Nuclear PDF Determination
via Markov Chain Monte Carlo Methods

Nasim Derakhshanian
Dr. Aleksander Kusina
IFJ PAN
Department of Theoretical Particle Physics

Peresenting at **Bialasowka** Seminar, AGH, Krakow, 26th Apr 2024

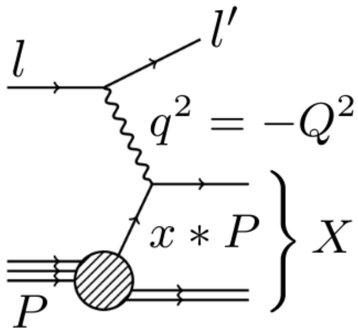
Parton Distribution Function (PDF):

The probability $\mathbf{f_{a/p}(x,\mu)}$ that a parton **a** carries fraction **x** of the proton's momentum

μ: Factorization scale

x: momentum fraction

- **Factorization** in case of Deep Inelastic Scattering (DIS)



$$\frac{d^2\sigma}{dx dQ^2} = \sum_{i=q,\bar{q},g} \int_x^1 \frac{dz}{z} f_i(z,\mu) d\hat{\sigma}_{il\to l'X}\left(\frac{x}{z}, \frac{Q}{\mu}\right)$$

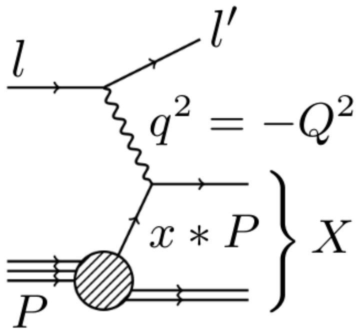
proton PDFs of parton i **parton level matrix element**

Parton Distribution Function (PDF):

The probability $\mathbf{f_{a/p}(x,\mu)}$ that a parton **a** carries fraction **x** of the proton's momentum

μ: Factorization scale

x: momentum fraction



➢ **Factorization** in case of Deep Inelastic Scattering (DIS)

$$\frac{d^2\sigma}{dx dQ^2} = \sum_{i=q,\bar{q},g} \int_x^1 \frac{dz}{z} f_i(z,\mu) d\hat{\sigma}_{il\to l'X}\left(\frac{x}{z}, \frac{Q}{\mu}\right)$$

proton PDFs of parton i **parton level matrix element**

PDF properties:

- Universal (independent of the process)
- Constrained through momentum and number sum rules
- μ^2-dependence governed by DGLAP evolution equations
- **Non-perturbative**: x-dependence of PDF is NOT calculable in pQCD

→ **Global PDF Fit**: using data at different scales and processes to extract PDFs

Nuclear PDFs (nPDFs):

nPDF describes the momentum distribution of partons (quarks and gluons) inside a nucleus

$$F_2^A(x) \neq Z F_2^p(x) + N F_2^n(x)$$

Nuclear PDFs (nPDFs):

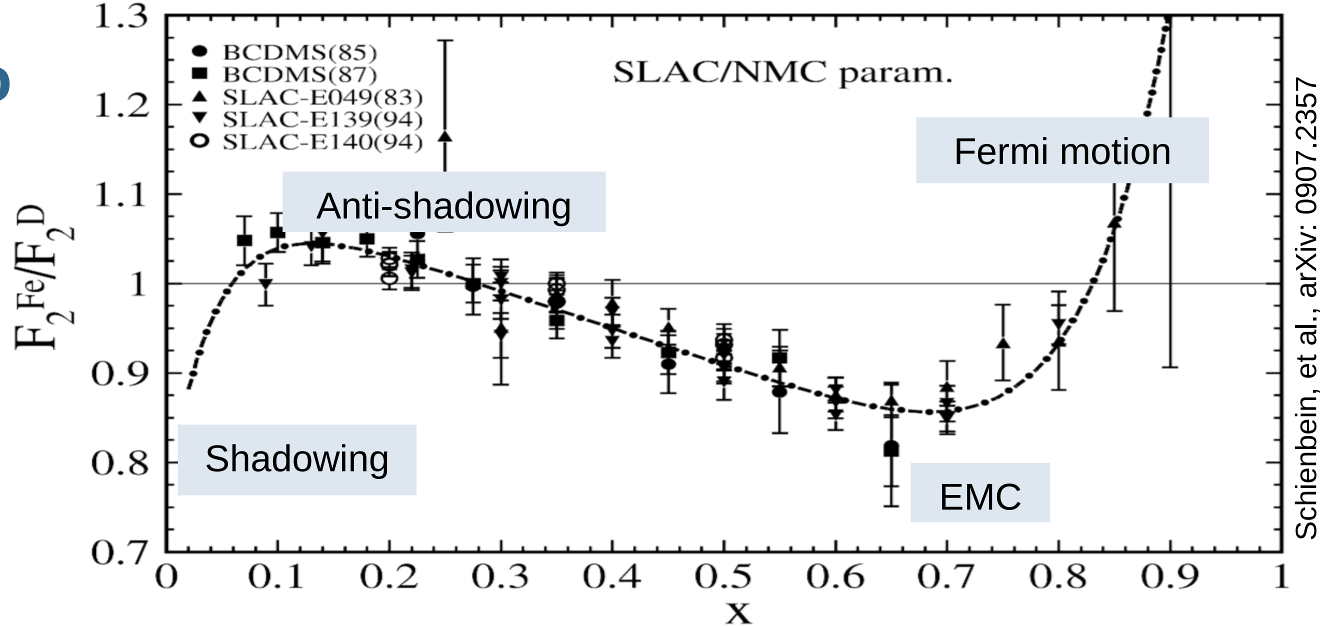
nPDF describes the momentum distribution of partons (quarks and gluons) inside a nucleus

$$F_2^A(x) \neq Z F_2^p(x) + N F_2^n(x)$$

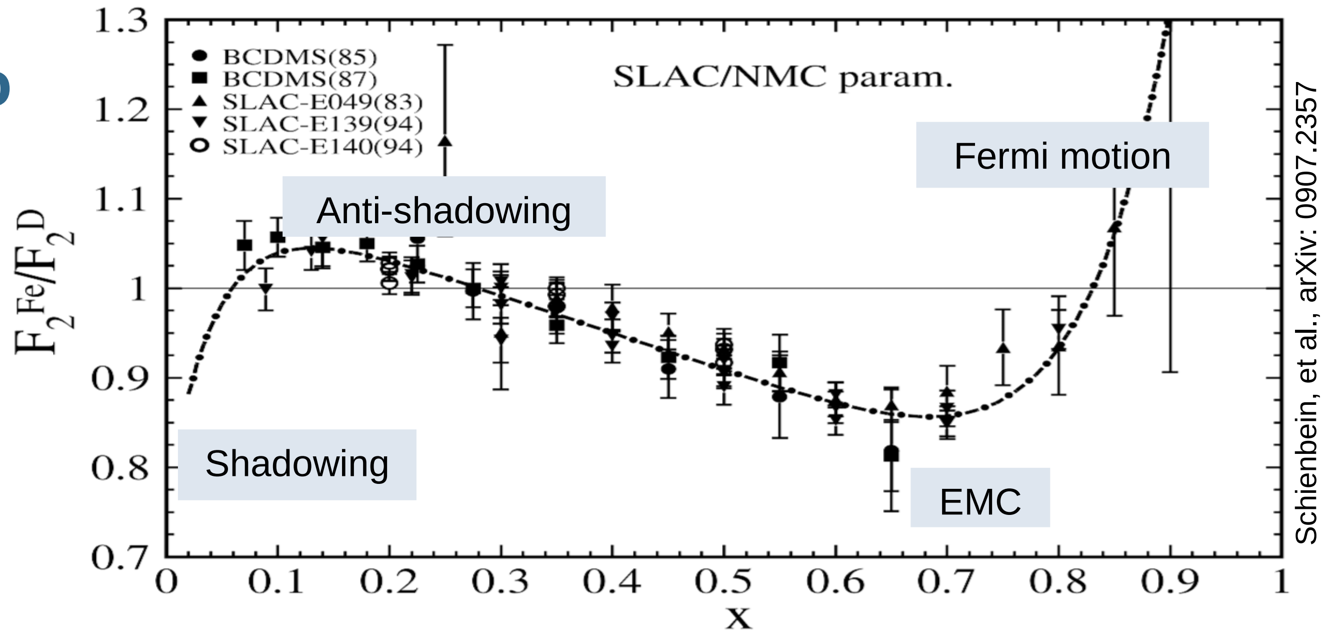
Where are nPDFs useful?

- **High-Energy Collider Physics (LHC & RHIC)**
  essential for predicting the outcomes of collisions involving nuclear targets
- **Neutrino Physics**
  Nuclei are used as targets in neutrino scattering experiments to increase the interaction probability
- **Nuclear Structure**
  provide a deeper insights into our understanding of nuclear matter.

Nuclear correction ratio



Nuclear correction ratio

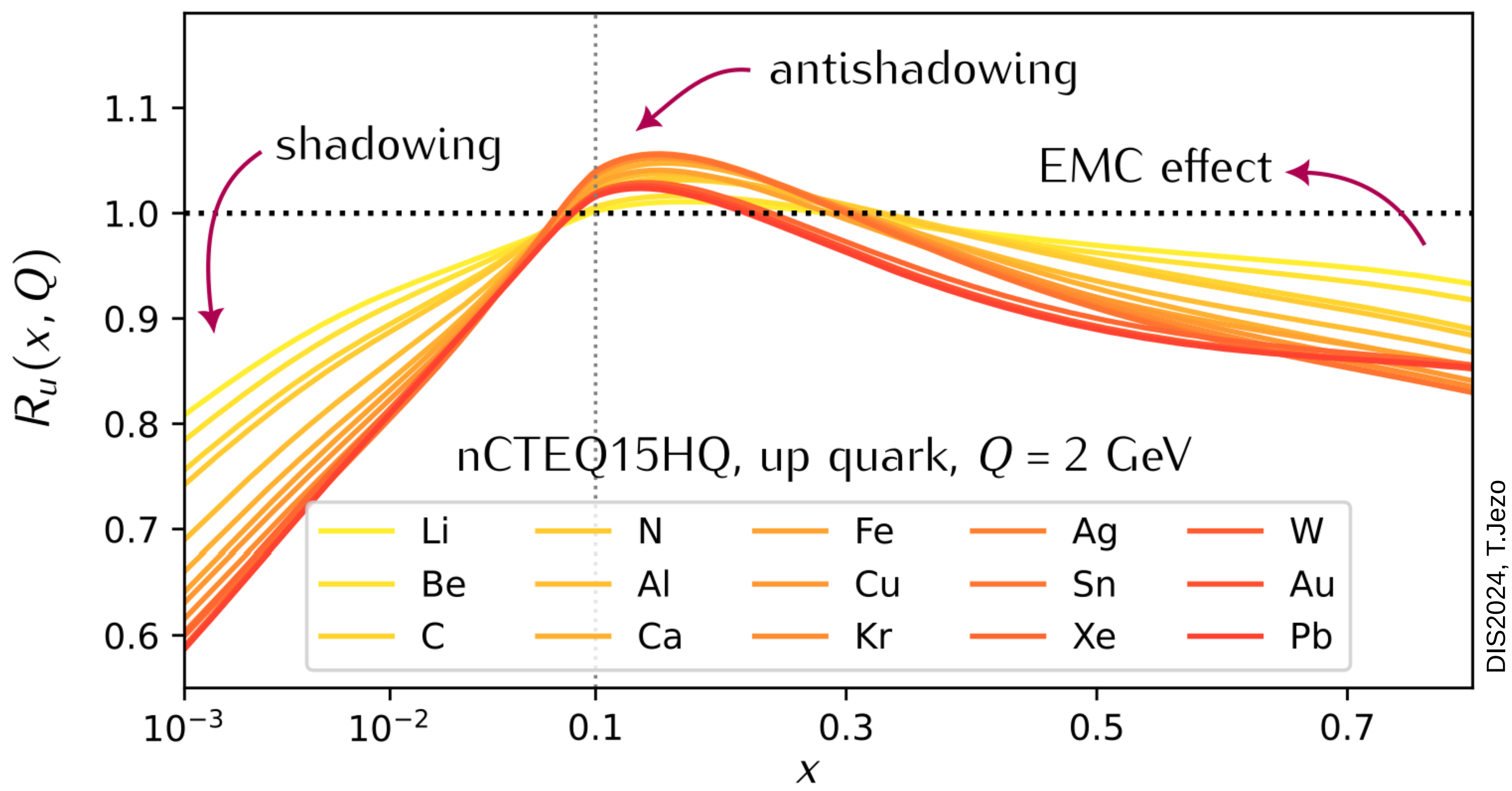


- **Shadowing**: a suppression due to the overlap of partons from different nucleons at low x which reduce the chance of interacting with the probe
- **Anti-Shadowing**: an enhancement of parton densities, compensates for shadowing based on the momentum sum rule.
- **EMC effect**: a reduction in parton densities due to nuclear binding, Pion Excess, quark clusters, Short-Range Correlations, etc.
- **Fermi motion**: an increase at high x, attributed to the intrinsic motion of nucleons within the nucleus

The underlying dynamics are still to be fully theoretically understood!

Nuclear correction ratio

$$R_i^A(x,Q^2) = f_i^{p/A}(x,Q^2)/f_i^p(x,Q^2)$$



Theoretical Framework for Nuclear PDF:

Nuclear modifications can be incorporated into the PDF framework:

1. Factorization*: We assume that the nuclear effects can be absorbed into the universal nPDFs.

$$\sigma_{pA\to X} = \underbrace{f^p(x_1,\mu^2)}_{\text{free proton PDF}} \otimes \underbrace{f^A(x_2,\mu^2)}_{\text{nuclear PDF}} \otimes \hat{\sigma}(x_1,x_2,\mu^2)$$

2. Bound proton PDF $f^{p/A}$ satisfies the same **evolution equations** and **sum rules** as free proton PDF.

3. Isospin symmetry: $f_{d,u}^{n/A} = f_{u,d}^{p/A}$

nuclear PDF:

$$\boxed{f_i^{(A,Z)} = \frac{Z}{A} f_i^{p/A} + \frac{A-Z}{A} f_i^{n/A}}$$

*Proof of factorization for nuclear collisions not yet available.

Global Analysis of nPDF

Input function at Q_0

Parameterize nuclear PDF at initial scale: 1.3GeV

DGLAP evolution

Compute theory predictions at scale Q by solving DGLAP equations

Construct χ^2 function

Calculate the goodness of fit in terms of theory predictions, data and uncertainties

Minimization

Minimize χ^2 function with respect to nPDF parameters

Experimental data

Choose experimental data (e.g. DIS, DY, W/Z, etc.) and apply kinematical cuts

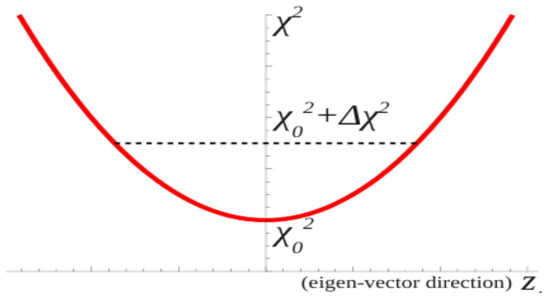
$$\chi^2(\{a_i\}) = \sum_j^N \left(\frac{(\mathrm{data}_j - \mathrm{theory}_j(\{a_i\}))}{\sigma_j} \right)^2$$

Uncertainties estimation

NPDF uncertainties estimation

The **Hessian** method is widely used for error estimation in both proton and nuclear PDFs.

It relies on the quadratic behavior of the χ^2 function near the minimum.



Shortcomings:

- Non-gaussian errors
- Global minima judgment
- Choice of χ^2 tolerance

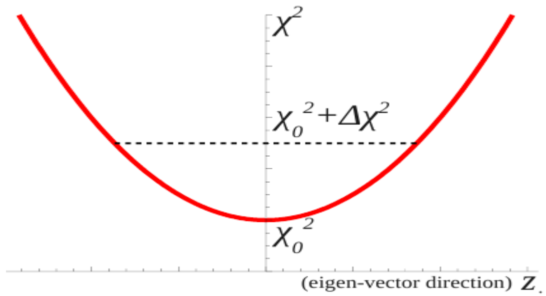
nPDF difficulties:

- Lacking data (range and precision of data for nuclei are generally lower than for proton)
- Complexity and nature of nuclear effects

NPDF uncertainties estimation

The **Hessian** method is widely used for error estimation in both proton and nuclear PDFs.

It relies on the quadratic behavior of the χ^2 function near the minimum.



Shortcomings:

- Non-gaussian errors
- Global minima judgment
- Choice of χ^2 tolerance

nPDF difficulties:

- Lacking data (range and precision of data for nuclei are generally lower than for proton)
- Complexity and nature of nuclear effects

Markov Chain Monte Carlo method

advanced statistical method as an alternative for Hessian

Global Analysis of nPDF

Input function at Q_0

Parameterize nuclear PDF at initial scale: 1.3GeV

DGLAP evolution

Compute theory predictions at scale Q by solving DGLAP equations

Construct χ^2 function

Calculate the goodness of fit in terms of theory predictions, data and uncertainties

Minimization

Minimize χ^2 function with respect to nPDF parameters

Experimental data

Choose experimental data (e.g. DIS, DY, W/Z, etc.) and apply kinematical cuts

$$\sum_j \left(\frac{(\mathrm{data}_j - \mathrm{theory}_j(\{a_i\}))}{\sigma_j} \right)^2$$

MCMC method

Uncertainties estimation

Markov Chain Monte Carlo (MCMC)

A sequence of random variables where the current value is dependent on the value of the prior variable (Memory-less property)

A technique for randomly sampling a probability distribution and approximating a desired quantity.

Bayes theorem:

$$\underbrace{p(\theta \mid \text{data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{data} \mid \theta)}^{\text{Likelihood}} \cdot \overbrace{p(\theta)}^{\text{Prior}}}{\underbrace{p(\text{data})}_{\text{Normalization}}}$$

Prior: initial belief about the parameter before considering the data.
Likelihood: probability of observing the data given a specific value of the parameter.
Posterior: updated belief about the parameter given the data.

- We aim to find the set of nPDF parameters that maximizes the posterior probability distribution given the experimental data.

Likelihood: $$p(data|\theta) \propto \exp\left(-\frac{\chi^2}{2}\right)$$

$$\chi^2(\{a_i\}) = \sum_j^N \left(\frac{(\mathrm{data}_j - \mathrm{theory}_j(\{a_i\}))}{\sigma_j}\right)^2$$

Statistical error
Correlated and uncorrelated systematic errors

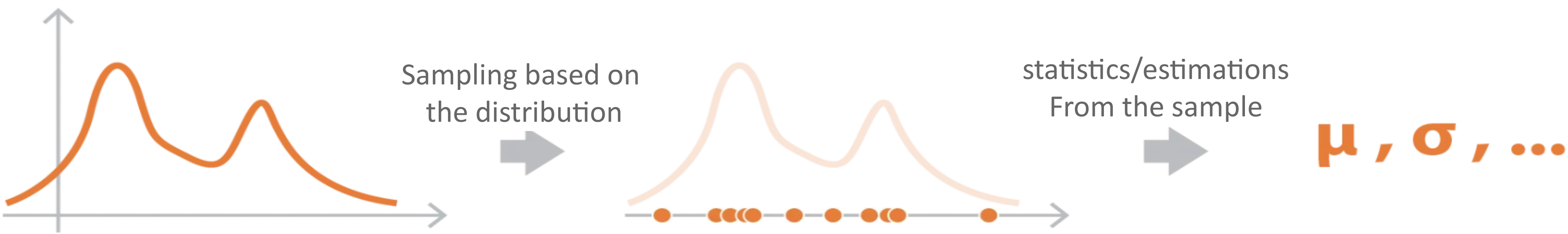
- We aim to find the set of nPDF parameters that maximizes the posterior probability distribution given the experimental data.

Likelihood: $$p(data|\theta) \propto \exp\left(-\frac{\chi^2}{2}\right)$$

$$\chi^2(\{a_i\}) = \sum_j^N \left(\frac{(\mathrm{data}_j - \mathrm{theory}_j(\{a_i\}))}{\sigma_j}\right)^2$$

Statistical error
Correlated and uncorrelated systematic errors

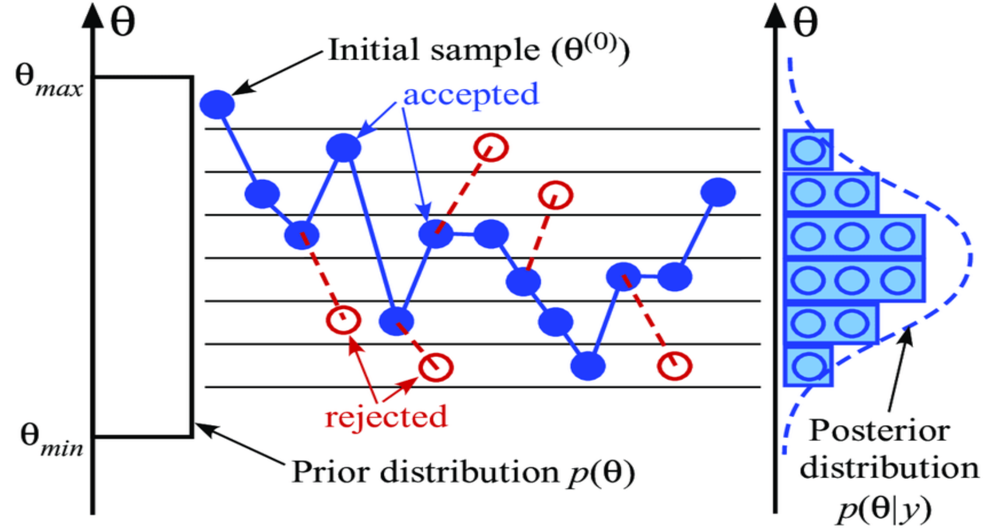
Bayesian inference → **MCMC algorithms**



Posterior distribution

samples

Metropolis algorithm:



Initialize parameters

for i=1 to i=N:

multiplicity =1

Proposing new parameters $\theta^* \sim q(\theta^*|\theta)$

Compute acceptance probability

$$\alpha = \min(p(\theta^*|D)/p(\theta|D), 1)$$

Sample from uniform distribution $u \sim \mathbf{U}(0,1)$

If $u < \min(1, \alpha)$ then $\theta_{i+1} = \theta^*$

Else $\theta_{i+1} = \theta$ (multiplicity +=1)

- **Multiplicity**: the number of consecutive rejections of proposed points before an acceptance occurs.
- Each point in the chain represents a vector of the posterior parameter values.

nPDF fit setup

Fit properties:

- fit **NLO** QCD predictions
- Kinematic cuts: Q > 2GeV, W > 3.5GeV, p_T > 3.0 GeV
- NC & CC DIS, W/Z boson and Heavy Quark
- 10 free parameters: 2 gluon, 6 valence, 2 sea
- Parameterization:

- **Pb PDF fit**
- Multiple nuclei PDF fit

CJ15

Functional form for bound protons at Q_0: $xf_i^{p/A}(x, Q_0) = c_0 x^{c_1}(1-x)^{c_2}(1 + c_3\sqrt{x} + c_4 x)$

Atomic number dependence: $c_k \rightarrow p_k + a_k \ln(A) + b_k \ln^2(A).$

$$\boxed{f_i^{(A,Z)} = \frac{Z}{A} f_i^{p/A} + \frac{A-Z}{A} f_i^{n/A}}$$

Accardi et al., arXiv:1602.03154

nPDF fit setup

$$xf_i^{p/A}(x,Q_0) \;=\; c_0 x^{c_1}(1-x)^{c_2}(1+c_3\sqrt{x}+c_4 x)$$

$$c_k \;\to\; p_k + a_k \ln(A) + b_k \ln^2(A).$$

$$xu_v \;\to\; a_1,\, a_2,\, a_3$$

$$xd_v \;\to\; a_1,\, a_2,\, a_3$$

$$x(\bar{d}+\bar{u}) \;\to\; a_1,\, a_2$$

$$xg \;\to\; a_1,\, a_2$$

CJ15

Functional form for bound protons at Q_0: $xf_i^{p/A}(x,Q_0) \;=\; c_0 x^{c_1}(1-x)^{c_2}(1+c_3\sqrt{x}+c_4 x)$

Atomic number dependence: $c_k \;\to\; p_k + a_k \ln(A) + b_k \ln^2(A).$

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

MCMC setup:

Adaptive MH algorithm setup:

- The algorithm starts with a normal random-walk MH phase until N_0 samples have been generated

  Proposal distribution: Multivariate Gaussian with fixed covariance C_0 $\mathbf{X}_{i+1} = \mathcal{N}(\mathbf{X}_i, C_0)$

- Then it switches to a self-learning proposal distribution

  Adaptive proposal distribution: Multivariate Gaussian with self learned covariance C_i (covariance from collected samples so far)

$$\mathbf{X}_{i+1} = (1-\beta)\mathcal{N}(\mathbf{X}_i, \frac{(2.4)^2}{d}.C_i) + \beta\mathcal{N}(\mathbf{X}_i, C_0)$$

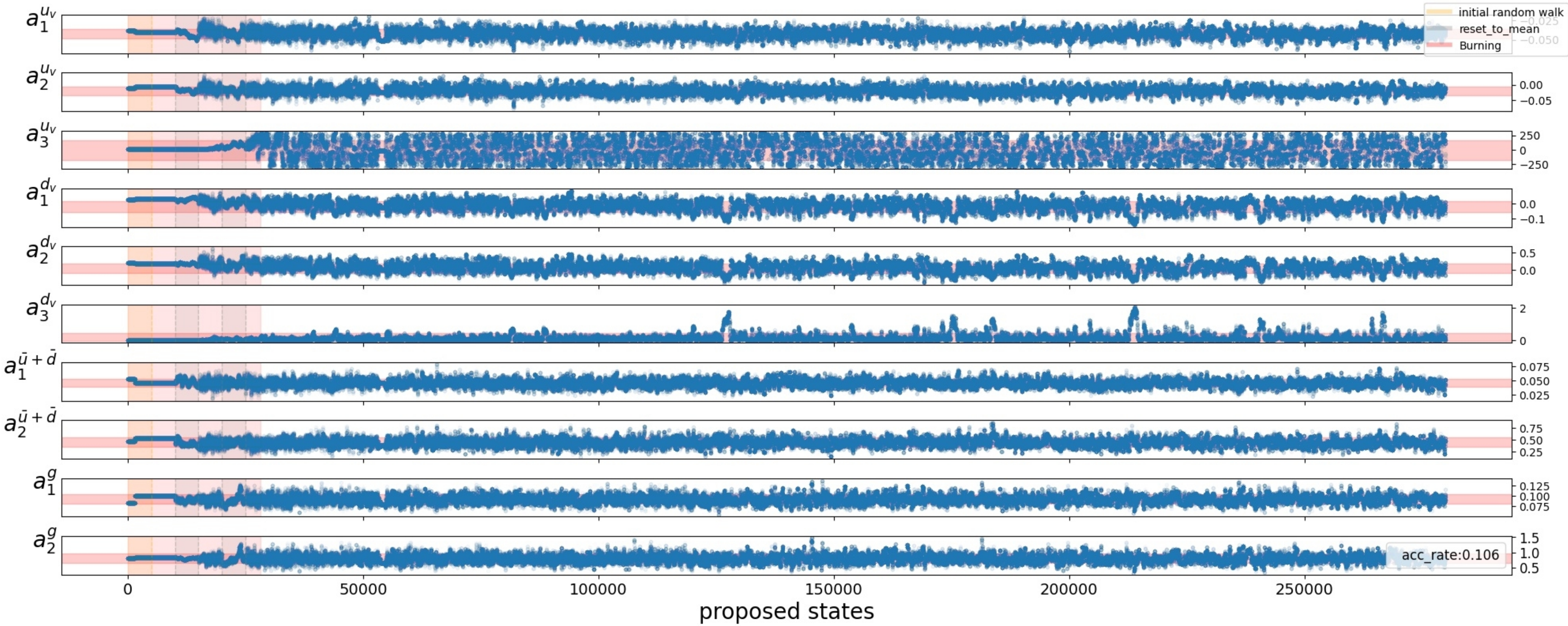
- To boost the convergence, the algorithm restarts from its current mean value*

---

*The fixed covariance matrix is first given by a fraction of initial parameter values and then after restarting, it adjusts to the fraction of diagonal elements in the current self-learned covariance C_i

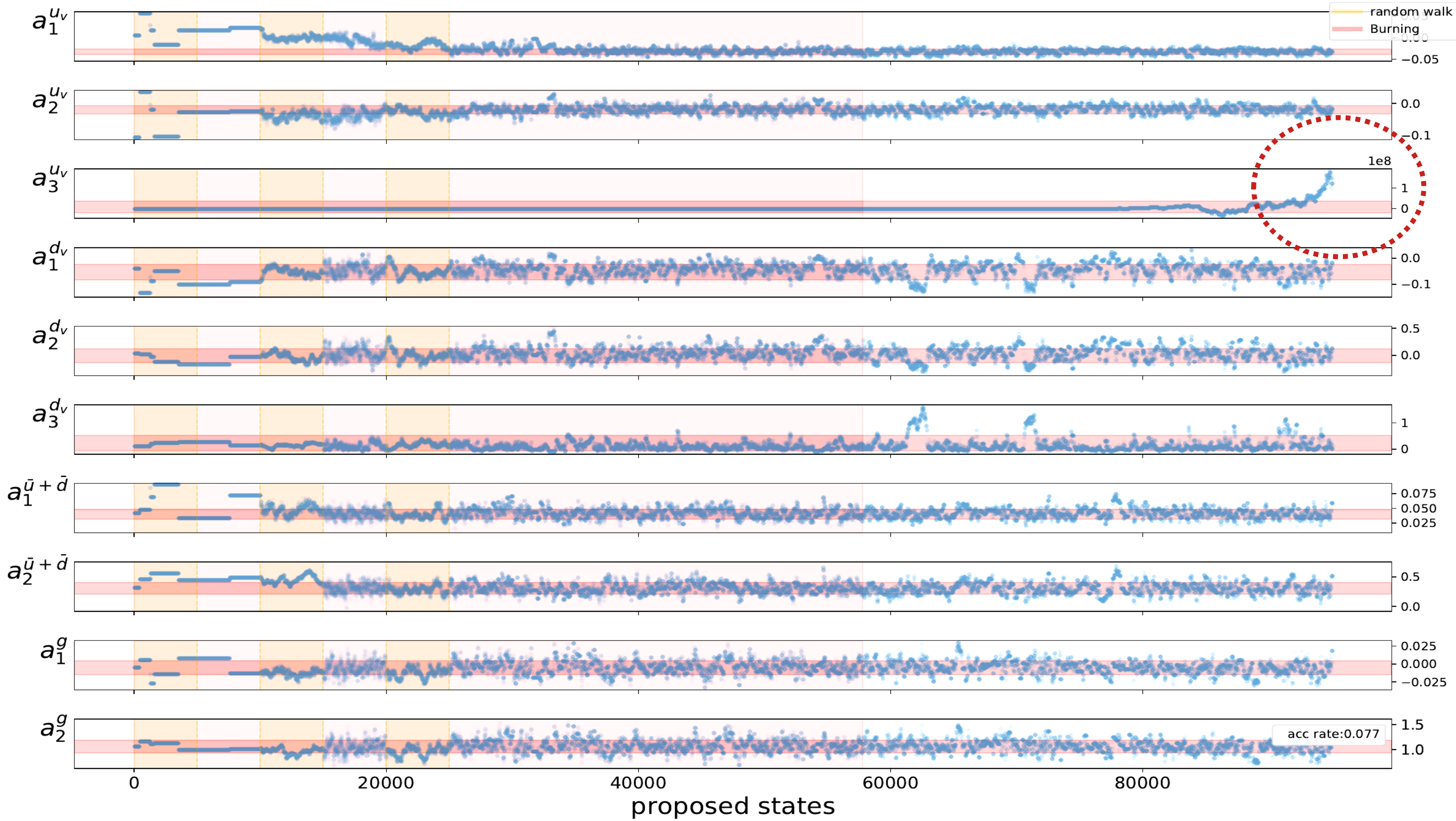
Preliminary results:

Markov chain generated for Pb PDF parameters (W/Z and Heavy Quark and ν-DIS(chorus); 1448 data)



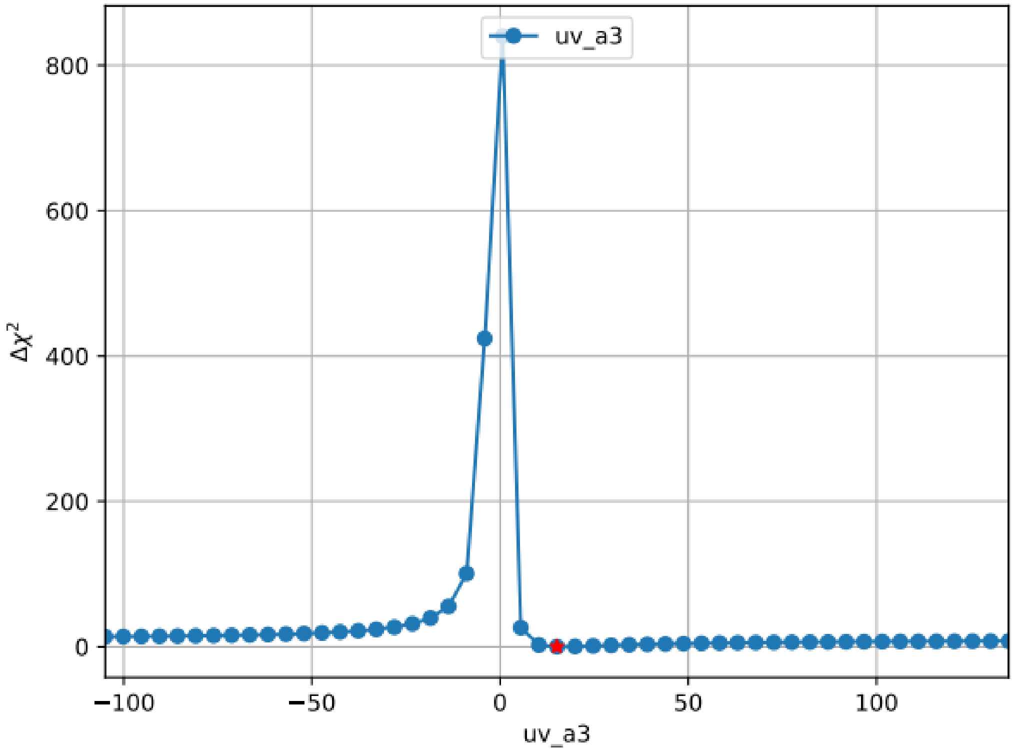
Generating this chain took about 20 days on 1 cpu

Markov chains without any prior



Prior setup:

Prior → we just use a uniform prior for the parameter: ${a_3}^{u_v} : U(-300, 300)$

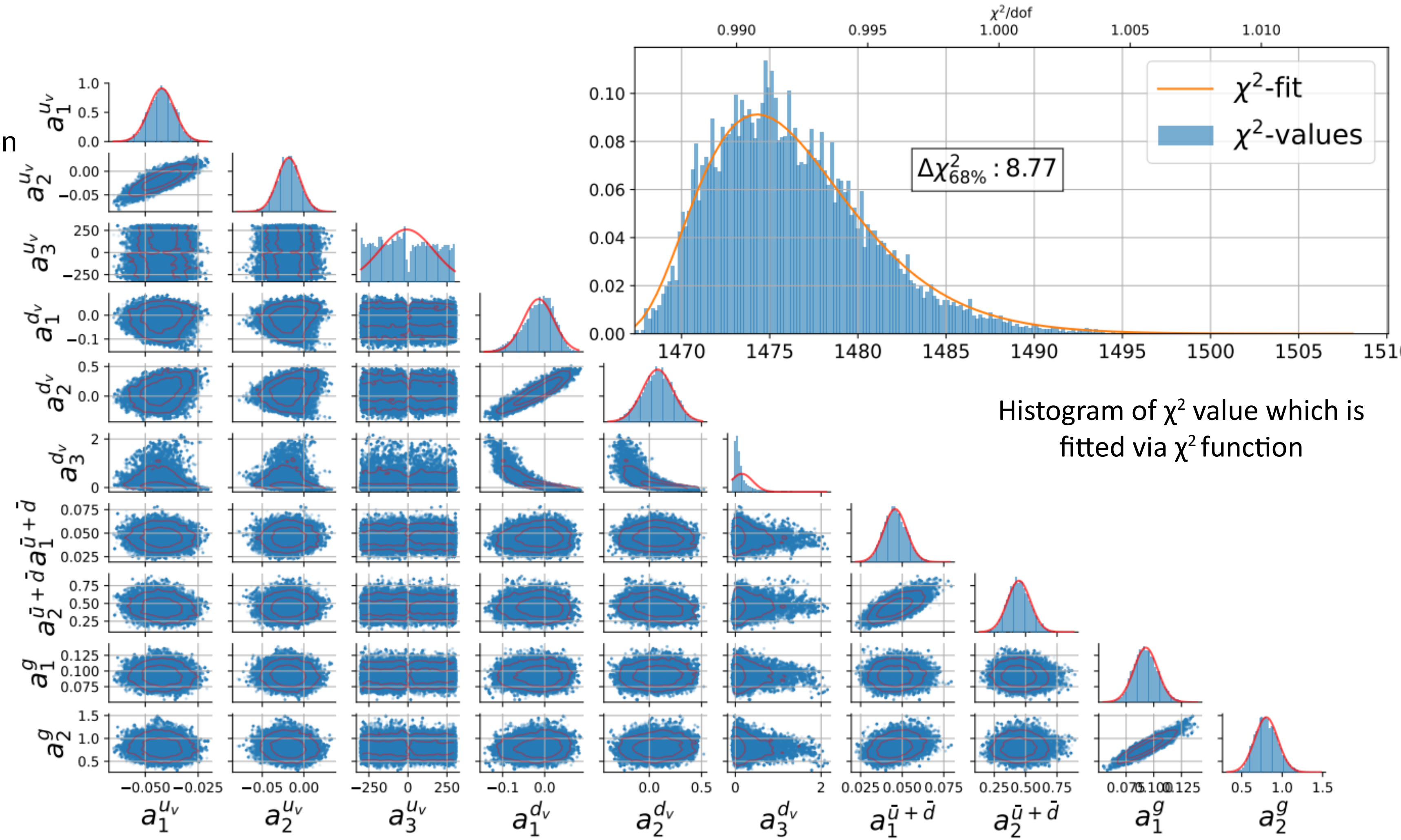


Scan of the χ^2 function along the nPDF parameters
(varying always one free parameter at a time while other parameters were left fixed at the global minimum)

Pairwise plot

MCMC can reveal non-Gaussian features of the underlying distribution

diagonal: histogram of each parameter
off-diagonal: 2D correlation plots between parameters



Histogram of χ^2 value which is fitted via χ^2 function

Error estimation:

Autocorrelation function (ACF) $\rho(k) = \frac{\mathrm{Cov}(k)}{\mathrm{Cov}(0)}$ $\mathrm{Cov}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(x_{t+k} - \bar{x})(x_t - \bar{x}),$

Integrated autocorrelation time $\tau_{int} = \frac{1}{2}\sum_{-\infty}^{+\infty}\rho(k)$ $\rightarrow$

Gamma-method

Estimating by analyzing the sum of autocorrelation up to a certain lag W_{opt}

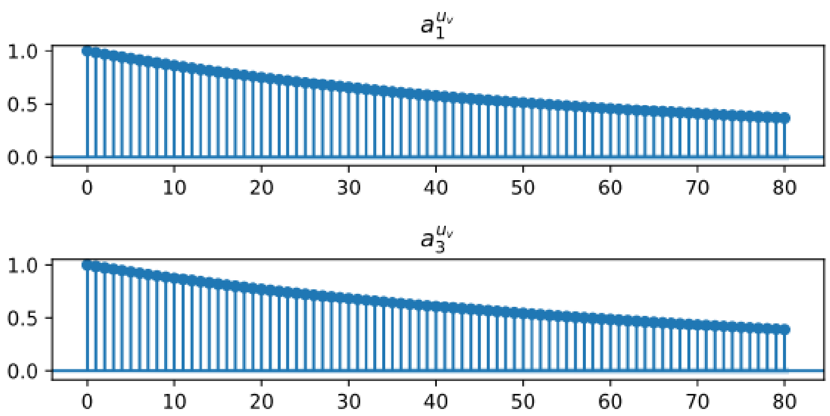
Monte Carlo error estimation (uncorrelated)

$$\sigma^2_{MC} = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \hat{\mu})^2$$

MCMC error estimation (correlated)

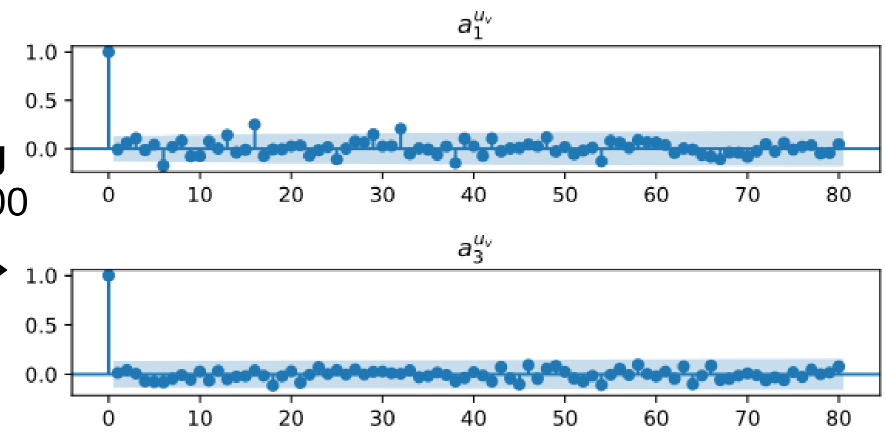
$$\sigma^2_{MCMC} = 2\,\tau_{int}\,\sigma^2_{MC}$$

Thinning method: keep only every k-th sample in the Markov chain and discard the rest



Autocorrelation function versus time interval

Thinning
by rate 100



Why Thinning?

- It provides an **uncorrelated** chain so we can use Monte-Carlo error estimation:

$$\sigma^2_{MCMC} = 2\,\tau_{int}\,\sigma^2_{MC} \quad \longrightarrow \quad \sigma^2_{MC} = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \hat{\mu})^2$$

- We aim to generate a set of PDF grids corresponding chain's units. Thinning the chain makes it more applicable.

Methodology:

- **Generating Multiple Chains**
  Each chain starts with random values from the Hessian fit results. Use different random seeds

- **Removing Burn-In Phase**
  Discard the initial segment of each chain, known as the burn-in or thermalization phase, which represents the period before the chain converges to the target distribution

- **Thinning Each Chain**
  Apply thinning to each chain to reduce the autocorrelation, aiming to retain only uncorrelated samples

- **Combining Uncorrelated Samples**
  Merge all the thinned, uncorrelated samples from the different chains into a single chain

- **Estimating Parameters and Uncertainties**
  Use the combined set of uncorrelated samples to estimate the values of nPDF parameters and their uncertainties.

- **generating an LHAPDF set**
  Construct nPDF corresponding to each unit of the combined chain and perform error estimation in the level of nPDF (Saving them in the standard LHAPDF format so that anyone can use such nPDFs)

Percentile method [nNNPDF]

After generating nPDF corresponding to each point of thinned Markov Chain, then we perform "percentile method" to estimate the nPDF uncertainty:

The percentile method involves:

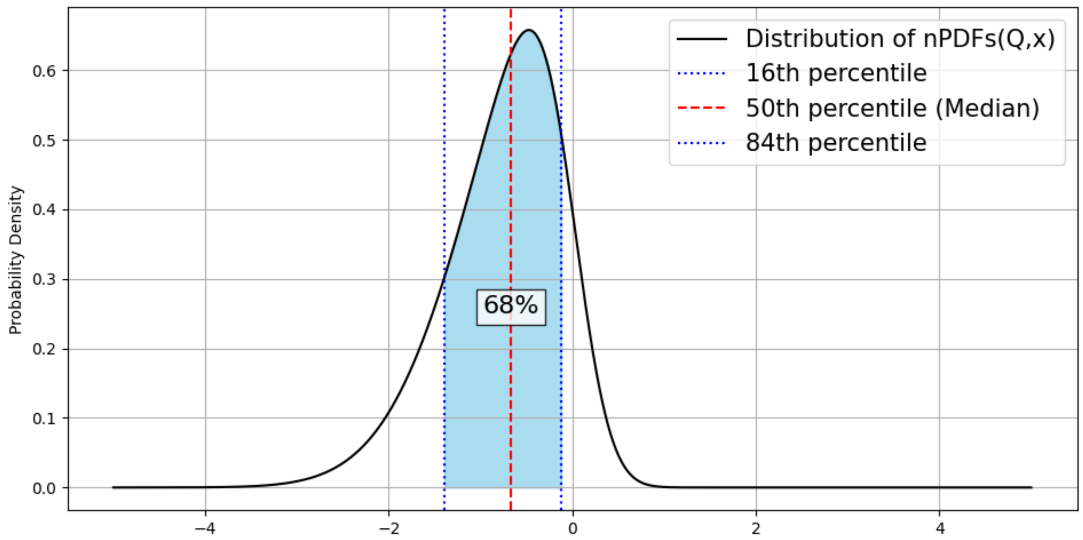
- generating a distribution of a statistic (in our case: distribution of nPDFs)
- then determining confidence intervals by identifying the appropriate percentiles of this distribution.

68% CI :
$\alpha = 1\text{-}0.68 = 0.32$

Lower Confidence Limit: $P_{\alpha/2} = \mathbf{P_{16th}}$

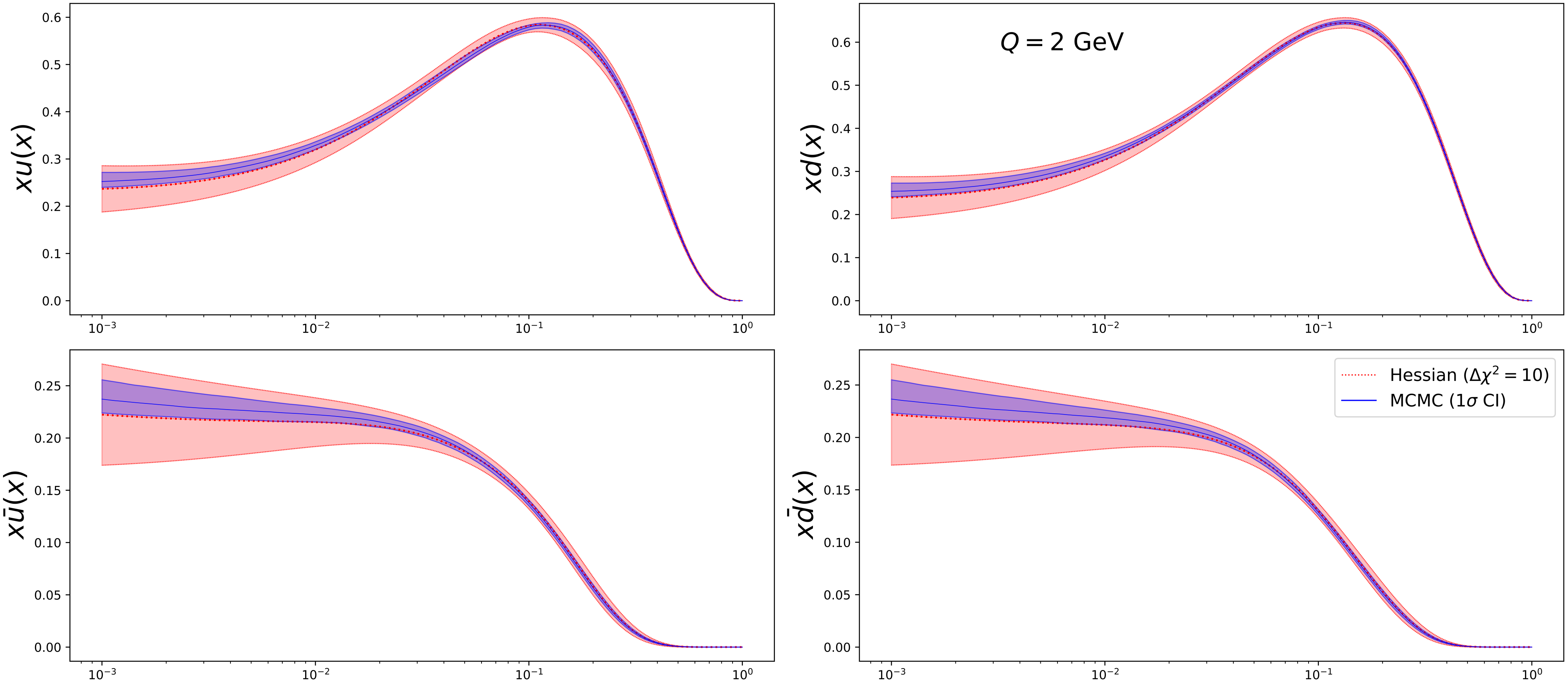
Upper Confidence Limit: $P_{(1-\alpha/2)} = \mathbf{P_{84th}}$

Central value: $\mathbf{P_{50th}}$



LHAPDF grids:

Pb^{208} PDF resulting from **MCMC**(percentile method to for uncertainty estimation) and **Hessian** methods



Conclusion:

- Despite the MCMC challenges (mainly computational cost), this method has become a powerful tool for determining nPDFs and so far we have obtained promising results (comparing with Hessian) for Pb PDF fit

- We would like to extend this approach for multiple nuclei PDF fits and investigate additional statistical methods for estimating Markov Chains uncertainty.

Acknowledgment:
This work was supported by Narodowe Centrum Nauki under grant no.\ 2019/34/E/ST2/00186.

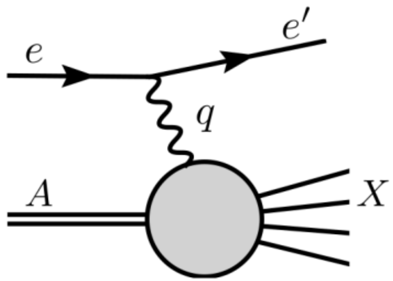
Backup

DIS variables for **nucleus**

$$q \equiv k' - k\,,\ Q^2 \equiv -q^2 \qquad x_A \equiv \frac{Q^2}{2p_A.q}$$

p_A : nucleus momentum

$x_A \in (0,1)$: fraction of the nucleus momentum carried by a nucleon



$$e(k) + A(p_A) \to e'(k') + X$$

DIS variables for **parton**

$x_N = Ax_A$: parton momentum fraction with respect to the average nucleon momentum p_N

$$p_N = \frac{p_A}{A}$$

$x_N \in (0, A)$

Sum rules:

$$\int_0^1 \mathrm{d}x_A \, \tilde{u}_v^A(x_A, Q^2) = 2Z + N \,,$$
$$\int_0^1 \mathrm{d}x_A \, \tilde{d}_v^A(x_A, Q^2) = Z + 2N \,,$$

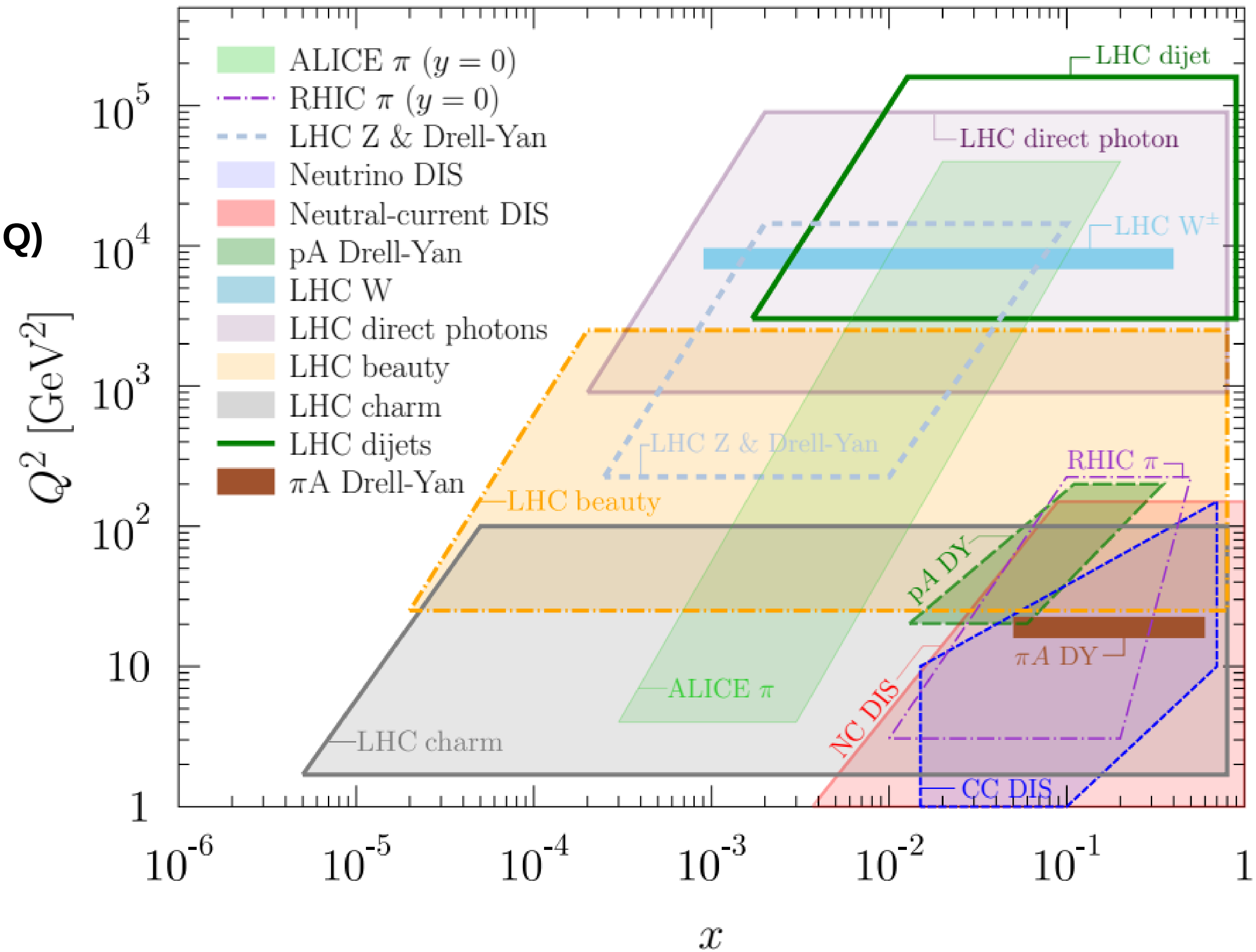
and the momentum sum rule

$$\int_0^1 \mathrm{d}x_A \, x_A \left[\tilde{\Sigma}^A(x_A, Q^2) + \tilde{g}^A(x_A, Q^2)\right] = 1 \,,$$

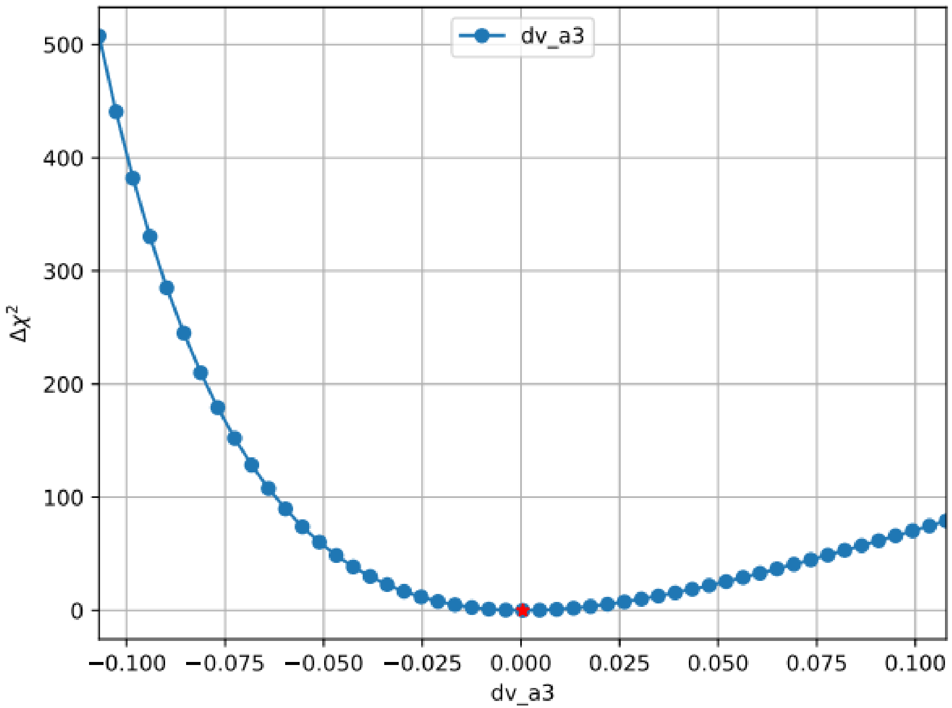
where $N = A - Z$ and $\tilde{\Sigma}^A(x_A) = \sum_i (\tilde{q}_i^A(x_A) + \tilde{\bar{q}}_i^A(x_A))$

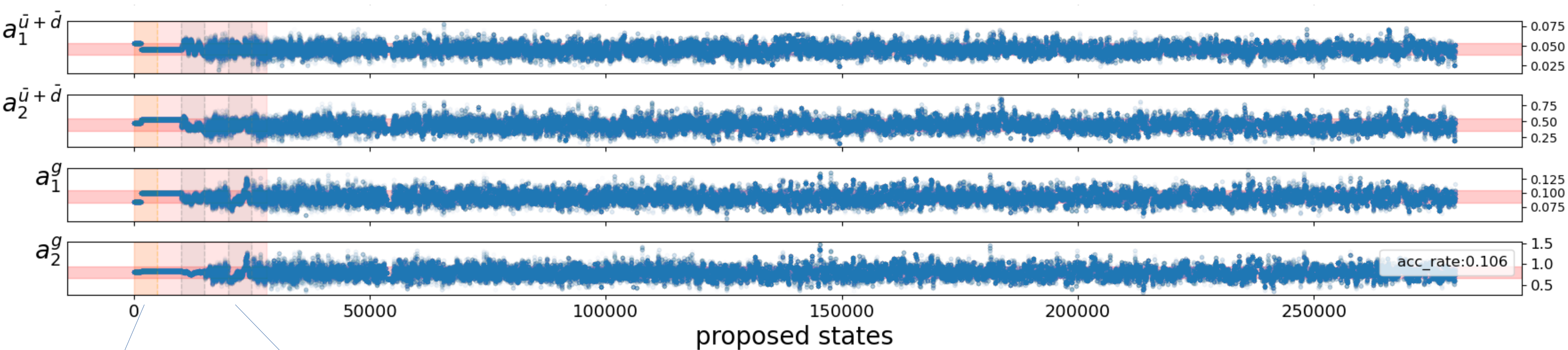
Experimental data:

- NC & CC DIS
- LHC W/Z production
- Heavy Quark production (HQ)



Scan of the χ^2 function along dv-a3 parameter



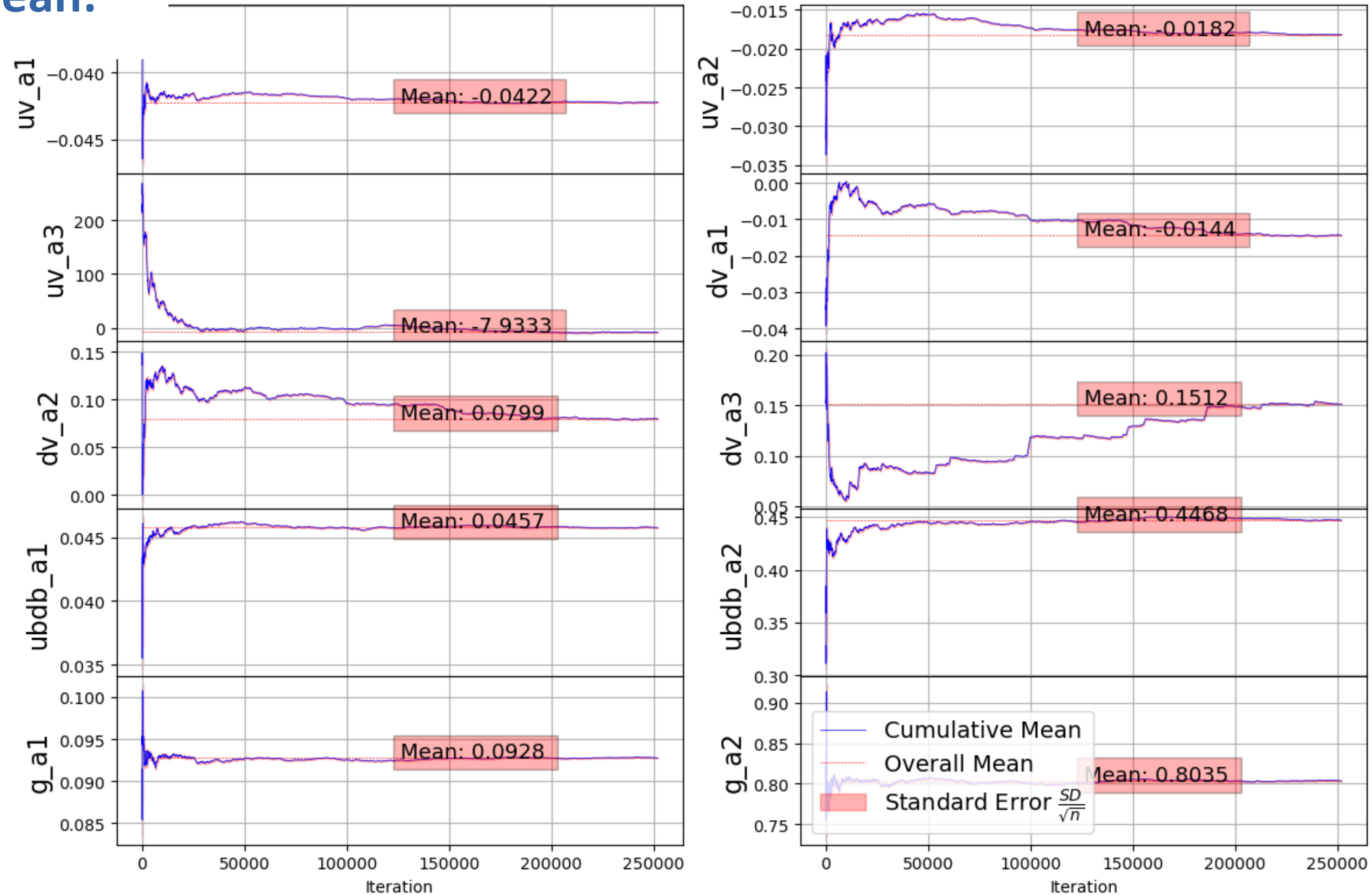


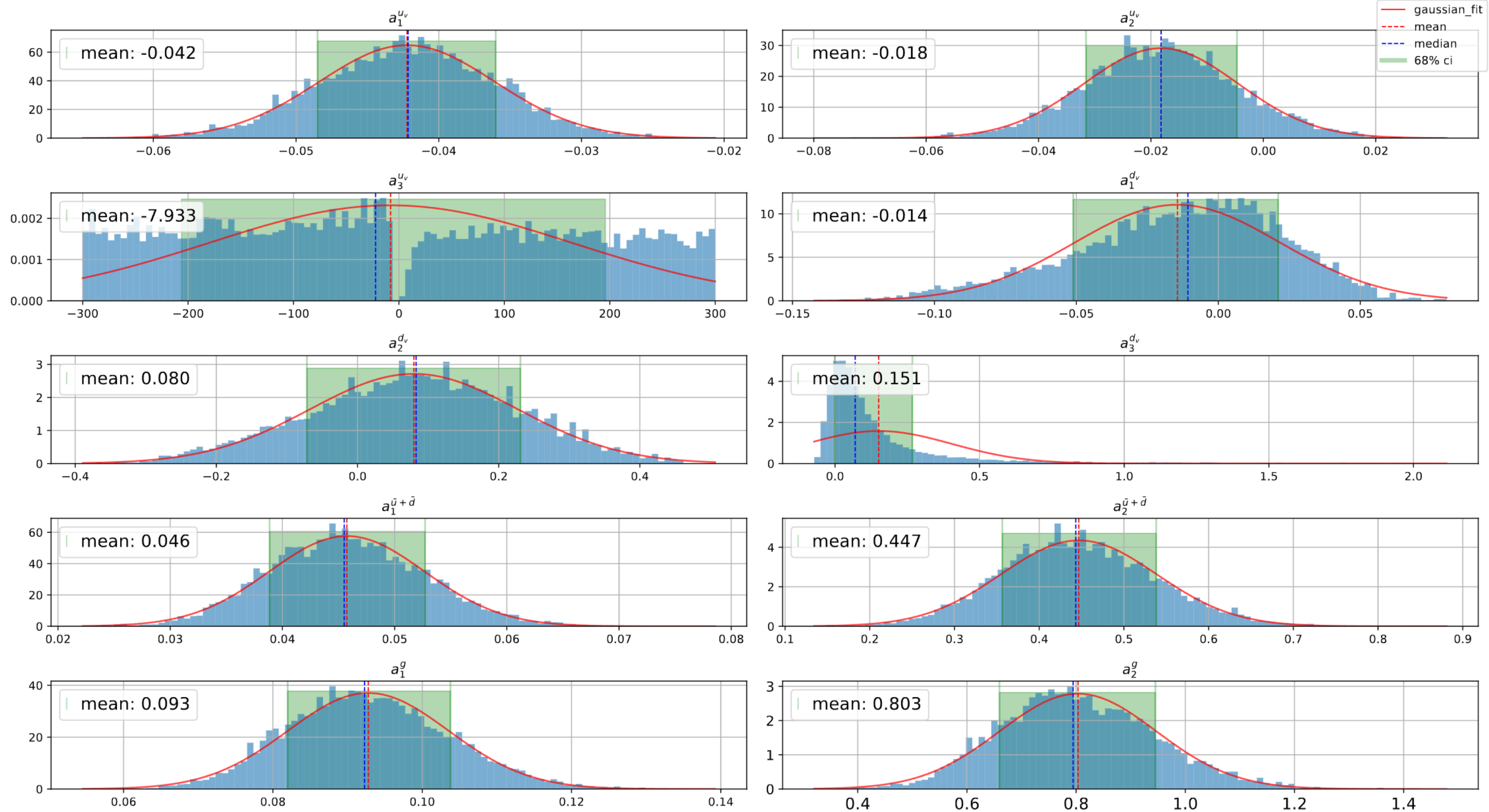
N_0 = 5000

Restarting the chain at
10,000 and 20,000

Starting point: global minimum from Hessian fit + Gaussian noise (width= 20 % of minimum value)
Thermalization (burn-in phase): removing first 8000 accepted points

Cumulative Mean:





gaussian_fit
mean
median
68% ci
$a_1^{u_v}$
mean: -0.042
$a_2^{u_v}$
mean: -0.018
$a_3^{u_v}$
mean: -7.933
$a_1^{d_v}$
mean: -0.014
$a_2^{d_v}$
mean: 0.080
$a_3^{d_v}$
mean: 0.151
$a_1^{\bar{u}+\bar{d}}$
mean: 0.046
$a_2^{\bar{u}+\bar{d}}$
mean: 0.447
a_1^{g}
mean: 0.093
a_2^{g}
mean: 0.803

MH vs adaptive MH