PROCEEDINGS OF SCIENCE

Small x extrapolation for parton distributions

Stefano Carrazza, Juan Cruz-Martinez and Roy Stegeman*

Tif Lab, Dipartimento di Fisica, Università di Milano and INFN, Sezione di Milano,
Via Celoria 16, I-20133 Milano, Italy
E-mail: stefano.carrazza@cern.ch, juan.cruz@mi.infn.it,
roy.stegeman@mi.infn.it

We present progress towards a new strategy to improve the precision of PDFs in the small x extrapolation region. In particular, we will show how a Gaussian Process can be used to model fake deep inelastic scattering data in the small x extrapolation region, and how, by treating this synthetic data as regular experimental data, it might be used in a PDF fit to control the uncertainties of PDFs in the extrapolation region. Finally, we will discuss current obstacles and possible solutions to these obstacles.

**** The European Physical Society Conference on High Energy Physics (EPS-HEP2021), ****
**** 26-30 July 2021 ****
**** Online conference, jointly organized by Universität Hamburg and the research center DESY ****

*Speaker

PoS(EPS-HEP2021)371

1. Introduction: the NNPDF methodology

Parton distribution functions (PDFs) are a critical component for precision physics at the LHC, and their accurate and precise determination is becoming increasingly important. Since PDFs describe the non-perturbative structure of the proton within the QCD factorization framework, they cannot be computed from first principles, and therefore they are determined using experimental data.

PDFs are determined by fitting a parametrized functional form at an input scale Q_0. A parametrization which is commonly used, is the following

$$xf_a(x, Q_0) = A_a x^{(1-\alpha_a)}(1-x)^{\beta_a}\mathcal{P}_a(x), \tag{1}$$

where the index a denotes the PDF flavours, x is the momentum fraction of the parton, and A_a is a normalization constant ensuring that the momentum and valence sum rules are satisfied. Finally, $\mathcal{P}_a(x)$ is a smooth functional form that is as general as possible, to prevent biasing the PDFs.

To achieve a parametrization that is as general as possible, PDF fitters have been using increasingly complex functional forms: starting from a simple polynomial with effectively $\mathcal{P}_a(x) = 1$ [1], to replacing $\mathcal{P}_a(x)$ with Chebychev [2], and Bernstein [3] polynomials, or, in the case of NNPDF, neural networks [4].

Here the $x^{(1-\alpha_a)}(1-x)^{\beta_a}$ polynomial prefactor controls the PDFs in the extrapolation region. The motivation for using this particular functional form for the prefactor has at the time of the first PDF determination been motived by theory arguments, in particular Regge theory which suggests a power-like behavior as $x \to 0$ [5], and the constituent counting rules which suggest a power-like behavior as $x \to 1$ [6]. However, it is important to note that the power-like behavior in these limits have not been determined from perturbative QCD calculations, which in fact suggest a logarithmic behavior at small x. Nevertheless, even if we assume that the prediction of power-like behavior is correct, it is unclear how the small or large x regions where this behavior holds are defined, and in particular whether they should hold also in the kinematic domain where the PDFs are provided by the PDF fitting collaborations. Furthermore, it is not clear at which scale Q this behavior should hold, and while it is enforced at the input scale Q_0, it is not preserved under Q evolution.

As opposed to many regression problems where the output of the regression model is directly compared to data, this is not the case for a PDF fit. Instead, the output of the neural networks correspond to the PDFs, and to compare them to experimental data a theoretical prediction of the corresponding observable needs to be calculated. In the NNPDF methodology this is done by performing a convolution of the PDFs with a `FastKernel` (FK) table [7, 8]. For deep inelastic scattering (DIS) processes this only requires a single convolution, while for Drell-Yan (DY) processes two convolutions are required, one for each colliding hadron. Finally, the figure of merit that is optimized during a fit is

$$\chi^2 = \sum_{i,j}^{N_{\text{dat}}} (D-P)_i C_{ij}^{-1} (D-P)_j, \tag{2}$$

where i, j denote the experimental datapoints, D_i the corresponding experimentally observed value, P_i the corresponding theoretical prediction, and C_{ij} the covariance between datapoints i and j, which can have both experimental and theoretical components [9].

The method suggested in this proceedings aims to control the uncertainties in the small x extrapolation region through the creation of synthetic data. Specifically, we employ a Gaussian process (GP) (see e.g. Ref. [10]) to model a selected subset of the experimental data, the prediction of which is then used to generate synthetic data in the small x extrapolation region. Finally, this synthetic data is treated as any other experimental dataset and included in a PDF fit.

2. Artificial experimental data in the extrapolation region

To test the accuracy of a PDF methodology in the data region, the NNPDF collaboration has developed Closure Tests (see Sect. 6.1 of [4]), which check whether the methodology is able to reproduce a known underlying truth. In the extrapolation region, however, it is impossible to test the accuracy of a methodology on experimental data and the most reliable operation to test whether the PDF uncertainties can accommodate unseen data. In the NNPDF context this is done by means of the so-called "future test" [11] by which the same methodology is applied to subsets of data in order to check the faithfulness of the PDF uncertainties, however this tell us nothing – by definition – on the accuracy of the PDF where no data is available. One could use the techniques of the Closure Test to generate new data for the Future Test but a Closure Test needs a prior PDF as truth which trivializes the result.

In this work we propose a methodology to increase the precision of the (nearby) extrapolation region by introducing fake data in the fit. A sort of data augmentation technique for PDF determination. While in the Closure Test method the fake data is generated from theory prediction we will instead generate data that follows the experimental trends (and, most importantly, uncertainties).

Within the NNPDF methodology it is assumed that the uncertainties of experimental measurements follow a Gaussian distribution. As a result, for the generation of synthetic data in the extrapolation region we need to provide a prediction of the central value and a Gaussian uncertainty. Let us denote the set of n experimental training datapoints as $(X, Y) = \{(x_i, y_i)|i = 1 \dots n\}$, where x_i denotes an input value in x and all input values are collected in the vector X, similarly y_i denotes the corresponding cross-section and all cross-sections in the dataset are collected in the vector Y. Because of the assumption that the experimental measurements follow a Gaussian distribution, we want to find a probability distribution $P(\bar{y}|X, Y; \bar{x})$ for $\bar{y}$ at a given $\bar{x}$ in the small x extrapolation region.

When using the polynomial prefactor such a probability is obtained as a result of controlling the extrapolation region (where the neural network saturates) with a functional form. Preferably, when modeling the experimental data, we would instead use a non-parametric method for the prediction in the extrapolation region. To this end, a particularly suitable tool is the Gaussian Process (GP). A GP is a jointly Gaussian distribution of random variables, meaning $\bar{y} \sim \mathcal{N}(\mu, \Sigma)$ with mean μ and covariance matrix Σ. In practice, a Gaussian Process is dictated by a covariance matrix, in this context commonly referred to as a kernel matrix, $\bar{y} = K(\bar{x}, X)$ which provides a description for the similarity between the training data and new data.

The kernel $k(x_i, x_j)$, a function of the points in input-space x_i, is a prior distribution providing a functional description for the elements of the kernel matrix K, different kernels may be tested to understand the impact of the choice of kernel on the generated data. Because of the properties of the covariance matrix of the experimental data, a suitable kernel with similar properties is the radial

basis function (RBF) (see e.g. Ref. [10]):

$$k(x_i, x_j) = \exp\left[-\frac{(x_i - x_j)^2}{2l^2}\right], \tag{3}$$

where l is a hyperparameter that is understood as the characteristic length-scale, and its value determined by maximizing the marginal likelihood $P(Y|X, l)$. To take into account also the uncertainties of the DIS data, the RBF kernel is extended by adding the values of the experimental covariance matrix.

As may be understood, the kernel provides a way of quantifying contiguity, providing a way to quantify a measure of proximity in the data space. We are particularly interested in applying this idea to the extrapolation region. Note that a loss of contiguity results in an increase in the spread of the uncertainty, which is exactly what we expect of a model for the extrapolation region.

Having understood the concept of a GP and why it might be of interest when addressing the problem of PDF extrapolation, we need to follow certain steps to include it in the NNPDF methodology. To begin with, we need to choose a DIS dataset of which the small x datapoints are on the edge of the global data region. A good example of such a dataset is HERA I+II inclusive NC e^+p 920 GeV [12]. Having identified a suitable dataset, the next step is to fit a GP kernel to a Q-bin (or multiple GPs to multiple Q-bins) of the selected dataset. The determined GP kernel can now be used to produce synthetic data on a grid in the extrapolation region. An example is shown in Fig. 1, where a GP with kernel parameters determined from the aforementioned dataset is used to produce synthetic data in the extrapolation region. The synthetic data needs corresponding theoretical predictions as encoded in FK tables, allowing for the fast evaluation of the corresponding observables during the training of the parameters. This FK table can then finally be used to include the dataset in an NNPDF fit.

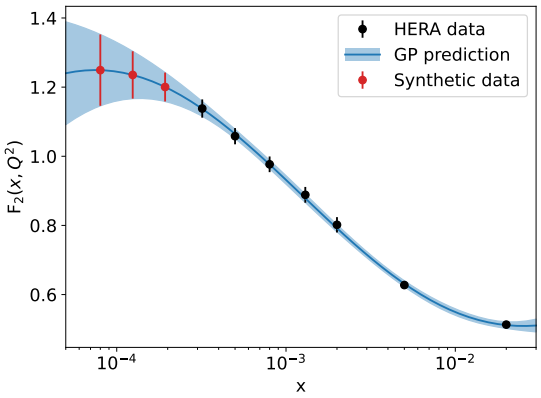

Figure 1: An example of a GP fitted to HERA data showing how synthetic data in the extrapolation region can be modeled using a GP. For the GP prediction the 1σ uncertainty band is shown, as well as the 1σ error bars of both the synthetic data and the experimental data.

3. Fits with artificial data

Here we study the results of a fit to the global NNPDF4.0 data data (see appendix B of Ref. [4]), to a fit using the same data extended with synthetic data in the extrapolation region $10^{-6} \lesssim x \lesssim 2 \cdot 10^{-5}$ produced as described in Sect. 2.

The results of these fits are shown in Fig. 2, where a comparison is provided both at the level of the PDF and at the level of the synthetic data. While it is clear that the synthetic data has been able to control the PDF uncertainty in the extrapolation, it can also be observed that the the prediction for both PDFs appear to follow the same trend. This suggests that the kinematic domain where the synthetic data is available is still controlled by the polynomial prefactor as shown in Eq. (1). This is possible if the neural network saturates in the range of the synthetic data, of which it has been confirmed that it is indeed the case by explicitly checking the values of the neural network parameters after training.

To understand why this happens we need to turn to the figure of merit Eq. (2), from which it is clear that as a result of the large uncertainties of the synthetic data observables, the impact of the synthetic data on the χ^2 is limited.

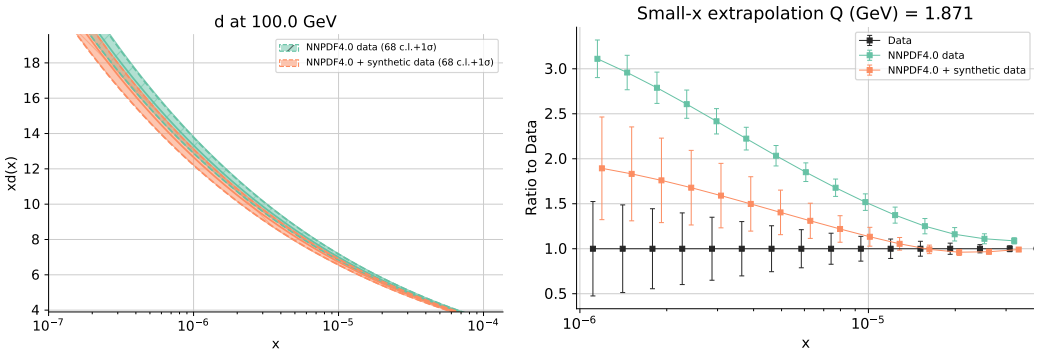

Figure 2: Comparison between a fit to NNPDF4.0 data (green) and a fit to the NNPDF4.0 dataset as well as the synthetic data produced using a Gaussian Process (orange). A comparison is shown between the down PDFs corresponding to each fit (left), and the predictions normalized to the synthetic data (right). Note that the down PDF is shown on a wider range of x than for which the synthetic data has been generated.

4. Outlook

A possible course of action to address the problem of large uncertainties is to improve the methodology used to generate the synthetic data. In particular a possible next step could be to adjust the RBF kernel, specifically by changing the prior distribution one can obtain a smaller uncertainty for the prediction of the GP in the extrapolation region. However, this is something that needs to be done with much care in order to prevent biasing the result.

Finally, it should also be noted that the exercise as presented in this work knows some limitations. Firstly, when done in this way, only DIS data can be generated in the extrapolation region. The reason for this is that it is the only process type for which the measurements depend on the kinematic variable x. Secondly, because the aim is to extrapolate into the small x region, the training data for

the GP should be the DIS data at the low x boundary of the data region. The suitable experimental data only cover a limited range of scales Q^2 at an order of magnitude of several GeV^2. Because of the limited diversity in the feature of the suitable data, ideally, the methodology would be extended to allow for the inclusion of more data in the training of the model.

Acknowledgments

SC thanks Jose I. Latorre for discussions on Gaussian Process models for PDF determination. The authors are supported by the European Research Council under the European Unions Horizon 2020 research and innovation Programme (grant agreement number 740006).

References

[1] R. McElhaney and S.F. Tuan, *Some consequences of a modified Kuti Weisskopf quark parton model*, *Phys. Rev. D* **8** (1973) 2267.

[2] S. Bailey, T. Cridge, L.A. Harland-Lang, A.D. Martin and R.S. Thorne, *Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs*, *Eur. Phys. J. C* **81** (2021) 341 [2012.04684].

[3] T.-J. Hou et al., *New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC*, *Phys. Rev. D* **103** (2021) 014013 [1912.10053].

[4] R.D. Ball et al., *The Path to Proton Structure at One-Percent Accuracy*, 2109.02653.

[5] H.D.I. Abarbanel, M.L. Goldberger and S.B. Treiman, *Asymptotic properties of electroproduction structure functions*, *Phys. Rev. Lett.* **22** (1969) 500.

[6] S.J. Brodsky and G.R. Farrar, *Scaling Laws at Large Transverse Momentum*, *Phys. Rev. Lett.* **31** (1973) 1153.

[7] V. Bertone, S. Carrazza and J. Rojo, *APFEL: A PDF Evolution Library with QED corrections*, *Comput. Phys. Commun.* **185** (2014) 1647 [1310.1394].

[8] V. Bertone, S. Carrazza and N.P. Hartland, *APFELgrid: a high performance tool for parton density determinations*, *Comput. Phys. Commun.* **212** (2017) 205 [1605.02070].

[9] NNPDF collaboration, *A first determination of parton distributions with theoretical uncertainties*, *Eur. Phys. J.* **C** (2019) 79:838 [1905.04311].

[10] C.K. Williams and C.E. Rasmussen, *Gaussian processes for machine learning*, MIT press Cambridge, MA (2006).

[11] J. Cruz-Martinez, S. Forte and E.R. Nocera, *Future tests of parton distributions*, *Acta Phys. Polon. B* **52** (2021) 243 [2103.08606].

[12] H1, ZEUS collaboration, *Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data*, *Eur. Phys. J. C* **75** (2015) 580 [1506.06042].