PROCEEDINGS OF SCIENCE

IFJPAN-IV-2024-11, MS-TP-24-18

# Estimating nPDF Uncertainties via Markov Chain Monte Carlo Methods

**N. Derakhshanian**,[a,*] **P. Risse**,[b] **T. Ježo**,[b] **K. Kovařík**[b] **and A. Kusina**[a]

[a] *Institute of Nuclear Physics Polish Academy of Sciences, ul. Radzikowskiego 152, 31-342 Kraków , Poland*

[b] *Institut für Theoretische Physik, Westfälische Wilhelms-Universität Münster, Wilhelm-Klemm-Straße 9, D-48149 Münster, Germany*

*E-mail:* nasim.derakhshanian@ifj.edu.pl

Nuclear Parton Distribution Functions (nPDFs) are critical for understanding nuclear structure and making heavy-ion collision predictions. nPDFs have been determined via 'global QCD analyses', which is a statistical approach based on fitting nPDF-dependent theoretical predictions to the relevant experimental data. One of the crucial aspects of nPDF determination is uncertainty estimation. Typically, the Hessian method is used to propagate experimental uncertainties into predictions for collisions of nuclei. However, due to the nature of nPDF fits (such as limited data constraints, non-gaussianity, and possible multiple minima), this method does not always provide reliable results. In this work, we introduce the application of Markov Chain Monte Carlo (MCMC) methods as a statistically sophisticated alternative for estimating nPDF uncertainties by sampling directly from the probability distribution of the nPDF parameters.

*31st International Workshop on Deep Inelastic Scattering*
*8–12 April 2024*
*Maison MINATEC, Grenoble, FRANCE*

*Speaker





PoS(DIS2024)058

## 1. Introduction

While the $Q$-dependence of parton distribution functions (PDFs) can be calculated using perturbative quantum chromodynamics, their dependence on the momentum fraction $x$ is non-perturbative and must be extracted using a 'global QCD analysis', where PDF-dependent predictions are fitted from experimental measurements. We similarly employ this approach to extract nuclear PDFs from experimental data. In both scenarios, global analyses rely on minimizing a statistical figure-of-merit to optimize the correspondence between theory and experiment. This figure-of-merit is typically the chi-squared function $\chi^2$, calculated for uncorrelated data points using the formula $\chi^2(\mathbf{a}) = \sum_i \frac{[D_i - T_i(\mathbf{a})]^2}{\sigma_i^2}$, where $D_i$ represents the observed data points, $T_i$ are the corresponding theoretical predictions, and $\sigma_i^2 \equiv \hat{\sigma}_i^2 + \bar{\sigma}_i^2$ is the sum of the statistical and systematic uncertainties, and $\mathbf{a}$ is the set of parameters that define the functional form of PDFs at the initial scale.
Several collaborations have performed global analyses of nuclear PDFs [1]. This proceedings paper presents the preliminary results of our study using the nCTEQ global analysis framework [2]. In this framework, assuming isospin symmetry, the PDF of the nucleus $f_i^{(A,Z)}$ is parameterized in terms of bound proton (neutron) PDF $f_i^{p(n)/A}$, as

$$f_i^{(A,Z)}(x,Q) = \frac{Z}{A} f_i^{p/A}(x,Q) + \frac{A-Z}{A} f_i^{n/A}(x,Q). \tag{1}$$

The CJ15 proton baseline [3] is the functional form that we use to parameterize the bound proton at the initial scale $Q_0 = 1.3$ GeV. This parameterization for $u_v, \bar{d}+\bar{u}, g, s+\bar{s}$ is as follows:

$$x f(x,Q_0) = c_0 x^{c_1} (1-x)^{c_2} \left(1 + c_3\sqrt{x} + c_4 x\right), \tag{2}$$

and $d_v$ is parameterized through

$$x d_v(x,Q_0) = c_0 \left[x^{c_1}(1-x)^{c_2}\left(1 + c_3\sqrt{x} + c_4 x\right) + c_5 x^{c_6} x u_v(x,Q_0)\right]. \tag{3}$$

In this framework, the nuclear $A$-dependence is represented by $c_j(A)$ coefficients, which are defined as $c_j(A) = p_j + a_j \ln A + b_j \ln^2 A$. In our study, we focus on fitting 10 parameters $a_j$, specifically $a_1$, $a_2$, and $a_3$ for the $u$-valence and $d$-valence, and $a_1$ and $a_2$ for both $\bar{d}+\bar{u}$ and gluons distributions. We keep $p_j$ parameters fixed based on the CJ15 proton PDF fit.
To estimate these nPDF parameters, we employ Bayesian inference, a statistical method that updates our beliefs or knowledge about the parameters based on observed data. Using Bayes theorem, the posterior distribution $P(\mathbf{a}|D)$, the probability of having a set of parameters $\mathbf{a}$ given the observed data $D$, is formulated in terms of the likelihood of the data given the parameters and the prior distribution of the parameters $P(\mathbf{a})$ as

$$P(\mathbf{a}|D) = \frac{1}{N} P(\mathbf{a}) \exp\left(-\frac{1}{2}\chi^2(D, T(\mathbf{a}))\right). \tag{4}$$

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from probability distributions in order to construct Markov chains in such a way that their stationary distribution converges to the target probability distribution. Although directly calculating the posterior is generally difficult and often impractical, especially in complex or high-dimensional problems, MCMC

enables the estimation of posterior probability densities for multi-dimensional models and provides reliable error estimatations. We obtain a distribution of all the probable values for each parameter by generating a Markov chain from the posterior distribution via a MCMC algorithm. It allows us to calculate statistical measures such as the mean and variance. However, in this context, we must carefully consider error estimation, as the samples can be highly correlated due to the Markovian property. The autocorrelation function measures how the correlation between any two samples changes with their time separation or lag. This correlation reduces the effective number of independent samples and complicates the error estimation. In the ideal scenario where samples are independent, the error estimation can be done straightforwardly by using the Monte Carlo standard error (MCSE), which is defined as the standard deviation of the chains divided by their effective sample size.

## 2. Motivation

Uncertainty estimation methods in nuclear PDFs are critical, as they play a significant role in understanding the precision and reliability of heavy-ion collision predictions. The Hessian method is most commonly used for estimating nPDF uncertainties [1]. It relies on $\chi^2$ function minimization to find the best fit and assumes a quadratic approximation of the $\chi^2$ distribution near this minimum to explore uncertainties. However, this method has certain shortcomings, such as difficulty with non-Gaussian error distributions, the potential for not identifying the global minimum, and the impact of the choice of the $\chi^2$ tolerance on the reliability of the uncertainty estimations. Particularly, the lack of sufficient data to constrain all nPDF flavors makes these issues more serious.
To address these limitations, we apply MCMC methods as a statistically sophisticated alternative for estimating nPDF uncertainties by sampling directly from the probability distribution of the parameters. MCMC has already been used in a simplified proton PDF fit [5], and we are also working towards a full scale proton fit using MCMC [6]. In our current study, we extend the application of MCMC to include a more realistic dataset, resulting in a comprehensive uncertainty estimation for nuclear PDFs and comparing these results with the standard Hessian approach.

## 3. Methodology

We aim to find a set of nPDF parameters that maximizes the posterior probability distribution given the experimental data. First, we need to construct the posterior probability distribution according to Eq. (4). The posterior is defined in terms of the $\chi^2$ likelihood and prior, where the $\chi^2$ function accounts for the correlated and normalization uncertainties of the data sets and includes the theoretical predictions calculated in dependence of the PDF parameters. Additionally, we need to set the prior distribution. In our setup, we don't use any priors (i.e., we don't consider any initial information about the parameters and formally use a constant prior), except for the prior for parameter $a_3$ of the up-valence distribution. If we run a chain without this prior, the parameter remains unconstrained due to the lack of constraining data and the specific form of the parametrization. To address this issue, we apply a uniform prior with the bounds $[-300, 300]$ specifically for the parameter $a_3$. The prior is constant within this interval and zero outside of it. To ensure that the $\chi^2$ function is independent of the prior choice, we perform a $\chi^2$-scan for this

parameter and observe that the $\chi^2$ value remains constant at the prior boundaries. The next step is running a MCMC algorithm to generate Markov chains and draw samples from the posterior distribution. For this purpose, We employ an adaptive Metropolis-Hastings (aMH) algorithm [7]. This algorithm starts with a random-walk phase where the proposal distribution is a multivariate Gaussian function with a fixed covariance matrix. After this initial phase, it switches to a self-learning phase where the covariance matrix of the Gaussian function is no longer fixed but is dynamically updated based on the collected samples.
We can summarize our approach in three steps:

- **Generation of MCMC Chains**: We generate multiple MCMC chains, each initialized with random values from Hessian fit results and using unique random seeds to ensure independence. We remove the initial segment known as the burn-in or thermalization phase of each chain to guarantee that subsequent samples reflect the equilibrium state of the chain.

- **Sample Optimization**: To address the issue of autocorrelation within each chain and improve error estimation, we apply a thinning method. In this procedure, we selectively keep only every $\eta$-*th* sample of the chain and discard the rest. This process significantly reduces the number of correlated samples and allows to obtained samples that are approximately uncorrelated (assuming the initial chains are long enough). Moreover, it is critical for creating a LHAPDF set of PDF grids [8], a standard PDF distribution format, since the number of chain units must be limited to make it practical and user-friendly.

- **Analysis and Output**: The uncorrelated samples from all chains are combined to form a comprehensive dataset, from which we estimate the nPDF parameters and their uncertainties. Given that these samples are effectively uncorrelated, we are able to employ standard Monte Carlo error estimation techniques, or alternatively, use more advanced methods. We then create a nPDF set in the standard LHAPDF format corresponding to each individual sample, facilitating further use and analysis by the research community.

## 4. Results

As a preliminary study, we generate Markov chains for 10 lead ($Pb^{208}$) PDF parameters (3 for $u$-valence, 3 for $d$-valence, 2 for light sea quarks, and 2 for gluons), as displayed in Fig. 1. Since MCMC method is inherently time-consuming, we have to optimize its computational efficiency. To speed up theory predictions and enhance efficiency, we restrict our datasets to $W/Z$ boson production in proton-lead collisions at the LHC, charged current (CC) neutrino DIS on lead from the CHORUS experiment, and heavy quark production in proton-lead collisions at LHC. After performing kinematic cuts (for CC DIS: $W^2 > 12.25\text{GeV}^2$ and $Q^2 > 4.0\text{GeV}^2$ ), we have 1448 data points to perform the fit for lead PDF extraction. All theoretical predictions are performed at next-to-leading order (NLO) in perturbative QCD and we consider sACOT scheme for handling heavy quarks [10]. In terms of our MCMC setup, we use the adaptive Metropolis-Hastings algorithm to collect approximately 800,000 points, after discarding the initial thermalization phase. Regarding computational cost, it takes one day for a single CPU to generate 20,000 points. Then, to reduce the autocorrelation of the chain, we apply a thinning process, selecting every 4000th sample ($\eta$=4000).

After ensuring that the thinned samples are uncorrelated, we construct the nuclear PDF for each sample, enabling us to estimate uncertainties directly at the nPDF level. This process provides a comprehensive evaluation of the confidence intervals for the distribution of nPDFs derived from the Markov chain. We use the percentile method to estimate the uncertainties to obtain $1\sigma$ confidence interval (CI), a technique also employed in the nNNPDF3.0 fit [9]. In this method, after arranging the nPDFs in ascending order, we determine the confidence interval directly by selecting the percentiles corresponding to the desired confidence level; in our case, the lower and upper bounds are 16th and 84th percentiles, and the central value is is identified as the 50th percentile.

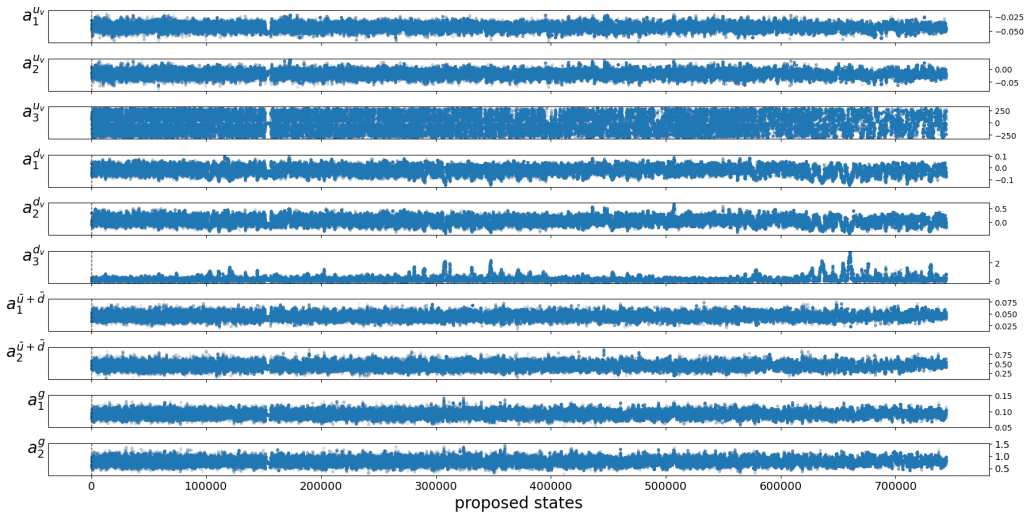


**Figure 1:** The time series of parameter values for the ($Pb^{208}$) PDF after removing the thermalization phase.

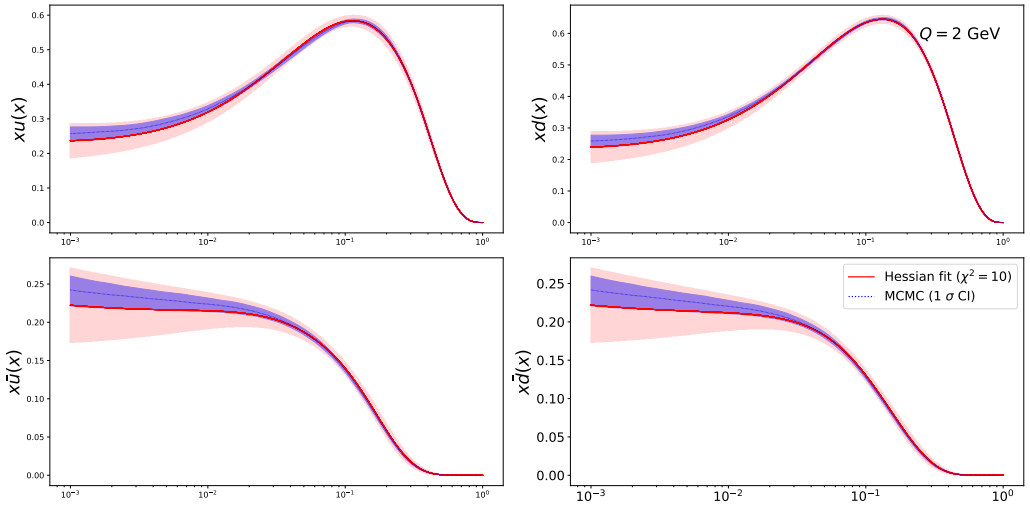


**Figure 2:** $u$, $\bar{u}$, $d$, and $\bar{d}$ PDFs in Lead resulting from the Hessian (red) and MCMC (blue) methods.

## 5. Conclusion

In conclusion, despite the MCMC methods challenges, particularly the computational demands, this approach could provide a robust and statistically rigorous framework for extracting nuclear PDFs, especially effective in addressing non-Gaussian features. Our preliminary results for the lead PDF are promising. We are currently generating additional Markov chains to collect larger statistics and improve the accuracy of our results. Furthermore, we aim to extend this approach for fits with multiple nuclei and explore additional statistical methods to refine uncertainty estimation in Markov chains.

## Acknowledgments

This work was supported by Narodowe Centrum Nauki under grant no. 2019/34/E/ST2/00186. We also acknowledge the computational resources provided by HPC cluster PALMA II of the University of Münster, subsidised by the DFG (INST 211/667-1). P.R., T.J. and K.K. acknowledge support of the DFG through the Research Training Group GRK 2149.

## References

[1] M. Klasen and H. Paukkunen, "Nuclear PDFs After the First Decade of LHC Data," [arXiv:2311.00450 [hep-ph]].

[2] K. Kovarik, *et al.*, "nCTEQ15 - Global analysis of nuclear parton distributions with uncertainties in the CTEQ framework," Phys. Rev. D 93, no. 8, (2016) 085037.

[3] A. Accardi, *et al.*, Phys. Rev. D 93, no. 11, (2016) 114017.

[4] N.T. Hunt-Smith, *et al.*, "Accelerating Markov Chain Monte Carlo sampling with diffusion models," Comput. Phys. Commun. 296 (2024) 109059.

[5] Yémalin Gabin Gbedo, Mariane Mangin-Brinet, "Markov Chain Monte Carlo techniques applied to Parton Distribution Functions determination: proof of concept," Phys. Rev. D 96, (2017) 014015.

[6] Risse, Peter and Derakhshanian, Nasim and Ježo, Tomas and Kovařík, Karol and Kusina, Aleksander, "A Markov Chain Monte Carlo determination of Proton PDF uncertainties at NNLO," arXiv:2407.12377.

[7] Heikki Haario, Eero Saksman, and Johanna Tamminen, "An Adaptive Metropolis Algorithm," Bernoulli 7.2 (2001): 223-242.

[8] A. Buckley, *et al.*, "LHAPDF6: parton density access in the LHC precision era," Eur. Phys. J. C 75 (2015) 132, arXiv:1412.7420 [hep-ph].

[9] R. Abdul Khalek, *et al.*, "nNNPDF3.0: evidence for a modified partonic structure in heavy nuclei," Eur. Phys. J. C **82**, no.6, (2022) 507, [arXiv:2201.12363 [hep-ph]].

[10] M. A. G. Aivazis, J. C. Collins, F. I. Olness and W. K. Tung, Phys. Rev. D **50**, 3102-3118 (1994), [arXiv:hep-ph/9312319 [hep-ph]].