investigation of nuclear PDFs

Seminar at Jefferson Lab

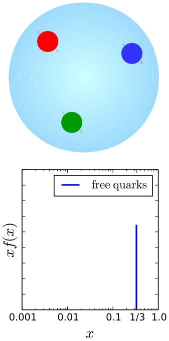


Peter Risse

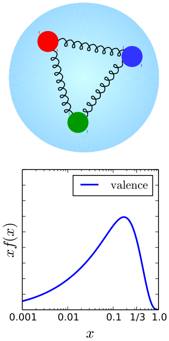
Contents

- overview on **Parton Distribution functions with nuclear effects**
- **reliable estimation** of errors
  - Hessian method
- the **advantages of Markov chain Monte Carlo** algorithms
  - sampling representation of the likelihood
  - autocorrelation: a bridge to lattice QCD
- **speed-up of theory predictions**
  - DIS and heavy quark schemes in `apfel++`
- a **proof of concept study**
  - proton valence PDFs from neutral current DIS

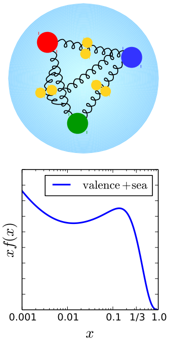
Free quarks



Bound quarks



Bound quarks + QCD effects



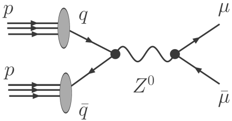
PDF $[f_{a/p}(x,\mu)]$: probability that a parton a carries fraction x of proton's momentum

$$x = \frac{\text{longitudinal parton momentum}}{\text{longitudinal nucleon momentum}} = \frac{p^+_{\text{parton}}}{p^+_{\text{nucleon}}}, \quad \text{where} \quad p^{\pm} = (p^0 \pm p^3)/\sqrt{2}$$

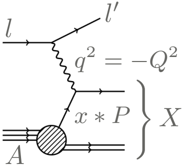
(valid at leading-order of QCD)

On the importance of (nuclear) PDFs

- information on the structure of proton/nucleus
- description of high-energy heavy ion experiments **LHC, RHIC** and **EIC**

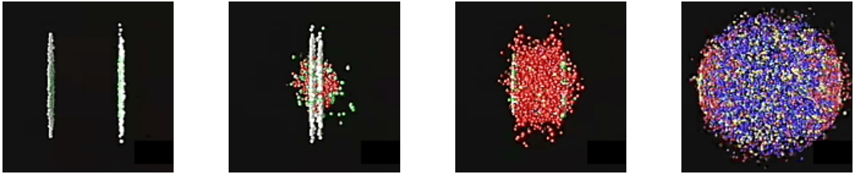


Drell-Yan lepton pair production (DY)

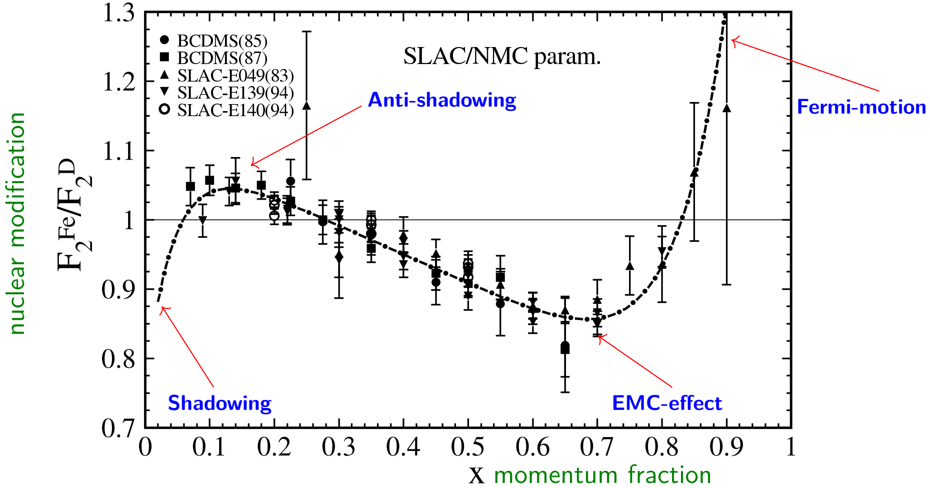


Deep Inelastic Scattering (DIS)

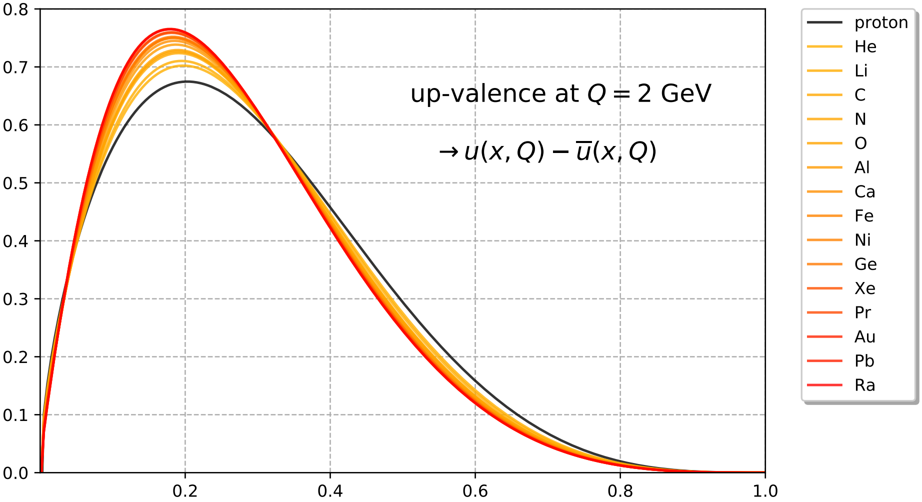
- key ingredient for perturbative probes of **Quark-Gluon-Plasma** (QGP)



Nuclear modification of F_2



Nuclear modification: free proton vs bound proton



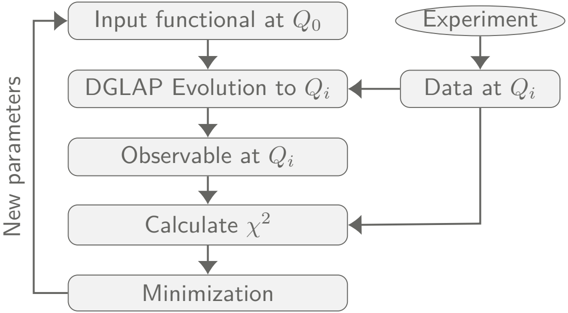
Determination of PDFs

- determine PDFs from experimental data
- the χ^2-function is defined as

$$\chi^2 = \sum_{ij}^{N} (D_i - T_i)\,(C^{-1})_{ij}\,(D_j - T_j)$$

- the covariance matrix is constructed from
  - total uncorrelated uncertainty σ_i^2
  - correlated systematic uncertainty $\overline{\sigma}_{i\alpha}$ from source α

$$C_{ij} = \sigma_i^2 \delta_{ij} + \sum_{\alpha}^{S} \overline{\sigma}_{i\alpha} \overline{\sigma}_{j\alpha}$$



nCTEQ nuclear PDFs parametrization

- define nuclear PDFs by extending the proton PDF parametrization to account for A-dependence.
- PDF of nucleus (A - mass, Z - charge, N - number of neutrons)

$$f_i^{(A,Z)}(x,Q) = \frac{Z}{A} f_i^{p/A}(x,Q) + \frac{N}{A} f_i^{n/A}(x,Q)$$

- bound proton PDFs are parametrized at Q_0

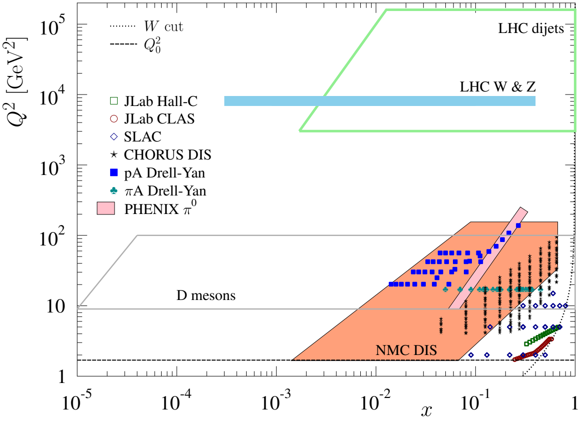
$$x f_i^{p/A}(x,Q_0) = c_0 x^{c_1} (1-x)^{c_2} e^{c_3 x} (1+e^{c_4} x)^{c_5}$$

- bound neutron PDFs are constructed assuming *isospin symmetry* from bound proton PDFs
- A - dependence

$$c_k \to c_k(A) \equiv p_k + a_k \left(1 - A^{-b_k}\right)$$

Available data sets

- one of the latest global analyses: **EPPS21 nuclear PDFs**
- good coverage at mid x
- low coverage at low x and high Q^2
- **fewer data points** compared to proton
  - decreased constraining power
  - have to rely on assumptions
- **assumptions limit the estimation of uncertainties**



K. Eskola et al., arXiv:2112.12462

Markov chain Monte Carlo

Estimation of Errors

Hessian method

- main approximation: likelihood is Gaussian around best fit $\mathbf{c}_0$

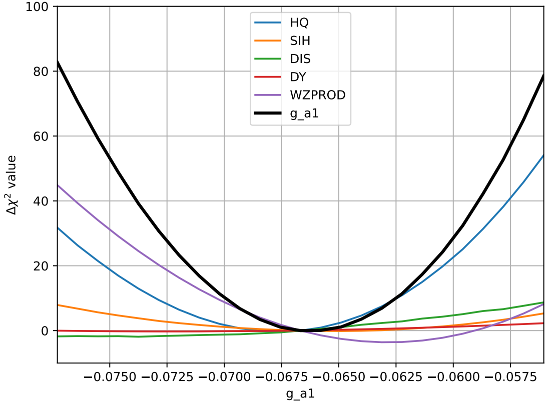
$$\mathcal{L}(\mathbf{c}; D) \propto \exp\left(-\frac{1}{2}\chi^2(\mathbf{c}, D)\right)\Bigg|_{\mathbf{c}_0} \approx \exp\left(-\frac{1}{2}\triangle\mathbf{c}^T H \triangle\mathbf{c}\right) \quad \Rightarrow \quad H_{ij} = \frac{1}{2}\frac{\partial^2\chi^2(\mathbf{c})}{\partial\mathbf{c}_i\partial\mathbf{c}_j}\Bigg|_{\mathbf{c}_0}$$

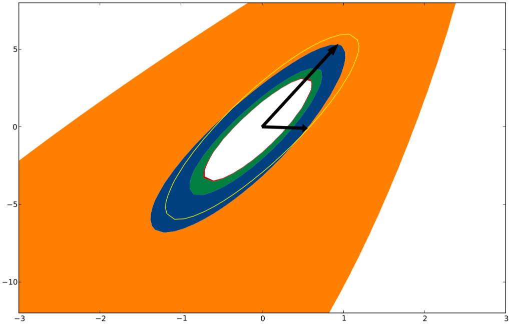
- find rescaled eigendirections of H
- allow **variation of parameters along eigendirections** up to some χ^2-increase of T

⇒ This defines the error envelopes.

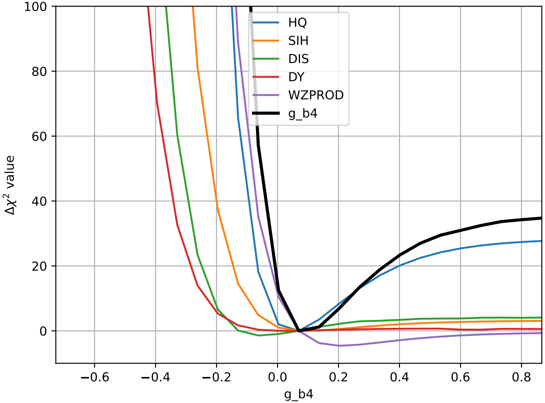
For a recent review see: N. T. Hunt-Smith et al., arXiv:2206.107782

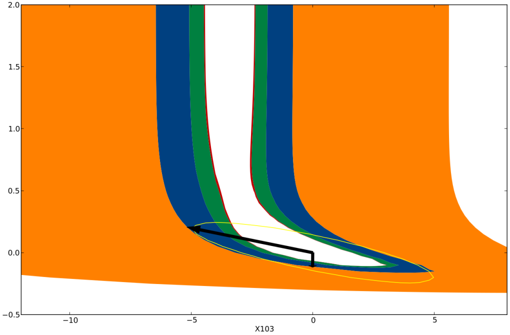
Gaussian likelihoods





non-Gaussian likelihoods





There is a better method...

Markov chain Monte Carlo representation of the likelihood

- **draw random samples** from the posterior function to get a **form independent representation**.

$$\mathrm{post}(\mathbf{c}|D) = \frac{1}{\mathcal{Z}} \exp\left(-\frac{1}{2}\chi^2(\mathbf{c}, D)\right) \rightarrow \{\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_n}\}$$

- **BUT** the samples have to be drawn in such a way that they **reproduce the expectation value and higher modes** of the likelihood

$$E\{\mathcal{O}(\mathbf{c})\} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{O}(\mathbf{c}_i) \qquad \overset{!}{=} \int \mathrm{d}\mathbf{c}\, \mathrm{post}(\mathbf{c}|D)\mathcal{O}(\mathbf{c})$$

$$V\{\mathcal{O}(\mathbf{c})\} = \frac{1}{n}\sum_{i=1}^{n} \left[\mathcal{O}(\mathbf{c}_i) - E\{\mathcal{O}(\mathbf{c})\}\right]^2 \quad \overset{!}{=} \int \mathrm{d}\mathbf{c}\, \mathrm{post}(\mathbf{c}|D)\left[\mathcal{O}(\mathbf{c}) - E\{\mathcal{O}(\mathbf{c})\}\right]^2$$

Markov chain Monte Carlo representation of the likelihood

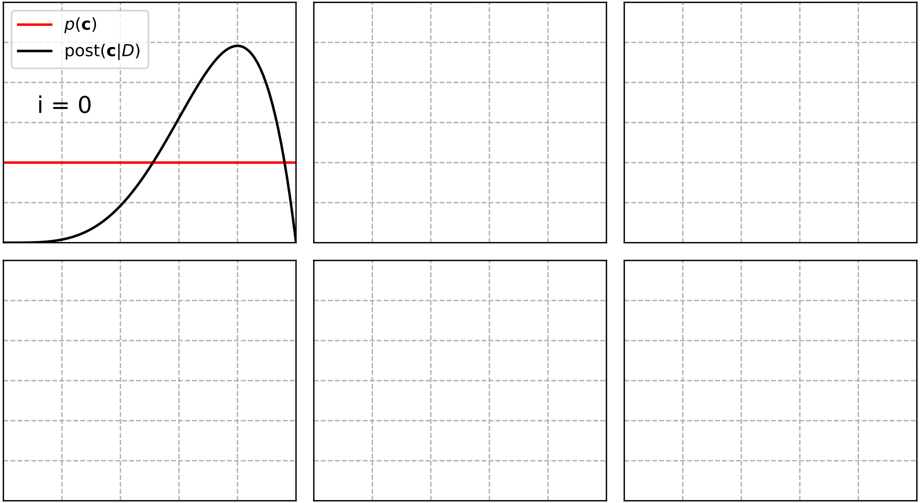
- posterior distribution too complicated to sample directly
  - need clever way to choose Monte Carlo samples
- construct the Monte Carlo samples via a Markov chain

$$\{\mathbf{c}_1 \rightarrow \mathbf{c}_2 \rightarrow \cdots \rightarrow \mathbf{c}_{n-1} \rightarrow \mathbf{c}_n\}$$

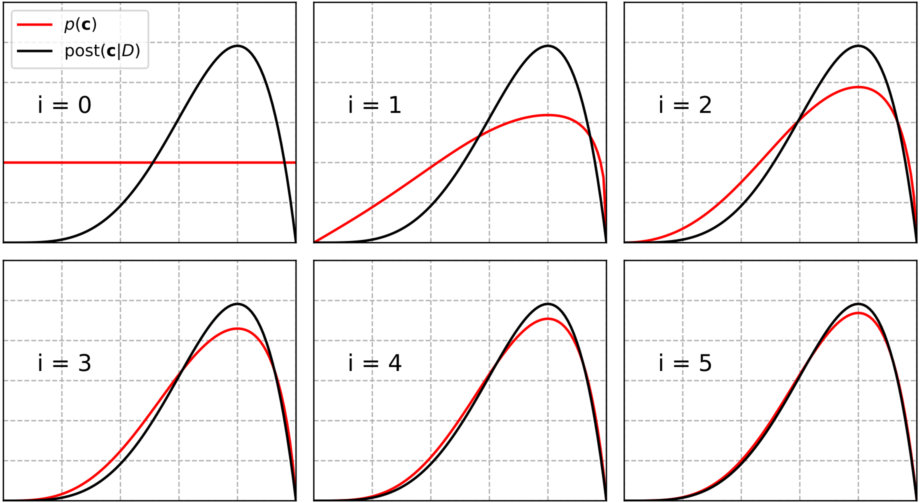
$$\text{with} \quad p_i(\mathbf{c}) = \int \mathrm{d}\mathbf{c}'\, p_{i-1}(\mathbf{c}') T(\mathbf{c}', \mathbf{c})$$

- with the **transition kernel** $T(\mathbf{c}, \mathbf{c}')$
  - has to transform the parameter distribution such that the set of **samples has the desired properties**

Andrey Andreyevich Markov



Time series of parameter distributions $p_i(\mathbf{c})$.



Time series of parameter distributions $p_i(\mathbf{c})$.

Reaching the invariant distribution

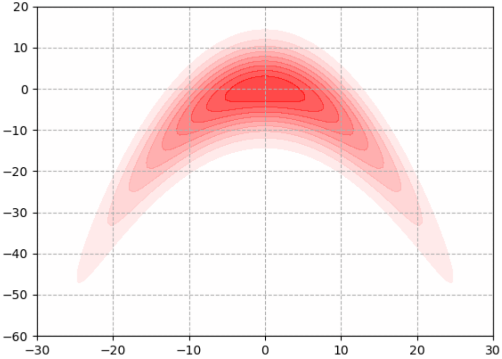
The **invariant distribution** has the property

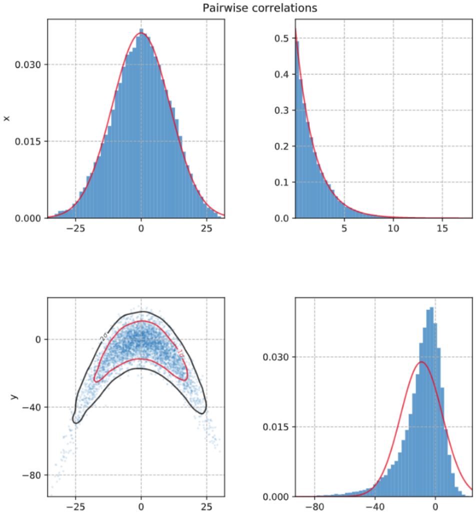
$$\mathrm{post}(\mathbf{c}|D) = \int \mathrm{d}\mathbf{c}' \, \mathrm{post}(\mathbf{c}'|D) T(\mathbf{c}', \mathbf{c})$$

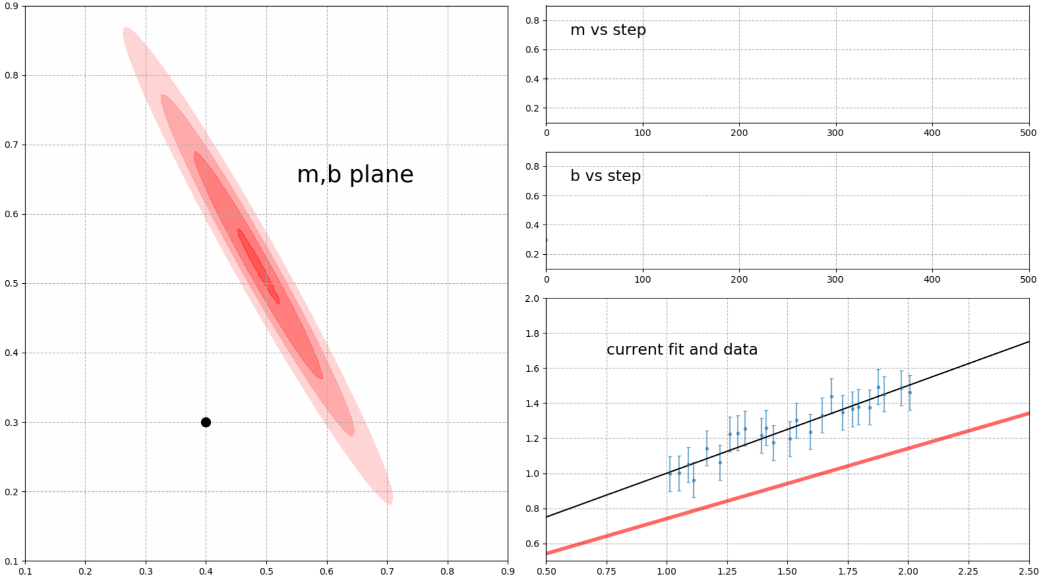
Example: Metropolis-Hastings algorithm

1. Start from state $\mathbf{c}_i$
2. **Propose** new state $\tilde{\mathbf{c}}$ from proposal distribution $q(\tilde{\mathbf{c}}, \mathbf{c}_i)$
   - usually a multidimensional Gaussian around the current state
3. **Accept** new state $\tilde{\mathbf{c}}$ with probability $a(\mathbf{c}_i, \tilde{\mathbf{c}}) = \min\left(1, \frac{\mathrm{post}(\tilde{\mathbf{c}}|D) q(\tilde{\mathbf{c}}, \mathbf{c}_i)}{\mathrm{post}(\mathbf{c}_i|D) q(\mathbf{c}_i, \tilde{\mathbf{c}})}\right)$
   - $T(\mathbf{c}, \tilde{\mathbf{c}}) = q(\mathbf{c}, \tilde{\mathbf{c}}) a(\mathbf{c}, \tilde{\mathbf{c}}) + \delta(\mathbf{c} - \tilde{\mathbf{c}}) \left[1 - \int \mathrm{d}\mathbf{c}' \, q(\mathbf{c}, \mathbf{c}') a(\mathbf{c}, \mathbf{c}')\right]$

Sampling a banana-distribution



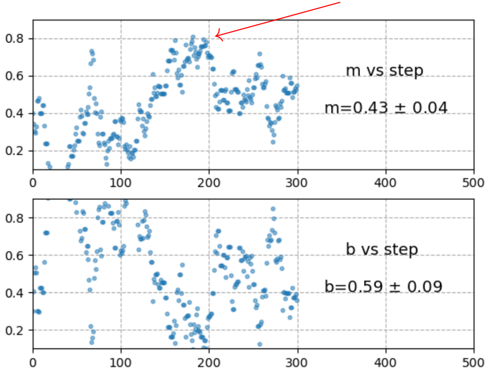



m,b plane
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
m vs step
b vs step
0
100
200
300
400
500
0.2
0.4
0.6
0.8
current fit and data
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
1.0
1.2
1.4
1.6
1.8
2.0

Autocorrelation

- we **cannot use the simple equations** to estimate variances and higher modes
  - these severely underestimate the true uncertainties
- since every new sample depends on the current **the gain in information is reduced**
- this is what is called **autocorrelation**
  - twice the **autocorrelation-time** τ estimates the number of links in the chain **until the next independent sample is drawn**

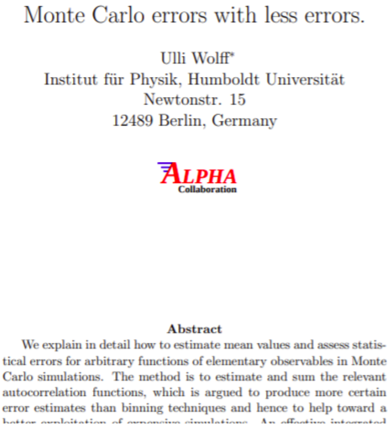
severe substructure



autocorrelation at full force

Bridge to Lattice QCD

- lattice QCD has several methods dealing with this problem
- one example is the **Gamma method**
  - this method estimates the autocorrelation time directly from the chain
  - used to **enlarge error estimates** as to eliminate bias
  - **or filter** the time series to get uncorrelated samples
- other methods: Bootstrap, Jackknife, binning ...



Monte Carlo errors with less errors.

Ulli Wolff*
Institut für Physik, Humboldt Universität
Newtonstr. 15
12489 Berlin, Germany

ALPHA Collaboration

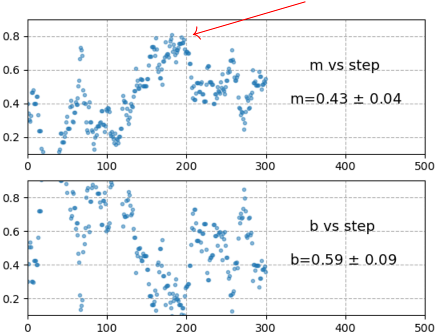
Abstract

We explain in detail how to estimate mean values and assess statistical errors for arbitrary functions of elementary observables in Monte Carlo simulations. The method is to estimate and sum the relevant autocorrelation functions, which is argued to produce more certain error estimates than binning techniques and hence to help toward a better exploitation of expensive simulations. An effective integrated

arXiv:hep-lat/0306017

Filtering based on the Gamma-method

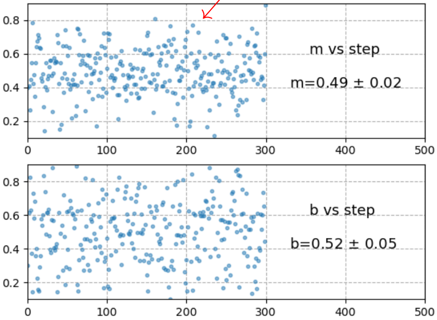
severe substructure



using 300 samples directly

⇒

uniform behaviour



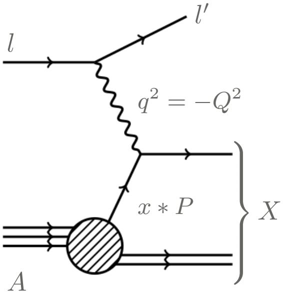
reducing 10^4 samples to a total of 300

speed-up of theory predictions

Optimizing DIS theory predictions

Factorization in DIS structure functions

$$F_\lambda(x,Q^2) = \sum_k C_k^\lambda \otimes f_k = \sum_k \int_\chi^1 \frac{\mathrm{d}\xi}{\xi} C_k^\lambda \left(\frac{\chi}{\xi}, \frac{Q}{\mu}, \frac{m_i}{\mu}, \alpha_s(\mu) \right) f_k(\xi,\mu)$$



- Wilson coefficients have a **complicated α_s** expansion
  - these are the hard scattering amplitudes
- **heavy quark mass effects** important at $Q \sim m_H$
- bulk of experimental data is from DIS
  - need fast theory predictions
- older implementations are **not well optimized**

DIS mass schemes

Zero Mass Variable Flavor Number Scheme (ZMVFNS)

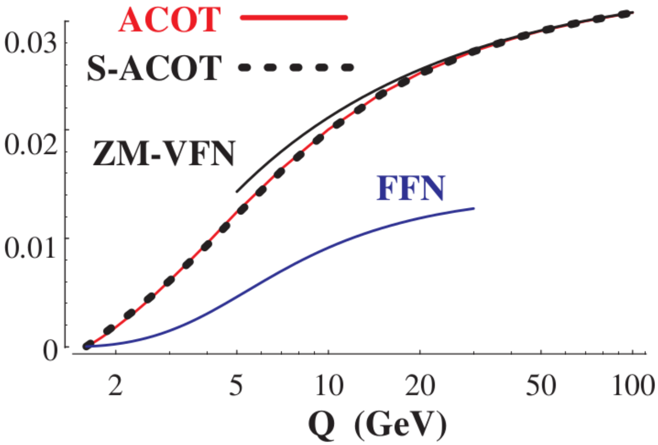
- consider only quarks below threshold: $m_q < Q$
- neglect all mass terms part of the Wilson coefficients
- do not take phase space constraints into account
- **simple but only works far from threshold $\mathbf{Q \gg m_q}$**

Fixed Flavor Number Scheme (FFNS)

- treat all quarks as massless except for the heaviest m_H
- this mass appears explicitly in the Wilson coefficients
- **good results for $\mathbf{Q \sim m_H}$ unreliable as Q becomes large**

General Mass Variable Flavor Number Schemes (GMVFNS)

- 'interpolating' between FFNS and ZMVFNS
- several choices can be made, resulting in different schemes:
  - **ACOT**: minimal extension of the $\overline{\mathrm{MS}}$ renormalisation scheme
  - **FONLL**: interpolating between schemes with a damping function
  - **TR-method**: requiring smooth transition at $Q = m_H$



T. Stavreva et al., arXiv: 1203.0282

APFEL++ – A PDF evolution library in c++

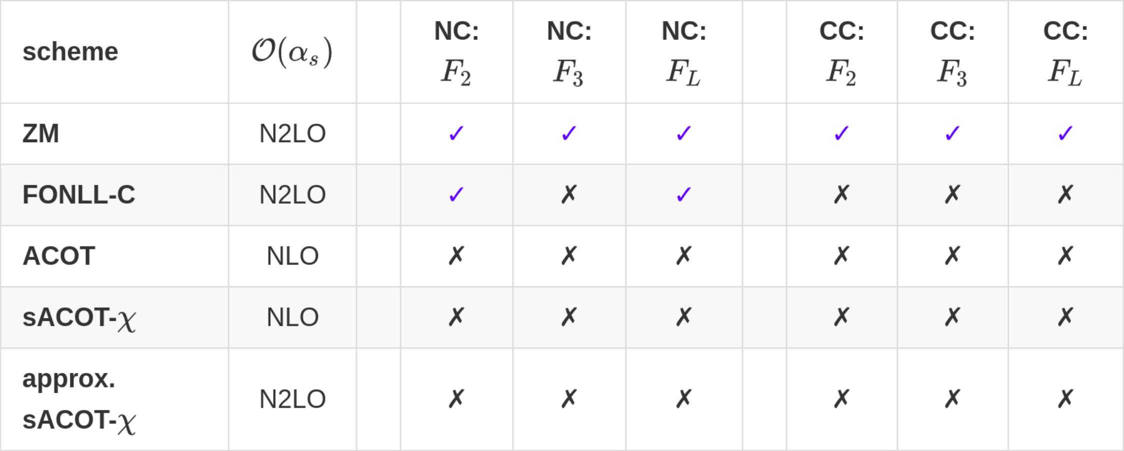
- main author: **V. Bertone**
- rewrite of the Fortran APFEL code
  - used by the NNPDF collaboration
- focus on **fast and memory efficient** implementations
- codes that use APFEL++
  - nCTEQ++
  - xFitter
  - NangaParbat
  - MontBlanc
  - PARTONS

Features:

- DGLAP evolution equations
- Deep Inelastic Scattering with or without mass effects
- single-inclusive-annihilation cross sections
- differential semi-inclusive DIS
- Drell-Yan cross sections

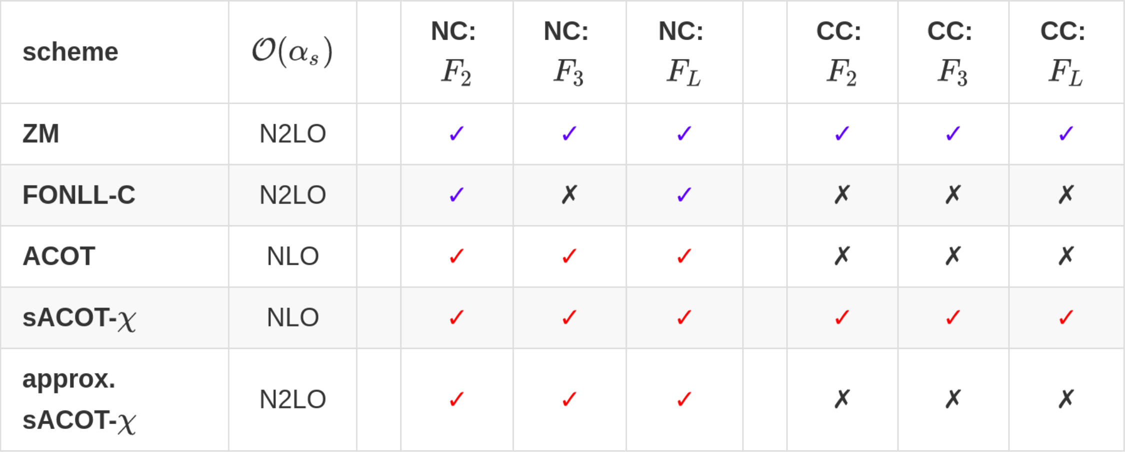
https://github.com/vbertone/apfelxx, arXiv: 1708.00911

Available schemes in APFEL++



scheme	$\mathcal{O}(\alpha_s)$		NC: F_2	NC: F_3	NC: F_L		CC: F_2	CC: F_3	CC: F_L
ZM	N2LO		✓	✓	✓		✓	✓	✓
FONLL-C	N2LO		✓	✗	✓		✗	✗	✗
ACOT	NLO		✗	✗	✗		✗	✗	✗
sACOT-χ	NLO		✗	✗	✗		✗	✗	✗
approx. sACOT-χ	N2LO		✗	✗	✗		✗	✗	✗

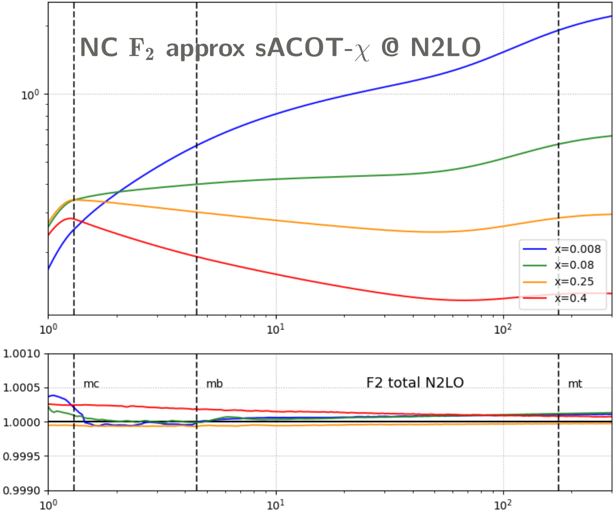
Available schemes in APFEL++ (new)



scheme	$\mathcal{O}(\alpha_s)$		NC: F_2	NC: F_3	NC: F_L		CC: F_2	CC: F_3	CC: F_L
ZM	N2LO		✓	✓	✓		✓	✓	✓
FONLL-C	N2LO		✓	✗	✓		✗	✗	✗
ACOT	NLO		✓	✓	✓		✗	✗	✗
sACOT-χ	NLO		✓	✓	✓		✓	✓	✓
approx. sACOT-χ	N2LO		✓	✓	✓		✗	✗	✗

Available schemes in APFEL++

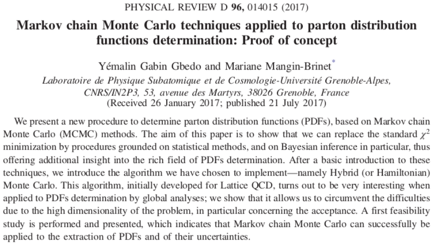
- **very good agreement** with old implementation in all kinematic regions
    - compared to the nCTEQ++ code
- speed-up to current implementation: $\mathcal{O}(100)$
- a chain with $\sim 10^4$ samples **is feasible in finite time**
- planned: make this available also via the xFitter code



Proof of concept study

Proof of concept: proton valence PDFs from HERA data

- 10 dimensional proton valence PDF-fit
- experimental data: **H1 and ZEUS data**
  - total: 537 points
- theory prediction: **ZMVFNS at NLO** from the `xFitter` code



PHYSICAL REVIEW D **96,** 014015 (2017)

Markov chain Monte Carlo techniques applied to parton distribution functions determination: Proof of concept

Yémalin Gabin Gbedo and Mariane Mangin-Brinet[*]

Laboratoire de Physique Subatomique et de Cosmologie-Université Grenoble-Alpes, CNRS/IN2P3, 53, avenue des Martyrs, 38026 Grenoble, France

(Received 26 January 2017; published 21 July 2017)

We present a new procedure to determine parton distribution functions (PDFs), based on Markov chain Monte Carlo (MCMC) methods. The aim of this paper is to show that we can replace the standard χ^2 minimization by procedures grounded on statistical methods, and on Bayesian inference in particular, thus offering additional insight into the rich field of PDFs determination. After a basic introduction to these techniques, we introduce the algorithm we have chosen to implement—namely Hybrid (or Hamiltonian) Monte Carlo. This algorithm, initially developed for Lattice QCD, turns out to be very interesting when applied to PDFs determination by global analyses; we show that it allows us to circumvent the difficulties due to the high dimensionality of the problem, in particular concerning the acceptance. A first feasibility study is performed and presented, which indicates that Markov chain Monte Carlo can successfully be applied to the extraction of PDFs and of their uncertainties.

Problems:

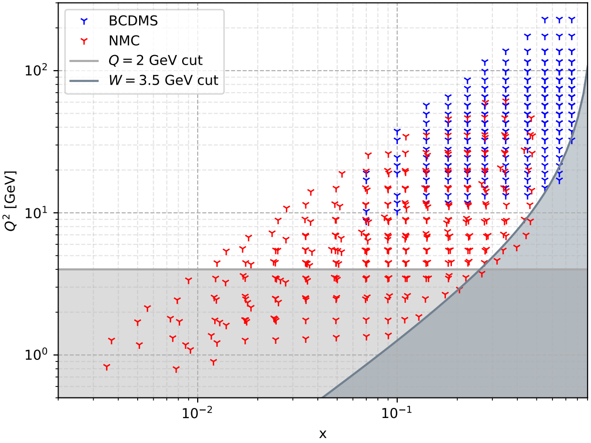
- not pursued by the authors
- only vague description of technical implementation

Experimental data sets

- Experimental measurements from the **BCDMS** and **New Muon Collaboration**
- F_2 measurements for proton and deuteron: **992 data points** after cuts
  - approximate deuteron as sum of proton and neutron
- relate proton and neutron PDFs via isospin symmetry

invariant mass

$$W^2 = M_p^2 + \frac{1-x}{x} Q^2 \geq 3.5\mathrm{GeV}$$



Fitting setup

up- and down-valence distributions: *valence* = *quark* - *anti-quark*

$$xf(x,Q_0) = c_0 x^{c_1}(1-x)^{c_2} e^{c_3 x}(1+e^{c_4}x)^{c_5}$$

$$u_v \;\rightarrow\; \{c_1, c_2, c_3, c_4, c_5\}$$

$$d_v \;\rightarrow\; \{c_1, c_2, c_3, c_4, c_5\}$$

- **992 experimental data points**, **10 parameters** to fit
- proposal algorithm:

?

Choosing the proposal distribution – Adaptive Metropolis-Hastings

1. Use **normal random walk Metropolis-Hastings** until N_0 samples have been obtained
   - proposal distribution: multivariate Gaussian

$$\tilde{\mathbf{c}}_{i+1} \quad \text{proposed from} \quad q(\tilde{\mathbf{c}}_{i+1}, \mathbf{c}_i) = \mathcal{N}(\mathbf{c}_i, C_0) \quad \text{with} \quad C_0: \ \text{covariance matrix from user input}$$

2. switch to a **self learning proposal distribution**

$$\tilde{\mathbf{c}}_{i+1} \quad \text{proposed from} \quad q(\tilde{\mathbf{c}}_{i+1}, \mathbf{c}_i) = (1-\beta)\mathcal{N}\left(\mathbf{c}_i, \text{scale} \cdot \overline{C}_i\right) + \beta\mathcal{N}(\mathbf{c}_i, C_0)$$

with self learned $\overline{\mathbf{C}}_{\mathrm{i}}$

   - $0 \leq \beta \leq 1$ controls the impact of the 'learned' proposal

3. reset self learned proposal distribution **to boost convergence**
   - this reduces the impact of the starting point

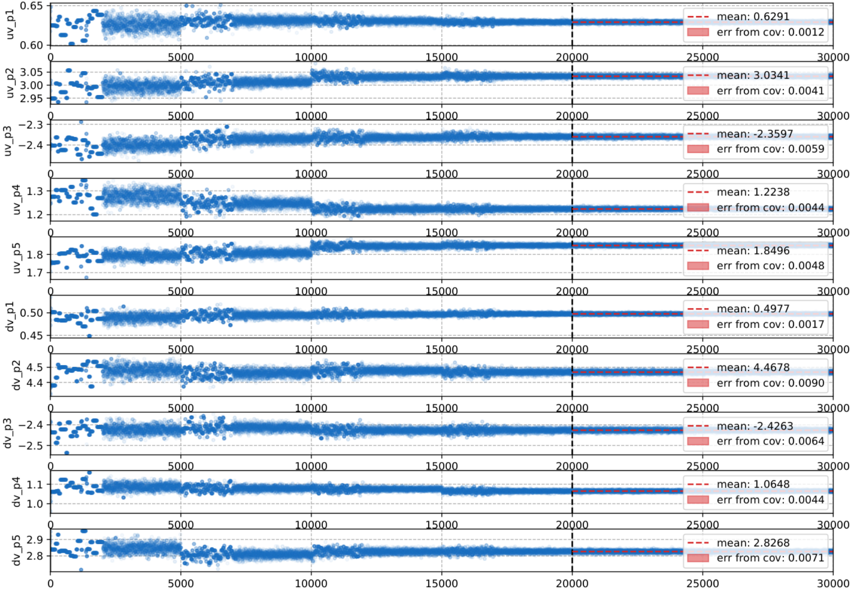
H. Haario et al.: “An adaptive Metropolis algorithm”, *Bernoulli* 7.2 (Apr. 2001)

Fitting setup

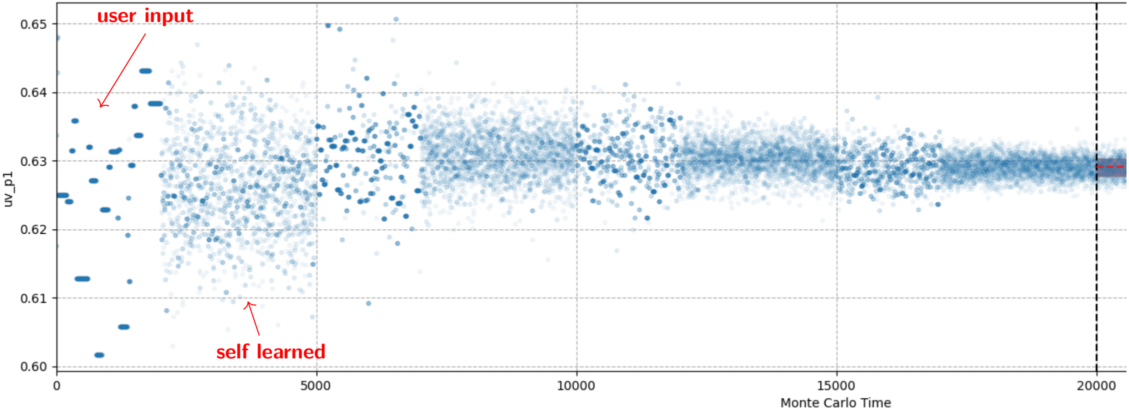
up- and down-valence distributions: *valence* = *quark* - *anti-quark*

$$xf(x,Q_0) = c_0 x^{c_1}(1-x)^{c_2} e^{c_3 x}(1+e^{c_4}x)^{c_5}$$
$$u_v \ \rightarrow \ \{c_1, c_2, c_3, c_4, c_5\}$$
$$d_v \ \rightarrow \ \{c_1, c_2, c_3, c_4, c_5\}$$

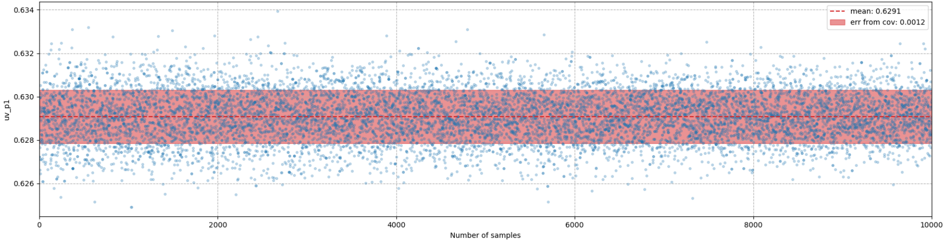
- **992 experimental data points**, **10 parameters** to fit
- proposal algorithm: Adaptive Metropolis-Hastings with 3x resets
- one very long run: 30,000 samples
  - convergence after 20,000 samples → **10,000 analysable samples**
  - computing time: 6.5 days
- implemented within the `nCTEQ++` code

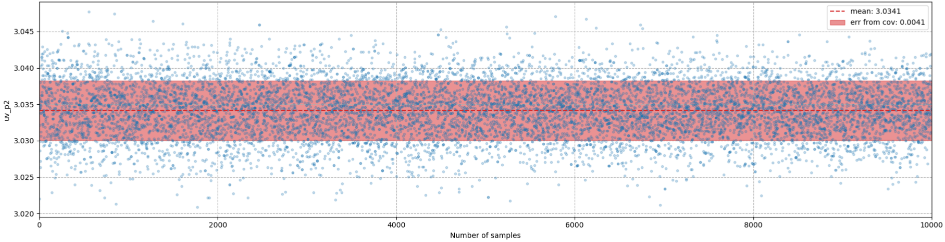

uv_p1
0.65
0.60
mean: 0.6291
err from cov: 0.0012
uv_p2
3.05
3.00
2.95
mean: 3.0341
err from cov: 0.0041
uv_p3
−2.3
−2.4
mean: -2.3597
err from cov: 0.0059
uv_p4
1.3
1.2
mean: 1.2238
err from cov: 0.0044
uv_p5
1.8
1.7
mean: 1.8496
err from cov: 0.0048
dv_p1
0.50
0.45
mean: 0.4977
err from cov: 0.0017
dv_p2
4.5
4.4
mean: 4.4678
err from cov: 0.0090
dv_p3
−2.4
−2.5
mean: -2.4263
err from cov: 0.0064
dv_p4
1.1
1.0
mean: 1.0648
err from cov: 0.0044
dv_p5
2.9
2.8
mean: 2.8268
err from cov: 0.0071
0
5000
10000
15000
20000
25000
30000

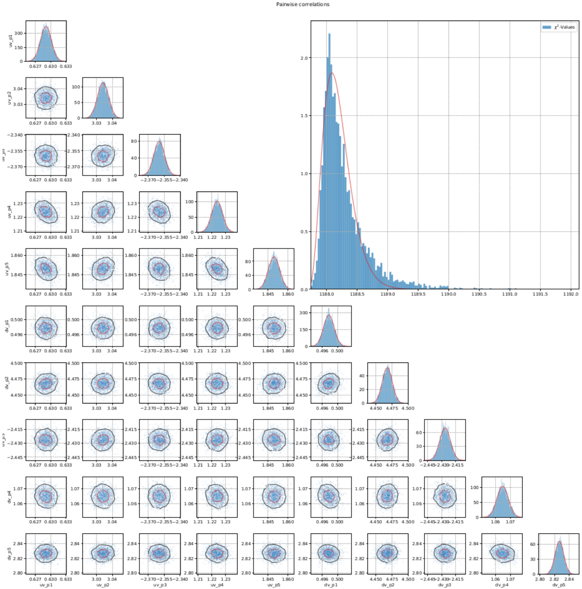
Burn-in phase



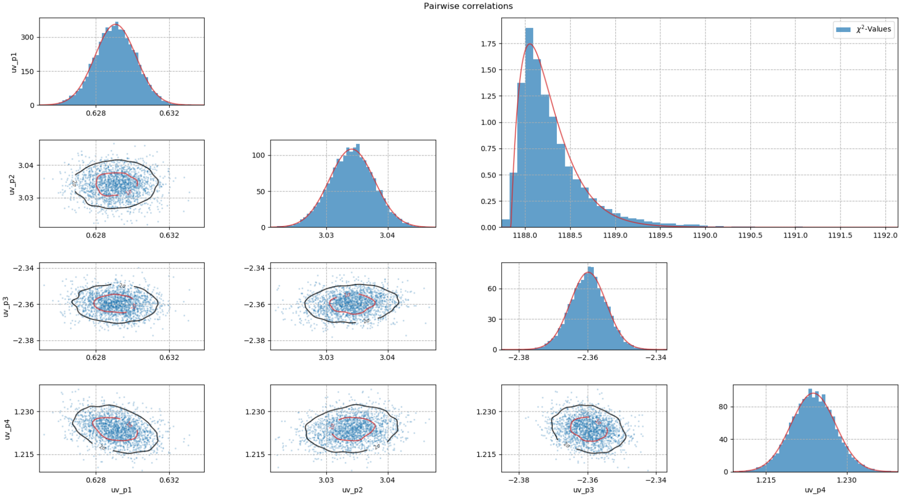
Converged part of the chain



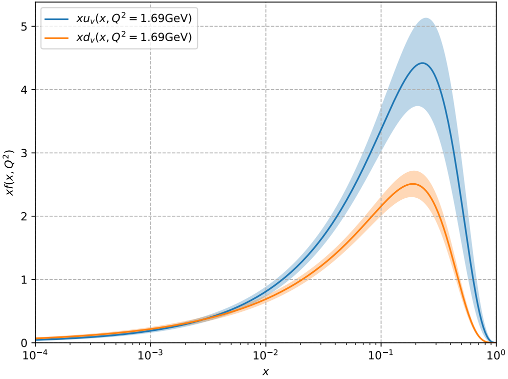



Pairwise correlations
χ²-Values
uv_p1
uv_p2
uv_p3
uv_p4
uv_p5
dv_p1
dv_p2
dv_p3
dv_p4
dv_p5
0.627 0.630 0.633
3.03 3.04
−2.370 −2.355 −2.340
1.21 1.22 1.23
1.845 1.860
0.496 0.500
4.450 4.475 4.500
−2.445 −2.430 −2.415
1.06 1.07
2.80 2.82 2.84
1188.0 1188.5 1189.0 1189.5 1190.0 1190.5 1191.0 1191.5 1192.0
0.0 0.5 1.0 1.5 2.0

Pairwise correlations

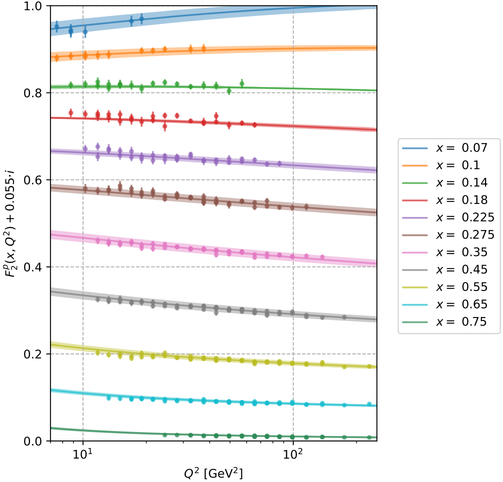


Results



valence-pdf results

BCDMS



Conclusion

- existing error **PDF estimation is limited**
  - parameter distributions **need to be close to a Gaussian**
- Markov Chain Monte Carlo algorithms are able to **access errors without any approximation** of the posterior distribution
- the **autocorrelation in the parameter samples** can be tackled
  - a **better proposal distribution**
  - the **Gamma-method** (from lattice QCD)
- a speed-up of calculations can be done by **griding the required observables** in beforehand
  - big **update on the heavy quark schemes** in `apfel++`
- a successful **proof of concept study**
  - 10 parameter fit for proton valence distributions
  - experimental data from DIS

backup

Estimation of Errors

Data resampling

- create a **set of pseudodata replicas**
  multidimensional Gaussian with mean and uncertainties from original data
- obtain **maximum likelihood estimation** (i.e. minimize χ^2) for each replica
- best estimate and standard deviation from

$$E\{\mathcal{O}(\mathbf{c})\} = \frac{1}{n_{\text{rep}}} \sum^{n_{\text{rep}}} \mathcal{O}(\mathbf{c}_{\text{rep}})$$

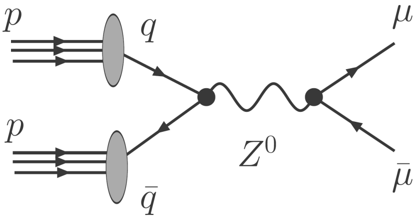
$$V\{\mathcal{O}(\mathbf{c})\} = \frac{1}{n_{\text{rep}}} \sum^{n_{\text{rep}}} \left[\mathcal{O}(\mathbf{c}_{\text{rep}}) - E\{\mathcal{O}(\mathbf{c})\}\right]^2$$

- works best **if the likelihood is Gaussian around the best fit**

For a recent review see: N. T. Hunt-Smith et al., arXiv:2206.107782

Speed-up of theoretical predictions – Hadron collider

$$\sigma_{pp\to X} = \sum_s^{partons} \sum_p \int \mathrm{d}x_1 \mathrm{d}x_2 \, \hat{\sigma}^{(s)(p)} \alpha_s^p(Q^2) F^{(s)}(x_1, x_2, Q^2) \ , \ F^{(s)} = \sum_{ij} f_i(x_1, Q^2) f_j(x_2, Q^2)$$



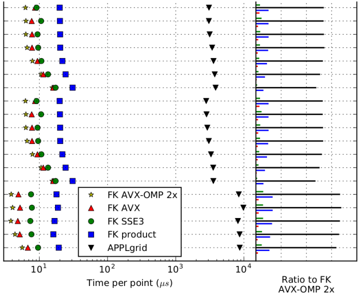
- computationally **expensive double integrals**
- increasing amount of experimental observables
- solution `APPLgrid`
  - interpolate the PDFs
  - precompute the integrals by including the interpolating functions as grids
  - now convolute grids with any pdf to get prediction

T. Carli, D. Clements et al., arXiv:0911.2985

Speed-up of theoretical predictions – Hadron collider

- `APPLgrid` is still too slow for several reasons
  - convolution of the grid with the PDFs is **not well optimized**
  - before one can convolute one has to compute the DGLAP evolution to get the PDFs at every Q

- solution **fast convolution tables** (FK-tables) by `APFELgrid`
  - combines `APPLgrid` tables with DGLAP-evolution tables
    - only need the PDFs at Q_0
  - well optimized by making use of vectorisation and multiprocessing
  - **possible speed-up** compared to `APPLgrid`: $\mathcal{O}(2) - \mathcal{O}(10^3)$



V. Bertone et al., arXiv:1605.02070