Università degli Studi di Milano

FACOLTÀ DI SCIENZE E TECNOLOGIE

Corso di Laurea Magistrale in Fisica

Studies of parametrization dependence and bias in the determination of parton distributions

Relatore:
Prof. Stefano Forte

Correlatori:
Dr. Stefano Carrazza
Zahari Kassabov

Tesi di Laurea di:
Carlo Mascaretti
Matr. 845347
Codice P.A.C.S.: 12.38.-t

Anno Accademico 2015-2016

Abstract

The purpose of this thesis is to investigate whether or not the choice of a specific functional form for PDFs (Parton Distribution Functions) biases the uncertainties resulting from their fits to experimental datasets. This task is performed exploiting the closure testing procedure, developed by the NNPDF group.

In the first chapter I briefly review the state of the art of PDFs, introducing them and trying to explain why they are important to determine along with the corresponding statistical uncertainties. A summary of the methodologies adopted by two specific groups, that is MSTW08 and NNPDF, follows and completes the introductory part.

In the second chapter I address the core of the problem, i.e. tolerance: that's the name given to the factor by which one has to increase the $\Delta\chi^2$ resulting from fits to experimental data in order to recover reliable confidence intervals. A brief review on possible explanations of this fine–tuning is then provided; the closure testing procedure is consequently dealt with, together with an explanation on how it provides us with a controlled environment to test the eventual parametrization bias.

Finally, in the third chapter I present the results of this procedure: using the MSTW08 parametrization in the context of closure testing, I compute statistical uncertainties on PDFs, compare them to the values present in literature and interpret the outcome.

Contents

Chapter 1

Parton distribution functions

"The question was, once we had this very complex quark theory,
why did Feynman's simple model still work so well?
No one has ever been able to figure out why.
In fact, the problem became so difficult that people just gave up"
K. G. Wilson

Parton distribution functions (PDFs) are needed to compute cross–sections of every process in which the internal structure of the hadron is relevant: I find it appropriate to introduce them, and discuss the method they are extracted from the data and their theoretical uncertainties are computed.

1.1 Introduction and definition

The first observation of the internal structure of nucleons was made possible by Deep Inelastic Scattering (DIS) experiments, in which leptons are scattered off nucleons: these experiments resulted in the discovery of more fundamental point–like particles, named partons. If the momentum transferred in the collision is high, the momentum of the parton interacting with the lepton is almost collinear to that of the nucleon, so that the target can be interpreted as a stream of partons, each carrying a fraction x of the longitudinal momentum. A PDF $f_i(x, Q^2)$ is the probability density for a parton of flavour i to carry a fraction x of the nucleon momentum at a squared energy scale Q^2. At high transferred momentum, there is an increasing number of quark–antiquark pairs carrying a low momentum fractions x: the sea quarks. A quite surprising discovery of DIS experiment is that quarks and antiquarks only carry about half of the total nucleon momentum, the remainder being carried by gluons.

PDFs take part in the computation of the cross section for a "hard" (i.e. highly–energetic and such that a perturbative description is allowed) process involving at least one hadron via the factorization formula [1]. Considering a generic inclusive hadroproduction process which depends on a single scale M_X^2, the statement above translates into:

$$\sigma(s, M_X^2) = \sum_{a,b} \int_{x_{\min}}^{1} \mathrm{d}x_1\, \mathrm{d}x_2\, f_{a/h_1}(x_1, M_X^2)\, f_{b/h_2}(x_2, M_X^2)\, \hat{\sigma}_{ab\to X}(x_1 x_2 s, M_X^2) \tag{1.1.1}$$

where s is the center–of–mass energy of the hadronic collision, $f_{a/h_i}(x_i, M_X^2)$ is the distribution of partons of type a in the i–th incoming hadron, $\hat{\sigma}_{ab\to X}$ is the partonic cross section for the production of the final state X. Partonic cross sections are computed in QCD using perturbation theory: they do not depend either on the incoming hadron or on PDFs. The formula (1.1.1) can be rewritten introducing

$\tau \equiv x_{\min} \leq x_i$:

$$\sigma(s, M_X^2) = \sum_{a,b} \sigma_{ab}^0 \int_\tau^1 \frac{\mathrm{d}x_1}{x_1} \int_{\tau/x_1}^1 \frac{\mathrm{d}x_2}{x_2} f_{a/h_1}(x_1, M_X^2)\, f_{b/h_2}(x_2, M_X^2)\, C_{ab}\left(\frac{\tau}{x_1 x_2}, \alpha_S(M_X^2)\right) \tag{1.1.2}$$

The hard coefficient function $C_{ab}(z, \alpha_S(M_X^2))$ depends on the scale M_X^2 and on the ratio of this scale to $\hat{s}$, that is:

$$z = \frac{M_X^2}{\hat{s}} = \frac{\tau}{x_1 x_2} \tag{1.1.3}$$

We have extracted a prefactor σ_{ab}^0 so that at leading perturbative order we have that:

$$\hat{\sigma}_{ab\to X} = \sigma_0\, C_{ab}(z, \alpha_S(M_X^2)) \qquad\qquad C_{ab}(z, \alpha_S(M_X^2)) = c_{ab}\delta(1-z) + \mathcal{O}(\alpha_S)$$

where c_{ab} depends on the specific process. The equation above implies that at leading order

$$\tau_{\mathrm{LO}} = x_1 x_2 \tag{1.1.4}$$

Equation (1.1.1) expresses the inclusive (i.e. lacking full information on the final state identity) cross section in terms of PDFs at the same scale M_X^2 at which the cross section is evaluated. PDFs at different scales are related by the DGLAP [2, 3, 4] integro–differential equations: their accuracy is thus limited by the perturbative order both of the coefficient function C_{ab} in formula (1.1.2) and of the splitting functions which enter the evolution equations. We say that PDFs are computed, as an example, at NLO if C_{ab} and the splitting functions are computed at least at NLO.

1.2 Determination of PDFs

Since parton distributions cannot be computed from first principles yet, they need to be determined by comparing experimental data with theoretical predictions which rely on the PDFs themselves, as shown above. This process is currently performed by several groups [5, 6]:

- MSTW [7];
- NNPDF [8];
- CT10 [9];
- HERAPDF;
- ABM;
- GJR.

NNPDF and MSTW are representatives of two different methods of determinating PDFs, Monte Carlo and Hessian respectively: for this reason I am going to review their approaches and differences.

Parton distributions are determined by comparing factorized expressions such as (1.1.1) with experimental data. Also in the case of PDFs extrapolation a measure of goodness–of–fit is needed to assign confidence intervals in the space of PDFs: this is a nontrivial (if not ill–posed) problem, since it involves defining a probability measure on a space of functions [10]. A determination of parton distributions consists in extrapolating at least seven[1] independent functions, i.e. three light quark and antiquark distribution

[1] There are in principle thirteen independent PDFs in a given hadron (six quarks and antiquarks and the gluon); however, in practice, charm and heavier quark PDFs in the nucleon are not independently determined in all current PDF sets, and are instead assumed only to be generated by QCD radiation. The (moderate) impact of introducing an independent (non–perturbative) charm PDF, so that charm does not vanish below the threshold for its radiation ("intrinsic" charm [11]) has been studied in references [12, 13].

and the gluon at some initial scale, from which PDFs at all other scales can be obtained using DGLAP equations [15]. The time–honoured [16, 17] method to address this problem is to project the infinite dimensional problem onto a finite dimensional parameter space by choosing a specific functional form for parton distributions. This is justified from the expectation that PDFs are smooth functions of x: since $0 \leq x \leq 1$, a finite–accuracy representation of PDFs must be possible on a finite basis of functions, i.e. with a finite number of parameters. The problem is now how to find the optimal parametrization, that is the one that, for given accuracy, minimizes the number of parameters without introducing a bias. Whatever the parametrization, a measure of goodness–of–fit is needed, may it be a χ^2 function or a likelihood function (for more details see [14]). Existing sets of PDFs are determined and delivered following two main strategies:

- a "Hessian" approach, in which the best fit result consists in the optimal set of parameters and in an error matrix centered on this optimal fit to compute uncertainties;
- a Monte Carlo approach, in which the best fit is obtained from the Monte Carlo sample by averaging and uncertainties are obtained as variances of the sample.

Available Hessian PDF sets are mostly based on a "standard" parametrization, suggested by various QCD arguments. The available full Monte Carlo PDF sets are based on a rather different form of parametrization, which adopts neural networks as interpolating functions in an attempt to reduce the bias related to the choice of functional form [18]. However, Monte Carlo studies based on other standard [19] and non-standard [20] parametrizations have also been presented. Here will be summarized the main features of both the Hessian and Monte Carlo approach: in each case I will discuss the parton parametrization which is most commonly used with each approach, and the way the best fit is determined in each case – which in turn requires a peculiar algorithm within the neural network approach.

In order to obtain meaningful predictions from the factorized QCD expressions like (1.1.1), the accurate determination of PDF uncertainties is necessary. PDFs are determined by comparing QCD predictions to the data: this immediatly implies that any uncertainty in the theory used to obtain these predictions will propagate onto the PDFs themselves. Such uncertainties include:

1. genuine theoretical uncertainties, such as lack of knowledge of higher–order perturbative corrections: these, generally, do not have a simple statistical interpretation (and in particular they are generally not gaussian);
2. lack of knowledge of parameters in the theory, in particular the value of the strong coupling constant α_S and the heavy quark masses m_c and m_b, which generally do follow gaussian statistics;
3. uncertainties which are related to the way the information contained in the data is propagated onto a PDF, which we will discuss in the following section.

I will refer to the first two kinds as "theoretical uncertainties" and to the third as "PDF uncertainties". The treatment of theoretical uncertainties is in principle straightforward, in the sense that all one has to do is propagate them onto the PDFs – their effect on PDFs is no different from their effect on the calculation of a physical observable, and PDFs do not entail any new problem. For example, if it is agreed that higher order corrections on cross sections can be conventionally estimated by varying renormalization and factorization scales in a certain range to be interpreted, say, as a 90% confidence level with flat distribution, the associate PDF uncertainty is simply found by repeating the PDF determination while performing this variation. The treatment of PDF uncertainties is, as before said, far less trivial: here I briefly review how this is performed following the two aforementioned approaches.

1.2.1 Hessian uncertainties and the MSTW08 approach

The standard approach to PDF determination consists in, at some reference scale Q_0, shaping the functional form of PDFs as to statisfy counting rules [21], which suggest a behaviour $f(x) \to (1-x)^\beta$ for $x \to 1$, and Regge theory [22], which suggests that $f(x) \to x^\alpha$ for $x \to 0$. These limiting behaviours are

only approximate, since perturbative evolution will correct them by logarithmic terms at other scales: that's why, even though theoretical grounds constrain the β and α exponents, they are taken as free fit parameters. The form of the typical PDF is thus assumed to be:

$$f_i(x, Q_0^2) = x^{\alpha_i}(1-x)^{\beta_i} g_i(x) \tag{1.2.1}$$

where $g_i(x)$ tends to a constant in the limits $x \to 0$ and $x \to 1$. Different choices are possible for the form of $g_i(x)$: the MSTW08 collaboration assume

$$g_i(x) = 1 + \epsilon_i \sqrt{x} + \gamma_i x + \delta_i x^2 \tag{1.2.2}$$

Given their relevance with this work, I find it meaningful to report here all the PDF parametrizations they use [11]:

$$xu_v(x, Q_0^2) = A_u\, x^{\eta_1}(1-x)^{\eta_2}(1 + \epsilon_u \sqrt{x} + \gamma_u\, x) \tag{1.2.3}$$
$$xd_v(x, Q_0^2) = A_d\, x^{\eta_3}(1-x)^{\eta_4}(1 + \epsilon_d \sqrt{x} + \gamma_d\, x) \tag{1.2.4}$$
$$xS(x, Q_0^2) = A_S\, x^{\delta_S}(1-x)^{\eta_S}(1 + \epsilon_S \sqrt{x} + \gamma_S\, x) \tag{1.2.5}$$
$$x\Delta(x, Q_0^2) = A_\Delta\, x^{\eta_\Delta}(1-x)^{\eta_S+2}(1 + \gamma_\Delta\, x + \delta_\Delta\, x^2) \tag{1.2.6}$$
$$xg(x, Q_0^2) = A_g\, x^{\delta_g}(1-x)^{\eta_g}(1 + \epsilon_g \sqrt{x} + \gamma_g\, x) + A_{g'}\, x^{\delta_{g'}}(1-x)^{\eta_{g'}} \tag{1.2.7}$$
$$x(s+\bar{s})(x, Q_0^2) = A_+\, x^{\delta_S}\,(1-x)^{\eta_+}(1 + \epsilon_S \sqrt{x} + \gamma_S\, x) \tag{1.2.8}$$
$$x(s-\bar{s})(x, Q_0^2) = A_-\, x^{\delta_-}(1-x)^{\eta_-}(1 - x/x_0) \tag{1.2.9}$$

where $Q_0 = 1\,\mathrm{GeV}$ and

$$q_v \equiv q - \overline{q} \qquad \Delta \equiv \overline{d} - \overline{u} \qquad S \equiv 2(\overline{u} + \overline{d}) + s + \overline{s}$$

The input PDFs listed in eqs. (1.2.3)–(1.2.9) are subject to three constraints from number sum rules:

$$\int_0^1 \mathrm{d}x\, u_v(x, Q_0^2) = 2 \qquad \int_0^1 \mathrm{d}x\, d_v(x, Q_0^2) = 1 \qquad \int_0^1 \mathrm{d}x\, s_v(x, Q_0^2) = 0$$

together with the momentum sum rule:

$$\int_0^1 \mathrm{d}x\, x\left[u_v(x, Q_0^2) + d_v(x, Q_0^2) + S(x, Q_0^2) + g(x, Q_0^2)\right] = 1$$

These sum rules determine A_g, A_u, A_d and x_0 in terms of the other parameters, so that there are potentially $34 - 4 = 30$ free PDF parameters in the fit, including α_S. In practice, there are only 28 of them: δ_- is fixed to 0.2 in (1.2.9) due to extreme correlation with A_- and η_- and the g' polynomial in (1.2.7) is omitted in LO fits due to the positive behaviour of the input gluon distribution at small x.

Now that a parametrization of PDFs is given, the problem is to determine best fit values and uncertainty ranges for the parameters $\mathbf{a}$: in a Hessian approach, this is done by minimizing a figure of merit such as

$$\chi^2(\mathbf{a}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}(D_i - T_i(\mathbf{a}))(V^{-1})_{ij}(D_j - T_j(\mathbf{a})) \tag{1.2.10}$$

where $i, j = 1, \ldots, N =$ number of the D_i data points, characterized by an experimental covariance matrix V_{ij}. T_i are the theoretical predictions, which are obtained by evolving the starting PDFs at any scale Q^2 using the DGLAP equations and then folding the result with known partonic cross sections according to the factorization theorem, such as in eq. (1.1.1). The χ^2 is thus a function of the $\mathbf{a}$ parameters through the predictions $T_i(\mathbf{a})$. It should be noted that the χ^2 in eq. (1.2.10) is normalized to the number of data

points, i. e. it is likely to be close to the χ^2 per degree of freedom as typical datasets consist of thousands of data, while the total number of parameters needed to describe accurately all PDFs with functional forms like (1.2.1) is likely to be rather lower than 100. An unnormalized χ^2 can be introduced as well:

$$\overline{\chi}^2 \equiv N\chi^2 \tag{1.2.11}$$

It should be noted that, since slightly different definitions of the χ^2 are used, the comparison of χ^2 values from different groups are only qualitatively significant: the main subtlety is related to the inclusion of normalization uncertainties, which cannot be simply introduced in the covariance matrix, as this would bias the fit [23].

Once the χ^2 is defined, the best fit set of parameters is identified with the absolute minimum of the χ^2 in parameter space. The variance of any observable f which depends on $\mathbf{a}$ (such as physical cross sections, or the PDFs themselves) is given by, assuming linear error propagation $f(\mathbf{a}) \approx f_0 + \mathbf{a} \cdot \nabla f|_0$:

$$\sigma_f^2 = \sigma_{ij}\partial_i f \partial_j f \tag{1.2.12}$$

where σ_{ij} is the covariance matrix for the parameters, given –assuming the χ^2 to be quadratic in the parameters around the minimum– by

$$\sigma_{ij} = \partial_i \partial_j \overline{\chi}^2\big|_{\min} \tag{1.2.13}$$

i.e. it is the matrix of second derivatives (Hessian) of the unnormalized χ^2, evaluated at its minimum. As explained in [14], the 68% confidence level (one–σ) interval for the parameters is the set in parameter space contained by the ellipsoid

$$\overline{\chi}^2 = \overline{\chi}^2_{\min} + 1 \tag{1.2.14}$$

In more realistic cases, as I will discuss in chapter 2, this argument may have to be modified, as to account for various effects (such as incorrect estimation of the covariance matrix of the data). For now, however, let us stick to the ideal case and make a couple of statements. The first is that we are free to make all the eigenvalues of σ_{ij} equal to one by adjusting the parametrization: this is simply done by diagonalizing the matrix and rescaling the eigenvectors by their corresponding eigenvalues, that is looking for new parameters $\mathbf{a}' = \mathbf{a}'(\mathbf{a})$ such that

$$\sigma_{ij}(a_i - a_i^0)(a_j - a_j^0) = \sum_{i=1}^{N} (a_i')^2 \tag{1.2.15}$$

which implies that eq. (1.2.12) becomes

$$\sigma_f^2 = |\nabla' f|^2 \tag{1.2.16}$$

where $\partial_i' \equiv \partial/\partial a_i'$. This means that the total contribution, being the length of a vector, to the uncertainty due to two different sources is simply found by adding the components (i.e. the different uncertainties) in quadrature, even when correlation is present (see [24] for more insight into this topic).
The second observations involves the statement that the one–σ interval is contained in the contour given by eq. (1.2.14). This must not be confused with the fact that the variance of a chi–squared distribution with k degrees of freedom is $2k$, which just means that, upon repetition of the experiment, the χ^2 (the $\overline{\chi}^2$ in our notation) is expected to be in the interval

$$[\chi^2_{\min} - \sqrt{2k}, \chi^2_{\min} + \sqrt{2k}]$$

representing thus a hypothersis–testing criterion [25]. The statement that $\Delta\overline{\chi}^2 = 1$ provides a parameter–fitting criterion, i.e. the range of parameter values which are compatible at one sigma for a given experimental result and theory.

1.2.2 Monte Carlo uncertainties and the NNPDF approach

The difference between the Monte Carlo approach and the Hessian one consists in the method the uncertainties on the observables are propagated from those in parameter space. A Monte Carlo propagation of uncertainties is convenient when dealing with a less manageable parametrization, for instance because of the complex functional form or the large number of parameters or, in general, whenever linear error propagation and the quadratic approximation to the χ^2 in parameter space are not advisable. As before, I'm going to briefly review the main aspects of the Monte Carlo approach and of the NNPDF method to determine uncertainties when using it.

If in a Hessian approach parameter are assumed to be gaussianly distributed with covariance matrix σ_{ij} given by the Hessian in eq. (1.2.13), in a Monte Carlo approach the probability distribution in parameter space is given by assigning a Monte Carlo sample of replicas.
Considering the example of the parametrization in eq. (1.2.1), one would simply give a list of $N_{\rm rep}$ replica copies of the parameters $\mathbf{a}$. Any observable f is then evaluated for all these parameter replicas: its central value is the average of these $N_{\rm rep}$ functions, and its standard deviation is the (square root of the) variance, and so on for all the moments of the probability distribution.
The problem is, now, how to sample the replica distribution in parameters space. The maximum likelihood method provides the answer, as it gives a way of mapping the probability distribution in data space onto the probability distribution in parameters space, as follows. Let D_i be the data with covariance matrix V_{ij} and D_i^k the $N_{\rm rep}$ ($k = 1, \ldots, N_{\rm rep}$) replicas, such that

$$\lim_{N_{\rm rep}\to\infty} \sum_{k=1}^{N_{\rm rep}} \frac{D_i^k}{N_{\rm rep}} = \lim_{N_{\rm rep}\to\infty} \overline{D}_i = D_i \qquad \lim_{N_{\rm rep}\to\infty} \sum_{k,l}^{N_{\rm rep}} \frac{(\overline{D}_i - D_i^k)(\overline{D}_j - D_j^l)}{N_{\rm rep} - 1} = V_{ij}$$

that is, if $N_{\rm rep} \to \infty$ the average of the replica sample D_i^k reproduces the initial data D_i and the covariance of the replica samples D_i^k and D_j^l is equal to the covariance matrix element V_{ij}.
Now, a best–fit parameter vector $\mathbf{a}^k$ can be determined, for each data replica, by minimizing the χ^2 (1.2.10) of the fit to the replica data D_i^k. We thus end up with a Monte Carlo set of best–fit parameter vectors $\mathbf{a}^k$, from which one can extract, by averaging over k, the best fit parameter vector $\mathbf{a}$ and, by computing the covariance of the i–th and j–th components of $\mathbf{a}^k$, the covariance matrix σ_{ij}.
It should be mentioned that within a Monte Carlo approach it is possible to sidestep the problem of choosing an adequate parametrization by using Bayesian inference. One starts from some prior Monte Carlo representation of the probability distribution based on some initial subset of data, or even on assumptions. The initial Monte Carlo set is then updated by including the information contained in new data, that is by changing the distribution of replicas: more or fewer copies are taken for those replicas which agree or respectively do not agree with the new data, in a way which is specified by Bayes' theorem. This applies to the extent that results do not depend too much on the choice of prior, which is often the case if the information used through Bayes' theorem is sufficiently abundant, and are then free of bias. Whereas the construction of a parton set fully based on this method has so far not been completed, preliminary results have been presented [27] on the inclusion of new data in an existing Monte Carlo fit using this methodology.
For a typical set of data used in a parton fit the numbers of replicas required turn out to be surprisingly small: for instance, the figure 1.1 represents a scatter plot of the averages vs. central values and variances vs. standard deviations for the set of 3372 data points included in the NNPDF1.2 [28] parton fit, computed using $N_{\rm rep} = 10$, 100, 1000 Monte Carlo Replicas. It is clear that the scatter plot deviates by just a few percent from a straight line already for $N_{\rm rep} = 10$ for central values, and for $N_{\rm rep} = 100$; $N_{\rm rep} = 1000$ replicas turn out to be only necessary in order to get percent accuracy on correlation coefficients.

The NNPDF group use the Monte Carlo approach in conjunction with neural network as a parton parametrization. Neural networks, like polynomials, have the feature that any (with suitable assumptions of continuity) function may be fitted in the limit of infinite number of parameters; unlike polynomials,

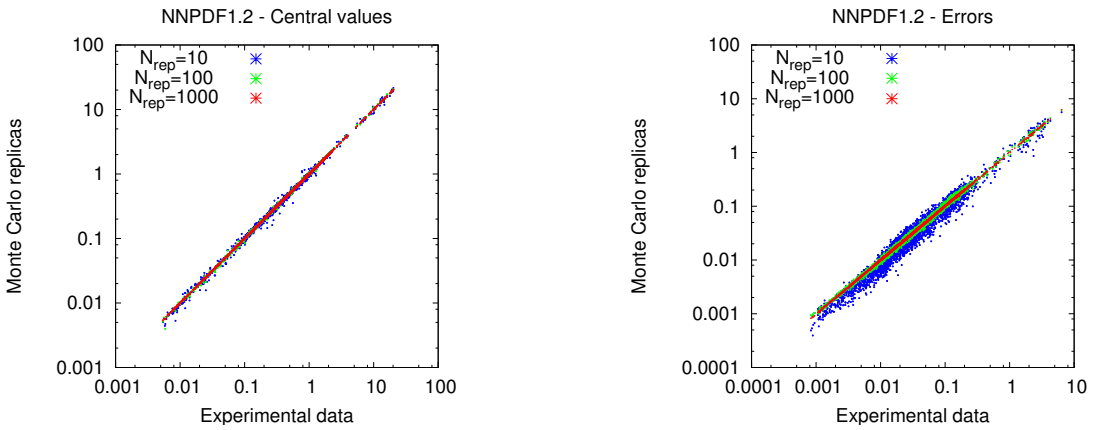

Figure 1.1: scatter plot of central values and uncertainties of a Monte Carlo sample compared to the data for the dataset of the NNPDF1.2 parton fit [28].

they are non–linear and "unbiased", that is, a finite–dimensional truncation of a neural network can fit a wide class of functions (both periodic and non periodic, for instance) without tuning the form of the parametrization to the considered problem. Each basis PDF at the reference scale Q_0 is parametrized in terms of a neural network (specifically a multi–layer feed–forward perceptron) times a preprocessing factor, whose form is chosen on the same theoretical grounds seen before:

$$f_i(x, Q_0) = A_i \hat{f}_i(x, Q_0) \qquad \hat{f}_i(x, Q_0) = x^{-\alpha_i}(1-x)^{\beta_i}\mathrm{NN}_i(x) \tag{1.2.17}$$

A_i is an overall normalization constant and the preprocessing term is there only to speed up the minimzation, without biasing the fit: for more details on preprocessing exponents and normalization constants, and on what I'm going to review hereafter, see [31]. An example of neural network is the following:

$$\mathrm{NN}(x) = \frac{1}{1 + e^{\theta_1^{(3)} - \frac{\omega_{11}^{(2)}}{1+e^{\theta_1^{(2)} - x\omega_{11}^{(1)}}} - \frac{\omega_{12}^{(2)}}{1+e^{\theta_2^{(2)} - x\omega_{21}^{(1)}}}}} \tag{1.2.18}$$

with $\theta_n^{(i)}$ and $\omega_{nm}^{(i)}$ free parameters – six, in this case. This is a 1–2–1 neural network: this refers to the way it is built (architecture), by iterating recursively the response function (sigmoid)

$$g(x) = \frac{1}{1 + e^{\theta - \beta x}}$$

on nodes arranged in layers which feed forward to the next layer, with the first (last) layer containing the input (output) variables. In refs. [10, 27, 28] the neural networks used to parametrize the PDFs have a 2–5–3–1 architecture, with 37 free parameters (the input has two variables x and $\ln x$, which are treated as two independent inputs to deacrease the redundancy of the parametrization). The six light flavours and antiflavours and the gluon are parametrized in this way, so that the total number of parameters is $37 \times 7 = 259$, rather larger than the numbers relative to parametrizations like (1.2.1). Even though so many free parameters reduce the risk of a parametrization bias, the determination of the best fit as the absolute minimum of the χ^2 would be a very inefficient way of proceeding – not to mention that one could end up fitting data fluctuations, that is, the output PDF set could fit perfectly the dataset used in the minimization process, but completely fail at fitting a different dataset. This noise–fitting could be, in principle, suppressed when averaging over Monte Carlo replicas, since the small differences of the replicas could make the final result less tailored to the particular dataset: however, it seems more meaningful to prevent this problem from occurring from the start. The best fit is determined using a cross–validation method, as represented in figure 1.2. The data are randomly (and differently for each replica) divided

into a training and a validation sample; the χ^2 is computed separately for the data in both samples, but only the training χ^2 is minimized, while the validation one is monitored during the minimization process. The best fit is defined as the point at which the validation χ^2 stops improving even if the training χ^2 keeps doing so: this would be the beginning of the noise–fitting process.

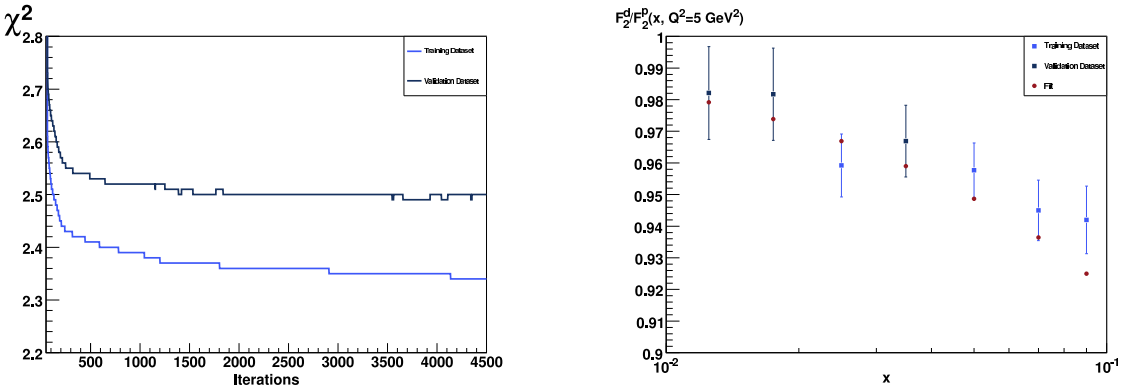

Figure 1.2: determination of the best-fit by cross-validation: the χ^2 of the fit to points in the training set (light blue) and the validation set (dark blue) are shown as a function of the number of iterations of the minimization algorithm (left plots), but only the χ^2 for training data is minimized. The data are shown in the right plot, along with the best fit (red points without error band).

In practice, in order to minimize the effect of random fluctuations in the data (or in the minimization algorithm), the stopping criterion needs to be imposed after a suitable averaging – for instance the average of the values of χ^2 in the last n iterations of the minimization algorithm.

However, a certain arbitrariness is introduced by such a method: one needs to choose the threshold within which these average values can fluctuate. In NNPDF3.0 this arbitrariness has been reduced by the following procedure. All replicas are minimized for the maximum number of generations M, while storing the initial validation χ^2 and PDF configurations: at the end of each generation, the validation χ^2 is computed again, and, if it is lower than the previous stored values, the PDFs are replaced with the current ones. The final PDFs result having the validation χ^2 corresponding to the end of its decreasing trend, and they are taken as the global best fit. This "look–back" method is completely objective, but it requires running all replicas up to M generations: M has to be big enough that the PDFs at the minimum do not depend on it.
Finally, at the end of the minimization, an a posteriori quality check on the resulting sample of Monte Carlo replicas is performed for each fit. These quality tests verify that the PDF generalized positivity has been successfully implemented, that no replica has a too unlikely arc–length L:

$$L = \int \mathrm{d}x \sqrt{1 + \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right)^2} \tag{1.2.19}$$

and that likewise no replica has a too unlikely value of the χ^2. As in the case of arc–length, if any given replica has a χ^2 whose value is more that four–sigma away the mean χ^2, it is automatically replaced by another replica that instead satisfies this condition.

Chapter 2

Uncertainties of PDFs

Anyone who considers arithmetical methods of producing random digits is, of course, in the state of sin.
John Von Neumann
(Remark made at a symposium on Monte Carlo methods)

In this chapter I discuss further the computation of PDF uncertainties: this will allow me to introduce the concept of tolerance and the main idea underlying my thesis work. I also review the closure testing method, which is a remarkable tool introduced in the NNPDF3.0 framework to compute theoretical uncertainties.

2.1 Tolerance

The first attempts to determine PDF sets which include PDF uncertainties immediately met with the difficulty that, as soon as wide enough datasets (see figure 2.1) are fitted, a standard statistical approach does not seem to be adequate.

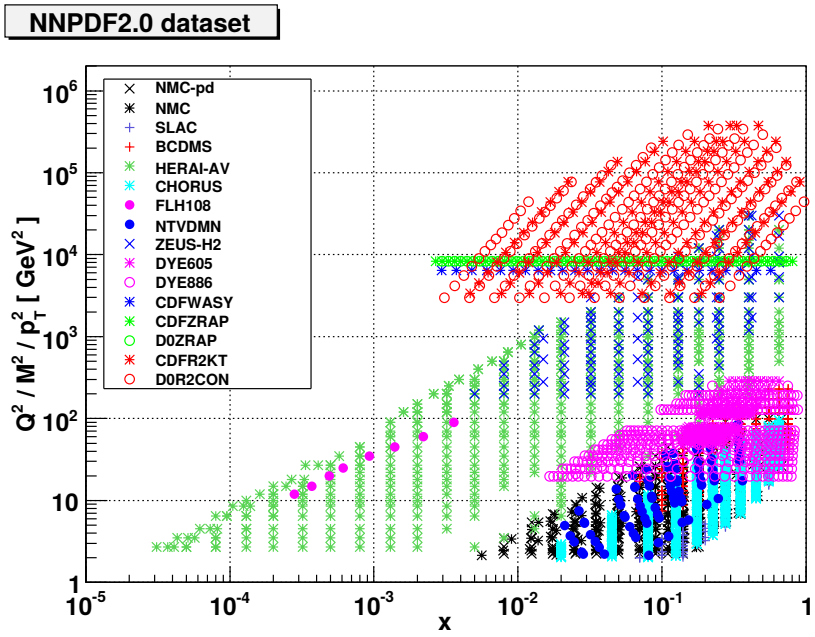

Figure 2.1: the data set used in a typical global PDF determination (NNPDF2.0 [8]).

Furthermore, predictions for hadronic collider processer are not always in complete agreement with each other. On both of these issues there has been considerable progress over the last several years. On the one hand, the understanding of statistical issues related to PDF uncertainties has advanced considerably; on the other hand, existing PDF determinations show a distinct convergence as various phenomenological and theoretical issues are addressed and understood. Fits, based on the Hessian approach and (1.2.1) parton parametrization, to wide enough datasets run into the difficulty that the minima relative to dataset are unexpectedly too distant from the global minimum. This patological behaviour is shown in figure 2.2: selecting, with Lagrange multipliers [25], a particular dataset to minimize (i.e. to reduce the $\Delta\chi_i^2$), one ends up in a minimum which corresponds in an increase $\Delta\chi^2_{\rm tot}$ of the total χ^2 larger than expected.

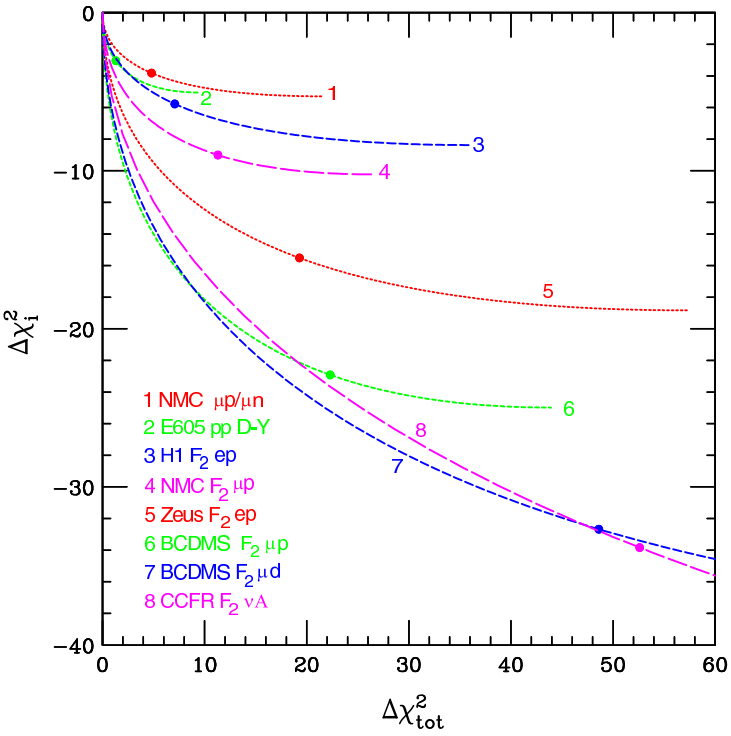

Figure 2.2: decrease of the χ^2 for various dataset entering a global fit plotted as a function of the increase in χ^2 of the global fit when moving away from the global minimum. See [25] for details and the meaning of the dots.

In fact, statistical fluctuations should cause around the 68% of the minima of the single datasets to lie $\Delta\chi^2_{\rm tot} \approx 1$ from the global minimum: looking at the figure above, not only the minima of the single datasets lie one order of magnitude further from the global minimun than expected, but also runaway directions seem to exist for some experiments.

This fact leads to consider that likelihood contours for the global fit can be determined only testing for the degree of agreement of all individual experiments (at the same time) with it: this is done introducing a "tolerance" parameter, as follows. After diagonalizing the Hessian matrix, the points in a neighborhood of the global minimum (in parameter space) which lie along one given eigenvector direction are candidate fits. The confidence ranges of these fits are evaluated with respect to the individual experiments: the range at, say, 90% c.l. of these fits is such that the best fit value lies within the 90% c.l. band of every experiment. This range is defined as the tolerance interval for the given eigenvector: the width of this interval can be measured in units of the variation of the χ^2 of the global fit, that is:

$$T^2 = \Delta\chi^2 \tag{2.1.1}$$

A result of this procedure is represented in figure 2.3, which refers to the eigenvector 13 as an example.

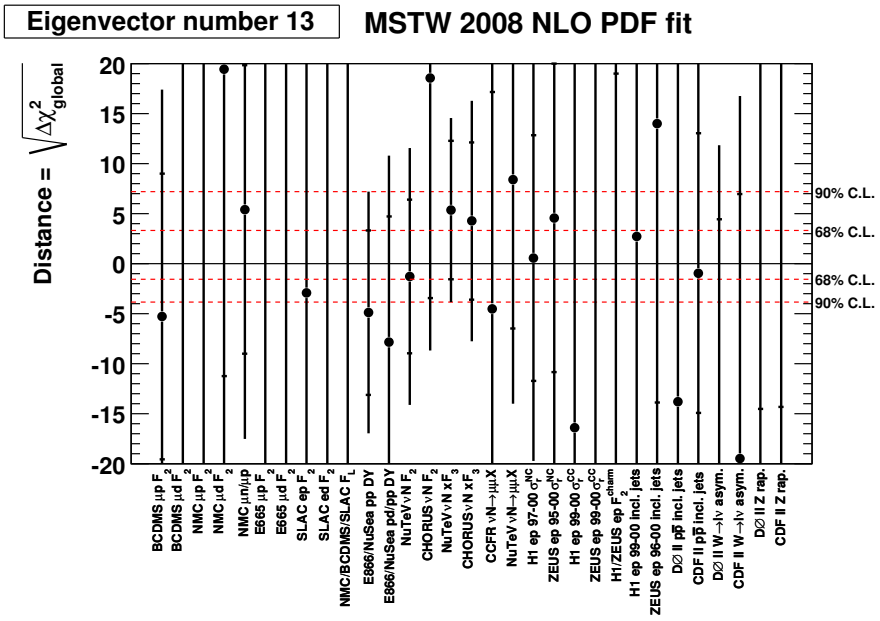

Figure 2.3: the dynamical tolerance interval for the 13th eigenvector, MSTW08 PDF fit. The inner and outer uncertainty bands correspond for each experiment to the 68% c.l. and 90% c.l. ranges.

The idea underlying this procedure is that PDFs should allow to obtain predictions for new processes at the desired confidence level. This means that if new experiments behave as the ones already included in the fit do on average, the result for a new experiment should have a 68% chance of falling into the so obtained one-σ band; if the one-σ band were defined only on standard statistical grounds, the chances of the measurements falling outside the band would be much higher.

In [29] was found that $T^2 = 100$ worked for all eigenvalues and experiments at 90% c.l., corresponding to $\Delta\chi^2 \approx 37$ at one-σ; in [30] a similar analysis led to a $T^2 = 50$ (90% c.l.) and $\Delta\chi^2 \approx 16$ (68% c.l.). The direct consequence of such analyses would be that all experimental uncertainties have been underestimated by a factor of about 6 and 4 respectively; while some uncertainty underestimation is acceptable, such a large factor is unreasonable and demands for further investigation. As explained before, one can perform a tolerance analysis for each eigenvector keeping trace of the individual experiment contribution instead of considering a global tolerance value: this is called "dynamical" tolerance [7]. This method provides insight on the relation between data and PDF parameters and their mutual consistency: as a result, a tolerance $T^2 \lesssim 43$, with most values in the range $4 < T^2 < 25$, is found, mitigating the large tolerance problem.

This sort of fine–tuning of confidence intervals isn't needed using the Monte Carlo approach together with neural network parametrization, as tested in [8, 28]. In the context of the Monte Carlo approach, the one-σ interval is just the standard deviation of the sample, and one could check if in this range lie the 68% of the distributions of PDF replicas, as shown for the gluon in figure 2.5. It is possible to verify a posteriori whether or not the fits are consistent with new data and their inclusion: to answer this question, statistical tests have been successfully performed in [8, 28, 31]. It should be noted, however, that computing one-σ intervals from the sample itself is not so trivial: as explained in 1.2.2, the $\overline{\chi}^2$ has fluctuations of the order of the number of data points, and these fluctuactions affect each replica, so that one needs a very large sample to determine the χ^2 accurately. In the following section I review how the issues which may be responsible for large tolerance values can be handled both in a Hessian and Monte Carlo approach.

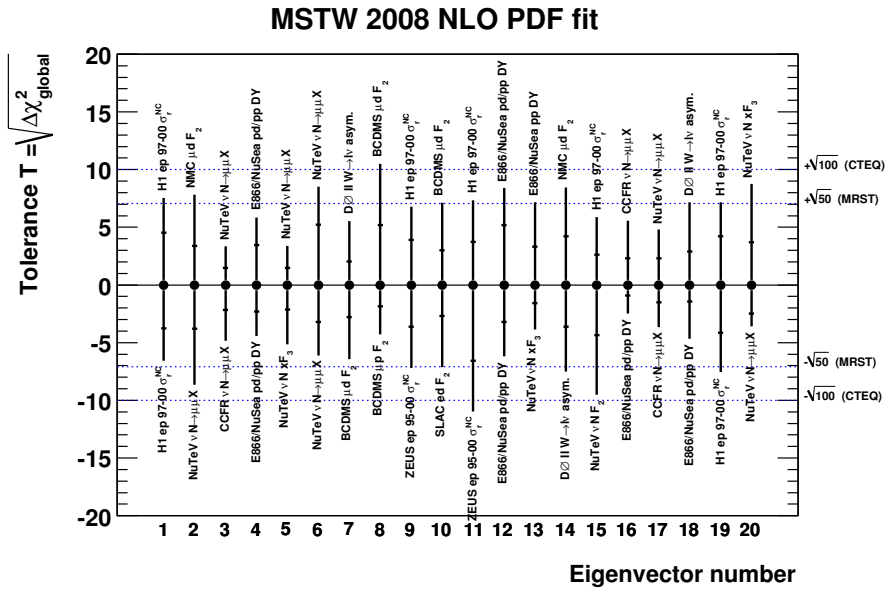

Figure 2.4: the tolerance interval for each eigenvector, along with the experiments which determine it; the $T^2 = 50$ and $T^2 = 100$ previously [29, 30] adopted are also shown (from [7]).

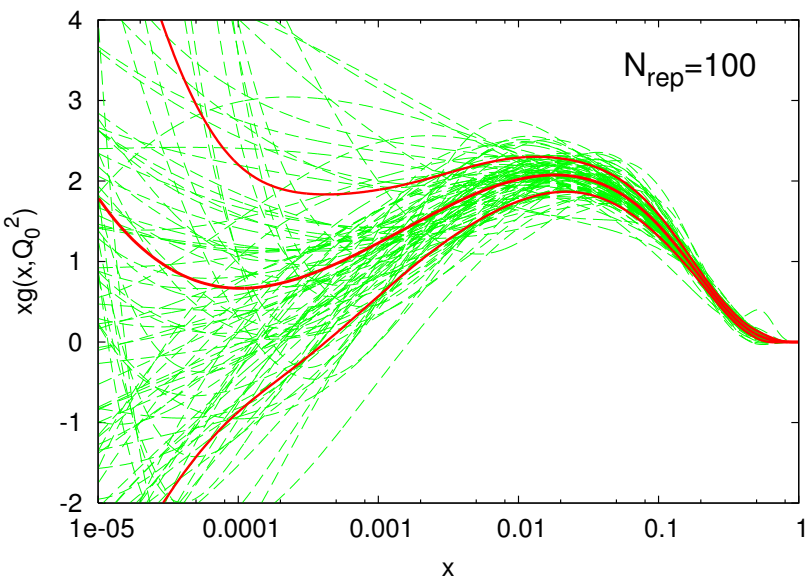

Figure 2.5: one–σ interval computed [28] from a distribution of 100 replicas of the gluon PDF.

2.2 Parametrization bias and data incompatibility

As seen in figure 2.3, the poor mutual compatibility of the experiments resulted in large corrections to the $\Delta\chi^2$ values which define uncertainty intervals: this correction is called "tolerance" $T^2 = \Delta\chi^2$. One can naturally ask which could be the sources of large tolerance values, and list them as follows:

1. theoretical inadequacy;
2. data incompatibility, as already mentioned;
3. parametrization bias.

Let us discuss them separately.
Theoretical inadequacy, that is, the theory not being correct enough, may be dealt with by considering that, despite the theoretical progress of the last ten years, the value of the tolerance has not sizably changed.

As for data incompatibility, here I review the procedure adopted to study it: for more details consult [32, 33] and references therein. In the Hessian method the dependence of χ^2 on the parameters $\mathbf{a}$ can be expanded about the minimum using a Taylor expansion of the second order. The eigenvectors of the Hessian matrix can be used as a basis to obtain a linear transformation to new coordinates for which

$$\chi^2 = \chi^2_{\text{min}} + \sum_{i=1}^{N} y_i^2 \tag{2.2.1}$$

Let χ^2_{S} be the contribution to the χ^2 from the data subset S: in a neighborhood of the global minimum, χ^2_{S} can be expanded through second order in the coordinates $\mathbf{y}$ using Taylor series. The eigenvectors of this second–derivative matrix can be used to provide another linear transformation which diagonalizes the relative quadratic form, without spoiling (2.2.1). This is known as the Data Set Diagonalization (DSD) method: it allows to separate the contribution to the χ^2 of a specific data subset from the rest. Data incompatibility is now measured as the distance of the minima of these two contributions in units of the corresponding standard deviation: if experimental uncertainties are correctly estimated, these distances should be gaussianly distributed. The results of this study are represented in the figure 2.6 below.

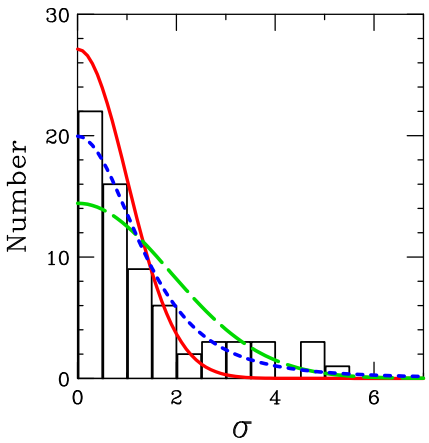

Figure 2.6: distribution of discrepancies between each experiment entering a global fit and the global best fit [32]. The solid red curve is the standard gaussian distribution, the dashed green curve is a gaussian distribution with σ rescaled by a factor 1.88, the dotted blue curve is a Lorentzian distribution.

As the plot above suggests, the distribution of discrepancies is not gaussian. If we still fit it to a gaussian (dashed green curve in the figure), its uncertainty should be rescaled by about a factor of 2: this suggests

uncertainty underestimation by a similar factor, which corresponds to $T^2 \sim 10$ for 90% c.l..
Of course a test on data inconsistency can be performed even with the Monte Carlo approach, by comparing the effect of subsequent inclusion of different datasets into a fit. For example, if the gluon distribution determined from jet is compatible with that extracted from DIS data but less so with that found from Drell–Yan experiments, inclusion of jet data in a pure DIS fit would have a different effect than their inclusion in a fit which contains both DIS and Drell–Yan. When this test is performed [8] no evidence for data incompatibility is found, as shown in figure 2.7 below.

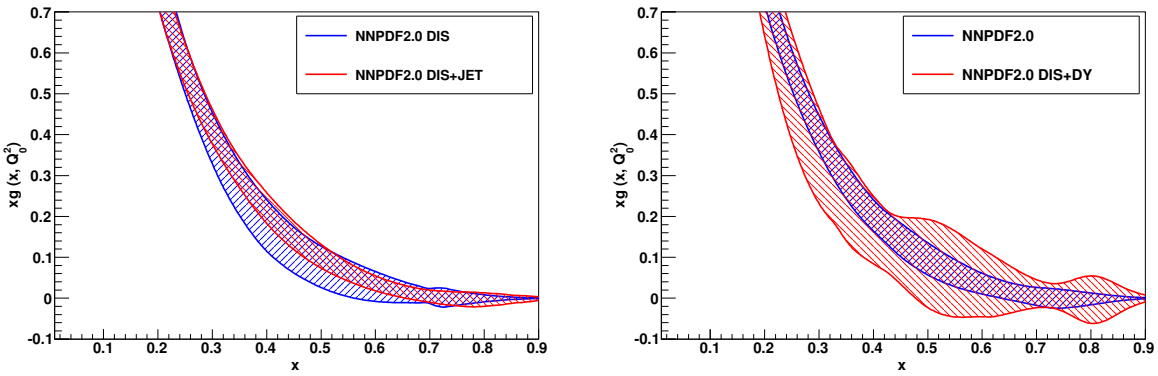

Figure 2.7: tests of data compatibility by changing the order of inclusion of data in fits with different datasets, based on the NNPDF2.0 [8] PDF determination: effect on the gluon distribution of the inclusion of jet data in a fit to DIS data only (left) or in a fit to DIS+Drell–Yan data (right).

Let us now deal with the hypothesis that parametrization bias may be responsible for this effect. In [34] is explained how this could happen in the Hessian approach, and here I provide a summary of that reference. Assume that a parameter y on which the PDFs depend is not fitted, but fixed at some value which is far from its best–fit. It is clear from figure 2.8 that the one-σ range for the other parameters corresponds to a greater variation of χ^2 than the case of y having the best–fit value.

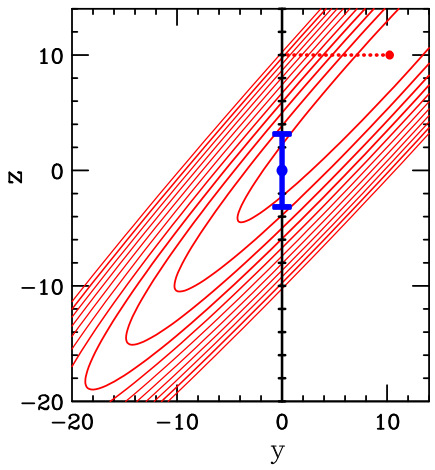

Figure 2.8: the one–σ range for parameter z when parameter y is kept fixed away from the best–fit value corresponds to a range which is larger than $\Delta\chi^2 = 1$.

In [34] a first estimate of the size of this effect was also provided by repeating the PDF fit of [9] with a much more general parametrization, based on the expansion of the gluon PDF on Chebyshev polynomials. The resulting fits whose χ^2 is similar or smaller than the best–fit χ^2 of the original parametrization are

found to suggest $T^2 \geq 10$, and this just to account for the bias imposed on the gluon PDF by the [9] parametrization.
One could as well wonder whether the gaussian distribution assumption for experimental data is a source of bias itself. This has been dealt with in [19], using a Monte Carlo approach together with a standard parton parametrization: data replicas based on the HERA DIS data have been generated following a gaussian or a lognormal distribution and used for a PDF fit based on the standard functional form (1.2.1). In each case, both Monte Carlo and Hessian uncertainties are computed, and the results are shown in figure 2.9.

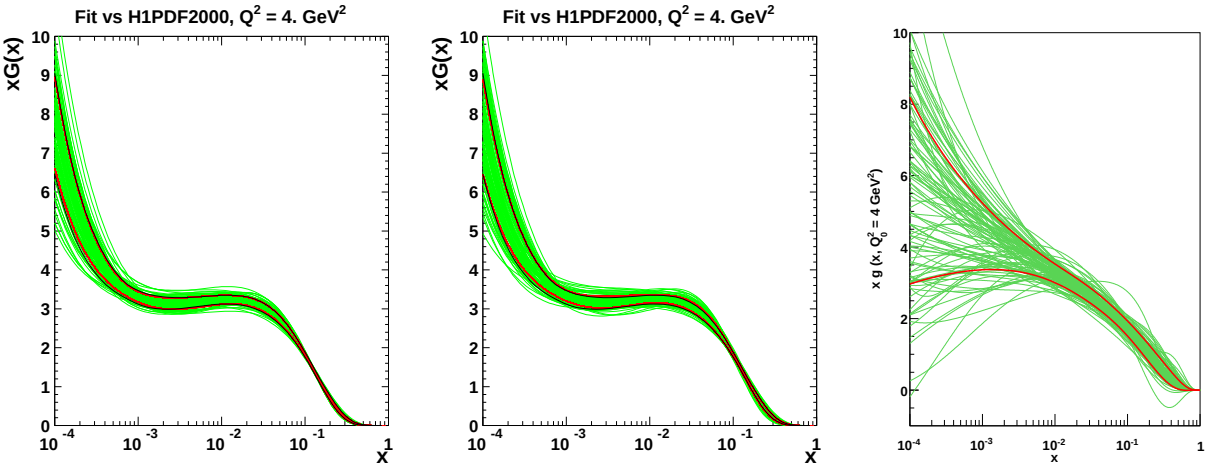

Figure 2.9: from left to right: gluon determined from lognormal data, gluon determined from gaussian data, gluon determined from the same gaussian data but using the neural network parametrization. See refs. [19, 31].

If on one hand the choice of probability distribution seems to play no major role –by the central limit theorem, with enough data everything looks gaussian–, on the other it is clear that the uncertainty is much wider in the case of the NNPDF parametrization: this suggests that it is indeed the form of the parametrization to play the dominant role.

Here lies the main idea this work is based on: if one were able to "control everything", from start to finish, generating data from the standard functional form and then fitting them with the standard PDFs and neural network–based PDFs, one could perform the following very important test. One could confront theoretical uncertainties, and check if they are smaller in the case of the standard parametrization: it would then be interesting to compare the corresponding tolerance value with the one present in literature, in order to finally establish whether tolerance is just a consequence of parametrization bias or not. The question is now how to "control everything": the recent introduction of the NNPDF collaboration, i.e. closure testing, is the answer, and that's what I'm going to deal with in the next section.

2.3 Closure Testing

2.3.1 Introduction and interpretation

Closure testing is the name of the strategy that has been adopted to validate the fitting methodology of the NNPDF3.0 framework [26]. The increase of experimental data and the increased precision of the resulting PDFs made it necessary to eliminate methodological uncertainties: a basic requirement for a successful methodology is to be able to fit very different datasets with the same procedure, i.e. without fine–tuning it to the dataset or the underlying theory.
The idea of the closure test is simple: given a PDF parametrization and a theoretical model, NNPDF generate a set of pseudo–data with known and realistic statistical properties by using the covariance matrices of the real datasets. These pseudo–data are said "perfect" in that they have known statistical properties, no internal inconsistencies and are consistent with the theoretical model used to produce them. Performing a fit to these pseudo–data should then return the assumed starting PDFs and their uncertainties.
As in [26], I will review three distinct categories (levels) of closure tests, which differ in the amount of stochastic noise the pseudo–data are generated with. This stochastic noise is based on the complete information contained in the experimental covariance matrix, so that the pseudo–data behave as realistically as possible. The three levels of closure test are set up as follows:

- Level 0
  Using a given set of input PDFs, pseudo–data are generated without any experimental uncertainties: N fits are then performed, each to the very same set of pseudo–data, minimizing the χ^2 per degree of freedom and using different seeds to initialize the random numbers used in the minimization. N replicas $\{f^k_{\text{fit}}\}$, $k = 1, \ldots, N$ are thus obtained.
  In Level 0 closure tests the fit quality can be arbitrarily good, if a flexible enough parametrization is used. In fact, by construction there are no inconsistencies and no noise, and thus perfect fits with vanishing χ^2 exist: the plural is due to the fact that there is an infinity of fits going through all data points with different in between interpolations.
  The Level 0 closure test serves as an assessment of the minimization procedure efficiency: the χ^2 for each replica should decrease monotonically towards 0 as the fit proceeds, if the fitted PDFs are flexible enough. The best–fit χ^2 (i.e. the χ^2 evaluated for the average of all replicas) should also go to zero.

- Level 1
  Stochastic fluctuations are added on top of the pseudo–data generated for the Level 0 closure tests D^0_o as follows:
  $$D^1_o = (1 + r^{\text{nor}}_o \sigma^{\text{nor}}_o) \left(D^0_o + \sum_{i=1}^{N_{\text{sys}}} r^{\text{sys}}_{o,i} \sigma^{\text{sys}}_{o,i} + r^{\text{stat}}_o \sigma^{\text{stat}}_o \right) \tag{2.3.1}$$
  o stands for an experimental observable, σ^{stat}_o, $\sigma^{\text{sys}}_{o,i}$ and σ^{norm}_o are the statistic, systematic and normalization uncertainties for each dataset, while r^{stat}_o, $r^{\text{sys}}_{o,i}$ and r^{norm}_o are random numbers generated with the appropriate distribution to reproduce the experimental covariance matrix [31].
  In Level 1 fits the N replicas are fitted to the very same pseudo–data generated as in (2.3.1) and, as before, a different random seed is used for initializing the minimization procedure. As a result experimental uncertainties are not propagated to those of the fitted PDFs, which are thus expected to underestimate the real uncertainties.
  Since the pseudo–data have fluctuated on average by one standard deviation away from the Level 0 value, the best–fit χ^2 per degree of freedom should be around 1. Also in this case perfect solutions for the minimization exist, i.e. such that $\chi^2[\{f_{\text{fit}}\}] = \chi^2[\{f_{\text{in}}\}]$; not all PDFs in $\{f_{\text{fit}}\}$ will be perfect, and their distribution will be, as explained later, too narrow.

- Level 2
  Fluctuations are propagated to fitted PDFs as follows: for $k = 1, \ldots, N$ a set of pseudo–data $\mathcal{D}^{2,k}$

is generated from the shifted pseudo–data in (2.3.1) via

$$D_o^{2,k} = (1 + r_o^{\mathrm{nor},k}\sigma_o^{\mathrm{nor}}) \left(D_o^1 + \sum_{i=1}^{N_{\mathrm{sys}}} r_{o,i}^{\mathrm{sys},k}\sigma_{o,i}^{\mathrm{sys}} + r_o^{\mathrm{stat},k}\sigma_o^{\mathrm{stat}} \right) \tag{2.3.2}$$

as before the set of random numbers is different replica by replica. The key point is that each replica is fitted to a different set of pseudo–data (and using the exaclty same procedure as in the case of real data), which should cause the propagation of the fluctuations on Level 1 pseudo–data to the fitted PDFs. The final error function of a Level 2 fit $\chi^2[\{f_{\mathrm{fit}}^k\}, \mathcal{D}^{2,k}]$ is then expected to be close to 2 (in a way, each replica contains two fluctuations), and the input PDFs f_{in} should lie within the one–σ band of the fitted PDFs with a probability of around 68%.

Let us now take a better look at the three levels of closure testing in order to separate the different sources of PDF uncertainties, and thus realize how clean an environment it provides. One can gain deeper insight on PDF uncertainties by comparing the three levels: in Level 0 fits, the only significant component is the interpolation and extrapolation uncertainty (extrapolation uncertainty for short), in Level 1 there is also the uncertainty due to the choice of functional form and in Level 2 one finally adds the uncertainties due to the fluctuations of the experimental data. The comparison of the three levels therefore allows to analyse how the total PDF uncertainties are made up of data, functional and extrapolation uncertainties. In Level 0, as already discussed, uncertainties are due to all the ways PDFs can interpolate the data points (interpolation) and fluctuate outside the data region (extrapolation): this is what it is referred to as the *extrapolation* uncertainty. It is difficult to precisely identify the extrapolation region, since the PDFs depend in a highly non–trivial way on data (i.e. measured cross–sections etc.): at small enough and large enough values of x non–negligible uncertainties for all PDFs are expected, but the interpolation freedom through datapoints generates uncertainties also at intermediate x. Moreover, some PDFs can even compensate each other to reproduce the same experimental data, and this accounts as a contribution to uncertainty. All this can be seen in figure 2.10: for increasing number of generations, uncertainties become smaller but do not go to zero, and even remain big at small and large x, where there is no experimental information – i.e. where extrapolation uncertainty kicks in.

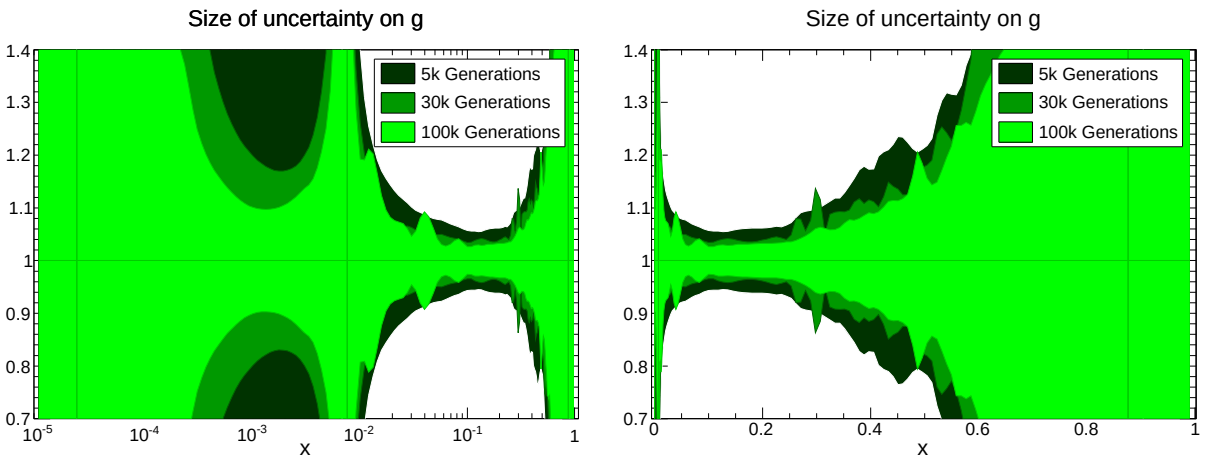

Figure 2.10: comparison of relative PDF uncertainties obtained from Level 0 closure test fits with MSTW2008 NLO as input set and three different training lengths. Results for the gluon (g) are shown at the input parametrization scale of $Q^2 = 1\,\mathrm{GeV}^2$, both in a logarithmic (left) and in a linear (right) scales [26].

In the case of the gluon, shown in figure 2.10, one can also see a component of uncertainty which does not vanish even for extremely long training length: that is the effect of the interpolation between data points. It must also be noted that in most of the data region the uncertainties at 30k generations and

100k generations are very close, and that the relative PDF uncertainty blows up at $x \sim 0.005$ because the MSTW08 gluon PDF has a node in that region.
In Level 1 closure fits the central values of the data fluctuate around the theoretical prediction, so that $\{f_{\rm fit}\} = \{f_{\rm in}\}$ no longer provides an absolute minimum for the χ^2: if the PDF parametrization is flexible enough, the minimization algorithm should provide a large number of different functional forms which result in $\chi^2 \approx 1$. That's why it is said that in Level 1, on top of the extrapolation uncertainty, *functional* uncertainty –i.e. the spread of these functional forms– is included. The look–back method discussed in section 1.2.2 ensures that, regardless of the length of the fit, the final χ^2 does not decrease beyond the overlearning point, and so functional uncertainty will survive even for infinite training length.
In Level 2 fits the starting point is Level 1 pseudo–data: starting from these, a set of N Monte Carlo replicas is fitted with the same algorithm to the different fluctuations added to Level 1 pseudo–data, resulting in an ensemble $\{f^k_{\rm fit}\}$ whose statistical properties are a faithful propagation of the fluctuations in the underlying dataset. The increase in uncertainty from Level 1 to Level 2 fits is thus pure *data* uncertainty.

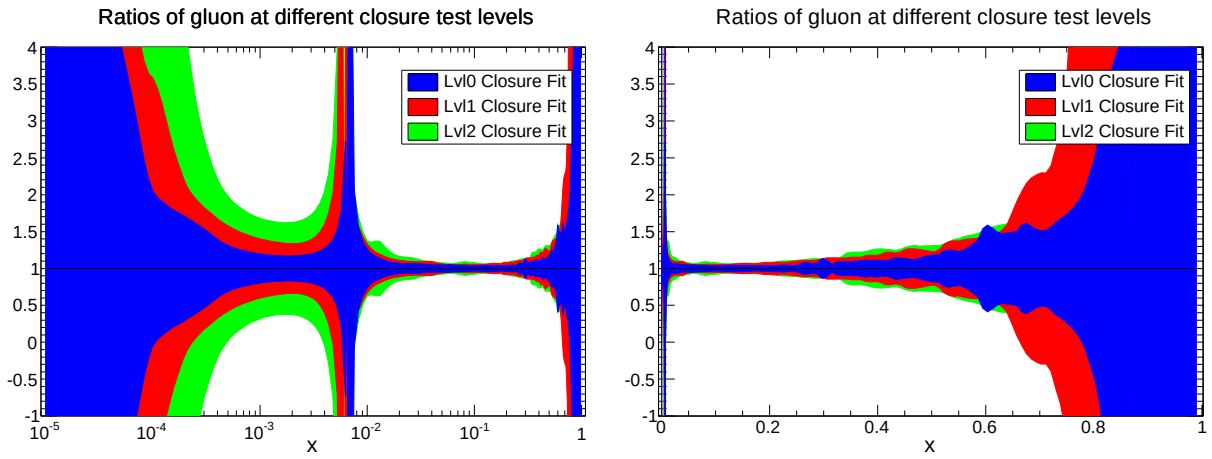

Figure 2.11: same as figure 2.10, but now comparing PDF uncertainties obtained from Level 0, Level 1 and Level 2 closure test fits. The plots show the 68% confidence level PDF uncertainty band for each of the fits, normalized to the corresponding central value of each fit. Results for the g are shown both in a logarithmic (left) and in a linear (right) scales [26].

One can now understand qualitatively what is shown in figure 2.11: firstly, it can be seen that Level 0 uncertainties are smaller than Level 1 ones, which are in turn smaller than those at Level 2, thus confirming that at each Level a new component of the total PDF uncertainty is included. One can also see that at small and large x the extrapolation uncertainty dominates, since the Level 2 PDF uncertainties are already reasonably reproduced by those of Level 0. As already noticed from 2.10, the Level 0 extrapolation uncertainty is non–negligible also in regions with experimental data, due to interpolation and degeneracies. Figure 2.11 shows that Level 0 uncertainty in the data region is around 5% or more, and it can go down to about 1% in small regions close to PDF peaks. One can see also that functional uncertainty is generally sizable, particularly when going from the data region to the extrapolation one – and especially at large x. A very important observation is that, in the experimentally well covered regions, the three components are roughly of similar size: the gluon at $x \sim 10^{-3}$ is well constrained by the high–precision HERA measurements, and there the extrapolation, functional and data uncertainties are all of similar size.
The consequence is clear: a proper estimate of the tohotal PDF uncertainty must include all these three components, and the estimates which include data uncertainties only will underestimate the overall PDF uncertainty. It is natural to assume that the tolerance method serves as an alternative way of supplementing the data uncertainty with these extra necessary components.

2.3.2 Quantitative validation

Only a qualitative discussion about the various components of the total PDF uncertainty has been provided, by comparing the results of the different Levels. In this section I provide a brief summary of the quantitative estimators of the successfulness and effectiveness of the closure tests, which I'll use in the next chapter to validate my closure test fits.

In Level 1 and Level 2 closure fits, the central χ^2 obtained from the average of the fitted PDFs is supposed to reproduce the one computed using the input PDFs, so that

$$\chi^2[\mathcal{T}[\overline{f}], \mathcal{D}_1] \approx \chi^2[\mathcal{T}[f_{\text{in}}], \mathcal{D}_1]$$

indicating with $\chi^2[\mathcal{T}[f]]$ the χ^2 obtained by fitting the theoretical predictions computed from a certain PDF set ($\mathcal{T}[f]$) to a certain set of data, in this case $\mathcal{D}_1$, i.e. pseudo–data generated in Level 1 closure fits. It is thus possible to test for a certain fit by using defining the following quantity:

$$\Delta_{\chi^2} = \frac{\chi^2[\mathcal{T}[\overline{f}], \mathcal{D}_1] - \chi^2[\mathcal{T}[f_{\text{in}}], \mathcal{D}_1]}{\chi^2[\mathcal{T}[f_{\text{in}}], \mathcal{D}_1]} \tag{2.3.3}$$

that is the % difference from the central (i.e. computed with the average $\overline{f}$ of the replicas) χ^2 of the closure test fit and the input PDFs computed with respect to the same closure test dataset. This estimator quantifies how well the closure fits reproduce the theoretical predictions of the input PDFs, and thus measures the success of the test. Moreover, since the χ^2 is a decreasing function of the training and the input PDFs are assumed to be the true ones, the sign of Δ_{χ^2} tells if there has been overlearning ($\Delta_{\chi^2} < 0$), underlearning ($\Delta_{\chi^2} > 0$) or perfect learning ($\Delta_{\chi^2} = 0$).

The second quantifier I review regards the accuracy with which PDF uncertainties are reproduced. Assuming gaussianity, the n–sigma intervals about the prediction can be interpreted as confidence levels for the true value, that is the true value must fall within the 1–sigma uncertainty band in the 68.3% of cases, in the 2–sigma band in the 95.5% of cases and so on. One can then define the following estimator:

$$\xi_{n\sigma} = \frac{1}{N_{\text{PDF}}} \frac{1}{N_x} \frac{1}{N_{\text{fits}}} \sum_{i=1}^{N_{\text{PDF}}} \sum_{j=1}^{N_x} \sum_{k=1}^{N_{\text{fits}}} \mathcal{I}_{[-n\sigma_{\text{fit}}^{i(k)}(x_j), n\sigma_{\text{fit}}^{i(k)}(x_j)]} \left(\overline{f}_{\text{fit}}^{i(k)}(x_j) - f_{\text{in}}^{i}(x_j) \right) \tag{2.3.4}$$

where N_{PDF} is the number of flavours (labelled with i), N_{fits} is the number of fits considered (labelled with k) and N_x is the number of points at which the PDFs are sampled (labelled with j), while

$$\mathcal{I}_{[a,b]}(x) = \begin{cases} 1 & x \in [a,b] \\ 0 & \text{else} \end{cases}$$

On the aforementioned statistical grounds one expects that, if the closure fits are successfull:

$$\xi_{1\sigma} \approx 0.68 \qquad \xi_{2\sigma} \approx 0.95 \qquad \xi_{3\sigma} \approx 0.99$$

Note that the $\xi_{n\sigma}$ should assume the values above (i.e. the distribution of deviations should be gaussian) even if only one PDF at one point, say $x = 0.1$ for the u quark distribution, were chosen to compute them: the only average that has true statistical significance is the one on the fits. In fact, all the other averages introduce correlations, biasing the indicators: all the PDFs are correlated with each other due to the momentum sum rule, and values of the same PDF at two points next to each other can be correlated. If, however, different PDFs are considered at different x-points, those in (2.3.4) can be considered statistically significant indicators. Computing these indicators with a certain degree of reliability would require generate many closure test fits, which is computationally expensive. It is possible then to approximate the mean of each fit $\overline{f}_{\text{fit}}^{i(k)}$ with the Level 1 replicas, over which the average is computed using the corresponding Level 2 uncertainties (it's a fact that the variation in the PDF uncertainties between different closure test fits is small, so this isn't a rough approximation).

Another indicator is the following:

$$d_{\mathrm{ct}}[f_{i,\mathrm{fit}}, f_{i,\mathrm{in}}](x,Q) = \sqrt{\frac{(\overline{f}_{i,\mathrm{fit}}(x,Q) - f_{i,\mathrm{in}}(x,Q))^2}{\sigma^2[f_{i,\mathrm{fit}}](x,Q)}} \tag{2.3.5}$$

which quantifies the distance, for fixed flavour i, between the central closure fitted PDF and the input PDF in units of the closure fitted PDF uncertainty.

Chapter 3

Testing for parametrization bias

"If googling stack-overflow has not helped you, the only recourse is to power cycle"
or, in other words:
"Have you tried restarting it?"
Mom

3.1 Introduction to my work

So far I've reviewed what PDFs are, the two main methodologies (Monte Carlo and Hessian) by which they are computed and the problem of redefining confidence intervals through the introduction of tolerance in Hessian fits. The aim of this work is to try to understand which could be the cause of this problem.

The idea of how to do this is inspired by the properties, seen in the last chapter, of the closure testing procedure. In fact, the features that make it the ideal tool for my purpose are that not only it allows to study in a controlled environment how fits proceed, but also it allows to use an arbitrary functional form as a fitting function. As already discussed, in the three Levels of closure testing different sources of uncertainties are accounted for, among which the functional uncertainty. This source of uncertainty is completely ignored by fits performed with a rigid functional form – which is the case of Hessian fits; that could possibly be the reason why tolerance is needed to enlarge the uncertainties obtained with the Hessian methodology.

Thanks to closure testing, it is possible to generate data (the D_o^0 in section 2.3.1) with one arbitrary functional form and then compare what happens if one uses, as a fitting functional form, a neural network and the data–generating functional form. If both these closure test fits are consistent, i.e. if they pass the validation procedure discussed in section 2.3.2, the uncertainties they produce are faithful; if these two sets of uncertainties differ in a sizable way, the reason is that in one case the underlying functional form is known, in the other it is not. It is expected that using the same rigid functional form both as data–generating and fitting function results into faithful uncertainties (i.e. the closure test is validated), which are smaller than those obtained from using a more general functional form as a fitting function. The next step, not included in my work, would be to perform a closure test with two different rigid functional forms as data–generating and fitting functions: in this case it is expected that, again, the uncertainties are still smaller than those obtained using a more general functional form as a fitting function, with the difference that in this case the closure test is not successfully validated.

In the following I am going to describe and motivate the steps I have taken to perform this study, the first of which consisted in the choice, implementation and test of the Hessian parametrization in the NNPDF framework.

3.2 Implementation of MSTW08 parametrization in the NNPDF framework

As my goal was to perform closure tests with a Hessian parametrization (and compare their results to those obtained from a closure test with a more general functional form), I needed to add to the NNPDF3.0 methodology and code the possibility of fitting another functional form. One could argue that I didn't really need to perform closure tests with a Hessian functional form, since there is already another way to find the best fit – that is, the Hessian methodology. However, I used the NNPDF3.0 methodology even for the Hessian functional forms because, besides the simplicity of using the same code for both parametrization, the procedure followed by the MSTW08 group (whose functional form I used, as explained below) to fit PDFs does not actually coincide with a standard Hessian procedure. Their methodology is more complicated, as they start from a larger number of parameters (34) of which four are fixed in a first fitting round and then another minimization is performed on the remaining. Since it is not a pure Hessian procedure, it might defeat the purpose of the standard Hessian minimization itself: that is why I need to use the closure testing with their parametrization as well. The results of these closure tests were compared to those obtained from closure fits with neural networks as a more general functional form, since they are native to the framework I am using and since there is no other well-established alternative when it comes to Monte Carlo PDF fitting.

To begin with, I needed to choose a functional form used in one of the Hessian fits: this functional form had to be complex enough that the problem of choice of functional form would be a serious one. In fact, if the data were described by a few–parameter distribution, there would not be the problem of ignoring the functional form, as the result would not show significative differences in the fits performed with said simple functional form and a general one. Only two PDF set parametrizations satisfy this requirement, which are those of MSTW08 and CTEQ: I arbitrarily chose the MSTW08 functional forms, but this whole study could be repeated with those of CTEQ without any conceptual difference.

The fist step of my work consisted in modifying the set of `C++` codes which constitute the NNPDF3.0 framework, to accommodate for the MSTW08 PDF parametrizations (1.2.3)–(1.2.9). In order to do this, I created a 9–parameter function class, so as to account for all their forms: the generic MSTW08 PDF has been codified as follows.

$$xf(x) = a_1 x^{a_2}(1-x)^{a_3}(1 + a_4\sqrt{x} + a_5 x + a_6 x^2) + a_7 x^{a_8}(1-x)^{a_9} \tag{3.2.1}$$

Each PDF corresponds to a particular choice of the $\{a_i\}$: in the table below is reported the best–fit value of the MSTW08 PDF (at the scale $Q^2 = 1\,\mathrm{GeV}^2$ and at Next–to–Leading perturbative Order) parameters – see the central column of Table 4 in [11].

PDF	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
xu_v	0.25871	0.29065	3.2432	4.0603	30.687	0	0	0	0
xd_v	12.288	0.96809	5.9435	−3.8911	6.0542	0	0	0	0
xS	0.31620	−0.21515	9.2726	−2.6022	30.785	0	0	0	0
$x\Delta$	8.1084	1.8691	11.2726	0	13.609	−59.289	0	0	0
xg	1.0805	−0.42848	3.0225	−2.2922	3.4894	0	−1.1168	−0.42776	32.869
xs^+	0.047915	−0.21515	9.7466	−2.6022	30.785	0	0	0	0
xs^-	−0.011629	0.2	11.261	0	−62.305	0	0	0	0

where $xs^{\pm} = x(s \pm \overline{s})$. After that, I performed a sanity check on this very first step, to make sure that I didn't introduce any bugs and that I could correctly reproduce the true MSTW08 PDFs. This was possible thanks to the `LHAPDF6` [36] library, which consists in a collection of highly–optimized PDF sets: these PDFs are interpolations that I could use as benchmark for the PDFs generated with my code.

Among these PDF sets there is the one of my concern, i.e. MSTW2008nlo68cl; to perform this control I sampled the ratio of "my" PDFs to the central value (that is, the best fit PDFs) of those stored in the library over the following x-points.

$$\{10^{-5}, 5\times10^{-5}, 10^{-4}, 5\times10^{-4}, 0.001, 0.003, 0.006, 0.01, 0.03, 0.06, 0.1, 0.3, 0.5\}$$

I computed the average and standard deviation of the ratios both dividing the single basis PDFs and considering them all together: the results are reported in the following table.

ratio	g	d_v	u_v	s^+	s^-	Δ	S	general
average	1.0000	1.0078	1.0002	0.9998	0.9998	0.9729	0.9995	0.9971
st. dev.	0.0012	0.0204	0.0008	0.0004	0.0008	0.8995	0.0011	0.3295

For the $x\Delta(x)$ PDF the average ratio differs from unity for about 3%, which is much larger than the second biggest relative discrepancy (which is xd_v, with 0.8% relative discrepancy): the ratio is in fact not well defined when sampled over the four smallest x-points, since $x\Delta(x) \lesssim 10^{-6}$ in that range. Avoiding sampling over the first four points, which would lead to the indeterminate form 0/0, I obtained the results listed in the following table.

ratio	g	d_v	u_v	s^+	s^-	Δ	S	general
average	1.0004	0.9997	0.9999	0.9998	0.9998	0.9964	0.99937	0.9993
st. dev.	0.0011	0.0017	0.0005	0.0004	0.0009	0.0320	0.0013	0.0117

It can be noted that now the ratio for the $x\Delta(x)$ PDF deviates from unity for less than 0.4%, while the ratios for the other PDFs have not undergone significative changes. Also the average ratio for all the PDFs combined has improved from 99.71% to 99.93%, with a much smaller standard deviation as well. Such agreement in the mostly data–constrained region, i.e. $10^{-3} \lesssim x \lesssim 10^{-1}$, convinced me that the code was producing MSTW08 PDFs with a satisfactory degree of agreement and that I could go on.

The other main implementation I had to work on was the Genetic Algorithm (GA). When I reviewed the Monte Carlo fitting methodology, I explained how each fit would run for the maximum number of generations and then a look–back method is adopted to find the best fit PDFs. I did not explain the technical feature that effectively contributes to the minimization of the χ^2, which is the Genetic Algorithm. The GA consists of three main steps: mutation, evaluation and selection. Firstly, many (80) mutant PDF sets are generated from a parent set from the previous generation: the goodness of fit of each mutant is computed and then the best–fit mutant is identified and passed on to the next generation, erasing the others. That's what happens going from a generation to the next: the GA generates a new set of functions, by mutating the parameters, from the previous ones and selects those that best fit the data. A Genetic Algorithm is used because, despite being slower than deterministic minimization methods (like Newton's, in which one follows the most direct route to the minima using information on the function's curvature) in the case of smooth functions, it has the advantage of being able (if the mutation is big enough) to escape from local minima and reach the global minimum – which is impossible for deterministic algorithms. This is a very desirable feature when dealing with complex figures of merit which depend on a large number of parameters.

There are many different forms of GAs: the fundamental idea consists in the introduction of mutations, which can be defined in various ways. In the NNPDF3.0 methodology it has been found to be advantageous to utilize the following nodal mutation: if a node is selected, its threshold and all of the weights (respectively, the θ_n^i and $\omega_{nm}^{(i)}$ parameters in (1.2.18)) are mutated according to [26]

$$\omega \to \omega + \frac{\eta r_\delta}{N^{r_{\text{ite}}}}$$

where η is the standard mutation size, r_δ is a uniform random number between -1 and 1, N is the number of generations elapsed and r_{ite} is another random number uniformly distributed between 0 and 1. The initial parameters of the neural networks are chosen at random between 0 and 1, and the mutations treat each node in the same way. However, as it is evident from the tabulated parameters previously shown, in the case of MSTW08 PDFs the parameters are very different to each other, spanning four orders of magnitude: this means that the mutation of each parameter cannot ignore its original size. I implemented a parameter mutation such that

$$a \to a + \frac{k|a|}{N^x} \tag{3.2.2}$$

where k is a random number between -1 and 1, N is the number of generations elapsed number and x is a random number between 0 and 1; both k and x are generated uniformly. In between a generation and the next, a parameter is chosen randomly (differently for each PDF flavour and for each of the 80 mutants) and modified accordingly to (3.2.2). After completing the implementation of the MSTW08 parametrization in the NNPDF3.0 framework, I was ready to perform the closure test fits: that is what I'm going to discuss in the next section.

3.3 Closure tests

In the following I discuss the methodology I followed to perform the closure tests: here I outline the main steps I took so as to simplify the task of keeping track of them and to explain their motivation.

1. Firstly, I reproduced the closure tests with the neural network as fitting functional form to make sure I could correctly use the NNPDF3.0 methodology and to have a benchmark for comparison for the results of the closure tests with the Hessian parametrization;

2. secondly, I produced and validated the closure tests with the Hessian parametrization;

3. after that I compared the two sets of uncertainties, which was the main test of the idea on which this work is based, i.e. that a closure fit with a rigid parametrization would produce smaller uncertainties than those obtained from a closure fit performed with a general functional form;

4. lastly, after finding out that I did get smaller uncertainties in the case of the closure fits with the MSTW08 parametrization, I tested whether or not this was due to a difference in the average $\Delta\chi^2$.

3.3.1 Shift of starting parameters

The first step after implementing MSTW08–like functional forms and a suitable GA in the NNPDF3.0 framework was to produce the first closure tests: I used the complete NNPDF3.0 dataset (see section 2.1 of [26]) to produce a closure test with the neural network as fitting functional form and ten closure tests with the MSTW08's one. These ten closure tests are different in the choice of a couple of random seeds, which I label by s_k, $k = 1, \ldots, 10$: one of these random seeds is needed in all the functions of the NNPDF3.0 framework which generate random numbers, the other is involved in the random division of the data into a training and validation sample (as mentioned in section 1.2.2). This means that the random shifts of the Level 0 data needed to get Level 1 (2.3.1) and Level 2 (2.3.2) pseudodata, as well as the way these pseudodata are partitioned into the training and validation samples, are different for the ten closure fits. Having ten different closure test fits allowed me to compute the indicators in (2.3.4) with statistical significance, as discussed in 2.3.2: a larger number would be of course better, but performing a large number of fits is very computationally expensive. I did not produce ten closure fits as well with the neural networks as parametrization because I needed only the uncertainties resulting from closure testing, which do not vary sizably between different closure test fits. In addition, the validation procedure for these closure test fits has already been performed, see [26].

Even though the GA has the property of being able, if the mutations are large enough, to reach the minimum even in the case of complex figures of merit with many parameters, it is still possible that the

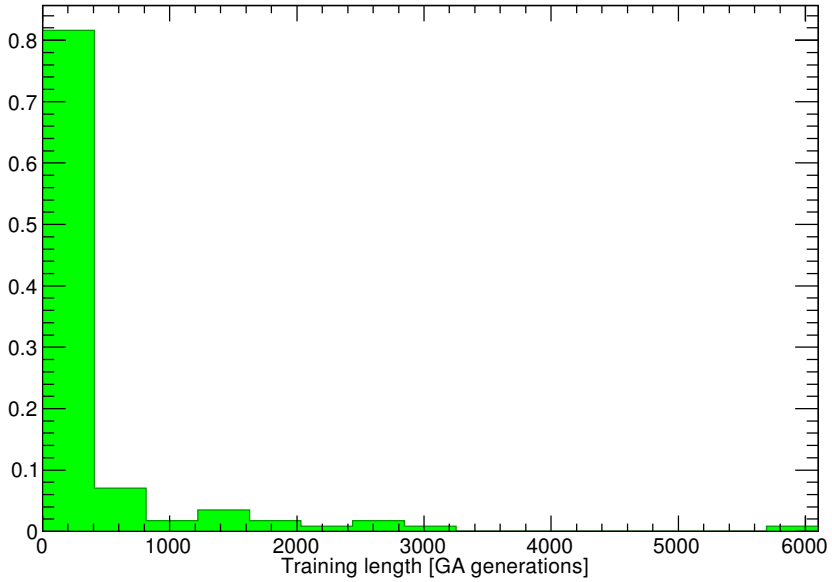

Figure 3.1: typical training length distribution of closure fits starting from the tabulate NLO best fit parameters. Such a small number (6k) of generations proved to provide the same result, in these cases, as 10 times longer fits.

outcome of the minimization procedure depends on the choice of the starting parameters. However, since the Level 2 pseudodata are obtained by shifting the "perfect" ones, I thought that starting from the NLO best—fit parameters would not particularly bias the minimization: unfortunately, this choice resulted in a complete failure. By this I mean that the best–fit PDFs resulting from Level 2 closure test with a particular choice of random seeds could fit pseudodata generated with different random seeds yielding a χ^2 equal (if not smaller) than that of the input PDF set. In practice, the PDF sets I obtained from a particular Level 2 closure test, corresponding to the choice of a particular couple of random seeds s_1, were such that, if $\{f_{s_1}\}$ are the resulting best–fit PDFs of this closure test, not only

$$\chi^2[\langle\{f_{s_1}\}\rangle, \mathcal{D}_2^{s_1}] = \chi^2[\langle\{f_{\mathrm{MSTW}}\}\rangle, \mathcal{D}_2^{s_1}]$$

where $\mathcal{D}_2^{s_1}$ are the Level 2 pseudodata obtained from the random seeds s_1, but also

$$\chi^2[\langle\{f_{s_1}\}\rangle, \mathcal{D}_2^{s_k}] = \chi^2[\langle\{f_{\mathrm{MSTW}}\}\rangle, \mathcal{D}_2^{s_k}]$$

with $s_k \neq s_1$ if $k \neq 1$. The statement that the closure test is not working is supported also by figure 3.1: the training length distribution for these fits shows that there is basically no improvement of the χ^2 as the fit proceeds. This is interpreted as the failure of the GA to lower the χ^2 by modifying the parameters in about the 80% of the cases, and as a result the output PDFs are very close to the input ones. What I did was then to modify the starting values of parameters: in case of reported uncertainty (see, as before, Table 4 in [11], central column), I would move a certain parameter from its central value of that quantity (arbitrarily adding or subtracting), otherwise I would add (or subtract) the square root of the modulus of the parameter itself.

In these closure fits, the maximum number of generations was set to 5k, as tests with 50k and 10k generations resulted in only about 10% of runs ending after 5k generations; see figure 3.2 as an example of training length distribution of these fits.

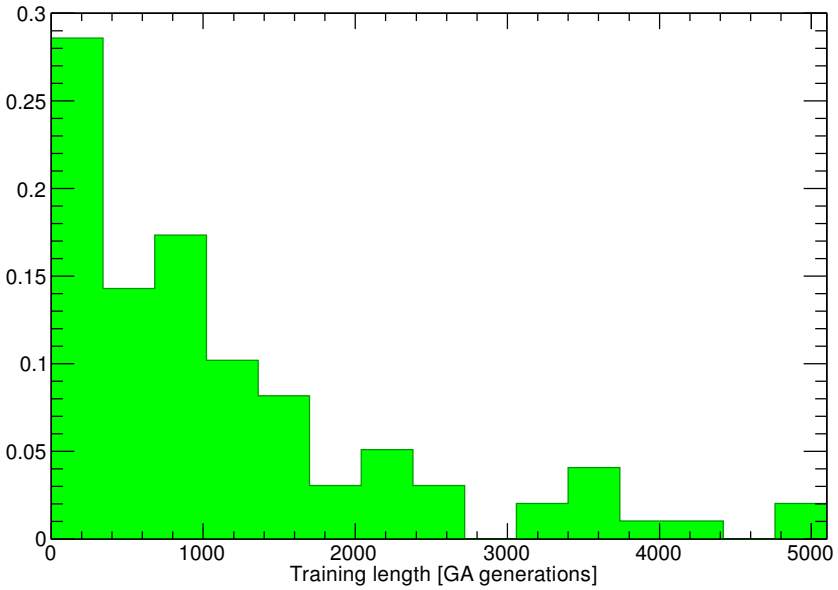

Figure 3.2: training length distribution of a Level 2 closure fit starting from the shifted parameters.

3.3.2 Validation of the closure tests

After performing the closure tests, I computed the main indicators to validate them: the results are listed in the table below.

Indicator	Level 2 closure test fit											Average
	1	2	3	4	5	6	7	8	9	10	NNPDF	(1–10)
χ^2_{central}	1.056	1.026	1.018	1.012	0.996	1.007	0.983	0.993	0.997	1.006	1.040	1.009
χ^2_{MSTW}	1.066	1.028	1.006	1.015	0.996	1.005	0.983	0.986	0.998	0.998	1.066	1.008
d^2_{ct}	0.976	1.900	0.623	0.872	0.932	0.539	0.602	0.957	0.716	0.828	0.386	0.894
$\langle \text{TL} \rangle$	4546	1973	2459	1656	1282	1784	2267	1468	1470	1159	22 097	2006
$\xi_{1\sigma}$	0.74	0.63	0.68	0.57	0.71	0.80	0.85	0.77	0.57	0.89	0.83	0.72
$\xi_{2\sigma}$	0.97	0.94	0.97	0.94	0.91	0.97	1.00	0.97	0.97	1.00	1.00	0.96
$\xi_{3\sigma}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

For each closure test fit the table reports:

- the χ^2 of the central PDF (the average of all the best–fit replicas) of both the fitted dataset (χ^2_{central}) and the MSTW08 `LHAPDF` PDF set (χ^2_{MSTW}) to the pseudodata;
- the squared value of the distance indicator defined in (2.3.5), averaged over seven (gluon and light quarks and antiquarks) flavours and over the x-points: $\{10^{-4}, 10^{-3}, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7\}$;
- the average training length $\langle \text{TL} \rangle$;
- the $\xi_{n\sigma}$ values defined in (2.3.4), averaged over the x-points: $\{0.005, 0.01, 0.05, 0.1, 0.3\}$.

The last column contains the average over the ten closure fits with the MSTW08 functional form of the various indicators. Note that only the last column contains the $\xi_{n\sigma}$ values with true statistical significance: as discussed in 2.3.2, the average over the fits is the only one that is really needed to compute the quantities in (2.3.4). However, I chose to compute these indicators also for the singular fits, to perform a qualitative

check: in order not to introduce too much correlations (and thus, bias) I averaged over a small number (five) of x–points, which have been chosen so as to cover the most important and meaningful region, i.e. the most data–constrained one. Combining the ten fits together I obtained:

$$\xi_{1\sigma} = 0.72 \qquad \xi_{2\sigma} = 0.96 \qquad \xi_{3\sigma} = 1.00$$

in satisfactory agreement with the expected values. Figure 3.3 shows an histogram with the distribution of the distances to the theory (that is, the `LHAPDF` MSTW08 PDFs) in units of sigma for the ten Level 2 closure tests combined compared to the expected standard normal distribution trend.

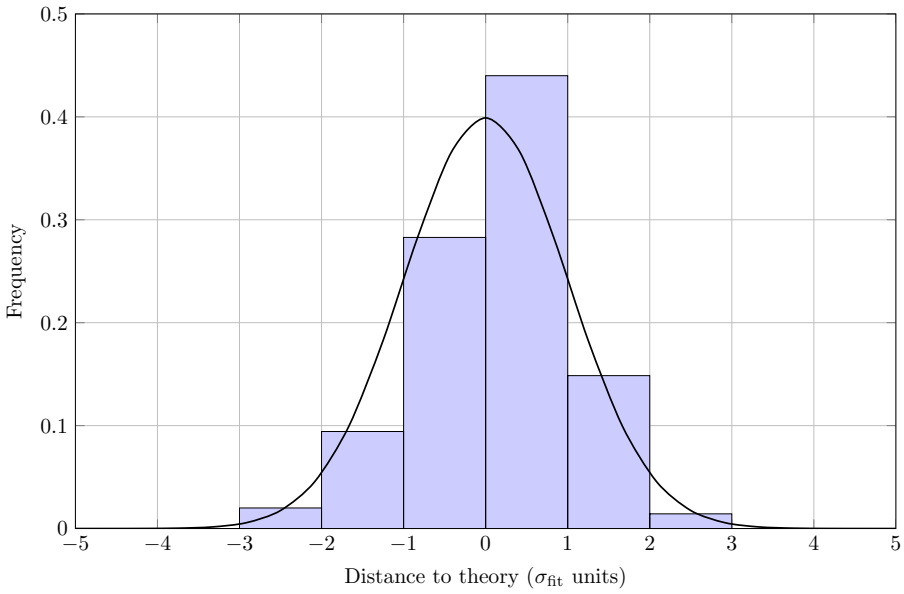

Figure 3.3: distribution of the distances to the theory in units of sigma for the ten Level 2 closure tests combined compared to a standard normal distribution.

3.3.3 Ratio of uncertainties

As said before, this work is based on the idea that the choice of a particular functional form biases the PDF uncertainties and leads to problems, as discussed in Chapter 2, ameliorated by artificially enlarging the confidence intervals and, thus, the PDF uncertainties. That is the reason why, after validating the closure tests, I computed the ratio of the uncertainties obtained from the Level 2 closure test with neural networks to those obtained from the closure fits with the Hessian functional form.

The results are shown in the following plots, in which it is clear that the closure tests were working as expected: the uncertainties of Level 0 result to be smaller than those at Level 1, which are in turn smaller than those at Level 2. In the following plots is shown the ratio of Level 0 (CT0), Level 1 (CT1) and Level 2 (CT2) closure fit (with MSTW08 functional form) uncertainties to those of the `LHAPDF` MSTW08 PDF set (σ_{MSTW}). It is also shown the ratio of the uncertainties obtained from the three Levels of closure testing with neural networks to the `LHAPDF` MSTW08 ones: NNPDFn labels the n–th Level closure test with neural networks. I chose to normalize the uncertainties obtained from the closure fits to those of the MSTW08 PDF set contained in the `LHAPDF` library: these uncertainties are of a completely different kind that those I computed, since they are obtained from fitting true experimental data and thus affected by all the sources of systematical uncertainties of the experiments. The uncertainties I computed, instead, come from closure tests in which all of these sources are eliminated; the point of this normalization is to compare the size of the uncertainties I get to that of the typical uncertainties of the true PDFs. In this way I compensate for the large fluctuations of the uncertainties and compare the behaviour of the uncertainties I get to a reasonable baseline. I show these ratios both in logarithmic (left) and linear (right) scale and for each flavour.

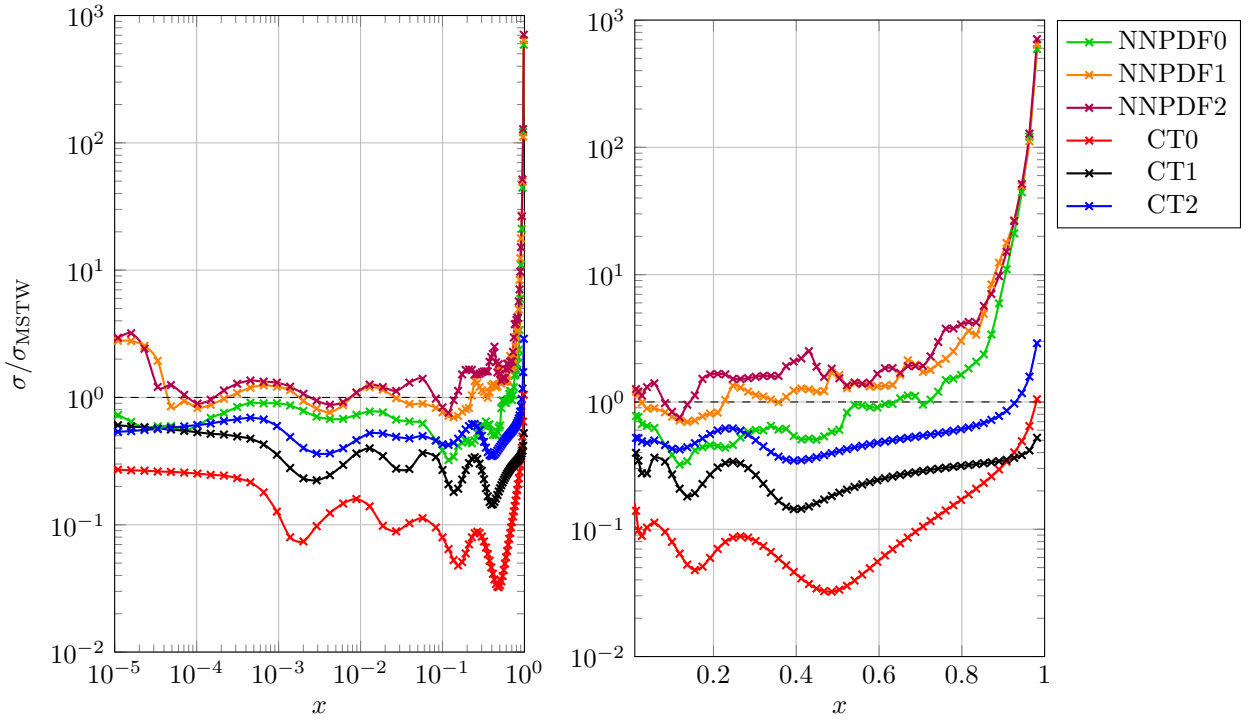

Figure 3.4: ratio of uncertainties on gluon PDFs, normalized to those of the `LHAPDF` PDF set MSTW2008nlo68cl, obtained from the closure tests I performed.

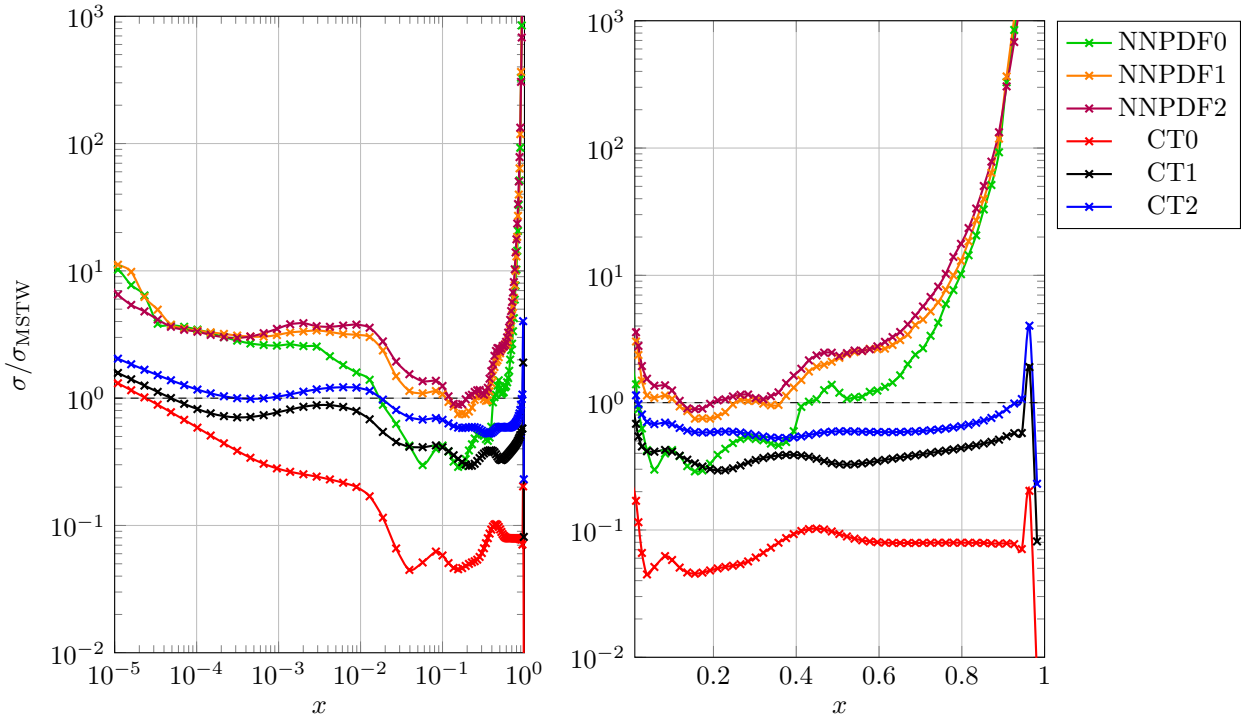

Figure 3.5: ratio of uncertainties on down PDFs, normalized to those of the `LHAPDF` PDF set MSTW2008nlo68cl, obtained from the closure tests I performed.

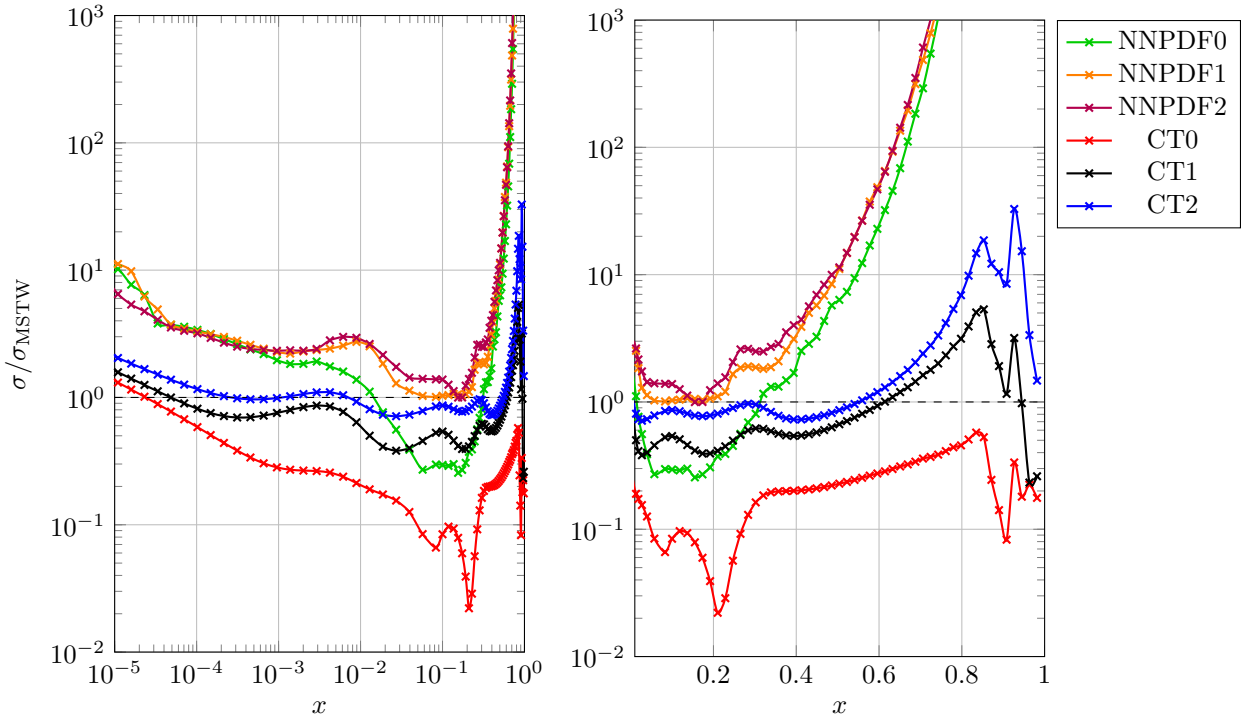

Figure 3.6: ratio of uncertainties on anti–down PDFs, normalized to those of the `LHAPDF` PDF set MSTW2008nlo68cl, obtained from the closure tests I performed.

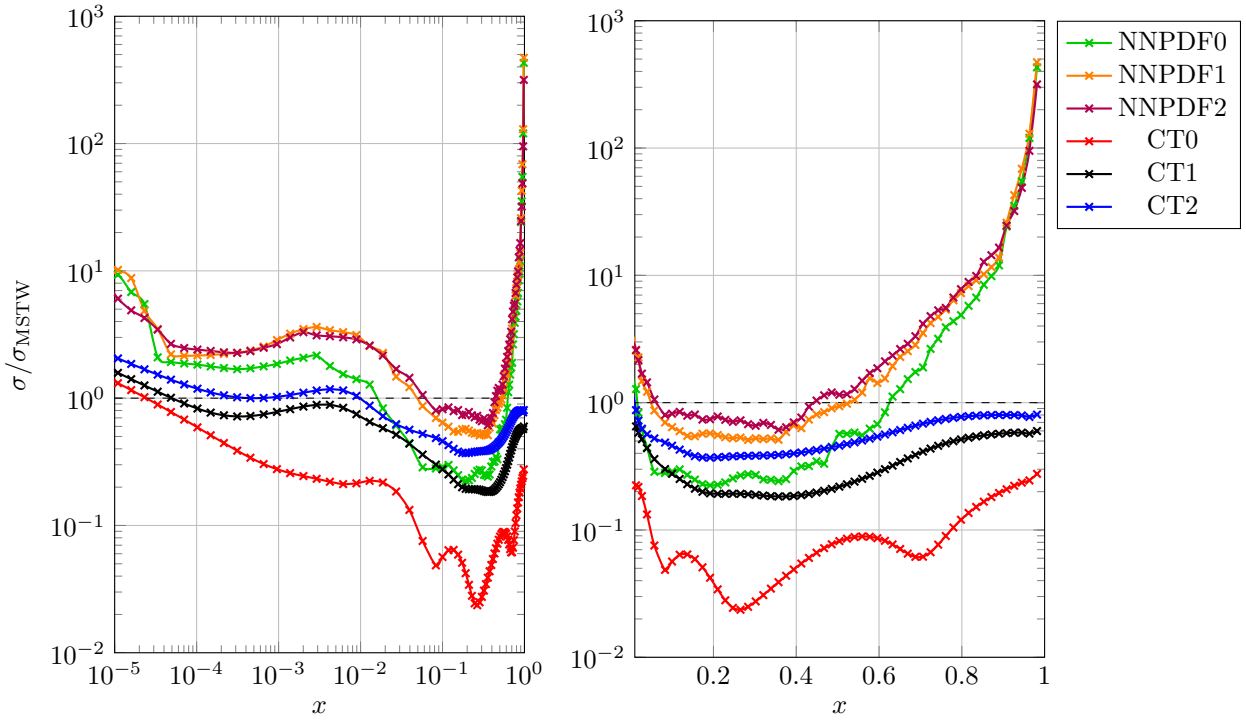

Figure 3.7: ratio of uncertainties on up PDFs, normalized to those of the `LHAPDF` PDF set MSTW2008nlo68cl, obtained from the closure tests I performed.

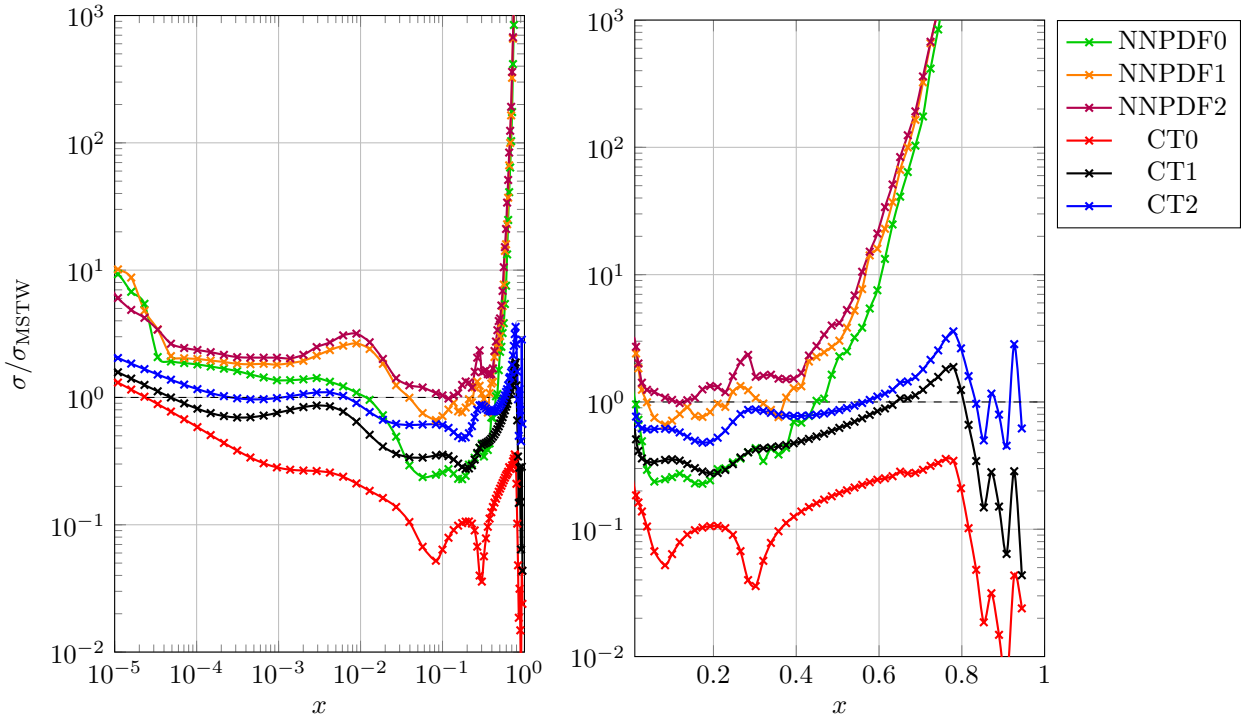

Figure 3.8: ratio of uncertainties on anti–up PDFs, normalized to those of the `LHAPDF` PDF set MSTW2008nlo68cl, obtained from the closure tests I performed.

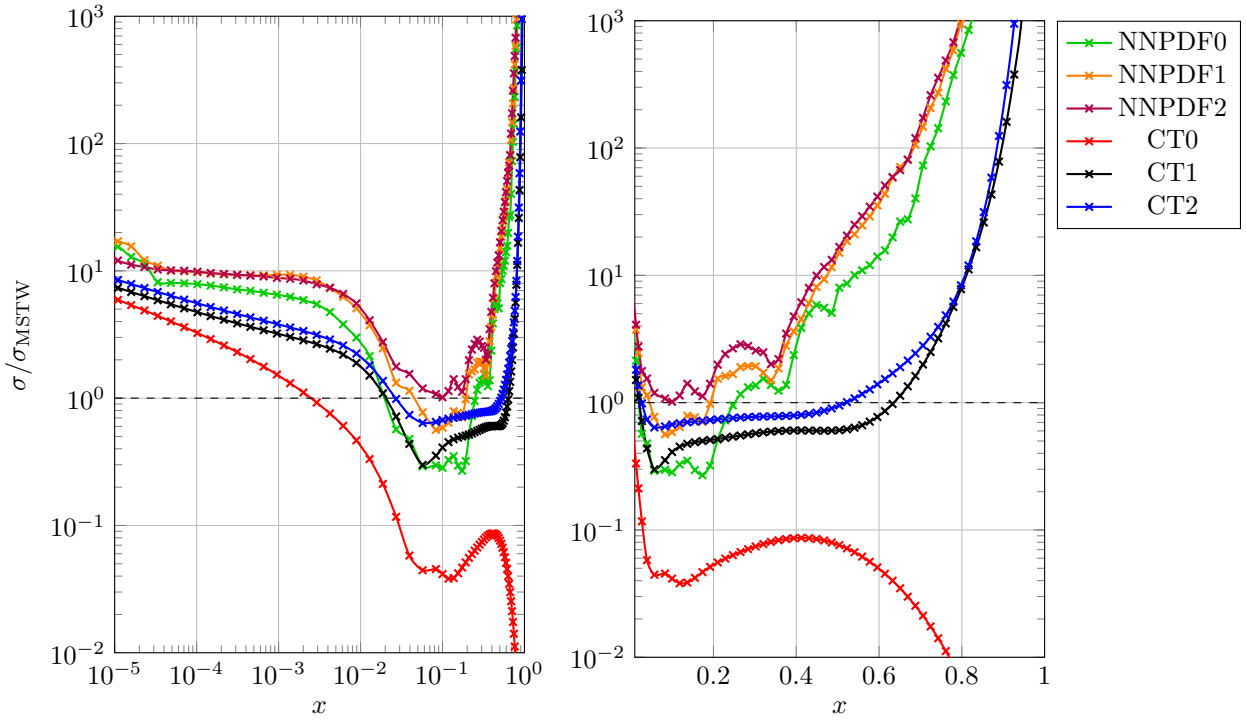

Figure 3.9: ratio of uncertainties on strange PDFs, normalized to those of the `LHAPDF` PDF set MSTW2008nlo68cl, obtained from the closure tests I performed.

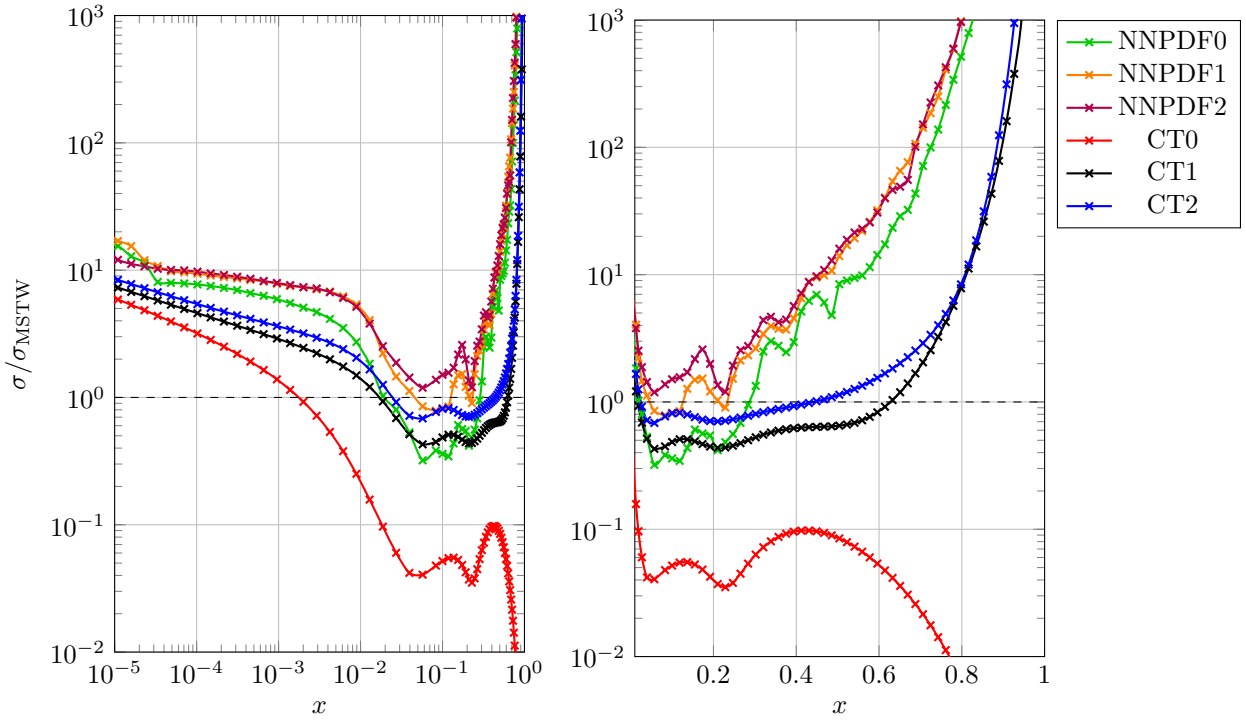

Figure 3.10: ratio of uncertainties on anti–strange PDFs, normalized to those of the `LHAPDF` PDF set MSTW2008nlo68cl, obtained from the closure tests I performed.

In the data–constrained range ($10^{-3} < x < 10^{-1}$) the uncertainties of the neural network fit are almost twice (1.9 times) the size of those of MSTW08. I computed this factor, as comparison, for the uncertainties of the NNPDF3.0 PDF set stored in the `LHAPDF` library: in this case they result to be 1.67 times those of MSTW08 in that region, compatibly enough to what I obtain with the neural network–based closure fit. It is interesting to note that, even in true fits where tolerance is accounted for, the NNPDF3.0 uncertainties are larger than MSTW08's ones, though similar.

It is evident from the plots in the figures 3.4–3.10 that the CT0 uncertainties are very small, much smaller than those of NNPDF0: it is clear that fitting a functional form to the pseudodata generated with itself will result into vanishing uncertainties. It can be noted that also at Level 1 there is a huge difference in the size of the uncertainties from MSTW08 closure fits to those obtained from NNPDF closure fits: this is another consequence of the choice of the MSTW08 functional form as a fitting function, which is not flexible enough to accommodate for functional uncertainties as well as neural networks manage to do. At Level 2, the uncertainties obtained from the closure fits with the two functional forms get closer: it is clear from the plots that the MSTW08 uncertainties grow much more than the NNPDF uncertainties. The fact that the largest part of the discrepancy between these sets of uncertainties is due to the Level 0 and Level 1, i.e. to the functional form, supports the idea on which this work is based: the choice of a rigid functional form biases the uncertainties.

To obtain quantitative results, I computed the ratio $\sigma_{\mathrm{NNPDF}}/\sigma_{\mathrm{CT2s}}$, where σ_{NNPDF} are the uncertainties obtained from the Level 2 closure test with neural networks and σ_{CT2s} are those obtained from the closure fits with the MSTW08 parametrization: I averaged this ratio over two distinct sets of points, which are

- small-x set: $\{0.003, 0.01, 0.03\}$;
- large-x set: $\{0.03, 0.1, 0.3\}$

and over the ten closure fits I performed. The results are shown in the following table:

x-range	g	u	d	s	$\overline{u}$	$\overline{d}$	$\overline{s}$	average
small-x	2.40	2.82	3.06	2.58	2.74	2.73	2.60	2.70
large-x	2.41	2.04	1.98	2.33	1.92	2.12	2.56	2.20

I kept the different flavour contributions separated to possibly highlight some pattern, which is not apparent: in fact, it can be noted that the ratio does not depend in a sizable way either on the x-range or on the PDF flavour. The tolerance factor is needed to recover the confidence intervals by enlarging the $\Delta\chi^2$ corresponding to the standard n-sigma bands, which in practice it can be seen as enlarging the sigma bands themselves; however, this uncertainty enlargement is not related to the tolerance in a clear way. However, a qualitative comparison with the tolerance values in 3.11 is possible: the uncertainties dilatation factor and the typical values of the tolerance are of similar size.

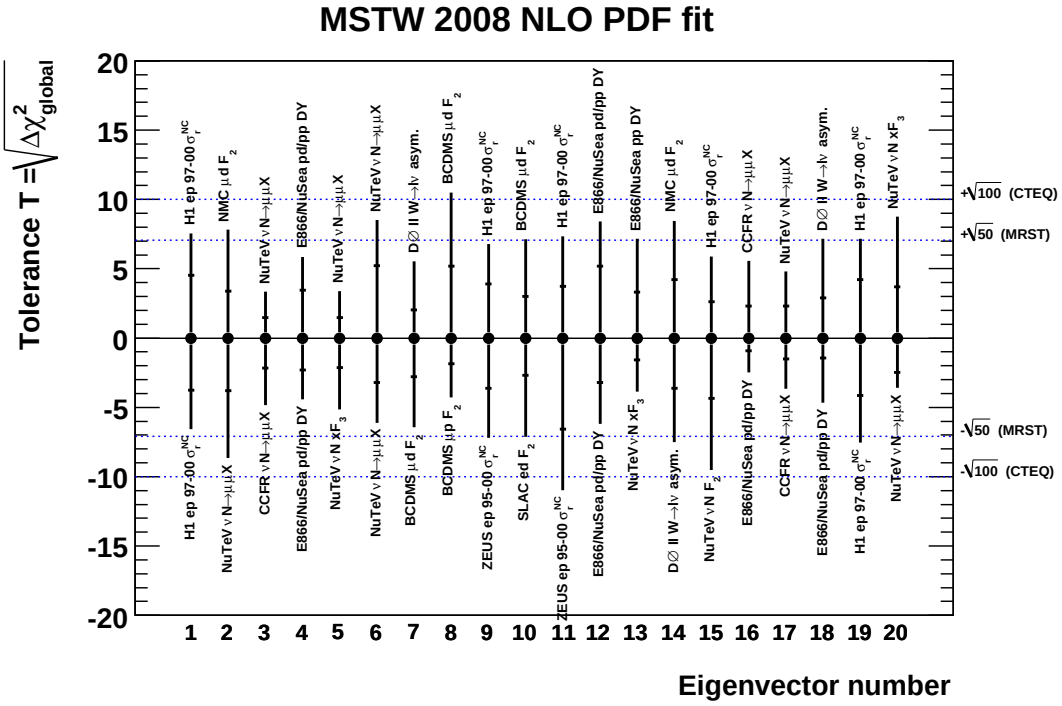

Figure 3.11: dinamically determined tolerance for each eigenvector direction: the inner error bars correspond to the 68% c.l. band, the outer ones to the 90% c.l. band. The labels give the name of the data set which sets the 90% c.l. tolerance for each eigenvector direction. [11]

3.3.4 Direct study of $\Delta\chi^2$

The fact that the uncertainties obtained from closure testing with neural networks are, on average, more than twice in size than those obtained from closure tests with the MSTW08 parametrization could be due to different values of the $\langle\Delta\chi^2\rangle$. It is expected the Monte Carlo procedure can correctly reproduce the Hessian scenario, in which, once the minimum has been reached, the hypersurface $\Delta\chi^2 = 1$ is identified as the one–sigma contour in the space of parameters around the minimum, corresponding to the 68% c.l. interval. What could happen is that the methodology I used has not reproduced the Hessian scenario: my procedure may not have converged to the minimum, producing $\Delta\chi^2$ larger than that obtained with neural networks; this serves as a sanity check of the procedure I followed. It could also be that I have correctly reproduced the Hessian procedure, with the difference in size of the uncertainties explained by a larger $\Delta\chi^2$ for the neural networks, for which there is no expected value *a priori*. A larger value of the

$\Delta\chi^2$ for the neural networks could be the direct consequence of using a general functional form along with the cross–validation methodology, while using a rigid functional form would imply $\Delta\chi^2 = 1$. I wanted to discover if the uncertainties are different despite the $\Delta\chi^2$ being the same or if they are different because the $\Delta\chi^2$ are different. Either way, the underlying cause of this difference in size of the uncertainties remains the choice of the functional form.

Thanks to the `mc2hessian` [37] framework, it was possible for me to compute the $\Delta\chi^2$ values of PDF sets in a direct way. Choosing the number of the eigenvectors of the output Hessian PDF set and the scale Q, this program can convert a Monte Carlo PDF set in a Hessian one. The steps of such methodology are, very roughly, the following:

1. a subset of replicas $\{\eta_\alpha^{(i)}\}$ ($\alpha = 2N_f + 1$, $i = 1, \ldots, N_{\text{rep}}$) of the starting Monte Carlo (MC) set is found, such that any replica $f_\alpha^{(k)}$ of the prior set can be written as a linear combination

$$f_\alpha^{(k)} \approx f_{H,\alpha}^{(k)} := f_\alpha^{(0)} + \sum_{i=1}^{N_{\text{eig}}} a_i^{(k)} \left(\eta_\alpha^{(i)} - f_\alpha^{(0)}\right)$$

   $f_\alpha^{(0)}$ is the central of the prior MC set and $a_i^{(k)}$ constant coefficients, i.e. independent of α, x, and Q: $f_{H,\alpha}^{(k)}$ is referred to as the Hessian representation of the original replica $f_\alpha^{(k)}$;

2. a covariance matrix and a figure of merit are defined in the PDF space to determine the best parameters $\{a_i^{(k)}\}$ by minimizing this figure of merit;

3. the covariance matrix in the space of these parameters is computed: the Hessian matrix corresponds to the inverse of this covariance matrix, and is diagonalized through a rotation matrix R_{ij};

4. the total one–sigma uncertainty is determined as the sum in quadrature of all the one–sigma uncertainty bands associated to each orthogonal direction:

$$\sigma_{H,\alpha}(x, Q^2) = \sqrt{\sum_{i=1}^{N_{\text{eig}}} \left[\sum_{j=1}^{N_{\text{eig}}} \frac{R_{ij}}{\sqrt{\lambda_i}} \left(\eta_\alpha^{(j)}(x, Q^2) - f_\alpha^{(0)}(x, Q^2)\right)\right]^2}$$

   where λ_n is the n–th eigenvalue of the Hessian matrix;

5. N_{eig} symmetric eigenvectors are defined as:

$$h_\alpha^{(i)}(x, Q^2) = f_\alpha^{(0)}(x, Q^2) + \sum_{j=1}^{N_{\text{eig}}} \frac{R_{ij}}{\sqrt{\lambda_i}} \left(\eta_\alpha^{(j)}(x, Q^2) - f_\alpha^{(0)}(x, Q^2)\right)$$

   so that the output Hessian set consists in all the $\{h_\alpha^{(i)}\}$ and the original central $f_\alpha^{(0)}$.

For more details and caveats refer to [37]: I verified is that $N_{\text{eig}} = 40$ was enough to faithfully reproduce the input PDFs and to compute the $\Delta\chi^2$ for each PDF set and, as always, I worked at the reference scale of $Q^2 = 1\,\text{GeV}^2$. The $\langle\Delta\chi^2\rangle$ corresponding to the 68% c.l. is computed as:

$$\langle\Delta\chi^2\rangle := \frac{1}{N_{\text{eig}}} \sum_{i=1}^{N_{\text{eig}}} \left|\chi^2_{(0)} - \chi^2_{(i)}\right| \qquad \chi^2_{(i)} := \frac{1}{2N_f + 1} \sum_{\alpha=1}^{2N_f+1} \chi^2[h_\alpha^{(i)}]$$

In the following table I report the outcome of this analysis, for each of the Level 2 closure fits I performed, including the one with the neural networks functional form.

Quantity	Hessianized Level 2 closure fit											Average
	1	2	3	4	5	6	7	8	9	10	NNPDF	
$\langle\Delta\chi^2\rangle$	14.17	20.94	18.49	15.12	16.74	15.11	17.61	15.68	16.15	14.17	16.09	16.39
$\sigma_{\langle\Delta\chi^2\rangle}$	15.43	17.09	14.27	12.81	16.59	14.54	16.07	13.71	13.02	12.13	14.91	14.69

On the one hand, it is surprising that for every of the closure fits based on the MSTW08 functional forms $\langle \Delta\chi^2 \rangle \approx 16 \gg 1$: statistical arguments lead us to expect, if linear error propagation holds (i.e. if the minimum is reached), that the 68% c.l. band should correspond to $\Delta\chi^2 \approx 1$. On the other hand, such value for the $\langle \Delta\chi^2 \rangle$ occurs also in the case of the closure fit performed with neural networks: this leads to the conclusion that the difference between uncertainties is not related to different $\langle \Delta\chi^2 \rangle$ values, rather on the parametrization choice alone.

The cause of this large value of the $\Delta\chi^2$ in the case of closure tests with Hessian parametrization could be that the Level 2 fluctuations on data cause the failure of linear approximation in a neighbourhood of the minimum: if this were true, reducing the size of this data fluctuations could recover $\Delta\chi^2 = 1$, but the sizes of the two sets of uncertainties would still be different.

Conclusions and outlook

The starting point of this work was the idea that the source of the problems in the Hessian procedure of computing PDFs and their uncertainties lies in the choice of a fixed parametrization. Such a choice would not take into account some sources of uncertainty and, thus, lead to a biased computation of uncertainties, which turn out to be underestimated. The closure testing methodology, developed by the NNPDF group, provides a very clean and controlled environment in which to test for the introduction of bias on behalf of a few–parameters functional form.

That is where my work started: I modified the NNPDF3.0 framework to accommodate for a Hessian functional form, which I chose to be the MSTW08 one, so that I could perform closure tests with this parametrization. I performed then ten different closure tests with this functional form and one with neural networks, and validated them by computing some relevant indicators.

After that, I addressed the core of the problem, comparing the uncertainties of a Level 2 closure fit with neural networks to those of my Level 2 closure fits with MSTW08–like parametrizations. The uncertainties obtained from the closure fits performed with the MSTW08 functional form result to be about 2.5 times smaller than those obtained from the neural network–based closure fits, and show no particular dependence either on flavour or on x-region – as long as it is near to the data constrained range. Moreover, comparison of the size of MSTW08 uncertainties in the three Levels of closure testing confirms the role of the functional form as a bias. A qualitative comparison of this uncertainties ratio to the typical size of the tolerance shows compatibility between these quantities.

The value of this ratio could be due to a difference in the $\Delta\chi^2$ of the two closure fits. In fact, while the $\Delta\chi^2$ is expected to be about 1 for closure fits with the Hessian parametrization, there is no particular value of this quantity to be expected *a priori* in the case of closure fits based on neural networks, and a larger $\Delta\chi^2$ (relative to the 68% c.l. band) leads to larger uncertainties. Thanks to the `mc2hessian` framework I was able to compute the average $\Delta\chi^2$ in both cases, and verified that I obtained the same value: 16. This allowed me to conclude that the choice of a rigid functional form reduces the uncertainties even without introducing differences in the $\Delta\chi^2$.

The size of the $\Delta\chi^2$ for the Hessian closure fit is, however, somewhat puzzling: in the future, it would be interesting to investigate the source of so large a value. If linear approximation holds, it is expected that the one–sigma contour is defined by $\Delta\chi^2 \approx 1$: can it then be that linear approximation fails? A way to understand this would be to sizably reduce the fluctuations on data introduced in the Level 2 of closure tests, but, unfortunately, there was not enough time for me to answer that question. What can be supposed is that, even in the case of smaller fluctuations on data, the sizes of the two sets of uncertainties of the rigidly–parametrized PDF set woud still be smaller than those of the more generally parametrized one.

Another exercise that could be performed to complete this work is the following: to perform a closure test with a rigid functional form to generate the pseudodata and with a different rigid functional form to fit them. It is expected that, in this case, the uncertainties are, once again, smaller than those that would be obtained from a closure test with neural networks, and that, in this case, the validation procedure results in failure.

Bibliography

[1] S. Forte and G. Watt, *Progress in the Determination of the Partonic Structure of the Proton*, Annual Review of Nuclear and Particle Science **63** (2013) 291–328.

[2] G. Altarelli and G. Parisi et al., *Asymptotic Freedom in Parton Language*, Nuclear Physics B **126** (1997) 298–318.

[3] Yuri L. Dokshitzer, *Calculation of the Structure Functions for Deep Inelastic Scattering and e^+e^- Annihilation by Perturbation Theory in Quantum Chromodynamics*, Sov. Phys. JETP **46** (1977) 641–653.

[4] V. N. Gribov and L. N. Lipatov, *Deep inelastic ep Scattering in Perturbation Theory*, Sov. J. Nucl. Phys. **15** (1972) 438–450.

[5] The **PDF4LHC** Working Group, *The PDF4LHC Working Group Interim Report*, (2011).

[6] R. Placakyte, *Parton Distribution Functions*, (2011).

[7] A. D. Martin, W. J. Stirling, R. S. Thorne, G. Watt, *Parton distributions for the LHC*, Eur. Phys. J. C **63** (2009) 189–285.

[8] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo and M. Ubiali, *A first unbiased global NLO determination of parton distributions and their uncertainties*, Nucl. Phys. B **838** (2010) 136–206.

[9] P. M. Nadolsky et al, *Implications of CTEQ global analysis for collider observables*, Phys. Rev. D **78** 013004 (2008).

[10] W. T. Giele, S. A. Keller and D. A. Kosower, *Parton distribution function uncertainties*, (2001).

[11] A. D. Martin, W. J. Stirling, R. S. Thorne and G. Watt, *Parton distributions for the LHC*, Eur. Phys. J. C **63** (2009) 189–285.

[12] S. Brodsky, P. Hoyer, C. Peterson, and N. Sakai, *The Intrinsic Charm of the Proton*, Phys. Lett. B **93** (1980) 451–455.

[13] J. Pumplin, H. Lai, and W. K. Tung, *The Charm Parton Content of the Nucleon*, Phys. Rev. D **75** 054029 (2007).

[14] G. Cowan, *Statistical data analysis*, Oxford Univ. Press, Oxford UK (1998).

[15] S. Forte, *Parton distributions at the dawn of the LHC*, Acta Phys. Polon. B **41** (2010) 2859–2920.

[16] M. Gluck and E. Reya, *Operator mixing and scaling deviations in asymptotically free field theories*, Phys. Rev. D **14** (1976) 3034–3070.

[17] P. W. Johnson and W. K. Tung, *Comparison of asymptotically free theories with high-energy deep inelastic scattering data*, Nucl. Phys. B **121** (1977) 270–276.

[18] S. Forte, L. Garrido, J. I. Latorre and A. Piccione, *Neural network parametrization of deep-inelastic structure functions*, JHEP 0205 **062** (2002).

[19] J. Feltesse, A. Glazov and V. Radescu, *Experimental error propagation*, Sect 3.2 in M. Dittmar et al., *Parton Distributions*, (2009).

[20] A. Glazov, S. Moch and V. Radescu, *Parton distribution uncertainties using smoothness prior*, Phys. Lett. B **695** issues 1–4 (2011) 238–241.

[21] S. J. Brodsky and G. R. Farrar, *Scaling laws at large transverse momentum*, Phys. Rev. Lett. **31** (1973) 1153–1156.

[22] H. D. I. Abarbanel, M. L. Goldberger and S. B. Treiman, *Asymptotic properties of electroproduction structure functions*, Phys. Rev. Lett. **22** (1969) 500–502.

[23] G. D'Agostini, *On the use of the covariance matrix to fit correlated data*, Nucl. Instrum. Meth. A **346** (1994) 306–311.

[24] H. L. Lai, J. Huston, Z. Li, P. Nadolsky, J. Pumplin, D. Stump and C. P. Yuan, *Uncertainty induced by QCD coupling in the CTEQ global analysis of parton distributions*, Phys. Rev. D **82** 054021 (2010).

[25] J. C. Collins and J. Pumplin, *Tests of goodness of fit to multiple data sets*, (2001).

[26] The NNPDF Collaboration, *Parton distributions for the LHC run II*, JHEP **040** (2015).

[27] M. Ubiali, Talk at "QCD at the LHC", Trento (2010); `http://indico.cern.ch/materialDisplay.py?contribId=49&sessionId=19&materialId=slides&confId=93790`

[28] The NNPDF Collaboration, *Precision determination of electroweak parameters and the strange content of the proton from neutrino deep-inelastic scattering*, Nucl. Phys. B **823** (2009) 195–233.

[29] J. Pumplin, D. R. Stump, J. Huston, H. L. Lai, P. M. Nadolsky and W. K. Tung, *New generation of parton distributions with uncertainties from global QCD analysis*, JHEP **0207** 012 (2002).

[30] A. D. Martin, R. G. Roberts, W. J. Stirling and R. S. Thorne, *Uncertainties of predictions from parton distributions. 1: Experimental errors*, Eur. Phys. J. C **28** (2003) 455–473.

[31] The NNPDF Collaboration, *A determination of parton distributions with faithful uncertainty estimation*, Nucl. Phys. B **809** (2009) 1–63 [Erratum-ibidem B **816** (2009) 293–356].

[32] J. Pumplin, *Experimental consistency in parton distribution fitting*, Phys. Rev. D **81** 074010 (2010).

[33] J. Pumplin, *Data set diagonalization in a global fit*, Phys. Rev. D **80** 034002 (2009).

[34] J. Pumplin, *Parametrization dependence and $\Delta\chi^2$ in parton distribution fitting*, Phys. Rev. D *82* 114020 (2010).

[35] L. Demortier, *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*, January 17–20, CERN, Geneva, (2011) see chapter *Open Issues in the Wake of Banff 2011.*

[36] A. Buckley, J. Ferrando, S. Lloyd, K. Nordstrom, B. Page, M. Ruefenacht, M. Schoenherr, G. Watt, *LHAPDF6: parton density access in the LHC precision era*, Eur. Phys. J. C75 3 (2015) 132–152.

[37] S. Carrazza, S. Forte, Z. Kassabov, J. I. Latorre, J. Rojo, *An Unbiased Hessian Representation for Monte Carlo PDFs*, Eur. Phys. J. C75 8 (2015) 1–20.

Acknowledgements

I would like to thank Stefano Forte, Stefano Carrazza and Zahari Kassabov for giving me the opportunity to work on this project and for helping me through the many difficulties I faced. Without you, I surely would not have learned this much about programming and about such a niche subject. I need to deeply thank prof. Forte, along with prof. Meroni, also for helping me find a future in the field of research.

There are no words which could ever be adequate enough to say thank you to my family. Without you, I would not be here: without your support, love and generosity I would never have achieved a result so important for me – even if with some delay. I still remember talking with mum about how the second grade of middle school would be difficult and how I did not feel quite ready for it. I still remember starting my physics studies in October 2010 with a clear sense of inadequacy and uncertainty – speaking of which, many thanks to prof. Andrea Belli for inspiring me and setting the bar so high. This may give a measure of the impact this achievement has on me: and it's all thanks to you, dad, mum and brother.

Of course all the professors and friends I met in these six years contributed to this achievement even by simply being there. I really do not know where to start, so I will try to go in chronological order: first comes Balizgrator, fellow of studies for ten years. You have been a rival, an ispiration and a friend – more or less in this order: keep doing what you do best, that is being a genius silly boy. The first guys I stuck with at university were Penny, Gebbiù, Rosy and Caccia: thanks for being friends since day one. The first year marked the start of a 4–year traidition, that is Tchoukball, and Tchoukball = Enrico: thank you, furry small man, for all your IT support during this thesis, along with Frà Dambrosio, and for sharing a huge number of experiences. The Tchoukball experience allowed me to get to know other fellow physicists, such as Da Col, Cristina, Baso, and super cool guys as La De 1/2, Irene, Bobo Merenda, Ruzzo, Ale & Pinna, Giò & Anna Zinetti: thanks everybody, I hold dear all the memories related to you.

I do not clearly remember when during the second semester of the first year I got to speak with some new guys, but I guess it was then that I met Pier and Federica, my two besties. My whole experience here would not have been the same without you: you were there in the hard times and in the good ones...I really do not know how to thank you, I hope we will stick together for ever. Then I got to know some of the cleverest guys I ever met: Fresta, Ghera, Tommy and Giuly (Bl4ck). You helped me grow a lot by discussing about physics and talking to me in general. Big shout out to my fellow particle physicists: Dia, Silvia, Baso, Ettore and Sara. You are fantastic guys and it was a pleasure for me to study with you. Special thanks and much love to SaraCerioli, she is one of the closest friends I have – and also she has the biggest (and most generous) father ever. Thanks also to Eleonora Mari and Eliana, kindest small women, Vales, the dual–delay mistress, Michelina, fellow swimmer, Enrico the Blond, doesitdj0nt?, Valvosauro, omg you are Legolas, Darione, maybe the only physicist with a voice louder than mine, Nico and Sguarao, amazing guys and superb musicians, CotroSilvia, fellow bodybuilder, Salvo Nome Danielao Carletto, genius imbruttiti, FedeMotti, poor lab fellow, Giorzio and Tommy: thank you all. Thanks also to the fallen comrades, who for some reason, decided that physics would not be their future: thanks to Cate, Jack and Riccà.

Special thanks to the Ce'sSO(3) '90 group -Treso, Beppe, Ragu, Fede, Chiara, Giulia, Cecia, Paolo, Fava, Jack, Buccio and Piottino- for letting me stick with them even though a year younger. Thanks to the

’92 group -Guido, Marghe, Minnie, Ben, Matteo, Alessio, Mattia, Luca, Nico, Lorenzo, Claudio- as well, for adopting me when I was left alone by my graduating fellows. Special mention for Marghe: you have been a friend and a disciple for a while, and now I feel like you need no more teaching. Something similar could be said about Laura, my favourite Bocconian: I am happy to see you fully exploiting your potential, thanks for all the laughs together. How could I not talk about my Erasmus experience? Thanks to Flo, Lili and Alejandro for being the best housemates I could hope for; thanks to Elmar, Karo and Sabine for being such nice and easy–going German people. It might be that my future is in Germany, in which case I will more than gladly see you again. Huge shout out to all the acquaintances I made in that period, with special mention for Ele, Anna, Linda, Federica. (the dot is part of your name now), Sere, Fabio, Kappei, Massimiliano and Fillo. Also big thanks to Jacopino, with whom I shared all the joy and sorrow of such an experience: we managed to get along with each other better than expected, I was lucky to have you as a fellow Erasmus friend.

For some reason my brain did not put you in the physicists’ group, dear Neiv, but I guess you really stand out in your own special way. Your friendship helped me learn things which were completely unknown to me. Thanks for all the night outs, your hospitality, the cazzate and for introducing me to the pale and for being there in general. Also thanks for letting me know people of the caliber of Emma and Pavel: in very different ways, I hold you very dear.

Last but not least, thank you Ilaria: your love, kindness and support make me feel the most important person on Earth. I’m really happy and fortunate to have you by my side.

P.S.: I’m really sorry to have forgotten some people (I know I must have forgotten somebody), it was not on purpose. Carlo out.