UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

Laurea Magistrale in Fisica

Optimized regression models for parton distribution functions determination using deep learning methods

Relatore:
Dott. Stefano Carrazza
Correlatore:
Dott. Juan Cruz-Martinez

Tesi di Laurea di:
Nicola Lambri
Matricola: 922003

Anno Accademico 2019/2020

To my parents.

Abstract

In the last two decades, machine learning algorithms have experienced a consistent increase in their use in particle physics research. The NNPDF collaboration has found great success in exploiting neural networks (NNs) to extract the Parton Distribution Functions (PDFs) from experimental data. The recent developments of more efficient libraries for training and hyperoptimization of neural networks have led to the reimplementation of the NNPDF framework in a new code, named `n3fit`. In this thesis we present a first analysis based on the Hessian representation of the Monte Carlo PDF sets obtained with this new methodology. We compare the predictions of the `n3fit` code with the latest release, NNPDF3.1, and quantify their goodness-of-fit from the study of a measure of fit quality, χ^2. The χ^2 values allow us to extract useful informations about potential inefficiencies in the fitting methodologies. This kind of analysis brings us to consider an experimental branch of `n3fit`, to repeat the χ^2 study and search for further improvements in the fitting procedure. We conclude with the extrapolation of an effective tolerance parameter from the Hessian conversions of the Monte Carlo sets obtained with these methodologies, from which we are able to determine the accuracy of their predictions. We finally suggest how the strategies adopted in this thesis could be included in future fits of the `n3fit` code.

Acknowledgments

I wish to express my sincere gratitude to my supervisor, Doctor Stefano Carrazza, who guided me with extreme patience and kindness through this long journey. I am immensely thankful for his persistent help and useful advice even during these unfortunate times.

I am infinitely grateful to my parents, who encouraged me since the very first day I took this path and helped me pursue my dreams with unconditional love. Their continuous support in all the aspects of my life is the most valuable gift I could receive.

My most sweet special thanks to Isabella, who has become so important to me in this last year. Without her, I am sure this work would have been a lot more difficult to accomplish.

Nicola Lambri

Contents

Introduction

Nowadays, the main focus of the research in theoretical particle physics is to provide predictions, based on the Standard Model (SM) of elementary particles, with enough precision that even a small deviation in the experimental measurements can be identified as a signal of new physics. The most important benchmarks for the SM are proton-proton collisions at the Large Hadron Collider (LHC), where two proton beams are accelerated at very high energies before they are made to collide. During these collisions, their fundamental constituents, quarks and gluons, interact to produce complicated multi-particle final states. The computation of any particle physics observable is then strictly related to a faithful description of the inner structure of the proton, which in turn cannot be obtained from first principles from the underlying theory of strong interactions, Quantum Chromodynamics (QCD). The proton structure can only be explained with a probabilistic interpretation in terms of Parton Distribution Functions (PDFs), probability densities describing the momentum distribution of quarks and gluons in the initial stage of a collision. The PDFs carry information on the non-perturbative structure of the proton, or hadrons in general, outside the domain of applicability of perturbative QCD.

The only way to determine the parton distributions is from an indirect analysis, based on comparing PDF-dependent predictions with experimental data. However, the problem gets complicated by the fact that PDFs are functions rather than simple parameters, and thus their extraction from a finite set of data will always result in some level of uncertainty. Nonetheless, the techniques exploited to determine their functional forms have greatly evolved from the first naive models of the '80s, where PDFs were parametrized by ad-hoc models and the uncertainties could not be estimated [1]. At present, the PDFs are determined from well established fitting procedures which rely on precise theoretical predictions, and take into account both the uncertainties of the input data and those related to the choice made for the parametrization of the PDFs. All this complex machinery is the result of the development of new methods to extract the PDFs and estimate their uncertainties, along with the continuous increase in computing power and efficiency.

During the last decade the NNPDF collaboration [2] (Neural Network PDF) has reached a new state-of-the-art in PDF determination with the introduction of machine learning tools for PDF analyses. In particular, NNPDF uses artificial neural networks to parametrize and find the best PDF estimate, while the uncertainties are propagated directly from the experimental data during the fitting procedure. A further step towards a new generation of PDFs predictions is currently under investigation within the N3PDF project [3] of the NNPDF

collaboration. Since neural networks themselves are not unique, and neither the algorithms used for their training, there is still some freedom in the choice of methodology. At present, the NNPDF framework has been reimplemented in a new code, named `n3fit` [4], capable to perform a semi-automatic hyperoptimization, that is, find the best combination of hyperparameters (neural network architecture, activation function, minimizer, etc.) given a specific input setup.

In this thesis we compare the predictions of the old methodology, namely NNPDF3.1, with the new one, `n3fit`, to quantify their goodness-of-fit. While NNPDF implements a "Monte Carlo" representation of the parton densities, we convert them to an equivalent "Hessian" representation, from which we can study a suitable figure of merit, χ^2. In a pure Hessian representation, the PDFs are parametrized with a fixed functional form and the best fit parameters are found from the minimization of the χ^2. The uncertainties are propagated by varying the parameters around the minimum, along directions specified by the eigenvectors of the χ^2 Hessian matrix. The confidence interval for each parameter is defined by the condition $\Delta\chi^2 = T^2$, where T is the so-called tolerance parameter. Therefore, the value of T is directly related to the uncertainties of the PDFs as it defines the region of acceptable fits. By converting a Monte Carlo set into a Hessian set, we obtain a further method to quantify the performance of the methodologies under examination, as we are able to introduce an effective tolerance for Monte Carlo sets.

Moreover, from the χ^2 study we can separate the contributions related to the inefficiencies of the methodologies, and search for further improvements in the `n3fit` procedure, thanks to the great flexibility of the new code. This analysis leads us to consider an experimental branch of `n3fit`, named `feature_scaling_test`, based on a different parametrization for the PDFs and a different treatment of the neural network input. From the Hessian representations of the Monte Carlo sets obtained with these three methodologies, NNPDF3.1, `n3fit`, and feature scaling, we estimate the corresponding tolerances to eventually conclude which one gives the most accurate predictions. We conclude with a short outlook about how this whole procedure can be included in future fits for potential improvements in the determination of parton distribution functions.

This thesis is organized as follows:

Chapter 1: The theoretical framework. We give an overview of QCD with particular attention on how the PDFs arise and their fundamental role for theoretical predictions. We begin from the concept of running coupling and the parton model description of Deep Inelasting Scattering. Then, we consider the next-to-leading order corrections to introduce the factorization theorem and its generalization to inclusive cross sections. We arrive at the DGLAP evolution equations with the treatment of heavy quarks, and conclude with the general properties expected for the proton PDFs.

Chapter 2: PDFs determination. PDFs predictions are the result of a complex fitting procedure which must take into account very different aspects: the selection of experimental data, the choice of parametrization for the PDFs to compute theoretical predictions during the fit, the minimization strategy to optimize the parameters that

describe their functional forms, and the method to propagate the uncertainties of the resulting parton distributions. In this chapter we give an overview of all these features and we focus on the two main methods used to represent the uncertainties: the Monte Carlo and Hessian method. We also introduce the Monte Carlo to Hessian conversion implemented in the `mc2hessian` code [5, 6], which is used to study the Monte Carlo sets considered in this thesis.

Chapter 3: PDFs from deep learning methods. We introduce in more detail the new `n3fit` code and argument how it can improve the old NNPDF framework. Then, we start our analysis from two "equivalent" Monte Carlo sets produced by these two methodologies, and we convert them into Hessian sets to study the χ^2 variations around the best PDF estimate. We consider a possible strategy, based on positive and negative variations of the χ^2, to determine kinematical regions where potential inefficiencies in the determination of the PDFs might appear. Guided by these observations we choose an experimental branch of `n3fit`, named `feature_scaling_test`, and repeat the same analysis to search for further improvements for the `n3fit` methodology.

Chapter 4: Tolerance for Monte Carlo sets. We continue our assessment of the Monte Carlo sets by considering a simple one-parameter model for the χ^2, which allows us to isolate the various contributions that define its shape. In particular, we investigate which set best converges to the model predictions. Finally, we introduce an effective tolerance for the equivalent Monte Carlo sets obtained from the three different methodologies under examination, to quantify the accuracy of their predictions.

Chapter 5: Conclusions and outlook. We conclude with a brief summary of the results obtained in this thesis. Moreover, we provide an outlook about further improvements that can be obtained in future fits from the inclusion of the strategies adopted in this thesis within the `n3fit` code.

Chapter 1

The theoretical framework

In this chapter we review the basics of QCD, starting from the running coupling to arrive at the derivation of PDFs and DGLAP evolution.

1.1 Fundamentals of QCD

The basic theory which describes the strong interactions is *Quantum Chromodynamics*, or QCD, a non-Abelian gauge theory with gauge group $SU(3)$. The fundamental constituents are *quarks*, fermions of fractional elementary charge (either $-1/3$ or $+2/3$), and *gluons*, which are massless gauge bosons. Quarks are grouped into three *families*, each one containing two of them with their corresponding antiquarks. These six types of quarks are often referred to as *flavours*.

The classical Lagrangian is given by the Yang-Mills Lagrangian density

$$\mathcal{L}_{\text{classical}} = \sum_{\text{flavours}} \bar{\psi}_a (i \not{D} - m)_{ab} \psi_b - \frac{1}{4} F^a_{\mu\nu} F^{\mu\nu a} \,. \tag{1.1}$$

These terms describe the interactions of quark fields ψ_a in the fundamental representation of the SU(3) *colour* group ($a = 1, 2, 3$), and gluons, whose fields lie inside $\not{D}$ and $F^a_{\mu\nu}$. Specifically, the latter term is the antisymmetric strength tensor derived from the gluon field A^a_μ,

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g f^{abc} A^b_\mu A^c_\nu \,, \tag{1.2}$$

where f^{abc} are the structure constants of $SU(3)$, and the indices a, b, c run over the eight colour degrees of freedom in the adjoint representation to which the gluon fields belong. Finally the term $\not{D}$ in eq. (1.1) is the contraction $\gamma_\mu D^\mu$ between the Dirac matrices and the covariant derivative acting on the quark fields

$$D^\mu = \partial^\mu + i g t^a A^{\mu a} \,, \tag{1.3}$$

where t^a are the eight generators of $SU(3)$ in the fundamental representation. In eqs. (1.2) and (1.3) g is the *bare* coupling of the theory which determines the strength of the interaction.

When the Lagrangian eq. (1.1) is quantized, additional terms related to gauge invariance are introduced and the Feynman rules of the theory can be derived. These rules allow us to calculate transition amplitudes as a perturbative series in the bare coupling g, or equivalently as a series in the coupling constant

$$\alpha_S = \frac{g^2}{4\pi} \,. \tag{1.4}$$

To make meaningful predictions, α_S should be small enough to guarantee the convergence of the perturbative series, and as we will see in the next section this is what happens for sufficiently high energy.

1.1.1 The running coupling of QCD

Typically in a quantum field theory, when calculating amplitudes using Feynman diagrams, infinite values arise whenever are present loops of virtual particles whose momenta are not bound (UV divergences).

Thanks to the general procedure called *renormalization*, these unphysical divergences can be reabsorbed by rescaling the fields and introducing renormalized (physical) parameters, such as a renormalized coupling. These new parameters are determined by a set of *renormalization conditions* applied at a certain momentum scale, or *renormalization scale*, μ_R. The renormalized parameters are then dependent on this quantity, and on the *scheme* adopted for the renormalization conditions.

The new scale μ_R was not mentioned in the starting Lagrangian, but its choice is required in order to define the theory at the quantum level. Furthermore it is arbitrary: we could choose another value μ'_R, which would change the renormalized parameters, but the predictions for the physical observables would remain the same. This means that the μ_R dependence of renormalized quantities must cancel out inside the expressions of physical observables. The mathematical formulation of the previous statement is given by the Callan-Symanzik equation [7, 8].

QCD is a renormalizable theory, in the sense that with a finite number of renormalization conditions all possible divergences can be cured. As any observable must be independent on μ_R, the introduction of the renormalization scale implies the definition of the *running* coupling $\alpha_S(Q^2)$, which depends on the process energy scale Q^2. The specific dependence is given by the renormalization group equation (RGE)

$$Q^2 \frac{\partial \alpha_S(Q^2)}{\partial Q^2} = \beta(\alpha_S(Q^2)) \,, \tag{1.5}$$

with the initial condition $\alpha_S(\mu_R^2) = \alpha_S$, the fixed renormalized coupling.

The β function has the perturbative expansion

$$\beta(\alpha_S) = -\alpha_S^2(\beta_0 + \beta_1\alpha_S + \beta_2\alpha_S^2 + \dots) \,, \tag{1.6}$$

where the leading order term is

$$\beta_0 = \frac{1}{12\pi}\,(33 - 2n_f) \,, \tag{1.7}$$

and n_f are the numbers of active flavours at the scale Q^2. Since in QCD $n_f < 17$ at any scale, β_0 is positive and the β function (1.6) stays negative. Accordingly, eq. (1.5) tells that the strength of the coupling decreases at a logarithmic rate as the energy of the interaction increases. This property of QCD is called *asymptotic freedom* [9]: for high enough energy the coupling is sufficiently small to allow perturbative calculations, with quarks and gluons treated as asymptotic states, or free particles.

More quantitatively, solving the differential equation (1.5) at leading order gives

$$\alpha_S(Q^2) = \frac{\alpha_S(\mu_R^2)}{1 + \beta_0 \alpha_S(\mu_R^2) \ln\left(Q^2/\mu_R^2\right)} , \tag{1.8}$$

and the denominator can be expanded in powers of $\alpha_S \ln\left(Q^2/\mu_R^2\right)$,

$$\alpha_S(Q^2) = \alpha_S(\mu_R^2) \left(1 - \beta_0 \alpha_S(\mu_R^2) \ln\left(Q^2/\mu_R^2\right) + \dots\right) . \tag{1.9}$$

Logarithmic corrections $(\alpha_S \ln\left(Q^2/\mu_R^2\right))^n$ typically arise in divergent amplitudes with n loops, and thus the RGE automatically resums these terms at all orders by using the running coupling.

Equation (1.8) gives a parametrization of the running coupling in terms of its value at the reference scale μ_R^2. Another possible parametrization is found by defining the QCD scale Λ as the zero of the denominator of eq. (1.8), which therefore is rearranged into the form

$$\alpha_S(Q^2) = \frac{1}{\beta_0 \ln(Q^2/\Lambda^2)} . \tag{1.10}$$

This formula is the clear statement that α_S decreases as $(\ln\left(Q^2\right))^{-1}$ for large Q. The QCD scale gives a rough estimate of the energy at which the coupling becomes so strong that a perturbative treatment of QCD is no longer justified. Experimental measurements yield a value of $\Lambda \simeq 200\,\text{MeV}$.

The parametrization (1.10) is not used for the estimation of $\alpha_S(Q^2)$ because at higher orders the quantity Λ does not directly arise. It has become standard practice to quote the value of α_S at the mass of the Z boson $M_Z \simeq 91.2\,\text{GeV}$, where $\alpha_S(M_Z^2) = 0.118$ [10]. Then, by solving eq. (1.5) at fixed order such as the leading order solution (1.8), the running of the coupling can be computed at any other scale.

1.2 Deep Inelastic Scattering

In the late '60s and early '70s, collision experiments were conducted to probe the inner structure of the proton and neutron, and provided direct experimental foundations of QCD as the theory of strong interactions. In those experiments [11, 12], a beam of leptons was accelerated and fired against fixed hadronic targets, such as atomic nuclei. Only the energy and direction of the scattered leptons were measured, leaving the final hadronic state, denoted by X, unknown.

The basic diagram[1] of the process, $\ell(k) + h(p) \to \ell'(k') + X$, is shown pictorially in fig. 1.1. If the target hadron remains intact, $X = h$, the reaction is an elastic scattering.

[1] All the diagrams which will be presented are obtained using the TikZ-Feynman package [13].

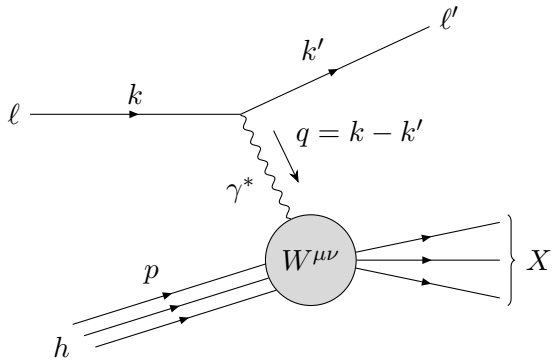

Figure 1.1: Hadronic level diagram of a DIS process. The incoming lepton ℓ scatters off a target hadron h by exchanging a virtual photon γ^* with momentum q. The hadronic tensor $W^{\mu\nu}$ is defined in eq. (1.17).

In case of large momentum transfer, the hadron fragments into many particles and the process is referred to as *deep inelastic* scattering, or DIS.

Assuming that the energy of the incoming lepton is much greater than its mass, the kinematic variables of DIS in the rest frame of the target hadron are:

$p = (M, 0, 0, 0)$	The initial momentum of the target hadron of mass M.
$k = (E, 0, 0, E)$	The lepton initial momentum of energy E.
$k' = (E', E' \sin\theta, 0, E' \cos\theta)$	The lepton final momentum of energy E'.
$q = k - k'$	The momentum transfer.
$\nu = M(E - E') = p \cdot q$	The energy loss of the lepton times the hadron mass.
$y = \nu / p \cdot k$	The fractional energy loss of the lepton.
$Q^2 = -q^2$	Since q is spacelike, it is convenient to define $Q^2 \geq 0$.
$x = \dfrac{Q^2}{2\nu} = \dfrac{Q^2}{2p \cdot q} = \dfrac{Q^2}{2MEy}$	The Bjorken variable, of crucial importance for DIS.

Conservation of baryonic number implies that the invariant mass of the final hadronic state X must be at least that of the initial nucleon

$$M_X^2 = (p+q)^2 \geq M^2 \quad \Leftrightarrow \quad M^2 + 2p \cdot q - Q^2 \geq M^2 \quad \Rightarrow \quad x \leq 1\,. \tag{1.11}$$

Since Q^2 and ν are both positive, x must be positive as well. Then the kinematically allowed interval for x is

$$0 \leq x \leq 1\,. \tag{1.12}$$

The unpolarized amplitude $\mathcal{M}$ of the process in fig. 1.1 is

$$i\mathcal{M} = -(ie)^2 \bar{u}(k')\gamma_\mu u(k) \frac{i}{Q^2} \langle X | \mathcal{J}_h^\mu | p \rangle\ , \tag{1.13}$$

where $\mathcal{J}_h^\mu$ is the hadronic electromagnetic current. The theoretical difficulty of the calculation arise from the fact that the hadronic states $|p\rangle$ and $|X\rangle$ are unknown, since they are characterized by non-perturbative effects.

To compute the cross section it is convenient to factor the modulus squared of the amplitude in eq. (1.13) into a leptonic and a hadronic piece,

$$d\sigma \sim L_{\mu\nu} W^{\mu\nu} \,. \tag{1.14}$$

The leptonic tensor, $L_{\mu\nu}$, is completely determined by QED,

$$\begin{aligned} L_{\mu\nu} &= e^2 \frac{1}{2} \sum_{spins} [\bar{u}(k')\gamma_\mu u(k) \bar{u}(k)\gamma_\nu u(k')] = e^2 \frac{1}{2} \mathrm{tr}\left[\not{k}'\gamma_\mu \not{k}\gamma_\nu\right] = \\ &= 2e^2 \left(k_\mu k'_\nu + k_\nu k'_\mu - g_{\mu\nu} k \cdot k'\right) \,, \end{aligned} \tag{1.15}$$

while the hadronic tensor, $W_{\mu\nu}$, contains all the information about the interaction of the photon and the target hadron

$$\begin{aligned} W^{\mu\nu} &= \frac{1}{4\pi} \sum_X \langle p|\mathcal{J}_h^{\mu\dagger}|X\rangle \, \langle X|\mathcal{J}_h^\nu|p\rangle \, (2\pi)^4 \delta^4(q + p - p_X) = \\ &= \frac{1}{4\pi} \int \mathrm{d}^4 z \, e^{iq\cdot z} \, \langle p|\mathcal{J}_h^{\mu\dagger}(z)\mathcal{J}_h^\nu(0)|p\rangle \,, \end{aligned} \tag{1.16}$$

where we used the completeness relation on the final states and the integral representation of the four dimensional delta function.

Due to our ignorance on the initial hadronic state, eq. (1.16) cannot be computed directly in QCD but symmetries and conservation laws can restrict its functional form. Conservation of the electromagnetic current, $\partial_\mu \mathcal{J}_h^\mu = 0$, implies that $q_\mu W^{\mu\nu} = 0$ and $q_\nu W^{\mu\nu} = 0$. Furthermore, by requiring the tensor to be symmetric under parity transformations, one can show that the most general form is

$$W^{\mu\nu} = \left(-g^{\mu\nu} + \frac{q^\mu q^\nu}{q^2}\right) F_1(x, Q^2) + \left(p^\mu + \frac{q^\mu}{2x}\right)\left(p^\nu + \frac{q^\nu}{2x}\right) \frac{1}{\nu} F_2(x, Q^2) \,, \tag{1.17}$$

where the functional parameters F_1 and F_2 are known as electromagnetic *structure functions*. To project out these structure functions, it is convenient to choose a reference frame where the struck hadron is moving very fast, such that we can neglect its mass, $p^2 = 0$. We can then introduce the lightlike vector n with the properties $n \cdot p = 1$ and $n \cdot q = 0$ to obtain the following relations:

$$\begin{aligned} \nu n^\mu n^\nu W_{\mu\nu} &= F_2 \,, \\ \frac{4x^2}{\nu} p^\mu p^\nu W_{\mu\nu} &= F_2 - 2xF_1 =: F_L \,, \end{aligned} \tag{1.18}$$

where the quantity in the second equation is called longitudinal structure function.

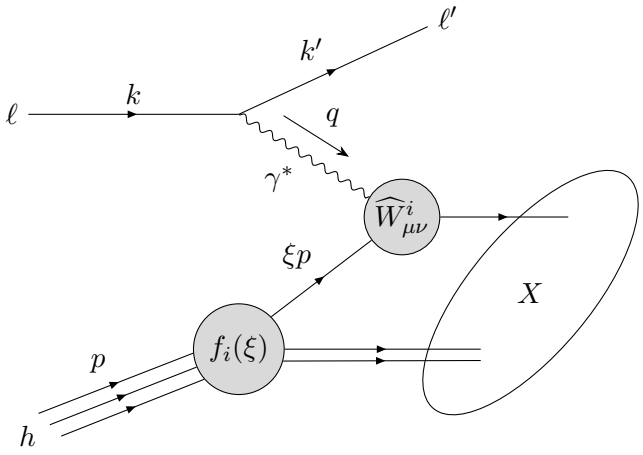

Figure 1.2: The DIS process in QCD. The virtual photon γ^* scatters off a parton with a fraction ξ of momentum of the struck hadron. The partonic tensor $\widehat{W}^i_{\mu\nu}$ describes the QCD corrections to the interaction vertex.

In this frame, we can write the differential cross section of the DIS process mediated by a virtual photon as:

$$\frac{d^2\sigma}{dxdQ^2} = \frac{4\pi\alpha^2}{Q^4}\left[\left(1+(1-y)^2\right)F_1 + \frac{(1-y)}{x}(F_2 - 2xF_1)\right], \tag{1.19}$$

and therefore F_1 and F_2 can be extracted from measurements of (1.19).

The *Bjorken limit* is defined as $Q^2, \nu \to \infty$ with x fixed. In this limit the structure functions were observed to obey an approximate *scaling law*, *i.e.* they depend only on the variable x:

$$F_i(x, Q^2) \xrightarrow[Q^2,\nu\to\infty]{} F_i(x)\,. \tag{1.20}$$

Bjorken scaling implies that the virtual photon scatters off pointlike constituents, since otherwise the structure functions would depend on the ratio Q/Q_0, with $1/Q_0$ some length scale characterizing the size of the constituents.

So far we made some assumptions that simplified the hadronic tensor structure of eq. (1.16). We are now ready to introduce the "naive" *parton model*, which successfully predicts the behaviour of eq. (1.20) and gives further understanding on the quantities in eqs. (1.17) and (1.18).

1.2.1 The parton model

One of the contributions that provided great support to QCD as the theory of strong interactions was the proposal of what is called "naive" parton model [14]. Naive because it was developed to explain the early results of electron/proton DIS experiments, without an actual basic theory as a foundation. Nevertheless, it is still a justifiable approximation

at high energies and its ideas are now part of QCD, fig. 1.2. The key points of the model can be summarized as follows:

- A hadron can be considered as a composition of point like constituents, called *partons*, which correspond to quarks and gluons.
- In a DIS process when Q^2 is sufficiently higher than the binding energy of the partons, the virtual photon becomes sensitive to the inner structure of the hadron and scatters incoherently off the quark constituents.
- If p is the momentum of the hadron in the fast moving frame then a parton carries a fraction of momentum ξp, with $0 < \xi < 1$. We are therefore neglecting the transverse components of the momentum of the partons, which are due to the internal soft dynamics, on the assumption of the high energy transfer in the direction of motion.
- The probability that the photon interacts with the i-th parton whose momentum fraction lies between ξ and $\xi + \mathrm{d}\xi$ is given by $f_i(\xi)\,\mathrm{d}\xi$, where $f_i(\xi)$ is called *Parton Distribution Function*, or PDF.

The probabilistic interpretation allows to rewrite the hadronic tensor (1.17) as a sum of parton-level tensors, $\widehat{W}^i_{\mu\nu}$, weighted by the corresponding distributions f_i,

$$W_{\mu\nu}(x, Q^2) = \sum_i \int_x^1 \frac{\mathrm{d}\xi}{\xi} f_i(\xi) \widehat{W}^i_{\mu\nu}\left(\frac{x}{\xi}, Q^2\right) , \tag{1.21}$$

where the index i runs over the partons which enter the process. The factor $1/\xi$ is needed to obtain the proper normalization $2p^0$ of the hadronic state in terms of that of quark states $2\xi p^0$.

The same arguments which brought us to (1.17) can be applied to the partonic level tensors, which therefore are

$$\widehat{W}^i_{\mu\nu} = \left(-g^{\mu\nu} + \frac{q^\mu q^\nu}{q^2}\right) \widehat{F}^i_1\left(\frac{x}{\xi}, Q^2\right) + \left(p^\mu + \frac{q^\mu}{2x}\right)\left(p^\nu + \frac{q^\nu}{2x}\right) \frac{\xi^2}{\nu} \widehat{F}^i_2\left(\frac{x}{\xi}, Q^2\right) , \tag{1.22}$$

where $\widehat{F}^i_1$ and $\widehat{F}^i_2$ are the parton structure functions, related to the electromagnetic structure functions by substituting eqs. (1.17) and (1.22) in eq. (1.21):

$$F_1(x, Q^2) = \sum_i \int_x^1 \frac{\mathrm{d}\xi}{\xi} f_i(\xi) \widehat{F}^i_1\left(\frac{x}{\xi}, Q^2\right) , \tag{1.23}$$

$$F_2(x, Q^2) = \sum_i \int_x^1 \mathrm{d}\xi\, \xi f_i(\xi) \widehat{F}^i_2\left(\frac{x}{\xi}, Q^2\right) . \tag{1.24}$$

The advantage is that now the partonic level tensor $\widehat{W}^i_{\mu\nu}$ is computable in perturbation theory from the averaged squared amplitude of the subprocess $q(\xi p) + \gamma^*(q) \to q(l)$, and

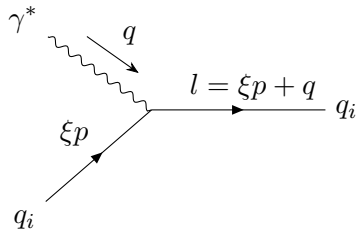

Figure 1.3: LO diagram of the process $q(\xi p) + \gamma^*(q) \to q(l)$.

the quantities $\widehat{F}^i$ can be extracted in a similar way as in eq. (1.18):

$$\frac{\nu n^\mu n^\nu}{\xi^2}\widehat{W}^i_{\mu\nu} = \widehat{F}^i_2\,, \tag{1.25}$$

$$\frac{4x^2}{\nu\xi^2}p^\mu p^\nu \widehat{W}^i_{\mu\nu} = \widehat{F}^i_2 - \frac{2x}{\xi^2}\widehat{F}^i_1 =: \widehat{F}^i_L\,. \tag{1.26}$$

The parton model treats the subprocess $q\gamma \to q$ at leading order (LO), shown in fig. 1.3, whose matrix element is

$$i\mathcal{M}^i_\mu = -ie_i\bar{u}(l)\gamma_\mu u(\xi p)\,. \tag{1.27}$$

During the computation we can neglect the quark masses and therefore the averaged modulus squared of $\mathcal{M}^i_\mu$ is the same as the leptonic tensor $L_{\mu\nu}$ of eq. (1.15) provided the substitutions $k \to \xi p$, $k' \to l$ and $e \to e_i$. The partonic tensor is similar to eq. (1.16) with an additional factor of integration over the final particle phase space:

$$\begin{aligned}\widehat{W}^i_{\mu\nu} &= \frac{1}{4\pi}\int\frac{\mathrm{d}^3 l}{(2\pi)^3 2E_l}\frac{1}{2}\sum_{spins}|\mathcal{M}^i|^2_{\mu\nu}(2\pi)^4\delta^{(4)}(\xi p + q - l) = \\ &= \frac{1}{4}\sum_{spins}|\mathcal{M}^i|^2_{\mu\nu}\delta(l^2)\,,\end{aligned} \tag{1.28}$$

where we used the identity of the Lorentz invariant integral

$$\int\frac{\mathrm{d}^3 l}{(2\pi)^3 2E_l} = \int\frac{\mathrm{d}^4 l}{(2\pi)^4}(2\pi)\delta(l^2)\Big|_{l^0>0}\,. \tag{1.29}$$

We can now extract the structure function (1.25),

$$\widehat{F}^i_2 = \frac{\nu n^\mu n^\nu}{\xi^2}\widehat{W}^i_{\mu\nu} = 2\nu e_i^2\delta(l^2) = e_i^2\delta(\xi - x)\,, \tag{1.30}$$

since

$$\delta(l^2) = \delta\left((\xi p + q)^2\right) = \delta(2\xi p\cdot q - Q^2) = \frac{1}{2\nu}\delta(\xi - x)\,. \tag{1.31}$$

Equation (1.31) states that at leading order the Bjorken variable x is actually the momentum fraction ξ of the scattered parton. By computing the left hand side of eq. (1.26) we find that $\widehat{F}^i_L = 0$ and then

$$\widehat{F}^i_1 = \frac{\xi^2}{2x}\widehat{F}^i_2 = e_i^2\frac{\xi^2}{2x}\delta(\xi - x)\,. \tag{1.32}$$

Finally we may write the electromagnetic structure functions of eqs. (1.23) and (1.24) as predicted by the parton model

$$\begin{aligned} F_1(x,Q^2) &= \sum_i \int_x^1 \frac{\mathrm{d}\xi}{\xi} f_i(\xi) e_i^2 \frac{\xi^2}{2x} \delta(\xi - x) = \frac{1}{2} \sum_i e_i^2 f_i(x)\,, \\ F_2(x,Q^2) &= \sum_i \int_x^1 \mathrm{d}\xi \xi f_i(\xi) e_i^2 \delta(\xi - x) = x \sum_i e_i^2 f_i(x)\,. \end{aligned} \tag{1.33}$$

In these equations there are two important features that were needed to explain the early experimental measurements of DIS:

(i) the structure functions F_1 and F_2 *scale*, which means they have no dependence on the momentum transfer Q^2;

(ii) $F_2(x) = 2xF_1(x)$, which is known as the Callan-Gross relation [15].

The first property was found by Bjorken [16] and inspired the development of the parton model, while the second implies that the longitudinal structure function vanish $F_L = F_2 - 2xF_1 = 0$, as we already noted at the partonic level where $\widehat{F}_L = 0$. The relation is a consequence of the specific process in fig. 1.3, and as such is evidence of the spin$-1/2$ nature of quarks.

1.2.2 Higher order corrections and scaling violation

The discovery of asymptotic freedom in QCD gave the theoretical foundation to explain the scaling feature of the structure functions. Since the constituents of the hadrons at high energies were expected to behave as quasi-free pointlike particles, the partons were then readily associated to quarks and gluons, and the predictions of the parton model related to the leading order subprocess $\gamma q \to q$ of fig. 1.3.

As said, the parton model was able to provide a valid description of the results of the first DIS experiments. However, to obtain more accurate predictions we cannot neglect the higher order corrections of QCD, starting at $\mathcal{O}(\alpha_S)$. When accounting for these corrections, virtual loops and real emissions of partons have to be considered. Since we are assuming the quarks to be massless, the virtual contributions suffer both UV and IR singularities, while real emission processes are IR divergent due to soft and collinear final states.

In fig. 1.4 are shown the next-to-leading order (NLO) *real* corrections, *i.e.* gluon emission from initial and final state, $\gamma^* q \to gq$, and the event in which the process is initiated by a gluon splitting into a pair of quark-antiquark, $\gamma^* g \to q\bar{q}$. These corrections will lead to a mild break of Bjorken scaling by the appearance of logarithms of Q^2 in the structure functions. The measurements of such scaling violations provided further evidence to establish QCD as the theory of strong interactions.

The squared amplitudes for gluon emission are shown in fig. 1.5. As stated above all these processes suffer from IR singularities that may be treated in several ways, for example by giving the gluon a small mass, or by using the dimensional regularization [17]. These methods can take care of both infrared and ultraviolet divergences: in particular, IR

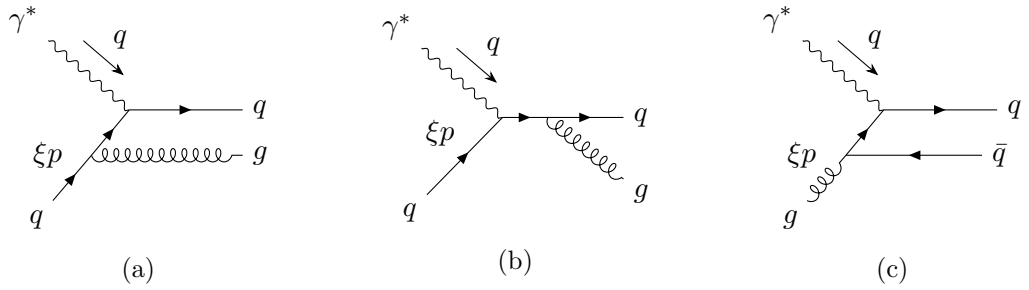

Figure 1.4: Feynman diagrams of the NLO contributions to a deep inelastic scattering. Real gluon emission (b) and (c), and gluon initiated process (d).

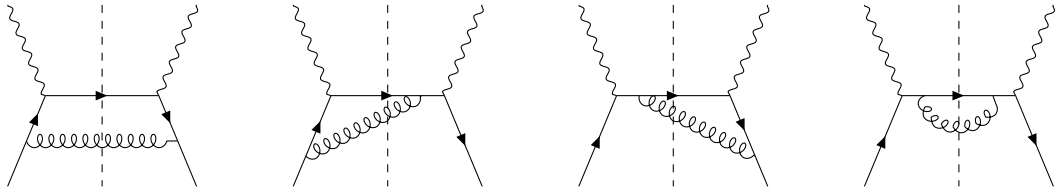

Figure 1.5: Contributions to the squared amplitude from real gluon emission at NLO.

singularities from virtual and *final* state emissions cancel each other out when combined. This result holds thanks to the Kinoshita-Lee-Nauenberg theorem [18, 19], which states that all completely inclusive processes in QCD are IR safe. However, the real emission from initial state, fig. 1.4a, is not subject to this cancellation and its divergence has to be treated separately (the same holds for the process of fig. 1.4c).

We can be more explicit with a different procedure [20], in which the computation is carried out in the light-cone gauge: the contributions from figs. 1.5b to 1.5d are finite, while the remaining singular term is therefore the diagram in fig. 1.5a. As in eqs. (1.25) and (1.28), we can extract the parton-level structure function $\widehat{F}_2^i$ and combine the result with the leading order term of eq. (1.30)

$$\widehat{F}_2^{i(q)}\left(\frac{x}{\xi}, Q^2\right) = e_i^2 \left\{ \delta(\xi - x) + \frac{\alpha_S}{2\pi} \frac{x}{\xi^2} \left[P_{qq}\left(\frac{x}{\xi}\right) \ln \frac{Q^2}{\kappa^2} + C_q\left(\frac{x}{\xi}\right) \right] + \mathcal{O}(\alpha_S^2) \right\} , \quad (1.34)$$

where the superscript q refers to the $\gamma^* q \to gq$ process. In general, the function P_{ab} describes the $b \to a$ parton splitting, so P_{qq} is known as the $q \to q$ *splitting function*,

$$P_{qq}(z) = \frac{4}{3} \frac{1+z^2}{1-z} , \quad (1.35)$$

while all the other finite contributions are in the C_q term. The logarithm $\ln\left(Q^2/\kappa^2\right)$ originates from the integration over the gluon transverse momentum spectrum

$$\int_{\kappa^2}^{Q^2} \frac{\mathrm{d}k_\perp^2}{k_\perp^2} = \ln \frac{Q^2}{\kappa^2} , \quad (1.36)$$

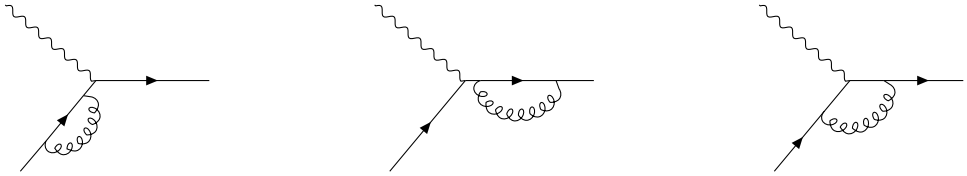

Figure 1.6: Virtual gluon emissions contributing at $\mathcal{O}(\alpha_S)$ to a deep inelastic scattering.

where we have introduced an infrared cut-off κ as a regulator to obtain a finite result, since the lower limit of integration should be set to zero.

This treatment is however incomplete, since we need also the virtual gluon diagrams shown in fig. 1.6. At $\mathcal{O}(\alpha_S)$ their contribution comes from the interference with the leading-order term of fig. 1.3. If we inspect eq. (1.35) we can see that it's divergent when $z \to 1$, which is due to gluon emission with energy that tends to zero (*soft* divergence). The virtual corrections singularities cancel exactly this divergent behaviour of P_{qq} by modifying its expression to

$$P_{qq}(z) = \frac{4}{3}\frac{1+z^2}{(1-z)_+} + 2\delta(1-z)\,, \tag{1.37}$$

where the *plus prescription* on the singular part refers to the distribution

$$\int_0^1 \mathrm{d}z \frac{f(z)}{(1-z)_+} = \int_0^1 \mathrm{d}z \frac{f(z)-f(1)}{1-z}\,. \tag{1.38}$$

In this way, this final result for gluon emission ensures the conservation of baryonic number, since

$$\int_0^1 \mathrm{d}z P_{qq}(z) = 0\,. \tag{1.39}$$

We can now compute the hadronic structure function $F_2^{(q)}$ using eq. (1.24),

$$F_2^{(q)}(x,Q^2) = x\sum_{i=q,\bar{q}} e_i^2 \left\{ f_i(x) + \frac{\alpha_S}{2\pi}\int_x^1 \frac{\mathrm{d}\xi}{\xi} f_i(\xi)\left[P_{qq}\left(\frac{x}{\xi}\right)\ln\frac{Q^2}{\kappa^2} + C_q\left(\frac{x}{\xi}\right)\right] + \mathcal{O}(\alpha_S^2)\right\}\,, \tag{1.40}$$

and we can see that beyond leading order it is Q^2 dependent, with Bjorken scaling broken by logarithms of Q^2.

Equation (1.40) is still ill defined since the cutoff κ has no physical meaning and eventually we should take the limit $\kappa^2 \to 0$. From eq. (1.36) we can see that this singularity arises when the gluon is emitted parallel to the quark, which is why it is called a *collinear divergence*. In this case there is no IR cancellation since the photon *can* distinguish between a quark and a collinear quark-gluon pair with the same overall momentum. The key to obtain a finite result is to realize that a collinear emission belongs to the long range or "soft" regime of the strong interaction, which we cannot compute in perturbation theory.

In the same way as the renormalization of the bare parameters of the Lagrangian, we can consider $f_i(\xi)$ in eq. (1.40) as unmeasurable bare distributions $f_i^{(0)}(\xi)$, and absorb the

collinear singularities into these functions at a momentum scale μ_F, called *factorization scale*. By splitting the divergent logarithm as

$$\ln \frac{Q^2}{\kappa^2} = \ln \frac{Q^2}{\mu_F^2} + \ln \frac{\mu_F^2}{\kappa^2}\,, \tag{1.41}$$

we can define the renormalized parton distributions

$$f_i(x,\mu_F^2) = f_i^{(0)}(x) + \frac{\alpha_S}{2\pi}\int_x^1 \frac{\mathrm{d}\xi}{\xi} f_i^{(0)}(\xi)\left[P_{qq}\left(\frac{x}{\xi}\right)\ln\frac{\mu_F^2}{\kappa^2} + C_q\left(\frac{x}{\xi}\right)\right] + \mathcal{O}(\alpha_S^2)\,, \tag{1.42}$$

and obtain a finite structure function independent from the infrared cutoff κ,

$$\begin{aligned} F_2^{(q)}(x,Q^2) &= x\sum_{i=q,\bar{q}} e_i^2\left[f_i(x,\mu_F^2) + \frac{\alpha_S}{2\pi}\int_x^1\frac{\mathrm{d}\xi}{\xi} f_i(\xi,\mu_F^2)P_{qq}\left(\frac{x}{\xi}\right)\ln\frac{Q^2}{\mu_F^2} + \mathcal{O}(\alpha_S^2)\right] = \\ &= x\sum_{i=q,\bar{q}} e_i^2\int_x^1\frac{\mathrm{d}\xi}{\xi} f_i(\xi,\mu_F^2)\left[\delta\left(1-\frac{x}{\xi}\right) + \frac{\alpha_S}{2\pi}P_{qq}\left(\frac{x}{\xi}\right)\ln\frac{Q^2}{\mu_F^2} + \mathcal{O}(\alpha_S^2)\right]. \end{aligned} \tag{1.43}$$

The distributions $f_i(x,\mu_F^2)$ cannot be computed from first principles in perturbation theory, but can be determined from measurements of the structure function at any scale, since by setting $\mu_F^2 = Q^2$ we have $F_2(x,Q^2) = x\sum_i e_i^2 f_i(x,Q^2)$.

During the factorization procedure the singular logarithmic terms are always absorbed inside the renormalized parton distributions, but there is an arbitrariness in how the finite parts are treated. In fact in eq. (1.42) we absorbed the whole term C_q, using the so called DIS scheme [21]. In general, the choice of the finite terms to include in $f_i(x,\mu_F^2)$ defines the *factorization scheme*. The most used one is the *Modified Minimal Subtraction*, or $\overline{\mathrm{MS}}$ scheme, in combination with the dimensional regularization of divergent amplitudes. In this case the absorbed finite terms are simply $\ln 4\pi - \gamma_E$, which appear in all calculations due to the dimensional regularization procedure. Then, with $\mu_F^2 = Q^2$, we have

$$F_2^{(q)}(x,Q^2) = x\sum_{i=q,\bar{q}} e_i^2\int_x^1\frac{\mathrm{d}\xi}{\xi} f_i(\xi,Q^2)\left[\delta\left(1-\frac{x}{\xi}\right) + \frac{\alpha_S}{2\pi}C_{\overline{\mathrm{MS}}}^q\left(\frac{x}{\xi}\right) + \mathcal{O}(\alpha_S^2)\right]. \tag{1.44}$$

Once a scheme gets fixed, it must be kept in the calculation of all the other quantities.

To conclude the computation of the $\mathcal{O}(\alpha_S)$ corrections, we must consider the remaining process $\gamma^* g \to q\bar{q}$ shown in fig. 1.4. This contribution is IR divergent due to collinear quark-antiquark splitting of the initial gluon state, and the resulting parton-level structure function is

$$\widehat{F}_2^{i(g)}\left(\frac{x}{\xi},Q^2\right) = e_i^2\frac{\alpha_S}{2\pi}\frac{x}{\xi^2}\left[P_{qg}\left(\frac{x}{\xi}\right)\ln\frac{Q^2}{\kappa^2} + C_g\left(\frac{x}{\xi}\right)\right] + \mathcal{O}(\alpha_S^2)\,, \tag{1.45}$$

where κ is an infrared cut-off, C_q contains all the finite contributions, the superscript g refers to the $\gamma^* g \to q\bar{q}$ process, and the $g \to q$ splitting function is

$$P_{qg}(z) = \frac{1}{2}\left[z^2 + (1-z)^2\right]. \tag{1.46}$$

We can finally compute the complete structure function F_2 by using eq. (1.24) and eq. (1.45) with a convolution of the gluon PDF $f_g(\xi)$, then add the result to eq. (1.40). In the $\overline{\text{MS}}$ scheme we may obtain the finite result:

$$\begin{aligned} F_2(x,Q^2) = x\sum_{i=q,\bar{q}} e_i^2 \int_x^1 \frac{\mathrm{d}\xi}{\xi} f_i(\xi,Q^2)\left[\delta\left(1-\frac{x}{\xi}\right)+\frac{\alpha_S}{2\pi}C^q_{\overline{\text{MS}}}\left(\frac{x}{\xi}\right)\right] \\ + x\sum_{i=q,\bar{q}} e_i^2 \int_x^1 \frac{\mathrm{d}\xi}{\xi} f_g(\xi,Q^2)\left[\frac{\alpha_S}{2\pi}C^g_{\overline{\text{MS}}}\left(\frac{x}{\xi}\right)\right] + \mathcal{O}(\alpha_S^2)\,, \end{aligned} \tag{1.47}$$

where $\mu_F^2 = Q^2$ as before.

During the whole computation we have left implicit the scale μ at which the running coupling α_S is evaluated. Since both the factorization and renormalization scales are arbitrary, they are often chosen to be equal $\mu_F = \mu$.

1.3 The factorization theorem

The generalization to all orders of the parton model assumption eq. (1.21) is given by *factorization theorem* [20, 22],

$$W_{\mu\nu}(x,Q^2) \;=\; \sum_{i=q,\bar{q},g}\int_x^1 \frac{\mathrm{d}\xi}{\xi} f_i(\xi,\mu_F^2)\widehat{W}^i_{\mu\nu}\left(\frac{x}{\xi},\frac{Q^2}{\mu_F^2},\alpha_S(\mu_F^2)\right)\,, \tag{1.48}$$

where now is present a dependence on the factorization scale μ_F.

As a consequence, the structure functions admit the most general decomposition

$$F_a(x,Q^2) \;=\; \sum_{i=q,\bar{q},g}\int_x^1 \frac{\mathrm{d}\xi}{\xi} f_i(\xi,\mu_F^2)\, C_{a,i}\left(\frac{x}{\xi},\frac{Q^2}{\mu_F^2},\alpha_S(\mu_F^2)\right) + \mathcal{O}\left(\frac{\Lambda^2}{Q^2}\right)\,, \tag{1.49}$$

where the final term denotes non-perturbative contributions, such as hadronizations processes, multiparton interactions, etc. For energies sufficiently higher than the QCD scale Λ, these effects are negligible, and the observable factorizes into

- *universal* parton densities, $f_i(\xi,\mu_F^2)$, which absorb the *long distance* collinear singularities. They contain informations about the soft internal dynamics of the hadron before the interaction and therefore are *process independent*, but not computable in perturbative QCD.
- (Wilson) coefficient functions, $C_{a,i}$, which describe the *short distance* subprocess. They are therefore calculable in perturbative QCD as a power series in α_S, but they depend on the particular observable F_a.

The factorization procedure described for DIS in section 1.2.2 is also valid for a large class of processes, where the IR divergent counterterms as in eq. (1.42) have always the same logarithmic structure, $\ln\mu_F^2$. Thus, for instance, one can measure the PDFs of a hadron from deep inelastic scattering experiments and then use them to predict cross sections of other type of processes.

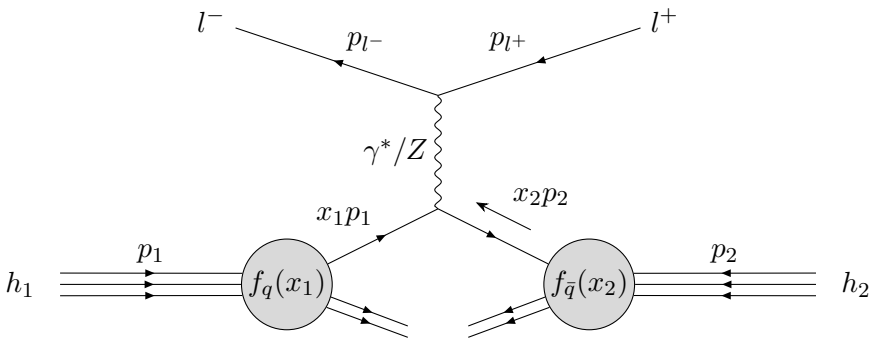

Figure 1.7: Pictorial representation of the neutral current Drell-Yan process for lepton pair production in the parton model.

1.3.1 Hadron-hadron collisions

The factorization theorem allows us to compute cross sections and observables of more complex processes, such as "hard" hadron-hadron collisions, which are of fundamental importance in colliders as the LHC.

A relevant outcome of hadronic collisions is the neutral-current Drell-Yan (DY) process, shown in fig. 1.7, where a quark-antiquark pair annihilates to produce a lepton-antilepton pair l^+l^- with large invariant squared-mass $M_{ll}^2 = (p_{l^+} + p_{l^-})^2 \gg 1\,\text{GeV}^2$. In the naive parton model, the total cross section is given by the subprocess cross section $\hat{\sigma}$ weighted by the respective PDFs of the colliding hadrons h_1 and h_2, summed over all quark-antiquark combinations

$$\sigma_{\text{DY}} = \sum_{q,\bar{q}} \int_\tau^1 \mathrm{d}x_1\,\mathrm{d}x_2\,[q(x_1)\bar{q}(x_2) + (q \leftrightarrow \bar{q})]\,\hat{\sigma}_{q\bar{q}\to l^+l^-}\,. \tag{1.50}$$

The formal domain of validity of eq. (1.50) is the asymptotic "scaling" limit: $M_{ll}^2, s \to \infty$ with $\tau = M_{ll}^2/s$ fixed, analogous to the Bjorken limit of DIS, where $s = (p_1 + p_2)^2$ is the squared center-of-mass energy of the colliding hadrons.

The factorization theorem ensures that the $\mathcal{O}(\alpha_S)$ corrections to the DY process have the same collinear singularities as those of the structure functions of deep inelastic scattering. Therefore we can absorb them into renormalized parton distributions which acquire a dependence on the factorization scale μ_F. The cross section in eq. (1.50) becomes

$$\sigma_{\text{DY}} = \sum_{q,\bar{q}} \int_\tau^1 \mathrm{d}x_1\,\mathrm{d}x_2\,\left[q(x_1, M_{ll}^2)\bar{q}(x_2, M_{ll}^2) + (q \leftrightarrow \bar{q})\right]\,\hat{\sigma}_{q\bar{q}\to l^+l^-}\,, \tag{1.51}$$

where the PDFs are evaluated at the relevant scale $\mu_F^2 = M_{ll}^2$. Particularly, $M_{ll}^2 = x_1x_2s$, and the variables x_1 and x_2 can be expressed as

$$x_1 = \frac{M_{ll}}{\sqrt{s}}\,e^y\,, \qquad x_2 = \frac{M_{ll}}{\sqrt{s}}\,e^{-y}\,, \tag{1.52}$$

where y is the rapidity of the virtual boson,

$$y = \frac{1}{2} \ln \frac{E + p_L}{E - p_L} \,, \tag{1.53}$$

written in terms of its energy E, and longitudinal momentum p_L relative to the collision axis.

The structure of eq. (1.51) finds confirmation in experimental measurements of a large variety of inclusive hard hadron-hadron collisions, $h_1(p_1) + h_2(p_2) \to H(Q, ...) + X$, where H denotes for instance a weak boson, a pair of jets, a Higgs boson, etc. The scale Q could be the invariant mass of H or the transverse momentum of a jet. Then, according to the factorization theorem, the related cross sections assume the general form

$$\sigma = \sum_{i,j=\{q,\bar{q},g\}} \int_{x_{min}}^{1} \mathrm{d}x_1 \, \mathrm{d}x_2 \, f_{i/h_1}(x_1, \mu_F^2) f_{j/h_2}(x_2, \mu_F^2) \, \hat{\sigma}_{ij}(x_1 x_2 \hat{s}, Q, \alpha_S, ...; \mu_F^2, \mu_R^2) \,, \tag{1.54}$$

where typically $x_{min} \gtrsim Q^2/s$. The dependence on the renormalization scale μ_R comes from the power expansion of $\hat{\sigma}_{ij}$ in the running coupling $\alpha_S(\mu_R^2)$. In practical applications it is usual to chose $\mu_F = \mu_R \sim Q$ and varying the scales near this value to estimate the theoretical uncertainties due to truncation of the perturbative series in α_S.

1.4 DGLAP evolution equations

The factorization scale μ_F is introduced to separate the long distance from the short distance physics of a hard scattering process. Since it is an arbitrary scale, at least for values greater than the QCD scale Λ, the observables can't depend from it. We can see for instance that the right hand side of eq. (1.49) for the structure functions contains a μ_F dependence, while the F_a depend only on x and Q^2. A similar observation can be made for the general cross section eq. (1.54). Therefore, the same arguments which led to the introduction of the running coupling eq. (1.5) can be applied in this case.

The Wilson Operator Product Expansion (OPE) provides the determination of the renormalization group equations, valid order by order in perturbation theory, for the PDFs and coefficients functions, which describe their scale dependence on μ_F:

$$\mu_F^2 \frac{\partial}{\partial \mu_F^2} f_i(x, \mu_F^2) = \frac{\alpha_S(\mu_F^2)}{2\pi} \sum_j \int_x^1 \frac{\mathrm{d}\xi}{\xi} P_{ij}\left(\frac{x}{\xi}, \alpha_S(\mu_F^2)\right) f_j(\xi, \mu_F^2) \,, \tag{1.55}$$

$$\mu_F^2 \frac{\partial}{\partial \mu_F^2} C_i\left(x, \frac{Q^2}{\mu_F^2}, \alpha_S(\mu_F^2)\right) = -\sum_j \int_x^1 \frac{\mathrm{d}\xi}{\xi} P_{ij}\left(\frac{x}{\xi}, \alpha_S(\mu_F^2)\right) C_j\left(\xi, \frac{Q^2}{\mu_F^2}, \alpha_S(\mu_F^2)\right) . \tag{1.56}$$

The integro-differential equations (1.55) are known as the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equations. They are of essential importance to compute the *evolution* of the PDFs as μ_F^2 varies, and require only the initial value of the PDFs at some

reference scale. As in eq. (1.5) for the running coupling α_S, where the renormalization group equation takes care of large logarithms, the DGLAP equations effectively sum leading powers of $\left[\alpha_S \ln(Q^2)\right]^n$. These contributions are generated by collinear parton emissions in a region of phase space where the momenta are *strongly ordered*, that is $Q^2 \gg k^2_{n,\perp} \gg \cdots \gg k^2_{2,\perp} \gg k^2_{1,\perp}$:

$$\int_{\kappa^2}^{Q^2} \frac{\mathrm{d}k^2_{n,\perp}}{k^2_{n,\perp}} \cdots \int_{\kappa^2}^{k^2_{3,\perp}} \frac{\mathrm{d}k^2_{2,\perp}}{k^2_{2,\perp}} \int_{\kappa^2}^{k^2_{2,\perp}} \frac{\mathrm{d}k^2_{1,\perp}}{k^2_{1,\perp}} \sim \frac{1}{n!} \ln^n \left(\frac{Q^2}{\kappa^2}\right) , \tag{1.57}$$

where κ^2 cuts off the infrared singularities.

We can see that both eqs. (1.55) and (1.56) are determined by the Altarelli-Parisi splitting functions P_{ij}, which have a perturbative expansion in the running coupling constant α_S and currently they have been computed to order $\mathcal{O}(\alpha_S^3)$ [23, 24]. The leading-order contributions are given by

$$P^{(0)}_{qq}(z) = C_F \left[\frac{1+z^2}{(1-z)_+} + \frac{3}{2}\delta(1-z)\right] , \quad C_F = 4/3 , \tag{1.58}$$

$$P^{(0)}_{qg}(z) = T_R \left[z^2 + (1-z)^2\right] , \quad T_R = 1/2 , \tag{1.59}$$

$$P^{(0)}_{gq}(z) = C_F \left[\frac{1+(1-z)^2}{z}\right] , \tag{1.60}$$

$$P^{(0)}_{gg}(z) = 2C_A \left[\frac{z}{(1-z)_+} + \frac{1-z}{z} + z(1-z)\right] \tag{1.61}$$

$$+ \delta(1-z)\frac{(11C_A - 4n_f T_R)}{6} , \quad C_A = 4 , \tag{1.62}$$

where the plus prescription was defined in eq. (1.38).

Because of charge conjugation invariance and $SU(n_f)$ flavour symmetry of QCD, the rank of the the matrix P_{ij} is not maximal. In fact, the following relations hold:

$$\begin{aligned} P_{q_i q_j} &= P_{\bar{q}_i \bar{q}_j} \\ P_{q_i \bar{q}_j} &= P_{\bar{q}_i q_j} \\ P_{q_i g} &= P_{\bar{q}_i g} \equiv P_{qg} \\ P_{g q_i} &= P_{g \bar{q}_i} \equiv P_{gq} , \end{aligned} \tag{1.63}$$

which means the splitting functions P_{qg} and P_{gq} are independent of the quark flavour and the same for quarks and antiquarks. At leading-order $P_{q_i q_j}$ is zero unless $q_i = q_j$, as can be seen from eq. (1.58). Hence, to solve the DGLAP equations it is convenient to define PDF combinations that make the evolution operator P_{ij} as diagonal as possible.

Given $n_f = 6$ flavours with $f_i = u, d, s, c, b, t$, we first introduce

$$f_i^{\pm} = f_i \pm \bar{f}_i , \tag{1.64}$$

then, we can write the sequence of *non-singlets* (NS) combinations composed of *valences* and *triplets* distributions,

$$\text{Valences:} \quad V_i \equiv f_i^- \tag{1.65}$$

$$\text{Triplets:} \quad \begin{cases} T_3 \equiv u^+ - d^+ \\ T_8 \equiv u^+ + d^+ - 2s^+ \\ T_{15} \equiv u^+ + d^+ + s^+ - 3c^+ \\ T_{24} \equiv u^+ + d^+ + s^+ + c^+ - 4b^+ \\ T_{35} \equiv u^+ + d^+ + s^+ + c^+ + b^+ - 5t^+ \end{cases} \tag{1.66}$$

which satisfy the decoupled evolution equations

$$\mu_F^2 \frac{\partial}{\partial \mu_F^2} f_{NS}(x,\mu_F^2) \;=\; \frac{\alpha_S(\mu_F^2)}{2\pi} \int_x^1 \frac{\mathrm{d}\xi}{\xi} P_{NS}\left(\frac{x}{\xi}, \alpha_S(\mu_F^2)\right) f_{NS}(\xi,\mu_F^2)\,, \quad f_{NS} = V_i, T_j\,. \tag{1.67}$$

The remaining combination of quark distributions is the *singlet* PDF

$$\Sigma(x,\mu_F^2) = \sum_i f_i^+(x,\mu_F^2)\,, \tag{1.68}$$

whose evolution is coupled to that of the gluon:

$$\mu_F^2 \frac{\partial}{\partial \mu_F^2} \begin{pmatrix} \Sigma(x,\mu_F^2) \\ g(x,\mu_F^2) \end{pmatrix} = \frac{\alpha_S(\mu_F^2)}{2\pi} \int_x^1 \frac{\mathrm{d}\xi}{\xi} \times \begin{pmatrix} P_{\Sigma\Sigma}\left(\frac{x}{\xi}, \alpha_S(\mu_F^2)\right) & 2n_f P_{\Sigma g}\left(\frac{x}{\xi}, \alpha_S(\mu_F^2)\right) \\ P_{g\Sigma}\left(\frac{x}{\xi}, \alpha_S(\mu_F^2)\right) & P_{gg}\left(\frac{x}{\xi}, \alpha_S(\mu_F^2)\right) \end{pmatrix} \begin{pmatrix} \Sigma(\xi,\mu_F^2) \\ g(\xi,\mu_F^2) \end{pmatrix}. \tag{1.69}$$

An alternative formulation of the DGLAP equations is in terms of the Mellin transforms, or moments of the parton distributions,

$$f_i(n,\mu_F^2) = \int_0^1 \mathrm{d}x\, x^{n-1} f_i(x,\mu_F^2)\,, \quad n \in \mathbb{C}\,, \tag{1.70}$$

which allow to simplify the convolution integrals into algebraic products, since

$$\begin{aligned} \int_0^1 \mathrm{d}x\, x^{n-1} \int_x^1 \frac{\mathrm{d}y}{y} g(x/y) f(y) &= \int_0^1 \mathrm{d}x\, x^{n-1} \left[\int_0^1 \mathrm{d}y \int_0^1 \mathrm{d}z f(y) g(z) \delta(x - yz) \right] \\ &= \int_0^1 \mathrm{d}y\, y^{n-1} f(y) \int_0^1 \mathrm{d}z\, z^{n-1} g(z) \\ &= f(n) g(n)\,. \end{aligned} \tag{1.71}$$

For instance, the non-singlet part reduces to

$$\mu_F^2 \frac{\partial}{\partial \mu_F^2} f_{NS}(n,\mu_F^2) \;=\; \frac{\alpha_S(\mu_F^2)}{2\pi}\, \gamma_{NS}(n, \alpha_S(\mu_F^2))\, f_{NS}(n,\mu_F^2)\,, \tag{1.72}$$

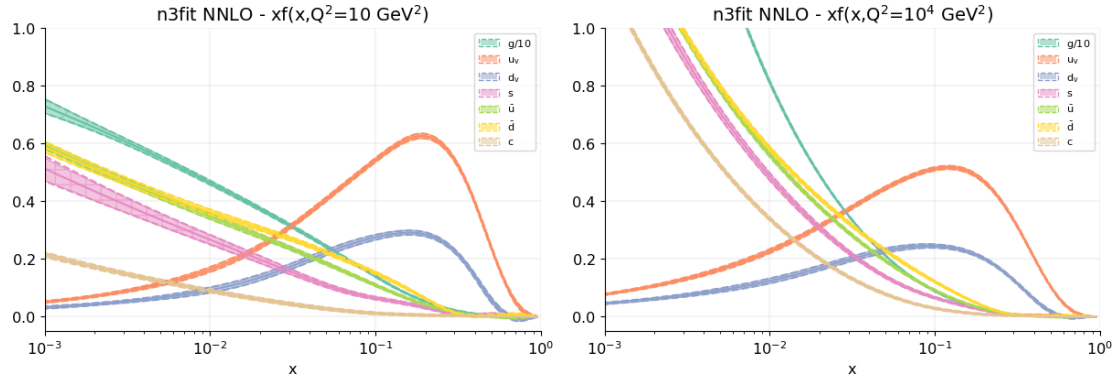

Figure 1.8: Example of proton PDFs evolution at $Q^2 = 10\,\text{GeV}^2$ (left plot) and $Q^2 = 10^4\,\text{GeV}^2$ (right plot), obtained from an `n3fit` NNLO global fit, with $\alpha_S(M_Z^2) = 0.118$. Plots generated with `reportengine` [29].

where γ_{NS} is the so-called anomalous dimension given by

$$\gamma_{NS}(n, \alpha_S(\mu_F^2)) = \int_0^1 \mathrm{d}x\, x^{n-1} P_{NS}(x, \alpha_S(\mu_F^2))\,. \tag{1.73}$$

The equations in the Mellin n-space can be solved analytically but the difficulty arises in returning to the x space, by taking the inverse Mellin transform,

$$f_i(x, \mu_F^2) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \mathrm{d}n\, x^{-n} f_i(n, \mu_F^2)\,. \tag{1.74}$$

In practice, the solution of the DGLAP equations are obtained with numerical codes either through direct integration in x space or by solving the differential equations in Mellin space, and then computing the inverse transformation eq. (1.74). The first method is used in algorithms such as HOPPET [25], QCDNUM [26] or APFEL [27], which is currently exploited in the NNPDF fits, while the second approach is implemented in the QCD-PEGASUS [28] code.

Throughout this thesis we will focus on the proton PDFs, since they are the most studied ones due to their crucial importance in colliders phenomenology, such as the LHC. Thanks to the DGLAP equations, PDF determination from data is much simpler because we can parametrize the distributions at an initial scale $\mu_0^2 = Q_0^2$, and evolve them to the energy scales at which the experimental measurements are performed. From now on we will denote the factorization scale in terms of the energy of the process, $\mu_F^2 = Q^2$.

An example of DGLAP evolution in x space with the code APFEL is shown in fig. 1.8, using the physical basis where $f_v = f^-$. The proton PDFs are evolved from $Q_0^2 = 2.7\,\text{GeV}^2$ to the scales $Q^2 = 10\,\text{GeV}^2$ (left plot) and $Q^2 = 10^4\,\text{GeV}^2$ (right plot).

(a) (b)

Figure 1.9: Lowest-order heavy quark production in FFN (left diagram) and ZM-VFN (right diagram) scheme in a DIS.

1.5 Heavy quarks

Our complete discussion on PDFs and their evolutions relies on the approximation that the quarks contributing in the processes are massless. This assumption is entirely reasonable for the three lightest quarks u, d, s since their mass is far below the QCD scale Λ. However, in order to provide reliable theoretical predictions at all energy scales, a careful treatment of the observables is needed to deal with terms that depend on heavy quark masses near their production threshold.

The specific approach to treat the quark masses is known as a *heavy quark scheme*, where in this case "scheme" refers to the particular approximation used rather than the definition of finite contributions as in the case of factorization or renormalization schemes. The choice of a heavy quark scheme can therefore lead to different results even in the limit of an all-orders calculation.

The heavy quark schemes specific for DIS processes have received a lot of attention due to their impact to the determination of PDFs, and correspondingly to collider physics. Since we are interested in studying the PDFs below and above the threshold production of heavy quarks, a heavy quark scheme is needed to provide a suitable interpolation between all the regimes considered.

Particularly, there are two kinematical regions which are treated differently, depending on the relation between the heavy quark mass m_h and the hard scale Q of the physical process. The first is $m_h^2 \gtrsim Q^2$, where the *Fixed Flavour Number Scheme* (FFNS) is used. In this scheme the heavy quark is treated as a purely final state particle, and the only partons that enter in the theory are the lighter quarks and the gluon. Therefore the heavy quark dependence enters only in the perturbative part of the computation: for instance, at lowest-order its production is due to gluon splitting into a $h\bar{h}$ pair fig. 1.9a. While accurate near and below the mass threshold, this scheme becomes unreliable when $Q^2 \gg m_h^2$, because powers of large logarithms, $\ln\left(Q^2/m_h^2\right)$, spoil the converge of the perturbative series.

The second limit is therefore $Q^2 \gg m_h^2$, where now the additional quark is treated as a massless parton, fig. 1.9b, with the introduction of an associated heavy quark PDF, which is set to zero below the mass threshold and evolved according to the DGLAP equations for scales greater than m_h^2. This treatment is then named as *Zero Mass Variable Flavour Number Scheme* (ZM-VFNS), since the new quark is considered massless. Symmetrically to the previous scheme, in this case the approximation becomes problematic near the mass

threshold, where powers of m_h^2/Q^2 are non-negligible.

For studies at energy scales that span several heavy quark thresholds, hybrid methods or *general mass* schemes are employed to combine the ZM-VF and FF treatments. Such schemes usually reduce to the previous ones at the corresponding energy scales, while the intermediate regimes are handled via some interpolating conditions. Therefore, the difficulty arises when matching the ZM-VF and FF in a unique scheme valid at all Q^2.

The basic idea of a *General Mass Variable Flavour Number Scheme* (GM-VFNS) is to switch the PDFs from a n_f-flavour FFNS ones to the (n_f+1)-flavour FFNS ones at the matching point $\mu = m_h$. The PDFs above and below the threshold are related order by order in α_S by

$$f_i^{\mathrm{VF}}(\mu \to m_h^+) \equiv f_i^{(n_f+1)\mathrm{FF}} = \sum_j^{n_f} A_{ij} \otimes f_j^{(n_f)\mathrm{FF}} \equiv \sum_j^{n_f} A_{ij} \otimes f_j^{\mathrm{VF}}(\mu \to m_h^-)\,, \tag{1.75}$$

where the transition matrix elements $A_{ij}(\mu/m_h)$ are known at NNLO [30, 31], the superscripts $\pm$ indicate the direction of the limits, and the symbol $\otimes$ is used as a shorthand for the usual convolution integral

$$(f \otimes g)(x) = \int_x^1 \frac{\mathrm{d}y}{y} g\left(\frac{x}{y}\right) f(y)\,. \tag{1.76}$$

The general assumption of a GM scheme is that the physical observables have to be continuous when passing through the mass threshold at all orders. Therefore, we may write the DIS structure functions in the vicinity of the matching point, $\mu^2 = Q^2 = m_h^2$, as

$$F_a(x,Q^2) = \sum_j^{n_f} C_{a,j}^-(m_h^2/Q^2) \otimes f_j^-(Q^2) = \sum_i^{n_f+1} C_{a,i}^+(m_h^2/Q^2) \otimes f_i^+(Q^2)\,, \tag{1.77}$$

where f_j^- corresponds to the n_f-flavour PDFs, while f_i^+ to the (n_f+1)-flavour ones. We can now insert eq. (1.75) in eq. (1.77) to arrive at

$$F_a(x,Q^2) = \sum_i^{n_f+1} \sum_j^{n_f} C_{a,i}^+(m_h^2/Q^2) \otimes A_{ij}(Q^2/m_h^2) \otimes f_j^-(Q^2)\,. \tag{1.78}$$

Finally, comparing eq. (1.78) with eq. (1.77) we may find that the coefficient functions must satisfy the transformation formula:

$$C_{a,j}^-(m_h^2/Q^2) = \sum_i^{n_f+1} C_{a,i}^+(m_h^2/Q^2) \otimes A_{ij}(Q^2/m_h^2)\,, \tag{1.79}$$

which define the minimal prescription of a GM-VFN scheme [32].

Several variants of GM-VFNS exist, due to the fact that eq. (1.79) does not completely define all the Wilson coefficients across the matching point, since the transition matrix

A is not a square matrix. Examples of different general mass schemes are S-ACOT [33], Thorne-Roberts [34], and the FONLL which was originally formulated for hadronic collisions [35] and then applied to DIS [36]. The FONLL scheme and its variants are currently in use in the NNPDF framework to account for charm initiated contributions and fit them explicitly [37].

1.6 General properties of the proton PDFs

So far we have reviewed the theoretical description of parton densities in the perturbative QCD framework, where the DGLAP evolution equations and heavy quark mass schemes allow to compute the PDFs at any relevant experimental energy scale Q.

As already mentioned, the main difficulty in the determination of parton densities is their functional dependence on the momentum fraction x at some initial scale Q_0. The number of independent PDFs to be determined depends on the choice of Q_0, as the heavier quarks may be generated perturbatively via the procedures described briefly in section 1.5. The choice adopted by the NNPDF collaboration since the latest release (3.1) is to consider $Q_0 = 1.65\,\text{GeV}$, such that the gluon, the three lightest quarks, u, d, s, and also the charm PDFs are independently parametrized (with the corresponding antiquarks).

Even though the six lightest quarks and gluon distributions are intrinsically related to the non-perturbative dynamics of the proton, some general statements can be made on their x-dependence, which should be valid at all energy scales. Particularly, the *momentum sum rule* (MSR) ensures that the momentum fraction carried by all the partons sum up to the momentum of the parent proton

$$\int_0^1 \mathrm{d}x\, x\left[\Sigma(x,Q^2) + g(x,Q^2)\right] = 1\,, \tag{1.80}$$

where the singlet distribution Σ was defined in eq. (1.68). Then, other constraints come from the number sum rules, which fix the quark distributions to match the observed quantum numbers of the proton:

$$\text{up-valence:} \quad \int_0^1 \mathrm{d}x\left[u(x,Q^2) - \bar{u}(x,Q^2)\right] = 2\,, \tag{1.81}$$

$$\text{down-valence:} \quad \int_0^1 \mathrm{d}x\left[d(x,Q^2) - \bar{d}(x,Q^2)\right] = 1\,, \tag{1.82}$$

$$\text{strange-valence:} \quad \int_0^1 \mathrm{d}x\left[s(x,Q^2) - \bar{s}(x,Q^2)\right] = 0\,. \tag{1.83}$$

Equation (1.80) suggests that the PDFs should vanish in the limit $x \to 1$, since otherwise their contribution would be too large to satisfy the momentum sum rule. The constraints eqs. (1.81) to (1.83) imply that the valence distributions V_i, defined in eq. (1.65), must be integrable over the whole x-range, whereas the same requirement can be imposed instead on the first momentum of the singlet and gluon distributions from eq. (1.80).

Combining these informations we may extract a simple parametrization of the small and large x dependence of valence-like and singlet-like PDFs:

$$f_V(x,Q^2) = N_V\, x^{\alpha_V}(1-x)^{\beta_V} p_V(x)\,,$$
$$f_\Sigma(x,Q^2) = N_\Sigma\, x^{\alpha_\Sigma}(1-x)^{\beta_\Sigma} p_\Sigma(x)\,. \tag{1.84}$$

The coefficients α and β control respectively the small and large x behaviour of the PDFs. Specifically, the values of β should ensure that the parton densities tend smoothly to zero as x approaches 1, while α should assume values for which the valences and the first moment of the singlet and gluon distributions are integrable. The normalizations N are determined from the corresponding sum rules.

Finally, the remaining terms $p(x)$ carry all the unknown dependence of the parton densities on the momentum fraction x. Therefore, their precise determination is the main focus in the research of the proton PDFs.

Chapter 2

PDFs determination

The precise understanding of the functional forms of the PDFs is crucial to provide accurate theoretical predictions of physical cross sections at hadron colliders. Since the parton distributions are related to the non-perturbative dynamics of the proton structure, they cannot be computed in the framework of perturbative QCD. Their determination relies on comparing experimental measurements with theoretical predictions based on a specific parametrization of the PDFs. The fundamental difficulty of this task is situated in the search space: a PDF fit must provide a function rather than a single parameter, despite having only a finite number of data points available. Therefore, one must attempt to find the optimal solution in an infinite-dimensional functional space. Moreover, a careful determination of the uncertainties of the PDFs is essential due to the high precision needed in their applications. The problem of PDF fitting is then to find a reliable estimator for a probability density in a space of functions.

All modern PDFs are given as a set of computer files in the LHAPDF format [38]: an LHAPDF set consists of a list of *members* where the PDFs values, sampled as a grid of points in (x, Q), are stored. The LHAPDF software can interpolate to obtain the value of the parton distributions at arbitrary points in (x, Q). The members of the PDF sets are then used to compute PDF dependent quantities and their uncertainties.

Nowadays, the problem of PDF determination is studied by several groups, each one providing its own sets of PDFs. The main differences between these sets are due to the different strategies adopted by each group, such as the experimental data included in the fit, the schemes used in the computation of theoretical predictions, the choice made for the PDFs parametrization, and finally the fitting algorithm. The most active collaborations involved are the ABM [39], CTEQ [40], HERAPDF [41], MMHT [42] and the NNPDF [43] collaboration.

2.1 Experimental data

The first step in the determination of parton distributions is the selection of the experimental datasets that are used in the fitting procedure. The most important aspect is to identify which dataset can offer precise and reliable data relative to a specific input PDF.

Usually, PDF fitting collaborations apply some specific kinematic cuts to the various processes considered. These cuts ensure that only data where the theoretical calculations are reliable is included.

We briefly present the main processes considered in PDF determination, and review some of the most relevant experiments that constitute the input data of a PDF fit, with particular attention to datasets used in the NNPDF global analyses.

2.1.1 Fixed-target and collider DIS

Deep inelastic scattering data provides the majority of the experimental points in PDF analyses. At leading-order, neutral current DIS measurements from a proton target directly probe the quark sea distributions $q_i + \bar{q}_i$, via the relative strength of interaction of each flavour with the bosons γ, Z. Moreover, charged current, and Z-mediated neutral current data can provide some constraint on flavour separation via the F_3 structure function.

Another important constraint can be obtained from scattering off deuteron targets, which provide important informations on the $u - d$ and u/d PDF combinations, under the assumption of isospin symmetry.

Since the gluon contribution starts at NLO, a DIS can only yield an indirect probing of the gluon PDF from scaling violations of the structure functions. However, the large variety of DIS measurements available at a wide range of scales provides a great deal of information in the determination of the gluon distribution.

DIS datasets can be presented either as experimental cross sections, or decomposed into structure functions. The main experiments which are usually included in he fits are NMC [44, 45], SLAC [46], BCDMS [47, 48], and the final HERA [49] combination. Important measurements for the charm and beauty structure functions are available again from HERA [50, 51, 52], which can be useful in specific applications as the determination of the beauty quark mass m_b.

A further subset of DIS measurements comes from neutrino scattering on nuclear targets, with the experiments CHORUS [53] and NuTeV [54, 55]. Typically, these datasets provide information on the valence quark distributions $q_i - \bar{q}_i$; moreover, since in the dimuon production $\nu N \to \mu\mu X$ the favoured initial state parton is the strange quark, this process is useful to probe the strangeness distribution, whose contribution is difficult to discern from global structure function measurements.

2.1.2 Drell-Yan and jet production

After DIS experiments, the second most important contribution in PDF fits is the dataset of Drell-Yan measurements. As described in section 1.3.1, at LO the neutral current DY process is controlled by the parton combinations $q(x_1)\bar{q}(x_2) + \bar{q}(x_1)q(x_2)$, and therefore can provide various constraints depending on the specific experimental setup. Particularly, eqs. (1.52) and (1.53) imply that high rapidity measurements probe the parton content of the proton at both small and large-x.

Additionally to the neutral current, a Drell-Yan process can occur also via the exchange of a charged weak boson, where the resulting final state is a pair lepton-(anti)neutrino:

$qq' \to W^{\pm} \to l^{\pm}\nu_l$. In this case, the presence of a neutrino from the decay of the W boson makes much more complicated the measurement of its rapidity. Therefore, data is often presented in terms of the pseudorapidity of the detected lepton

$$\eta = -\ln\tan\theta\,, \tag{2.1}$$

defined in terms of the angle θ between the final state lepton and the collision axis, and independent on the particle mass and momentum. A further type of experimental result in charged current DY process is the lepton asymmetry, defined in terms of $W^{\pm} \to l^{\pm}\nu_l$ differential cross sections $\mathrm{d}\sigma_{l^{\pm}}/\,\mathrm{d}\eta_l$ as

$$\mathcal{A}_W^l = \frac{\mathrm{d}\sigma_{l^+}/\,\mathrm{d}\eta_l - \mathrm{d}\sigma_{l^-}/\,\mathrm{d}\eta_l}{\mathrm{d}\sigma_{l^+}/\,\mathrm{d}\eta_l + \mathrm{d}\sigma_{l^-}/\,\mathrm{d}\eta_l}\,, \tag{2.2}$$

which benefit from cancellations of shared systematic uncertainties and provide constraints for the light quark sea $q_i + \bar{q}_i$.

Fixed-target DY of proton and deuteron targets constrain the u/d combination, and the datasets typically used come from the Tevatron experiments E605 [56] and E866/NuSea [57, 58, 59]. The outcome of these measurements are usually affected by nuclear corrections due the low energies that could be reached in fixed target collisions.

The cleanest environmental setup to probe the parton distributions from a DY process is found in high energy colliders data: the Tevatron collaborations, D0 and CDF, provided the earliest measurements with $p\bar{p}$ collisions. Specifically, D0 carried out measurements of W electron/muon asymmetries [60, 61], which provide important information on the quark flavor separation at large-x, and Z rapidity distribution $\mathrm{d}\sigma_Z/\,\mathrm{d}y_Z$ [62].

In the same way, the CDF collaboration considered the Z rapidity distribution [63], but treated also one-jet inclusive cross sections [64], where a narrow cone of hadrons and other particles (jet) is produced as a result of the hadronization of quarks and gluons radiated after the fragmentation of the colliding protons. These are the most important processes for the determination of the gluon distribution, particularly in the large-x region, where the high energy allows gg initiated diagrams to be the most dominant contributions for the production of inclusive jet or di-jet events.

Experimentally, jets are reconstructed with some clustering algorithm that can identify jet structures starting from final hadronic states, ensuring that collinear and infrared safety of QCD are satisfied. The most used jet reconstruction algorithms are the k_t [65], anti-k_t [66], and the Cambridge-Aachen [67] algorithms, which have become popular also within experimentalists thanks to the speed up introduced by the FastJet [68] package.

Cross sections computations for the inclusive jet and di-jet production in hadron-hadron collisions are available at NLO in QCD [69], while the NNLO corrections have been recently studied by the NNLOjet collaboration [70].

2.1.3 LHC data

The current LHC datasets used in PDF determination includes measurements from Run I, taken at center-of-mass energies of 2.76 TeV, 7 TeV and 8 TeV. These measurements include the results from all the three main collaborations, ATLAS, CMS, and LHCb:

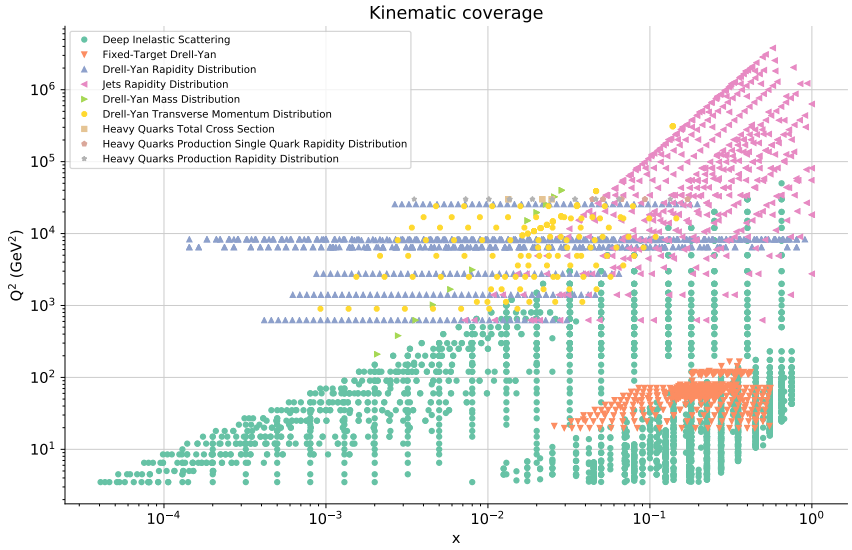

Figure 2.1: Kinematic coverage of datasets included in NNPDF3.1 which will be adopted for the fits presented in this thesis.

- W, Z boson production rapidity and pseudorapidity distributions;
- Z boson double differential cross sections (p_T^Z, y_Z) and (p_T^Z, M_{ll});
- W electron/muon asymmetry distributions, $\mathcal{A}_W^l$;
- jet production data;
- top-quark pair production normalized $y_{t\bar{t}}$ distributions, $(1/\sigma_{t\bar{t}})\,\mathrm{d}\sigma_{t\bar{t}}/y_{t\bar{t}}$;
- total inclusive $t\bar{t}$ cross-sections, $\sigma_{t\bar{t}}$.

Particularly the top pair production at the LHC is a primary probe for the gluon PDF through the $gg \to t\bar{t}$ subprocess.

For the full list of datasets included in the latest release NNPDF3.1, and which will be used in the following, see Ref. [43]. The kinematic range covered is shown in fig. 2.1 and grouped by type of process.

2.2 PDF fit methodology

2.2.1 Parametrization

Once the experimental datasets are given, one must choose a convenient and effective parametrization of the parton distributions in order to compare (indirectly) their predictions to data.

The six quarks, six antiquarks and the gluon give a total of 13 PDFs, but, since typically the quarks c, b, t are determined perturbatively, the total number reduces to 7 independent PDFs to be fitted. The parton parametrization basis is chosen to facilitate both fitting and calculation of the perturbative evolution, as discussed in section 1.4. The final choice is however dictated by the necessity to avoid fitting quantities that are poorly determined by the experimental datasets. We shall now describe as an example the similar PDF parametrization strategies adopted by MMHT and CTEQ, and then focus on the unique one in use within the NNPDF collaboration.

A brief example: MMHT and CTEQ

MMHT2014 [42] uses the following basis for their fits:

$$\begin{aligned} &g\,,\\ u_V &= u - \bar{u}\,,\\ d_V &= d - \bar{d}\,,\\ \Delta &\equiv \bar{d} - \bar{u}\,,\\ S &\equiv 2(\bar{u} + \bar{d}) + s + \bar{s}\,,\\ s^+ &\equiv s + \bar{s}\,,\\ s^- &\equiv x(s - \bar{s})\,, \end{aligned} \tag{2.3}$$

which can parametrize all the 7 distributions to be determined. Then, a functional form in x is chosen for each of these PDF combinations, at the input scale $Q_0^2 = 1\,\text{GeV}^2$. Usually, all groups use the general decomposition eq. (1.84), while the main differences are found in the choice made for the unknown functions $p(x)$. For example, the MMHT gluon parametrization is given by

$$xg(x, Q_0^2) = A_g(1-x)^{\eta_g} x^{\delta_g} \left[1 + \sum_{i=1}^{2} a_{g,i} T_i^{\text{Ch}}(y(x))\right] + A_{g'}(1-x)^{\eta_{g'}} x^{\delta_{g'}}\,, \tag{2.4}$$

where $T_i^{\text{Ch}}(y(x))$ are Chebyshev polynomials in $y(x) = 1 - 2\sqrt{x}$, while the latest CTEQ PDF set [40] uses a polynomial in $y = \sqrt{x}$,

$$\begin{aligned} xg(x, Q_0^2) &= x^{a_1 - 1}(1-x)^{a_2} P_a^g(y)\,,\\ P_a^g(y) &= a_3(1-y)^3 + a_4 3y(1-y)^2 + a_5 3y^2(1-y) + y^3\,. \end{aligned} \tag{2.5}$$

The coefficients $\{A, \eta, \delta, a_{g,i}\}$ of eq. (2.4) and $\{a_i\}$ of eq. (2.5) are the parameters to be determined in the fit. The criterion that guides the choice of the number of parameters used in the fits is to obtain the most flexible parametrizations that at the same time can avoid overfitting, as expanded parametrizations attempt to describe statistical noise. Taking into account the number and momentum sum rules, some parameters can be expressed in terms of the others: the MMHT set has a total of 37 free parameters, while CTEQ has slightly less freedom with a total of 29. The problem of PDF determination reduces then to find the optimal set of parameters which minimizes a suitable measure of the fit quality. The main definitions for such quantity will be discussed in the next sections.

NNPDF

The NNPDF collaboration adopts a completely independent approach to PDF fits respect to the general procedure described above. The first main difference being the choice of the parametrization: NNPDF exploits neural networks to obtain the x-dependence of the parton densities. Neural networks (NNs) allow more flexible and unbiased parametrizations, since they are not limited by specific input functional forms such as a normal parametric model.

A neural network can be viewed as a directed graph where each node is either an *input* or an *activation* node. Each activation node has an associated *activation function*, whose result is used as the input for the next connected node. Each node i has a corresponding threshold θ_i, and each edge connecting the output of the node j to the input of node i has a corresponding weight, w_{ij}. Thresholds θ_i and weights w_{ij} constitute the parameters of a neural network which have to be optimized during a PDF fit.

Let's consider a *feed-forward* neural network, which is the type used by NNPDF. In this case, the NN graph is restricted to be acyclic. Therefore, each node can only belong to one element of an ordered list of *layers*, and the edges can only connect nodes of adjacent layers. The first layer is denoted as the *input layer*, which is directly connected to the input data to be fitted, while the last layer is the *output layer*, and represents the output of the network. Any layer in between is called a *hidden layer*. In this configuration, a node is said *fully connected* when it is connected to all the nodes of the previous and next layer.

The activation function $g(x)$ of a given node i in a layer l takes as input the weighted sum of the nodes outputs $\xi_j^{(l-1)}$ of the previous layer, to produce the result

$$\xi_i^{(l)} = g\left(\sum_{j=1}^{\text{inputs}} w_{ij}^{(l)} \xi_j^{(l-1)} + \theta_i^{(l)}\right), \tag{2.6}$$

where the indexes i, j run over the nodes of layer l and $(l-1)$ respectively.

With NNPDF methodology, each PDF of the fit basis is parametrized with a feed-forward, fully connected, multi-layer perceptron, with architecture 2-5-3-1 shown in fig. 2.2. The first layer contains two nodes of input x and $\ln(x)$. The two hidden layers of 5 and 3 artificial neurons use the most common activation function,

$$g(x) = \frac{1}{1+e^{-x}}, \tag{2.7}$$

which is the sigmoid or *logistic* activation. The output layer instead has a simple linear activation, $g(x) = x$, allowing the final result of the network to acquire values outside the range $(0, 1)$.

Since the latest release NNPDF 3.1, the total charm distribution c^+ has been independently parametrized, which extends the fit basis to 8 independent parton densities. With the architecture described above used for each PDF, the total number of free parameters reaches 296, much more than what can be used with fixed functional forms as eqs. (2.4) and (2.5). Thanks to the large flexibility offered by neural networks, in this case the choice

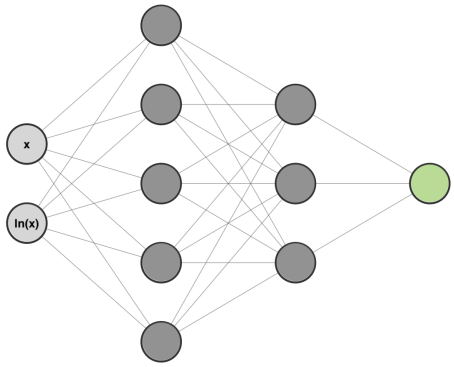

Figure 2.2: Default neural network architecture used by NNPDF: a feed-forward, fully connected, multi-layer perceptron, where the input layer (light gray) is composed of two nodes of input x and $\ln(x)$. The two hidden layers (dark gray) have logistic activation functions, while the output layer (green) has a linear activation.

of fit basis has little effect on the resulting PDFs. The default basis is then one that diagonalizes as much as possible the DGLAP evolution equations:

$$\begin{aligned} & g\,, \\ & \Sigma = u^+ + d^+ + s^+ \\ & V = u^- + d^- + s^-\,, \\ & V_3 = u^- - d^-\,, \\ & V_8 = u^- + d^- - 2s^-\,, \\ & T_3 = u^+ - d^+\,, \\ & T_8 = u^+ + d^+ - 2s^+\,, \\ & c^+\,, \end{aligned} \tag{2.8}$$

where $f_i^{\pm} = f_i \pm \bar{f}_i$ as defined in eq. (1.64).

Each PDF combination eq. (2.8) is parametrized at the input scale $Q_0 = 1.65\,\text{GeV}$ as

$$f_i(x, Q_0) = A_i x^{1-\alpha_i}(1-x)^{\beta_i}\text{NN}_i(x)\,, \tag{2.9}$$

where A_i is a normalization constant used to enforce the values of the sum rules, NN_i denotes the output of the neural network parametrization, and $x^{1-\alpha_i}(1-x)^{\beta_i}$ is a *preprocessing factor* used to speed up the convergence of the fit. This simple polynomial is used to ensure that the NN predictions don't deviate too much from the expected behaviour of the PDFs at small and large-x, while the exponents α_i and β_i provide the correct integrability of the distributions. Their values are therefore randomly chosen within an optimized range at the beginning of the fit and then kept fixed. The preprocessing factor is in practice a mandatory theoretical constraint, as it forces the PDF predictions in the small- and large-x regions to follow a polynomial behaviour, where the lack of experimental data leaves too much freedom in parameter space.

2.2.2 Measure of fit quality and minimization

After the selection of the datasets and the choice made for the parametrization of the PDFs, a meaningful fitting procedure requires the definition of a measure of fit quality, which is usually denoted as χ^2. In general, a global fit quality is defined as the sum over the fit quality of individual datasets d,

$$\chi^2 = \sum_d \chi^2_d \,. \tag{2.10}$$

In the NNPDF approach the full covariance matrix of data, $(\text{cov})_{ij}$, is used to construct the χ^2, considering correlations within and between all the datasets included (the explicit form will be given below). A natural candidate to measure the fit quality is then given by

$$\chi^2 = \sum_{i,j}^{N_{\text{dat}}} (D_i - T_i)(\text{cov}^{-1})_{ij}(D_j - T_j) \,, \tag{2.11}$$

where T are the theoretical predictions computed from the neural network parametrization, D the corresponding experimental measurements, and $(\text{cov}^{-1})_{ij}$ the inverse of the covariance matrix between data points i and j. Within NNPDF determinations, a dataset may be included in a fit if the full experimental correlations are available.

Other groups adopt an alternative, yet numerically equivalent, definition of the fit quality. For instance, when the individual sources of correlated errors are provided, the MMHT collaboration uses the following expression

$$\chi^2 = \sum_i^{N_{\text{dat}}} \left(\frac{D_i + \sum_k^{N_{\text{corr}}} r_k \sigma^{\text{corr}}_{i,k} - T_i}{\sigma^{\text{uncorr}}_i} \right)^2 + \sum_k^{N_{\text{corr}}} r_k^2 \,. \tag{2.12}$$

The term σ^{uncorr}_i is built as the sum in quadrature of the statistical and uncorrelated systematic errors. Instead, systematic uncertainties associated with N_{corr} sources may induce correlated variations (shifts) in the experimental data points. Their effect is then modeled by allowing the data D_i to shift by some multiple r_k of the correlated systematic uncertainties, $\sigma^{\text{corr}}_{i,k}$, in order to give the best fit. By a common assumption, each r_k follows the standard normal distribution, while its deviation from $r_k = 0$ incurs a penalty contribution r_k^2 to χ^2, as defined by the rightmost term of eq. (2.12). In particular, the correlated errors are combined multiplicatively, that is $\sigma^{\text{corr}}_{i,k} = \beta^{\text{corr}}_{i,k} T_i$, where $\beta^{\text{corr}}_{i,k}$ are the percentage errors.

The optimal shifts of data points are solved analytically by minimizing the χ^2 with respect to r_k, while the input PDF parameters must then be determined by numerical minimization of the χ^2. A similar measure of fit quality is also used by the CTEQ group.

Multiplicative uncertainties

The value of the χ^2 estimator depends on the assumed functional form in the presence of experimental systematic uncertainties. Even with the same χ^2 measure, the treatment of

multiplicative uncertainties may result in substantial deviations when different definitions of the covariance matrix are used.

The full experimental uncertainty information is encoded in a sum of three contributions for each data point: uncorrelated errors σ_i^{uncorr}, constructed by adding the statistical and uncorrelated systematic uncertainties in quadrature; correlated additive systematic errors $\sigma_{i,k}^{\text{add}}$; correlated multiplicative systematic errors $\sigma_{i,k}^{\text{mul}}$. The covariance matrix used in eq. (2.11) can then be defined as

$$(\text{cov})_{ij} = \delta_{ij}\sigma_i^{\text{uncorr}}\sigma_j^{\text{uncorr}} + \sum_k^{N_{\text{add}}} \sigma_{i,k}^{\text{add}}\sigma_{j,k}^{\text{add}} + \left(\sum_k^{N_{\text{mul}}} \sigma_{i,k}^{\text{mul}}\sigma_{j,k}^{\text{mul}}\right) D_i D_j \,, \tag{2.13}$$

and its value is unambiguously defined by the experimental results.

However, this *experimental* prescription [71] is unreliable for direct use within a fitting procedure. In fact, it is known that the theoretical values determined from the minimization of the χ^2 with the covariance matrix eq. (2.13) are systematically shifted below the true value. This effect, known as the *D'Agostini bias* [72], is due to the overall normalization uncertainties associated with each experiment. Normalization uncertainties are usually multiplicative, in the sense that each data point within a set has a normalization uncertainty proportional to the measurement at that point. Using the complete covariance matrix leads to a substantial bias in the fitted values due to the fact that smaller data points are assigned a smaller uncertainty than larger ones. In particular, the resulting systematic underestimate of the fit worsens as the number of points that share the same multiplicative error increases.

A general treatment of multiplicative uncertainties which is always free from any bias was developed by the NNPDF collaboration and adopted since the NNPDF2.0 release [73]. With this method the covariance matrix is built with the so-called t_0-prescription,

$$(\text{cov}_{t_0})_{ij} = \delta_{ij}\sigma_i^{\text{uncorr}}\sigma_j^{\text{uncorr}} + \sum_k^{N_{\text{add}}} \sigma_{i,k}^{\text{add}}\sigma_{j,k}^{\text{add}} + \left(\sum_k^{N_{\text{mul}}} \sigma_{i,k}^{\text{mul}}\sigma_{j,k}^{\text{mul}}\right) T_i^{(0)} T_j^{(0)} \,, \tag{2.14}$$

where $T_i^{(0)}$ (the theory prediction for the associated data point) is used to define the normalization contribution to the χ^2. Since the theory predictions are not subject to the same fluctuations as the data, the definition eq. (2.14) has the advantage of avoiding the D'Agostini bias when performing a PDF fit. In practice, the t_0 set is determined consistently via an iterative procedure that updates the covariance matrix eq. (2.14) by taking the central values of the previous fit as the new t_0 set.

Minimization

With the measure of fit quality constructed, a PDF determination is now an optimization problem whose solution is the optimal set of free parameters of the PDF basis which minimize the figure of merit χ^2. Even for those groups utilizing a fixed parametrization, this task is numerically challenging due to the combination of a large amount of data points, $\mathcal{O}(10^3)$, and the moderate number of free parameters, $\sim 50/60$. A common choice

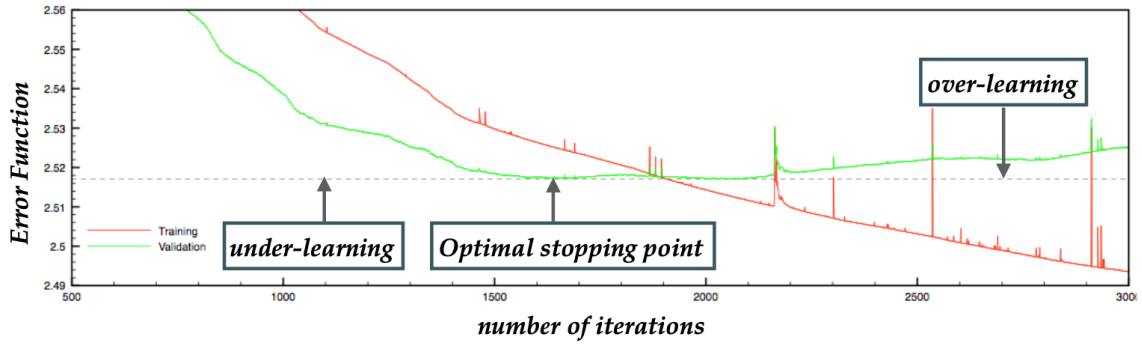

Figure 2.3: Schematic representation of the look-back cross-validation stopping criterion used in the NNPDF fits. The optimal stopping point corresponds to the iteration where the validation loss reaches its minimum value. Figure taken from Ref. [75].

to perform the minimization is the MINUIT [74] package, as done by the CTEQ group in the CT18 analysis. It involves numerical calculations of the first- and second-order derivatives of χ^2, combined with sequential minimum searches along fixed directions in the PDF parameter space.

In the NNPDF case, the minimization is complicated by the very large number of parameters available and the non-linear relation between the fitted experimental data and the input PDFs, making conventional methods based on gradient descent impractical[1]. These difficulties are overcome with the implementation of *genetic algorithms* (GAs) for the minimization procedure, which are particularly useful to explore complex parameter spaces and require the knowledge of the χ^2 local values only.

A further issue to be addressed during the minimization is the possibility that one might end up fitting point-to-point fluctuations, due to the highly redundant parametrization offered by the neural network shown in fig. 2.2. This phenomenon of learning the statistical fluctuations of an input set rather than the underlying law that produced it is known as *overfitting* or *overlearning*, and is a problem often encountered while training large neural networks. Within NNPDF determinations, the strategy adopted to identify and avoid overlearning is the *look-back cross-validation stopping* criterion, illustrated schematically in fig. 2.3.

With this method, the input experimental measurements are randomly divided into two separate sets: the *training set*, used for the actual minimization of the χ^2 (called *error/loss function* in machine learning applications), and the *validation set*, which plays the role of a control sample to monitor and validate the training progress. At each iteration of the GA minimization, the error function between the NN predictions and both sets is computed. During the early stages, both values of the error functions should decrease, as the neural network is learning the underlying law. However, after a certain number of iterations, the

[1]The development of new technologies has now made possible a deterministic minimization using gradient descent. A new methodology implemented within the NNPDF collaboration will be presented in section 3.1.

χ^2 calculated to the training set may continue to decrease while the value computed to the validation set has stopped decreasing or even begun to increase. This behaviour is indeed indicating that what is being learned from the training sample is not present in the validation one (namely the fluctuations). Then, the optimal stopping point is defined as the global minimum of the χ^2 validation set, computed over a large fixed number of iterations.

Adoption of this stopping criteria has been made possible with greater computing efficiencies, since NNPDF3.0 [76] (2015). The method in fact requires reaching the maximum number of iterations for all replicas, out of which the absolute minimum is determined. This maximum must be chosen to be large enough that the absolute minimum is always reached, and it therefore leads on average to longer training.

2.2.3 Error propagation

Once the minimization procedure has produced the "best prediction" of the true PDFs, a method to estimate their uncertainties is needed for a meaningful interpretation of the measured observables. Ideally, one would like to determine a representation of the PDF probability distribution in the whole functional space. That is, given a dataset d, we would like to find the probability density, $\mathcal{P}(f|d)$, of a certain PDF candidate f such that our fitted PDF central value is given by

$$\langle f \rangle (x) = \int \mathcal{D}f \, f(x) \mathcal{P}(f|d) \,, \tag{2.15}$$

with a variance

$$\mathrm{Var}[f](x) = \int \mathcal{D}f \, \left[f(x) - \langle f \rangle (x) \right]^2 \mathcal{P}(f|d) \,. \tag{2.16}$$

The generalization for an observable $\mathcal{O}[f]$ is then straightforward.

The probability density $\mathcal{P}(f|d)$ is however a difficult quantity to calculate. In the following, we will describe the two main methods used to provide an estimate of PDF uncertainties.

The Hessian method

The Hessian method is the most widely used method of error propagation. In a nutshell, this approach is based on examining the variations of the χ^2 induced by displacements of the fit parameters $\vec{\boldsymbol{a}}$ near the optimal values, denoted as $\vec{\boldsymbol{a}}_0$, that minimize the χ^2. A tolerance for the χ^2 variation is then chosen, and the errors on the observables are calculated from PDFs obtained using the displaced parameters based on the selected value of the tolerance.

More quantitatively, the basic assumption of the Hessian method is a quadratic approximation of the χ^2 in the neighborhood of the minimum $\chi^2(\vec{\mathbf{a}}_0)$,

$$\Delta\chi^2(\vec{\mathbf{a}}) := \chi^2(\vec{\mathbf{a}}) - \chi^2(\vec{\mathbf{a}}_0) = \sum_{i,j=1}^{n} (a - a_0)_i H_{ij} (a - a_0)_j \,, \tag{2.17}$$

where $(a - a_0)_i$ is the i-th component of the displacement from the best parameter set $\vec{\boldsymbol{a}}_0$. The symmetric matrix H_{ij} in eq. (2.17) is the Hessian matrix,

$$H_{ij} = \frac{1}{2} \left. \frac{\partial^2 \chi^2}{\partial a_i \partial a_j} \right|_{\vec{\boldsymbol{a}}=\vec{\boldsymbol{a}}_0} , \tag{2.18}$$

evaluated at the minimum of $\chi^2(\vec{\boldsymbol{a}})$. Early Hessian uncertainty estimates were based upon standard linear error propagation,

$$(\Delta X)^2 = T^2 \sum_{i,j=1}^{n} \frac{\partial X}{\partial a_i} C_{ij} \frac{\partial X}{\partial a_j} , \tag{2.19}$$

where X is a generic quantity that depends on the PDFs, $T^2 = \Delta\chi^2$ is the tolerance of the χ^2 variation and $C = H^{-1}$ the covariance matrix in parameter space, equivalent to the inverse Hessian matrix. This procedure is however numerically inefficient since it requires the computation of the partial derivatives of X with respect to the fit parameters. To overcome the issue a different geometrical method [77, 78] was developed by the CTEQ collaboration.

Within this method, the eigenvectors of the Hessian matrix are exploited to obtain the estimate of the uncertainties. The Hessian matrix eq. (2.18) has in fact a complete set of orthonormal eigenvectors $\{\vec{\boldsymbol{v}}_k\}_{k=1,\dots,n}$ defined by

$$H\vec{\boldsymbol{v}}_k = \epsilon_k \vec{\boldsymbol{v}}_k , \tag{2.20}$$

$$\vec{\boldsymbol{v}}_k \cdot \vec{\boldsymbol{v}}_j = \delta_{kj} , \tag{2.21}$$

where ϵ_k are the correpsonding eigenvalues. It is convenient to define the rescaled eigenvectors $\vec{\boldsymbol{e}}_k = \vec{\boldsymbol{v}}_k / \sqrt{\epsilon_k}$ and expand the displacement $\vec{\boldsymbol{a}} - \vec{\boldsymbol{a}}_0$ in this new basis, in order to write it in the simple form

$$\vec{\boldsymbol{a}} - \vec{\boldsymbol{a}}_0 = \sum_{k=1}^{n} \vec{\boldsymbol{e}}_k z_k , \tag{2.22}$$

where z_k are the expansion coefficients. Then, substituting eq. (2.22) in the quadratic approximation eq. (2.17), the $\Delta\chi^2$ reduces to

$$\Delta\chi^2(\vec{\boldsymbol{a}}) = \chi^2(\vec{\boldsymbol{a}}) - \chi^2(\vec{\boldsymbol{a}}_0) = \sum_{k=1}^{n} z_k^2 . \tag{2.23}$$

Equation (2.23) defines the interior of a hypersphere of radius $\sqrt{\Delta\chi^2}$ centered in $\vec{\boldsymbol{a}}_0$ in the parameter space generated by the rescaled eigenvector basis, which corresponds to the variation of the parameters consistent with the tolerance $T = \sqrt{\Delta\chi^2}$ in the quadratic approximation. Therefore, the particular choice of T defines the region of acceptable fits, since eqs. (2.17) and (2.23) imply that a shift in the best fit parameters may induce an increase $\Delta\chi^2$ that can be at most T^2.

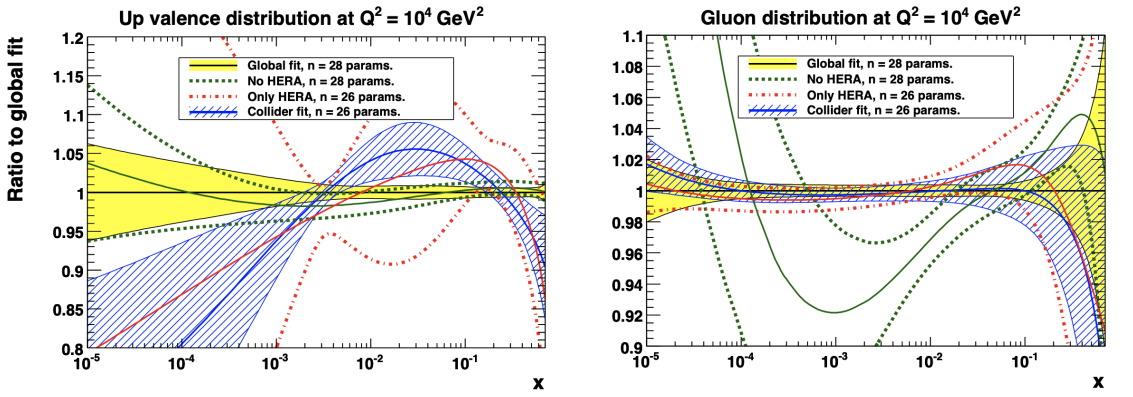

Figure 2.4: Ratio to global fit (yellow) of the up-valence (left plot) and gluon (right plot) distributions obtained from different datasets, HERA only (red), no HERA (green), Collider only (blue), by MSTW2008. Figures taken from Ref. [80].

It is now possible to construct $2n$ eigenvector PDF sets, $S_k^{\pm}$, to span this hypersphere, with parameters specified by the boundaries of the hypersphere volume,

$$\vec{a}(S_k^{\pm}) = \vec{a}_0 \pm t\,\vec{e}_k\,, \tag{2.24}$$

i.e. each parameter of the $S_k^{\pm}$ set is displaced by an amount t "up" or "down" the direction of the eigenvector $\vec{e}_k$. In the quadratic approximation $t = T$, while an iterative procedure is applied to obtain the target $\Delta\chi^2$ when this approximation is not valid. Finally, the error on a generic quantity X that depends on the PDFs is given by

$$(\Delta X)^2 = \frac{1}{2}\sum_{i=1}^{n}\left(X(S_k^+) - X(S_k^-)\right)^2\,. \tag{2.25}$$

Equation (2.25) is also valid when X is a PDF, allowing therefore to compute the errors of the parton densities themselves.

The choice of the tolerance value follows from the "parameter fitting" criterion [79]: in the ideal case of consistent datasets and Gaussian errors, the 68% confidence interval of a parameter is given by the value which induces a variation of one unit in the χ^2, that is $T = 1$. However, in the context of global PDF fits, this value often leads to an underestimation of the uncertainties. In practice, larger values of tolerance are usually used, such as $T = \sqrt{50}$ or even $T = \sqrt{100}$ for the 90% confidence intervals, even if Gaussian statistics would require $T = \sqrt{2.7}$.

The motivations for larger values come from two distinct aspects: the first is the incompatibility between PDFs fitted from different datasets [81], which could be due to unknown systematic uncertainties in the experiments, or to theoretical errors as missing higher order uncertainties. For instance, in fig. 2.4 the up-valence and gluon distributions of the MSTW2008 set show that the uncertainty bands from fits to subsets do not always overlap with a global fit. The second problem is the parametrization bias, which originates in

representing unknown functions as the PDFs in terms of expressions that depend only on a finite number of free parameters [82]. In this case, more generic parametrizations are exploited, as outlined in section 2.2.1 for the MMHT and CTEQ collaborations.

In the MSTW2008 [83] set, the use of a global tolerance T was refined by choosing separately for each eigenvector direction a value T_k with a dynamic determination according to the weaker "hypotesis testing" criterion [79]. With this criterion, a fit is judged to be "good" if each data set d, consisting of N data points, has $\chi^2_d \simeq N \pm \sqrt{2N}$. More precisely, ranges of χ^2_d corresponding to a 90% c.l. limit, for example, can be calculated, then the value of the tolerance $T = \sqrt{\Delta\chi^2}$ can be chosen to ensure that each data set is described within its 90% C.L. limit. On average, the dynamic tolerance reduced the values to estimate the 68% and 90% confidence intervals to $T \approx 3$ and $T \approx 6$ respectively. The same approach is then inherited by the MMHT set.

More specialized fits, such as ABM or the HERAPDF, based upon relatively small datasets, may use the standard tolerance of $\Delta\chi^2 = 1$, thanks to the restricted number of experimental points leading to fewer conflicts that would require an inflation of the tolerance.

In general, the uncertainties produced via the Hessian procedure are difficult to examine in a statistical sense, due to the large deviation from the expected value $\Delta\chi^2 = 1$ and the approximations made in the procedure. It is therefore difficult to find a representation in the Hessian approach of the full probability distribution $\mathcal{P}(f|d)$.

The Monte Carlo method

The second method of PDF error propagation is the Monte Carlo method, designed to faithfully represent the uncertainties present in the initial data, and to propagate the errors in a way that does not assume anything of the nature of the error propagation.

In the Monte Carlo procedure, for each data point in the fit, an ensemble of N_{rep} artificial data, called *pseudo-data replicas*, is generated according to the probability distribution of the input data. Usually, in NNPDF fits this distribution is a multi-Gaussian defined as

$$D_i^{(\text{art})(k)} = D_i^{(\text{exp})} + \sum_{j=1}^{N_{\text{dat}}} [\text{Chol}(\text{cov}_{t_0})]_{ij}\, \mathcal{N}(0,1)\,, \tag{2.26}$$

where $\text{Chol}(\text{cov}_{t_0})$ is the transpose of the Cholesky decomposition of the covariance matrix based on the t_0-prescription eq. (2.14), and $\mathcal{N}(0,1)$ is a random number sampled from a standard normal distribution to generate fluctuations of the artificial data around the experimental central value $D_i^{(\text{exp})}$. It was shown in Ref. [84] that $\mathcal{O}(1000)$ replicas is needed to reproduce the mean values, the variances, and the correlations of the original experimental data at the percent level of accuracy.

Therefore, when a fair amount of pseudo-data is sampled, instead of performing just one fit to the data, each of the N_{rep} replicas is independently fitted to minimize the measure of fit quality eq. (2.11) with respect to the neural network parameters, that is the weights and biases (see section 2.2.1). In this specific case, the measure of fit quality may be written

as:

$$\chi^{2(k)} = \frac{1}{N_{\rm dat}} \sum_{i,j}^{N_{\rm dat}} \left(D_i^{({\rm art})(k)} - T_i^{({\rm NN})(k)} \right) (\mathrm{cov}_{t_0}^{-1})_{ij} \left(D_j^{({\rm art})(k)} - T_j^{({\rm NN})(k)} \right) , \tag{2.27}$$

where $T_i^{({\rm NN})(k)}$ is the theoretical prediction obtained from the NN parametrization for the i-th data point of the replica k, and cov_{t_0} is again the covariance matrix in the t_0-prescription. Theory uncertainties (such as missing higher order uncertainties) can also be included as presented in Ref. [85], but this has only been done in preliminary PDF sets so far.

The outcome of the fitting procedure is an ensemble of $N_{\rm rep}$ PDFs, $\{f^{(k)}\}_{k=1,\ldots,N_{\rm rep}}$, named *PDF replicas* or simply *replicas*, which faithfully describe the probability distribution of the PDFs based upon the original experimental uncertainties. Central value and uncertainty of a quantity X which depends on the PDFs are now calculable by mere average and standard deviation over the replica sample,

$$\langle X \rangle = \frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} X^{(k)} , \tag{2.28}$$

$$\sigma^2[X] = \frac{1}{N_{\rm rep} - 1} \sum_{k=1}^{N_{\rm rep}} (X^{(k)} - \langle X \rangle)^2 , \tag{2.29}$$

where $X^{(k)}$ denotes the quantity X evaluated to the PDF replica k. Similarly to eq. (2.25), the quantity X can also be a PDF itself, so that central values and uncertainties of the PDFs can be computed directly from the replica values $f^{(k)}$.

The Monte Carlo method gives therefore a discrete representation of the underlying probability distribution $\mathcal{P}(f|d)$, so that the uncertainties from the experimental data to the PDFs are propagated without the need for a linear error propagation assumption, or the introduction of a tolerance in order to define the region of acceptable fits. Figure 2.5 presents a Monte Carlo ensemble of 1000 replicas for the gluon distribution: the left plot shows all the PDF replicas and the central value of the set, while the right plot shows the uncertainty bands.

2.3 Monte Carlo to Hessian conversion

2.3.1 Introduction

Unlike the Hessian approach, where one assumes that the distributions around the best fit are Gaussian and the uncertainties can be calculated from linear error propagation, the Monte Carlo approach can provide a representation of the PDFs uncertainties even when non-Gaussian effects become important, in regions where there is lack of experimental data points (namely at small- and large-x). A practical way to inspect whether a set of replicas present non-Gaussian errors is to compare the one standard deviation band and

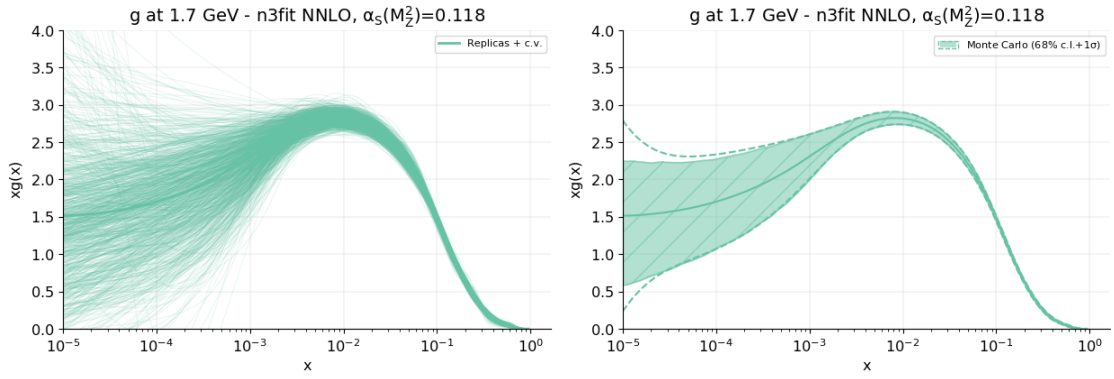

Figure 2.5: A Monte Carlo ensemble of 1000 gluon replicas obtained from a "global" fit, namely using all datasets included in NNPDF3.1, at NNLO and $\alpha_S(M_Z^2) = 0.118$. Each line on the left plot shows a different replica, while the nearly visible bold line is the central value of the set. The right plot shows, in addition to the central value, the standard deviation (dashed line), and the 68% c.l. band.

the 68% c.l. band. The latter is defined as the symmetric interval which contains the 68% of the replica sample around the central value of the set. Since a Gaussian distribution would require these two bands to coincide, a discrepancy between them is a symptom of non-Gaussian effects. For instance, the right plot of fig. 2.5 shows that, in this particular global fit, non-Gaussian errors in the gluon distribution are found only in the region of extrapolation $x \lesssim 10^{-4}$.

Even though the Hessian and Monte Carlo representations provide uncertainties in a very different manner, they should nonetheless lead to the same description of the parton content of the proton, and in fact it was firstly shown it is possible to convert an Hessian PDF set into a Monte Carlo representation [80]. Basically, the Hessian parameter space is randomly sampled by allowing Gaussian deviations around the standard values eq. (2.24). Then, a set of PDF replicas is built from the set of Hessian parameters using the fixed Hessian parametrization. The reverse operation, namely the conversion of a Monte Carlo set into an Hessian representation requires a more careful consideration.

The advantage of the Hessian methodology is that the errors can be interpreted in terms of continuous parameters variations. In fact, the eigenvectors of the Hessian matrix may be treated as nuisance parameters to quantify how much a subset of parameters affects the fit to a dataset or the predictions of an observable. On the other hand, with the Monte Carlo approach there is no need for any assumption on the statistical distribution followed by the parameters, and also, PDFs can be parametrized with more general unbiased functional forms, such as neural networks, with a large number of parameters.

However, when the Gaussian approximation is reasonably accurate, that is, when PDF uncertainties are small and driven by abundant experimental data, it should be possible to obtain a meaningful Hessian representation of a Monte Carlo set. This goal was indeed achieved with the implementation of the `mc2hessian` code [5]. In the following will be

described the Singular Value Decomposition (SVD) version, combined with the Principal Component Analysis (PCA) [6]

2.3.2 The SVD + PCA method

The Hessian conversion of a Monte Carlo set is based on the construction of a multi-Gaussian covariance matrix in PDF space, whose eigenvectors give directly the Hessian set members. The fundamental assumption of this procedure is that the central value of the resulting Hessian set coincides with the prior Monte Carlo set.

All the necessary information is encoded in the $N_x N_f \times N_{\rm rep}$ rectangular matrix X which samples the difference between each replica, $f_\alpha^{(k)}(x_i, Q)$, and the central value, $f_\alpha^{(0)}(x_i, Q)$, of the set:

$$X_{lk}(Q) := f_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q)\,, \tag{2.30}$$

where α runs over the N_f independent flavours at the energy scale Q, i runs over the N_x x-grid points where the PDFs are sampled, $l = N_x(\alpha - 1) + i$ runs over the $N_x N_f$ grid points $x \times flavour$, and k runs over the $N_{\rm rep}$ replicas. The covariance matrix in PDF space is then built as

$$\mathrm{cov}(Q) = \frac{1}{N_{\rm rep} - 1} X X^t\,. \tag{2.31}$$

Assuming $N_{\rm rep} > N_x N_f$, the eigenvectors of the $N_x N_f \times N_x N_f$ covariance matrix eq. (2.31) can be represented as linear combinations of the $N_{\rm rep}$ replicas, by using the SVD of the sampling matrix eq. (2.30):

$$X = U \Sigma V^t\,, \tag{2.32}$$

where U and V are orthogonal matrices, with dimensions $N_x N_f \times N_x N_f$ and $N_{\rm rep} \times N_{\rm rep}$ respectively. Σ is a diagonal positive semi-definite matrix with dimension $N_x N_f \times N_{\rm rep}$ and whose entries, called singular values of X, are the square roots of the eigenvalues of XX^t. From eq. (2.32) follows that

$$X X^t = U(\Sigma \Sigma^t) U^t = U \Sigma^2 U^t\,, \tag{2.33}$$

which means the columns of U are the orthogonal eigenvectors of the covariance matrix that are needed to construct the Hessian set members. Then, the matrix $Z = U\Sigma$ yields the representation of the multigaussian covariance matrix in terms of the original PDF replicas, since

$$Z Z^t = X X^t\,, \tag{2.34}$$

and also

$$Z = XV\,, \tag{2.35}$$

which means that V_{kj} is the expansion coefficient of the j-th eigenvector along the k-th replica. From now on the singular values of Σ can be assumed in decreasing order starting from the first diagonal entry.

Since the number of Hessian eigenvectors $N_{\rm eig} = N_x N_f$ is generally large, it is convenient to retain only $\widetilde{N}_{\rm eig}$ principal components, *i.e.* the eigenvectors relative to the largest

singular values. This procedure is implemented by the PCA optimization, which consists in replacing U and Σ by their submatrices u and σ with dimensions $N_x N_f \times \widetilde{N}_{\rm eig}$ and $\widetilde{N}_{\rm eig} \times N_{\rm rep}$ respectively, where $\widetilde{N}_{\rm eig} < N_{\rm eig}$. Now σ has only $\widetilde{N}_{\rm eig}$ non vanishing diagonal entries and therefore only the $N_{\rm rep} \times \widetilde{N}_{\rm eig}$ submatrix of V contributes in eq. (2.32). This is referred to as principal submatrix P of V, and so the optimized set of eigenvectors is found by using P in place of V in eq. (2.35). The final $\widetilde{N}_{\rm eig}$ eigenvectors are given by

$$\tilde{f}_\alpha^{(k)}(x_i, Q) = f_\alpha^{(0)}(x_i, Q) + \frac{1}{\sqrt{N_{\rm rep} - 1}} (XP)_{lk} \,, \quad k = 1, \ldots, \widetilde{N}_{\rm eig} \,, \tag{2.36}$$

and the uncertainties of the resulting Hessian set may be computed similarly to eq. (2.25),

$$\sigma_{H,\alpha}^{PDF}(x_i, Q) = \sqrt{\sum_{k=1}^{\widetilde{N}_{\rm eig}} \left(\tilde{f}_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q) \right)^2} \,. \tag{2.37}$$

This procedure allows to reduce the typically large number of replicas of the starting Monte Carlo set into a smaller set without significant loss of accuracy. Also, the conversion does not depend on the energy scale, since the QCD evolution equations of the Monte Carlo set are automatically satisfied by the Hessian PDFs thanks to linearity.

As a check of this method, in fig. 2.6 is shown a comparison between the starting Monte Carlo representation with $N_{\rm rep} = 1000$ replicas and the final Hessian representation with $N_{\rm eig} = 100$ eigenvectors of the published set `NNPDF31_nnlo_as_0118_1000`. On the left plot is shown the gluon distribution and on the right plot the singlet one, both at $Q^2 = 10^2\,{\rm GeV}^2$ and normalized to the central PDF. It is evident that the Hessian representation faithfully describe the prior Monte Carlo set, with differences in the one-σ uncertainty bands of few percents only at very small- and large-x.

2.3.3 $\Delta\chi^2$ variations of converted Monte Carlo sets

The method presented above offers an unbiased Hessian representation for Monte Carlo PDF sets. If the procedure used to obtain the prior Monte Carlo set has really found the "best true" PDF, and the Hessian conversion introduces only a small loss of information, the resulting Hessian set should be comprised of PDFs which describe positive variation of the χ^2 around the virtual minimum corresponding, by definition, to the Monte Carlo central PDF. Thus, the $\Delta\chi^2$ distribution may be studied from the variation computed for each of the $N_{\rm eig}$ eigenvectors extracted during the Hessian conversion. If from such analysis negative variation in the χ^2 appear, then the optimization procedure which led to the prior Monte Carlo set was actually not the optimal one, as some of the Hessian eigenvectors are describing negative variations $\Delta\chi^2 < 0$. Moreover, a further test of the reliability of the Monte Carlo predictions comes from the expected increment in the χ^2, which, as discussed in section 2.2.3, should be equal to one unit for the 68% c.l. interval. The presence of positive values of $\Delta\chi^2$ but larger or smaller than one are therefore indicators of under- or over-estimation of the uncertainties in a Monte Carlo set.

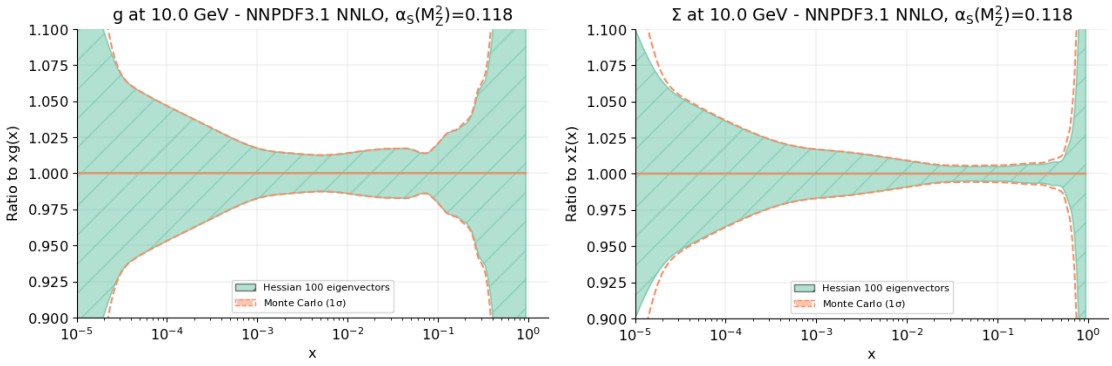

Figure 2.6: Comparison between the gluon (left plot) and singlet (right plot) at $Q = 10\,\text{GeV}$ normalized to the central PDF. The dashed orange line shows the 1σ value of the NNPDF3.1 Monte Carlo set of $N_{\text{rep}} = 1000$ replicas, while the teal band the final Hessian representation of $N_{\text{eig}} = 100$ eigenvectors. The prior distributions are obtained from a global fit with $\alpha_S(M_Z^2) = 0.118$.

In the next chapter, a new fitting methodology (`n3fit`) based on deep learning methods and implemented by the NNPDF collaboration will be introduced. Then, the basic idea of studying χ^2 variations in the context of a Monte Carlo to Hessian conversion will be applied to examine the reliability of the predictions of the old and new methodology.

Chapter 3

PDFs from deep learning methods

The AI-based approach to PDF determination of NNPDF has been developed to eliminate potential source of bias, particularly those related to the choice of functional form. Neural networks are the optimal candidates in such sense, since they provide universal function approximators. However, neural networks themselves are not unique, and neither the algorithms used for their training. The methodology created by NNPDF is an ongoing effort that started more than ten years ago, and is the result of a series of improvements based on trial and error. Nevertheless, the human intervention to tackle the obstacles faced in this complex problem might have been a source of bias. Therefore, the next step towards a new generation of parton density determinations is to extend an unbiased methodology with an unbiased *choice* of fitting methodology. This goal has been recently achieved with the implementation of a new procedure that can optimize the methodology itself, by a so-called hyperoptimization scan. In the following, this new approach will be introduced, while a deeper description can be found in Refs. [4, 86].

3.1 A new approach to the NNPDF fitting methodology

As largely discussed, the NNPDF methodology is based on the Monte Carlo treatment of experimental data, the parametrization of PDFs with artificial neural networks, and the minimization strategy on genetic algorithms. Starting from the release NNPDF3.0 [76], all the code is implemented in C++ and relies on a very small set of external libraries. However, the complex structure of the codebase impairs the study of novel architectures or the introduction of modern machine learning techniques which could lead to an enhancement of the methodology. Furthermore, the use of a GA minimization algorithm is computationally demanding and represent a significant limitation to perform systematic scans in order to optimize the fitting methodology. This problem has been examined in the last two years within the N3PDF project [3], and resulted in the reimplementation of the NNPDF regression model from scratch in a python-based framework.

The neural network capabilities of this new framework are provided by Keras [87] and Tensorflow [88], which are some of the most used and well documented NN libraries. The code can also abstract any dependence on them, so that other machine learning tools may

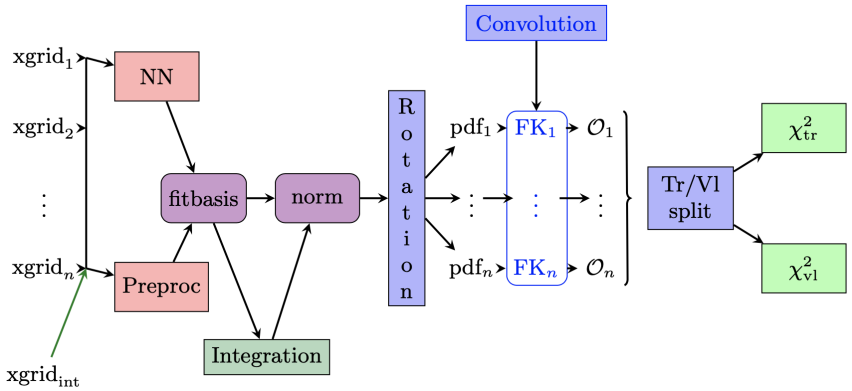

Figure 3.1: Scheme of the `n3fit` code. Each box represents an independent operation, while the red squared ones, namely the neural network and the preprocessing factor, contain the parameters of the PDF fit. Figure taken from Ref. [4].

be implemented. As mentioned before, neural networks are not unique, and the space of hyperparameters (number of layers, nodes per layer, activation functions, optimizer, etc.) is big enough that finding the best choice becomes an overwhelming task. To this purpose, the entire framework is enclosed in hyperoptimization scan routines, implemented with the hyperopt library [89], which allows to systematically scan over many different combinations of hyperparameters and find the optimal configuration for the neural network, given a specific input setup.

The new framework implements gradient descent (GD) methods to replace the previously used genetic algorithm. Thanks to the current technologies, this change reduces the computing cost of a fit while achieving similar or better results in the goodness-of-fit. Particularly, the optimizers which are found most well suited for the fits are Adadelta [90], Adam [91], and RMSprop [92]. The GD methods produce more stable fits than their GA counterparts, and, given the possibility of performing hyperoptimizations, there is no longer a risk of ending up in architecture-dependent local minima.

Concerning the neural network employed in this new methodology, to be sensible to cross-correlation between the different PDFs, all the eight functional forms are now parametrized with a single densely connected network, instead of the single net for each flavour shown in fig. 2.2. As previously done, the first layer is fixed to split the input data into the pair $(x, \ln(x))$, while the PDF basis is still $\{g, \Sigma, V, V_3, V_8, T_3, T_8, c^+\}$. In this case, the output layer is composed of 8 nodes, one per flavour, with linear activation functions. The NN architecture (number of layers and nodes per layer) is now hyperoptimized, rather than being fixed.

In fig. 3.1 is shown a schematic view of the full new methodology which will be referred to as `n3fit`. The vectors $\text{xgrid}_1, \ldots, \text{xgrid}_n$ contain the x-inputs for each dataset entering the fit, and therefore are used to compute both the value of the neural network and preprocessing factor. The normalizations A_i of the PDFs are computed at each step of

the fit with the $\text{xgrid}_{\text{int}}$ points, yielding the normalized distributions eq. (2.9). Since the PDF basis employed is a combination of flavours, to compute physical observables a basis rotation is applied to obtain the physical one $\{\bar{s}, \bar{d}, \bar{u}, g, u, d, s, c(\bar{c})\}$.

Unlike in many standard regression problems, in which during the optimization procedure the model is compared directly to the input data, in PDF fits the data are compared to theoretical predictions for physical observables of the form eq. (1.54). Any observable depends on the PDFs through a number of convolution integrals between the PDFs at the initial scale of parametrization Q_0, the (DGLAP) evolution factors that take them to scale Q and the partonic cross sections. In practice, these convolutions are turned into multiplication of pre-computed tables, called FastKernel (FK) tables, by projecting on suitable basis functions, as discussed in Refs. [76, 93]. In this sense, the FK tables implement a separation between theory and fitting procedure, since the PDFs at the initial scale are varied in order to minimize the χ^2 while the FK tables are always kept fixed and treated as an external input. The predictions of the network for the initial PDFs (the pdf_i of fig. 3.1) are contracted with the FK_i-table, and the resulting observable $\mathcal{O}_i$ is used to compute the effective figure of merit χ^2, eq. (2.27), to eventually update the NN parameters.

To avoid overlearning, the cross-validation of input data is used in combination with a more refined stopping criterion, which calls a patience algorithm after the validation stops improving, and waits for a number of iterations before raising the stopping action. However, since the hyperopt framework is actually performing a higher level optimization, on top of the NN training, it happens that the selection of the best model is affected by the correlations between training and validation sets. This problem is avoided with the introduction of a further quality control for the hyperoptimization scan, which evaluates the NN predictions on test sets never seen before or with a more refined procedure using k-foldings.

Table 3.1 lists some of the main differences between the old and new methodology. In summary, there are two principal reasons that make `n3fit` much more efficient and appealing: firstly, the semi-automatic hyperoptimization performed by the hyperopt library, which is in turn made possible by the second feature, that is, limited computing resource usage thanks to GD optimizers (see Ref. [94]). This efficiency is in fact exploited to test hundreds of architectures in the same amount of time that it takes for a single fit of the old methodology to complete.

Component	NNPDF3.1	n3fit
Neural Net	fixed architecture, per flavour	**single net, flexible architecture**
Preprocessing	random fixed	random fixed, **fitted in range**
Optimizer	genetic optimizer	**gradient descent**
Stopping	look-back	**patience**
Fine tuning	manual	**semi-automatic**
Model selection	closure test	closure test, **hyper optimization**

Table 3.1: Component by component comparison of the different strategies adopted in the old and new methodology.

3.2 $\Delta\chi^2$ analysis for NNPDF3.1 and `n3fit`

We shall now consider the Monte Carlo to Hessian conversion of two equivalent sets of NNPDF3.1 and `n3fit` to compare the $\Delta\chi^2$ distributions, as explained in section 2.3.3. With "equivalent sets" we mean same theory parameters (initial scale, quark masses and scheme, coupling value, ...) and same fraction of training/validation split for the datasets which enter in the fits. Also, even though `n3fit` allows to vary the preprocessing exponents α_i, β_i during the fit, we have left them fixed as in NNPDF3.1.

We therefore consider the following two Monte Carlo sets, both comprised of $N_{\text{rep}} = 1000$ replicas at NNLO and with $\alpha_S(M_Z^2) = 0.118$:

- `NNPDF31_nnlo_as_0118_1000`;
- `PN3_Global_nonfittedprepro_1000`.

The first set was published in the latest release of the NNPDF collaboration [43], while the second has been obtained by running a 1000 replica fit of `n3fit` with a hyperoptimized configuration[1].

3.2.1 Gaussian error deviation

Since we want to obtain a Hessian representation of these Monte Carlo sets, the question arises whether we may carry out the conversion if the prior uncertainties are not Gaussian. It is thus necessary to quantify the deviations from a Gaussian behaviour in order to make sure that the procedure can be consistently applied. This is achieved by considering the simplest indicator, that is the second moment of the probability distribution, and thus we compare the one-sigma and 68% c.l. intervals. In figs. 3.2 and 3.3 is shown a comparison of these intervals for some PDFs at $Q = 1.7\,\text{GeV}$ from the NNPDF3.1 and `n3fit` set, respectively. We observe that typically the intervals do not coincide in regions of extrapolation, namely at small- and large-x, being the uncertainties determined by

[1] In appendix A we present a brief comparison between the Monte Carlo sets we use in this thesis, along with their Hessian representations.

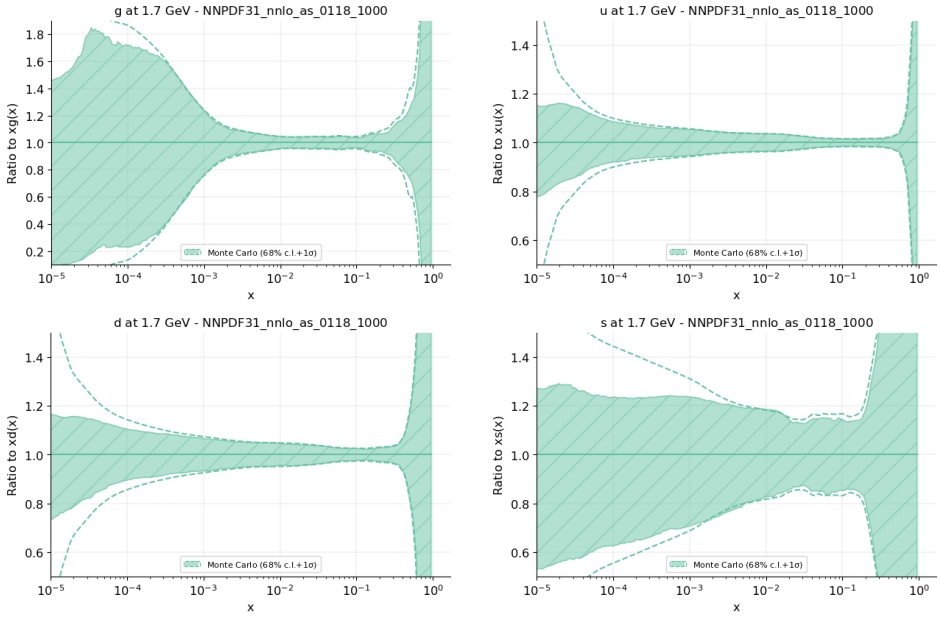

Figure 3.2: Comparison of one-sigma (dashed line) and 68% c.l. (teal band) interval for the gluon and the three lightest quarks, from left to right and top to bottom, for the NNPDF3.1 Monte Carlo set. The values are normalized to the central PDF.

theoretical constraints (sum rules and cross-section positivity) due to lack of experimental data. At first sight, we can state that the `n3fit` PDFs are less subject to large uncertainties in these regions and, moreover, non-Gaussian effects appear to be smaller than those of NNPDF3.1.

More precisely, to take into account deviations from a Gaussian probability distribution, the following figure of merit is defined:

$$\epsilon_\alpha(x_i, Q) = \frac{|\sigma_\alpha(x_i, Q) - \sigma_\alpha^{68\%}(x_i, Q)|}{\sigma_\alpha^{68\%}(x_i, Q)}\,, \quad \alpha = 1, 2, \dots N_{\text{flv}}\,, \tag{3.1}$$

where $\sigma_\alpha(x_i, Q)$ and $\sigma_\alpha^{68\%}(x_i, Q)$ are respectively the one-sigma and 68% c.l. intervals evaluated at x_i and scale Q for the α-th flavour. Thus, a threshold value ϵ, independent from x, is chosen such that all points for which $\epsilon_\alpha(x_i, Q) > \epsilon$ are discarded. After that, the Hessian conversion may be carried out from the sampling matrix eq. (2.30) without these outliers. Of course, the choice of ϵ is dictated by the compromise to include only grid points for which the Gaussian approximation is valid, without loosing too much information.

In fig. 3.4 is shown the estimator $\epsilon_\alpha(x_i, Q)$ for both Monte Carlo sets[2] at the scale $Q = 1.7\,\text{GeV}$ at which the conversion is accomplished. The horizontal line marks the threshold value $\epsilon = 0.25$ chosen for the `n3fit` set, while the value used instead for NNPDF3.1 is slightly greater, $\epsilon = 0.30$ (not shown), just to ensure that some of the grid points corresponding to the sharp peaks at large-x in the gluon and anti-up quark distributions could

[2]The blue lines are the results for another `n3fit` set that will be described in section 3.2.2.

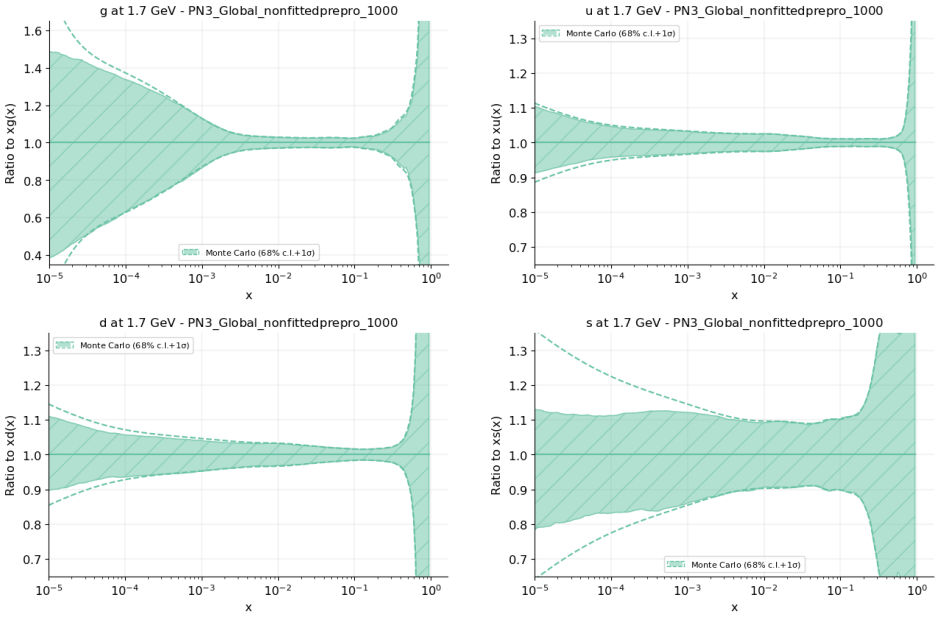

Figure 3.3: Same as fig. 3.2 but for the `n3fit` set.

be included. The charm quark shows in both sets the largest deviations from Gaussian-like errors for $x \lesssim 10^{-3}$. The threshold values chosen allow in any case to retain grid points in the bulk of experimental data where the charm distribution is more likely to be well behaved. The observation we made before by looking at the plots of figs. 3.2 and 3.3 is then confirmed by these results: the NNPDF set is more inclined to present non-Gaussian effects, with a peculiar sharp structure in $\epsilon_\alpha(x_i, Q)$ at large-x for the gluon distribution.

We have then produced the Hessian conversion of the two Monte Carlo sets by using the SVD+PCA method implemented in the `mc2hessian` code (see section 2.3.2). Since the method provides a great compression of information, a faithful representation of the prior sets may be obtained from $N_{\rm eig} = 100$ eigenvectors. We can now study the χ^2 variations around the best PDF estimate for each eigenvector produced. The χ^2 variation for an eigenvector is simply computed as the difference between the χ^2 eq. (2.11) evaluated with the eigenvector prediction (with real data and t_0-prescription for the covariance matrix), and the same χ^2 evaluated with the central value of the Hessian set which, by definition, is equal to the central value of the prior Monte Carlo set.

The $\Delta\chi^2$ distribution are presented in fig. 3.5: on the left, a bar plot shows the χ^2 variations for each of the $N_{\rm eig} = 100$ eigenvectors, while on the right is shown the distribution of these values. The first evident observation for both sets is that a fair number of eigenvectors, almost 30, describe negative variations of the χ^2, and so we denote them as "negative" eigenvectors from now on. Moreover, the spread of the variations are quite far from the expected value $\Delta\chi^2 = 1$ for the 68% c.l. interval: in the Hessian NNPDF set $-6 \lesssim \Delta\chi^2 \lesssim 20$, while in the `n3fit` one $-14 \lesssim \Delta\chi^2 \lesssim 15$. Apart for the large negative variations of the first two eigenvectors of the `n3fit` set, the overall range of fluctuations

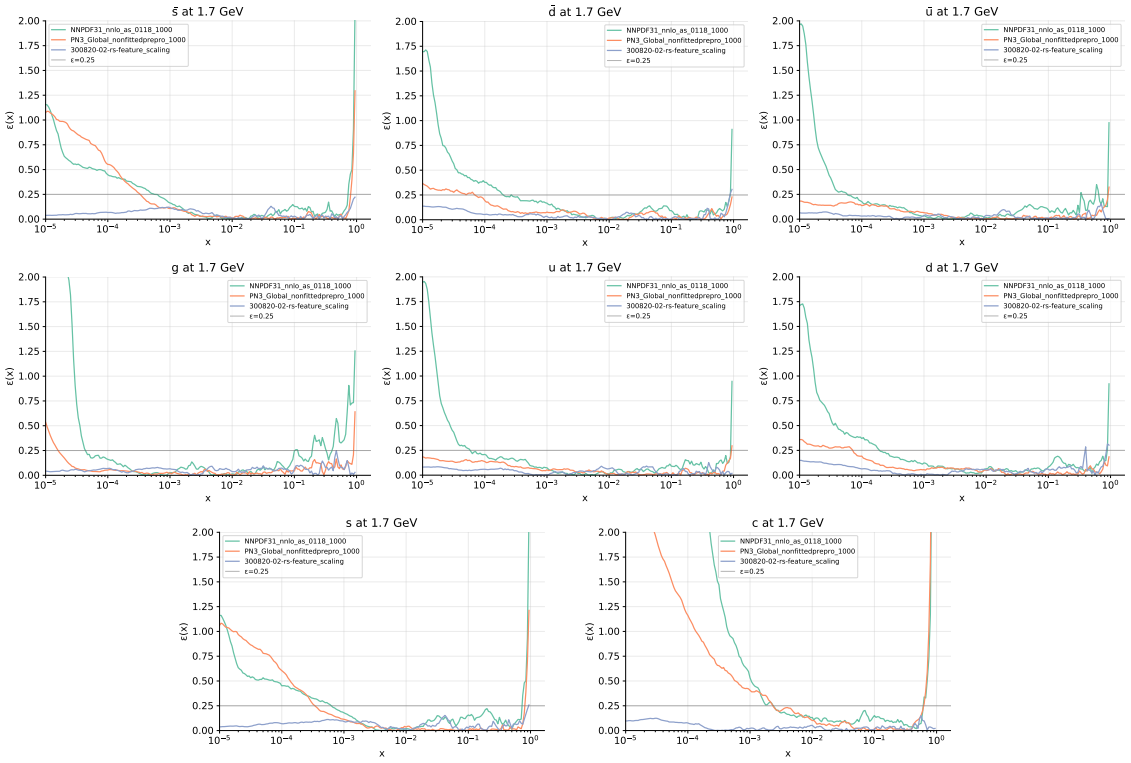

Figure 3.4: Gaussian error estimator $\epsilon_\alpha(x_i, Q)$ as a function of x at $Q = 1.7\,\text{GeV}$. The results for the `n3fit` set are presented with orange lines, while in teal those of NNPDF3.1. The blue lines represent the values for another `n3fit` set which will be considered in section 3.2.2. From left to right and top to bottom are shown the results for $\bar{s}, \bar{d}, \bar{u}, g, u, d, s, c$. The thin horizontal line represents the threshold value $\epsilon = 0.25$ chosen for the conversion of the `n3fit` set.

is similar, which may suggest that the two methodologies give similar results in the description of PDFs uncertainties. We may also note that the minimum of the χ^2 reached by NNPDF is smaller, $\chi^2_0 = 5045$, while for `n3fit` $\chi^2_0 = 5120$, even though we cannot know which one is closer to the real global minimum. Nevertheless, the presence of negative eigenvectors has to be interpreted as a symptom of inefficiency in the fitting methodologies, even though the results of fig. 3.5 cannot tell us where the inefficiency comes from. We shall now describe a possible strategy to address this problem.

3.2.2 Eigenvector decomposition and the feature scaling branch

By inspecting the $\Delta\chi^2$ distributions, we may consider the "eigenvector decomposition" of the Hessian sets, namely divide each of them into two disjointed subsets of positive and negative eigenvectors, relative to positive and negative variations of the χ^2 respectively. We can then compare these sets to study where the negative eigenvectors influence the

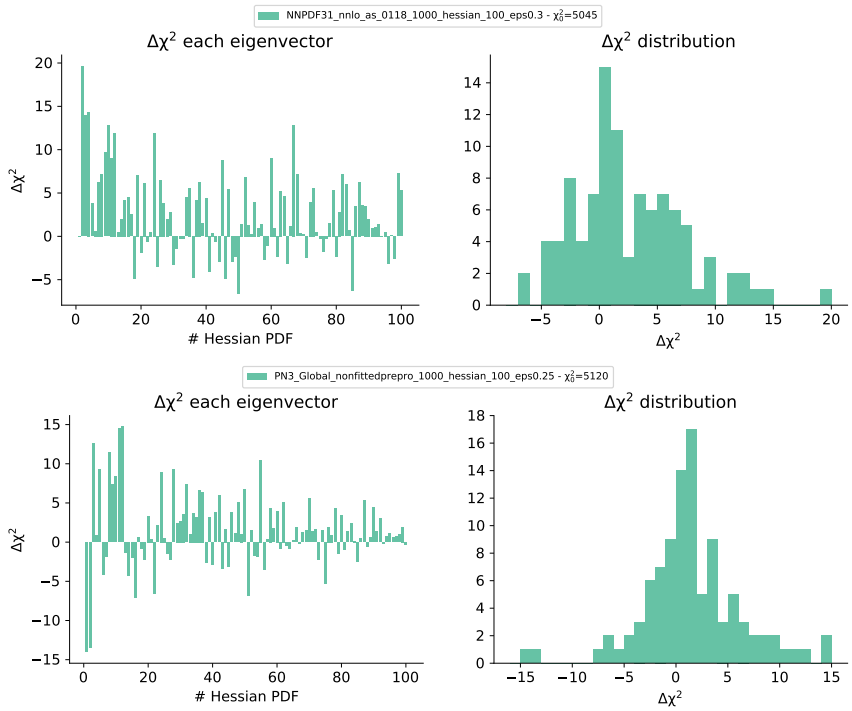

Figure 3.5: $\Delta\chi^2$ distribution for the Hessian sets NNPDF3.1 (top) and `n3fit` (bottom) with $N_{\text{eig}} = 100$ eigenvectors. The left plot shows the variation $\Delta\chi^2$ for each eigenvector, while the right plot shows their distribution with bins of unit length.

most the uncertainties of the PDFs.

The results for the NNPDF3.1 and `n3fit` sets for the lightest flavours and gluon are shown in figs. 3.6 and 3.7: the positive (orange) and negative (blue) eigenvector subsets are compared with the corresponding parent Hessian set (teal), with values normalized to the (same) central PDF. Overall, the uncertainties of the quark distributions are described by the positive eigenvectors, except in particular regions where the negative subset gives a similar contribution, for instance at large-x. The most intriguing results are found for the gluon distribution, where the uncertainties of the negative subset dominate over the positive ones at $x \lesssim 10^{-3}$, particularly in the `n3fit` set.

Since the largest uncertainties of the negative eigenvector subsets are found at small- and large-x, this analysis might be hinting to an inefficiency in the PDFs determination in regions of extrapolation. As explained in section 2.2.1, the PDFs parametrization in those regions is given by the preprocessing factor $x^{1-\alpha_i}(1-x)^{\beta_i}$, while the neural network has little to no effect due to lack of data points. It could be possible that this simple polynomial affects the PDFs functional form found by the NN towards incorrect values at the boundary of the regions of extrapolation. Thus, we are induced to think that the presence of negative eigenvectors is a consequence of the non optimal parametrization of the PDFs at small- and large-x.

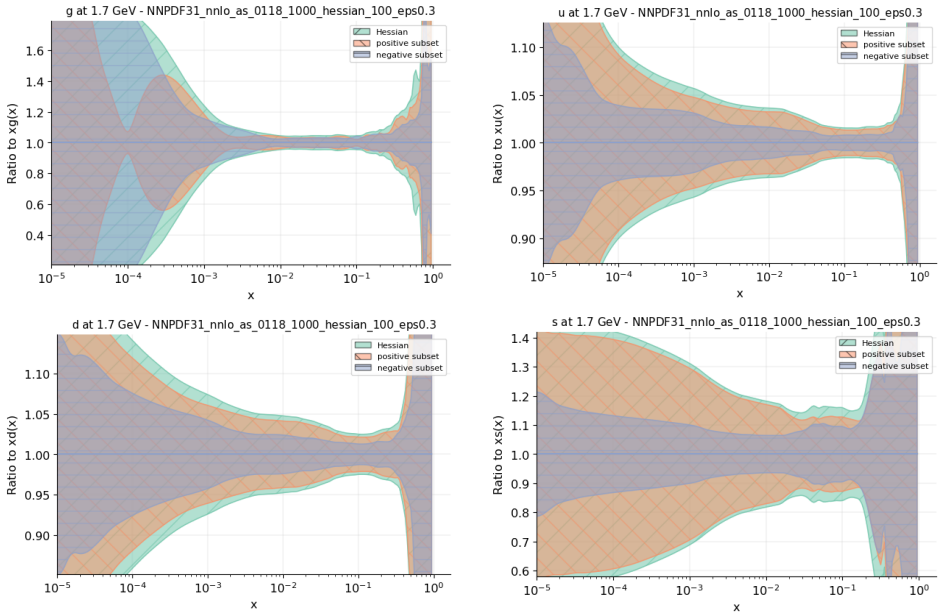

Figure 3.6: Comparison of positive (orange) and negative (blue) eigenvector subsets with the parent NNPDF3.1 Hessian set (teal) at the scale $Q = 1.7\,\text{GeV}$. Values are normalized to the central PDF, which is always the same. From left to right and top to bottom are shown the gluon, and then up, down, strange quarks.

Guided by these observations, we considered an experimental branch of `n3fit`, named `feature_scaling_test`, in which the following changes have been made:

1. the PDFs are parametrized by the neural network only,

$$f(x, Q_0^2) = \text{NN}(x) - \text{NN}(x = 1)\,, \tag{3.2}$$

   where the second term ensures that the PDFs vanish in $x = 1$, while at small-x the network output may saturate;

2. instead of $(x, \ln x)$, the input layer is just x. However, while in the usual `n3fit` $x \in [0, 1]$, here the input domain is properly smeared in the interval $x \in [-1, 1]$.

We have then run a 1000 replica fit[3] to produce the new Monte Carlo set `300820-02-rs-feature_scaling`. From the same analysis outlined in the previous section, we study the Gaussian error estimator eq. (3.1): remarkably, $\epsilon_\alpha(x_i, Q) \lesssim 0.25$ for all x_i, as can be seen in fig. 3.4 (blue line), and thus we have chosen a threshold value $\epsilon = 0.25$ which can accommodate in practice all grid points. We eventually carry out the Monte Carlo to Hessian conversion with $N_{\text{eig}} = 100$ eigenvectors at $Q = 1.7\,\text{GeV}$.

First, we should look at the $\Delta\chi^2$ distribution of this new set, shown in fig. 3.8, from which we can accomplish the eigenvector decomposition. If compared to the previous ones

[3]Same theory parameters, datasets, and training/validation split as in the previous sets.

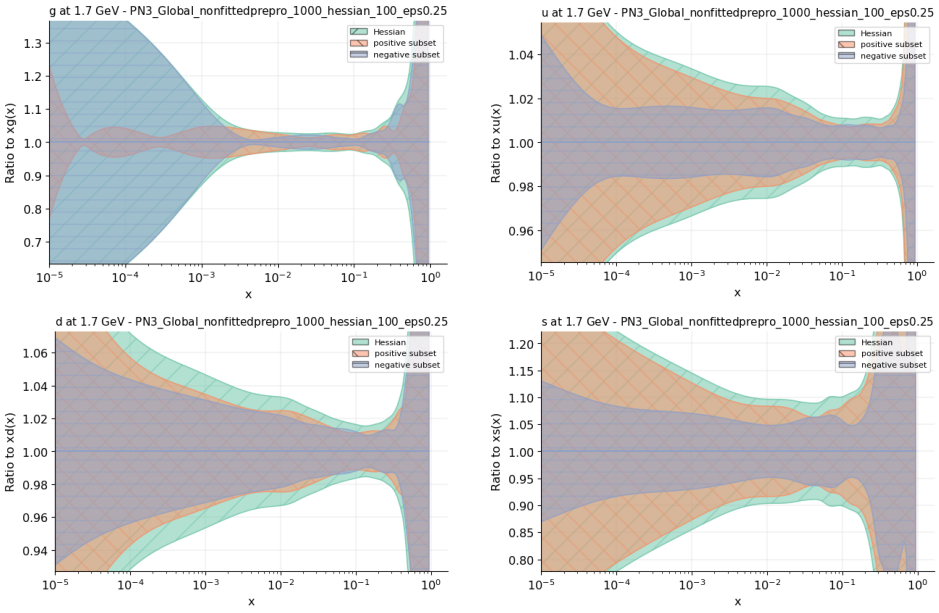

Figure 3.7: Same as fig. 3.6 but for the `n3fit` Hessian set.

in fig. 3.5, we observe an increase in the number of negative eigenvectors, now almost 40, whereas the χ^2 variations are doubled $-22 \lesssim \Delta\chi^2 \lesssim 42$. We may also notice the minimum found is the lowest of the three sets, $\chi^2_0 = 5015$, though we cannot know if we are getting closer to the real global minimum.

The results of the eigenvector decomposition are shown in fig. 3.9, where we provide the analogous plots of figs. 3.6 and 3.7. The most obvious difference from the previous `n3fit` set is in the gluon distribution, where now the uncertainties at small-x are described by the positive subset (note also the reduced range of the y-scale). However, the uncertainties of the quarks distributions are now dominated by the negative subset, particularly at small-x for the down and strange quarks. Based on the criterion of how the negative eigenvectors affect the uncertainties of a Hessian set, we may conclude the feature scaling improves the determination for the gluon distribution, but at the same time the predictions for the quarks get worse. Since the PDF parametrization adopted in the feature scaling is free of preprocessing factor, we may conclude the functional form of the gluon at $x \lesssim 10^{-3}$ could be more complicated than the simple $x^{1-\alpha}$, whereas for the quarks the usual term $x^{1-\alpha}(1-x)^{\beta}$ is appropriate for their description in the regions of extrapolation.

Given also the $\Delta\chi^2$ distribution of fig. 3.8, it is evident that the claimed improvement in the gluon predictions cannot correspond to a global improvement of the fitting methodology. Nevertheless, we shall consider this feature scaling set in our future discussion in chapter 4 to compare its predictions with both `n3fit` and NNPDF3.1.

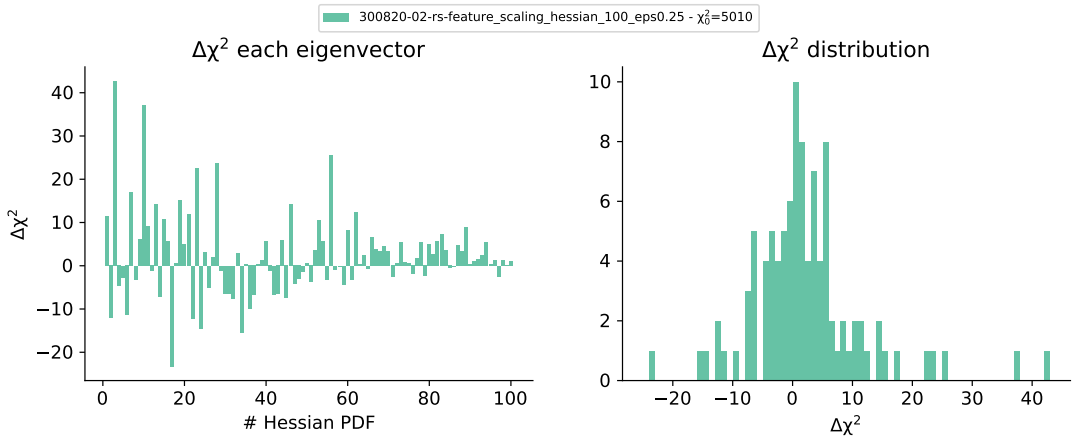

Figure 3.8: Same as fig. 3.5 but for the Hessian representation of the feature scaling Monte Carlo set.

3.3 Different prescription for the eigenvector decomposition

The χ^2 variation we studied so far was the difference between the χ^2 evaluated using the eigenvector eq. (2.36), and the central value of the Hessian set. As discussed in section 2.3.2, the starting point for the Monte Carlo to Hessian conversion is the construction of a *symmetric* covariance matrix in PDF space eq. (2.31), which consequently assumes a symmetric distribution in the resulting Hessian set. From this hypothesis we expect that if we consider the opposite eigenvector directions, namely flip the "+" into a "−" in eq. (2.36), we should find the same variations $\Delta\chi^2$.

We have therefore applied the same Hessian conversions to all previous Monte Carlo sets by considering instead the opposite directions to construct the eigenvectors. In fig. 3.10 we compare for each set the resulting $\Delta\chi^2$ distributions with the previous ones already shown in figs. 3.5 and 3.8. The values from the usual computation are labeled as "+ direction" while the new results as "− direction". We see that, contrary to our assumption, for the large part of the eigenvectors a positive variation is matched to a negative one in the opposite direction, and vice versa. Even though the χ^2 variations actually never match, the right plots show a rather unexpected symmetry in the distribution of these values.

The results presented here should therefore be included within the eigenvector decomposition analysis by applying a new prescription: we define a negative eigenvector when the corresponding χ^2 variation is negative in *at least* one direction. We may therefore reconsider the previous Hessian sets and see whether this new prescription can give further insights to possible sources of inefficiencies in the methodologies. Particularly, due to the asymmetry of the χ^2 in the eigenvector directions, we shall expect an increase of the uncertainties for the negative eigenvector subsets due to the increase in the number of negative eigenvectors, and conversely for the positive ones.

This observation is actually confirmed by the plots of fig. 3.11, which show the results of the decomposition for the three lightest quarks of each set, namely NNPDF3.1, `n3fit` and

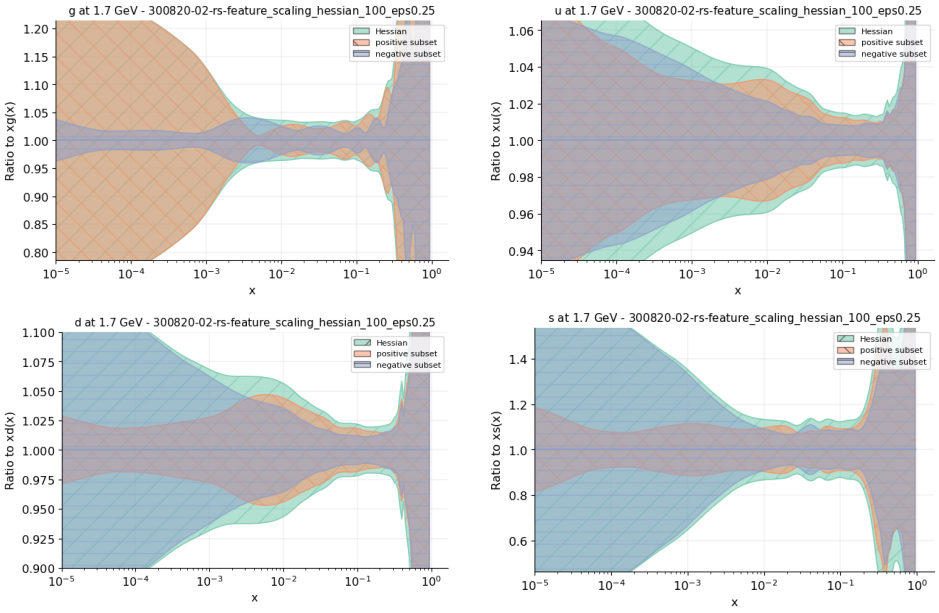

Figure 3.9: Same as fig. 3.6 but for the feature scaling Hessian set.

feature scaling (from top to bottom), at $Q = 1.7\,\text{GeV}$. The new prescription affects mostly the feature scaling set (third row), where the uncertainties are now completely described by the negative eigenvector subset. Furthermore, from the eigenvector decomposition for the gluon distribution, shown in fig. 3.12, it is clear that the uncertainties of the Hessian sets are now determined by the negative eigenvector subsets for all x, even in the feature scaling set.

Given this further informations, we are forced to discard our hypothesis that the pre-processing factor is a source of inefficiency in the determination of the PDFs. Nevertheless, the introduction of this kind of Hessian analysis in future fits should be a correct way to assess whether new Monte Carlo sets are actually giving the correct description of parton densities. Particularly, the improvement of the feature scaling fitting methodology could lead to an unbiased parametrization in the extrapolation regions at small- and large-x, where there is no theoretical constraint on the *exact* functional form of the PDFs.

The presence of negative eigenvectors in the Hessian representation of a Monte Carlo set might be the result of other effects, besides inefficiency, that we shall investigate in the following chapter to introduce an effective tolerance parameter for all three Monte Carlo sets considered in this thesis.

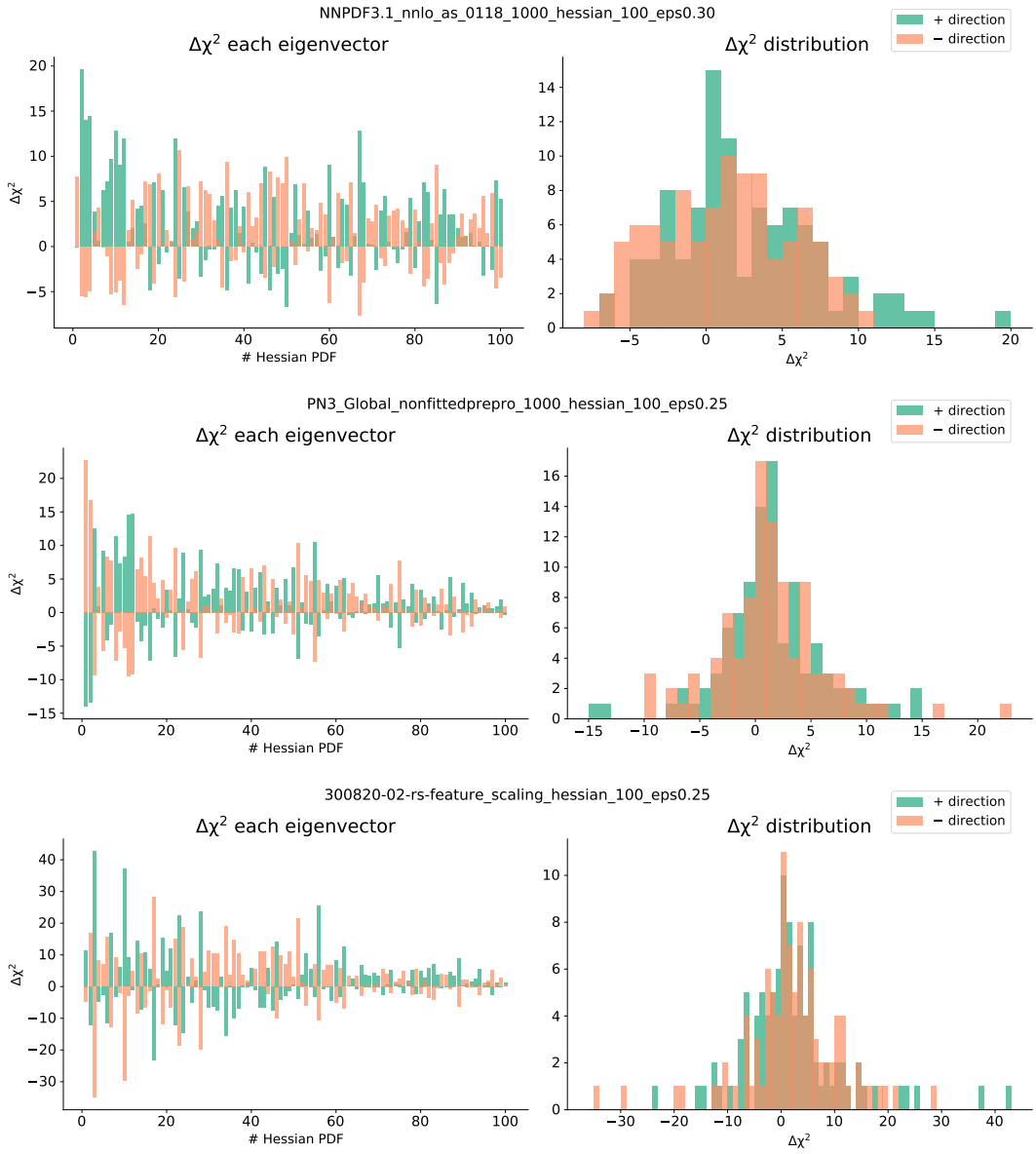

Figure 3.10: Comparison of $\Delta\chi^2$ distributions obtained from "+" (green) and "−" (orange) directions, as described in the text. As usual, on the left is shown the χ^2 variation for each eigenvector, while on the right the distribution of these values with bins of unit length. From top to bottom are shown the results for NNPDF3.1, `n3fit` and feature scaling.

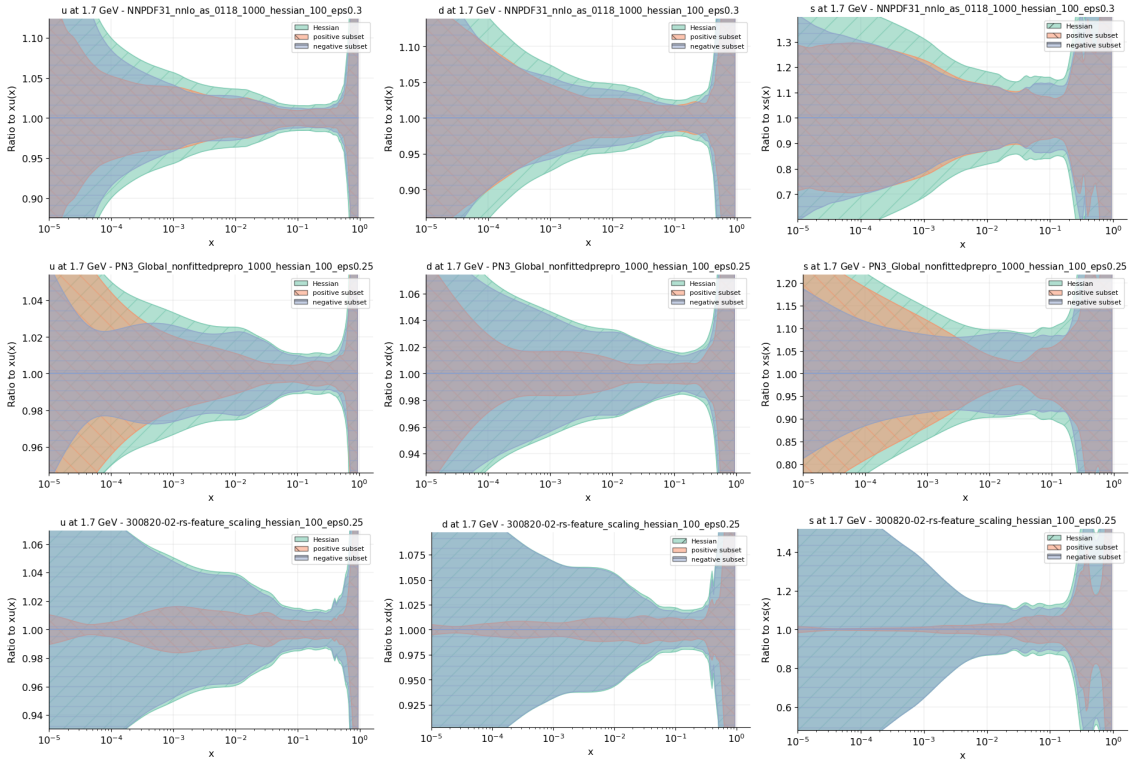

Figure 3.11: Comparison between the parent Hessian set (teal) with the positive (orange) and negative (blue) eigenvector subsets at $Q = 1.7\,\text{GeV}$ obtained from the *new prescription*. The first, second and third row correspond to the results from NNPDF3.1, `n3fit` and feature scaling respectively, while from left to right are shown the up, down, and strange quark distributions normalized to the central PDF.

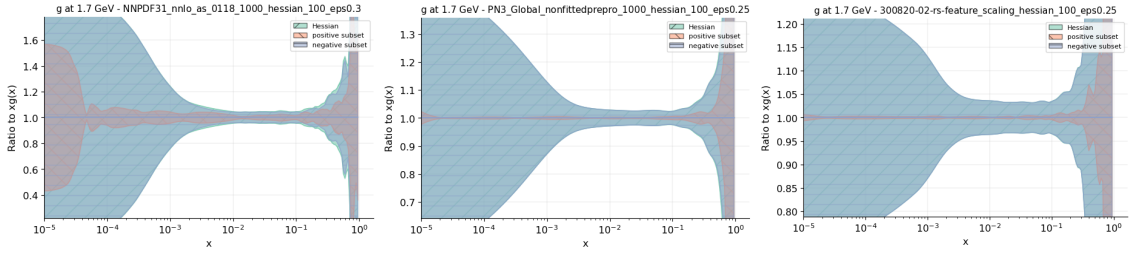

Figure 3.12: Same as fig. 3.11 but for the gluon distribution. From left to right are shown the results of NNPDF3.1, `n3fit` and feature scaling.

Chapter 4

Tolerance for Monte Carlo sets

As largely discussed, the presence of negative eigenvectors from the Hessian representation of a Monte Carlo set suggests the fitting procedure might have potential flaws. As a matter of fact, negative eigenvectors tell us that the central value of the prior Monte Carlo set does not correspond to the real global minimum of the χ^2, contrary to what is required by the Hessian method. There are however different aspects which determine the χ^2 shape:

- **Non-Gaussianity**: for a Hessian representation to be meaningful, the errors should follow a Gaussian probability distribution. If the grid points from the prior Monte Carlo set do not show Gaussian errors, they could spoil the conversion. As discussed in section 3.2.1, this problem is tackled by choosing a threshold value ϵ to discard points deemed to be non-Gaussian.

- **Finite size effects**: a Monte Carlo set provides a discrete representation of the underlying PDF probability distribution, and allows to obtain an estimate of PDF central values and uncertainties propagated from the experimental data. Thus, the "real" values for these quantities can be found only in the limit $N_{rep} \to \infty$.

- **Parabolic deviation**: the estimated uncertainties of a Monte Carlo set, when represented by the Hessian conversion, could be inadequate for a quadratic approximation of the χ^2 (see eq. (2.17)) around the supposed unique global minimum. In that case, larger variations than the expected $\Delta\chi^2 = 1$ or even negative variations might appear.

- **Inefficiency**: the χ^2 shape could be complicated by local or even degenerate minima. If the fitting methodology used to construct a Monte Carlo set is not optimal, it can lead to the wrong results. A potential error is therefore propagated from the prior set to the Hessian representation.

The combination of all these factors can thus produce the χ^2 distributions observed in the plots of fig. 3.5.

In the following we will describe and apply a procedure, firstly explained in the MSc. thesis of Ref. [95], which was used to extrapolate and quantify the contributions listed

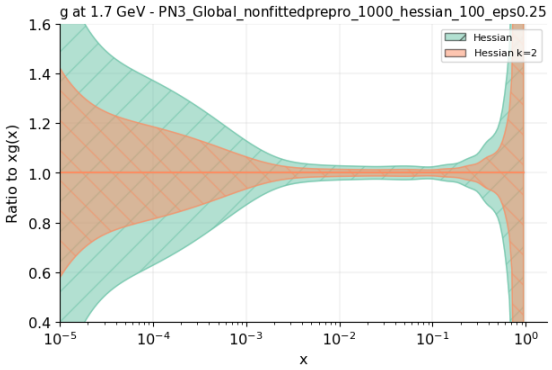

Figure 4.1: Comparison between the usual Hessian representation (teal) of the `n3fit` Monte Carlo set with the analogous representation obtained from a conversion with sigma-fraction $k = 2$ (orange) for the gluon distribution at $Q = 1.7\,\text{GeV}$. Values are normalized to the central PDF.

above only for the set `NNPDF31_nnlo_as_0118_1000`. We have then reimplemented the framework to repeat the same analysis for the NNPDF3.1 set and extend it to the `n3fit` and feature scaling sets. Thanks to the hyperopt framework of `n3fit`, this whole procedure could be incorporated in future fits as a further measure of goodness-of-fit, to exploit the different yet equivalent Hessian representation of a Monte Carlo set.

4.1 Monte Carlo to Hessian with sigma-fraction

In order to explore the χ^2 shape globally around the central value of a Monte Carlo set, we can build the Hessian representation using different sizes for the uncertainties. Specifically, this is achieved multiplying the covariance matrix in PDF space eq. (2.31) with a constant value $1/k^2$. The term k, named sigma-fraction, defines the new size of the fluctuations: in fact, this operation corresponds to rescale the sampling matrix X (see eq. (2.36)) by an amount $1/k$, and thus the same rescaling applies to the calculation of the uncertainties of the Hessian set eq. (2.37).

An illustrative example is shown in fig. 4.1, where the usual `n3fit` Hessian set of $N_{\text{eig}} = 100$ eigenvectors is compared to the same Hessian representation, but obtained from a conversion with $k = 2$, namely with uncertainties halved.

4.1.1 One-parameter model of χ^2

Since this method allows us to probe χ^2 variations around the best fit, it should also provide a quantitative information regarding the inefficiency of the fitting methodology. In particular, signals of inefficiency are alway found from negative eigenvectors, although they may be related to two different aspects:

1. $\boldsymbol{\nabla}\chi^2\big|_{\text{min}} = 0$. The minimization might have found the global minimum. A negative

variation in the χ^2 would then reveal that it was either a local minimum where the parabolic approximation for the uncertainties of the prior Monte Carlo set is not valid, or that it was a saddle point. In both cases, higher order terms must be taken into account.

2. $\boldsymbol{\nabla}\chi^2\big|_{\text{min}} \neq 0$. The minimization has failed, thus the errors of the prior Monte Carlo set do not even describe a region of uncertainty around a minimum.

From the Monte Carlo to Hessian procedure described in section 2.3.2, it is understood that the resulting Hessian set is comprised of N_{eig} *independent* eigenvectors. It is reasonable to consider all these PDFs as given by independent Hessian parameters following the same underlying Gaussian distribution. We may then introduce a one parameter model for the χ^2 which can be used to isolate the various contributions that determine its shape. Since one of the Hessian assumptions is a quadratic approximation of the χ^2 around its minimum, we may write its functional form as a Taylor expansion up to fourth order, to study the effects described above:

$$\begin{aligned}\Delta\chi^2(\theta) &= \chi^2(\theta) - \chi^2(\theta_0) = \\ &= a\,\frac{(\theta-\theta_0)}{\sigma} + b\,\frac{(\theta-\theta_0)^2}{\sigma^2} + c\,\frac{(\theta-\theta_0)^3}{\sigma^3} + d\,\frac{(\theta-\theta_0)^4}{\sigma^4}\,,\end{aligned} \tag{4.1}$$

where θ is the Hessian parameter distributed according to $\mathcal{N}(\theta_0, \sigma^2)$, while the coefficients a, b, c, d are proportional to the derivatives of χ^2 and therefore carry the information of its true shape. Particularly, since a is related to the first derivative it quantifies the inefficiency of the fitting procedure. The term b is analogous to a tolerance parameter, while c and d both describe parabolic deviation.

If a Monte Carlo representation of the parameter θ is sampled from $\mathcal{N}(\theta_0, \sigma^2)$, then, Hessian conversions for different values of sigma-fraction k may be carried out. It can be shown (see appendix B.1) that the finite size of the prior Monte Carlo set and the different sizes of the uncertainties can all be taken into account to rewrite eq. (4.1) as an expectation value for the χ^2 variation,

$$\left\langle \Delta\chi^2 \right\rangle = a_N\, x + b_N\, x^2 + c_N\, x^3 + d_N\, x^4\,, \quad x = 1/k\,. \tag{4.2}$$

Equation (4.2) is therefore a polynomial of fourth degree in $x = 1/k$, namely the inverse of the sigma-fraction, while the average coefficients a_N, b_N, c_N, d_N[1] carry the dependence on the number of replicas, N_{rep}, used in the Monte Carlo sampling. Their values and uncertainties are reported in appendix B.1, and more importantly the following relations hold:

$$a_N \sim a\,, \quad b_N \sim b\,, \quad c_N \sim c\,, \quad d_N \sim d\,, \quad \text{as } N_{\text{rep}} \to \infty\,, \tag{4.3}$$

which means the coefficients a, b, c, d in eq. (4.1) describe indeed the true χ^2 shape without any finite size effect, and we can extract them from their N_{rep}-dependent counterparts of eq. (4.2).

[1] For simplicity of notation, for these coefficients the subscript N denotes the number of replicas N_{rep}.

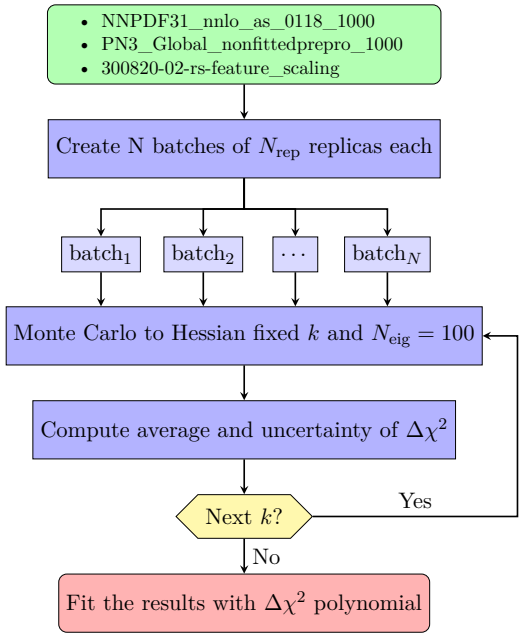

Figure 4.2: Flowchart describing the steps to extract the coefficients a_N, b_N, c_N, d_N in eq. (4.2). The whole procedure is repeated for all the group of batches considered as described in the text.

We can now turn the attention to our PDF Monte Carlo sets. We want to fit the $x = 1/k$ dependence of the expectation value $\langle \Delta\chi^2 \rangle$ to extract a_N, b_N, c_N, d_N and use the model predictions to extrapolate the Taylor coefficients a, b, c, d. The whole procedure applied is sketched in fig. 4.2. We start from the three Monte Carlo sets we have examined thus far. At first, we are interested in the coefficients a_N, b_N, c_N, d_N and to their dependence on the number of replicas. Since all the Monte Carlo sets we use are made of 1000 replicas, to increase the statistics we create the maximum number of batches given a fixed number of replicas we want to use. Specifically, we consider subsets comprised of $N_{\rm rep} = \{100, 125, 150, 175, 200, 250, 300, 350, 400, 500, 750, 1000\}$ replicas, thus the number of batches is respectively $N = \lfloor 1000/N_{\rm rep} \rfloor = \{10, 8, 6, 5, 5, 4, 3, 3, 2, 2, 1, 1\}$. All these twelve groups of batches are made up of subsets with the same number of replicas.

Then, we repeat the following steps for each group: we carry out the Monte Carlo to Hessian conversion of the subsets at fixed sigma-fraction k, with $N_{\rm eig} = 100$ eigenvectors and threshold ϵ unchanged (see sections 3.2.1 and 3.2.2). From these Hessian subsets we compute average and uncertainty of the usual χ^2 variation over all the eigenvectors, since we assumed they all follow the same underlying distribution. The result is therefore the expectation value $\langle \Delta\chi^2 \rangle$, with related uncertainty, at fixed k and $N_{\rm rep}$. We must then iterate the Hessian conversions over several values of k in order to finally fit the computed values of $\langle \Delta\chi^2 \rangle$ with a fourth degree polynomial in $x = 1/k$, as predicted by eq. (4.2). Specifically, we considered $k = \{0.125, 0.143, 0.166, 0.2, 0.25, 0.333, 0.4, 0.5, 0.66, 1, 2, 3, 4, 6, 7\}$ and corresponding negatives.

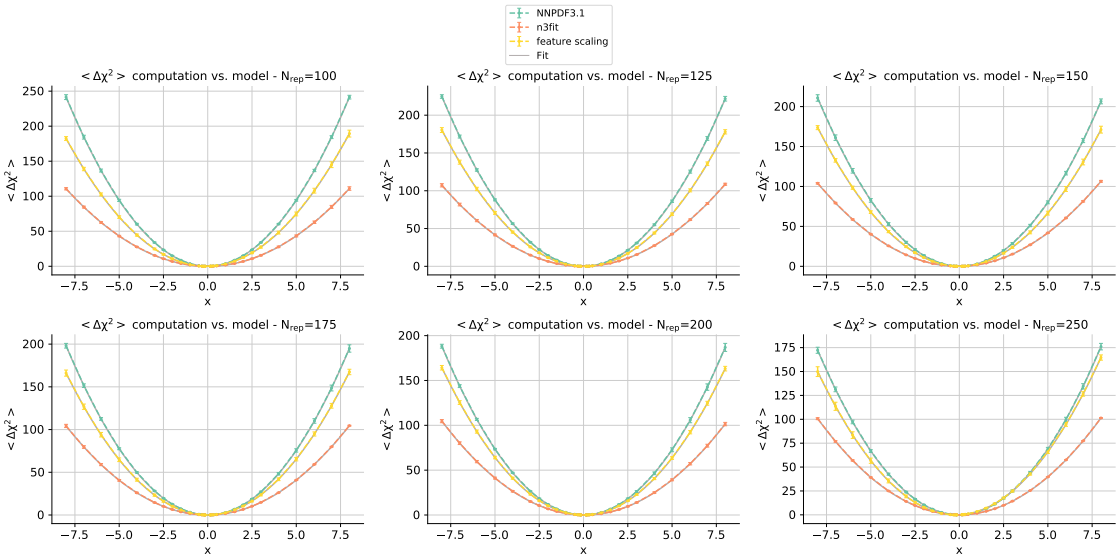

Figure 4.3: Comparison of the computed values of $\langle \Delta\chi^2 \rangle$ (error bars) as a function of $x = 1/k$ from the Monte Carlo sets NNPDF3.1 (teal), `n3fit` (orange), and feature scaling (yellow). From left to right and top to bottom are shown the results for $N_{\text{rep}} = \{100, 125, 150, 175, 200, 250\}$. The fit results to a fourth degree polynomial (gray line) always overlap with the coloured lines of the computed values.

Figure 4.3 shows the plots of $\langle \Delta\chi^2 \rangle$ as a function of $x = 1/k$ for the lowest number of replicas considered, $N_{\text{rep}} = \{100, 125, 150, 175\}$, from top to bottom and left to right[2]. The three paraboloids in each plot correspond to the results computed from the Monte Carlo sets NNPDF3.1 (teal), `n3fit` (orange) and feature scaling (yellow): the error bars mark central values and uncertainties of the χ^2 variation at fixed $x = 1/k$, or equivalently, at fixed sigma-fraction. Since the `n3fit` results are systematically smaller, we may conclude this is the Monte Carlo set that provides the most reliable predictions for the PDFs.

We then fit these values with a fourth degree polynomial in x as predicted by the model, eq. (4.2), to extract the N_{rep}-dependent coefficients. In practice, the computed values (coloured lines) perfectly match a fourth degree polynomial, as the solid gray line used to represent the fit always overlap with them. From each of these fits, we obtain the array of coefficients (a_N, b_N, c_N, d_N), relative to the N_{rep} value considered. Using the model predictions eqs. (B.8) to (B.11), from (a_N, b_N, c_N, d_N) we can calculate the corresponding Taylor coefficients (a, b, c, d). The final results are given as an average over the arrays of N_{rep}-*independent* coefficients, with uncertainties propagated according to the formulas being used. The values extrapolated with this procedure are listed in table 4.1, from which we may notice that $c, d \ll a \ll b \sim \mathcal{O}(1)$ in all three Monte Carlo set.

We may now use these values to compare the N_{rep}-dependent coefficients obtained from the previous analysis with the model predictions. In fig. 4.4 are shown the results

[2]See appendix B.2 for the complete results.

	Taylor coefficients					
	NNPDF3.1		`n3fit`		feature scaling	
a	−0.0936	± 0.0005	0.024	± 0.001	0.0462	± 0.0007
b	2.7331	± 0.0003	1.5749	± 0.0005	2.4487	± 0.0004
c	−0.000 10	± 0.000 03	−0.000 07	± 0.000 07	−0.000 05	± 0.000 03
d	0.000 184	± 0.000 007	0.000 08	± 0.000 01	0.000 12	± 0.000 08

Table 4.1: Taylor coefficients extrapolated with the model predictions for the Monte Carlo sets NNPDF3.1, `n3fit`, and feature scaling.

for NNPDF3.1 (left), `n3fit` (right) and feature scaling (bottom). The orange lines represent the fit results for a_N, b_N, c_N, d_N and the green bands the model predictions. From this comparison we observe that the fluctuations of the coefficients a_N and c_N are highly underestimated by the model, while for d_N the results are compatible. A separate discussion is needed for the dominant coefficient, b_N: although the model predicts $b_N = b$ (see eq. (B.9)), this coefficient exhibits a decreasing behaviour as the number of replicas increases. We should therefore interpret this trend as the result of finite size effects not considered by the model. Particularly, for the NNPDF3.1 and feature scaling sets we observe at $N_{\text{rep}} = 100$ the initial values $b_N \simeq 3.8$ and $b_N \simeq 2.8$, respectively, but similar final values $b_N \simeq 2$ at $N_{\text{rep}} = 1000$. This could indicate that in the limit $N_{\text{rep}} \to \infty$ the PDF predictions of these two sets should be subject to similar fluctuations. On the other hand, in the `n3fit` case the results are more stable and compatible with the model predictions, which suggests the minimization procedure of `n3fit` is more efficient, as it converges faster to the real value of b.

4.1.2 Dependence on the number of eigenvectors

We have then investigated the dependence of the fit to the N_{eig} number of eigenvectors exploited in the conversions. We considered five disjointed subsets of eigenvectors for the computation of the coefficients a_N, b_N, c_N, d_N. Specifically, instead of calculating the expectation value $\langle \Delta\chi^2 \rangle$ over the complete set of $N_{\text{eig}} = 100$ eigenvectors, we use the eigenvectors 1-20, 21-40, 41-60, 61-80, 81-100 and also the cumulative subsets (1-20), 1-40, 1-60, and 1-80. In fig. 4.5 are shown the results for NNPDF3.1 (left), `n3fit` (right), and feature scaling (bottom): the dashed lines represent the values obtained from the disjointed subsets of eigenvectors, the solid lines those from the cumulative subsets, whereas the solid light green lines in each plot correspond to the previous analysis with $N_{\text{eig}} = 100$.

Overall, a_N and c_N show similar fluctuations around zero in all three sets, particularly $-3 \lesssim a_N \lesssim 2$ and $-0.1 \lesssim c_N \lesssim 0.1$. Concerning the coefficient d_N, we may observe that it is essentially independent to the number of replicas, and by increasing the number of eigenvectors its value lowers towards zero for all Monte Carlo sets. Finally, the coefficient b_N shows a clear dependence on the *number of eigenvectors* in all sets, in addition to its dependence on the number of replicas (see fig. 4.4). We thus provide the same plots for this coefficient in fig. 4.6, but only for the values obtained from the cumulative subsets

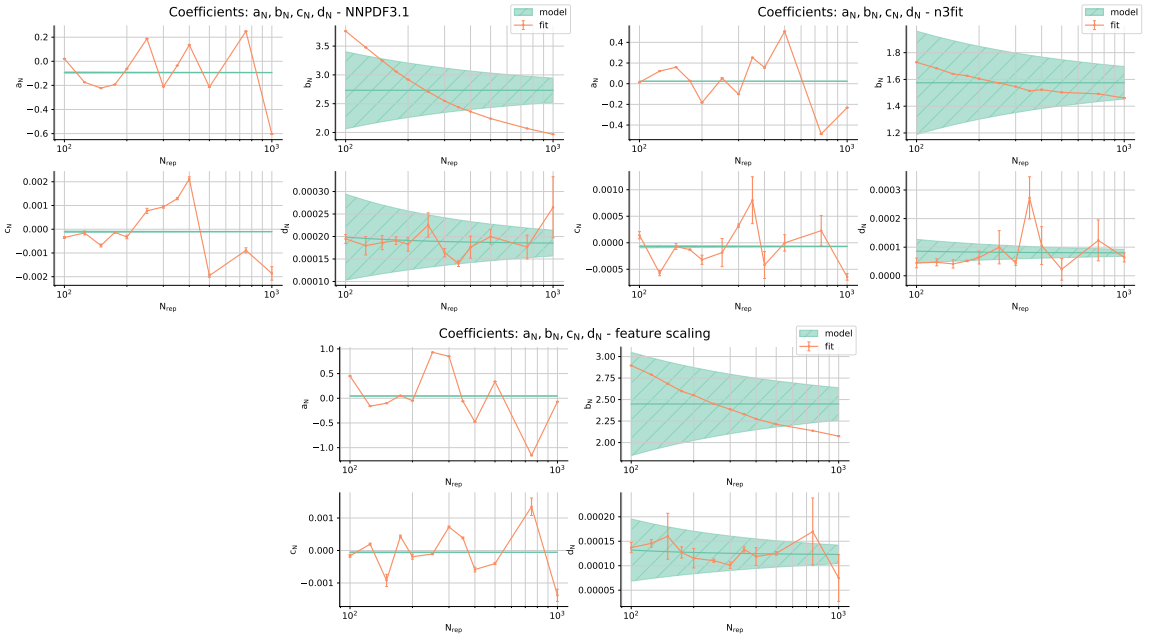

Figure 4.4: Comparison between the fit results (orange lines) and the model predictions (green bands) for the N_{rep}-dependent coefficients a_N, b_N, c_N, d_N from the Monte Carlo to Hessian conversions of NNPDF3.1 (left), `n3fit` (right), and feature scaling (bottom).

in order to highlight the N_{eig} dependence. In particular, for NNPDF3.1 and `n3fit` b_N always decreases as the *number of eigenvectors* increases at fixed N_{rep}, whereas in the feature scaling set we observe this behaviour only at $N_{\text{rep}} < 250$. However, it is clear that as we use more eigenvectors to represent the prior Monte Carlo sets, the coefficient b_N becomes less subject to finite size effects, and approaches the (constant) Taylor coefficient b as predicted by the model. Since b describes the real shape of the χ^2 without finite size effects, it should be possible to extrapolate its value in the limit $N_{\text{rep}} \to \infty$ from the functional forms describing the curves in fig. 4.6.

In conclusion, `n3fit` Monte Carlo sets comprised of $N_{\text{rep}} \geq 100$ replicas can be faithfully described using $N_{\text{eig}} = 100$ eigenvectors: since the *dominant* coefficient b_N is approximately constant, the χ^2 variations around the central value are on average equal, and thus the predictions for the PDFs will show similar fluctuations. Instead, for the NNPDF3.1 and feature scaling sets the finite size effects are a limiting factor, as reducing the fluctuations in the χ^2 (therefore the value of b_N) requires a very large number of replicas $\sim \mathcal{O}(1000)$. In the following section, we will exploit the coefficients extracted from this analysis to estimate an effective tolerance parameter T for these three Monte Carlo sets, providing a further measure of goodness-of-fit for them.

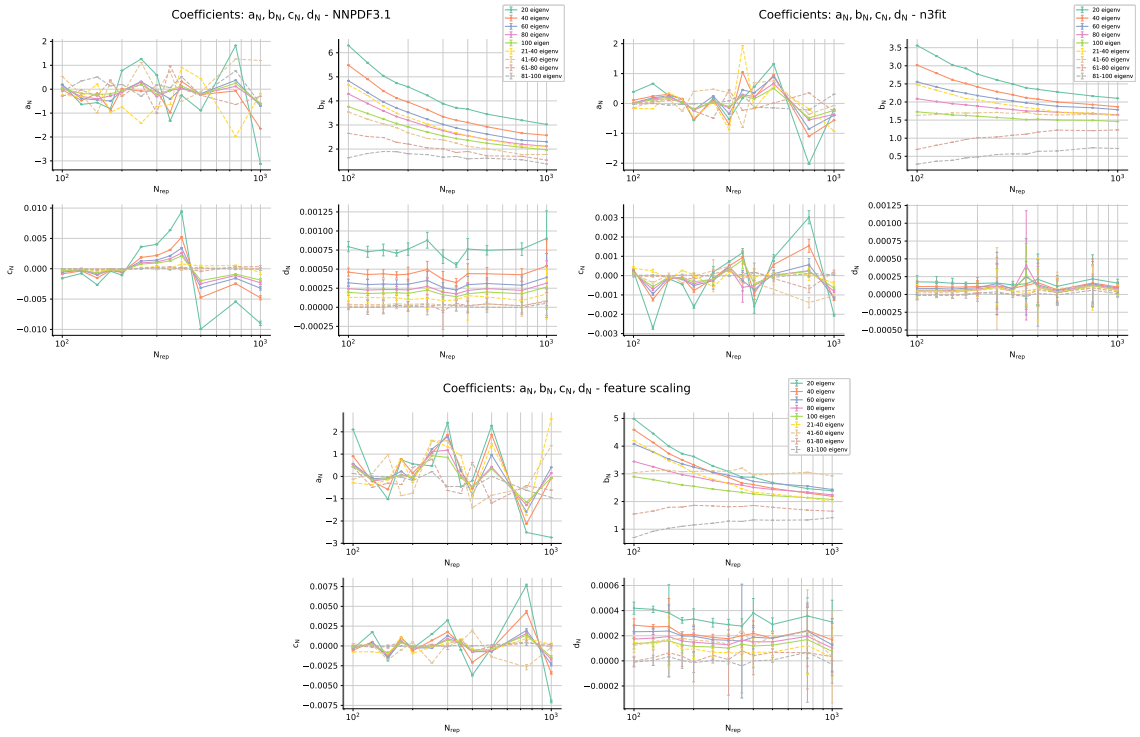

Figure 4.5: N_{rep}-dependent coefficients a_N, b_N, c_N, d_N obtained from disjointed (dashed lines) and cumulative (solid lines) subsets of eigenvectors from the Monte Carlo to Hessian conversions of NNPDF3.1 (left), `n3fit` (right), and feature scaling (bottom).

4.2 Tolerance parameter

In order to provide a quantitative evaluation for the predictions of the Monte Carlo sets studied in this thesis, we should eventually extrapolate from their Hessian representations an effective tolerance parameter $T = \sqrt{\Delta\chi^2}$. This is our final test to determine whether the `n3fit` methodology improves the determination of PDFs, as a smaller value for the tolerance implies more precise predictions (see section 2.2.3).

As we observed in figs. 4.5 and 4.6, the coefficients a_N, b_N, c_N, d_N depend on the number of eigenvectors used to represent the prior Monte Carlo sets. Particularly, when we increase N_{eig}, a_N and c_N become less subject to wide fluctuations around zero, while b_N and d_N decrease towards some limit value. However, as the model predictions are not reliable (see fig. 4.4) we cannot use eqs. (B.8) to (B.11) to extract the Taylor coefficients a, b, c, d. We then need a different procedure for the extrapolation of a tolerance.

Since the results shown in fig. 4.3 are in perfect agreement with the quartic polynomial eq. (4.2), we must conclude that this is indeed the real functional form of the variation of the χ^2, and therefore we can trust the values of the N_{rep}-dependent coefficients we extracted. Moreover, all these results have been obtained from the computation with $N_{\text{eig}} = 100$ eigenvectors, and therefore we shall use them to compute the tolerance parameter for the

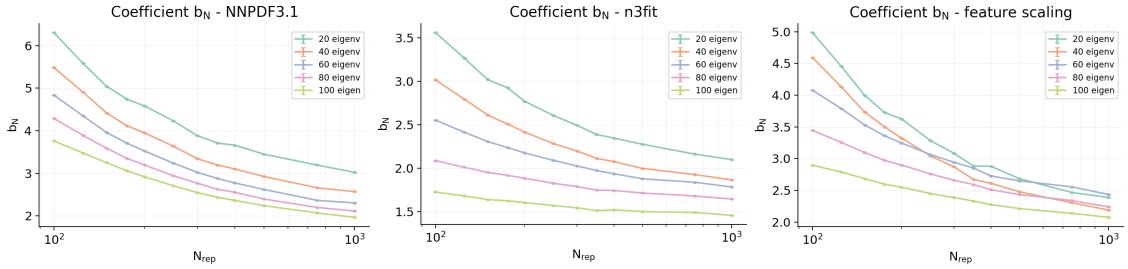

Figure 4.6: Coefficient b_N obtained from the cumulative subsets of eigenvectors, namely with $N_{\text{eig}} = \{20, 40, 60, 80, 100\}$. From left to right are shown the results from the Monte Carlo to Hessian conversions of NNPDF3.1, `n3fit` and feature scaling.

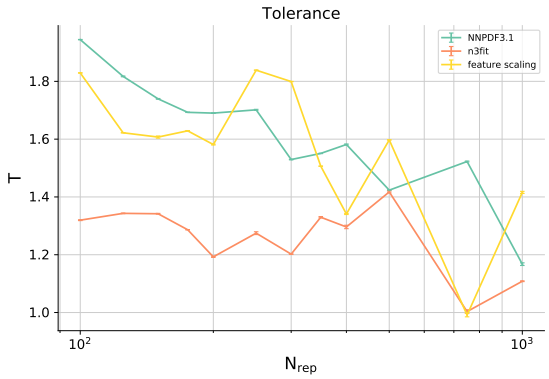

Figure 4.7: Comparison between the tolerance parameter $T = \sqrt{\langle \Delta\chi^2 \rangle}$ computed at different values of N_{rep} replicas from the Monte Carlo sets NNPDF3.1 (teal), `n3fit` (orange), and feature scaling (yellow). The results are computed with eq. (4.4) using the coefficients a_N, b_N, c_N, d_N obtained from Hessian conversions with $N_{\text{eig}} = 100$ eigenvectors.

Monte Carlo sets.

Thus, we apply the following "model independent" approach: we evaluate eq. (4.2) in $x = 1/k = 1$, in order to obtain

$$\langle \Delta\chi^2 \rangle = a_N + b_N + c_N + d_N \,, \tag{4.4}$$

which corresponds to the χ^2 variation for the Hessian representation of the prior Monte Carlo set without any rescaling in the uncertainties, since $k = 1$. We can then use eq. (4.4) to estimate an effective tolerance parameter $T = \sqrt{\langle \Delta\chi^2 \rangle}$, for each N_{rep} number of replicas we considered. In fig. 4.7 we show the resulting values of T as a function of N_{rep} for the three sets NNPDF3.1 (teal), `n3fit` (orange) and feature scaling (yellow).

As could be expected, the dominant coefficient b_N determines the dependence of the tolerances to the number of replicas. In fact, for the NNPDF3.1 set T decreases as the number of replicas increases, from $T \simeq 1.9$ at $N_{\text{rep}} = 100$ to $T \simeq 1.2$ at $N_{\text{rep}} = 1000$. For the feature scaling set, at first the tolerance fluctuates around $T \simeq 1.7$ and then decreases

similarly to NNPDF3.1 as soon as the number of replicas exceeds $N_{\text{rep}} = 300$, whereas for the `n3fit` set T is more stable in the range $1 \lesssim T \lesssim 1.4$. From the comparison of the tolerance values at fixed N_{rep} we can conclude that with the same number of replicas the predictions of the `n3fit` methodology are always more accurate.

Finally, we may give an estimate for a global tolerance for the three Monte Carlo sets by computing mean and standard deviation over all the values at different N_{rep}, which should be interpreted as an average of goodness-of-fit for the three methodologies. We thus find the following results:

$$\text{NNPDF3.1:} \quad T = 1.6 \pm 0.2\,, \tag{4.5}$$

$$\texttt{n3fit:} \quad T = 1.3 \pm 0.1\,, \tag{4.6}$$

$$\text{feature scaling:} \quad T = 1.6 \pm 0.2\,. \tag{4.7}$$

Overall, the values for the tolerance parameters are close to what expected from the parameter fitting criterion, $T = \sqrt{\Delta\chi^2} = 1$. Thus, the uncertainties in the Monte Carlo sets considered in this thesis[3] do not impose a large inflation of the tolerance parameter as instead is required in direct Hessian analyses (see section 2.2.3). The values in eqs. (4.5) to (4.7) may then be related to small incompatibilities of datasets or missing higher order uncertainties in theoretical predictions. However, these contributions should be similar for all three Monte Carlo sets because we considered equivalent sets (same theory parameters, datasets and training/validation split) and so we expect that differences in the tolerances are mainly due to the different strategies adopted in the fitting procedures.

We can finally conclude that the predictions of the `n3fit` methodology are indeed more accurate, while the NNPDF and feature scaling methodologies are expected to perform equally well on average. This suggests that further improvements in the feature scaling procedure should lead to results similar to `n3fit`, but at the same time provide an unbiased description of the PDFs functional forms without the need for a preprocessing factor (see section 3.2.2).

[3] `NNPDF31_nnlo_as_0118_1000`, `PN3_Global_nonfittedprepro_1000`, `300820-02-rs-feature_scaling`.

Chapter 5

Conclusions and outlook

In this thesis we examined the accuracy of predictions between the current state-of-the-art methodologies for PDFs determination in use within the NNPDF collaboration. We compared the latest version NNPDF3.1 with the new code, `n3fit`, developed in the last two years by the N3PDF project.

We started from two equivalent Monte Carlo sets (see section 3.2):

- `NNPDF31_nnlo_as_0118_1000`;
- `PN3_Global_nonfittedprepro_1000`.

The first is the NNPDF3.1 set of 1000 replicas published in the latest release of the NNPDF collaboration, while the second was obtained from a 1000 replica fit using `n3fit`, with a hyperoptimized configuration. Our analysis has been carried out by converting these sets into the equivalent Hessian representations, with the `mc2hessian` code. From the covariance matrix in the PDF space of replicas, the Hessian parton distributions are extracted by singular value decomposition as the dominant eigenvectors of this matrix, to provide a basis in the vector space spanned by the replicas (see section 2.3.2). While in the pure Hessian method the PDFs correspond to positive variations from the minimum of the χ^2, we found that both sets present a fair number of PDFs relative to negative variations. Furthermore, we observe much larger fluctuations than the textbook value $\Delta\chi^2 = 1$ for the 68% c.l. interval (see section 3.2.1).

Since the negative variations in the χ^2 should be interpreted as inefficiency in the fitting methodology, we introduced a strategy to isolate the negative contributions and find regions where the determination of the PDFs is not optimal. The largest tensions were found in the extrapolation regions, namely small- and large-x (see section 3.2.2). Thus, we searched for potential improvements in the `n3fit` fitting methodology using the experimental branch `feature_scaling_test`. We created a third Monte Carlo set, with a 1000 replica fit, `300820-02-rs-feature_scaling`, and repeated the same analysis. We observed an improvement in the determination of the gluon PDF at small-x, but a deterioration for the quark distributions. Furthermore, the variations of the χ^2 for this new set showed larger fluctuations. We extended the prescription to examine the contributions

of the negative variations by including in the analysis the χ^2 variations in the opposite directions defined by the Hessian eigenvectors (see section 3.3). We were forced to discard our previous considerations, and conclude that the PDFs relative to negative variations have a significant role to determine the uncertainties of the Monte Carlo sets, especially in regions of extrapolation.

We continued our investigation with the introduction of a simple one-parameter model of the χ^2, for which we assume a quartic expansion near the minimum (see section 4.1.1). We were able to extrapolate an effective tolerance for the three Monte Carlo sets, which can be interpreted as an average goodness-of-fit of the three methodologies under examination. We found that the value related to the `n3fit` set is the lowest, $T = 1.3$, and we concluded that the new framework improves the parton distribution functions determination. Concerning NNPDF3.1 and feature scaling, we obtained the same value $T = 1.6$, therefore we expect that further improvements in the feature scaling methodology could lead to similar results to `n3fit`.

Finally, we suggest the strategies presented in this work could be implemented in the `n3fit` code to improve the determination of the PDFs in future fits. Specifically, the effective tolerance of a Monte Carlo set might be used to introduce the loss function

$$\chi^2_{\text{loss}} = T - 1\,, \tag{5.1}$$

which should be minimized during the hyperoptimization scan. Moreover, the analysis of the negative variations of the χ^2 can always provide further tests about the reliability of the predictions of a fitting procedure.

Appendix A

Monte Carlo sets

In this appendix we present briefly the resulting parton distributions from the Monte Carlo sets considered in this thesis (see section 3.2):

- `NNPDF31_nnlo_as_0118_1000`;
- `PN3_Global_nonfittedprepro_1000`;
- `300820-02-rs-feature_scaling`,

and we also provide a comparison with their Hessian representations. The three sets are "equivalent", in the sense that theory parameters, datasets and training/validation split are always the same. Particularly, they are all obtained from NNLO calculations with $\alpha_S(M_Z^2) = 0.118$. In this way, we are able to compare the outcome of the different strategies adopted for the determination of the PDFs.

In fig. A.1 is shown the comparison between the PDFs from NNPDF3.1 (teal), `n3fit` (orange), feature scaling (blue), at the scale $Q = 1.7\,\text{GeV}$. Overall, the distributions are compatible within the uncertainty bands: the most evident differences are found for the gluon, for which `n3fit` and feature scaling do not predict a steep rise at small-x as NNPDF3.1. We may also notice the strange and anti-strange distributions of the feature scaling set which are almost flat at $x \lesssim 10^{-3}$. This could be due to the difficulty of the neural network alone to extrapolate in a region where the experimental constraints on these PDFs are poorer respect to the others.

In fig. A.2 we show again the same distributions but in ratio to the central PDFs of the NNPDF3.1 Monte Carlo set. From these plots we can also appreciate the relative difference in predictions at large-x of these methodologies.

Finally, we present in figs. A.3 to A.5 the comparison between the Monte Carlo PDFs with their Hessian representations. The figures show respectively the results for NNPDF3.1, `n3fit` and feature scaling, with values in ratio to the central PDF in order to highlight potential differences in the uncertainties. We may observe that the `n3fit` and feature scaling PDFs are always well reproduced by the Hessian conversions, figs. A.4 and A.5, while the discrepancies between the uncertainties at small-x and large-x for NNPDF3.1, fig. A.3, are due to the non-Gaussian grid points that had to be discarded to obtain a faithful Hessian representation (see section 3.2.1).

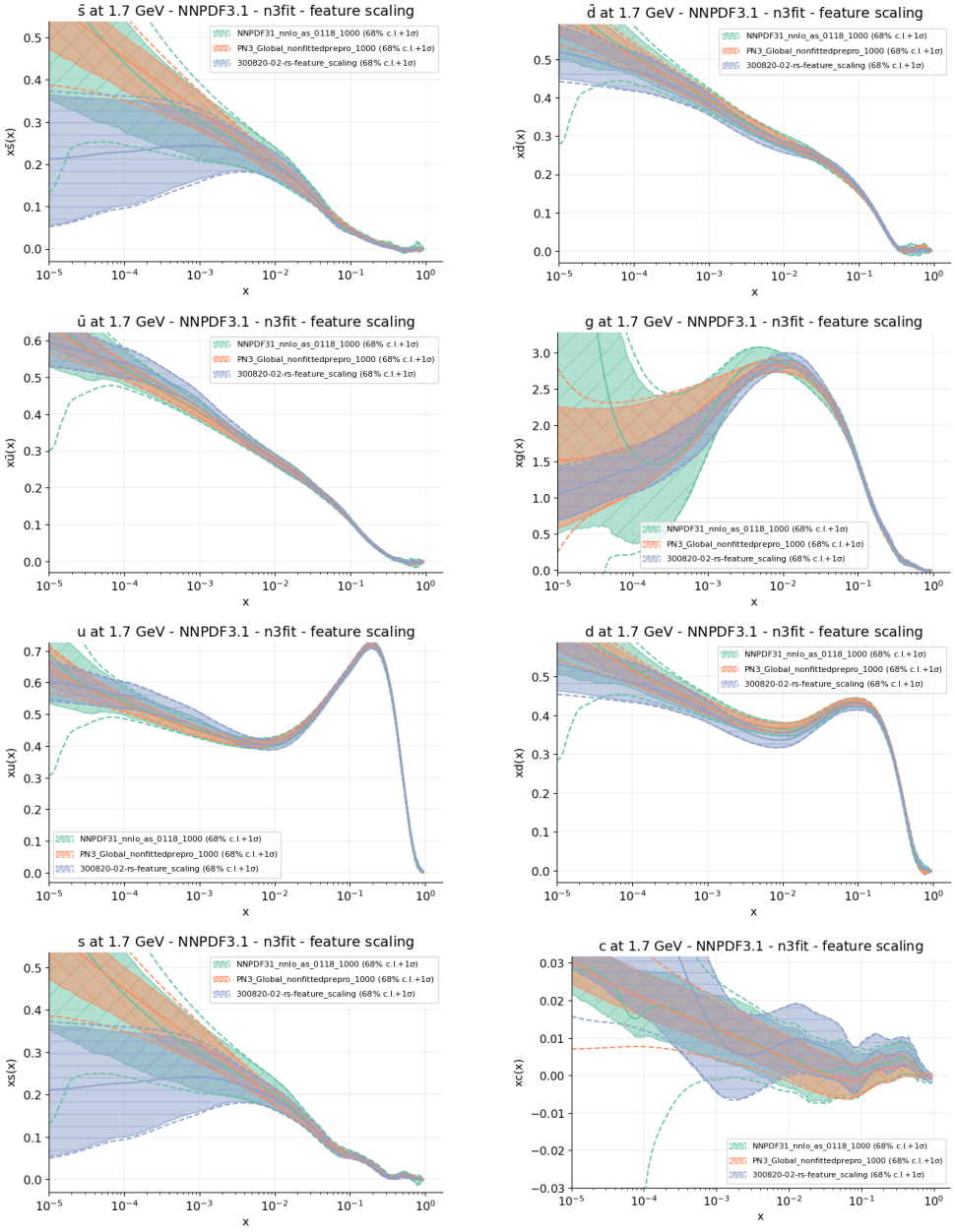

Figure A.1: Comparison between the PDFs from the Monte Carlo sets we considered in this thesis: NNPDF3.1 (teal), `n3fit` (orange), feature scaling (blue). The dashed lines represent the one-sigma uncertainty, while the coloured bands the 68% c.l. intervals. From left to right and top to bottom are shown the $\bar{s}, \bar{d}, \bar{u}, g, u, d, s, c$ distributions at $Q = 1.7\,\text{GeV}$.

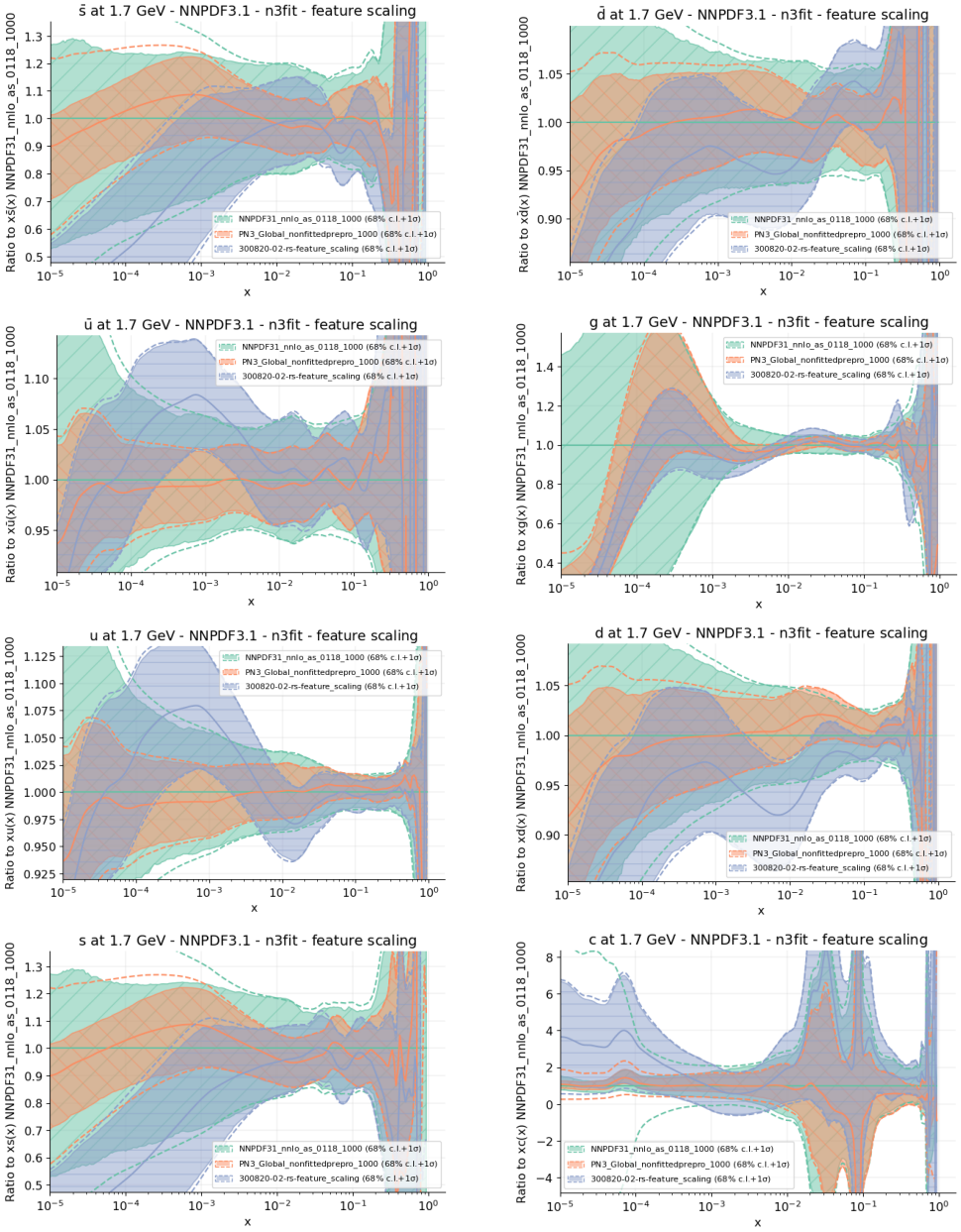

Figure A.2: Same as fig. A.1 but with values in ratio to the central PDFs of the NNPDF3.1 set.

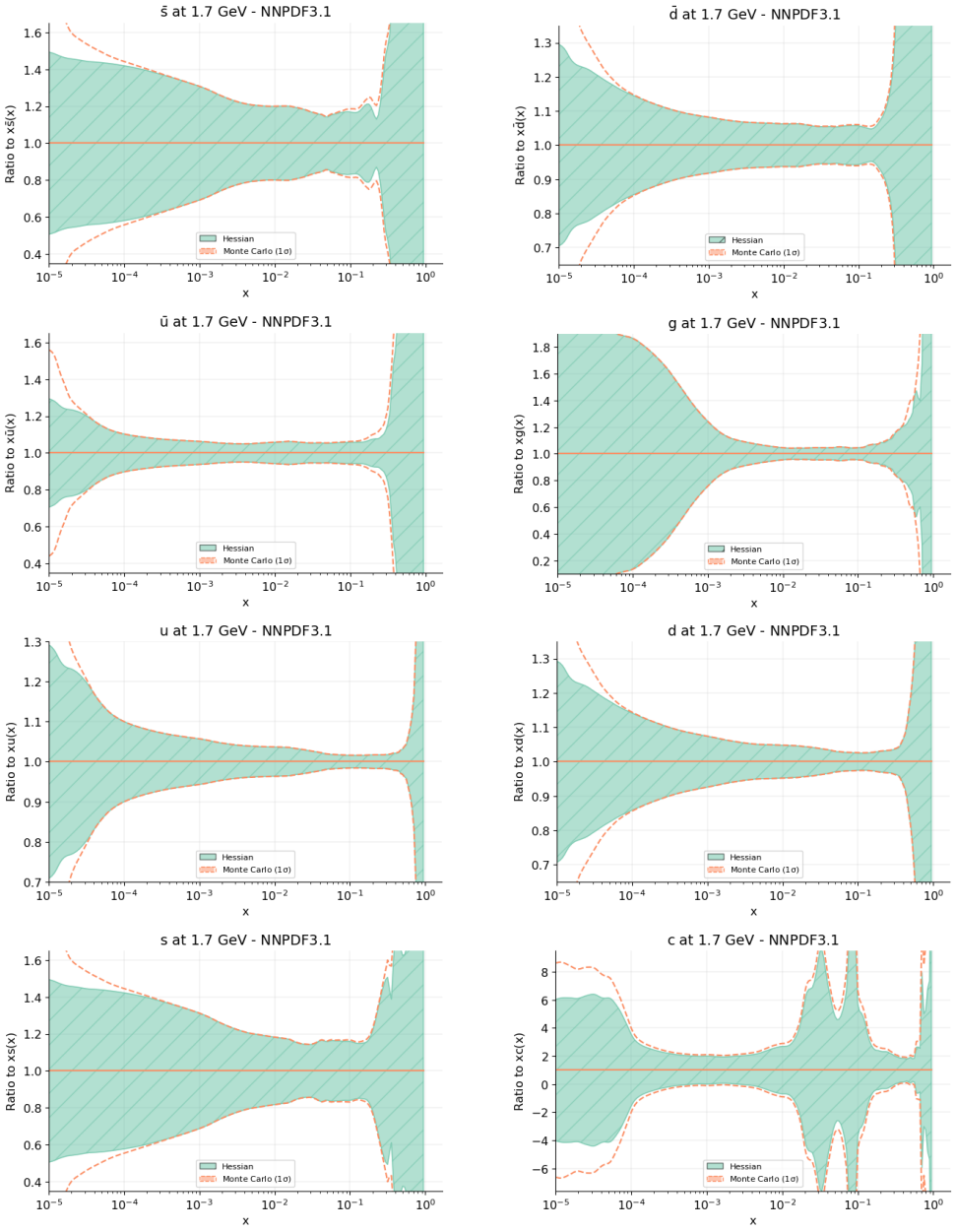

Figure A.3: Comparison between the PDFs from the NNPDF3.1 Monte Carlo set with the corresponding Hessian representation at $Q = 1.7\,\text{GeV}$ and $N_{\text{eig}} = 100$. Values are normalized to the (same) central PDF. The dashed orange lines correspond to the one-sigma uncertainty of the Monte Carlo set, while the teal bands to the uncertainties of the Hessian set. From left to right and top to bottom are shown the $\bar{s}, \bar{d}, \bar{u}, g, u, d, s, c$ distributions.

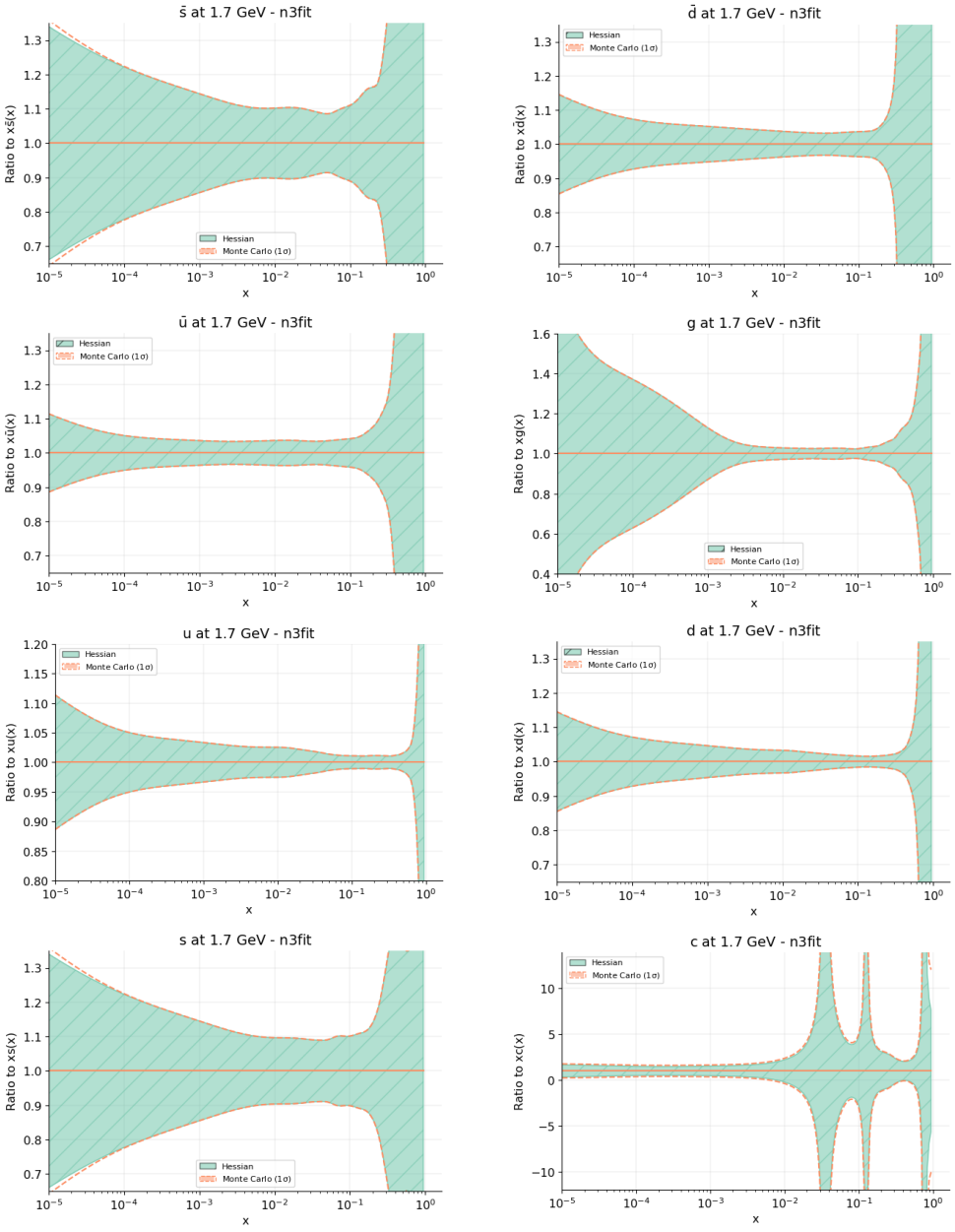

Figure A.4: Same as fig. A.3 but for the `n3fit` set.

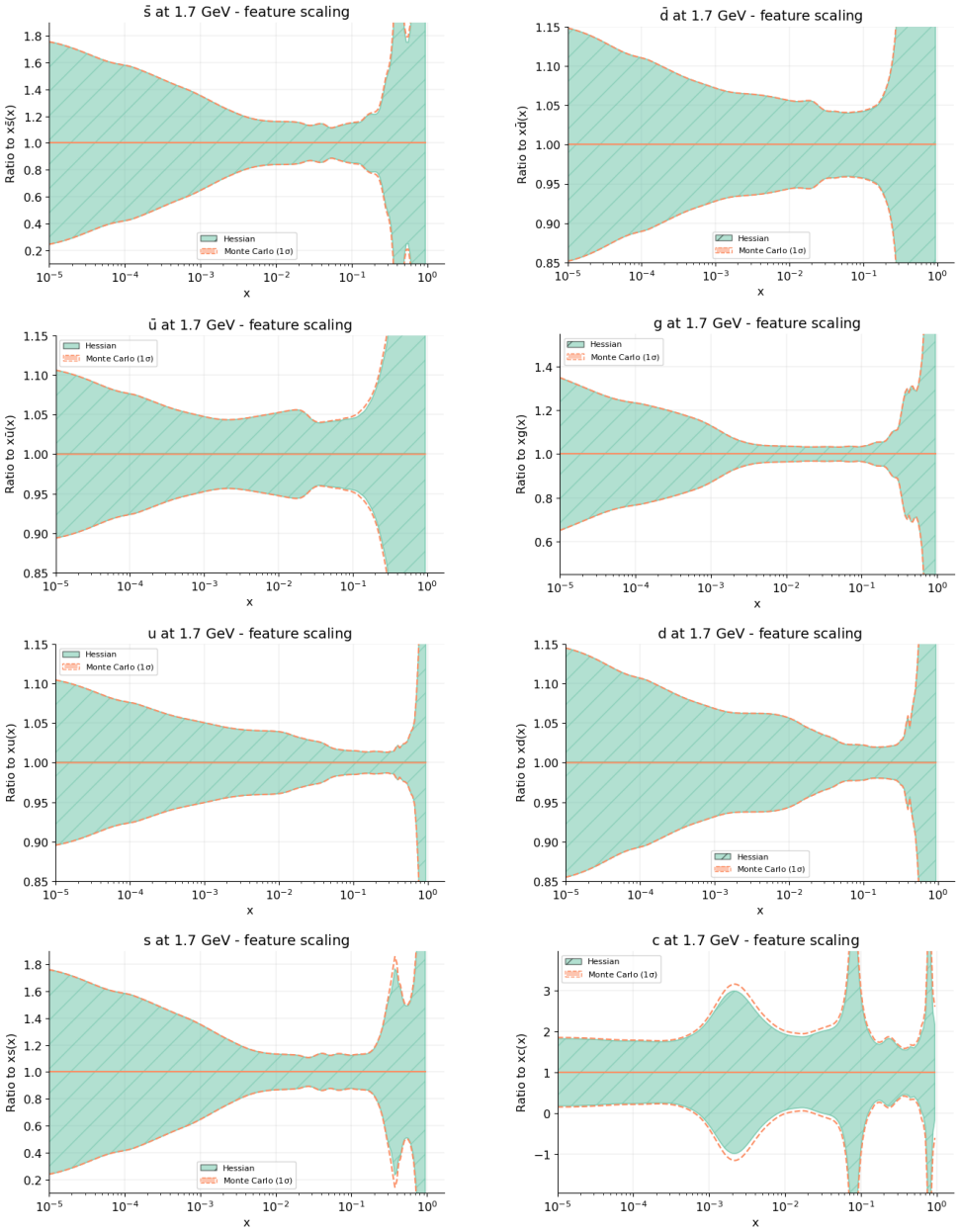

Figure A.5: Same as fig. A.3 but for the feature scaling set.

Appendix B

One-parameter model of χ^2

Here we discuss in more detail the calculation of the model predictions for the expectation value $\langle \Delta\chi^2 \rangle$, which is used to extract the tolerance parameter for the Monte Carlo sets considered in this thesis. We also show the fit results for all the number of replicas N_{rep} considered.

B.1 Model coefficients

As explained in section 4.1.1, the starting point in the Taylor expansion eq. (4.1) near the minimum of the χ^2:

$$\Delta\chi^2(\theta) = a\,\frac{(\theta-\theta_0)}{\sigma} + b\,\frac{(\theta-\theta_0)^2}{\sigma^2} + c\,\frac{(\theta-\theta_0)^3}{\sigma^3} + d\,\frac{(\theta-\theta_0)^4}{\sigma^4}\,, \tag{B.1}$$

where θ is a Hessian parameter distributed according to $\mathcal{N}(\theta_0, \sigma^2)$. By drawing a Monte Carlo sample for the parameter θ from this distribution, we may obtain the set of N_{rep} replicas[1] $\{\theta^{(k)}\}$ which, due to the finite size of the set, have central value μ and variance s^2 instead of the true θ_0 and σ. By applying a Hessian conversion with sigma-fraction k, the resulting Hessian set of θ's is distributed according to a Gaussian centered in μ and with variance s^2/k^2. The sample mean μ is a random variable, and specifically

$$\mu \sim \mathcal{N}\left(\theta_0, \frac{1}{N}\frac{s^2}{\sigma^2}\right) \tag{B.2}$$

while the sample variance s^2 can be expressed as

$$s^2 = \frac{\sigma^2}{k^2}\frac{x}{N-1}\,, \tag{B.3}$$

where x is distributed according to the χ^2 probability density function with $m = N-1$ degrees of freedom,

$$f(x;m) = \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)} x^{m/2-1}\exp^{-x/2}\,. \tag{B.4}$$

[1] For compact notation from now on we denote the number of replicas N_{rep} simply as N.

Since the Hessian conversion assumes the quadratic behaviour of the χ^2 near the minimum, we find that

$$\Delta\chi^2_{\rm samp}(\theta) = \frac{(\theta-\mu)^2}{s^2}\,. \tag{B.5}$$

The 68% c.l. interval for θ is then given by the parameter fitting criterion $\Delta\chi^2_{\rm samp} = 1$, which corresponds to the values $\theta = \mu \pm s$. This shift is analogous to the displacement from the best fit parameters $\vec{\boldsymbol{a}}_0$ along the eigenvector directions of the Hessian matrix.

However if the real χ^2 is given by eq. (B.1), the values $\theta = \mu \pm s$ induce instead a different increase, given by $\Delta\chi^2 = \chi^2(\mu \pm s) - \chi^2(\mu)$. From this difference we can now compute average and uncertainty of the χ^2 variation as we know the statistical distributions of both μ and s^2. Specifically, the calculation involves expectation values of powers of $(\mu - \theta_0) \sim \mathcal{N}\left(0, s^2/(\sigma^2 N)\right)$, and

$$\begin{aligned}\left\langle \left(\frac{s}{\sigma}\right)^n \right\rangle &= \left\langle \left(\frac{x}{k^2(N-1)^{n/2}}\right) \right\rangle = \frac{1}{k^n(N-1)^{n/2}} \left\langle x^{n/2} \right\rangle = \\ &= \frac{2^{n/2}}{k^n(N-1)^{n/2}} \frac{\Gamma\left(\frac{n}{2}+\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)} = \frac{1}{k^n} G_N(n)\,,\end{aligned} \tag{B.6}$$

where $G_N(n)$ is a shorthand for the gamma functions and power terms.

The result of this calculation is the expected value of the χ^2 eq. (4.2),

$$\left\langle \Delta\chi^2 \right\rangle = a_N\, x + b_N\, x^2 + c_N\, x^3 + d_N\, x^4\,, \quad x = 1/k\,, \tag{B.7}$$

where the coefficients are given by

$$a_N = a\, G_N(1)\,, \tag{B.8}$$

$$b_N = b\,, \tag{B.9}$$

$$c_N = c\left[G_N(3) + 3/N G_N(1)\right]\,, \tag{B.10}$$

$$d_N = d\,\frac{N^2+7N-6}{N(N-1)}\,, \tag{B.11}$$

along with their variances,

$$\sigma_a^2 = a^2\left(1 - G_N^2(1)\right)\,, \tag{B.12}$$

$$\sigma_b^2 = b^2\,\frac{6N-4}{N(N-1)}\,, \tag{B.13}$$

$$\sigma_c^2 = c^2\left(\frac{N^3+19N^2+3N-15}{N(N-1)^2} - \frac{c_N^2}{c^2}\right)\,, \tag{B.14}$$

$$\begin{aligned}\sigma_d^2 = d^2\Bigg(&\frac{(N+1)(N+3)(N+5)}{(N-1)^3} + 28\frac{(N+1)(N+3)}{N(N-1)^2} \\ &+ 140\frac{N+1}{N^2(N-1)} + 540\frac{1}{N^3} - \frac{d_N^2}{d^2}\Bigg)\,.\end{aligned} \tag{B.15}$$

We may note that in the limit $N \to \infty$ the coefficients eqs. (B.8) to (B.11) become N-independent and approach the Taylor coefficients eq. (B.1), while the variances vanish. Particularly, the model predicts a constant coefficient b_N and equal to b, the analogous to the tolerance.

B.2 $\Delta\chi^2$ fit results

As explained in section 4.1.1 we compute the expectation value $\langle\Delta\chi^2\rangle$ for different values of sigma-fraction k at fixed number of replicas N. In fig. B.1 are shown the results of this computation. The paraboloids are fitted to a polynomial of fourth degree in $x = 1/k$ to extract the coefficients a_N, b_N, c_N, d_N. The large error bars in the last two plots on the bottom right are due to the use of only one batch in the calculation.

We may observe, as we already did in section 4.1.1 for the lowest number of replicas, that the χ^2 variations for `n3fit` (orange) are always smaller than those of NNPDF3.1 (teal) and feature scaling (yellow). This will end up in lower values for the tolerance parameter, and thus confirms the stability and accuracy of the predictions of the `n3fit` methodology. A final comment for the feature scaling results: initially, the χ^2 variations are lower than those of NNPDF3.1, but once the number of replicas reaches $N = 300$ they show similar values. In particular, for $N \geq 250$ the $\langle\Delta\chi^2\rangle$ shape for the feature scaling is slightly asymmetric, which implies that inefficiency and parabolic deviation effects are becoming relevant as they are related to the odd terms of eq. (B.7). Since the increase in the number of replicas corresponds to an overall decrease in the expectation value of $\Delta\chi^2$, we may deduce that the predictions of NNPDF3.1 scale better with respect to the number of replicas than those of feature scaling.

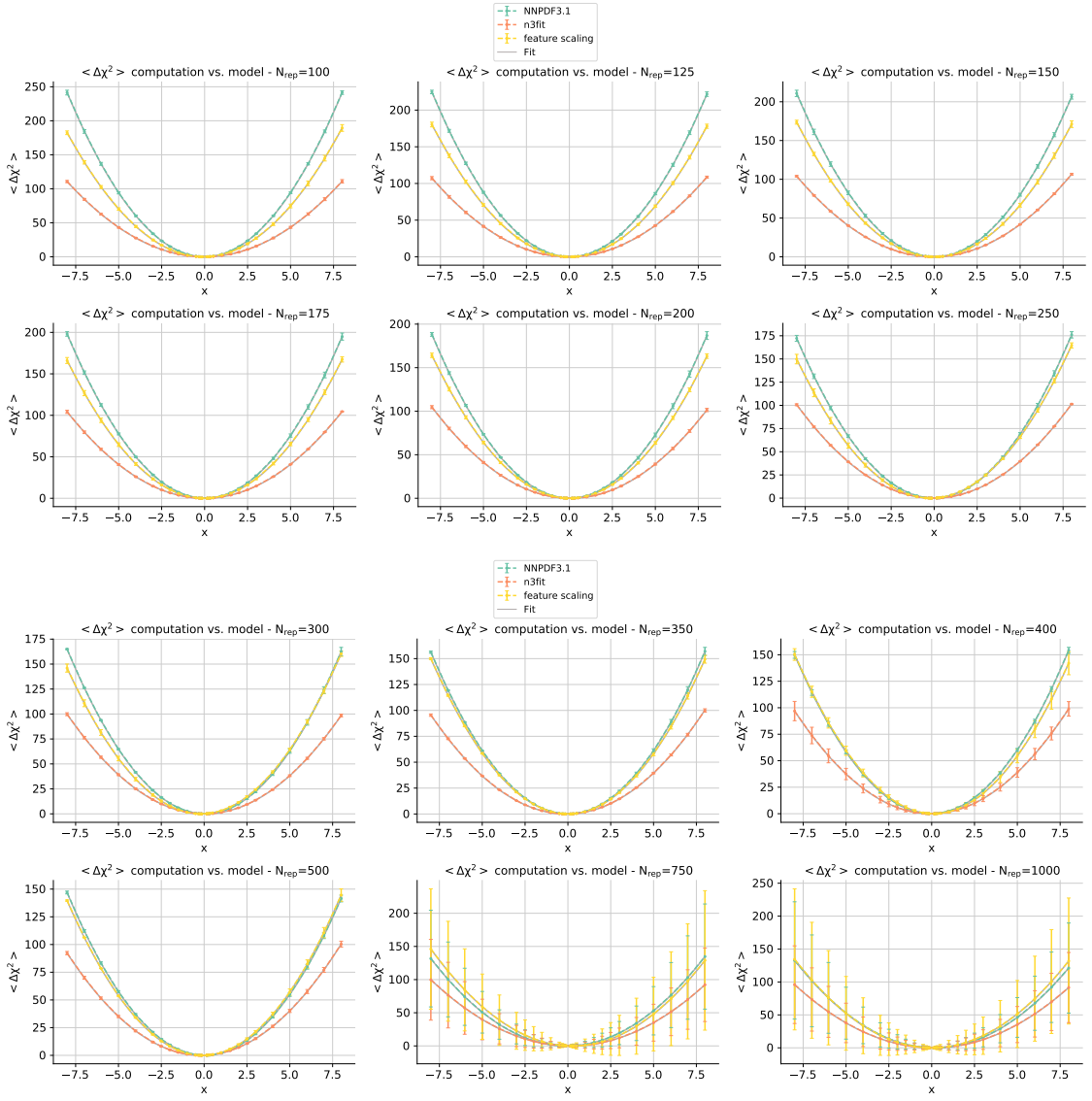

Figure B.1: Comparison between the computed values of $\langle \Delta\chi^2 \rangle$ (error bars) as a function of $x = 1/k$ from the Monte Carlo sets NNPDF3.1 (teal), `n3fit` (orange) and feature scaling (yellow). From top to bottom and left to right are shown the results for increasing number of replicas, namely $N_{\text{rep}} = 100, 125, 150, 175, 200, 250, 300, 350, 400, 500, 750, 1000$.

Bibliography

[1] E. Eichten, I. Hinchliffe, Kenneth D. Lane, and C. Quigg. Super Collider Physics. *Rev. Mod. Phys.*, 56:579–707, 1984. [Addendum: Rev.Mod.Phys. 58, 1065–1073 (1986)].

[2] NNPDF website. http://nnpdf.mi.infn.it.

[3] N3PDF webite. http://n3pdf.mi.infn.it.

[4] Stefano Carrazza and Juan Cruz-Martinez. Towards a new generation of parton densities with deep learning models. *The European Physical Journal C*, 79(8), Aug 2019.

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

[6] Stefano Carrazza, Stefano Forte, Zahari Kassabov, and Juan Rojo. Specialized minimal PDFs for optimized LHC calculations. *The European Physical Journal C*, 76(4), Apr 2016.

[7] Curtis G. Callan. Broken Scale Invariance in Scalar Field Theory. *Physical Review D*, 2(8):1541–1547, 1970.

[8] K. Symanzik. Small distance behaviour in field theory and power counting. *Communications in Mathematical Physics*, 18(3):227–246, Sep 1970.

[9] David J. Gross and Frank Wilczek. Ultraviolet Behavior of Non-Abelian Gauge Theories. *Phys. Rev. Lett.*, 30(26):1343–1346, Jun 1973.

[10] Particle Data Group. Review of Particle Physics. *Progress of Theoretical and Experimental Physics*, 2020(8), 08 2020. 083C01.

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

[12] M. Breidenbach et al. Observed Behavior of Highly Inelastic Electron-Proton Scattering. *Physical Review Letters*, 23(16):935–939, 1969.

[13] Joshua P. Ellis. TikZ-Feynman: Feynman diagrams with TikZ. *Computer Physics Communications*, 210:103–123, Jan 2017.

[14] Richard P. Feynman. Very High-Energy Collisions of Hadrons. *Physical Review Letters*, 23(24):1415–1417, 1969.

[15] C. G. Callan. High-Energy Electroproduction and the Constitution of the Electric Current. *Physical Review Letters*, 22(4):156–159, 1969.

[16] J. D. Bjorken. Asymptotic Sum Rules at Infinite Momentum. *Phys. Rev.*, 179:1547–1553, 1969.

[17] R. D. Field. *Applications Of Perturbative QCD (Frontiers in Physics)*. Addison-Wesley Publishing Company, 1989.

[18] T. Kinoshita. Mass singularities of Feynman amplitudes. *J. Math. Phys.*, 3:650–677, 1962.

[19] T.D. Lee and M. Nauenberg. Degenerate Systems and Mass Singularities. *Phys. Rev.*, 133:B1549–B1562, 1964.

[20] R.Keith Ellis, W.James Stirling, and B.R. Webber. *QCD and collider physics*, volume 8. Cambridge University Press, 2 2011.

[21] G. Altarelli, R.K. Ellis, and G. Martinelli. Leptoproduction and Drell-Yan processes beyond the leading approximation in chromodynamics. *Nuclear Physics B*, 143(3):521 – 545, 1978.

[22] Alan D. Martin. Proton structure, Partons, QCD, DGLAP and beyond, 2008.

[23] A. Vogt, S. Moch, and J.A.M. Vermaseren. The three-loop splitting functions in QCD: the singlet case. *Nuclear Physics B*, 691(1-2):129–181, Jul 2004.

[24] S. Moch, J.A.M. Vermaseren, and A. Vogt. The three-loop splitting functions in QCD: the non-singlet case. *Nuclear Physics B*, 688(1-2):101–134, Jun 2004.

[25] G.P. Salam and J. Rojo. A Higher Order Perturbative Parton Evolution Toolkit (HOPPET). *Computer Physics Communications*, 180(1):120–156, Jan 2009.

[26] M. Botje. QCDNUM: Fast QCD evolution and convolution. *Computer Physics Communications*, 182(2):490–532, Feb 2011.

[27] Valerio Bertone, Stefano Carrazza, and Juan Rojo. APFEL: A PDF evolution library with QED corrections. *Computer Physics Communications*, 185(6):1647–1668, Jun 2014.

[28] A. Vogt. Efficient evolution of unpolarized and polarized parton distributions with QCD-Pegasus. *Computer Physics Communications*, 170(1):65–92, Jul 2005.

[29] Zahari Kassabov. Reportengine: A framework for declarative data analysis (Version v0.27). http://doi.org/10.5281/zenodo.2571601, Feb 2019.

[30] M. Buza, Y. Matiounine, J. Smith, R. Migneron, and W.L. van Neerven. Heavy quark coefficient functions at asymptotic values $Q^2 \gg m^2$. *Nuclear Physics B*, 472(3):611–658, Jul 1996.

[31] M. Buza, Y. Matiounine, J. Smith, and W. L. van Neerven. Charm electroproduction viewed in the variable-flavour number scheme versus fixed-order perturbation theory. *The European Physical Journal C*, 1(1-2):301–320, Mar 1998.

[32] R. S. Thorne and W. K. Tung. PQCD Formulations with Heavy Quark Masses and Global Analysis, 2008.

[33] Marco Guzzi, Pavel M. Nadolsky, Hung-Liang Lai, and C.-P. Yuan. General-mass treatment for deep inelastic scattering at two-loop accuracy. *Phys. Rev. D*, 86:053005, Sep 2012.

[34] R. S. Thorne. Variable-flavor number scheme for next-to-next-to-leading order. *Physical Review D*, 73(5), Mar 2006.

[35] Matteo Cacciari, Mario Greco, and Paolo Nason. The p_T spectrum in heavy-flavour hadroproduction. *Journal of High Energy Physics*, 1998(05):007–007, May 1998.

[36] Stefano Forte, Eric Laenen, Paolo Nason, and Juan Rojo. Heavy quarks in deep-inelastic scattering. *Nuclear Physics B*, 834(1-2):116–162, Jul 2010.

[37] The NNPDF Collaboration, Richard D. Ball, et al. A Determination of the Charm Content of the Proton, 2016.

[38] Andy Buckley, James Ferrando, Stephen Lloyd, Karl Nordstrom, Ben Page, Martin Ruefenacht, Marek Schoenherr, and Graeme Watt. LHAPDF6: parton density access in the LHC precision era, 2014.

[39] S. Alekhin, J. Blümlein, and S. Moch. The ABM parton distributions tuned to LHC data. *Physical Review D*, 89(5), Mar 2014.

[40] Tie-Jiun Hou et al. New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC, 2019.

[41] Zhiqing Zhang. HERA Inclusive Neutral and Charged Current Cross Sections and a New PDF Fit, HERAPDF 2.0, 2015.

[42] L. A. Harland-Lang, A. D. Martin, P. Motylinski, and R. S. Thorne. Parton distributions in the LHC era: MMHT 2014 PDFs. *The European Physical Journal C*, 75(5), May 2015.

[43] The NNPDF Collaboration, Richard D. Ball, et al. Parton distributions from high-precision collider data, 2017.

[44] M. Arneodo et al. Measurement of the proton and deuteron structure functions, F_2^p and F_2^d, and of the ratio. *Nuclear Physics B*, 483(1-2):3–43, Jan 1997.

[45] M. Arneodo, A. Arvidson, B. Badełek, M. Ballintijn, G. Baum, J. Beaufays, I.G. Bird, P. Björkholm, M. Botje, C. Broggini, and et al. Accurate measurement of F_2^d/F_2^p and R_d – R_p. *Nuclear Physics B*, 487(1-2):3–26, Mar 1997.

[46] L.W. Whitlow, E.M. Riordan, S. Dasu, S. Rock, and A. Bodek. Precise measurements of the proton and deuteron structure functions from a global analysis of the SLAC deep inelastic electron scattering cross sections. *Physics Letters B*, 282(3):475 – 482, 1992.

[47] A.C. Benvenuti et al. A high statistics measurement of the proton structure functions $F_2(x, Q^2)$ and R from deep inelastic muon scattering at high Q^2. *Physics Letters B*, 223(3):485 – 489, 1989.

[48] A.C. Benvenuti et al. A high statistics measurement of the deuteron structure functions $F_2(x, Q^2)$ and R from deep inelastic muon scattering at high Q^2. *Physics Letters B*, 237(3):592 – 598, 1990.

[49] H1 and ZEUS Collaborations. Combination of Measurements of Inclusive Deep Inelastic $e^{\pm}p$ Scattering Cross Sections and QCD Analysis of HERA Data, 2015.

[50] H1 Collaboration and ZEUS Collaboration. Combination and QCD Analysis of Charm Production Cross Section Measurements in Deep-Inelastic ep Scattering at HERA, 2012.

[51] F. D. Aaron et al. Measurement of the charm and beauty structure functions using the H1 vertex detector at HERA. *The European Physical Journal C*, 65(1-2):89–109, Nov 2009.

[52] H. Abramowicz et al. Measurement of beauty and charm production in deep inelastic scattering at HERA and measurement of the beauty-quark mass. *Journal of High Energy Physics*, 2014(9), Sep 2014.

[53] G. Onengut et al. Measurement of nucleon structure functions in neutrino scattering. *Phys. Lett. B*, 632:65–75, 2006.

[54] M. Goncharov et al. Precise measurement of dimuon production cross sections in $\nu\mu$Fe and $\bar{\nu}\mu$Fe deep inelastic scattering at the Fermilab Tevatron. *Physical Review D*, 64(11), Nov 2001.

[55] David Alexander Mason. *Measurement of the strange - antistrange asymmetry at NLO in QCD from NuTeV dimuon data*. PhD thesis, Oregon U., 2006.

[56] G. Moreno et al. Dimuon production in proton-copper collisions at $\sqrt{s} = 38.8$ GeV. *Phys. Rev. D*, 43:2815–2835, May 1991.

[57] R. S. Towell et al. Improved measurement of the $\bar{d}/\bar{u}$ asymmetry in the nucleon sea. *Physical Review D*, 64(5), Aug 2001.

[58] FNAL E866/NuSea Collaboration. Absolute Drell-Yan Dimuon Cross Sections in 800 GeV/c pp and pd Collisions, 2003.

[59] Jason C. Webb. Measurement Of Continuum Dimuon Production In 800-GeV/C Proton-Nucleon Collisions, 2003.

[60] V. M. Abazov et al. Measurement of the muon charge asymmetry in $p\bar{p} \to W + X \to \mu\nu + X$ events at $\sqrt{s} = 1.96$TeV. *Physical Review D*, 88(9), Nov 2013.

[61] V.M. Abazov et al. Measurement of the electron charge asymmetry in $p\bar{p} \to W + X \to e\nu + X$ decays in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$TeV. *Physical Review D*, 91(3), Feb 2015.

[62] V. M. Abazov et al. Measurement of the shape of the boson rapidity distribution for $p\bar{p} \to Z/\gamma^* \to e^+e^- + X$ events produced at $\sqrt{s}$ of 1.96 TeV. *Physical Review D*, 76(1), Jul 2007.

[63] T. Aaltonen et al. Measurement of $d\sigma/dy$ of Drell-Yan e^+e^- pairs in the Z Mass Region from $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV. *Physics Letters B*, 692(4):232–239, Sep 2010.

[64] A. Abulencia et al. Measurement of the Inclusive Jet Cross Section using the k_T algorithm in $p\bar{p}$ Collisions at $sqrts = 1.96$ TeV with the CDF II Detector, 2007.

[65] S. Catani, Yuri L. Dokshitzer, M.H. Seymour, and B.R. Webber. Longitudinally invariant K_t clustering algorithms for hadron hadron collisions. *Nucl. Phys. B*, 406:187–224, 1993.

[66] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, Apr 2008.

[67] CMS collaboration. A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging. 7 2009.

[68] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet user manual. *The European Physical Journal C*, 72(3), Mar 2012.

[69] S. Catani and M.H. Seymour. A general algorithm for calculating jet cross sections in NLO QCD. *Nuclear Physics B*, 485(1-2):291–419, Feb 1997.

[70] F. Krauss T. Gehrmann, A. Gehrmann-De Ridderand et al. MC@NNLO - NNLOjet collaboration. https://mcatnnlo.org.

[71] Richard D. Ball et al. Parton Distribution Benchmarking with LHC Data, 2012.

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

[73] Richard D. Ball et al. Fitting parton distribution data with multiplicative normalization uncertainties. *Journal of High Energy Physics*, 2010(5), May 2010.

[74] Fred James and Matthias Winkler. MINUIT User's Guide. 6 2004.

[75] Juan Rojo. Machine Learning tools for global PDF fits, 2018.

[76] Richard D. Ball et al. Parton distributions for the LHC run II. *Journal of High Energy Physics*, 2015(4), Apr 2015.

[77] J. Pumplin, D. R. Stump, and W. K. Tung. Multivariate fitting and the error matrix in global analysis of data. *Physical Review D*, 65(1), Dec 2001.

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

[79] J. C. Collins and J. Pumplin. Tests of goodness of fit to multiple data sets. 2001.

[80] G. Watt and R. S. Thorne. Study of Monte Carlo approach to experimental uncertainty propagation with MSTW 2008 PDFs. *Journal of High Energy Physics*, 2012(8), Aug 2012.

[81] Jon Pumplin. Experimental consistency in parton distribution fitting. *Physical Review D*, 81(7), Apr 2010.

[82] Jon Pumplin. Parametrization dependence and delta chi-square in parton distribution fitting. *Physical Review D*, 82(11), Dec 2010.

[83] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt. Parton distributions for the LHC. *The European Physical Journal C*, 63(2):189–285, Jul 2009.

[84] Richard D. Ball et al. A first unbiased global NLO determination of parton distributions and their uncertainties. *Nuclear Physics B*, 838(1-2):136–206, Oct 2010.

[85] Rabah Abdul Khalek et al. A First Determination of Parton Distributions with Theoretical Uncertainties, 2019.

[86] Stefano Forte and Stefano Carrazza. Parton distribution functions, 2020.

[87] François Chollet et al. Keras. https://keras.io, 2015.

[88] Martín Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. Software available from tensorflow.org.

[89] J. Bergstra, D. Yamins, and D. D. Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages I–115–I–123. JMLR.org, 2013.

[90] Matthew D. Zeiler. ADADELTA: An Adaptive Learning Rate Method, 2012.

[91] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[92] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[93] Valerio Bertone, Stefano Carrazza, and Nathan P. Hartland. APFELgrid: A high performance tool for parton density determinations. *Computer Physics Communications*, 212:205–209, Mar 2017.

[94] Juan M Cruz-Martinez, Stefano Carrazza, and Roy Stegeman. Studying the parton content of the proton with deep learning models, 2020.

[95] Luca Talon. Optimization of parton density uncertainties. Master's thesis, University of Milan, 2019.