A New Avenue to Parton Distribution Functions Uncertainties: Self-Organizing Maps

Abstract

Neural network algorithms have been recently applied to construct Parton Distribution Functions (PDFs) parametrizations as an alternative to standard global fitting procedures. We propose a different technique, namely an *interactive* neural network algorithm using Self-Organizing Maps (SOMs). SOMs enable the user to directly control the data selection procedure at various stages of the process. We use all available data on deep inelastic scattering experiments in the kinematical region of $0.001 \leq x \leq 0.75$, and $1 \leq Q^2 \leq 100$ GeV^2, with a cut on the final state invariant mass, $W^2 \geq 10$ GeV^2. Our main goal is to provide a fitting procedure that, at variance with standard neural network approaches, allows for an increased control of the systematic bias. Controlling the theoretical uncertainties is crucial for determining the outcome of a variety of high energy physics experiments using hadron beams. In particular, it is a prerequisite for unambigously extracting from experimental data the signal of the Higgs boson, that is considered to be, at present, the key for determining the theory of the masses of all known elementary particles, or the Holy Grail of particle physics.

PACS numbers: 13.60.-r, 12.38.-t, 24.85.+p

I. DEFINITION OF THE PROBLEM AND GENERAL REMARKS

In high energy physics experiments one detects the **cross section**, σ_X, for observing a particle X during a given collision process. σ_X is proportional to the number of counts in the detector. It is, therefore, directly **"observable"**.

Consider the collision between two **hadrons** (a hadron is a more general name to indicate a proton, neutron...or a more exotic particle made of **quarks**). σ_X is expressed as:

$$\begin{aligned} \sigma_X &= \sum_{a,b} \int dx_1 dx_2 \, f_{a/h_1}(x_1, Q^2) \, f_{b/h_2}(x_2, Q^2) \, \hat{\sigma}_{ab}(x_1, x_2, Q^2) \\ X &= W, Z, H, \ldots \end{aligned} \tag{1}$$

where:

- $\hat{\sigma}_{ab}(x_1, x_2, Q^2)$, is the cross section for the elementary collision process between two quarks in the initial hadrons, producing a particle X. It is calculated theoretically.

- $X = W, Z, H, \ldots$ defines the kind of particle produced in the high energy process (each particle is defined by a letter, *e.g.* H = Higgs boson). The processes we are interessted in will be measured at the Large Hadron Collider in CERN (LHC).

- $f_{a/h_1}(x_1, Q^2)$, $f_{b/h_2}(x_2, Q^2)$ are the **Parton Distribution Functions (PDFs)**. They are the *unknowns* we are after. They cannot be calculated theoretically, and they need to be extracted from experiment simultaneously to the cross section producing the signal of *e.g.* a Higgs particle.

- "Parton" is a term invented by Richard Feynman [1] to denominate the constituents of the proton, the *quarks* and *gluons*. The notation for a parton distribution function, $f_{a/h_1}(x_1, Q^2)$ is as follows:
  - a/h_1 refers to the distribution of parton of type a in hadron h_1. The Standard Model of particle physics predicts 13 types of partons: 6 quarks (up, down, strange, charm, bottom, top); 6 anti-quarks (anti-up, anti-down, etc...), and 1 gluon. The gluon is a special type of particle that carries the interaction among the quarks, and it is therefore more elusive in experiments.
  - x_1 and Q^2 define the *kinematics*. They are in turn functions of the momenta, angles and energies of the hadrons detected in the experiment. They are therefore observables.
  - x_1 is the longitudinal component of the momentum fraction taken by parton a in h_1.
  - Q^2 is the energy scale for the process. The range of kinematical variables $x_1 \equiv x$ and Q^2 (GeV^2), measured in different experiments at different laboratories worldwide is shown in Fig.1.

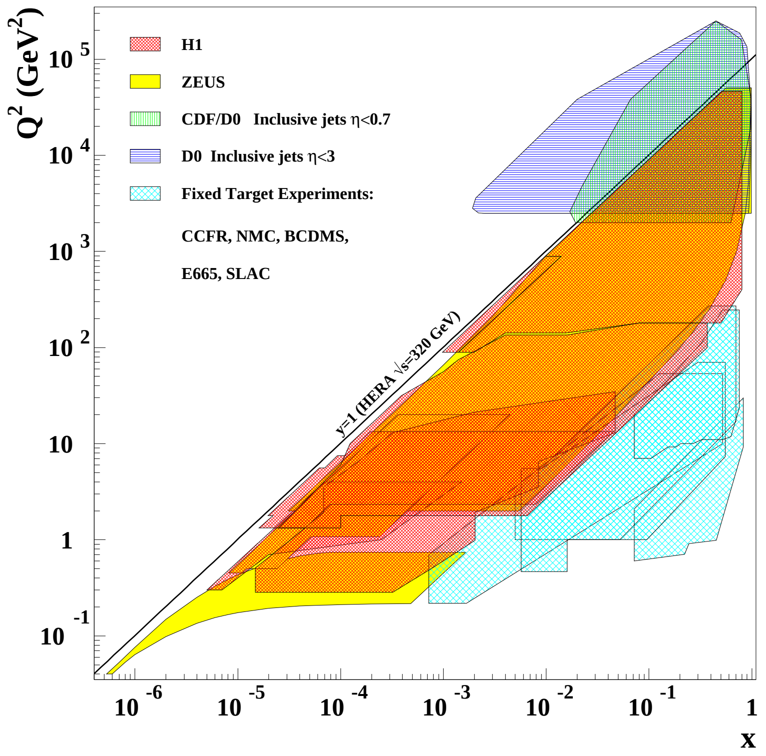

FIG. 1: Kinematical range covered in deep inelastic scattering experiments worldwide.

- PDFs have the property of **"universality"**. That is, since they describe more fundamental objects, the quarks and gluons, we can measure them in one particular experiment at a given scale, $Q^2 \equiv Q_1^2$, and "plug" them in as contributions to another experiment, at another scale, $Q^2 \equiv Q_2^2$. Here we will consider PDFs extracted from a wide cathegory of experimets called Deep Inelastic Scattering (DIS), and we will study their prediction for LHC observables.

- The connection between PDFs at different scales is provided by a set of coupled integro-differential equations called **DGLAP** evolution equations (**D**okshitzer, **G**ribov, **L**ipatov, **A**ltarelli, **P**arisi). Although the structure of DGLAP is known theoretically, they depend on a number of parameters and assumptions. This is how an evolution equation (not coupled, for simplicity) looks like:

$$\frac{d f_{a/h_1}(x, Q^2)}{d \ln Q^2} = \frac{\alpha_S(Q^2)}{2\pi} \int_x^1 \frac{dy}{y} f_{a/h_1}(y, Q^2) P_{qq}(x/y, \alpha_S(Q^2)) \qquad (2)$$

  (the parameters are embedded in the functions: α_S, and in P_{qq}).

Notice that PDFs are not directly observable, but that they contribute to the cross section through a structure function, $F_2(x,Q^2)$ defined as the following convolution:[1]

$$\sigma_{DIS} \propto F_2(x,Q^2) = x\sum_a e_a^2 \int_x^1 \frac{dy}{y} f_a(y,Q^2)\, C_a^{\overline{MS}} + x\int_x^1 \frac{dy}{y} g(y,Q^2)\, C_g^{\overline{MS}} \tag{3}$$

(the parameters are embedded in the functions: α_S, and in P_{qq}). The functions $C_{a(g)}^{\overline{MS}}$ are calculable theoretically (g indicates a gluon, it is a parton distribution but we write it separately because of its different features, e_a is the known quark's charge).

The detection of the Higgs particle and of new physics embedded in σ_X (Eq.(1)) depends crucially on our ability to predict and extract PDFs from experiment.

II. METHOD

There exist two broadly defined approaches to attack this problem. One is the "conventional" approach summarized in [3]. An alternative one, developed more recently, is the Neural Network approach (NNPDF Collaboration [4]).

The goal is to determine the PDFs and their uncertainties from the known region, or the region where data exist (Fig.1), and to extrapolate them to the region to be measured at the LHC through the DGLAP evolution equations.

The kinematical region covered at the LHC is much larger than the one presented in Fig.1. We show it in Fig.2 (compare the blue triangle with the green one).

A Conventional Approaches

The conventional approach proceeds by parametrizing the various PDFs as a function of the variable x, at a given initial scale: $Q^2 \equiv Q_o^2$. An initial parametrization looks like:

$$f_a(x,Q_o^2) = A_1 x^{A_2}(1-x)^{A_3} P(A_4, A_5, ...) \tag{4}$$

where A_i are unknown parameters. The mathematical form is chosen in such a way that it can reproduce the experimental behavior of F_2 (Eq.(3)). The parameters are fine tuned by plugging (4) as the initial condition for the integro-differential DGLAP equations, by stopping the evolution at the different scales Q^2 where the experimental data exist, and by performing the fit at each scale. Theoretical errors on the PDFs are introduced at each step. The main sources are:

- the functional forms necessarily cover ranges of x where data do not exist: these affect the final outcome because DGLAP requires an integration in the variable x. In order to understand the role of the contributions from different regions of x we are

[1] We will assume by default that all calculations are done in $\overline{MS}$ scheme, see *e.g.* Ref.[2] for a review.

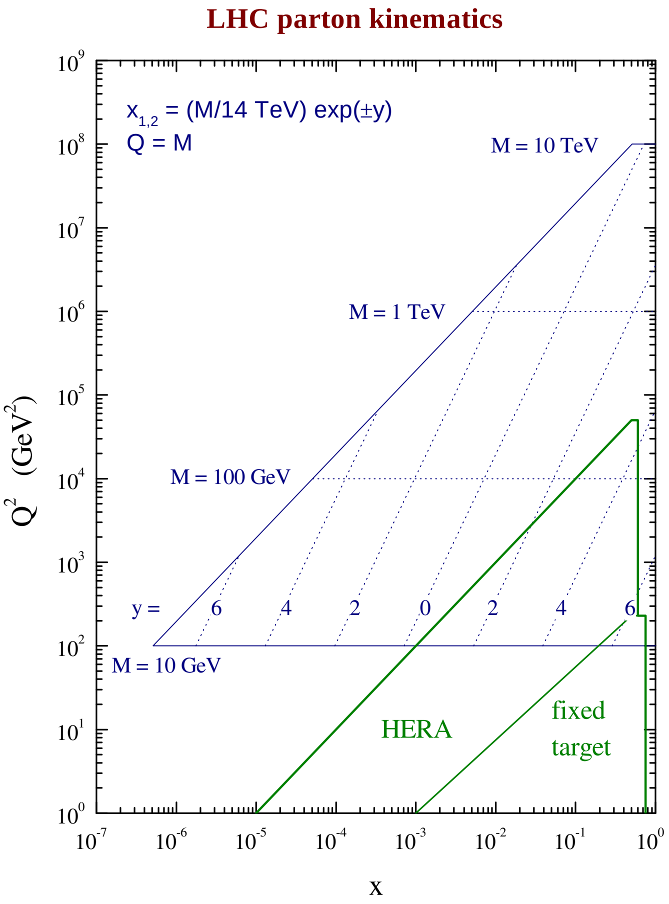

FIG. 2: Kinematical range covered at the Large Hadron Collider (LHC).

varying the PDFs by $1-2\%$ at fixed values of x and studying how the perturbation propagates (in what direction in the x and Q^2 plane, and what PDFs other than the varied one are affected, as a consequence of the PDFs being couple equations) (*figure in preparation*).

- even if within a given fit and choice for f_a one reproduces the data for F_2, a similar fit using different forms for f_a can give an equivalently good fit because the PDFs contribute to the obsevables, F_2, in an under-constrained linear combination.

  The error from the different PDFs contributions can be reduced by considering a wide variety of experiments (global fitting). However this also introduces other types of theoretical uncertainties because by "global fitting" the PDFs need to be extracted from different types of cross sections, which is not as clean and controlled as in DIS. In Fig.3 we show the accuracy with which one evaluates F_2 *e.g.* in the small x region (similar graphs exist for larger values of x).

- corrections inherent in DGLAP. These are in better control (see later).

An example of the uncertainties on PDFs is shown in Fig.4. Two approaches have been developed to treat these uncertainties: the Lagrange Multiplier method, and the Hessian

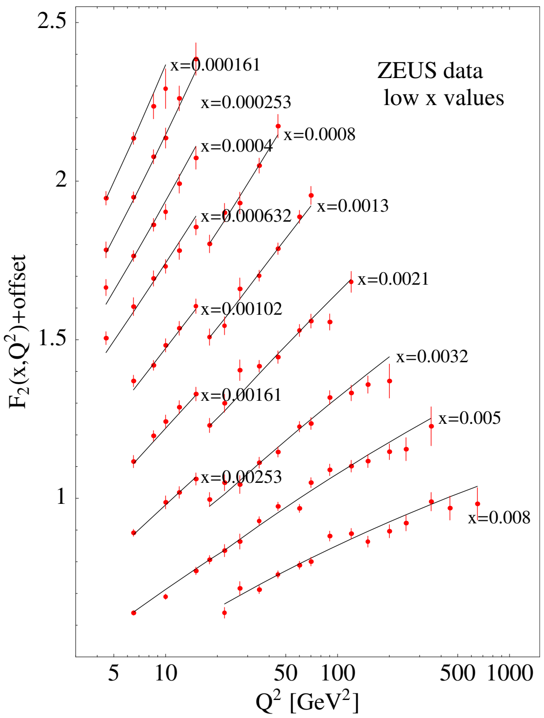

FIG. 3: Structure function, $F_2(x, Q^2)$ plotted vs. Q^2 as a function of x.

method (these are described in [6] and will not be discussed here).

B NNPDFs and SOMPDFs

In the NN approach one can in principle eliminate the model dependence implicit in Eq.(4) by starting from a randomized input, and by training a NN using the experimental data. Initial approaches did not use DGLAP (see however [5]). The process is fully automated. It does not offer flexibility to include *e.g.* other data in order to perform a global analysis, or variations of the initial assumptions.

Our approach (Self-Organizing-Map PDFs, SOMPDFs) was developed with the specific goal to let the "users" interact and control the so-called flexible points. (A detailed description of SOMPDFs is not given here, see paper in preparation). Flexible points are defined as: *"opportunities for adapting and fine tuning"* a simulation.

We identified the following ones:

List of Flexible Points

1. Construction of the "initial state". This is made by an "envelope" of curves displayed as a function of x, at a fixed initial scale Q_o^2. The curves are obtained from the randomized contribution of all available functional forms (similar to Eq.(4)) of the PDFs. The PDFs resulting from the randomization process are subsequently smoothed out, in order to apply DGLAP, with an interpolation procedure.

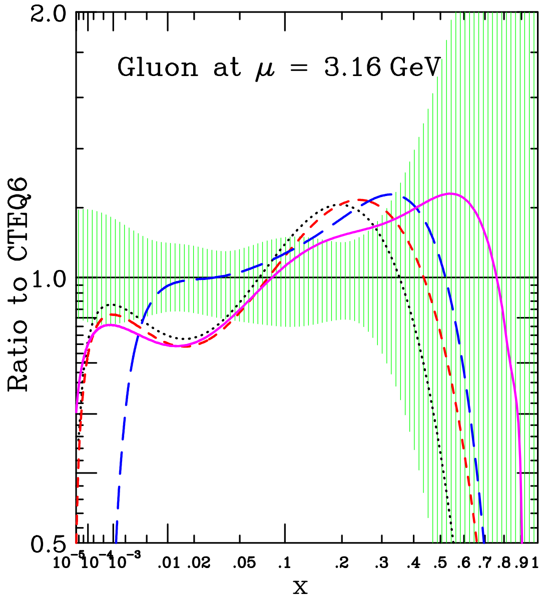

FIG. 4: Uncertainties on the gluon distribution calculated at the scale $\mu^2 = Q^2 = 10$ GeV^2 [6]. The hatched area is the uncertainty band evaluated for a particular PDF set (CTEQ6, see y-axis label). The curves correspond to the ratios of variations of the PDF fits to the nominal one. One can observe that for this particular comparison, the curves lie outside the uncertainty bands at small x. This difference is due to assumptions made on the parametrization shape.

2. Size of the SOM.
3. Selection process within the SOM: how many PDFs are selected at each iteration, and from what regions in the map.
4. Criterion of selection process: it is now based on a global χ^2 and on "similarity" among the curves
5. Role of different data sets covering different kinematical regions: what happens if different data sets are inserted or taken out.
6. "Tweaking" DGLAP:
   - role of various parameters: $\alpha_S(M_Z^2)$, heavy quarks thresholds,...
   - structural changes: introduction of non-linear terms
   - extra sources of Q^2 depedence beyond DGLAP type evolution (Higher Twists)

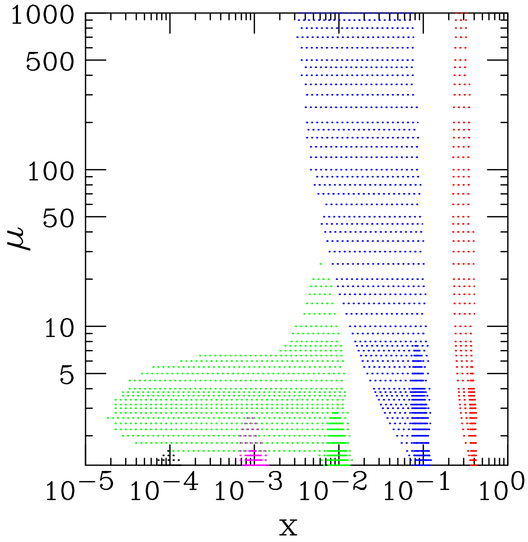

FIG. 5: Possible interaction of flexible points. Adapated from Ref.[6]

Interaction of Flexible Points

We expect the flexible points listed above to interact considerably, and to produce correlations. For example, modifying one PDF will produce a "ripple effect" within the behavior of the same PDF, but in different kinematic domains, and on all the others. This is illustrated in Fig.5 where the impact of a small variation on the x axis is shown to propagate and to therefore affect the PDF behavior in Q^2. We would like to explore systematically and quantitatively these types of correlations.

Emerging Behavior

An important goal is to determine whether one can single out a particular behavior of PDFs with different characteristics, for instance:

- whether a larger gluon distribution is preferred over a flatter one, etc., in the determination of the cross section (1).
- whether gluon distributions can be better determined in some types of experiments with respect to others (this would help defining a list of preferred reactions)
- whether different PDFs "track" each others behavior by establishing proportionality relationships among them
- whether DGLAP breaks down in particular kinematical limits that will be accessible in the wide domain explored at LHC (Fig.2)

REFERENCES

[1] R.P. Feynman, *"Photon-Hadron Interactions"*, Addison Wesley Longman, Reading, Mass., 1972.

[2] R.K. Ellis, W.J. Stirling, and B. Webber, *"QCD and Collider Physics"*, Cambridge University Press, Cambridge, United Kingdom, 1996.

[3] J. M. Campbell, J. W. Huston and W. J. Stirling, arXiv:hep-ph/0611148.

[4] S. Forte, L. Garrido, J. I. Latorre and A. Piccione, JHEP **0205**, 062 (2002)

[5] A. Piccione, L. Del Debbio, S. Forte, J. I. Latorre and J. Rojo [NNPDF Collaboration], arXiv:hep-ph/0607199.

[6] J. Pumplin, AIP Conf. Proc. **792**, 50 (2005)