UNIVERSITÀ DEGLI STUDI DI MILANO

Laurea Triennale in Fisica

Overfitting and Gaussianity of Parton Distribution Functions

Relatore:
Prof. Stefano Forte

Correlatore:
Dott. Juan Manuel Cruz-Martinez

Tesi di Laurea di:
Francesco Paolo Guerci

Matricola:
917465

A.A.
2020/2021

Acknowledgements

Innanzitutto intendo ringraziare il Prof. Stefano Forte. Partecipare anche in minuscola parte ad un progetto tanto complesso quanto ambizioso è stato per me motivo di grande orgoglio. Inoltre, intendo sottolineare la mia ammirazione per l'eccelsa competenza con cui svolge il ruolo d'insegnante, senza la quale non avrei avuto la possibilità di apprezzare le idee rivoluzionarie della Meccanica Quantistica che hanno cambiato irreparabilmente il modo in cui osservo le cose.

Un sentito ringraziamento va al Dott. Juan Cruz-Martinez, il quale durante il periodo di tesi mi ha fornito una disponibilità totale, colmando ogni mia carenza teorica e soprattutto informatica, intevenendo anche in prima persona per sbrigare problemi la cui risoluzione mi avrebbe richiesto anche diversi giorni di patimento.

Queste righe sono per i miei genitori, Alessandro e Giuliana, i quali sono stati i pilastri fondamentali su cui si sono saldamente costruite la mia vita e la mia carriera, e su cui si baserà ogni impresa che riuscirò a compiere nel futuro. Intendo ringraziare mio padre, tra le tante cose, in particolare per aver visto più lontano di molti altri mostrandomi sin dal primo istante il suo entusiasmo per un percorso di studi così inconsueto. Senza le sue parole oggi non sarei dove sono. A mia madre, tra gli altrettanti motivi di gratitudine, devo la serenità con la quale mi è stato possibile affrontare questi lunghi anni, permettendomi di impegnarmi nelle mie cose senza distrazioni. Le piccole attenzioni quotidiane e le parole di consiglio, o di conforto, sono state molto più preziose di quanto abbia mostrato. Vi amo, il vostro orgoglio è il motore delle mie azioni.

Ai miei fratelli Elisabetta ed Edmondo, per avermi sostenuto in questo lungo arco, trattandomi sempre come se fossi di più di tutti in tutto, per la felicità del mio ego spropositato. Vi voglio bene, mi mancate e questo piccolo traguardo è nostro.

A Silvia, Ilaria e Valeria, delle persone meravigliose che hanno sempre trovato del tempo da dedicarmi e per cui non potrò mai essere all'altezza dell'opinione che hanno di me.

Un ringraziamento più ampio ai miei zii, dei secondi genitori i quali anche se lontani trovano sempre il modo di dimostrarmi il loro affetto incondizionato.

Ai ragazzi del tennis ed in particolare a Fabio, un ringraziamento speciale per il bene che mi ha dimostrato costantemente negli anni senza pretendere niente in cambio. Anche se non sembra, ho apprezzato e continuo ad apprezzare tutti i gesti senza mai dare niente per scontato.

Alla mia famiglia di Cirella: Andrea, Vittorio, Dario, Laura, Giacomo, Tommaso, Davide, Fulvia, Salvatore, Carolina, Francesca, Maurizio, Fabiano, Chiara... Ogni giorno maledico gli undici mesi che ci separano, non aspettando altro che rivedrvi. Quando siamo insieme è una delle poche occasioni in cui sento di essere nel posto giusto al momento

giusto.

Ai ragazzi di Napoli: Davide, Alessio, i molteplici Francesco, Augusto, Giovanni, Antonio, Andrea, Gennaro, Giuseppe, Maurizio, Vincenzo... Il tempo passato insieme mi regala sempre storie incredibili. A voi che mi fate sempre restare con i piedi per terra ricordandomi di prendere le cose con più leggerezza.

Ai miei colleghi di università: Miriam, Stefano, Jacopo, Riccardo, Carlo, John, Marta, gli Andrea, Tiziano, Francesca ... Non credo sinceramente che avrei conseguito questo traguardo senza di voi, mi avete insegnato tanto ed è stato bello svolgere questo cammino con voi. Una parte di me è un vostro riflesso.

A Dylan ed Edoardo, con cui ho vissuto anni importanti e nei quali ho trovato dei veri amici.

A Fabrizia e Federica che, in questi anni trascorsi lontano da casa, sono state preziose compagne di sorprendenti avventure.

A Giorgia, Allegra, Anna, Sofia e Martina che in così poco sono diventate dei piacevoli punti di riferimento.

Abstract

This thesis focuses on the study of a suitable figure of merit to check the quality of a Monte Carlo set of parton distribution functions (PDFs). A Monte Carlo set consists of several random replicas of the initial set of PDFs. The number of replicas is chosen in such a way that this set reproduces certain features of the initial one, such as the mean value and the uncertainties. The figure of merit is the statistical estimator χ^2, the minimization of which provides the best-fit configuration for each replica of PDFs. Following the method used by the NNPDF collaboration, this minimization process is carried out through the use of neural networks as an universal impartial interpolator. The χ^2 of the Monte Carlo set is then calculated by averaging over all the χ_i^2 of the individual replicas. The problem involved in this work concerns the observation that the assumption of Gaussianity on the PDF probability distributions around the central limit allowed the possibility of lower values of the χ^2 of the Monte Carlo set to be found by the neural networks. This observation implies that the central limit of the Monte Carlo set does not correspond with the PDF best fit. In order to eliminate the possibility of obtaining better χ^2 values, minimization has been made more aggressive forcing the neural network to perform overfitting. With this method, it was observed that the possibility of obtaining minor χ^2 values did not decrease. Rather, a systematic increase was seen as the intensity of overlearning increased. For this reason, it was deduced that these χ^2 values were not related to the possibility of finding better PDF fits. Using the PDFs resulting from this process, a significant loss of Gaussianity of probability distributions was also observed. From these observations it was possible to conclude that the possibility of χ^2 minors is due to the impossibility of applying the Gaussianity hypothesis in the determination of the PDFs, which evidently follow a different distribution pattern.

Contents

1 Theoretical framework

Around 1960, hundreds of new particles were discovered. These were initially supposed to be elementary, i.e. without an inner structure. This spectroscopy was possible thanks to the study of resonances in cross-section trends of scattered particles. Protons and neutrons were no longer considered the only hadrons: particles interacting by the strong nuclear force. These new particles were found with different electron charges, masses, and spins.

At that time, the main goal of particle physics was sorting and classifying this chaotic ensemble to get predictions about the spectrum and the behavior of hadrons. For this purpose, there was the need for a theoretical structure.

1.1 from Quark model to partons

The observation of several conservation laws and symmetries, such as the independence of strong interactions from electron charge, leads to consider protons and neutrons as the same objects, i.e. the nucleons, elements of symmetry group SU(2) of conserved isospin quantum number. Furthermore, when the conservation of strangeness was found, hadrons became elements of the symmetry group $SU(3)$.

Therefore, particles were sorted according to these numbers in multiplets, i.e. ordered schemes that allowed discovering new particles associating them with gaps in the pattern. In 1964 Gell-Mann ([1]) and Zweig ([2]) separately found that all occupied multiplets could be described by a combination of at least two fundamental representations of $SU(3)$, each of which was associated with an elementary entity, the quark.

This scheme made it possible to consider mesons made of a quark-antiquark pair and baryons made of three quarks. So, these new hypothetical point-like particles were proposed with a half-integer spin and fractional charge: $\pm 1/3, \pm 2/3$ in units of electron charge. In the beginning, just three quark types, or flavors, were needed to reproduce every multiplet: up, down, and strange quarks. Then, with the observation of new conservation laws, more quarks were added.

Thus, the Quark model, built on experimentally observed conservation laws derived by comparing hadronic initial and final states in collision experiments, was particularly efficient in predicting hadron spectroscopy. However, it was unable to describe the strong interaction behavior. Furthermore, the lack of experimental evidence of these entities leads one to consider this model just as a mathematical structure and quarks were not assumed to be real physical particles.

Two main discoveries brought these to be regarded as tangible bodies beyond reasonable doubts. The first one was performed at SLAC (Stanford Linear Accelerator Center) in Deep Inelastic Scattering (DIS) experiments ([3]). Here, high-energy leptons like elec-

trons were accelerated up to colliding against hadrons. Studying the deflection angles it was noticed that some proprieties of the cross-section could be explained assuming that scattering was performed on free point-like particles with fractional electron charge. This behavior was justified by the second discovery, namely Wilczek, Gross, and Politzer's formulation of a theory describing why strong interactions at high energy become weak ([4], [5]). This new theory goes under the name of Quantum Chromodynamics (QCD), a renormalizable quantum field theory describing the strong nuclear interactions, no longer experienced by hadrons but rather by their constituents, the partons. (for a more complete historical introduction see: [6])

1.2 Quantum Chromodynamics

QCD is the state of the art in describing the internal structure of hadrons and strong interactions. As a quantum field theory, this assumes that quarks interact through the exchange of new gauge bosons called gluons, also observed experimentally ([7]). Similarly to how the photon is the carrier of the electron charge in QED (Quantum Electrodynamics), the gluon is assumed to be the carrier of a new charge type in QCD, the color.

By the Pauli exclusion principle, a bound state isn't allowed to consist of three identical particles, like in the Omega baryon case. Hence, unlike the electrical charge that can appear only in two different natures, the colored one is necessarily conjugated into three types. So quarks now not only come in different flavors but also in different colors: red, green and blue, which attract each other, ensuring the hadronic stability. QCD furthermore, in agreement with time-energy Heisenberg's uncertainty principle, enables extracting the sea quarks from within the hadron, a quark-antiquark pair from the vacuum, without altering the total quantum numbers. This behavior is responsible for observing more flavors than the three assumed to make up the hadron, referred to as valence quarks.

In QCD, the main feature that differs from QED is that gauge bosons must have a charge. Indeed, while photons are electrically neutral, gluons must have a color to mediate interactions between different colored quarks. Hence, QCD allows gluons to interact with each other too, becoming a non-Abelian gauge theory. This structure is supposed to justify how chromodynamic force becomes stronger with increasing the observation length scale. If studying strong interactions on small-length scales with high-energy probes, quarks appear almost like free particles. This is known as asymptotic freedom. Instead, on the long-length scale with low-energy probes, QCD predicts strong interacting quarks. This is called infrared slavery.

This scale dependence is explainable assuming QCD as a renormalization theory of color charge. Gluons allow renormalizing color charge making it stronger or weaker according to the length scale. The impossibility of observing isolated quarks could be ex-

plained by the fact that by reaching length scales comparable to the size of the proton, it might be energetically convenient to extract sea quarks from the vacuum, which in turn will bind within new hadrons.

1.3 Parton Distribution Functions

Nowadays, the benchmark for particles physics is played by collisions experiments at the Large Hadron Collider (LHC), the largest and most powerful particles accelerator on the Earth. Since its activation in 2012, an intensive program was performed to investigate the Standard Model (SM) with more increasing accuracy. This way, even a small unexpected deviation could be considered as evidence of new physics.

The main experiments are the high-energy proton-proton collisions, where measurements of cross sections are carried out for the final states of initial interacting hadrons. All the related observables are then measured using these experimental values. From QCD, we can describe the cross-sections of two incoming hadrons h_1, h_2 producing a colorless, large-mass $(M_X >> m_p)$ final state X, as follows:

$$\sigma_X(S, M_X^2) = \sum_{a,b}^{n_f} \int_{x_{min}}^{1} dx_1\, dx_2\, f_{a/h_1}(x_1, M_X^2)\, f_{b/h_2}(x_2, M_X^2)\, \hat{\sigma}_{ab\to X}(x_1 x_2 S, M_X^2) \qquad (1)$$

In (1), the cross-section depends on: S, the center-of-mass energy squared of the particles' system $\left(S \simeq 13\,TeV^2\,at\,LHC\right)$, and M_x is the hadrons' mass of final state X. The convolution is made over the x_i variables, the momentum fractions carried by the parton a of h_i hadron. The range goes from a $x_{min} = \frac{M^2}{S}$ to 1, the value whereby the parton carries all the momentum of the initial hadron. Then, the summation is executed over all possible parton interactions, so a and b run over the different n_f flavors involved in the process.

The cross-section equation in (1) is the factorization theorem, it implies that the cross-section is given by two independent contributions. The first one is $\hat{\sigma}_{ab\to X}$, i.e. the *hard partonic cross-section* due to the small-distance interactions of two incoming partons a, b within the hadrons. With *hard*, we mean that there is a large physical scale such as M_X, and it implies that $\hat{\sigma}$ is computable in QCD perturbation theory thanks to Feynman diagrams, as in QED. On the other hand, the factors containing the long-distance information of the proton structure are $f_{a/h_i}(x_i)$, i.e. the Parton Distribution Functions (PDFs). If the hard partonic cross-section is computed at Leading Order (LO) of perturbation series for strong coupling factor α_s, PDFs are the probability density of extracting a parton a with fraction momentum x_i from the hadron h_i. This picture breaks down at higher orders, so they are not probability density because they can become negative.

These functions describe the non-perturbative regime of hadron interactions, and at

low energies, it's not possible to use Feynman diagrams to determine their x dependence. Since QCD doesn't allow one to compute the proton's wave function, PDFs cannot be calculated from first principles, so we have to extract them from experimental data. Once obtained the PDFs for some reference energy scale Q_0, the energy dependence is computable in perturbation theory with DGLAP equations:

$$Q^2 \frac{\partial^2 f_i(x,Q^2)}{\partial Q^2} = \sum_{j=1}^{n_f} P_{i,j}(x,\alpha_s(Q^2)) \otimes f_j(x,Q^2) \quad i = 1,...,n_f \tag{2}$$

In (2), $P_{i,j}$ are the perturbative kernels. With this set of integro-differential equations, it is possible to evolve the known PDFs set $\{f_i(x,Q_0^2)\}$ from Q_0 to any other $Q \neq Q_0$.
Additionally, from QCD we have some constraints on the PDFs. The conservation of momentum:

$$\int_0^1 dx\, x \left(\sum_{i=1}^{n_f} f_{q_i}(x,Q^2) + f_{\bar{q}_i}(x,Q^2) + f_g(x,Q^2) \right) = 1 \tag{3}$$

This is due to the conservation of energy. Here, we add over all parton PDFs: quark f_q, antiquark $f_{\bar{q}}$, and gluons f_g.
Two other sum rules are given by the conservation of flavor quantum numbers:

$$\int_0^1 dx \, \left(f_u(x,Q^2) - f_{\bar{u}}(x,Q^2) \right) = 2 \tag{4}$$

$$\int_0^1 dx \, \left(f_d(x,Q^2) - f_{\bar{d}}(x,Q^2) \right) = 1 \tag{5}$$

The (4) and (5) ensure that valence quarks provide a constant flavor contribution to the proton, for any Q and x there are always two up quarks and one down quark.

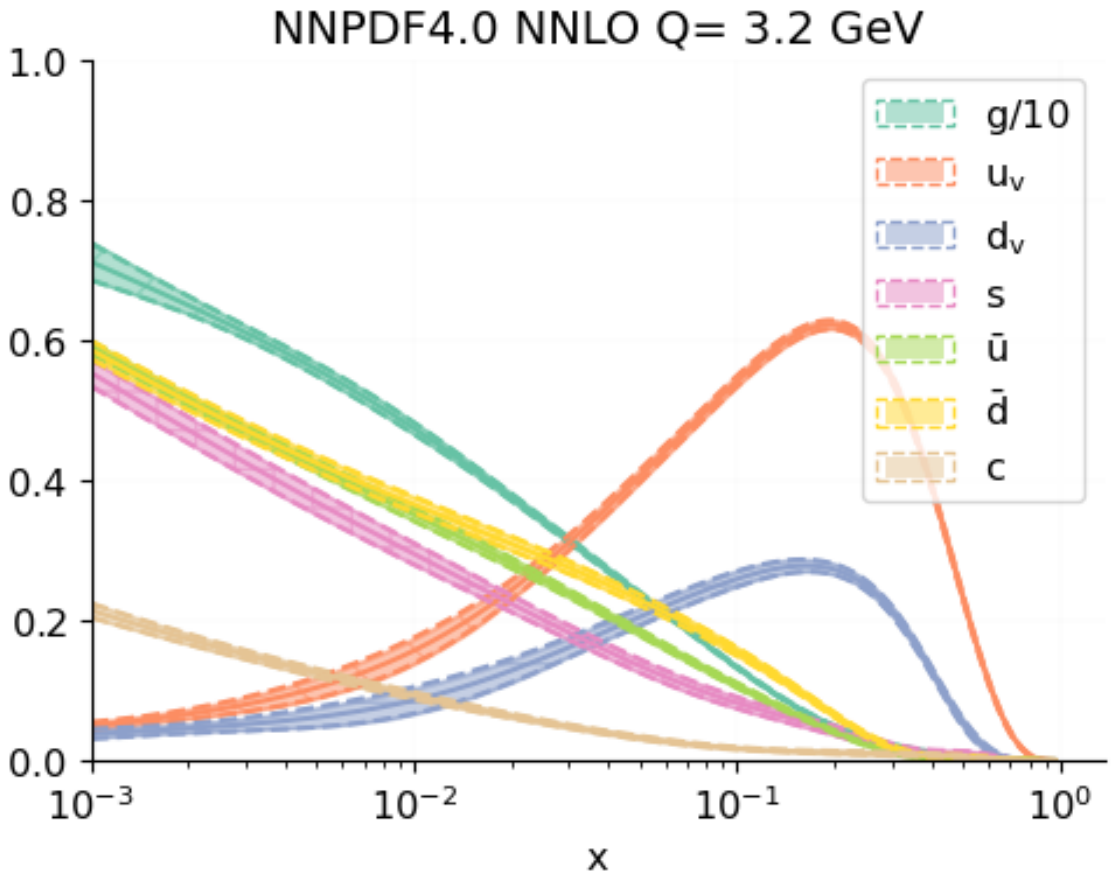


Figure 1: Parton distribution functions of proton obtained from NNPDF NNLO global PDF analysis, at $Q = 3.2GeV$ and $\alpha_s = 0.118$. (see [8])

PDFs play a special role in proton structure determination due to their universality. They are the same for different hadronic processes. Because of this we can separate the experiment-dependent contribution, given by the hard partonic cross-section, and the hadronic structure-dependent factors, the PDFs.

2 Fitting methodology

The PDF determination was a difficult task from the very beginning because of the need to extrapolate them from experimental data. This procedure raised several issues since we are searching for a set of continuous output functions starting from a finite set of data points. The impossibility to convert a finite to an infinite quantity of information can be viewed as setting the problem into an infinite-dimensional functional space with infinite solutions. For each PDF we search for the best fit and its uncertainties: the probability density in this functional space.

NNPDF collaboration solves this problem by employing two different features, both built-in the current NNPDF methodology and implemented in the `NNPDF4.0` code, allowing to convert the datasets into a probability functional in the PDFs space. The first one is a Monte Carlo approach for the density probability: the datasets involved in the fit are replicated many times randomly extracting data points around their true values with a Gaussian probability density. Then, neural networks are used on each datasets replica for finding acceptable solutions by minimization of a suitable figure of merit: the χ^2. (for a more general introduction see [8], [9])

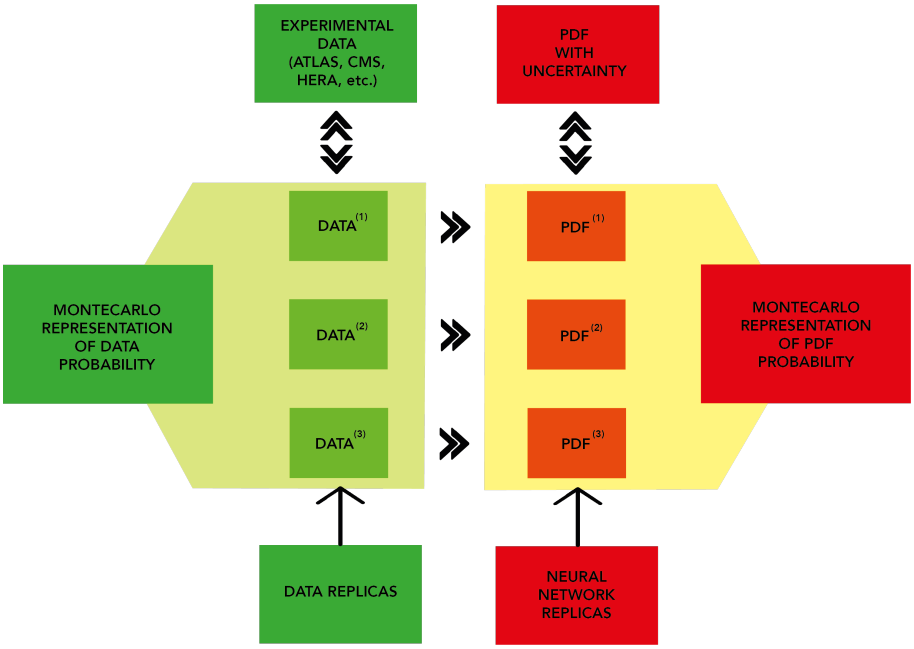


Figure 2: General approach of NNPDF collaboration

2.1 Experimental data

Before starting the fitting procedure, it is necessary to select the datasets used during the PDF determination. A dataset typically collects measurements of differential cross-sections, total cross-sections, or some related observables.

A fit requires comparing experimental measurements with theoretical predictions based on the PDF parameterization, so several cuts are made to the datasets, and data points without accurate enough theoretical predictions are removed, e.g. for missing higher orders in both QCD and electroweak perturbation theory.

Different datasets could describe different features of PDFs or add information in poorly known kinematic regions of the (x, Q^2) grid, so it's necessary to introduce redundancy in the kinematic coverage with different experiments. By observing the same features with many experiments, the peculiarities of PDFs can be freed from their initial process-dependence.

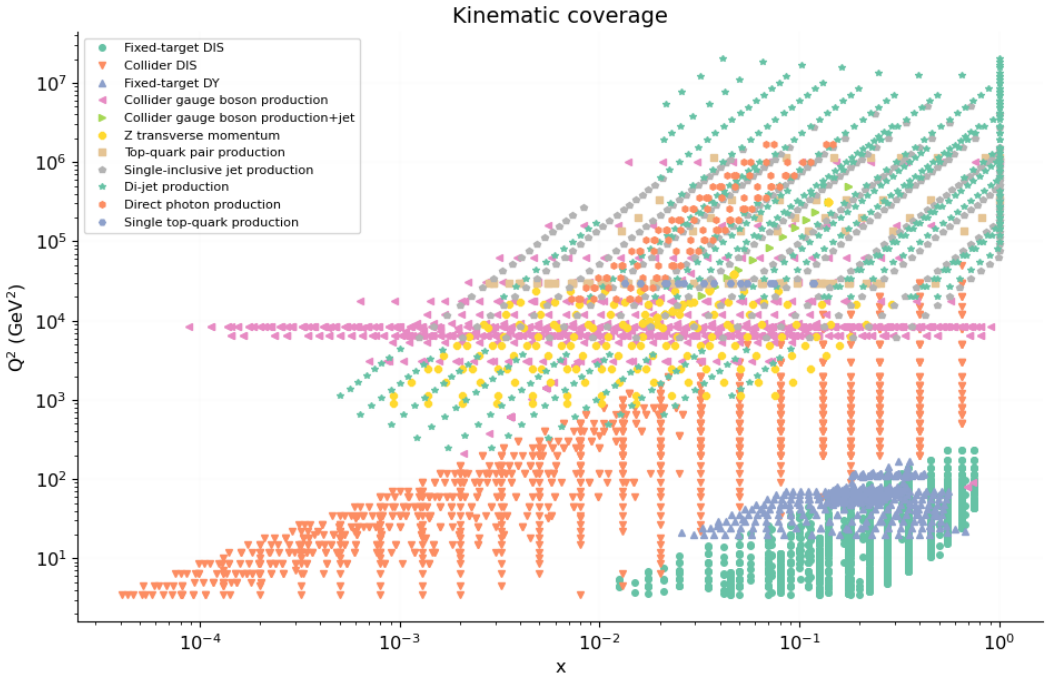


Figure 3: Current kinematic coverage of PDFs experimental datasets

A global fit is achieved by taking into account all available datasets, and after verifying

that they meet the characteristics described above. The experiments available for a global fit are):

- Fixed-target neutral-current deep-inelastic scattering (NC DIS): NMC , SLAC and BCDMS
- Fixed-target charged-current deep-inelastic scattering (CC DIS): CHORUS, NuTeV and NOMAD.
- Collider neutral- and charged-current DIS: HERA.
- Fixed-target Drell-Yan (DY): E866 (NuSea), E605 and E906 (SeaQuest).
- Collider gauge boson production: CDF, D0, ATLAS, CMS, LHCb.
- Collider gauge boson production with jets: ATLAS, CMS.
- Z boson transverse momentum production: ATLAS, CMS.
- Single-inclusive jet and dijet production: ATLAS, CMS.
- Direct photon production: ATLAS.
- Top-quark pair production: ATLAS, CMS.
- Single top-quark production: ATLAS, CMS.

(for a complete description: [10])

2.2 Neural networks

In the very beginning, different research groups tried to cast PDFs in a fixed functional form. This choice became doubtful when an uncertainty estimation was tried for the first time. Indeed, the parameter uncertainties obtained through minimization of the least-squares method and propagating errors were rather smaller then expected.([11], [12])

With the beginning of the LHC era, the PDF parametrization was expanded to accommodate the features brought by the new data collected, but the uncertainties became ever-increasing. Therefore, the choice of a fixed form seemed to be linked to this undesirable behavior of uncertainties, so the NNPDF collaboration tackled the problem by employing an artificial intelligence acting as an unbiased interpolator. Neural networks allow avoiding the introduction of arbitrariness in PDFs pattern choice.

A neural network can be represented as a graph made by interconnected nodes: input-output objects arranged on several layers. These are divided into *input* nodes or *activation* nodes. The former are at the start of connections chain and are used to provide the input value to the algorithm. The latter type has an associated *activation function* $f(x)$, which provides an output that is taken as input by the subsequently connected nodes. By considering an i activation node in a specific l layer, this gives the following output:

$$\xi_i^{(l)} = f\left(\sum_{j=1}^{N^{(l-1)}} w_{ij}^{(l)}\, \xi_j^{(l-1)} + \theta_i^{(l)}\right) \tag{6}$$

In (6), the function $f(x)$ takes as input the weighted sum of the outputs of all nodes in the previous layer $(l-1)$. In fact, each node i has its own threshold θ_i and each link connecting the output of a node j with the input of a node i, has an associated *weight* w_{ij}.

Training a neural network consists of optimizing all weights and thresholds, randomly initialized at the beginning of procedure. If we are taking into account a *feed-forward* neural network, the architecture is structured in single layer belonging nodes, and connections are allowed just for adjacent layers. NNPDF uses a feed-forward *fully connected* nodes, where each activation node is connected to all those in the previous and next layer.

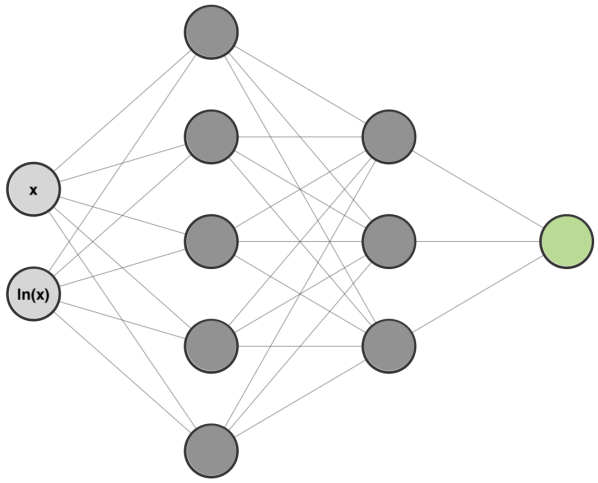


Figure 4: Old picture of a NNPDF neural network taking as input a pair $(x, \log x)$ and reporting as output a PDF value

2.3 Architecture and parameterization

The state-of-the-art NNPDF fitting methodology is the `NNPDF4.0` code, this will be described below.

PDFs are determined in their x dependence by neural networks at a fixed energy scale Q_0 and then evolved to higher energies via DGLAP equations. It is convenient considering the initial set of PDFs as the basis that diagonalizes these evolution equations, so we have:

$$
\begin{aligned}
g \\
\Sigma &= u^+ + d^+ + s^+ \\
V &= u^- + d^- + s^- \\
V_3 &= u^- - d^- \\
V_8 &= u^- + d^- - 2s^- \\
T_3 &= u^+ - d^+ \\
T_8 &= u^+ + d^+ - 2s^+ \\
T_{15} &= u^+ + d^+ + s^+ + 3c^+
\end{aligned}
$$

where $f_i^{\pm} = f_i \pm \bar{f}_i$ is a symmetric or anti-symmetric combination of quark-antiquark distributions.

Once the datasets to be used are selected, each of them is converted into a n dimensional vector $xgrid_i = \{x_i^{(1)}, x_i^{(2)}, ..., x_i^{(n)}\}$, made out of the single data points $x_i^{(k)}$ in the dataset. These vectors are taken, point by point, as input by the first layer of the neural network. The `NNPDF4.0` code uses a single densely connected network: a feed-forward multi-layer perceptron with an architecture: $2-25-20-8$. The input layer has two nodes that take the pair: $(x, logx)$. This choice is due to taking into account the different behavior of PDFs in the accessible physical region $10^{-4} \lesssim x \lesssim 0.5$. We have a linear regime in the $0.03 \lesssim x \lesssim 0.5$ domain and a logarithmic regime in the $10^{-4} \lesssim x \lesssim 0.03$ region. The two hidden layers in the middle use the $f(x) = \tanh x$ activation function. The output layer, with linear activation function $f(x) = x$, is made of 8 different nodes corresponding to the 8 flavor combinations of the basis, so all PDF combinations are determined at the same time.
Each of the basis elements is parameterized as:

$$
x\, f_i(x,Q) \; = A_i\, x^{1-\alpha_i}\, (1-x)^{\beta_i}\, NN_i(x) \tag{7}
$$

$NN_i(x)$ is the continuous output of the $i-th$ node in the final layer of the neural network and corresponds to the central contribution to the unnormalized combination of PDFs. The

constant A_i is an overall normalization factor that guarantees the validity of the conservation constraints. The $x^{1-\alpha_i}\ (1-x)^{\beta_i}$ is a preprocessing factor to control the trends of PDFs in the small and large x-regimes. The coefficients α_i and β_i are randomly selected with uniform distribution in a different range for each PDF. This range is determined self-consistently: the 68% confidence range is found for each PDFs combination, then the fit is repeated assuming that parameters can be distributed in a range twice the previous one. The procedure is iterated till the range stops changing.

Once all PDFs are obtained, it's possible to get the single flavor PDF through a rotation in their own space. To achieve the complete PDFs determination it's necessary to describe the minimization procedure performed by the neural network to get the $NN_i(x)$ factor.

2.4 The *loss function*

Unlike common A.I. (Artificial Intelligence) recognition problems, in PDFs determination is not possible to provide an input-output pair for each data point. Data are not direct instances of the functions, but rather each data point provides the measurement of some observable that depends in a non-linear way on distributions through a set of convolution integrals of all the PDFs in a range starting from a x_{min}.

The statistical estimator χ^2 is used to check the goodness of agreement between experimental data and theoretical predictions. For a fit comprising N_d datasets, the total χ^2 is given by:

$$\chi^2 \ = \ \sum_{d=1}^{N_d} \chi_d^2 \tag{8}$$

Assuming a fit with N_{dat} data, the χ^2 expressed in terms of the individual data points takes the following explicit form:

$$\chi^2 \ = \ \sum_{i,j}^{N_{dat}} \left(D_i - T_i\right) \left(cov^{-1}\right)_{ij} \left(D_j - T_j\right) \tag{9}$$

D_i are the data points corresponding to the experimental measurements of different datasets. T_i are the theoretical predictions. These are obtained through a set of convolution integrals of neural network PDFs. These computations are greatly accelerated by employing the FastKernel method ([13]). Considering a rank-4 luminosity tensor:

$$\mathscr{L}_{i\alpha\, j\beta} \ = \ f_{i\alpha}\, f_{j\beta} \tag{10}$$

where $(i\,,\ j)$ are the flavor indices and $(\alpha\,,\beta)$ labels the x grids. We can calculate the phys-

ical observables contracting a rank-5 FastKernel table FK for each separated datasets:

$$\mathscr{O}_n \;=\; FK^n_{i\alpha\, j\beta}\, \mathscr{L}_{i\alpha\, j\beta} \tag{11}$$

So the FastKernel tensor contains PDFs information about the convolutions with partonic cross-section and DGLAP energy evolution. These are stored in the form of pre-computed multiplicative factors. This process allows converting the neural network PDFs into a set of N observables, each associated with one of the N data points of the dataset. Then, the χ^2 compares the data directly with the theoretical predictions obtained with the neural network PDFs.

In (9) we have $(cov^{-1})_{ij}$, this is the inverse of covariance matrix correlating data points i and j within the same dataset or belonging to two different datasets. In NNPDF methodology, all the available information on uncertainties and their correlations is included. The covariance matrix has the following form:

$$(cov)_{ij} \;=\; \delta_{ij}\, \sigma_i^{uncorr} \sigma_j^{uncorr} \;+\; \sum_k^{N_{add}} \sigma_{i,k}^{add}\, \sigma_{j,k}^{add} \;+\; \left(\sum_k^{N_{mul}} \sigma_{i,k}^{mul}\, \sigma_{j,k}^{mul} \right) D_i D_j \tag{12}$$

The single data pair covariance is made by three different contributions: the first is given by the uncorrelated errors σ_i^{uncorr} obtained by adding the statistical and systematical uncorrelated uncertainties in quadrature. The second comes from correlated additive systematic errors $\sigma_{i,k}^{add}$, and the last is the correlated systematic multiplicative errors $\sigma_{i,k}^{mul}$ contribution.

2.5 Deterministic minimization

Having described all the parts that compose the χ^2, we can introduce the learning procedure. NNPDF collaboration employed the *Gradient Descent* (GD) method as *deterministic* minimization of χ^2. The `Adadelta` optimizer was chosen ([14]), with the `glorot_normal` initialization procedure ([15], [16]) for network parameters.

To allow the neural network to learn the PDFs patter while avoiding fitting noise requires the so-called cross-validation method. All the data points that make up the complete dataset used for the fit have to be divided into two subsets. The division is done in such a way that both sets have all the information necessary to describe the different features of the PDFs. The two subsets are the Validation and Training sets. The key idea exploited by this method is the following: while the two sets share the same underlying truths, the statistical fluctuations respectively associated with them are uncorrelated. For this reason only the Training set is used as input to train the neural network, i.e. to minimize the figure

of merit of this set: referred to as χ^2_{tr}. During this procedure, at each training epoch, we use the updated parameters of the network to evaluate the χ^2_{val} of the Validation set. The optimization of the training figure leads to an improvement of the validation one.

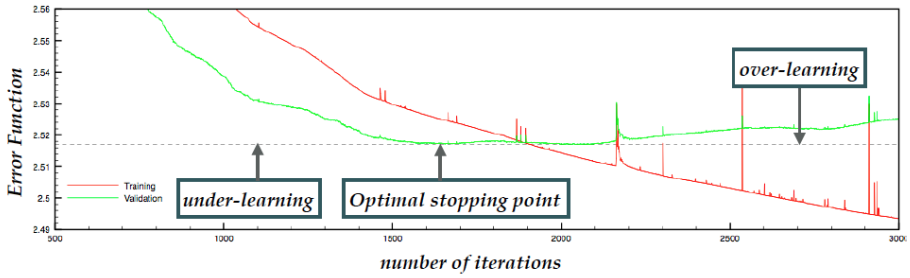


Figure 5: Typical trends of χ^2_{tr} and χ^2_{val} through learning epochs of NNPDF fitting method.

The χ^2_{tr} keeps ever-improving during the learning procedure, while after a certain time the χ^2_{val} systematically starts to worsen. This behavior is a symptom of overlearning. In fact, from this point onwards, the network starts to learn the noises associated with the training data in addition to the common underlying truths. Since the noises of the validation set are uncorrelated with those of the training one, the χ^2_{val} gets worse.

To avoid overfitting, `NNPDF4.0` provides the following stopping criterion: χ^2_{val} is monitored during minimization, when this stops improving, a patience algorithm is launched waiting for several epochs before quitting the learning. The only stopping points accepted are the ones where PDFs produce positive predictions for many experiments in different kinematic regions.

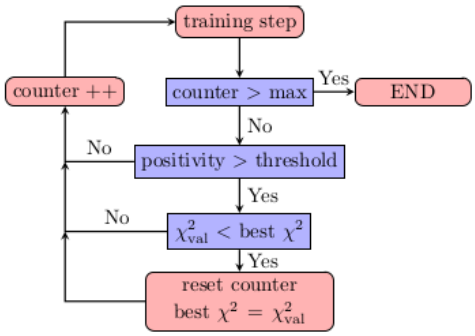


Figure 6: Flow scheme of the `NNPDF4.0` stopping criterion algorithm.

2.6 Monte Carlo method

Here it is presented the second main feature of `NNPDF4.0`, which allow getting a probability density profile in a functional space. Since these distributions are extracted from the data, their uncertainties have to be too. NNPDF collaboration tackled this problem with a Monte Carlo approach.

Starting from the experimental dataset $\{D_1, D_2, ..., D_{N_{dat}}\}$ containing N_{dat} data, we consider each data point D_i as the mean value of a stochastic variable with Gaussian distribution. These variables are correlated with each other via the covariance matrix $(cov_{t_0})_{ij}$. It is possible to construct a multi-Gaussian probability density in a $N_{dat}-$dimensional data space where each point is randomly extracted and corresponds to a dataset replica. For instance, the $k-$th point in this space is a random dataset: $\{D_1^{(rand)(k)}, D_2^{(rand)(k)}, ..., D_{N_{dat}}^{(rand)(k)}\}$. By producing many dataset replicas, we can reproduce the mean value of multi-distribution, i.e. our sample is statistically equivalent to the starting one, the experimental dataset.

Once a sufficiently high number of replicas N_{rep} has been generated ($N_{rep} \sim O(1000)$ to a percent level accuracy), the fitting methodology is applied to each replica, so for the $k-$th replica we have the following figure of merit:

$$\chi^{2(k)} = \sum_{i,j}^{N_{dat}} (D_i^{(rand)(k)} - T_i^{(k)}) \, (cov^{-1})_{ij} \, (D_j^{(rand)(k)} - T_j^{(k)}) \tag{13}$$

where $T_i^{(k)}$ are the theoretical prediction of the $k-$th neural network replica. At the end of the process there are N_{rep} PDFs set, so for each PDF we have a functional distribution in the PDF space, as shown in Fig.7.

Assuming that each PDFs set replica allows to computing an observable $X^{(k)}$, it's possible to get the mean value as:

$$< X > = \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} X^{(k)} \tag{14}$$

While the standard deviation is given by:

$$\sigma_X = \sqrt{\frac{1}{N_{rep}-1} \sum_{k=1}^{N_{rep}} \left(X^{(k)} - < X >\right)^2} \tag{15}$$

Furthermore, it can be noted in Fig.8 that a 100 replicas fit is sufficient for percent-level accuracy on uncertainties.

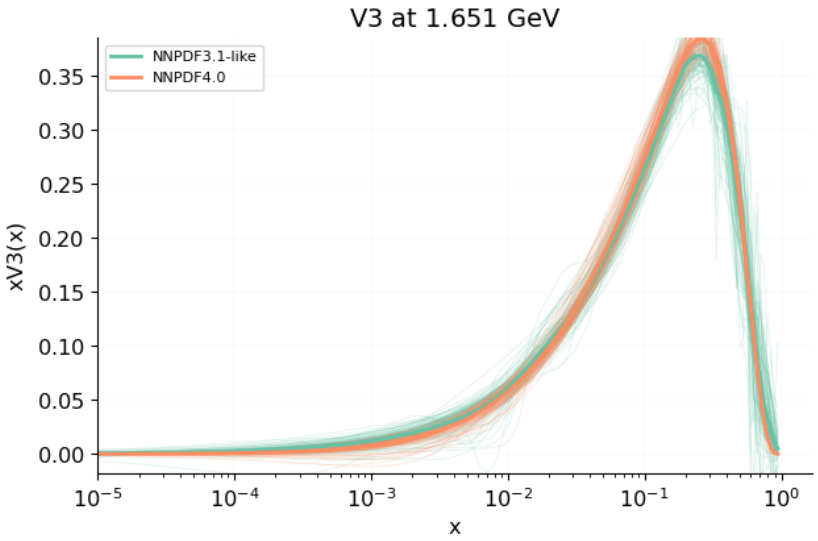


Figure 7: Comparison of the V_3 PDFs for the `NNPDF4.0` and NNPDF3.1 versions. Each set is made up of 1000 Monte Carlo replicas at $Q = 1.651 GeV$.

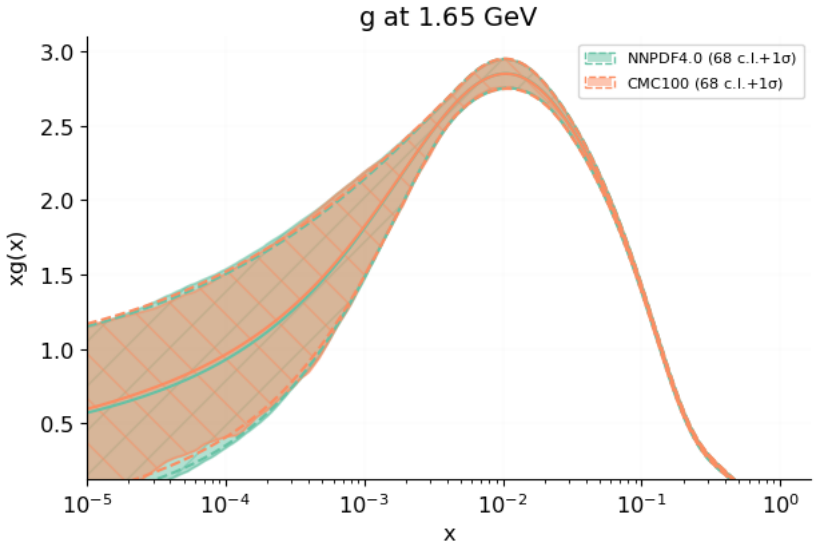


Figure 8: Comparison of gluon PDFs for the 1000 and 100 replicas Monte Carlo NNPDF4.0 global fit at $Q = 1.65\,GeV$.

2.7 Hyperoptimization

Hyperoptimization consists of hyper-scan many aspects of the fitting methodology, looking for the best parameter values ([17]). Hence, the object of this special optimization is the fitting method itself. Thanks to the great computational power brought with the introduction of the `NNPDF4.0`, it is possible to analyze in great detail many aspects of the fitting, e.g. the neural network structure: the number of layers, the number of nodes for each layer, the activation functions, the initialization function, etc. The fit option can be tuned too: the initial learning rates, the optimizer, the number of epochs during minimization, the stopping patience algorithm, etc.

The goal of the hyper-scan is to find the best methodology configuration to get better χ^2_{val} during the `NNPDF4.0` procedure. The minimization of χ^2_{val} performed by the hyperoptimization could lead to overlearning, due to the correlations between training and validation sets. In this framework, quality control is needed. This task is performed by the *k-fold* cross-validation. Data are divided into k partitions, each of which contains the main features of the full dataset. The fit is performed k times, and in each of them, a different partition is excluded from the procedure. This allow producing many loss functions that can be chosen for optimization, such as the mean value of the loss over the excluded partitions.

2.8 Hessian method

The Hessian approach is another method for the determination of uncertainties in PDF fits (see [18]). This is applied in the framework of fitting PDFs with fixed parameterization, and allows one to study the variation of the χ^2 by varying with continuity the fitting parameters.

Let be the PDFs parameterization set $\{p_1\,, p_2\,, ...\,, p_N\}$ composed by N parameters, this can be converted into a point $\vec{p} = (p_1\,, p_2\,, ...\,, p_N)$ of a $N-$dimensional parameter space. In this domain, we can consider the $\chi^2 = \chi^2(p)$, a function $\chi^2 : \mathbb{R}^N \to \mathbb{R}$. Let be $\vec{p_0}$ the point corresponding to the optimized parameter set: the one which minimizes χ^2, i.e. the global minimum of $\chi^2(p)$ in the parameter space. The Hessian method consists in assuming a quadratic approximation of χ^2 in a neighborhood of the $\vec{p_0}$.

With a Taylor expansion truncated at second order, for a point $\vec{p}$ sufficiently close to $\vec{p_0}$, ignoring higher orders we have:

$$\chi^2(\vec{p}) \;=\; \chi^2(\vec{p_0}) + \nabla\chi^2(\vec{p_0})\cdot(\vec{p}-\vec{p_0}) + \frac{1}{2}\sum_{i,j=1}^{N}(\vec{p}-\vec{p_0})_i \left.\frac{\partial^2\chi^2}{\partial p_i \partial p_j}\right|_{\vec{p}=\vec{p_0}} (\vec{p}-\vec{p_0})_j \quad (16)$$

In the global minimum $\vec{p}_0$ we have $\nabla\chi^2(\vec{p}_0) = 0$, so rewriting the $N \times N$ Hessian matrix:

$$H_{i,j} \;=\; \frac{1}{2}\frac{\partial^2\chi^2}{\partial p_i \partial p_j}\bigg|_{\vec{p}=\vec{p}_0} \tag{17}$$

Finally, the χ^2 variation in a neighborhood ca be written as:

$$\Delta\chi^2(\vec{p}) \;:=\; \chi^2(\vec{p}) - \chi^2(\vec{p}_0) \;=\; \sum_{i,j=1}^{N} (\vec{p}-\vec{p}_0)_i \, H_{i,j} \, (\vec{p}-\vec{p}_0)_j \tag{18}$$

To get standard deviations for an PDFs-dependent observable X, the Hessian method uses error propagation:

$$(\Delta X)^2 \;=\; T^2 \sum_{i,j=1}^{N} \frac{\partial X}{\partial p_i} (H^{-1})_{i,j} \frac{\partial X}{\partial p_j} \tag{19}$$

$T^2 = \Delta\chi^2$ is the *tolerance* parameter and was introduced after observing a systematically under-estimation of uncertainties. By choosing $T^2 = 1$ we get the Gaussian variance σ_X^2, the squared error related to the confidence interval for which we have at most $\Delta\chi^2 = 1$, with the confidence level of $\sim 68\%$. Hence, this parameter is used to set the confidence interval to get the same confidence level.

The inverse of the Hessian H^{-1} is the covariance matrix in the parameter space. This requires the computation of partial derivatives with respect to fitting parameters. Let be $\{\vec{e}_i\}_i^N$ a complete set of orthonormal eigenvectors for H_{ij}:

$$H\, e_i \;=\; e_i \tag{20}$$

$$e_i \cdot e_j = \; \delta_{i,j} \tag{21}$$

Expanding the difference $\vec{p} - \vec{p}_0$ on this vector basis:

$$\vec{p} - \vec{p}_0 \;=\; \sum_{k=1}^{N_{rep}} c_k \vec{e}_k \tag{22}$$

for some set $\{c_1, ..., c_N\}$ of real coefficients. By substituting in (19):

$$\Delta\chi^2(\vec{p}) \;=\; \sum_{i=1}^{N} c_i^2 \tag{23}$$

This equation defines an ellipsoid in a $N-$dimensional space of radius $\sqrt{\Delta\chi^2(\vec{p})}$ centered in $\vec{p}_0$, whose axes are the eigenvectors of H. By setting the T parameter, we determine the region of acceptable fits where a variation of $\vec{p}$ can produce a $\Delta\chi^2$ at most equal to T^2. It's possible to construct $2N$ eigenvectors spanning boundaries of the ellipsoid in the parameter space:

$$\vec{v}_k^{\pm} = \vec{p}_0 \pm T\,\vec{e}_i \tag{24}$$

To each eigenvector $\vec{v}_k{}^{\pm}$, we associate a set of parameters $F_k^{\pm}$ used to compute a set of PDFs. Once we have found the $2N$ sets of PDFs, the error of a generic observable is given by:

$$(\Delta X)^2 = \frac{1}{2}\sum_{i=1}^{N}\left(X(F_k^+) - X(F_k^-)\right)^2 \tag{25}$$

Even if for a Gaussian error we should have $T = 1$, in PDF determinations much bigger values are chosen to avoid underestimation of uncertainties. This could be due to minimizing a finite number of parameters during the fit, or perhaps to the incompatibility of PDFs fitted with different datasets, which could be affected by unknown correlations or missing higher-order theoretical computation.

2.9 From Monte Carlo to Hessian PDFs set

By construction, the Hessian method, with the quadratic approximation on the PDF parameter distribution, assumes that errors are Gaussianly distributed and calculates them with error propagation. The Monte Carlo method instead, allows checking the non-Gaussianity of a PDF replica set. A test of Gaussianity is for instance achieved by comparing the 1-σ band with the confidence interval corresponding to the 68% confidence level, and verifying that these two coincide. Even if the Monte Carlo method is more general, the Hessian method is useful because of the possibility of interpreting errors in terms of continuous parameter variations.

For this purpose, we are interested in the conversion from the initial Monte Carlo PDF set, to the Hessian one. This is made possible thanks to the `mc2hessian` code ([19]). We want to construct a multi-Gaussian covariance matrix in PDF space, where the central value of the final Hessian set coincides with the prior one. The Monte Carlo set can be viewed as N_{rep} PDFs sets:

$$\{f_\alpha^{(k)}(x_i,Q)\} \quad \begin{cases} k \in \{1,\dots,N_{rep}\} \\ \alpha \in \{1,\dots,N_f\} \\ i \in \{1\dots,N_{dat}\} \end{cases} \tag{26}$$

N_{rep} is the number of dataset replicas, N_f is the number of PDF flavors, and N_{dat} is the number of data points. Let be X the $N_{dat}N_f \times N_{rep}$ matrix with:

$$X_{l_{(\alpha,i)},k}(Q) \;=\; f_\alpha^{(k)}(x_i,Q) - f_\alpha^{(0)}(x_i,Q) \tag{27}$$

where $f_\alpha^{(0)}(x_i,Q)$ is the central value of the N_{rep} replicas of f_α evaluated in the x_i data point. The $l_{\alpha,i} = N_{dat}(\alpha-1)+i$, runs over the $x-$point and the flavors. The covariance matrix is then built as:

$$cov(Q) \;=\; \frac{1}{N_{rep}-1} X \cdot X^T \tag{28}$$

Let be $N_{rep} > N_{dat}\,N_f$, we can describe the eigenvectors of the sub-matrix $N_{dat}\,N_f \times N_{dat}\,N_f$ as linear combination of the N_{rep} replica. Using the *Singular Value Decomposition*, we can rewrite:

$$X \;=\; U\,\Sigma\,V^T \tag{29}$$

Where U and V are orthogonal matrices respectively: $N_{dat}\,N_f \times N_{dat}\,N_f$ and $N_{rep} \times N_{rep}$. The Σ is a diagonal and positive-definite matrix $N_{dat}\,N_f \times N_{rep}$. The values on the diagonal are the square roots of the eigenvalues of $X \cdot X^T$. By substituing:

$$X \cdot X^T \;=\; U\,(\Sigma^2)\,U^T \tag{30}$$

The columns of U are the eigenvectors of the covariance matrix. Let be $Z = U\Sigma$, then:

$$Z \cdot Z^T \;=\; X \cdot X^T \tag{31}$$

and:

$$Z \;=\; X \cdot V \tag{32}$$

The X matrix is quite large: $N_{eig} = N_{dat}N_f$. However, the eigenvectors with smallest eigenvalues give a negligible contribution, so we may consider just the $\tilde{N}_{eig} < N_{eig}$ largest eigenvalues. Now U and Σ are replaced by their submatrices u and σ with dimension respectively: $N_{dat}N_f \times \tilde{N}_{eig}$ and $\tilde{N}_{eig} \times N_{rep}$. This substitution leads to considering the principal submatrix P of V.
This allows to find the Hessian set as follow:

$$\tilde{f}_\alpha^{(k)}(x_i, Q) \;=\; f_\alpha^{(0)}(x_i, Q) + \frac{1}{N_{rep} - 1}\,(XP)_{l_{\alpha,i},k} \quad k \in \{1, \ldots \tilde{N}_{eig}\} \tag{33}$$

The uncertainties are given by:

$$\sigma_\alpha(x_i, Q) \;=\; \sqrt{\sum_{k=1}^{\tilde{N}_{eig}} \left(\tilde{f}_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q)\right)^2} \tag{34}$$

From a direct comparison between the initial PDFs Monte Carlo set and the Hessian conversion PDFs set, we can observe that using $\tilde{N}_{eig} = 100$ eigenvalues for the Hessian PDFs set, the lack of information is negligible, as shown below:

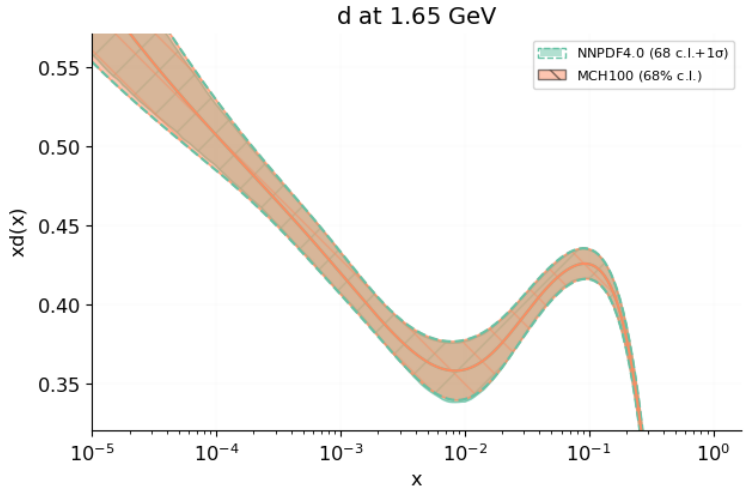


Figure 9: Comparison of the down quark PDF between the Monte Carlo `NNPDF4.0` set with 1000 replicas and the Hessian format with $\tilde{N}_{eig} = 100$, at $Q = 1.65GeV$.

3 Results

Now that the theoretical framework has been presented and the different computational methods employed in the fit have been introduced, we now focus on the main problem tackled in this work.

In order to understand the problem that we want to study, we perform a Hessian conversion with $\tilde{N}_{eig} = 100$ eigenvectors of a 1000 replicas Monte Carlo set of the global NNPDF4.0 fit. In Fig.10 we show the decomposition of the 100 largest eigenvalues of the $\Delta\chi^2$ computed from the Hessian PDF set. It can be seen that there are directions corresponding to negative eigenvalues $\Delta\chi^2 < 0$. This means that the central value of the PDFs does not correspond to the global minimum of χ^2.

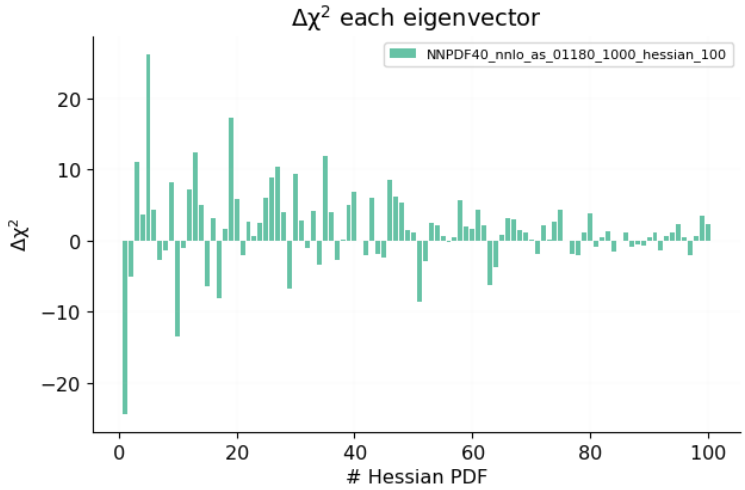


Figure 10: Decomposition of the 100 largest eigenvalues of the $\Delta\chi^2$ for the Hessian set converted from the 1000 replicas Monte Carlo set of the global NNPDF4.0 fit.

Actually, this might not be so surprising, in fact in section 2.5 we saw that cross-validation stops the minimization of the χ^2 for each replica to prevent the neural network from also learning the noise associated with the data. For this reason, negative directions could be associated with overlearning directions of the network. Despite these considerations, negative eigenvalues can have different origins. The aim of this thesis is precisely to study the cause of these negative directions.

3.1 Overlearning

In order to test whether the negative directions correspond to overlearning, we try to increase the amount of overfitting performed by the neural network and check the behavior of the eigenvalues after the conversion to the Hessian set. To this purpose, we change the settings of the cross-validation that controls overlearning, and in particular by varying the fraction of data in the training and validation sets. Moving data from the validation set to the training set, there are fewer and fewer constraints on the χ^2_{tr} minimization. In the limit of zero validation data, all the data are in the training set and the minimization algorithm continues to run until the maximum number of epochs is reached.

All the computations were carried out using the last version available of NNPDF parton distribution functions, the NNPDF4.0 with NNLO QCD calculations, and NLO electroweak corrections and nuclear uncertainties (for more details [8]). For a preliminary analysis, the 100 replicas Monte Carlo set of the global NNPDF4.0 fit was taken into account and the data fraction of the sets has been varied. It is possible to verify that the network had performed overfitting as an increase in the data fraction of training set resulted in a systematic reduction in χ^2, as shown below:

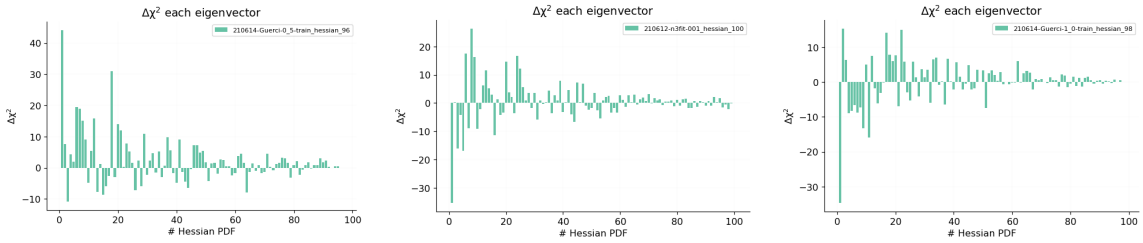
χ^2 of a 100 replicas Monte Carlo set of the global NNPDF4.0 fit		
Training data fraction: 50%	Training data fraction: 75%	Training data fraction: 100%
1.16702	1.16223	1.15571

By converting the three fits into the Hessian format and computing the $\Delta\chi^2$, in Fig.11 it can be seen that the negative directions not only do not decrease but rather increase.

This can also be observed through the eigenvalue distributions in Fig.12, in which there is a tendency to shift towards the negative eigenvalue region as the training fraction is increased.

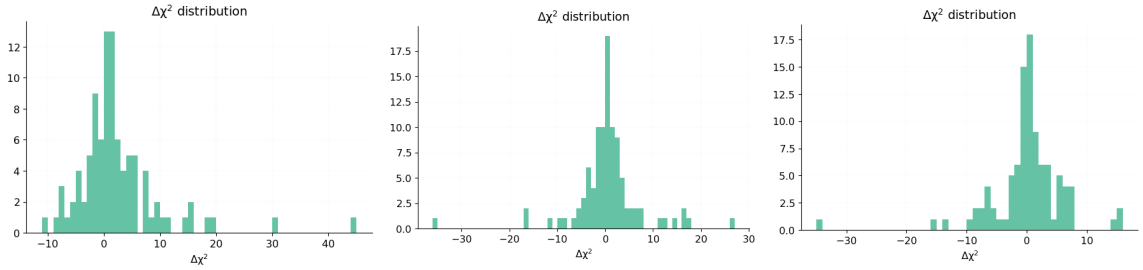
In order to check that this behavior is not due to a statistical fluctuation, it is necessary to move on to a more quantitative analysis. To do so, we change the dataset. We now consider the DIS-only (Deep Inelastic Scattering) dataset, which being smaller requires less computational time, allowing more fits to be made.

In addition, two other methods have been added to force overlearning. The first ex-



50% training data 75% training data 100% training data

Figure 11: Eigenvalue decomposition of $\Delta\chi^2$ for a Hessian set with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the global NNPDF4.0 fit.



50% training data 75% training data 100% training data

Figure 12: Eigenvalue distribution of $\Delta\chi^2$ for a Hessian set with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the global NNPDF4.0 fit.

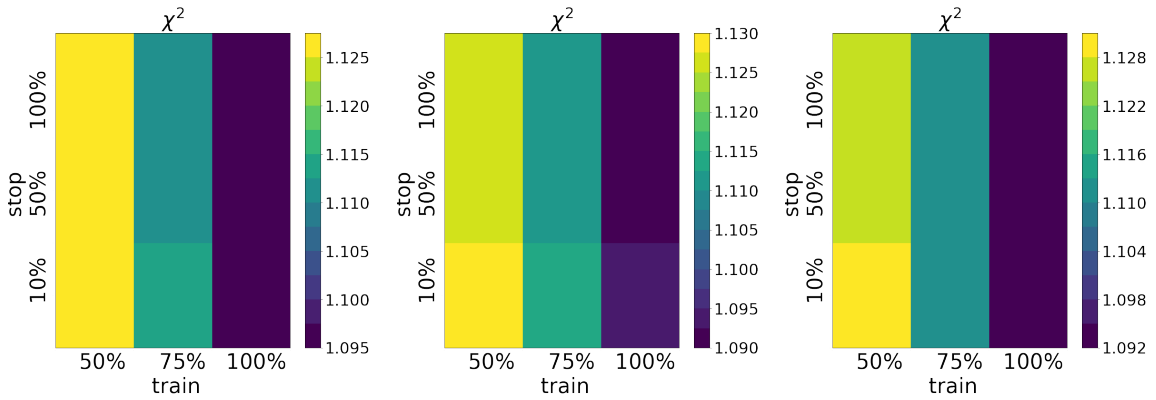
ploits the stopping algorithm of cross-validation by varying the parameter that determines after how many epochs the minimization must be stopped once the global minimum of the χ^2_{val} is reached. This because as the network continues to train, it increases the possibility that statistical fluctuations will cause the χ^2_{val} to end in a new minimum. The variable associated with the stop criterion will be called `stop` and when stop $=$ 100% the minimization reaches the maximum number of epochs. The last overfitting method directly affects the parameter optimization carried out by the Gradient Descent. Specifically, when we fix the value of the `clipnorm` we are setting the maximum value that the gradient can acquire in the parameter space. Increasing this leads to a more aggressive minimization, that can lead to local minima.

A systematic scan of the eigenvalues was carried out by combining these three overlearning methods. Each of these is associated with a variable and during the analysis. Each variable was sampled on three different values.

- training data fraction (**train**): $\{50\%\,;\,75\%\,;\,100\%\}$
- no stopping criterion (**stop**): $\{10\%\,;\,50\%\,;\,100\%\}$
- clipnorm (**clip**): $\{6.073\cdot10^{-6}\,;\,6.073\cdot10^{-3}\,;\,6.073\cdot10^{-1}\}$

All different combinations were tried, leading to 27 Monte Carlo sets with 100 replicas of the DIS-only NNPDF4.0 fit.

By studying the χ^2 of these fits, in Fig.13 it can be seen that the joint variation of `train` and `stop` with a fixed `clip` leads to a systematic decrease mainly dependent on the `train`, and to a lesser extent also dependent on `stop`. The `clip` dependency is even lower.



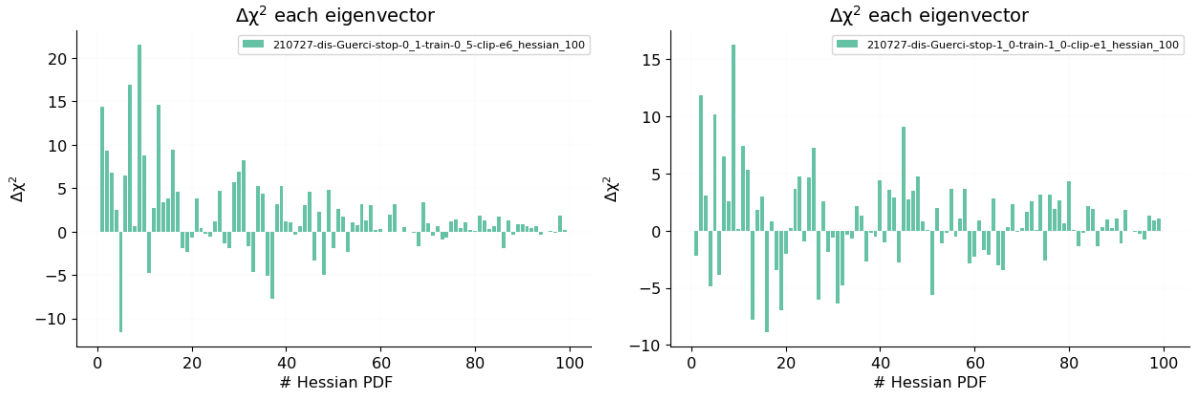
$clip = 6.073\cdot10^{-6}$ $clip = 6.073\cdot10^{-3}$ $clip = 6.073\cdot10^{-1}$

Figure 13: χ^2 grids of the 27 Monte Carlo sets with 100 replicas of the DIS-only NNPDF4.0 fit. Each grid contains 9 values of χ^2 with a fixed *clip*.

In order to check the eigenvalue behavior, the fits with the most and least overlearning are considered, i.e. respectively those with bigger and smaller parameter values:

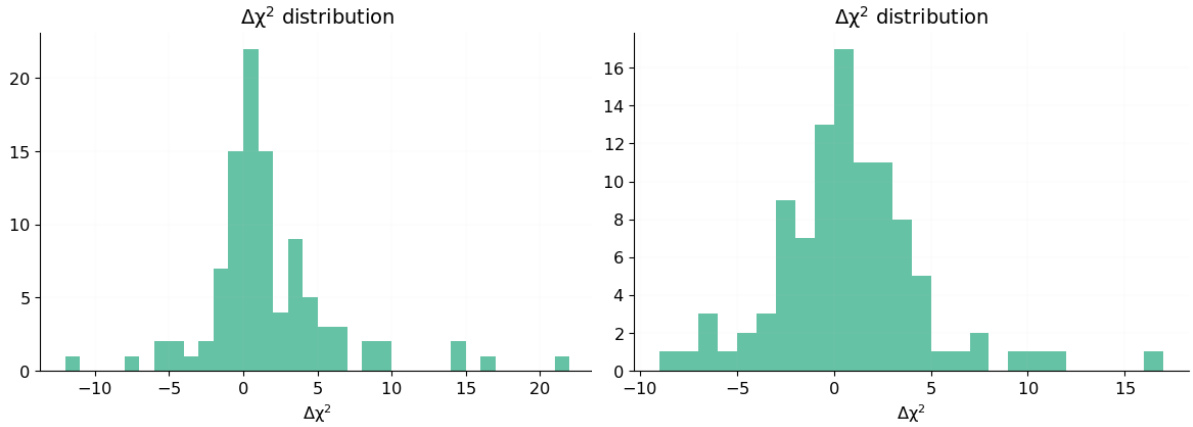
- (least overfitting) smaller parameter values:
  $\{\texttt{train}, \texttt{stop}, \texttt{clip}\} = \{50\%\,;\,10\%\,;\,6.073\cdot10^{-6}\}$
- (most overfitting) bigger parameter values:
  $\{\texttt{train}, \texttt{stop}, \texttt{clip}\} = \{100\%\,;\,100\%\,;\,6.073\cdot10^{-1}\}$

By converting the two fits into the Hessian format with 100 eigenvectors and computing the $\Delta\chi^2$, it is again observed in Fig.14 that the negative directions increase by intensifying overlearning.



(least overfitting) smaller parameter values (most overfitting) bigger parameter values

Figure 14: Eigenvalues decomposition of $\Delta\chi^2$ for a Hessian set with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the DIS-only NNPDF4.0 fit.



(least overfitting) smaller parameter values (most overfitting) bigger parameter values

Figure 15: Eigenvalues decomposition of $\Delta\chi^2$ for a Hessian set with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the DIS-only NNPDF4.0 fit.

As can be seen in Fig.15, the distribution still shifts towards the region of negative eigenvalues and is less scattered as expected, as we are getting closer and closer to the global minimum. This analysis allows us to conclude that the negative directions are not overlearning directions, in fact by forcing the overfitting of the neural network they increase.

3.2 Uncertainty distribution

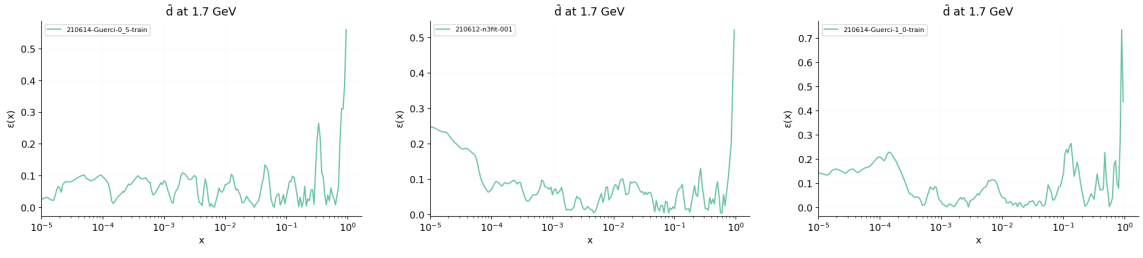
In order to understand the origin of the negative directions, it is possible to carry out a test of the hypothesis of Gaussianity of PDF probability distribution assumed by Hessian method. Thus, for the method to be usable, the PDFs must be Gaussianly distributed around the central limit in the PDF space. Although it is not possible to establish whether the distribution is Gaussian, it is still possible to check whether the necessary condition is met whereby the 1σ band around the central limit must correspond with the confidence interval associated with the 68% confidence level.

To compare these intervals we calculate the following percentage changes of 1-σ band with respect to the confidence interval:

$$\varepsilon(x) = \frac{|1\sigma(x) - 68\%c.i.(x)|}{68\%c.i.(x)} \tag{35}$$

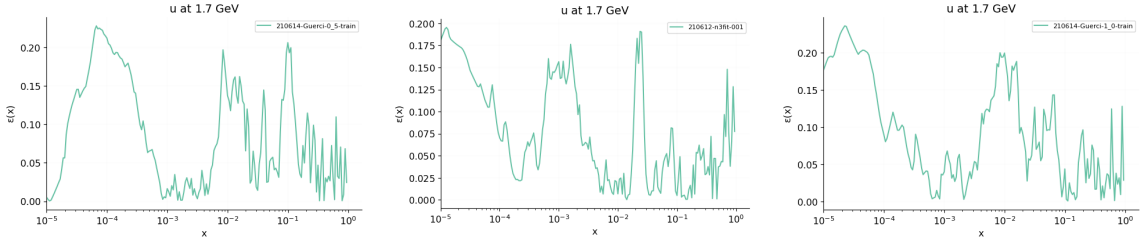
By carrying out these measurements on fits with and without overlearning, it is possible to determine whether the approximation of the Hessian method can be valid in the overfitting regime.

Following the same procedure as in the previous section, we initially consider the Hessian set with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the global NNPDF4.0 fit, varying the training data fraction:



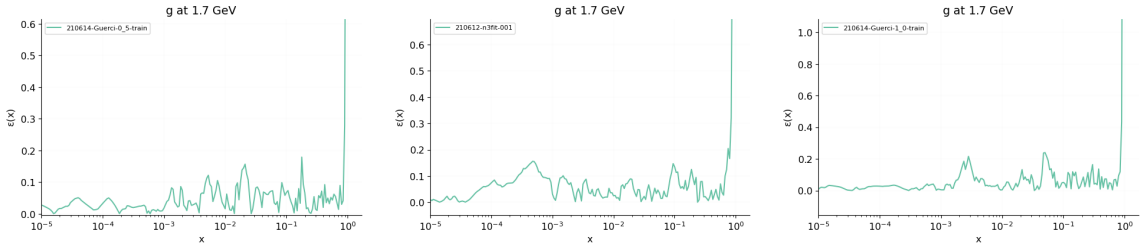
50% training data 75% training data 100% training data

Figure 16: Measurement of Gaussianity for the $\bar{d}$ Hessian PDF with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the global NNPDF4.0 fit.



50% training data 75% training data 100% training data

Figure 17: Measurement of Gaussianity for the $\bar{u}$ Hessian PDF with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the global NNPDF4.0 fit.



50% training data 75% training data 100% training data

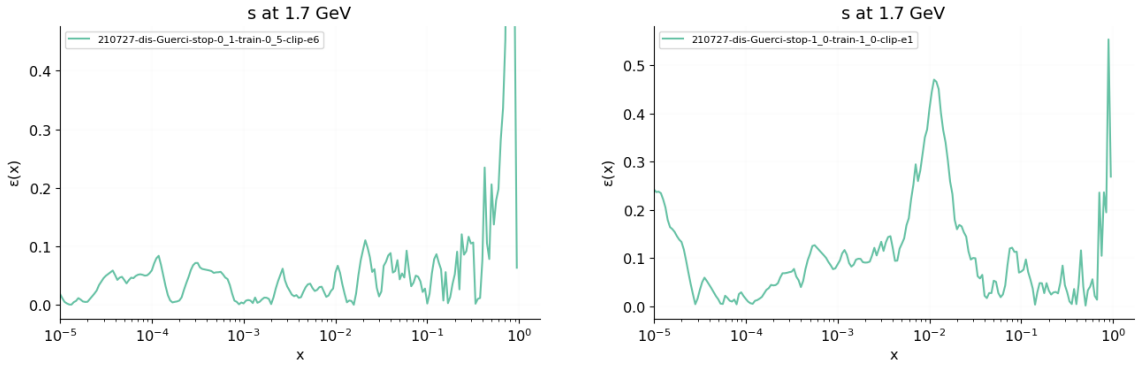
Figure 18: Measurement of Gaussianity for the g Hessian PDF with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the global NNPDF4.0 fit.

Some PDFs have a slight increase as the fraction of training data increases, such as the gluon PDF. Other PDFs have no significant variations and still others have variations that do not follow a pattern.

To try to capture more significant variations we again use the dataset composed of the DIS-only. We again consider the two fits that were found to be overfitting and non-overfitting in the previous section: the Hessian sets with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the DIS-only NNPDF4.0 fit, each with its own set of parameters $\{\texttt{train}, \texttt{stop}, \texttt{clip}\}$.

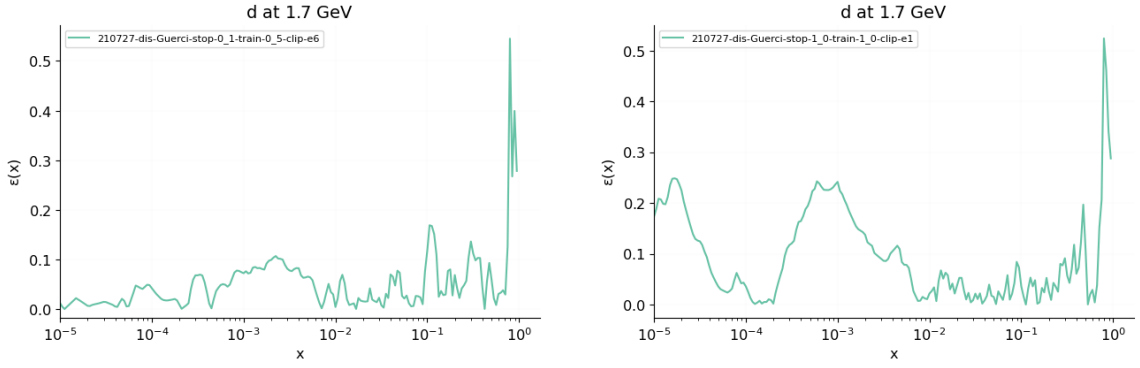
- (least overfitting) smaller parameter values:
  $\{\texttt{train}, \texttt{stop}, \texttt{clip}\} = \{50\% \,; 10\% \,; 6.073 \cdot 10^{-6}\}$
- (most overfitting) bigger parameter values:
  $\{\texttt{train}, \texttt{stop}, \texttt{clip}\} = \{100\% \,; 100\% \,; 6.073 \cdot 10^{-1}\}$

Comparing the two fits we can see that there are important deviations from the Gaussian distribution.



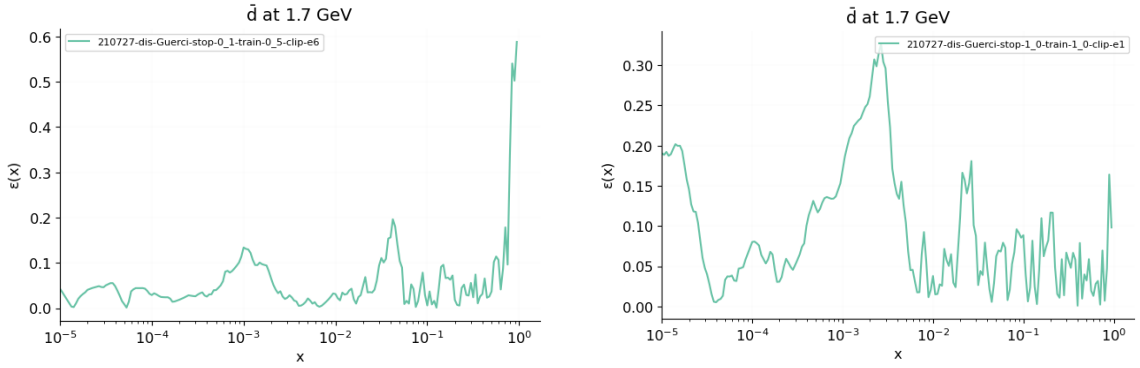
(least overfitting) smaller parameter values (most overfitting) bigger parameter values

Figure 19: Measurement of Gaussianity for the s Hessian PDF with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the DIS-only NNPDF4.0 fit.



(least overfitting) smaller parameter values (most overfitting) bigger parameter values

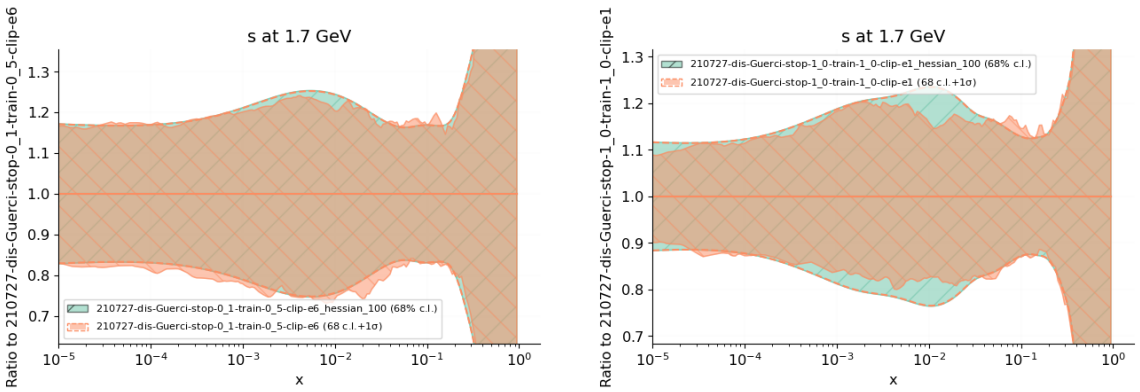
Figure 20: Measurement of Gaussianity for the d Hessian PDF with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the DIS-only NNPDF4.0 fit.



(least overfitting) smaller parameter values (most overfitting) bigger parameter values

Figure 21: Measurement of Gaussianity for the $\bar{d}$ Hessian PDF with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the DIS-only NNPDF4.0 fit.

These behaviours can also be seen by direct comparison of the uncertainty intervals corresponding to the 68% confidence level of the Monte Carlo PDF set (coinciding with 1σ) and the Hessian PDF set, as it is shown in Fig.22:



(least overfitting) smaller parameter values (most overfitting) bigger parameter values

Figure 22: Comparison between the 1σ band with the confidence interval associated to the 68% confidence level for the s Hessian PDF with 100 eigenvectors converted from a 100 replicas Monte Carlo set of the DIS-only NNPDF4.0 fit.

It can be seen that in the case of overlearning we have a significant loss of Gaussianity.

4 Conclusions

The problem of the existence of negative eigenvalues of $\Delta\chi^2$ was addressed by focusing on two distinct observations. The increase in the number of negative directions with the intensification of overfitting has made it possible to deduce that the existence of negative variations of the χ^2 is not caused by a lack of efficiency in the PDF fitting method. If this had been the case, then optimizing the minimization would have led to ever smaller values of χ^2, so it would have been possible at the limit to approach the global minimum, corresponding to the best fit in the PDFs space, and at that point, it would no longer be possible to have such eigenvalues.

To describe this behavior, the Gaussianity hypothesis of the probability distributions of PDFs, on which the Hessian method is based, was studied. By comparing the 1σ band with the confidence interval associated to the confidence level of 68% we saw a significant loss of Gaussianity with increasing the overfitting of the neural network.

These deviations allowed us to conclude that the presence of negative eigenvalues of the matrix $\Delta\chi^2$ was due to the forcing of the probability distribution of PDFs in an inappropriate Gaussian pattern. Thus, it is impossible to match the PDF best fit with the global minimum of χ^2 and we find incorrect negative directions. When we intensify the overfitting of the neural network, we are just increasing the discrepancies between the two distributions. That is, the point of PDF best fit moves away from the global minimum of χ^2, reaching a random point of the space. This point is completely independent of the global minimum predicted by the Gaussian trend, and here the sign of the variations along the various directions becomes a stochastic variable, so the negative directions increase.

References

[1] Gell-Mann, M. (1964). *A Schematic Model of Baryons and Mesons*. Phys. Lett., 8:214–215.

[2] Zweig, G. (1964). *An SU(3) model for strong interaction symmetry and its breaking*. Version 2. In Lichtenberg, D. and Rosen, S. P., editors, DEVELOPMENTS IN THE QUARK THEORY OF HADRONS. VOL. 1. 1964 - 1978, pages 22–101.

[3] Bloom, E. D. et al. (1969). High-Energy Inelastic e p Scattering at 6- Degrees and 10-Degrees. Phys. Rev. Lett., 23:930–934.

[4] Gross, D. J. and Wilczek, F. (1973). Ultraviolet Behavior of Non- abelian Gauge Theories. Phys. Rev. Lett., 30:1343–1346. [,271(1973)].

[5] Politzer, H. D. (1973). Reliable Perturbative Results for Strong Interactions? Phys. Rev. Lett., 30:1346–1349. [,274(1973)].

[6] G. D. Coughlan, J. E. Dodd and B. M. Gripaios. *The Ideas of Particle Physics*. Cambridge University Press, 2006.

[7] Brandelik, R. et al. (1979). Evidence for Planar Events in e+ e- Anni- hilation at High-Energies. Phys. Lett., 86B:243–249.

[8] Ball, R. et al. (2021). The Path to Proton Structure at One-Percent Accuracy. arXiv: https://arxiv.org/abs/2109.02653

[9] S. Forte, S. Carrazza (2020). Parton distribution functions. arXiv: https://arxiv.org/abs/2008.12305

[10] Ball, R. et al. (2021). An open-source machine learning framework for global analyses of parton distributions. arXiv: https://arxiv.org/abs/2109.02671

[11] 21. D. Stump, J. Pumplin, R. Brock, D. Casey, J. Huston, J. Kalk, H. Lai, and W. Tung, Uncertainties of predictions from parton distribution functions. 1. The Lagrange multiplier method, Phys. Rev. D 65 (2001) 014012, arXiv:hep-ph/0101051

[12] A. D. Martin, R. G. Roberts, W. J. Stirling, and R. S. Thorne, Uncertainties of predictions from parton distributions. I: Experimental errors. ((T)), Eur. Phys. J. C28 (2003) 455, arXiv:hep-ph/0211080.

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

[14] M. D. Zeiler, ADADELTA: An Adaptive Learning Rate Method, arXiv:1212.5701 [cs.LG].

[15] X. Glorot and Y. Bengio, "understanding the difficulty of training deep feedforward neural networks", in In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics. 2010.

[16] Y. Bengio and X. Glorot, "understanding the difficulty of training deep feed forward neural networks", International Conference on Artificial Intelligence and Statistics (01, 2010) 249.

[17] 41. J. Bergstra, D. Yamins, and D. D. Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures, in Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ICML'13. JMLR.org, 2013. http://dl.acm.org/citation.cfm?id=3042817.3042832.

[18] J. Pumplin et al. Uncertainties of predictions from parton distribution functions. II. The Hessian method. Physical Review D, 65(1), Dec 2001.

[19] Stefano Carrazza, Stefano Forte, Zahari Kassabov, José Ignacio Latorre, and Juan Rojo. An unbiased Hessian representation for Monte Carlo PDFs. The European Physical Journal C, 75(8), Aug 2015.

[20] Nicola Lambri. Optimized regression models for parton distribution functions determination using deep learning methods. Master's thesis, University of Milan, 2020.