NUCLEAR PDF DETERMINATION USING MARKOV CHAIN MONTE CARLO METHODS*

N. Derakhshanian[a], P. Risse[b], T. Ježo[b], M. Klasen[b]
K. Kovařík[b], A. Kusina[a], F.I. Olness[c], I. Schienbein[d]

[a]Institute of Nuclear Physics Polish Academy of Sciences, 31342 Kraków, Poland
[b]Institut für Theoretische Physik, Westfälische Wilhelms-Universität Münster
Wilhelm-Klemm-Straße 9, 48149 Münster, Germany
[c]Southern Methodist University, Dallas, TX 75275, USA
[d]Laboratoire de Physique Subatomique et de Cosmologie
53 Rue des Martyrs Grenoble, France

Received 28 April 2023, accepted 1 May 2023,
published online 6 September 2023

Nuclear parton distribution functions (nPDFs) are crucial in studying nuclear structure and high-energy nuclear collisions. nPDFs have been determined via 'global QCD analyses', in which the nPDF-dependent predictions for a given process are compared with their actual measurements. One of the challenging parts of nPDF extractions is the estimation of uncertainties. The most common approach for this purpose is the Hessian method, which, however, has certain shortcomings, especially in the case of weaker data constraints. Here, we will show a case study for an alternative approach where nPDF uncertainties are estimated using Markov Chain Monte Carlo (MCMC) methods.

DOI:10.5506/APhysPolBSupp.16.7-A33

1. Introduction

Parton distribution functions (PDFs) are an essential part of predictions of hadronic observables. They characterize the quark and gluon content of nucleons. So far, we cannot reliably compute PDFs from first principles, and we need to determine them by comparing (fitting) PDF-dependent predictions with the corresponding experimental measurements in a process called 'global QCD analysis' [1, 2]. Such an approach is possible due to the factorization property of quantum chromodynamics (QCD), which provides us with a framework to calculate such PDF-dependent predictions. When

* Presented by N. Derakhshanian at the 29th Cracow Epiphany Conference on *Physics at the Electron–Ion Collider and Future Facilities*, Cracow, Poland, 16–19 January, 2023.

describing the structure of a nucleus, analogous reasoning can be used. However, the PDFs of a nucleus are modified compared to a simple combination of free nucleon PDFs. Such modified quantities, known as the nuclear PDFs (nPDFs), also need to be determined in global QCD fits. Several collaborations provide nPDFs [3–5] obtained using different fitting frameworks. In this contribution, we present the preliminary results of a study using the nCTEQ global analysis framework. One of the key elements of this framework is the parameterization of the PDF of the nucleus $f_i^{(A,Z)}$ [6]

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

where A is the atomic number, Z is the number of protons in a given nucleus, and $f_i^{p(n)/A}$ is the effective bound proton (neutron) PDF, which at the initial scale Q_0 is parametrized in the following way:

$$x f_i^{p/A}(x,Q_0) = c_0 x^{c_1} (1-x)^{c_2} e^{c_3 x} \left(1 + e^{c_4} x\right)^{c_5}\,. \tag{2}$$

The PDF of a parton i in a nucleus A depends on a fraction x of the average nucleon momentum and Q the factorization scale representing the hard scale of the collision. Additionally, the dependence on the nucleus mass, A, is introduced in the c_k coefficients as

$$c_k \to c_k(A) \equiv p_k + a_k \left(1 - A^{-b_k}\right)\,, \qquad k = \{1,\ldots,5\}\,. \tag{3}$$

The Markov Chain Monte Carlo (MCMC) method is a powerful tool to sample complex probability distributions [7]. It is based on building a Markov chain in which each state is made by drawing a sample from a proposal distribution and accepting or rejecting the sample based on a certain criterion — the most commonly used one is based on the Metropolis–Hastings (MH) algorithm [8]. The algorithm proceeds by iteratively generating a sequence of samples that converge to the target distribution, allowing for efficient exploration of the parameter space. In the context of standard Monte Carlo sampling, the error estimation can be calculated by

$$\sigma_{\mathrm{MC}}^2 = \frac{1}{N-1} \sum_{t=1}^{N} \left[\mathcal{O}\left(\{c_k\}^t\right) - \mu\left(\mathcal{O}\right)\right]^2\,, \tag{4}$$

where σ_{MC} is the standard deviation, μ is the corresponding mean value, $\mathcal{O}$ represents any observable/function depending on the random sample $\{c_k\}$, and N is the number of units in the sample. However, in the MCMC method, due to the Markovian property, the units of a sample are correlated, which

can be measured by the autocorrelation function (ACF)[1]. In this case, we need to take into account the autocorrelations via the autocorrelation time, τ_{int}, which can be estimated as the sum of ACF over all lags [9][2]

$$\sigma^2_{\text{MCMC}} = 2\tau_{\text{int}}\sigma^2_{\text{MC}} + \mathcal{O}\left(\frac{\tau_{\text{int}}}{N}\right) . \tag{5}$$

2. Methodology

The goal of this work is to find the set of nPDF parameters that maximizes the posterior probability distribution given the experimental data. For this purpose, we first construct a likelihood function regarding the nCTEQ parametrization that incorporates experimental uncertainties. Using Bayes' theorem, the posterior function is defined in terms of likelihood and prior. The MCMC algorithm is then used to generate a Markov chain of the nPDF parameters that samples from the posterior probability distribution. The nPDF parameters are varied during the MCMC iterations to improve the agreement between the model predictions and the experimental data. The algorithm explores the parameter space by generating new sets of parameters based on the previous ones, with a probability that depends on the posterior. The uncertainty estimation is done through the analysis of the Markov chain generated by the algorithm. It can be challenging, particularly when dealing with high-dimensional parameter spaces or complex models. To ensure accurate uncertainty estimates, it is important to carefully evaluate the Markov chain and choose appropriate convergence diagnostics and statistical measures such that one is confident the chain has converged to the posterior distribution.

Aside from the high computational cost, one of the challenges with using MCMC to find PDFs is dealing with autocorrelations. Since the samples generated in chains can be highly correlated, they do not effectively explore the entire parameter space. This can lead to slow convergence and inefficient sampling, making it difficult to obtain accurate parameter estimates. Thinning is a method used in the MCMC algorithms to reduce autocorrelation in the generated samples [10, 11]. The basic idea of thinning is to keep only every l^{th} sample in the Markov chain and discard the rest. The choice of the thinning parameter l should be carefully tuned to balance the reduction in autocorrelation with the loss of efficiency/information.

[1] A low autocorrelation between samples is desired in MCMC since each sample provides independent information about the target distribution, and hence the uncertainty estimates are reliable.

[2] Lag refers to the distance between two points in a chain.

3. Results

To determine nPDFs using the MCMC method, we first perform a simplified test with a restricted number of parameters and data to keep the efficiency under control and be able to test different algorithms and settings. We used the adaptive MH algorithm [12, 13] to generate a Markov chain for 10 nPDF parameters (3 for u-valence, 3 for d-valence, 2 for light sea quarks, and 2 for gluons), see Fig. 1. In this preliminary study, we restrict ourselves only to data from deep-inelastic scattering (DIS) experiments (including NMC, JLAB, and SLAC) which cover more than 15 different nuclei. In the next step, we performed thinning. This was done for two reasons: Firstly, to reduce the autocorrelation (see Fig. 2) and to be able to estimate the uncertainty using the standard prescription for Monte Carlo errors from Eq. (4) rather than MCMC error estimation (including autocorrelation time). Secondly, to generate an LHAPDF set of PDF grids [14] (a standard format for distributing PDFs), we need to limit the number of chain units to make it feasible and user-friendly. In Fig. 3, we compare the u-valence nPDF from the thinned chain with different thinning rates l. As we can see, for a higher thinning rate (200, represented by the red/black curve), we have a larger uncertainty. Whereas uncertainties obtained with thinning rates equal to 100 and 50 are basically identical. This indicates that thinning with $l = 200$ removes too much information, and a lower rate needs to be used to faithfully represent the uncertainties.

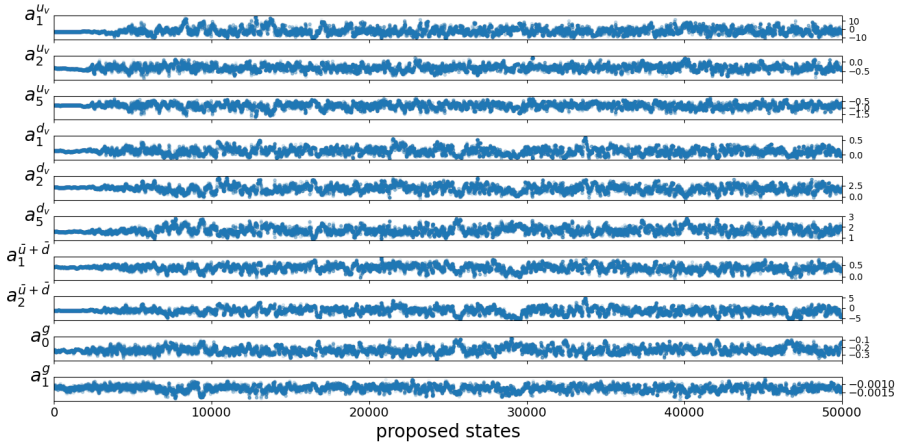

Fig. 1. Chains representing a time series of parameter values (10 nPDF parameters).

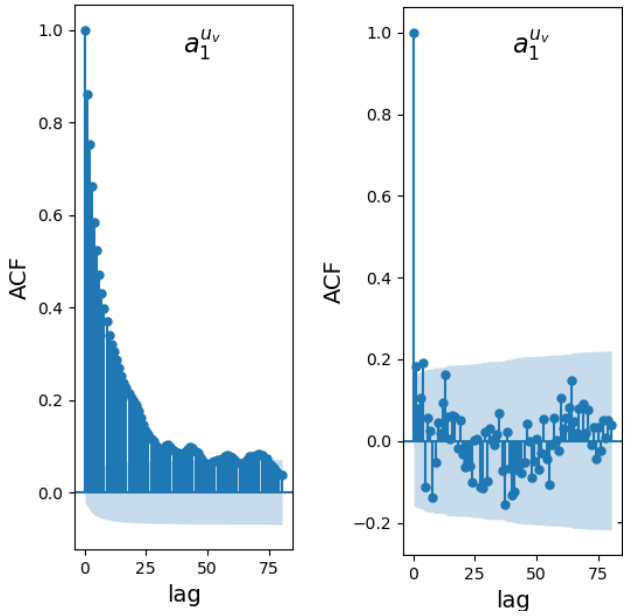

Fig. 2. Autocorrelation function (ACF) *versus* lag, before (left) and after (right) thinning. Here the thinning rate is 50.

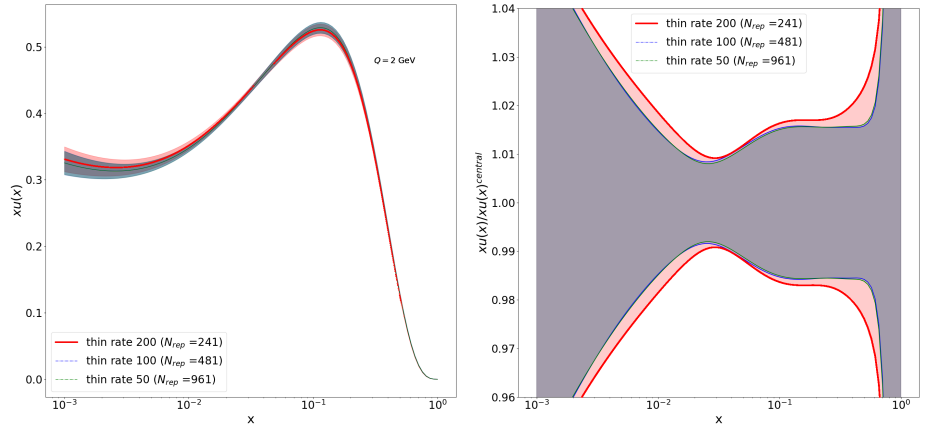

Fig. 3. (Color online) Full-PDF (left) and ratio-PDF (right) for thinned chain of Fig. 1. We compare the generated PDFs with different thinning parameter l: (200, 100, 50). The replicas are LHAPDF members. Each thinned chain unit corresponds to a replica PDF.

4. Conclusion

Despite the challenges, MCMC has proven to be a powerful tool for determining PDFs. It provides a robust and statistically rigorous framework for extracting PDFs from experimental data. The promising results that we obtained were for a simplified analysis, and now they need to be scaled up for more parameters and more data. Furthermore, we will also perform the analysis using the standard Hessian method and compare the results of the two approaches.

This work was supported by the National Science Centre, Poland (NCN) under grant No. 2019/34/E/ST2/00186. We also acknowledge the computational resources provided by PLGrid under the grant number PLG/2022/016037.

REFERENCES

[1] K. Kovařík, P.M. Nadolsky, D.E. Soper, *Rev. Mod. Phys.* **92**, 045003 (2020), arXiv:1905.06957 [hep-ph].

[2] J.J. Ethier, E.R. Nocera, *Annu. Rev. Nucl. Part. Sci.* **70**, 43 (2020), arXiv:2001.07722 [hep-ph].

[3] R. Abdul Khalek *et al.*, *Eur. Phys. J. C* **82**, 507 (2022), arXiv:2201.12363 [hep-ph].

[4] P. Duwentäster *et al.*, *Phys. Rev. D* **105**, 114043 (2022), arXiv:2204.09982 [hep-ph].

[5] K.J. Eskola, P. Paakkinen, H. Paukkunen, C.A. Salgado, *Eur. Phys. J. C* **82**, 413 (2022), arXiv:2112.12462 [hep-ph].

[6] K. Kovařík *et al.*, *Phys. Rev. D* **93**, 085037 (2016), arXiv:1509.00792 [hep-ph].

[7] W.R. Gilks, S. Richardson, D. Spiegelhalter, «Markov Chain Monte Carlo in Practice», *CRC Press*, 1995.

[8] W.K. Hastings, *Biometrika* **57**, 97 (1970).

[9] ALPHA Collaboration (U. Wolff), *Comput. Phys. Commun.* **156**, 143 (2004), arXiv:hep-lat/0306017; *Erratum ibid.* **176**, 383 (2007).

[10] W.A. Link, M.J. Eaton, *Methods Ecol. Evol.* **3**, 112 (2012).

[11] M. Riabiz *et al.*, *J. R. Stat. Soc. B* **84**, 1059 (2022), arXiv:2005.03952 [stat.ME].

[12] H. Haario, E. Saksman, J. Tamminen, *Bernoulli* **7**, 223 (2001).

[13] G.O. Roberts, J.S. Rosenthal, *J. Appl. Probab.* **44**, 458 (2007).

[14] A. Buckley *et al.*, *Eur. Phys. J. C* **75**, 132 (2015), arXiv:1412.7420 [hep-ph].