Parton Distribution Functions for Precision LHC Phenomenology

Francesco Cerutti

Universitat de Barcelona

Facultat de Física

Departament d'Estructura i Constituents de la Matèria

Parton Distribution Functions for Precision LHC Phenomenology

Advisor:
Dr. José Ignacio Latorre Sentís

Ph.D. Thesis of:
Francesco Cerutti

Academic Year
2012-2013

“A few observation and much reasoning lead to error;
many observations and a little reasoning to truth.”

Alexis Carrel

With gratitude
to my dear friends
Alfredo and Mariona,
Diego and Silvia.

List of Publications

This thesis, except where otherwise stated, is based on the following papers:

The NNPDF Collaboration: R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, N. P. Hartland, J. I. Latorre, J. Rojo, M. Ubiali
"Reweighting and Unweighting of Parton Distributions and the LHC W lepton asymmetry data"
Nucl. Phys. **B855**, 608 (2012), arXiv:1108.1758.

The NNPDF Collaboration: R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, M. Ubiali
"Unbiased global determination of parton distributions and their uncertainties at NNLO and at LO"
Nucl. Phys. **B855**, 153 (2012), arXiv:1107.2652.

S. Lionetti, R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, M. Ubiali
"Precision determination of α_s using an unbiased global NLO parton set"
Phys. Lett. **B701**, 346 (2011), arXiv:1103.2369.

The NNPDF Collaboration: R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, M. Ubiali
"On the Impact of NMC Data on Parton Distributions and Higgs Production at the Tevatron and the LHC"
Phys. Lett. **B704**, 36 (2011), arXiv:1102.3182.

The NNPDF Collaboration: R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, M. Ubiali
"Impact of Heavy Quark Masses on Parton Distributions and LHC Phenomenology"
Nucl. Phys. **B849**, 296 (2011), arXiv:1101.1300.

The NNPDF Collaboration: R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, M. Ubiali
"Reweighting NNPDFs: the W lepton asymmetry"
Nucl. Phys. **B849**, 112 (2011), arXiv:1012.0836.

Contents

Introduction

The last four decades of the twentieth century saw the planning and realization of several particle colliders. They all were mainly constructed to gain introspection into the "particle zoo" up to then observed and to refine and test that highly predictive theory that is the Standard Model (SM). Their design allowed for lepton-lepton (as LEP at CERN), lepton-hadron (as HERA at DESY), and hadron-hadron (as Tevatron at FermiLab) collisions. The huge information collected along the years brought to an extremely precise determination of many Standard Model parameters and to the assessment of perturbative Quantum ChromoDynamics (QCD). The priceless experience earned during those years is the basis that made possible the construction of the largest and by far more precise working machine that humanity ever built: the Large Hadron Collider (LHC) at CERN, near Geneva.

The two proton beams that cross each other in four points of the 27 kilometers-long ring previously occupied by LEP are bended by more than 1600 superconductive magnets, each positioned with a displacement resolution of 50 micrometers. Many features of the LHC and of the experiments connected to it are achieved using the latest technology. The correct and harmonic functioning of all the components involved implies a huge, challenging effort of thousands of experts.

The large amount of data produced each year ($\sim$25 petabytes) needs to be analyzed into detail in order to test new theories and to look for possible hints of new physics. Indeed, an increased accuracy on experimental measurements calls for a much stricter control over theoretical uncertainties. The results of the present work go in that direction.

At hadron colliders, even if the beams contain protons as it is the case for LHC, what really interact during a collision are proton's constituents: quarks and gluons. As a consequence, all the measured hadronic observables depend on the behaviour of these constituents, that cannot be isolated and accelerated up to a known energy before their collision because of quark confinement. The strong coupling constant has an opposite behaviour if compared to the coupling of Quantum ElectroDynamics (QED), i.e. it has large values for low energy scales, breaking the perturbative regime: large couplings imply strong interactions that no more can be considered as perturbations.

However, one of the consequences of the non-Abelian nature of QCD, asymptotic freedom, allows for a perturbative treatment of the strong interaction for sufficiently high energy scales. In fact, increasing the resolution the α_s value gets smaller. The observables factorize into a non-perturbative part, Parton Distribution Functions (PDFs), and hard-scattering matrix elements that can instead be determined from perturbation theory. Parton sets must be extracted from data, and are then used to compute theoretical predictions and to perform phenomenological studies.

These theoretical considerations, preceded by a brief historical introduction, are presented into more detail in Chapter 1. There, the main features of perturbative QCD are discussed. Special attention is paid to that features and tools on which mostly rely the results here discussed, as are the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) evolution equations or the treatment of heavy quark effects. It is very important, in order to deliver a competitive parton set for precision LHC phenomenology, to address all theoretical and procedural issues that may introduce non-negligible uncertainties on top of the experimental ones. Indeed, these would affect all the analyses that relies on such a PDF set. All the efforts of the groups that produce PDF fits from available data attempt to address these issues.

The various procedures followed to this extent by the various groups are discussed in Chapter 2. The work presented in this thesis is realized within the Neural Network Parton Distribution Functions (NNPDF) collaboration and its methodology is presented in Chapter 3. As it is not known which functional form may be suitable for the correct description of a set of parton distributions, the parametrization used must be very flexible. Commonly, a polynomial functional form is used, but in this way the fit is likely to be biased because of lack of flexibility. In NNPDF methodology, a combination of Monte Carlo sampling in the space of data with a non-linear parametrization given by neural networks is used. In this way it is possible to use a much larger parameter set ($\sim$ ten times), minimized through a properly tuned Genetic Algorithm (GA). To correctly determine the best set of parameters, the fit needs to stop its minimization before the minimum possible χ^2 value is reached. This is done using dynamical stopping based on a cross-validation method. This avoids to enter the overlearning regime, i.e. the fitting of statistical fluctuations of datapoints.

In Chapter 4, at first are discussed results relative to the NNPDF2.1 NLO release, and then also the LO and NNLO determinations. For the energies and the performance reached at the LHC, heavy quark effects must be included and the impact of the neglected higher perturbative orders must be kept under control. This is done through the comparison with results from the same NNPDF collaboration, using previous releases or *ad hoc* modifications of them made to assess the impact of each single new feature introduced,

but also through the comparison with parton sets from other groups. This benchmark is also performed computing predictions for a set of LHC standard candles in Chapter 5. It is important to check consistency of the results computed with different parton sets, both among them and with the first available LHC measurements.

As one of the major sources of theoretical uncertainty is included in the dependence on α_s, an estimate of $\alpha_s(M_Z)$ using NNPDF methodology is finally presented. A parabolic fit of the χ^2 profile given by a set of PDF sets with varying $\alpha_s(M_Z)$ values is performed. The profile is composed by the values for total χ^2 on the global fit for each different $\alpha_s(M_Z)$ value and their errors. Also, the same kind of fit is performed on subsets of the main global data as on single experiments. Hence, not only the best value for $\alpha_s(M_Z)$ is determined but is also performed a complete study of the role that each single set plays inside the fit.

Chapter 1

Perturbative Quantum Chromodynamics

In this Chapter, I will discuss the main properties of quantum chromodynamics. In the first part the evidences that first hinted at and then confirmed the existence of quarks and gluons are presented. Then, the properties of asymptotic freedom and confinement are introduced. In the second part of the Chapter, a general discussion on Deep-Inelastic Scattering (DIS) is used to gain insight on parton model and its improvements, DGLAP equations, and heavy quarks treatment. The discussion here developed follows mainly the arguments presented in Refs. [1–3], along with the cited references throughout the Sections.

1.1 Quarks and Gluons

The second half of last century was characterized by the discovery of an astonishing number of new particles. This fact, together with the peculiar structure emerged in organizing these particles (hadrons) into groups, suggested that they were not elementary: on the contrary, they were composed by other more fundamental constituents. A scheme that grouped together particles with the same spin, the SU(3) symmetry, was introduced in 1961 by Gell-Mann [4] and Ne'eman [5]. Two groups of mesons, with spin-0 and spin-1 respectively, and two of baryons, with spin-1/2 and spin-3/2 respectively. The feeling that this symmetry was the correct one was enhanced by the discovery of new particles, predicted by this model. Gell-Mann [6] and Zweig [7] in 1964 first proposed the existence of smaller, three-flavored particles inside hadrons: the quarks. But still the Δ^{++} baryon, that accordingly to the quark model is composed of three up quarks with parallel spins, showed that another ingredient was missing. Three identical quarks cannot form an antisymmetric state, as is the case for a fermion. A year later Han, Nambu, and Greenberg

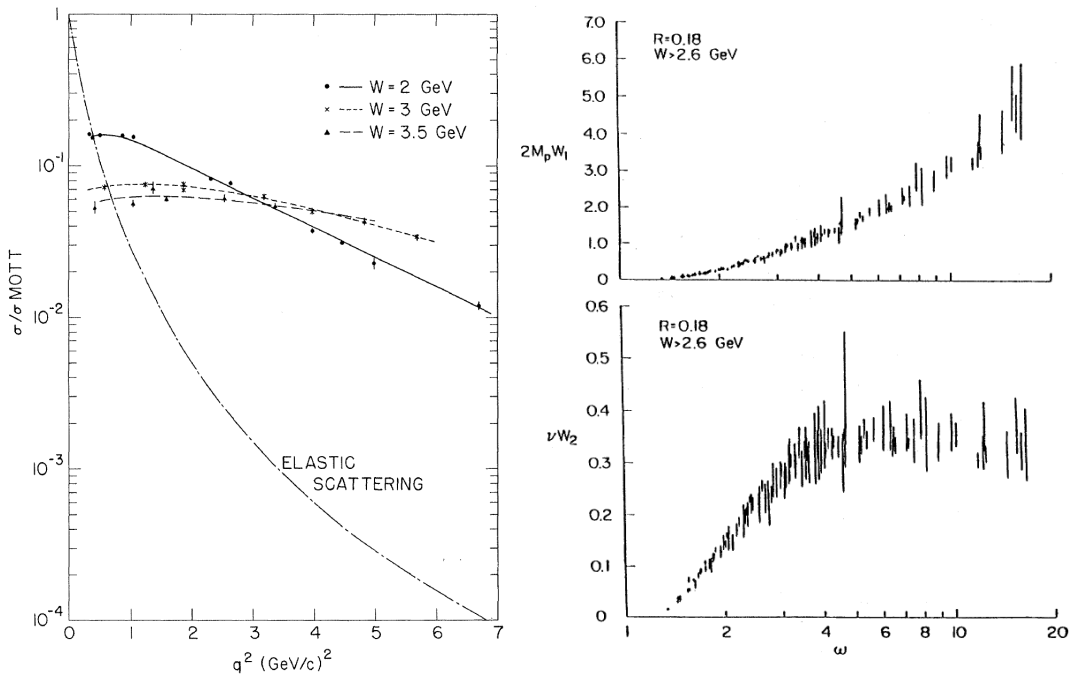

Figure 1.1: Left: ratio σ/σ_{Mott} vs q^2 for constant values of the invariant mass of the recoiling target system $W = 2$, 3, and 3.5 GeV. Also shown is the elastic $e - p$ scattering cross-section divided by σ_{Mott} [9]. Right: $2MW_1$ and νW_2 for the proton as functions of ω for $W > 2.6$ GeV, $q^2 > 1$ (GeV/c)^2 and $R = 0.18$ [10].

solved this problem by proposing that quarks possess an additional SU(3) gauge degree of freedom: the color charge.

It took several years to experimentally prove the real existence of these approximately point-like components of the hadrons. Deep-inelastic scattering experiments performed at SLAC [8] showed that electrons were hitting some kind of hard core inside the target protons. As discussed in Ref. [9] and summarized in Ref. [10], two unexpected hints of this were the weak q^2 dependence of the ratio σ/σ_{Mott} for that process and scaling, where $q^2 = 2EE'(1 - \cos\theta)$ is the momentum transfer of the proton with E the energy of the incident electron, E' the energy of the scattered electron, and θ the scattering angle.
It was clear that σ/σ_{Mott}, with

$$\sigma_{Mott} = \frac{e^4}{4E^2}\frac{\cos^2\frac{\theta}{2}}{\sin^4\frac{\theta}{2}} \quad , \tag{1.1}$$

decreased much more slowly with q^2 than the same quantity for elastic scattering (Fig. 1.1). This was a similar situation to the one that brought Ernest Rutherford to the discovery of the atomic nucleus [11], where the probability of large-angle alpha particles scattering from gold atoms was found to be far larger than anticipated by J.J. Thompson's model.

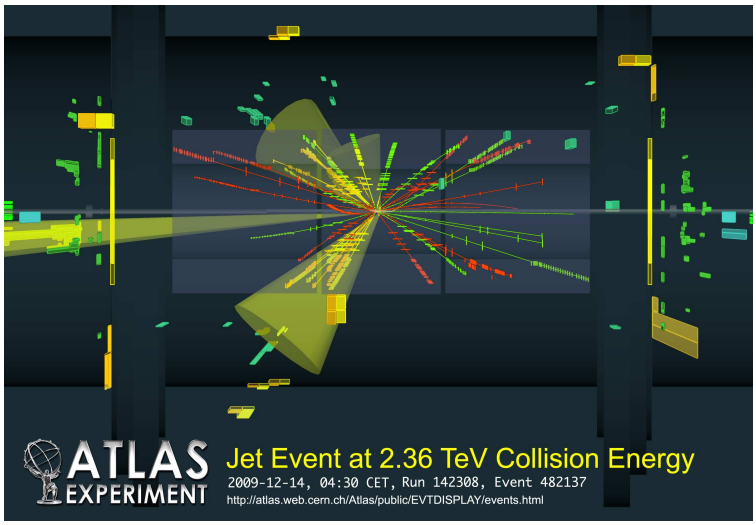

Figure 1.2: Di-jets event detected inside the ATLAS experiment [13].

Moreover, introducing the general expression for the differential cross-section for unpolarized electrons scattering from unpolarized nucleons [12] as

$$\frac{d^2\sigma}{d\Omega dE'} = \sigma_{Mott}\left[W_2 + 2W_1 \tan^2\frac{\theta}{2}\right] \ , \tag{1.2}$$

where W_1 and W_2 are the structure functions of the target, these two expressions are expected to be functions of both q^2 and ν, where the latter is the energy loss of the scattered electron, $E - E'$. Bjorken suggested in a private communication that the quantities νW_2 and $2M_P W_1$ become functions only of the ratio $\omega = 2M_P\nu/q^2$ in the limit of high q^2 and ν, that is:

$$2M_P W_1(\nu, q^2) \rightarrow F_1(\omega) \tag{1.3}$$

$$\nu W_2(\nu, q^2) \rightarrow F_2(\omega) \tag{1.4}$$

where M_P is the proton mass (M from now on). In later contributions the ω variable was replaced by its inverse, x. The scaling behaviour of the structure functions is shown in Fig. 1.1.

Another important result is related to sum rules: it is possible to write

$$F_2^p(x) = \nu W_2^p(x) \tag{1.5}$$

$$= x\left[Q_u^2\left(u_p(x) + \bar{u}_p(x)\right) + Q_d^2\left(d_p(x) + \bar{d}_p(x)\right)\right] \tag{1.6}$$

where functions u, d, $\bar{u}$, and $\bar{d}$ are the momentum distributions of the up and down quarks and the respectively antiquark distributions (more details on these distributions will be

discussed in the next Chapter), and Q_u, Q_d are their electrical charges.
Considering the sum rule for neutrons and protons

$$I_1 = \int_1^\infty \nu W_2(\omega) \frac{d\omega}{\omega} \tag{1.7}$$

using charge symmetry it reads:

$$\frac{1}{2}\int_0^1 [F_2^p(x) + F_2^n(x)]dx = \frac{Q_u^2 + Q_d^2}{2}\int_0^1 x[u_p(x) + \bar{u}_p(x) + d_p(x) + \bar{d}_p(x)]dx. \tag{1.8}$$

If quarks and antiquarks carries the nucleon's total momentum, the integral on the right hand side of Eq. (1.8) should equal 1. So, assuming this, it should be

$$\frac{1}{2}\int_0^1 [F_2^p(x) + F_2^n(x)]dx = \frac{Q_u^2 + Q_d^2}{2} = \frac{1}{2}\left[\frac{4}{9} + \frac{1}{9}\right] = \frac{5}{18} = 0.28 \ . \tag{1.9}$$

The evaluations of the experimental sum from proton and neutron results over the entire kinematic range studied yielded

$$\frac{1}{2}\int_0^1 [F_2^p(x) + F_2^n(x)]dx = 0.14 \pm 0.005 \ , \tag{1.10}$$

thus suggesting that half of the nucleon's momentum is carried by neutral constituents, the gluons, which do not interact with the electron. To identify the constituents of the nucleon as quarks, some electron scattering results played a crucial role during the mid 70s. Mainly, measurements of the ratio between longitudinal and transversal components $R = \sigma_L/\sigma_T$ (determining the spin 1/2 of the constituents), measurements of σ_n/σ_p ratio between neutron and proton results (excluding purely diffractive models and discarding identical momentum distributions for the constituents), and evaluations of sum rules (that confirms that the constituents has fractional electric charge and, as anticipated, that half of the nucleon's momentum is carried by neutral gluons).

In that same period, among the facts that definitely convinced physics community that quarks and gluons really existed it is worth citing the discovery of J/Ψ particles in 1974 at Brookhaven National Laboratory [14], that cannot be explained without adding to the up, down and strange quarks also a fourth *charm* quark (c), the observation in 1975 of hadron jets [15] from high energy electron-positron collisions (detailed analysis indicated that these jets were the footprints of individual spin-1/2 particles), the discovery in 1977 at Fermilab [16] of the *bottom* quark (b), and the observation of additional jets emerging from electron-positron collisions at DESY in 1979 [17], that gave visible evidence of gluons existence. In particular, the production and observation of multi-jets events has

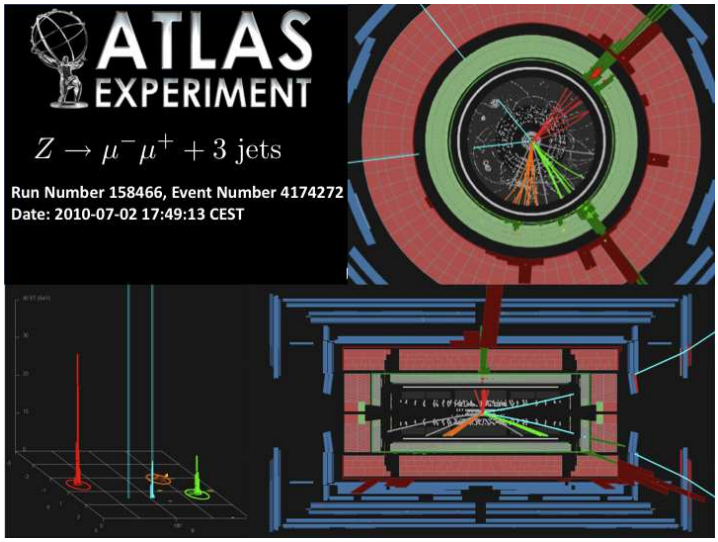

Figure 1.3: A three-jets event detected inside the ATLAS experiment [13].

nowadays become a frequent and usual fact into LHC detectors (Figs. 1.2-1.3).

1.2 Asymptotic Freedom and Confinement

As introduced in the previous Section, the 70s have been crucial years for the development of QCD. That years paved the way for high energy collider physics of today. As a matter of fact, without a precise and complete description of strong interaction it would be impossible today to study complex events as, for example, a Higgs boson decay. A turning point in understanding and describing strong interaction was reached in 1973. Up to that moment, one of the unanswered questions was about the possibility of observing free *partons*, as R.P. Feynman called nucleon's constituents. Whatever the energy of collision in the experiments, these particles were never directly observed.

In the summer of 1973, D.J. Gross and F. Wilczek [18] and H.D. Politzer [19] demonstrated asymptotic freedom for QCD. The coupling of strong interaction gets smaller increasing the energy scale of the interaction and it is clear that asymptotic freedom and quark confinement are two different consequences of α_s behaviour. The strong interaction, among the fundamental forces, has the peculiarity of increasing its strength with the distance between partons. This explains why all the attempts of tearing apart a free parton from a nucleon failed. As soon as the energy of the interaction grows enough, a quark-antiquark pair is generated. This mechanism also explains the phenomenon of *hadronization*, and jets are its detectable manifestation. If quark confinement makes it impossible to have free partons, asymptotic freedom guarantees the possibility of a per-

turbative approach to QCD for sufficiently large Q scales.

In the following, asymptotic freedom is approached in a more detailed way. QCD is a renormalizable Lagrangian field theory of the strong interaction. It is the study of the SU(3) Yang-Mills theory of color-charged fermions (quarks) and of a set of spin 1 gauge fields (gluons) which mediate the interaction among quarks as well as among themselves. The four-spinor describing the quark (antiquark) field is denoted by q_j^a ($\bar{q}_j^a$). The index j refers to flavor while index a to the three color degrees. A vector gauge field A_μ^a, which transforms under the adjoint representation of SU(3), is introduced to describe the flavorless colored octet of gluons. The index a is again a color index but this time it runs over eight values instead of only three. The Feynman rules required for a perturbative analysis of QCD can be derived from a Lagrangian density which is given by

$$\mathcal{L} = \mathcal{L}_{classical} + \mathcal{L}_{gauge-fixing} + \mathcal{L}_{ghost}. \tag{1.11}$$

The minimal locally gauge invariant Lagrangian density implied by this SU(3) symmetry is

$$\mathcal{L}_{classical} = -\frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu} + \sum_j \bar{q}_j^a (i\gamma^\mu D_\mu - m_j)_{ab} q_j^b. \tag{1.12}$$

These terms describe the interaction of spin-1/2 quarks of mass m and massless spin-1 gluons, where

$$G_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + g f^a{}_{bc} A_\mu^b A_\nu^c \tag{1.13}$$

is the field strength tensor derived from the gluon field A_μ^a, with f_{abc} the structure constants of the SU(3) color group, and

$$D_\mu = \partial_\mu - ig\frac{1}{2}\lambda_a A_\mu^a \tag{1.14}$$

is the covariant derivative, with g the coupling constant of the interaction, that is unique and universal, and with λ_a the SU(3) color matrices:

$$[\lambda_a, \lambda_b] = i f_{abc} \lambda_c \,. \tag{1.15}$$

It is impossible to define a gluon field propagator without choosing a gauge. For this reason is needed a gauge-fixing term $\mathcal{L}_{gauge-fixing}$ that allows, for a specific value of the gauge parameter α, to fix the class of covariant gauges:

$$\mathcal{L}_{gauge-fixing} = -\frac{1}{2\alpha} (\partial^\mu A_\mu^a)^2. \tag{1.16}$$

To avoid that unphysical degrees of freedom propagate in covariant gauges, a non-Abelian

theory such as QCD must be supplemented by a ghost Lagrangian, which is given by

$$\mathcal{L}_{ghost} = \partial_\mu \eta^{a\dagger}(D^\mu_{ab}\eta^b), \tag{1.17}$$

where η^a are massless fields that obey to Fermi statistics.

This is a renormalizable theory. As a consequence, a physical observable perturbatively computed also depends on the scale μ at which the divergences of the perturbative series are subtracted. Considering as an example a dimensionless physical observable F which depends on a single energy scale Q, when this is calculated as a perturbative expansion in the coupling $\alpha_s = g^2/4\pi$ the perturbative series requires renormalization to remove ultraviolet divergences. So, in general, the observable F also depends on the ratio Q^2/μ^2. Another consequence of this procedure is that also the renormalized coupling α_s depends on the choice made for the subtraction point μ. Physical observables must give a description of the phenomenology independent of the method used to obtain them. This means that they cannot depend on the choice of the renormalization scale. It is possible to express this condition in a mathematical way: given that F is dimensionless, it can only depend on the ratio Q^2/μ^2 and α_s, so it must be

$$\mu^2 \frac{d}{d\mu^2} F(Q^2/\mu^2, \alpha_s) = \left[\mu^2 \frac{\partial}{\partial \mu^2} + \mu^2 \frac{\partial \alpha_s}{\partial \mu^2}\frac{\partial}{\partial \alpha_s}\right] F = 0 \ . \tag{1.18}$$

To solve this differential equation, the running coupling constant $\alpha_s(Q^2)$ is introduced. As a first step, Eq. (1.18) is rewritten in a more compact way as

$$\left[-\frac{\partial}{\partial t} + \beta(\alpha_s)\frac{\partial}{\partial \alpha_s}\right] F(e^t, \alpha_s) = 0 \tag{1.19}$$

defining

$$t \equiv \ln\left(\frac{Q^2}{\mu^2}\right) \ , \ \beta(\alpha_s) \equiv \mu^2 \frac{\partial \alpha_s}{\partial \mu^2} \, . \tag{1.20}$$

Starting from Eq. (1.20), the running coupling $\alpha_s(Q^2)$ is implicitly defined as

$$t = \int_{\alpha_s}^{\alpha_s(Q^2)} \frac{dx}{\beta(x)} \ , \ \alpha_s = \alpha_s(\mu^2) \tag{1.21}$$

and differentiating with respect to t and α_s the first part of Eq. (1.21) it follows

$$\frac{\partial \alpha_s(Q^2)}{\partial t} = \beta(\alpha_s(Q^2)) \ , \ \frac{\partial \alpha_s(Q^2)}{\partial \alpha_s} = \frac{\beta(\alpha_s(Q^2))}{\beta(\alpha_s)} \ . \tag{1.22}$$

Computing the observable F in $\mu^2 = Q^2$ it gives

$$F(e^t, \alpha_s)|_{\mu^2=Q^2} = F(1, \alpha_s(Q^2)) \,, \tag{1.23}$$

and taking into account the expressions of Eq. (1.22) it is easy to see that this last expression of F, with all the scale dependence included in the running coupling, is a solution of Eq. (1.19).

In QCD, a way to determine the running of the coupling constant is to expand perturbatively the β function as [18, 20]

$$Q^2 \frac{\partial \alpha_s}{\partial Q^2} = \beta(\alpha_s) = -b\alpha_s^2(1 + b'\alpha_s + b''\alpha_s^2 + O(\alpha_s^3)) \;\; , \tag{1.24}$$

where

$$b = \frac{11C_A - 2n_f}{12\pi} = \frac{33 - 2n_f}{12\pi} \tag{1.25}$$

$$b' = \frac{17C_A^2 - 5C_A n_f - 3C_F n_f}{2\pi(11C_A - 2n_f)} = \frac{153 - 19n_f}{2\pi(33 - 2n_f)} \tag{1.26}$$

$$b'' = \frac{2857C_A^3 + (54C_F^2 - 615C_F C_A - 1415C_A^2)n_f + (66C_F + 79C_A)n_f^2}{288\pi^2(11C_A - 2n_f)} \tag{1.27}$$

$$= \frac{77139 - 15099n_f + 325n_f^2}{288\pi^2(33 - 2n_f)} \tag{1.28}$$

with

$$C_F = \frac{4}{3} \qquad , \qquad C_A = 3 \tag{1.29}$$

as is the case for SU(3), and n_f the number of active light flavors.
Contrary to what happens in the case of QED, in QCD the first coefficient of the β function is negative for $n_f < 16$, thanks to the gluon-gluon interaction of this non-Abelian theory. As briefly discussed here below, this is a crucial feature of the theory.

If both $\alpha_s(\mu^2)$ and $\alpha_s(Q^2)$ are in the perturbative region, it is possible to truncate the expansion in Eq. (1.24) and solve Eq. (1.21):

$$t = \int_{\alpha_s}^{\alpha_s(Q^2)} \frac{dx}{-bx^2} = \left[\frac{1}{b\alpha_s(Q^2)} - \frac{1}{b\alpha_s(\mu^2)} \right]. \tag{1.30}$$

Recalling that $t = \ln(Q^2/\mu^2)$ it can be written

$$\alpha_s(Q^2) = \frac{\alpha_s(\mu^2)}{1 + \alpha_s(\mu^2)bt}. \tag{1.31}$$

For large Q^2, t also grows and so this expression tells us that $\alpha_s(Q^2)$ decreases. This

property is called *asymptotic freedom* and it depends on the sign of the b coefficient. In QED the sign is the opposite and so α_{QED}'s behaviour is also the opposite. This means that to a large Q^2 corresponds a large coupling, thus leading outside the perturbative region. Then, of course, are small Q^2 that in QCD leads outside perturbative region.

The threshold at which the perturbative approach is spoiled out is easily introduced with the parameter Λ^2. When the term $\alpha_s(\mu^2)bt \simeq 1$ then $\alpha_s(Q^2)$ is no more as small as required by perturbation theory.
The parameter Λ^2 is defined as

$$\Lambda^2 = \mu^2 e^{1/\alpha_s(\mu^2)b} \tag{1.32}$$

and it follows

$$\alpha_s(Q^2) = \left.\frac{\alpha_s(\mu^2)}{1+\alpha_s(\mu^2)b(\ln\frac{Q^2}{\Lambda^2}+\ln\frac{\Lambda^2}{\mu^2})}\right|_{\mu^2=\Lambda^2} = \frac{\alpha_s(\Lambda^2)}{1+\alpha_s(\Lambda^2)b\ln\frac{Q^2}{\Lambda^2}}. \tag{1.33}$$

The value of Λ^2 refers to the energy scale at which partons cannot be considered as free particles, and are strongly bound to each other forming a hadronic state. There are several possible approximate solutions for the renormalization group Eq. (1.24) and an alternative to the use of such approximated expressions is to exactly solve Eq. (1.24) in a numerical way. In these cases Λ^2 is not defined and in determinations of the strong coupling it is a common practice to quote α_s at a given scale rather than fixing a Λ^2 parameter. The typical scale for α_s is the mass of the Z boson M_Z. Asymptotic freedom is a fundamental property that allows for precision QCD phenomenology. In the next Section the case of deep-inelastic scattering will be considered and discussed to illustrate the main features of the parton model and perturbative QCD.

1.3 The Parton Model

The parton model can be considered as the leading order approximation of perturbative QCD. As represented in Fig. 1.4, the DIS process here considered is a lepton-hadron scattering. The incoming and outgoing four-momenta of the lepton are labeled by k^μ and k'^μ respectively. As a consequence, the momentum transfer is $q^\mu = k^\mu - k'^\mu$. Assuming to have as hadronic target a proton, its momentum is labeled by P^μ. The process here considered is a high-energy scattering, and so lepton masses can be neglected.

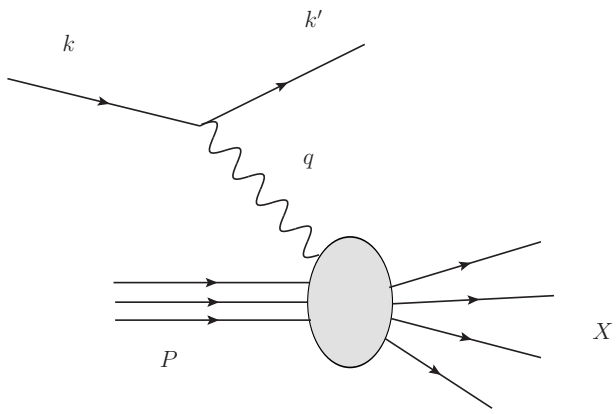

Figure 1.4: Diagram of a DIS process.

The standard deep inelastic variables are defined by

$$Q^2 \equiv -q^2 \tag{1.34}$$

$$\nu \equiv \frac{P \cdot q}{M} \tag{1.35}$$

and also

$$x = \frac{1}{\omega} = \frac{Q^2}{2M\nu} \tag{1.36}$$

$$y = \frac{q \cdot P}{k \cdot P} = 1 - \frac{E'}{E}\,. \tag{1.37}$$

Notice that only two variables among x, y, and Q^2 are independent of each other as the center of mass energy of the lepton-nucleon system is $s = Q^2/(xy) + M^2$.

It is possible to determine a general expression for the cross-section of such a process paying attention to some peculiarity. To write down a Feynman diagram and compute the matrix element in this case is nontrivial as the interaction here is not between two point-like particles as is the case for example in an electron-muon scattering. Here, instead of a muon, a more complicated hadronic object has to be considered.

In general it is still possible to write

$$d\sigma \sim L_{\mu\nu} W^{\mu\nu}\,, \tag{1.38}$$

where $L_{\mu\nu}$ represents the lepton tensor while $W^{\mu\nu}$ parametrizes the unknown form of the current relative to the proton system in the matrix element. As $L_{\mu\nu}$ is symmetric, the hadronic tensor must also be symmetric. Moreover, the most general form of $W^{\mu\nu}$ must include terms constructed out of $g^{\mu\nu}$ and the independent momenta P and q.

It is possible to write

$$W^{\mu\nu} = -F_1 g^{\mu\nu} + \frac{F_2}{M^2} P^\mu P^\nu + \frac{F_4}{M^2} q^\mu q^\nu + \frac{F_5}{M^2}(P^\mu q^\nu + q^\mu P^\nu)\,, \tag{1.39}$$

and as at the hadronic vertex must be verified, as for the leptonic part, current conservation, that is

$$q_\mu W^{\mu\nu} = q_\nu W^{\mu\nu} = q^\mu L_{\mu\nu} = q^\nu L_{\mu\nu} = 0\,, \tag{1.40}$$

imposing these conditions on $W^{\mu\nu}$ allows to express F_4 and F_5 in terms of F_1 and F_2, obtaining

$$W^{\mu\nu} = F_1\left(-g^{\mu\nu} + \frac{q^\mu q^\nu}{q^2}\right) + \frac{F_2}{P\cdot q}\left(P^\mu - \frac{P\cdot q}{q^2}q^\mu\right)\left(P^\nu - \frac{P\cdot q}{q^2}q^\nu\right). \tag{1.41}$$

In this expression the F_3 contribution is not considered: this term appears as a parity-violating structure function, in the case in which the electron is substituted by a neutrino. In the present treatment electroweak contribution will not be considered, as the result can be easily generalized. Including the phase space factor for the outgoing electron and the flux factor, the inclusive differential cross-section for this process is obtained as

$$d\sigma = \frac{1}{4(k\cdot P)}\left[\frac{4\pi e^4}{q^4} L_{\mu\nu} W^{\mu\nu}\right]\frac{d^3k'}{2E'(2\pi)^3}\,, \tag{1.42}$$

where the leptonic tensor is easy to compute and it gives

$$L_{\mu\nu} = 2\Big[k_\mu k'_\nu + k'_\mu k_\nu - \frac{Q^2}{2} g_{\mu\nu}\Big]\,. \tag{1.43}$$

As for the hadronic tensor, terms associated with parity violation are neglected.
The 4π factor that multiplies the electric charge appears because of the normalization of $W^{\mu\nu}$:

$$W^{\mu\nu} = \frac{1}{2}\frac{1}{4\pi}\sum_X \langle h|J^{\dagger\mu}|X\rangle\langle X|J^\nu|h\rangle (2\pi)^4\delta(P_X - k - P)\,. \tag{1.44}$$

For small-wavelength virtual photons (large momentum transfer) the nucleon internal structure is resolved. A simple model is considered, in which the virtual photon scatters off a point-like quark constituent inside the proton carrying a fraction ξ of the proton momentum, so that: $p^\mu = \xi P^\mu$. If an $e^+e^- \to q\bar{q}$ process is now considered, crossing symmetry relates it to the partonic process $e^- q \to e^- q$, so that is possible to easily write the spin- and color-averaged squared matrix element summed over the final colors and

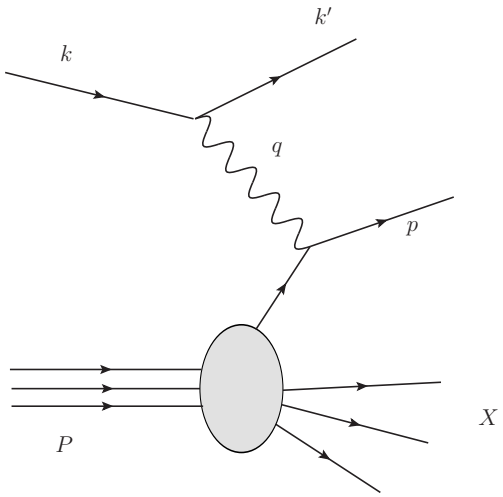

Figure 1.5: Diagram of a DIS partonic process.

spins in terms of the Mandelstamm [1] invariants

$$\overline{\sum}|\mathcal{M}|^2 = 2e_q^2 e^4 \frac{\hat{s}^2 + \hat{u}^2}{\hat{t}^2} \tag{1.45}$$

with

$$\hat{s} = (k+p)^2 = \xi Q^2/xy \tag{1.46}$$
$$\hat{t} = (k-k')^2 = -Q^2 \tag{1.47}$$
$$\hat{u} = (p-k')^2 = \hat{s}(y-1) \tag{1.48}$$

that substituted into the standard result for the massless $2 \to 2$ scattering gives

$$\frac{d\hat{\sigma}}{dQ^2} = \frac{2\pi\alpha^2 e_q^2}{Q^4}[1 + (1-y)^2]\,. \tag{1.49}$$

An important consequence comes out of the mass-shell constraint for the outgoing quark, in fact

$$p'^2 = (p+q)^2 = q^2 + 2p\cdot q = -2P\cdot q(x-\xi) = 0\,, \tag{1.50}$$

from which it follows necessarily that $x = \xi$.
In other words, the fraction of the momentum carried by the struck parton is equal to the

[1] The ^symbol is used for variables and observables that refers to the partonic level of the process.

Bjorken-x. Finally, it is possible to write

$$\frac{d^2\hat{\sigma}}{dxdQ^2} = \frac{4\pi\alpha^2}{Q^4}[1+(1-y)^2]\frac{1}{2}e_q^2\delta(x-\xi)\,. \tag{1.51}$$

Going back to Eq. (1.42), this expression can be rewritten using the results for lepton and hadron tensors Eqs. (1.41)-(1.43) and introducing standard variables defined in Eqs. (1.34)-(1.37).
It reads

$$\begin{aligned}\frac{d^2\sigma}{dxdQ^2} = \frac{y}{Q^2}\frac{d^2\sigma}{dxdy} &= \frac{e^4}{4\pi Q^4}\left[y^2F_1(x,Q^2)+(1-y)\frac{F_2(x,Q^2)}{x}\right] \\ &= \frac{4\pi\alpha^2}{Q^4}\left[[1+(1-y)^2]F_1+\frac{1-y}{x}(F_2-2xF_1)\right] \end{aligned} \tag{1.52}$$

and by comparison with Eq. (1.51) it follows that

$$\hat{F}_1(x) = \frac{1}{2}e_q^2\delta(x-\xi)\,, \tag{1.53}$$

$$\hat{F}_2(x) - 2x\hat{F}_1(x) = 0\,. \tag{1.54}$$

Partons inside the nucleon not always bring the same fraction of the nucleon momentum. This fact is observed looking at experimental results: structure functions appear to be distributions in x rather than a delta function. To take this into account

$$\begin{aligned}2xF_1(x) = F_2(x) &= \sum_i \int_0^1 d\xi f_i(\xi)\hat{F}_2(\xi) \\ &= \sum_i \int_0^1 d\xi f_i(\xi)xe_i^2\delta(x-\xi) = \sum_i e_i^2 x f_i(x)\,, \end{aligned} \tag{1.55}$$

where $f_i(\xi)$ are probability distributions that give a weight to the quark structure functions. The fact that $F_2 = 2xF_1$ is directly related to the spin-1/2 property of the quarks and is called Callan-Gross relation. It confirms that the longitudinal structure function is zero, as a spin-1/2 quark cannot absorb a longitudinally polarized vector boson. A spin-0 quark would not absorb transversely polarized vector bosons, giving as a consequence that $F_1 = 0$ and $F_L = F_2$.

1.4 Perturbative Corrections to Parton Model

As already said, the parton model is a leading order approximation of QCD. Taking into account next to leading order effects, gluons start playing a fundamental role. In fact, while in the parton model p_T is zero, if gluon radiation is considered this is no more true. In this case it can be that $p_T \sim Q^2$, as the transverse momentum is not restricted to be small. This is a direct consequence of a real gluon emission from the quark, before or after its interaction with the virtual boson (Fig. 1.6). Another consequence of the inclusion of next to leading order contributions is that structure functions lose their scaling property. This feature here described for parton model is broken in QCD by logarithms of Q^2. This means that structure functions are functions of both x and Q^2. Let us reconsider the expression of Eq. (1.55): this relates the structure functions defined in the proton frame with the ones defined in the parton frame.
Introducing

$$z = \frac{Q^2}{2p_i \cdot q} = \frac{x}{y} \tag{1.56}$$

as a new variable for the parton frame, it is possible to rewrite the relation between the structure functions in a more general way as

$$F_2(x, Q^2) = \sum_i \int_0^1 dz \int_0^1 dy f_i(y) \delta(x - zy) \hat{F}_2(z, Q^2)\,, \tag{1.57}$$

where F_2 and $\hat{F}_2$ are used as an example, x is fixed, z and y obey the constrain $x = zy$, and $f_i(y)$ are the parton distribution functions for a parton of momentum p_i.
Integrating over z gives

$$F_2(x, Q^2) = \sum_i \int_x^1 \frac{dy}{y} f_i(y) \hat{F}_2(x/y, Q^2)\,, \tag{1.58}$$

from which it is easy to recover the result of Eq. (1.55) when $\hat{F}$ is computed at leading order.

In this Section the $O(\alpha_s)$ corrections to the parton model are considered. The contributions that need to be taken into account are represented in Figs. 1.6-1.7. The tree-level contribution, the one-loop correction, the real gluon emission, and the gluon-boson fusion need to be considered. Divergences may appear due to soft gluons (i.e. almost zero energy gluons) and collinear gluons (i.e. gluons emitted parallel to the incoming or outgoing quark). The main aim of this Section and of the following one is to show how it is possible to "cure" these singularities. Dimensional regularization is a standard method used to perform ultraviolet regularization in the context of renormalization, and is the same

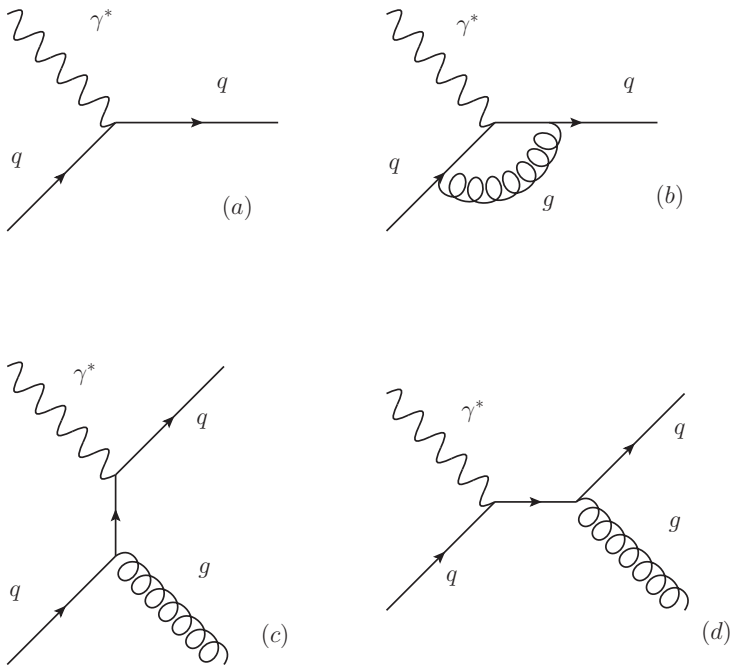

Figure 1.6: The Feynman diagrams for the tree-level (a), the one-loop (b), and the gluon emission (c-d) processes.

method applied in this case. Looking at the loop contribution, the gluon is simultaneously soft and collinear. An effective vertex [1] can be written, describing both the tree-level $q\gamma^* \to q$ and the one-loop diagrams as

$$i\Gamma^\mu = -iee_q\Big[1 - \frac{\alpha_s}{4\pi}C_F\Big(\frac{4\pi\mu^2}{Q^2}\Big)^\epsilon \frac{\Gamma(1-\epsilon)}{\Gamma(1-2\epsilon)}\Big(\frac{2}{\epsilon^2} + \frac{3}{\epsilon} + 8 + \frac{\pi^2}{3} + O(\epsilon)\Big)\Big]\,, \qquad (1.59)$$

where $d = (4 - 2\varepsilon)$. It is possible to explicitly check that soft and collinear singularities at the origin of the double pole in ε are removed by the real gluon contribution (Fig. 1.6). To do this, the hadronic tensor is considered in the form

$$W_{\mu\nu}(P,q) = \sum_i \int_0^1 \frac{dy}{y} f_i(y)\hat{W}^{(i)}_{\mu\nu}(yP,q)\,, \qquad (1.60)$$

where $\hat{W}^{(i)}_{\mu\nu}$ is the partonic tensor for the parton i. This expression can be written thanks to the factorization theorem [21,22]. It is useful to split the hadronic tensor in its longitudinal

and transverse components, writing

$$W_T = \sum_i \int_0^1 \frac{dy}{y} f_i(y) \hat{W}_T^{(i)}(yP, q) \tag{1.61}$$

$$W_L = \sum_i \int_0^1 \frac{dy}{y^3} f_i(y) \hat{W}_L^{(i)}(yP, q)\,. \tag{1.62}$$

Using these expressions, the structure functions are

$$\begin{aligned} \frac{F_2(x,Q)}{x} &= \frac{1}{1-\varepsilon} W_T + \frac{3-2\varepsilon}{1-\varepsilon}\frac{4x^2}{Q^2} W_L \\ F_1(x,Q) &- \frac{F_2(x,Q)}{2x} = -\frac{4x^2}{Q^2} W_L\,. \end{aligned} \tag{1.63}$$

For massless fermions the longitudinal projection vanishes and the Callan-Gross relation is recovered. The transverse component for the virtual contribution is easily determined from the expression of the effective vertex, and it is

$$\hat{W}_{T,virt} = e_q^2(1-\varepsilon)\delta(1-\xi)\left\{1 - \frac{\alpha_s}{2\pi} C_F \Big(\frac{4\pi\mu^2}{Q^2}\Big)^{\epsilon} \frac{\Gamma(1-\epsilon)}{\Gamma(1-2\epsilon)} \left(\frac{2}{\epsilon^2} + \frac{3}{\epsilon} + 8 + \frac{\pi^2}{3}\Big)\delta(1-\xi)\right\} \tag{1.64}$$

while, to obtain the real contribution for the transverse component, the square amplitude of the sum between the s- and t-channels of the process $\gamma^* q \to q' g$ (Fig. 1.6) needs to be computed. The result is

$$\begin{aligned} \hat{W}_{T,real} &= e_q^2 \frac{\alpha_s}{2\pi} C_F \Big(\frac{4\pi\mu^2}{Q^2}\Big)^{\epsilon} (1-\varepsilon) \frac{\Gamma(1-\epsilon)}{\Gamma(1-2\epsilon)} \\ &\left\{ \left(\frac{2}{\epsilon^2} + \frac{3}{2\epsilon} + \frac{7}{2}\Big)\delta(1-\xi) - \frac{1}{\varepsilon}\frac{1+\xi^2}{(1-\xi)_+} + (1+\xi^2)\left(\frac{\log(1-\xi)}{1-\xi}\right)_+ \right. \\ &\left. -\frac{1+\xi^2}{1-\xi}\log\xi - \frac{3}{2}\frac{1}{(1-\xi)_+} + 3 - \xi + O(\varepsilon)\right\}. \end{aligned} \tag{1.65}$$

The plus-prescription is a distribution defined as

$$[f(z)]_+ = f(z) - \delta(1-z)\int_0^1 d\xi f(\xi). \tag{1.66}$$

The double pole of the virtual contribution (that corresponds to a soft and collinear singularity) is exactly cancelled by an identical negative double pole from the real contribution. This fact is not accidental: suitably defined inclusive observables are free of singularities in the massless limit [23–25]. As can be seen immediately by a comparison between

Eq. (1.64) and Eq. (1.65), a single pole is still present. This is due to the emission of a collinear gluon. Before discussing how to get rid of this kind of singularities, it is better to look in a more detailed way at their physical origin.

A quark radiates a gluon before interacting with a virtual photon. The region of interest is the collinear kinematic one. In this region it can be used the Weizsäker-Williams approximation [26] [27]. The fermionic propagator has momentum $(p-k)$. Let's call this quantity p'. Then it is possible to write

$$\frac{\not{p}'}{p'^2} = \frac{u(p')\bar{u}(p')}{p'^2}\Big(1+O(|k_T|^2)\Big). \tag{1.67}$$

Because of the collinearity of the gluon the fermionic propagator can be approximated with an on-shell fermion. Using the Sudakov kinematics this fact is obvious: a four-vector can be written on a basis composed by the momentum p, a four-vector n, and a transverse vector k_T so that

$$p^2 = n^2 = n\cdot k_T = p\cdot k_T = 0 \;\;,\;\; 2p\cdot n = 1 \tag{1.68}$$

with

$$p^\mu = (P,0,0,P) \tag{1.69}$$
$$n^\mu = (\frac{1}{2P},0,0,-\frac{1}{2P}). \tag{1.70}$$

So the k^μ four-vector can be written as

$$k = p(1-z) + yn - k_T\,. \tag{1.71}$$

Reminding that $k^2 = 0$, from the previous expression it follows

$$y = \frac{|k_T|^2}{1-z}\,. \tag{1.72}$$

For $|k_T|^2 \to 0$, then, also $y \to 0$ and as

$$p-k = zp - yn + k_T \tag{1.73}$$
$$(p-k)^2 = -zy - |k_T|^2 = -\frac{|k_T|^2}{1-z} \tag{1.74}$$

it is clear from the last expression that $p' \to 0$.

Thanks to this approximation the matrix element factorizes like this:

$$|\mathcal{M}|^2 = |\mathcal{M}_{qg}|^2 \frac{1}{p'^4} |\mathcal{M}_{\gamma q}|^2 \Big(1 + O(|k_T|^2)\Big). \tag{1.75}$$

Focusing on the partonic cross-section for this process, it can be computed

$$\hat{\sigma}(\gamma^* q \to q'g) = \frac{1}{K} \int d\Phi \frac{|\mathcal{M}_{qg}|^2 |\mathcal{M}_{\gamma q}|^2}{p'^4} \tag{1.76}$$

where K is the flux factor and $d\Phi$ the one body phase space that can be written as

$$d\Phi = \frac{d^3k}{(2\pi)^3 2E_k} = \frac{\pi dz}{2(2\pi)^3(1-z)} d|k_T|^2. \tag{1.77}$$

The factorized matrix elements are easy to compute, and gathering the results finally gives

$$\begin{aligned} \hat{\sigma}(\gamma^* q \to q'g) &= \frac{\alpha_s}{2\pi} \int_{m^2}^{\hat{s}/4} \frac{d|k_T|^2}{|k_T|^2} \int_0^1 dz \left(P_{qq}(z) e_q^2 \delta(z-x) + O(|k_T|^2) \right) \\ &= e_q^2 \frac{\alpha_s}{2\pi} P_{qq}(x) \log\left(\frac{Q^2}{m^2}\right) + \dots \end{aligned} \tag{1.78}$$

where $P_{ij}(x)$ is a splitting function that can be considered as the probability of finding a parton j with momentum xp inside a parton i with momentum p. This contribution is not subject to the theorems already cited in this Section that eliminate the double pole singularity of the one loop diagram. The limit $|k_T|^2 \to 0$ corresponds to a long-range part of the strong interaction which is not calculable in perturbation theory. After applying the Weizsäker-Williams approximation, the divergent part is isolated into the $\mathcal{M}_{qg}$ term and going back to the expression of Eq. (1.58) for the structure functions follows

$$F_2(x,Q^2) = x \sum_{q,\bar{q}} e_q^2 \left\{ q_0(x) + \frac{\alpha_s}{2\pi} \int_0^1 \frac{dz}{z} q_0\Big(\frac{x}{z}\Big) \left[P(z) \log\left(\frac{Q^2}{m^2}\right) + C(z) \right] + \dots \right\}, \tag{1.79}$$

where the parton distribution function q_0 is considered as a "bare" distribution, that can now be used to absorb collinear singularities at a factorization scale μ writing

$$q(x,\mu) = q_0(x) + \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} q_0\Big(\frac{x}{z}\Big) \left[P(z) \log\left(\frac{\mu^2}{m^2}\right) + C(z) \right] + \dots \tag{1.80}$$

The factorization scale μ is introduced by simply splitting the logarithm in Eq. (1.78) as

$$\log\frac{Q^2}{m^2} = \log\frac{Q^2}{\mu^2} + \log\frac{\mu^2}{m^2}. \tag{1.81}$$

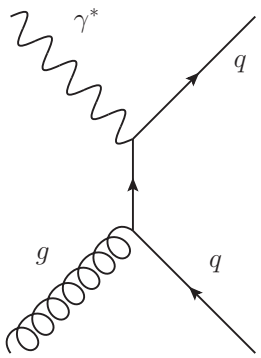

Figure 1.7: The Feynman diagram for $\gamma g \to qq$ process.

It is possible then to write Eq. (1.79) as

$$F_2(x,Q^2) = x\sum_{q,\bar{q}} e_q^2 \Bigg\{ q(x,\mu^2) + \frac{\alpha_s}{2\pi}\int_0^1 \frac{dz}{z} q_0\Big(\frac{x}{z}\Big)\left[P(z)\log\Big(\frac{Q^2}{\mu^2}\Big) + C(z)\right] + \ldots \Bigg\}. \tag{1.82}$$

Looking at the result, it is clear that the unphysical scale m^2 has been pushed inside the PDF and now the structure function only presents the factorization scale μ^2 in its right hand part. In the next Section will be explicitly discussed F_2 independence from this scale.

The last contribution that needs to be considered is the gluon initiated process $g\gamma^* \to q\bar{q}$. To obtain the corresponding result, the procedure is pretty much the same, and so for gluon

$$\hat{F}_2^g(x,Q^2) = x\sum_{q\bar{q}} e_q^2 \frac{\alpha_s}{2\pi}\left(P_{qg}(x)\log\frac{Q^2}{m^2} + C_g(x)\right). \tag{1.83}$$

Using the same arguments as before, the structure functions in the $\overline{MS}$ scheme can be written as

$$\begin{aligned} F_2(x,Q^2) &= x\sum_{q\bar{q}} e_q^2 \int_x^1 \frac{dz}{z} q(\frac{x}{z},Q^2)\left[\delta(1-z) + \frac{\alpha_s}{2\pi}C_q^{\overline{MS}}(z) + \ldots\right] \\ &\quad + x\sum_{q\bar{q}} e_q^2 \int_x^1 \frac{dz}{z} g(\frac{x}{z},Q^2)\left[\frac{\alpha_s}{2\pi}C_g^{\overline{MS}}(z) + \ldots\right] \qquad (1.84) \\ F_L(x,Q^2) &= x\sum_{q\bar{q}} e_q^2 \int_x^1 \frac{dz}{z} q(\frac{x}{z},Q^2)\left[\frac{\alpha_s}{2\pi}2C_F z + \ldots\right] \\ &\quad + x\sum_{q\bar{q}} e_q^2 \int_x^1 \frac{dz}{z} g(\frac{x}{z},Q^2)\left[\frac{\alpha_s}{2\pi}4T_R z(1-z) + \ldots\right] \qquad (1.85) \end{aligned}$$

with

$$C_q^{\overline{MS}}(z) = C_F\Bigg[2\Big(\frac{\log(1-z)}{1-z}\Big)_+ - \frac{3}{2}\Big(\frac{1}{1-z}\Big)_+ - (1+z)\log(1-z) - \frac{1+z^2}{1-z}\log z + 3 + 2z - \Big(\frac{\pi^2}{3}+\frac{9}{2}\Big)\delta(1-z)\Bigg] \tag{1.86}$$

$$C_g^{\overline{MS}}(z) = T_R\Bigg[\Big((1-z)^2+z^2\Big)\log\Big(\frac{1-z}{z}\Big) - 8z^2 + 8z - 1\Bigg]. \tag{1.87}$$

Gathering the results obtained up to now and generalizing to the case of the exchange of any vector boson, the structure functions F_j with $j = 1, 2, 3$ at NLO can be written as

$$\frac{F_j(x,Q^2)}{x} = \int_x^1 \frac{d\xi}{\xi}\sum_{i=q,\bar{q}} g_{Vi}^2 f_i^{\overline{MS}}\Big(\frac{x}{\xi},\mu^2\Big)\Big[\delta(1-\xi) + \frac{\alpha_s}{2\pi}\Big(P_{qq}(\xi)\log\frac{Q^2}{\mu^2} + C_j^{Vi,\overline{MS}}(\xi)\Big)\Big] + g^{\overline{MS}}\Big(\frac{x}{\xi},\mu^2\Big)\frac{\alpha_s}{2\pi}\Big[P_{qg}\log\frac{Q^2}{\mu^2} + C_j^{Vg,\overline{MS}}(\xi)\Big], \tag{1.88}$$

where g_{Vi} is the coupling between the vector boson and the quark.

1.5 DGLAP Equations

A collinear emission of the kind treated in the previous Section can occur repeatedly n times before the parton interacts with the virtual boson. Because of the non-Abelian nature of QCD, the parton that finally participate in the interaction may be the result of a great number of previous splittings. The parton here probed with a DIS process seems point-like at low energy but as a short-range energy is reached, the parton appears itself surrounded by a cloud of other partons. These partons indeed share the proton's momentum and so, as rising the scale causes the number of elements inside the proton to increase, a scale dependence arouses not only in structure functions but also in PDFs. Inside the evolution of these non-perturbative objects all of these collinear singularities are absorbed, order by order, using the well-known DGLAP equations. Even if perturbation theory cannot be applied for PDFs determination, dependence on the factorization scale μ^2 can be calculated perturbatively through these equations. Looking at Eq. (1.88) it is clear that the structure function must be independent from the factorization scale μ^2, and so taking the logarithmic derivative $\partial/\partial(\log\mu^2)$ of both sides of that equation gives a differential equation for the μ^2 dependence of $q(x,\mu^2)$ that reads

$$\mu^2\frac{\partial}{\partial\mu^2}q(x,\mu^2) = \frac{\alpha_s}{2\pi}\int_x^1\frac{d\xi}{\xi}\left(P_{qq}(\xi,\mu^2)q\Big(\frac{x}{\xi},\mu^2\Big) + P_{qg}(\xi,\mu^2)g\Big(\frac{x}{\xi},\mu^2\Big)\right). \tag{1.89}$$

A more exact treatment based on the operator product expansion [28] and the renormalization group equation [29] gives

$$\frac{\partial}{\partial \log \mu^2}\begin{pmatrix} q_i \\ g \end{pmatrix}(x,\mu^2) = \frac{\alpha_s}{2\pi}\int_x^1 \sum_{j=q,\bar{q}} \frac{d\xi}{\xi}\begin{pmatrix} P_{ij}\left(\frac{x}{\xi},\alpha_s\right) & P_{ig}\left(\frac{x}{\xi},\alpha_s\right) \\ P_{gj}\left(\frac{x}{\xi},\alpha_s\right) & P_{gg}\left(\frac{x}{\xi},\alpha_s\right) \end{pmatrix}\begin{pmatrix} q_j \\ g \end{pmatrix}(\xi,\mu^2)\,, \tag{1.90}$$

where $\alpha_s \equiv \alpha_s(\mu^2)$, q_i is the generic quark distribution function, P_{ij} are the Altarelli-Parisi kernels. Equations (1.90) represent a system of $2n_f+1$ coupled integro-differential equations, with n_f the number of active flavors. It is the generalization of Eq. (1.89) to higher orders in perturbation theory. The splitting functions only depend on the scale through the strong coupling constant, as it can be seen from their N^{m}LO approximation. In fact, defining $a_s \equiv \alpha_s/2\pi$, it gives for the splitting functions $P_{ij}(x,\mu^2)$

$$P_{ij}^{N^mLO}(x,\mu^2) = \sum_{k=0}^{m} a_s^{k+1}(\mu^2) P_{ij}^{(k)}(x). \tag{1.91}$$

In the case here considered, the splitting functions are known up to NNLO and their explicit expressions are found in Refs. [30–32].

In the following, to describe the solution to the DGLAP evolution equations the discussion will refer to the treatment given in Ref. [33]. The Mellin space is considered. A theorem in fact states that given a function $f = g \otimes h$, where the symbol $\otimes$ refers to a convolution as the one considered in Eq. (1.90) between hard coefficient functions and PDFs, it follows that the Mellin transform of f

$$M[f](N) = \int_0^1 d\xi \xi^{N-1} f(\xi) \tag{1.92}$$

is given by the product of the Mellin transform of g and h as

$$M[f] = M[g]M[h]\,. \tag{1.93}$$

From considerations based on charge conjugation and flavor symmetry it is possible to rewrite the system of equations (1.90) as $2n_f - 1$ equations

$$\mu^2 \frac{\partial}{\partial \mu^2} q_{NS}^{\pm,v}(x,\mu^2) = P_{NS}^{\pm,v} \otimes q_{NS}^{\pm,v}(x,\mu^2) \tag{1.94}$$

describing the independent evolution of the non-singlet quark asymmetries

$$\begin{aligned} q^{\pm}_{NS,ij} &= q_i \pm \overline{q}_i - (q_j \pm \overline{q}_j) \\ q^{v}_{NS} &= \sum_{i=1}^{n_f} (q_i - \overline{q}_i) \end{aligned} \tag{1.95}$$

and a system of 2 equations describing the coupled evolution of the singlet and gluon parton distributions

$$\mu^2 \frac{\partial}{\partial \mu^2} \begin{pmatrix} \Sigma \\ g \end{pmatrix} (x,\mu^2) = \begin{pmatrix} P_{qq} & P_{qg} \\ P_{gq} & P_{gg} \end{pmatrix} \otimes \begin{pmatrix} \Sigma \\ g \end{pmatrix} (x,\mu^2)\,, \tag{1.96}$$

where the singlet combination, Σ, is defined as

$$\Sigma = \sum_{i=1}^{n_f} (q_i + \overline{q}_i)\,. \tag{1.97}$$

At LO $P^{(0),+}_{NS} = P^{(0),-}_{NS} = P^{(0),v}_{NS} = P^{(0)}_{qq}$. At NLO $P^{(0),-}_{NS} = P^{(0),v}_{NS}$ while all the other splitting functions are different. Starting from $\mathcal{O}(\alpha_s^2)$ all splitting functions are different from each other.

The evolution of the individual quark distributions with the scale can be computed by introducing the following set of non-singlet distributions:

$$\begin{aligned} V &= u^- + d^- + s^- + c^- + b^- + t^- \\ V_3 &= u^- - d^- \\ V_8 &= u^- + d^- - 2s^- \\ V_{15} &= u^- + d^- + s^- - 3c^- \\ V_{24} &= u^- + d^- + s^- + c^- - 4b^- \\ V_{35} &= u^- + d^- + s^- + c^- + b^- - 5t^- \\ T_3 &= u^+ - d^+ \\ T_8 &= u^+ + d^+ - 2s^+ \\ T_{15} &= u^+ + d^+ + s^+ - 3c^+ \\ T_{24} &= u^+ + d^+ + s^+ + c^+ - 4b^+ \\ T_{35} &= u^+ + d^+ + s^+ + c^+ + b^+ - 5t^+ \end{aligned} \tag{1.98}$$

where $q^{\pm}_i = q_i \pm \overline{q}_i$, and u, d, s, c, b, t are the various flavor distributions.

The combinations V_j and T_j evolve according to Eq. (1.96) with P^{-}_{NS} and P^{+}_{NS} respec-

tively, while the total valence V evolves with the P_{NS}^{v} kernel.

To find the solution of the DGLAP equations for the coupled singlet-gluon equations and for the non-singlet ones, the structure of

$$q_i(x,Q^2) = \sum_j \Gamma_{ij}(x,a_s,a_0) \otimes q_j(x,Q_0^2)\,, \tag{1.99}$$

is exploited, introducing the Mellin transforms for the functions Γ and P as

$$\Gamma_{ij}(N,a_s,a_0) = \int_0^1 dx x^{N-1} \Gamma_{ij}(x,a_s,a_0) \tag{1.100}$$

$$\gamma_{ij}(N,a_s) = \int_0^1 dx x^{N-1} a_s x P_{ij}(x,a_s)\,, \tag{1.101}$$

where γ_{ij} are the anomalous dimensions that can be written in a series of a_s as follows:

$$\gamma_{ij}(N,a_s) = \sum_{n=0}^{\infty} a_s^n \gamma_{ij}^{(n)}\,. \tag{1.102}$$

It has been pointed out before that the splitting functions (and therefore the anomalous dimensions) depend on the scale only through the coupling constant. It is then natural to choose a_s as evolution variable and rewrite the DGLAP evolution equation for the evolution kernels Γ, in Mellin-N space, as

$$a_s \frac{\partial}{\partial a_s} \Gamma_{ij}(N,a_s,a_0) = -\sum_k R_{ik} \Gamma_{kj}(N,a_s,a_0)\,, \tag{1.103}$$

where the matrix $\mathbf{R}$ has the following perturbative expansion

$$\mathbf{R} = \mathbf{R}_0 + a_s \mathbf{R}_1 + a_s \mathbf{R}_2 + \ldots \tag{1.104}$$

with

$$\mathbf{R}_0 \equiv \frac{\boldsymbol{\gamma}^{(0)}}{\beta_0} \qquad \mathbf{R}_k \equiv \frac{\boldsymbol{\gamma}^{(k)}}{\beta_0} - \sum_{i=1}^{k} \frac{\beta_i}{\beta_0} \mathbf{R}_{k-i}\,, \tag{1.105}$$

where the $\boldsymbol{\gamma}$ stands for the matrix of anomalous dimensions.

The solution of the singlet evolution equation at leading order is then

$$\mathbf{q}_{LO}(x,Q^2) = \mathbf{L}(a_s,a_0,N) \mathbf{q}_{LO}(x,Q_0^2)\,. \tag{1.106}$$

The leading order evolution operator $\mathbf{L}$ is written in terms of the eigenvalues of the leading order anomalous dimension matrix

$$\lambda_\pm = \frac{1}{2\beta_0}\left[\gamma_{qq}^0 + \gamma_{gg}^0 \pm \sqrt{\left(\gamma_{qq}^0 - \gamma_{gg}^0\right)^2 + 4\gamma_{qg}^0\gamma_{gq}^0}\right] \tag{1.107}$$

and the corresponding projector matrices

$$\mathbf{e}_\pm = \frac{\pm 1}{\lambda_+ - \lambda_-}(\mathbf{R}^{(0)} - \lambda_\mp \mathbf{I})\,, \tag{1.108}$$

in the following form:

$$\mathbf{L}(a_s, a_0, N) = \mathbf{e}_- \left(\frac{a_s}{a_0}\right)^{-\lambda_-(N)} + \mathbf{e}_+ \left(\frac{a_s}{a_0}\right)^{-\lambda_+(N)}\,. \tag{1.109}$$

The solution of the evolution Eq. (1.103) is expressed as a perturbative expansion around the LO solution $\mathbf{L}(a_s, a_0, N)$

$$\begin{aligned}\begin{pmatrix} \Sigma \\ g \end{pmatrix}(N, a_s) &= \left[\mathbf{I} + \sum_{k=1}^{\infty} a_s^k \mathbf{U}_k(N)\right] \mathbf{L}(a_s, a_0, N) \left[\mathbf{I} + \sum_{k=1}^{\infty} a_0^k \mathbf{U}_k(N)\right]^{-1} \begin{pmatrix} \Sigma \\ g \end{pmatrix}(N, a_0) \\ &\equiv \mathbf{\Gamma}_S(N, a_s, a_0) \begin{pmatrix} \Sigma \\ g \end{pmatrix}(N, a_0)\,.\end{aligned} \tag{1.110}$$

The $\mathbf{U}$ matrices introduced in the previous equation are defined by the following commutation relations

$$\begin{aligned}\left[\mathbf{U}_1, \mathbf{R}_0\right] &= \mathbf{R}_1 + \mathbf{U}_1 \\ \left[\mathbf{U}_2, \mathbf{R}_0\right] &= \mathbf{R}_2 + \mathbf{R}_1\mathbf{U}_1 + 2\mathbf{U}_2 \\ &\vdots \\ \left[\mathbf{U}_k, \mathbf{R}_0\right] &= \mathbf{R}_k + \sum_{i=1}^{k-1} \mathbf{R}_{k-i}\mathbf{U}_i + k\mathbf{U}_k \;\equiv\; \widetilde{\mathbf{R}}_k + k\mathbf{U}_k\end{aligned} \tag{1.111}$$

as

$$\mathbf{U}_k = -\frac{1}{k}\left[e_+\widetilde{\mathbf{R}}_k e_+ + e_-\widetilde{\mathbf{R}}_k e_-\right] + \frac{e_+\widetilde{\mathbf{R}}_k e_-}{\lambda_- - \lambda_+ - k} + \frac{e_-\widetilde{\mathbf{R}}_k e_+}{\lambda_+ - \lambda_- - k} \tag{1.112}$$

where

$$\widetilde{\mathbf{R}}_k = \mathbf{R}_k + \sum_{i=1}^{k-1} \mathbf{R}_{k-i}\mathbf{U}_i\,. \tag{1.113}$$

Solving recursively Eqs. (1.112)-(1.113) and the NLO approximation of Eq. (1.105) gives

$$\mathbf{R}_0 \equiv \frac{\boldsymbol{\gamma}^{(0)}}{\beta_0} \qquad \mathbf{R}_k \equiv -b_1\mathbf{R}_{k-1} + \mathcal{O}(\text{NNLO}) \tag{1.114}$$

and the NLO exact solution for the evolution factor in the singlet case can be linearized as

$$\mathbf{\Gamma}_S(N) = \left[\mathbf{L} + a_s\mathbf{U}_1\mathbf{L} - a_0\mathbf{L}\mathbf{U}_1 + a_s^2\mathbf{U}_2\mathbf{L} - a_s a_0\mathbf{U}_1\mathbf{L}\mathbf{U}_1 + a_0^2\mathbf{L}(\mathbf{U}_1^2 - \mathbf{U}_2)\right]. \quad (1.115)$$

Looking at the non-singlet combinations for the distributions of quarks, it can be seen that Eq. (1.111) also holds in this case of a scalar evolution, but as the involved objects are no more matrices the right-hand sides vanish. Consequently, it is possible to write at LO the solution for the evolution factor as

$$\Gamma^{\pm,v}_{NS,LO}(N, a_s, a_0) = \left(\frac{a_s}{a_0}\right)^{-R_0^{\rm ns}}, \quad (1.116)$$

and at NLO non-singlet solutions can be written down in a compact closed form both as an iterated solution

$$\Gamma^{\pm,v}_{\rm NS,NLO}(N, a_s, a_0) \;=\; \exp\left\{\frac{U_1^{\pm,v}}{b_1}\ln\left(\frac{1+b_1 a_s}{1+b_1 a_0}\right)\right\}\left(\frac{a_s}{a_0}\right)^{-R_0^{\rm ns}} \quad (1.117)$$

or a truncated solution

$$\Gamma^{\pm,v}_{\rm NS,NLO}(N, a_s, a_0) \;=\; \left(1 - U_1^{\pm,v}(a_s - a_0)\right)\left(\frac{a_s}{a_0}\right)^{-R_0^{\rm ns}}. \quad (1.118)$$

The x-space parton distributions are obtained by taking the inverse Mellin transforms of the solutions obtained in Eq. (1.110) which, making use of the convolution theorem, can be written as

$$q^{\pm,v}_{NS}(x, Q^2) \;=\; \int_x^1 \frac{dy}{y}\Gamma_{qq}(y, a_s, a_0)\, q^{\pm,v}_{NS}\left(\frac{x}{y}, Q_0^2\right) \quad (1.119)$$

$$\begin{pmatrix} \Sigma \\ g \end{pmatrix}(x, Q^2) \;=\; \int_x^1 \frac{dy}{y}\mathbf{\Gamma}_S(y, a_s, a_0)\begin{pmatrix} \Sigma \\ g \end{pmatrix}\left(\frac{x}{y}, Q_0^2\right). \quad (1.120)$$

The evolution kernels $\Gamma(x)$ are defined as the inverse Mellin transforms of the evolution factors introduced in Eq. (1.110)

$$\Gamma_S(x, a_s, a_0) = \int_{c-i\infty}^{c+i\infty}\frac{dN}{2\pi i}x^{-N}\Gamma_S(N, a_s, a_0). \quad (1.121)$$

Note however that all splitting functions, except the off-diagonal entries of the singlet matrix, diverge when $x = 1$. This implies that the evolution kernels $\Gamma(x)$ will likewise be divergent in $x = 1$.

In the following is shown that, like the splitting functions, the evolution factors can be

defined as distributions. To this purpose consider the generic evolution factor Γ such that (omitting the explicit dependence of Γ on the coupling a_s)

$$f(x,Q^2) = \int_x^1 \frac{dy}{y}\Gamma(y) f\left(\frac{x}{y}, Q_0^2\right) . \tag{1.122}$$

Defining the distribution

$$\Gamma_+(x) = \Gamma(x) - \gamma\delta(1-x)\,, \qquad \text{where} \quad \gamma = \int_0^1 dx\Gamma(x)\,, \tag{1.123}$$

Eq. (1.122) can then be rewritten as

$$\begin{aligned} f(x,Q^2) &= \gamma f(x,Q_0^2) + \int_x^1 \frac{dy}{y}\Gamma_+(y) f\left(\frac{x}{y}, Q_0^2\right) \\ &= \gamma f(x,Q_0^2) + \int_x^1 \frac{dy}{y}\Gamma(y)\left[f\left(\frac{x}{y}, Q_0^2\right) - y f\left(x, Q_0^2\right)\right] - f(x,Q_0^2)\int_0^x dy\Gamma(y)\,. \end{aligned} \tag{1.124}$$

Due to the subtraction Eq. (1.123), all integrals on the r.h.s of Eq. (1.124) converge and can be evaluated numerically. This expression, then, can be used to compute the parton distribution functions in x-space, determining Γ numerically from Eq. (1.121) and γ as

$$\gamma = \int_0^1 dx \int_{c-i\infty}^{c+i\infty} \frac{dN}{2\pi i} x^{-N}\Gamma(N) = \int_{c-i\infty}^{c+i\infty} \frac{dN}{2\pi i}\frac{\Gamma(N)}{1-N}\,. \tag{1.125}$$

In this singlet case, however, this prescription has been slightly modified because $\Gamma(N)\Big|_{N=1}$ is indeed infinite. So Eq. (1.124) is rewritten in another equivalent form. Let us define

$$f^{(1)}(x,Q^2) = x\, f(x,Q^2) \qquad \Gamma^{(1)}(x,Q_0^2,Q^2) = x\Gamma(x,Q_0^2,Q^2)\,.$$

Thus

$$\begin{aligned}
f^{(1)}(x,Q^2) &= x\,f(x,Q^2) = \int_x^1 \frac{dy}{y}\,\Gamma(y,Q_0^2,Q^2)\,x\,f\left(\frac{x}{y},Q_0^2\right) \\
&= \int_x^1 \frac{dy}{y}\,\Gamma^{(1)}(y,Q_0^2,Q^2)\,f^{(1)}\left(\frac{x}{y},Q_0^2\right) \\
&= \int_x^1 \frac{dy}{y}\,\Gamma^{(1)}(y,Q_0^2,Q^2)\,\left(f^{(1)}\left(\frac{x}{y},Q_0^2\right) - y f^{(1)}(x,Q_0^2)\right) \\
&+ \int_x^1 \frac{dy}{y}\,y\Gamma^{(1)}(y,Q_0^2,Q^2)\,f^{(1)}(x,Q_0^2) \\
&= \int_x^1 \frac{dy}{y}\,\Gamma^{(1)}(y,Q_0^2,Q^2)\,\left(f^{(1)}\left(\frac{x}{y},Q_0^2\right) - y f^{(1)}(x,Q_0^2)\right) \\
&+ f^{(1)}(x,Q_0^2)\left[\int_0^1 dy\,y\Gamma(y,Q_0^2,Q^2) - \int_0^x y\Gamma(y)\right] \\
\Rightarrow f(x,Q^2) &= \int_x^1 \frac{dy}{y}\,y\Gamma(y,Q_0^2,Q^2)\,\left(\frac{1}{y}f\left(\frac{x}{y},Q_0^2\right) - y f(x,Q_0^2)\right) \\
&+ f(x,Q_0^2)\left[\Gamma(N,Q_0^2,Q^2)\Big|_{N=2} - \int_0^x y\Gamma(y,Q_0^2,Q^2)\right] \qquad (1.126)
\end{aligned}$$

that is, finally, a regularized expression free of infinite contributions.

1.6 Heavy Quarks: the FONLL Scheme

In this Section the FONLL scheme is introduced, paying particular attention, as in previous Sections, to DIS structure functions and the impact that the inclusion of heavy quark effects has on them. The case of neutral current (NC) interactions will be considered first, and subsequently the case of charged ones will be discussed at NLO. Then, the LO and NNLO treatment of heavy quark mass effects is also discussed.

1.6.1 Neutral Current Structure Functions

The idea behind the FONLL general-mass scheme is simple and allows for a consistent combination of terms from different schemes. The name of this method comes from the original work [34], where a fixed order calculation (second order) is combined with a next-to-leading log one. But the method is general and a significant feature is that it can be used consistently to combine a fixed order with a resummed calculation to any order of either. The FONLL method only relies on standard QCD factorization and calculations with massive quarks in the CWZ [35] decoupling scheme in which the heavy quark is subtracted at zero momentum (it decouples for scales much below its mass) and massless quarks in the $\overline{\text{MS}}$ scheme in which the heavy flavor is treated as another massless parton

and so it is included in the all-order resummation of collinear logarithms up to a suitable chosen logarithmic order. The method was generalized in [36] to deep-inelastic structure functions.

In this specific case here studied, a Fixed-Flavor Number (FFN) scheme massive term up to order α_s is combined with a NLO Zero-Mass (ZM) computation. This scheme is called FONLL-A in [36] and it is equivalent to the S-ACOT scheme [37]. As shown in Ref. [38], this turns out to be identical to the S-ACOT [37] scheme. A near-threshold suppression is also adopted for the subleading terms as in [36, 39] through a damping factor. A benchmark of this method against the χ-scaling method is performed in [38].

The explicit expression for the $F_{2,h}$ heavy quark structure function [2] in the FONLL-A scheme is given by the sum of two terms:

$$F_{2,h}^{\rm FONLL}(x,Q^2) = F_{2,h}^{(n_l)}(x,Q^2) + \theta\left(Q^2-m_h^2\right)\left(1-\frac{m_h^2}{Q^2}\right)^2 F_{2,h}^{(d)}(x,Q^2)\,. \qquad (1.127)$$

The first contribution on the right-hand side of Eq. (1.127) is the massive-scheme heavy quark structure function at $\mathcal{O}\left(\alpha_s\right)$

$$F_{2,h}^{(n_l)}(x,Q^2) = x\int_x^1 \frac{dy}{y} C_{2,g}^{(n_l)}\left(\frac{x}{y},\frac{Q^2}{m_h^2},\alpha_s(Q^2)\right) g^{(n_l+1)}(y,Q^2)\,. \qquad (1.128)$$

The heavy quark gluon coefficient function is given by

$$C_{2,g}^{(n_l)}\left(z,\frac{Q^2}{m_h^2},\alpha_s(Q^2)\right) = \frac{\alpha_s(Q^2)}{2\pi} 2e_h^2 C_{2,g}^{(n_l),1}\left(z,\frac{Q^2}{m_h^2}\right) + \mathcal{O}\left(\alpha_s^2\right)\,. \qquad (1.129)$$

The $\mathcal{O}\left(\alpha_s\right)$ coefficient is

$$\begin{aligned} C_{2,g}^{(n_l),1}\left(z,\frac{Q^2}{m_h^2}\right) &= \theta\left(W^2-4m_h^2\right)\times T_R[(z^2+(1-z)^2+4\epsilon z(1-3z)-8\epsilon^2 z^2)\log\frac{1+v}{1-v} \\ &\quad +(8z(1-z)-1-4\epsilon z(1-z))v\,]\,, \end{aligned} \qquad (1.130)$$

where

$$\epsilon \equiv m_h^2/Q^2, \quad v \equiv \sqrt{1-4m_h^2/W^2}, \qquad (1.131)$$

and the partonic center of mass energy $W^2 = Q^2(1-z)/z$.

[2] See Ref. [36] for the discussion on the FONLL expressions for the longitudinal structure functions.

The second term on the right-hand side of Eq. (1.127) is the "difference" contribution

$$\begin{aligned} F_{2,h}^{(d)}(x,Q^2) &= x\int_x^1 \frac{dy}{y}\left[C_{2,q}^{(n_l+1)}\left(\frac{x}{y},\alpha_s(Q^2)\right)\left[h^{(n_l+1)}(y,Q^2)+\bar{h}^{(n_l+1)}(y,Q^2)\right]\right. \\ &+ \left.\left(C_{2,g}^{(n_l+1)}\left(\frac{x}{y},\alpha_s(Q^2)\right)-B_{g,h}^{(0)}\left(\frac{x}{y},\frac{Q^2}{m_h^2},\alpha_s(Q^2)\right)\right)g^{(n_l+1)}(y,Q^2)\right], \end{aligned} \tag{1.132}$$

where $h, \bar{h}$ are the heavy quark parton distributions. At first-order in α_s, $B_{g,h}^{(0)}$ is given by

$$B_{g,h}^{(0),1}\left(z,\frac{Q^2}{m_h^2}\right) = 2e_h^2 C_{2,g}^{(n_l,0),1}\left(z,\frac{Q^2}{m_h^2}\right), \tag{1.133}$$

and the massless limit of the massive coefficient function is

$$C_{2,g}^{(n_l,0),1}\left(z,\frac{Q^2}{m_h^2}\right) = T_R\left[(z^2+(1-z)^2)\log\frac{Q^2(1-z)}{m_h^2 z}+(8z(1-z)-1)\right], \tag{1.134}$$

which in the limit $Q^2 = m_h^2$ reproduces as required the usual massless scheme coefficient function.

It is possible to suppress the difference term Eq. (1.132) because it is of higher order near the threshold of the heavy quark and it is possible to check it easily as all terms in Eq. (1.127) PDFs and α_s are expressed in the same $n_f = 3$ decoupling scheme.

The expression in Eq. (1.127) interpolates smoothly between the massive scheme at small Q^2 and the massless scheme suitable at large Q^2. As an illustration of the differences between various schemes for the heavy quark structure functions, in Fig. 1.8 the $F_{2,c}$ and the $F_{L,c}$ charm structure functions are compared for various schemes: ZM, FONLL-A and the FFN scheme as a function of Q^2 for different values of x. It is clear that FONLL-A interpolates smoothly between the FFN scheme near threshold and the massless scheme at large Q^2 (also thanks to the use of a damping factor in Eq. (1.127)). For this comparison, PDFs and other settings, like the value of m_c, are identical to those of the Les Houches heavy quark benchmark comparison [38]. The comparison for the longitudinal structure function $F_{L,c}$ shows that mass effects are much larger than in $F_{2,c}$, so the ZM computation is completely unreliable.

Further quantifications of the impact of heavy quark mass effects in DIS structure functions are performed in [40]: for the phenomenologically more relevant case of F_2^p, heavy quark mass effects can be as large as $\sim 10\%$, decreasing fast for increasing x and Q^2. As in the case of Fig. 1.8, the Les Houches heavy quark benchmark settings have been used.

The $\mathcal{O}(\alpha_s)$ massive scheme heavy quark coefficient function, Eq. (1.130), was first

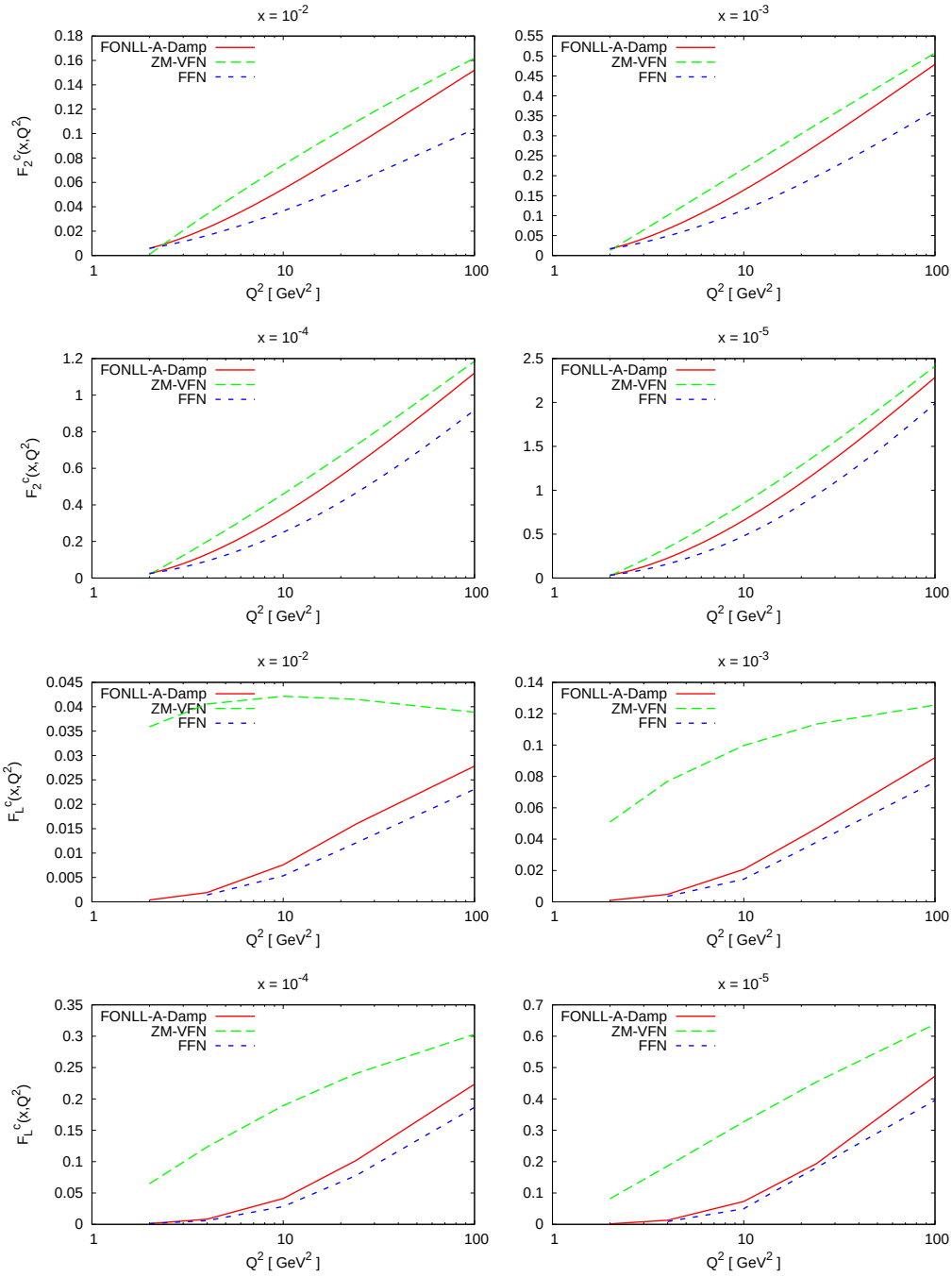

Figure 1.8: The charm structure functions $F_{2,c}(x, Q^2)$ and $F_{L,c}(x, Q^2)$ as a function of Q^2 for different values of x from $x = 10^{-5}$ to $x = 10^{-2}$ in various heavy quark schemes, computed using the FastKernel method: FONLL-A, ZM-VFN and the FFN scheme. The PDFs and settings are identical to those of the Les Houches heavy quark benchmark comparison.

computed in Refs. [41–43], while its Mellin transform is presented in Appendix A. The accuracy of the FONLL implementation has been assessed computing the Les Houches heavy quark benchmark tables [38], showing that the accuracy is sufficient for precision

PDF determination.

1.6.2 Charged Current Structure Functions

The case of charged current (CC) structure functions is treated according to the FONLL-A scheme. The coefficient functions needed are known up to $O(\alpha_s)$ and so other schemes that need higher order computations cannot be implemented.

This treatment is necessary especially to describe neutrino DIS charm production. In fact, in this case data lie in a region where the scale Q^2 is close to the charm mass threshold.

The analytic computation of the Mellin transforms of the $\mathcal{O}(\alpha_s)$ charged current heavy quark coefficient functions [44] is given into detail in Appendix A. For simplicity, the assumption that $|V_{\rm cs}| = 1$ is made, and the rest of the CKM matrix elements are set to zero. The generalization to realistic CKM elements is straightforward. A single heavy quark, the charm quark with mass m_c, is also assumed. The factorization scale is set to be equal to $\mu_F^2 = Q^2$. Finally, only neutrino induced charm production is considered, as the antineutrino case is again straightforward.

In the FFN massive scheme, the charged current charm production $F_{2,c}^{\rm CC}$ structure function for neutrino induced scattering has been computed in x-space in [44, 45] as

$$\begin{aligned} F_{2,c}^{(n_l),{\rm CC}}(x,Q^2) &= 2\xi s\left(\xi,Q^2\right) \\ &+ 2\xi\frac{\alpha_s\left(Q^2\right)}{2\pi}\Bigg\{\int_\xi^1 \frac{dz}{z}\Bigg[C_{2,h}^{(n_l),1}(z,Q^2,\lambda)s\left(\frac{\xi}{z},Q^2\right) \\ &+ C_{2,g}^{(n_l),1}(z,Q^2,\lambda)g\left(\frac{\xi}{z},Q^2\right)\Bigg]\Bigg\}\,, \end{aligned} \tag{1.135}$$

where

$$\xi = x\left(1+\frac{m_c^2}{Q^2}\right)\,, \qquad \lambda \equiv \frac{Q^2}{Q^2+m_c^2}\,. \tag{1.136}$$

In Eq. (1.135), $C_{2,g}^{(n_l),1}$ includes the contributions in which the gluon splits into a s and a $\bar{c}$ quark, both of which contribute to $F_{2,c}^{(n_l),{\rm CC}}$ at NLO. The Feynman diagrams for the LO and NLO gluon-induced subprocesses are shown in Figs. 1.9-1.10.

The x-space expressions for the $\mathcal{O}(\alpha_s)$ charged current coefficient functions in Eq. (1.135) are given in Refs. [44, 45]. The quark coefficient function can be separated into a delta function piece, a regular piece, and a singular piece regulated with the usual plus prescription,

$$C_{2,h}^{(n_l),1}(z,Q^2,\lambda) = C_{h,\delta}^{(n_l)}(\lambda)\,\delta(1-z) + C_{h,{\rm r}}^{(n_l)}(\lambda,z) + \left[C_{h,{\rm s}}^{(n_l)}\left(\lambda,z,Q^2\right)\right]_+\,. \tag{1.137}$$

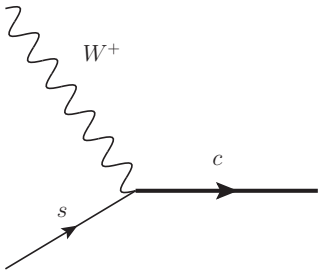

Figure 1.9: Feynman diagram for the LO contribution to $F_{2,c}^{(n_l),\mathrm{CC}}$ in the FFNS scheme. Thick solid lines indicate a heavy quark (charm) and thin solid lines a light quark (strange).

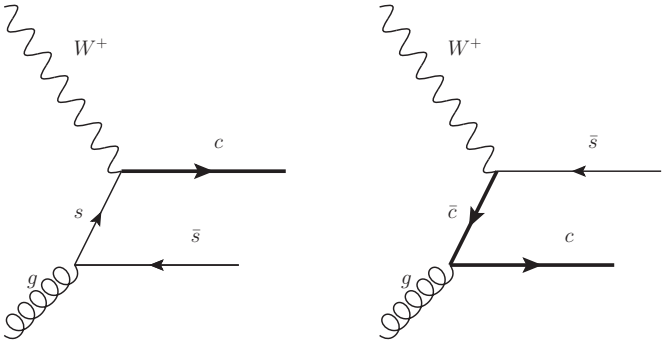

Figure 1.10: Feynman diagrams for the NLO ($\mathcal{O}(\alpha_s)$) gluon-induced contribution to $F_{2,c}^{(n_l),\mathrm{CC}}$ in the FFNS scheme.

It is possible to write the explicit expressions for the different pieces contained in Eq. (1.137). For the delta term

$$C_{h,\delta}^{(n_l)}(\lambda) = -C_F\left(4 + \frac{1}{2\lambda} + \frac{\pi^2}{3} + \frac{1+\lambda}{2\lambda}K_A\right) , \tag{1.138}$$

$$K_A = (1-\lambda)\ln(1-\lambda)/\lambda . \tag{1.139}$$

The regular piece can be written as

$$\begin{aligned} C_{h,\mathrm{r}}^{(n_l)}(\lambda,z) &= C_F\Bigg[-(1+z)\left(2\ln(1-z) - \ln(1-\lambda z)\right) - \frac{(1+z^2)\ln z}{1-z} \\ &+ \left(2z+2-\frac{2}{z}\right) + \left(\frac{2}{z} - 1 - z\right)\frac{1}{1-\lambda z}\Bigg] , \end{aligned} \tag{1.140}$$

and finally the singular piece reads

$$C_{h,\mathrm{s}}^{(n_l)}\left(\lambda,z,Q^2\right) = C_F\Bigg[-\frac{1+z^2}{1-z}\ln\frac{Q^2+m_c^2}{Q^2}+2\frac{2\ln(1-z)-\ln(1-\lambda z)}{1-z} - \frac{2}{1-z}+\frac{1}{2}\frac{1-z}{(1-\lambda z)^2}\Bigg], \tag{1.141}$$

where the first term is the contribution that depends on the factorization scale and is proportional to the qq splitting function. Separating the massive quark coefficient functions into the various contributions is important to properly evaluate their Mellin transforms, as will be discussed below.

Finally, the expression for the FFN gluon coefficient function is given. In this case there are no singular terms and it reads

$$\begin{aligned} C_{2,g}^{(n_l),1}(z,Q^2,\lambda) &= \Bigg[T_f\left(z^2+(1-z)^2\right)\left(\ln\frac{1-\lambda z}{(1-\lambda)z}+\ln\frac{Q^2+m_c^2}{Q^2}\right) \\ &+ T_f\left(z^2+(1-z)^2\right)\left(2\ln(1-z)-\ln(1-\lambda z)-\ln z\right) \\ &+ \left(8-18(1-\lambda)+12(1+\lambda)^2\right)z(1-z)+\left(\frac{1-\lambda}{1-\lambda z}-1\right) \\ &+ (1-\lambda)\,z\ln\frac{1-\lambda z}{(1-\lambda)z}\left(6\lambda-12\lambda^2 z\right)\Bigg]. \end{aligned} \tag{1.142}$$

Again, the last term in the first line is the scale-dependent contribution and is proportional to $P_{qg}^{(0)}$. Note that both diagrams shown in Fig. 1.10 contribute [44]. Analogous expressions for the charged current $F_{3,c}$ and $F_{L,c}$ structure functions can be found in Refs. [44, 45].

As in the case of neutral currents, the massless limit of the FFN structure functions is easily obtained. For the massive $F_{2,c}^{\mathrm{CC}}$ structure function it has the structure

$$\begin{aligned} F_{2,c}^{(n_l,0),\mathrm{CC}}(x,Q^2) &= 2xs\left(x,Q^2\right) \\ &+ 2x\frac{\alpha_s\left(Q^2\right)}{2\pi}\Bigg\{\int_x^1\frac{dz}{z}\Bigg[C_{2,h}^{(n_l,0),1}(z,Q^2,\lambda)s\left(\frac{x}{z},Q^2\right) \\ &+ C_{2,g}^{(n_l,0),1}(z,Q^2,\lambda)g\left(\frac{x}{z},Q^2\right)\Bigg]\Bigg\}, \end{aligned} \tag{1.143}$$

where

$$C_{2,h}^{(n_l,0),1}(z,Q^2,\lambda)=C_{h,\delta}^{(n_l,0)}\delta(1-z)+C_{h,\mathrm{r}}^{(n_l,0)}(z)+\left[C_{h,\mathrm{s}}^{(n_l,0)}(z)\right]_+, \tag{1.144}$$

with

$$C_{h,\delta}^{(n_l,0)} = -C_F \left(\frac{9}{2} + \frac{\pi^2}{3} \right) , \tag{1.145}$$

$$C_{h,\mathrm{r}}^{(n_l,0)}(z) = C_F \left[-(1+z)\ln(1-z) - \frac{(1+z^2)\ln z}{1-z} + 3 + 2z \right] , \tag{1.146}$$

$$C_{h,\mathrm{s}}^{(n_l,0)}(z) = C_F \left[2\left(\frac{\ln(1-z)}{1-z} \right) - \frac{3}{2}\left(\frac{1}{1-z} \right) \right] , \tag{1.147}$$

and for the gluon

$$\begin{aligned} C_{2,g}^{(n_l,0),1}(z,Q^2) &= 2T_f \left[\left(z^2 + (1-z)^2\right) \ln\frac{1-z}{z} + 8z(1-z) - 1 \right] \\ &+ T_f \left(z^2 + (1-z)^2\right) \ln\frac{Q^2}{m_c^2} . \end{aligned} \tag{1.148}$$

For completeness, the ZM Variable Flavor Number (ZM-VFN) scheme quark coefficient functions for quarks and gluons is also provided,

$$\begin{aligned} C_{2,h}^{(n_l+1),1}(z) = C_F \Bigg[& 2\left(\frac{\ln(1-z)}{1-z} \right)_+ - \frac{3}{2}\left(\frac{1}{1-z} \right)_+ - (1+z)\ln(1-z) \\ & - \frac{(1+z^2)\ln z}{1-z} + 3 + 2z + \delta(1-z)\left(-\frac{\pi^2}{3} - \frac{9}{2} \right) \Bigg] , \end{aligned} \tag{1.149}$$

$$C_{2,g}^{(n_l+1),1}(z) = T_F \left[\left(z^2 + (1-z)^2\right) \ln\frac{1-z}{z} + (8z(1-z) - 1) \right] . \tag{1.150}$$

Note that the above gluon coefficient function, Eq. (1.150), is defined according to the notation of [46], that is, it corresponds to the production of a single quark or antiquark.

Comparing the FFNS0 and ZM-VFN coefficient functions it follows that for the gluon piece holds the relation

$$C_{2,g}^{(n_l,0),1}(z,Q^2) = 2C_{2,g}^{(n_l+1),1}(z) + T_f \left(z^2 + (1-z)^2\right) \ln\frac{Q^2}{m_c^2} , \tag{1.151}$$

where the overall factor 2 is due to the fact that the ZM coefficient function, Eq. (1.150), has been defined for a single quark, while in Eq. (1.142) the gluon coefficient function accounts for the production of two quarks (s and $\bar{c}$). Note also the presence of the usual collinear logarithm. For the quark piece

$$C_{2,h}^{(n_l,0),1}(z) = C_{2,h}^{(n_l+1),1}(z) , \tag{1.152}$$

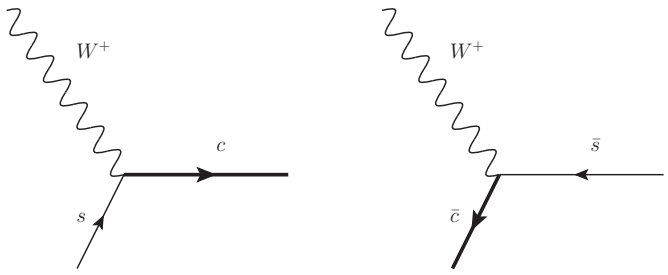

Figure 1.11: Feynman diagrams that contribute to $F_{2,c}^{(n_l+1),\mathrm{CC}}$ in the ZM-VFN scheme at leading order. The NLO diagrams are the same as in the FFNS scheme.

without any collinear logarithm.

The definition of the heavy CC structure function in the ZM scheme is not unique: here it is defined as the contribution to the structure function which includes all contributions to the inclusive structure function which survive when all CKM elements but $|V_{\mathrm{cs}}|$ are set to zero. With this definition, both the leading-order processes $cW^+ \to s$ and $\bar{c}W^+ \to \bar{s}$ contribute to it (see Fig. 1.11). This definition coincides with the experimental one because the struck charm antiquark must be accompanied by an (observed) charm quark, and it is free of mass singularities. The gluon initiated NLO contributions remain those shown in Fig. 1.10. The structure function in the massless scheme above charm threshold is then given by

$$\begin{aligned} F_{2,c}^{(n_l+1),\mathrm{CC}}(x,Q^2) &= 2x\left(s\left(x,Q^2\right)+\bar{c}\left(x,Q^2\right)\right) \\ &+ 2x\frac{\alpha_s\left(Q^2\right)}{2\pi}\left\{\int_x^1 \frac{dz}{z}\left[C_{2,h}^{(n_l+1),1}(z,Q^2,\lambda)\left(s\left(\frac{x}{z},Q^2\right)+\bar{c}\left(\frac{x}{z},Q^2\right)\right)\right.\right. \\ &+ \left.\left. 2C_{2,g}^{(n_l+1),1}(z,Q^2,\lambda)g\left(\frac{\xi}{z},Q^2\right)\right]\right\}. \end{aligned} \tag{1.153}$$

The ZM-VFN massless coefficient functions have been defined in Eqs. (1.149)-(1.150). Note the factor two in front of the gluon coefficient function, to account for the production of two quarks in the two NLO subprocesses of Fig. 1.10.

Finally, the various schemes can be combined to construct the FONLL-A structure functions. As in the NC case, the FONLL structure function is defined as

$$F_{2,c}^{(\mathrm{FONLL}),\mathrm{CC}}(x,Q^2) \equiv F_{2,c}^{(n_l),\mathrm{CC}}(x,Q^2) + \theta\left(Q^2-m_c^2\right)\left(1-\frac{m_c^2}{Q^2}\right)^2 F_{2,c}^{(\mathrm{d}),\mathrm{CC}}(x,Q^2) \tag{1.154}$$

$$F_{2,c}^{(\mathrm{d}),\mathrm{CC}}(x,Q^2) = F_{2,c}^{(n_l+1),\mathrm{CC}}(x,Q^2) - F_{2,c}^{(n_l,0),\mathrm{CC}}(x,Q^2)\,, \tag{1.155}$$

where the damping factor is used as default threshold prescription.

Using the explicit expressions derived in the previous Section for the difference between the ZM and FFNS0 coefficient functions, Eqs. (1.152) and (1.151), the difference term can be written as

$$F_{2,c}^{(\mathrm{d}),\mathrm{CC}} = 2x\bar{c}\left(x,Q^2\right) - 2x\frac{\alpha_s}{2\pi}\ln\frac{Q^2}{m_c^2}\int_x^1\frac{dz}{z}T_f\left(z^2+(1-z)^2\right)g\left(\frac{x}{z},Q^2\right)+\mathcal{O}(\alpha_s^2)\,, \tag{1.156}$$

where the fact that the heavy quark distribution is $\mathcal{O}(\alpha_s)$ has been used. Now, it is easy to see explicitly that, in the region where $L \equiv \ln Q^2/m_c^2$ is not large, the difference term is of order $\mathcal{O}(\alpha_s^2)$: to first order in α_s the FONLL expression coincides with the massive-scheme one also for charged current scattering. The use of the leading-order QCD evolution equations immediately leads to

$$c(x,Q^2) = \bar{c}(x,Q^2) = \frac{\alpha_s(Q^2)}{2\pi}\ln\frac{Q^2}{m_c^2}\int_x^1\frac{dz}{z}T_f(z^2+(1-z)^2)g\left(\frac{x}{z},Q^2\right)+\mathcal{O}(\alpha_s^2)\,. \tag{1.157}$$

Inserting this expansion in Eq. (1.156), it is trivial to check the explicit cancellation of the $\mathcal{O}(\alpha_s)$ terms, that is, near the heavy quark threshold the difference term is of order $F_{2,c}^{(\mathrm{d}),\mathrm{CC}} = \mathcal{O}(\alpha_s^2)$.

The final FONLL-A expression for the charged current charm production structure function $F_{2,c}^{\mathrm{CC}}$ is given by

$$\begin{aligned}
F_{2,c}^{(\mathrm{FONLL}),\mathrm{CC}}(x,Q^2) &= 2\xi s\left(\xi,Q^2\right)+\theta\left(Q^2-m_c^2\right)\left(1-\frac{m_c^2}{Q^2}\right)^2 2x\bar{c}\left(x,Q^2\right)\\
&+ 2\xi\frac{\alpha_s\left(Q^2\right)}{2\pi}\Bigg\{\int_\xi^1\frac{dz}{z}\Bigg[C_{2,h}^{(n_l),1}(z,Q^2,\lambda)\left(s\left(\frac{\xi}{z},Q^2\right)+\theta\left(Q^2-m_c^2\right)\left(1-\frac{m_c^2}{Q^2}\right)^2 2x\bar{c}\left(\frac{\xi}{z},Q^2\right)\right)\\
&\qquad +C_{2,g}^{(n_l),1}(z,Q^2,\lambda)g\left(\frac{\xi}{z},Q^2\right)\Bigg]\Bigg\}\\
&- \theta\left(Q^2-m_c^2\right)\left(1-\frac{m_c^2}{Q^2}\right)^2 2x\frac{\alpha_s\left(Q^2\right)}{2\pi}\int_x^1\frac{dz}{z}T_f(z^2+(1-z)^2)g\left(\frac{x}{z},Q^2\right)\,.
\end{aligned} \tag{1.158}$$

It can be easily verified that Eq. (1.158) reduces to the FFN scheme Eq. (1.135) at the heavy quark threshold $Q^2 = m_c^2$, and to the ZM-VFN expression Eq. (1.153) in the asymptotic region $Q^2 \gg m_c^2$.

The above derivation generalizes straightforwardly to the other relevant charged current structure functions $xF_{3,c}^{\mathrm{CC}}$ and $F_{L,c}^{\mathrm{CC}}$, as well as to the case with a general CKM quark mixing matrix. Note that in all the results shown below the standard CKM mixing has been assumed, with the CKM matrix elements set to their PDG values [47].

Now that the FONLL-A General Mass (GM) scheme has been defined for charged

current structure functions, it is possible to compare the various schemes (ZM, FFNS, FONLL-A) in the kinematic region that is most relevant in the global PDF analysis, namely the region covered by the NuTeV dimuon measurements [48]. In Fig. 1.12 the results of such a comparison between various schemes for charm production in neutrino-induced charged current scattering are shown. Results are compared at the level of the phenomenologically relevant charm production reduced cross-section, defined as [49]

$$\begin{aligned}\tilde{\sigma}^{\nu(\bar{\nu}),c}(x,y,Q^2) &\equiv \frac{1}{E_\nu}\frac{d^2\sigma^{\nu(\bar{\nu}),c}}{dx\,dy}(x,y,Q^2) \\ &= \frac{G_F^2 M_N}{2\pi(1+Q^2/M_W^2)^2}\left[\left(\left(Y_+ - \frac{2M_N^2x^2y^2}{Q^2} - y^2\right) + y^2\right)F_{2,c}^{\nu(\bar{\nu})}(x,Q^2)\right. \\ &\qquad \left. - y^2 F_{L,c}^{\nu(\bar{\nu})}(x,Q^2) \pm Y_-\, xF_{3,c}^{\nu(\bar{\nu})}(x,Q^2)\right] \end{aligned} \tag{1.159}$$

with $Q^2 = 2M_N E_\nu xy$ and $Y_\pm = 1 \pm (1-y)^2$. This expression for the NuTeV dimuon cross-section has been corrected with respect to Eq. (33) of Ref. [40]. A spurious factor of $(1 + m_c^2/Q^2)$ was present there, and it was pointed out by S. Alekhin. The impact of this correction was studied in Ref. [50] and found to be not relevant. In Fig. 1.12 the various schemes are compared in some representative bins of the NuTeV dimuon kinematics [48]. Parton distribution functions and other settings are those of the Les Houches heavy quark benchmark comparison [38]. It is observed that in the kinematic region of neutrino data (both inclusive CHORUS data and dimuon NuTeV data), the FONLL-A result is very close to the FFN scheme computation, and it only begins to differ from it at the highest energies, where resummation of charm mass collinear logarithms begins to become relevant.

Even if the differences between the FFN and FONLL-A schemes for charged current scattering in the NuTeV kinematic region are moderate, as shown in Fig. 1.12, they become rather more important at small-x and medium-large Q^2, where the charm and gluon PDFs become larger. To illustrate this, in Fig. 1.13 the charged current charm structure function $F_{2,c}^{\rm CC}$ as a function of Q^2 is compared for two different values of x. Notice in particular that at very small-x the FONLL-A expression is essentially the massless result. However, producing dimuons at $x \sim 10^{-3}$ and $Q^2 \sim 10$ GeV^2, where differences are larger, requires a fixed target neutrino experiment with a neutrino beam with energy in the multi-TeV range, which is not foreseen in the near future. Therefore one can conclude that any reasonable general-mass scheme for charged current scattering will be very close to the FFNS in the region of experimental data.

The FONLL-A calculation of charged current structure functions has been implemented in a x-space code, FONLLdisCC, that will be used for benchmarking purposes.

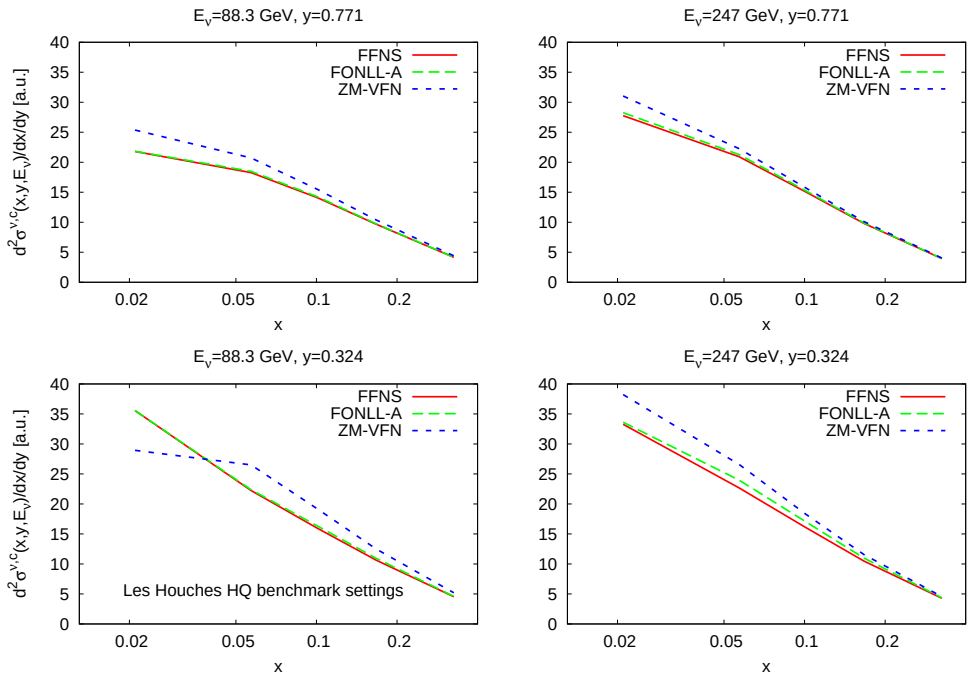

Figure 1.12: Comparison of different schemes for charm production in neutrino-induced DIS. The kinematic range is representative of the NuTeV dimuon data range. The ZM-VFN, FFN and FONLL-A schemes are compared at the level of the neutrino induced charm production cross-section, Eq. (1.159). The settings are the same as those of the Les Houches heavy quark benchmark comparison [38].

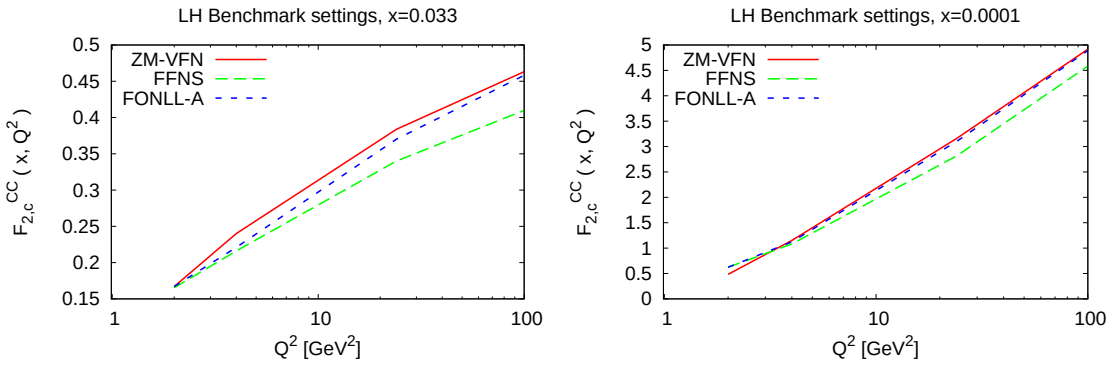

Figure 1.13: Comparison of different schemes for charm production in neutrino-induced DIS. The $F_{2,c}^{\rm CC}$ structure function in the massless, massive and FONLL-A schemes is shown; in this case the FONLL-A expression is given by Eq. (1.158). The settings are the same as those of the Les Houches heavy quark benchmark comparison [38].

This is the analogue of the FONLLdis code for neutral currents [51], however is rather simpler since the unknown $\mathcal{O}\left(\alpha_s^2\right)$ massive coefficient functions do not have to be implemented. The NNPDF implementation of the FFNS calculations has been benchmarked with the corresponding results of the MSTW08 code [52], finding perfect agreement.

The FONLL-A and MSTW08 general-mass schemes for charged currents have been also compared, finding qualitative agreement but some quantitative difference.

In Ref. [53], where a LO and a NNLO PDFs sets are presented, the heavy quark mass effects are included using the so-called FONLL-C GM-VFN scheme of Ref. [36]. The main idea is the same of FONLL-A but combining NNLO massless perturbative evolution with the $\mathcal{O}(\alpha_s^2)$ massive coefficient functions. The FONLL-C scheme interpolates smoothly between the $O\left(\alpha_s^2\right)$ massive scheme (FFN) near the heavy quark threshold, and the $O\left(\alpha_s^2\right)$ massless scheme (ZM-VFN) at large Q^2. Mass effects are much larger for the longitudinal structure function $F_{L,c}$ than for $F_{2,c}$, as shown in Ref. [53], so there the ZM-VFN computation is completely unreliable. A benchmark against different GM-VFN schemes was performed in Ref. [38] for DIS structure functions, with common input toy PDFs and common choices of all other settings. This treatment is applied to neutral current structure functions, while in the case of charged current DIS a full implementation of this scheme cannot be performed because only the asymptotic $Q^2 \to \infty$ limit is known [54] for the massive $O(\alpha_s^2)$ heavy quark coefficient functions. For this reason in the charged sector of FONLL-C GM-VFN scheme the $O(\alpha_s^2)$ massive contribution is set to zero and the other components (PDFs, ZM structure functions, and α_s) are evaluated at NNLO. The computation of the neutral current massive coefficient functions in the Mellin space at $O(\alpha_s^2)$ is given in Appendix B, while benchmarking of the numerical accuracy of the FONLL-C implementation is discussed in Appendix C. In that Appendix, also a comparison with the FONLL-B scheme (that combines $O(\alpha_s^2)$ massive contribution with a NLO massless perturbative evolution) is performed.

The case of LO QCD is much simpler as here both neutral and charged current massive coefficient functions for DIS vanish and the only difference with a standard ZM scheme is the presence of the damping factor. The only massive contribution present in this case is for a heavy quark produced from a struck light quark. Then the FONLL expression is reduced to the parton-model ($\mathcal{O}(\alpha_s^0)$) massive coefficient function.

Chapter 2

Determination of Parton Distribution Functions

The introduction of parton distribution functions was needed since the first steps in the formulation of the parton model and modern QCD, as seen in the previous Chapter. The inclusion of a PDF set in a specific analysis is necessary for the study and prediction of all that observables based on some hadronic initial state as also to have reliable results from the Monte Carlo event generators commonly used for experimental simulations.

Looking at the last forty years before the Large Hadron Collider (LHC) of Geneva was operative, it is possible to distinguish three main "seasons" for parton fitting: the first one was intended to demonstrate the compatibility of the partonic interpretation of first QCD processes with experimental data [55–57], the second was dedicated to the extraction of parton sets from the first hadron colliders data available, allowing for qualitative QCD and slowly progressing towards the third stage, which mainly thanks to DIS NMC [58,59] and HERA [60] data has permitted QCD precision physics. Nowadays we are facing a new era, strictly related to new physics [61, 62] at the LHC.

In this Chapter the main elements needed for a PDF set determination are sketched. After introducing the general strategy for extracting a parton set from data, the error determination technique is discussed focusing in particular on the treatment of multiplicative uncertainties as for example normalization uncertainties. Finally, the main fitting strategy is analyzed.

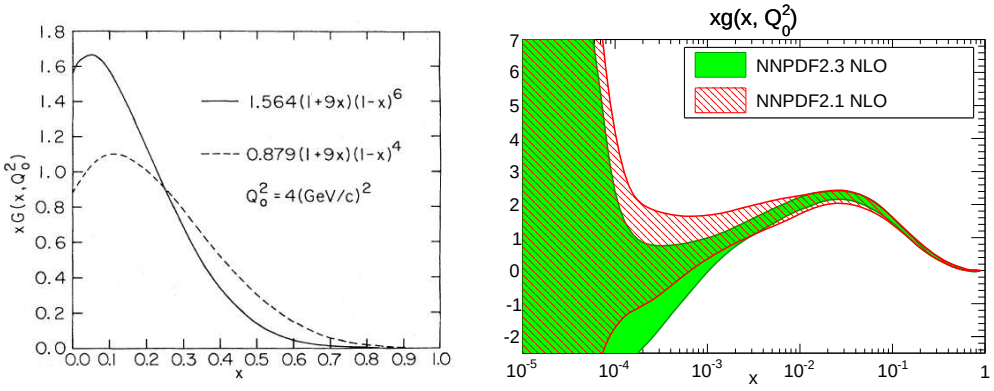

Figure 2.1: A gluon PDF estimate of 1984 [63](left) and a comparison between gluon PDFs from NNPDF2.3 [50] and NNPDF2.1 [40] analyses (right).

2.1 General Strategy

Parton distribution functions, as already stated, are non-perturbative, universal quantities, because of their general independence on the specific process. In order to release a PDF set, experimental data involving hadronic initial states are needed, as these quantities only may be extracted from data. Thanks to their universality, it is possible to use data from different processes to better determine PDF's behaviour. Only in recent years these quantities started to be released together with an estimation of the experimental error they may be affected by. Before then, the determination of the sole central value was enough for the level of precision of the first QCD analyses (Fig. 2.1). In fact the impact of PDF's errors was widely considered to be negligible in comparison to the uncertainty due to contributions from other theoretical and experimental sources.

The determination of a set of parton distribution functions is a rather delicate exercise. Several theoretical, experimental, and statistical issues need to be addressed in order to provide reasonable results. From a theoretical point of view, precision physics of modern colliders as Tevatron and LHC requires the implementation of hard scattering matrix elements and evolution kernels for DGLAP evolution equations computed at the appropriate perturbative order. Heavy quark mass effects has to be considered and also the choice of the parametrization used in the fit may affect the final result. On the experimental side, even if a wide number of different experiments are included in a global analysis, some distributions may be poorly constrained because of lack of experimental information and also cuts must be performed on some kinematic regions. In fact, where theory loses reliability it is not possible to compare predictions with data. Moreover, a proper use of the uncertainties and correlations delivered by experimental groups together with the development and application of refined statistical tools are necessary.

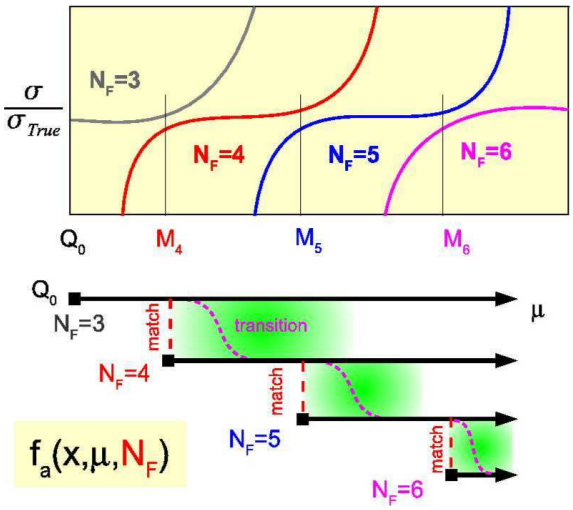

Figure 2.2: The upper figure schematically represents how each calculation with a set number of flavors N_F has a region of applicability. The transition from the $N_F - 1$ scheme to the N_F scheme should be in the vicinity of the m_{N_F} mass, but need not occur exactly at $\mu = m_{N_F}$. The lower figure illustrates that multiple PDFs can coexist for $\mu \geq m_{N_F}$ with matching performed at $\mu = m_{N_F}$. Image taken from Ref. [64].

The general strategy for extracting PDFs from data it is almost the same for all collaborations involved in parton fitting: thanks to factorization theorem, the observables are computed theoretically at a certain initial scale Q_0^2 as the convolution between coefficient functions and PDFs. The coefficient functions are perturbative objects computed directly from partonic Feynman diagrams. For all relevant processes they are available at LO, NLO, and in some cases also at NNLO perturbative orders. On the contrary, PDFs are parametrized and randomly initialized. Each observable has its own definition in terms of a linear combination of partonic distributions and for this reason a specific observable can constrain or disentangle some distributions and not all of them. Several different observables are needed to determine all the $2n_f + 1$ independent parton components inside the nucleon. This lack of information forced for a long time to make general assumptions on the unconstrained PDFs.

As each data point is measured at a physical scale Q^2, to compute the error function that then will be minimized in the fit it is necessary to evolve the observables to that scale solving the DGLAP evolution equations. Again, the solution of these equations can be performed with different methods and at different perturbative orders of the evolutive kernels. Also, the error function, that contains the covariance matrices, the measured observables and the same observables computed as explained above, is minimized accordingly to different techniques (mainly log-likelihood minimization but also genetic algorithms).

On top of the general fitting strategy described above, other issues has to be considered. One of these is related to nuclear corrections: even if it can be assumed that the

scattering with individual partons inside the nucleus is incoherent, corrections may be needed for different targets in some experiments. These corrections are model-dependent and may affect theoretical calculations. Another issue is the implementation of heavy quark mass effects. Different schemes may be used to take them into account. In the following I briefly describe the main characteristics of each scheme.

The theoretical framework proposed by Collins-Wilczek and Zee [35, 65] is the basis for the schemes considered here. Due to kinematic and dynamical issues it translates in different schemes. The simplest one is the fixed flavor number scheme: the heavy quarks are considered in the massive coefficient functions with a fixed number of light flavors n_l. All the flavors below the heavy quark mass are treated as massless. This scheme becomes increasingly unreliable for scales larger than the heavy quark mass threshold and so the most reliable results are: for $n_l = 3$ when $Q \sim m_c$, for $n_l = 4$ when $m_c \leq Q \leq m_b$, for $n_l = 5$ when $m_b \leq Q \leq m_t$, and for $n_l = 5$ when $Q \geq m_t$. On this basis it is ideally straight-forward to define a variable flavor number scheme by matching the previous schemes in their respective regions of validity, but as this matching can be performed in slightly different ways there are several possible equivalent schemes deriving from this one. Since a few years ago the most commonly implemented VFNS was the ZM-VFNS. All quarks are massless and the heavy quarks are radiatively generated only above the m_H threshold. The number of light flavors n_l changes going through thresholds, changing as a consequence the anomalous dimensions and the QCD β function. The ZM-VFNS lose accuracy near threshold because it neglects terms proportional to powers of m_H^2/Q^2 and also the approximate treatment of phase space introduces inconsistencies. A solution is proposed in Ref. [39] with the Improved-ZM-VFNS (I-ZM-VFNS) by defining a new lower integration limit in PDF convolution with hard scattering matrix elements as:

$$\chi(x, Q^2) = x(1 + M_f^2/Q^2), \tag{2.1}$$

where M_f^2 is the total mass of the final state.

Schemes as the ACOT, the Thorne Robert (TR), and the FONLL (already introduced in the previous Chapter) are also called general mass VFNS (GM-VFNS) as they all exploit the accuracy of the FFNS near threshold and of the ZM-VFNS at large Q^2. They combine the two schemes going smoothly from a FFNS treatment to a ZM-VFNS one as the scale Q^2 is increased. The ACOT scheme [66] introduces the subtraction term to remove double counting due to LO and NLO contributions overlap, giving an inclusion of mass effects at all energy scales. On this are based variants as the S-ACOT [37] and the ACOT-χ [67, 68] schemes. The TR scheme [69] is similar to ACOT (a comparison is performed between the two schemes in Ref. [64]) with more emphasis on matching conditions treatment. A generalization is represented by TR' [70]. For details on the FONLL

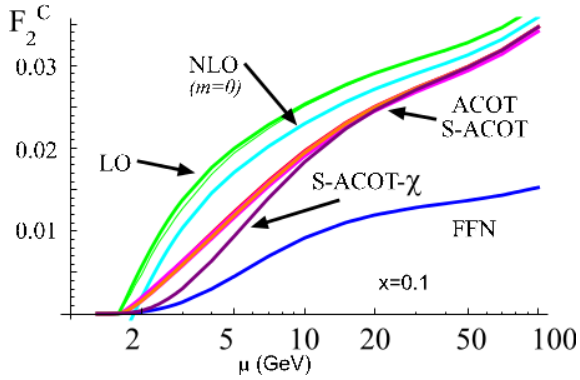

Figure 2.3: Calculation of DIS heavy quark production for a variety of schemes. Image taken from Ref. [64].

scheme see Section 1.6.

Among the collaborations that actively produce global partonic sets of distributions it is worth citing the CTEQ-TEA [71–76], MRST-MSTW [77–82], NNPDF [83–88] [40], and Alekhin-ABKM [89–92] groups. The CTEQ collaboration used a ZM-VFNS for several analyses and the ACOT scheme [66] for specific studies with heavy quarks [93] and also in [73]. The MRST global analysis used the Thorne Roberts (TR) scheme [69] up to MRST 2004 [79], then the TR' scheme [70] began to be used in the MRST 2006 release [80] and in all the MSTW analyses. The ABKM collaboration uses a FFNS with $n_f = 3, 4, 5$ [94]. Finally, the NNPDF collaboration uses the FONLL GM-VFN scheme [34, 36] already introduced in the previous Chapter since the NNPDF2.1 release [40].

2.2 Statistical and Systematic Errors

Monte Carlo Sampling

To determine PDFs error from experimental information contained in covariance matrices nontrivial. The attempt here is to extract from a finite number of data points the necessary information to describe the error on a set of functions, that are infinite dimensional objects. A way to solve this difficulty was proposed in Ref. [95] and developed in Ref. [96]. It is necessary to build a probability density in the space of functions, in order to give an estimation of a generic observable $\mathcal{F}$ that depends on a PDF set $\{f\}$ as

$$\left\langle \mathcal{F}_{\{f\}} \right\rangle = \int \mathcal{F}_{\{f\}} \mathcal{P}[\{f\}]\mathcal{D}[f] \tag{2.2}$$

where $\mathcal{P}[\{f\}]\mathcal{D}[f]$ is the density probability measure in PDFs space. Of course, the set $\{f\}$ depends on a set of parameters $\mathbf{a} \equiv (a_1, a_2, \ldots, a_n)$ with n the number of parameters

that determine $\{f\}$ given the fixed functional form used in the fit. The main idea is to reduce Eq. (2.2) to a sum over a number of replicas generated by Monte Carlo (MC) sampling of the parameter space obtaining a random set $\{\mathbf{a}\}^{(k)}$ of sets of parameters. To this number N_{rep} of random parameter sets corresponds an equal number of parton sets $\{f\}^{(k)}$, with $k = 1, 2, \ldots, N_{rep}$ and it is possible to write

$$\left\langle \mathcal{F}_{\{f\}} \right\rangle \simeq \frac{1}{N_{rep}} \sum_{i=1}^{N_{rep}} \mathcal{F}_{\{f\}^{(i)}}. \tag{2.3}$$

A problem with this approach is that in several directions in parameter space there are flat regions, thus a problem when generating MC replicas rises: shallow directions are related to almost unconstrained parameters, and this results in very large values in the samples of replicas. The NNPDF collaboration also uses the MC sampling technique, but at the level of experimental data, avoiding this problem (Sect. 3.2).

Hessian Approach

The Hessian formalism [77, 97], among the PDF-fitter collaborations, is the most commonly used method for PDF error determination. The χ^2 function is quadratically expanded about its global minimum[1]. It can be written as

$$\Delta\chi^2 = \chi^2 - \chi^2_0 = \sum_{i=1}^{d} \sum_{j=1}^{d} H_{ij}(a_i - a_i^0)(a_j - a_j^0) \tag{2.4}$$

with $\chi^2_0 = \chi^2(S_0)$, $\{\mathbf{a}^0\}$ respectively the χ^2 and the set of parameters corresponding to the best estimate S_0 for the PDFs set $\{f\}$. The value H_{ij} is the Hessian matrix element defined as:

$$H_{ij} = \frac{\partial^2 \chi^2(\mathbf{a})}{\partial a_i \partial a_j}. \tag{2.5}$$

Moving the parameters around their best value, a shift is observed in the χ^2 function, $\Delta\chi^2$, and can be defined

$$\Delta\chi^2 \leq T^2 \tag{2.6}$$

as the region of "acceptable fits", with T the tolerance parameter: all uncertainties are proportional to this parameter. Going back to the Hessian introduced in Eq. (2.5), this matrix

[1] To avoid the quadratic approximation the Lagrange multiplier method can be applied [98, 99].

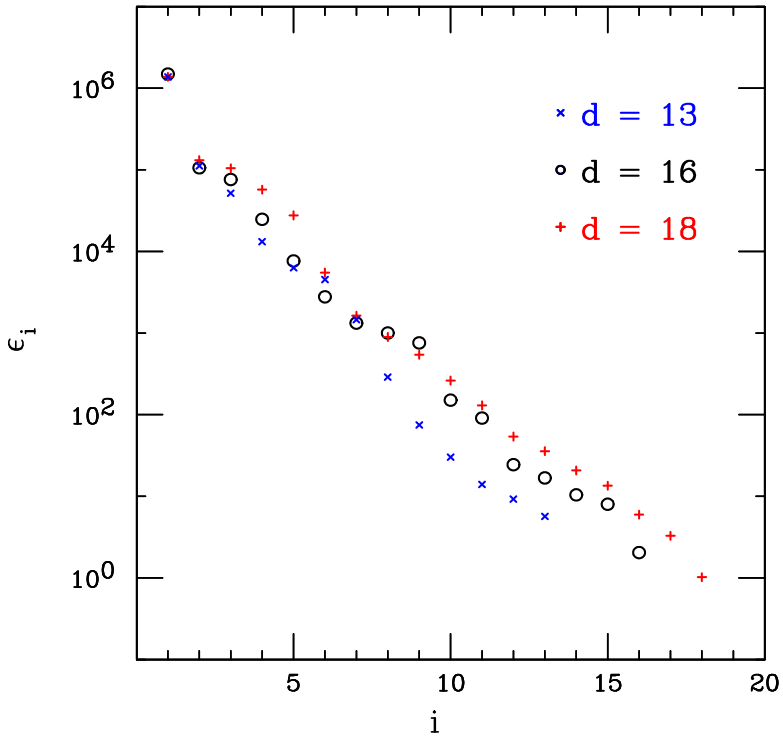

Figure 2.4: Distribution of eigenvalues defined in Eq. (2.7) of the Hessian matrix, taken from Ref. [97].

has a complete set of N_{par} orthonormal eigenvectors v_{ik} with eigenvalues ε_k defined by

$$\sum_{j=1}^{d} H_{ij} v_{jk} = \varepsilon_k v_{ik} \tag{2.7}$$

$$\sum_{i=1}^{d} v_{il} v_{jk} = \delta_{lk}. \tag{2.8}$$

Each eigenvector determines a direction in parameter space along which the χ^2 variation can be quick or slow, and as a direction can be far quicker than another a consequence is that the eigenvalues ε_k are distributed over a wide range that covers many orders of magnitude (Fig. 2.4).

The eigenvectors can be related directly to the variation into the space of parameters by

$$a_i - a_i^0 = \sum_{k=1}^{d} v_{ik} s_k z_k. \tag{2.9}$$

The normalization of z_k such that $\Delta\chi^2 = \sum_{k=1}^{d} z_k^2$ is obtained through the introduction of s_k factors. Thanks to the transformation of Eq. (2.9), graphically illustrated in Fig. 2.5, it is possible to notice that the region of acceptable fits around the global minimum is con-

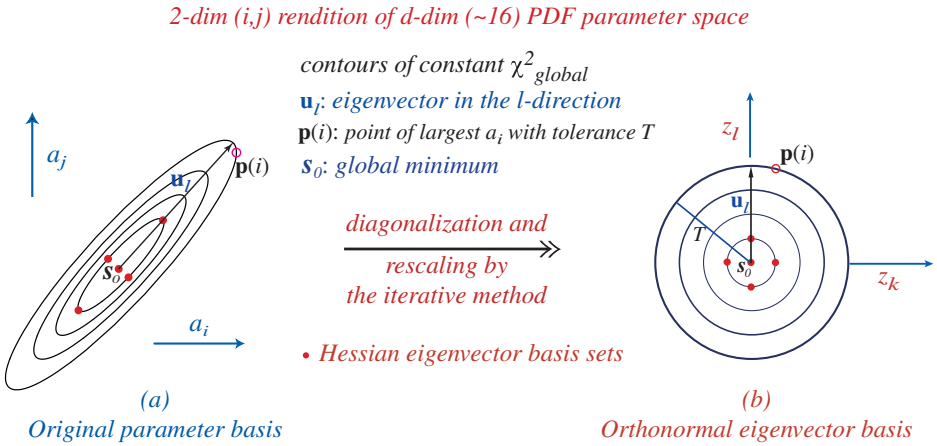

Figure 2.5: Representation of the transformation of Eq. (2.9) from the original parameter basis to the orthonormal eigenvector basis, taken from Ref. [97]. To each point in the graph corresponds a set of N_{par} parameters associated to a PDFs set.

tained inside a hypersphere of radius $\sqrt{\Delta\chi^2}$ and in case of an ideal quadratic behaviour of the χ^2 function the scale factors ε_k would be equal to $\sqrt{1/\varepsilon_k}$.

Coming back to the generic observable $\mathcal{F}$, the error on it and on the PDFs on which it depends can be estimated using the formula

$$\Delta\mathcal{F} = \sum_i (\mathcal{F}(S_i^+) - \mathcal{F}(S_i^-))^2 \tag{2.10}$$

where $S_i^\pm$ are the $2N_{rep}$ sets of PDFs computed at the two points defined by

$$z_i^\pm = \pm\frac{T}{2} \tag{2.11}$$

on the edge of the N_{rep}-dimensional hypersphere in the z parameter space. Together with S_0 they form a set of $2N_{rep}+1$ sets of PDFs, that are the ones needed to compute PDFs errors on S_0, besides observables error from Eq. (2.10).

However, the tolerance criteria for $T^2 > 1$ lose a statistically rigorous interpretation and together with the assumption of the validity for linearized approximation in error propagation introduce weaknesses in the error treatment through the Hessian method.

2.3 Normalization Errors

In this Section the normalization uncertainties introduced in Eq. (2.43) are discussed. After a first introductory part which briefly explains the Hessian and Monte Carlo methods, the discussion moves towards the issues at the origin of the biases [100–102] associated

with overall multiplicative uncertainties as normalization errors in global fits. Finally two methods for the definition of the covariance matrix are presented: the *penalty trick* and the t_0 method.

Hessian and Monte Carlo Methods

Given a number n of experimental measurements m_i, one may want to extract from them the expectation value and the variance of a theoretical quantity t. Information about experimental uncertainties is contained into a covariance matrix $(\mathrm{cov})_{ij}$ that takes the common form

$$(\mathrm{cov})_{ij} = \delta_{ij}\sigma_i^2 + \sum_{k=1}^{n} \bar{\sigma}_{ik}\bar{\sigma}_{kj}. \tag{2.12}$$

The uncertainties are supposed to be Gaussian. The values σ_i and $\bar{\sigma}_{ik}$ are referred respectively to the uncorrelated and correlated uncertainties. It follows that, using a Hessian approach, the best value for t after the minimization of the χ^2 function is

$$t = \frac{\sum_{i,j=1}^{n}(\mathrm{cov}^{-1})_{ij}m_j}{\sum_{i,j=1}^{n}(\mathrm{cov}^{-1})_{ij}} \tag{2.13}$$

and its variance V_{tt} is

$$V_{tt} = \left(\tfrac{1}{2}\frac{\partial^2\chi^2}{\partial t^2}\right)^{-1} = \frac{1}{\sum_{i,j=1}^{n}(\mathrm{cov}^{-1})_{ij}}, \tag{2.14}$$

for a χ^2 function defined as

$$\chi^2(t) = \sum_{i,j=1}^{n}(t-m_i)(\mathrm{cov}^{-1})_{ij}(t-m_j). \tag{2.15}$$

An equivalent method, that is particularly useful in the case in which the t quantity is connected to another underlying theoretical quantity like PDFs, is the Monte Carlo method. Each data point is related to a random variable M_i that gives a Gaussian distribution centered in m_i and spread around this value according to the covariance matrix $(\mathrm{cov})_{ij}$. From these n random variables it is possible to generate N_{rep} replicas of the original set of n data points. Calling $\{T\}$ the ensemble of replicas generated in this way, it is possible to determine for each of them the best value minimizing the same χ^2 function given in Eq. (2.15) for T and obtain

$$T = \frac{\sum_{i,j=1}^{n}(\mathrm{cov}^{-1})_{ij}M_j}{\sum_{i,j=1}^{n}(\mathrm{cov}^{-1})_{ij}}, \tag{2.16}$$

and given that

$$\mathrm{E}[t] = \langle T \rangle, \quad \mathrm{Var}[t] = \langle T^2 \rangle - \langle T \rangle^2 \tag{2.17}$$

it follows

$$\begin{aligned} \mathrm{E}[t] &= \frac{\sum_{i,j=1}^{n} (\mathrm{cov}^{-1})_{ij} m_j}{\sum_{i,j=1}^{n} (\mathrm{cov}^{-1})_{ij}} && (2.18) \\ \mathrm{Var}[t] &= \frac{\sum_{i,j,k,l=1}^{n} (\mathrm{cov}^{-1})_{ij} (\mathrm{cov}^{-1})_{kl} (\langle M_j M_l \rangle - \langle M_j \rangle \langle M_l \rangle)}{\left(\sum_{i,j=1}^{n} (\mathrm{cov}^{-1})_{ij} \right)^2} \\ &= \frac{1}{\sum_{i,j=1}^{n} (\mathrm{cov}^{-1})_{ij}}. && (2.19) \end{aligned}$$

This is the same result found with the Hessian approach given in Eqs. (2.13)-(2.14) Of course the results depend on the choice of the error function to minimize.

D'Agostini Bias

As studied into detail in Ref. [103], normalization uncertainties can be added to the simple case considered above to illustrate the Hessian and Monte Carlo methods. Considering the MC method, it follows that the results for a single experiment read

$$\begin{aligned} E[t] &= \frac{\sum_{i,j=1}^{n} (\mathrm{cov}^{-1})_{ij} m_j}{\sum_{i,j=1}^{n} (\mathrm{cov}^{-1})_{ij}} && (2.20) \\ \mathrm{Var}[t] &= \frac{1 + s^2}{\sum_{i,j=1}^{n} (\mathrm{cov}^{-1})_{ij}} + s^2 \mathrm{E}[t]^2 && (2.21) \end{aligned}$$

as expected, where s is the overall normalization uncertainty of this single experiment. On the contrary, considering a case in which each measurement m_i comes from a different experiment along with its independent normalization uncertainty s_i, gives the results

$$\begin{aligned} E[t] &= \Sigma^2 \sum_{i=1}^{n} \frac{m_i}{\sigma_i^2} && (2.22) \\ \mathrm{Var}[t] &= \Sigma^2 + \Sigma^4 \sum_{i=1}^{n} s_i^2 (m_i^2 + \sigma_i^2)/\sigma_i^4 && (2.23) \end{aligned}$$

with

$$\frac{1}{\Sigma^2} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}. \tag{2.24}$$

It is clear that the result of Eq. (2.22) cannot be correct: this result does not depend on the normalization uncertainty s, and if is imagined a single data point m_i having a huge value of s_i its effect would not be taken into account by the expression here determined.

Here the realistic situation in which correlations between different measurements from independent experiments are not available is assumed. However their inclusion does not affect the result.

Considering now the Hessian method, things goes still worse: again, an error function that includes normalization uncertainties is used, with the simple choice

$$\chi^2_m(t) = \sum_{i,j=1}^{n} (t - m_i)(\mathrm{cov}_m^{-1})_{ij}(t - m_j) \tag{2.25}$$

where

$$(\mathrm{cov}_m)_{ij} = (\mathrm{cov})_{ij} + s^2 m_i m_j. \tag{2.26}$$

For the simple case of a single experiment with only two data points the covariance matrix is

$$(\mathrm{cov}_m)_{ij} = \begin{pmatrix} \sigma_1^2 + s^2 m_1^2 & s^2 m_1 m_2 \\ s^2 m_1 m_2 & \sigma_2^2 + s^2 m_2^2 \end{pmatrix} \tag{2.27}$$

and as a consequence the χ^2 Eq. (2.25) is

$$\chi^2_m(t) = \frac{(t - m_1)^2(\sigma_2^2 + m_2^2 s^2) + (t - m_2)^2(\sigma_1^2 + m_1^2 s^2) - 2(t - m_1)(t - m_2)m_1 m_2 s^2}{\sigma_1^2\sigma_2^2 + (m_1^2\sigma_2^2 + m_2^2\sigma_1^2)s^2}. \tag{2.28}$$

The result of the minimization of this expression with respect to t is

$$t = \frac{m_1/\sigma_1^2 + m_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2 + (m_1 - m_2)^2 s^2/\sigma_1^2\sigma_2^2}. \tag{2.29}$$

That this expression is biased can be better seen by assuming the simple case $\sigma_1 = \sigma_2 = \sigma$. Then, defining $\bar{m} \equiv \frac{1}{2}(m_1 + m_2)$ and $r \equiv \frac{m_1 - m_2}{m_1 + m_2}$, the above expression for t reduces to

$$t = \frac{\bar{m}}{1 + 2r^2 s^2 \bar{m}^2/\sigma^2} = \bar{m}(1 - 2r^2 s^2 \bar{m}^2/\sigma^2 + O(r^4)). \tag{2.30}$$

It is possible to notice a downward bias, also present in the variance for t

$$V_{tt} = \frac{\Sigma^2 + s^2 w^2(1 + r^2)}{1 + r^2 s^2 w^2/\Sigma^2}, \tag{2.31}$$

where

$$w \equiv \Sigma^2 \sum_{i=1}^{n} \frac{m_i}{\sigma_i^2}. \tag{2.32}$$

It is common to refer to this as the *d'Agostini bias*, from the name of the author of Ref. [100, 102].

Penalty Trick and t_o Method

As already seen, to use the most intuitive error function containing normalization uncertainties leads to biased results. A new function to be minimized that could at the same time take into account multiplicative errors and produce unbiased results has to be defined. Considering the Hessian method, a standard way of (partially) solving the problem is to include as parameters into the fit also the normalizations of the data, n_i, and to add penalty terms in the error function to fix the estimated value of these parameters close to one and their variance close to s_i^2.

In the simple case of a single experiment with s^2 the variance of its overall normalization uncertainty, the error function in the penalty trick approach is defined as

$$E_{\text{Hess}}(t,n) = \sum_{i=1}^{n} \frac{(t/n - m_i)^2}{\sigma_i^2} + \frac{(n-1)^2}{s^2}\,. \tag{2.33}$$

The last term is the penalty term: minimizing with respect to t gives $t = nw$, where w is defined in Eq. (2.32), while minimizing with respect to n gives $n = 1$ and inverting the respective Hessian matrix it follows that the covariance matrix

$$V_{tt} = \Sigma^2 + s^2 w^2 \tag{2.34}$$

is the same as the result easily obtained from Eq. (2.21) in the Monte Carlo approach by considering uncorrelated uncertainties σ_i (apart from a negligible term of cross-correlation between variances). The result for a single experiment is unbiased, but considering the case of more data points from different experiments with independent normalizations $n_i = 1 \pm s_i^2$ things again goes wrong. In this case

$$E_{\text{Hess}}(t,n_i) = \sum_{i=1}^{n} \frac{(t/n_i - m_i)^2}{\sigma_i^2} + \sum_{i=1}^{n} \frac{(n_i-1)^2}{s_i^2}\,, \tag{2.35}$$

and the minimum for this more complex error function is determined by $n+1$ equations for t and n_i:

$$t = \frac{\sum_{i=1}^{n} \frac{m_i}{n_i \sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{n_i^2 \sigma_i^2}}, \tag{2.36}$$

$$n_i = 1 + \frac{s_i^2 t}{n_i^2 \sigma_i^2}\left(\frac{t}{n_i} - m_i\right)\,. \tag{2.37}$$

These equations cannot be solved easily because of their complexity and non-linearity. In certain special cases, solutions can be found [103]. It is clear from these special cases that

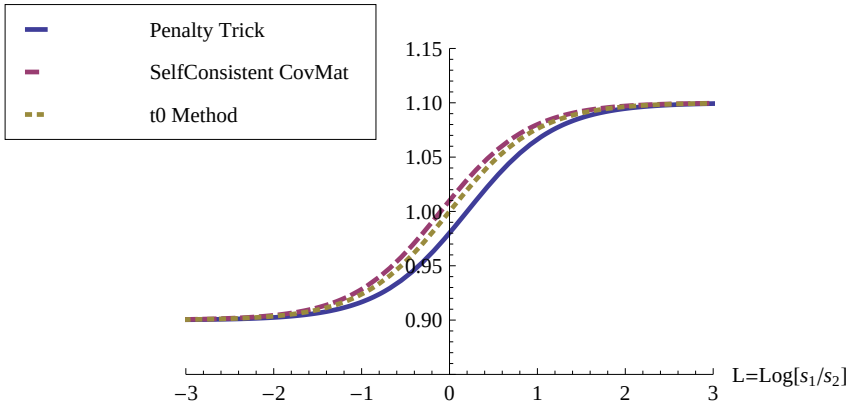

Figure 2.6: Dependence of the central value t on the ratio of normalization uncertainties s_1 and s_2 for a pair of measurements with central values $m_1 = 0.9$, $m_2 = 1.1$ and negligible uncertainties σ_i. The unbiased result must be symmetric about the point $L = 0$ (i.e. $s_1 = s_2$): only the t_0 curve is unbiased. The image is taken from Ref. [103].

both the central value and the variance are again biased. In summary, with the penalty trick the problem is only solved for the case of a single experiment.

In Ref. [100] d'Agostini proposed an alternative method but in Ref. [103] it is shown that also this method, in the case of more than one experiment, is biased. There this method is called self-consistent covariance matrix method, and relies on a definition of the covariance matrix that reads

$$(\mathrm{cov}_t)_{ij} = (\mathrm{cov})_{ij} + s^2 t^2 \ . \tag{2.38}$$

As for the penalty trick, for a single experiment everything works fine, but for several experiments biased results and also multiple solutions are found.

The main problem is that including a t dependence inside the covariance matrix or anyway adding terms that make the error function no more quadratic in t brings as a consequence that the distribution $\exp[-\frac{1}{2}\chi^2_t(t)]$ is not a Gaussian. Also the self-consistent covariance matrix method, while the t parameters inside the covariance matrix varies, present this problem. Considering here once again the Hessian approach, in Ref. [103] the so-called t_0-method is discussed for both the Hessian and the MC methods. The idea is to fix the covariance matrix by fixing its t dependence to an arbitrary value t_0

$$(\mathrm{cov}_{t_0})_{ij} = (\mathrm{cov})_{ij} + t_0^2 s^2, \tag{2.39}$$

in the case of a single experiment or

$$(\mathrm{cov}_t)_{ij} = (\sigma_i^2 + s_i^2 t_0^2)\delta_{ij}, \tag{2.40}$$

for several ones and then to determine iteratively the right value of t_0 that gives stability over subsequent iterations. In this way the problem is solved both for a single experiment and for more experiments. In all cases the results are unbiased for t and for its variance. The procedure converges quite quickly: in a couple of iterations, as shown in Ref. [103], stability is achieved. The dependence on t_0 is quite weak (it determines uncertainties and so a variation of it can be considered a second order effect) and so even with a bad first choice of its value the impact is mild. In the same reference above this is quantitatively proven. In Fig. 2.6 an interesting comparison among the penalty, self-consistent matrix, and t_0 methods is shown.

2.4 PDFs Fitting

A first difficulty in performing a fit is represented by the total ignorance on the shape each PDF has. When fitting a set of data points that obeys a known physical law, the problem of determining the best function that goes through the points is reduced to the determination of a finite set of parameters. In the present case it shouldn't be decided *a priori* a fixed functional form, because actually which is the best form for a parton distribution is not known. For this reason parton distributions need to be described with a sufficiently loose parametrization, in order to reduce as much as possible the bias on results introduced by fixing a functional form. Practical reasons don't allow for a too large set of parameters, and so what almost all the parton-fitter collaborations do to approach this problem is to exploit some well-known property a parton set must satisfy.

The usual parametrization for a single parton distribution has the form

$$f_i(x, Q_0^2) = ax^b(1-x)^c P(d_1, d_2, \ldots, x) \tag{2.41}$$

where Q_0^2 is the initial scale at which PDFs are determined. The desired scale Q^2 at which PDFs are convoluted with coefficient functions to compute theoretical predictions is reached using DGLAP evolution equations, as already treated. The parameters a, b, c are somehow constrained respectively by imposing momentum and valence sum rules, by Regge interpretation at small-x, and by constraining to zero parton distributions at $x = 1$. In general, a Padé expansion [104] is used to derive Eq. (2.41). Through this technique it is often possible to approximate a function with a rational function with a better precision than with a truncated Taylor expansion. The $P(\vec{d}, x)$ expression above is a smooth polynomial and the number of d parameters in the fit must be high enough to avoid any possible functional tension but at the same time it cannot be too high for practical reasons. A radically different approach to this problem has been developed

inside the NNPDF collaboration. More details on this are presented in the next Chapter.

A figure of merit is defined to determine the best set of parameters assuming that all the uncertainties are Gaussian. The quality of the fit can be described using a χ^2 function

$$\chi^2 = \sum_{i,j=1}^{N_{dat}} (F_i^{(\text{exp})} - F_i^{(\text{th})})[\text{cov}^{-1}]_{ij}(F_j^{(\text{exp})} - F_i^{(\text{th})}). \tag{2.42}$$

This function is minimized during the fitting procedure, and its dependence on the set of parameters is all contained into the theoretical predictions $F_i^{(\text{th})}$. The indices i and j run over the N_{dat} number of experimental points in the fit, $F_i^{(\text{exp})}$ are the measured central values of the observables. If in the case of a polynomial parametrization to a smaller χ^2 corresponds a better fit, in the case of a parametrization through artificial neural networks, as for NNPDF analyses, things are different. The best fit does not corresponds to the smaller possible χ^2, and as discussed in Sect. 3.6 a stopping criterion must be defined to determine the best set of parameters. The covariance matrix is generally defined as

$$[\text{cov}]_{ij} = \Big(\sum_{i,j=1}^{N_c} \sigma_{i,l}\sigma_{j,l} + \sum_{i,j=1}^{N_a} \sigma_{i,n}\sigma_{j,n} + \sum_{i,j=1}^{N_r} \sigma_{i,n}\sigma_{j,n} + \delta_{ij}\sigma_{i,s}^2 \Big) F_i^{(\text{exp})} F_j^{(\text{exp})} \tag{2.43}$$

where $\sigma_{i,l}$ are the N_c correlated systematic uncertainties, $\sigma_{i,n}$ the N_a (N_r) absolute (relative) normalization uncertainties, and $\sigma_{i,s}$ are the statistical uncertainties. In Sect. 2.2 the treatment of normalization and in general of multiplicative uncertainties was already discussed. A proper inclusion of these contributions that avoids systematic biases is in fact nontrivial.

The lack of experimental data that could constrain and disentangle the various partonic components may call for the introduction of some assumptions. The number of independent components is $(2n_f + 1)$, to be extracted from experimental data involving linear combinations of PDFs. However, in several nowadays obsolete analysis the assumption $\bar{u} = \bar{d}$ was used, due to the impossibility of disentangle the two distributions for lack of information. Also, not so long ago, the assumption of setting to zero the strange valence distribution was commonly imposed in parton fits. Nowadays all the $(2n_f + 1)$ components can be determined separately, with different degrees of constraint depending on the considered PDF. This is achieved thanks to the universality of PDFs: this feature allows for a global QCD analysis where, as anticipated in Sect. 2.1, the results proceeding from many different experiments are combined into a same analysis as described in a more detailed way in Sect. 3.1.

A very important role in PDF determination is played by benchmarks against results obtained using different techniques and methods. To this extent there is a constant effort

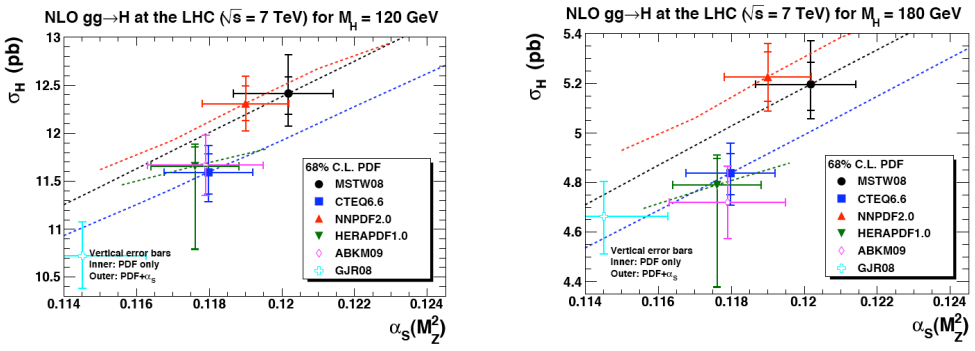

Figure 2.7: Cross-section predictions at 7 TeV for a Higgs boson (gg fusion) for a Higgs mass of 120 GeV (left) and 180 GeV(right). Plot by G. Watt [112], PDF4LHC.

within the scientific community that was translated in the series of HERALHC [105] workshops between years 2004-2008 and nowadays in the ongoing PDF4LHC series [106, 107] that since 2008 is intended not only to perform benchmark exercises but also to provide guidance on PDFs to LHC experiments and phenomenology.

Another important contribution to the development in parton fitting is represented by the Les Houches Accord PDFs (LHAPDF), first conceived in 2001 [108–110], that provides a library with all the main and most recent PDF sets from all the delivering collaborations [111], along with precise instructions on the usage of each set, and which has favoured useful discussion among different groups to establish baseline standards to PDF determinations (data inclusion, methodology, theoretical issues).

More details regarding the NNPDF approach to parton fitting and specifically on PDFs determination are extensively given in the next Chapter.

2.5 Summary of Available PDF Sets

Several parton sets have been delivered during the years, superseding obsolete ones. New sets are usually updated with the inclusion of new data, theoretical or statistical features. In this Section, the main characteristics and the status of the most commonly used PDF sets are discussed and compared to each other, leaving the discussion around NNPDF sets for the next Chapters.

ABM11

The ABM11 PDF set [113] is based on a dataset composed by a series of DIS inclusive data (HERA [114, 115], BCDMS [116, 117], NMC [58], SLAC [118–122]), Drell-Yan

data (E605 and E866 [123, 124]), and DIS dimuonic data (NuTeV and CCFR [115]). The parton set is determined both at NLO and at NNLO QCD perturbative orders for six independent PDFs. The parametrization is obtained through a polynomial function. The whole set of PDFs is determined fitting 25 parameters, among which the α_s coupling constant is also fitted. Several sets are delivered with α_s variations: the strong coupling is varied between $\alpha_s = 0.110$ and $\alpha_s = 0.130$ over 21 members (steps of $\Delta\alpha_s = 0.001$). The heavy quark mass effects are treated in a fixed flavor number scheme, delivering sets corresponding to $n_f = 3, 4, 5$ for both NLO and NNLO. The α_s variations are instead performed only for $n_f = 5$. To determine the errors over PDFs, experimental uncertainties are treated through the Hessian method, setting a tolerance $T^2 = \Delta\chi^2 = 1$.

These sets are publicly available, and for each ABM11 set 29 replicas are delivered: the central value plus 28 replicas to compute the errors. In the case of α_S variations the delivered replicas are respectively 21 (20+1) at NLO and 17 (16+1) at NNLO.

CJ12

The CJ12 PDF set [125] is based on a global dataset composed by DIS data (BCDMS [116], NMC [58, 59], SLAC [126], JLab [127], HERA [128]), Drell-Yan data (E866 [129]), W asymmetry (CDF [130–132], D0 [133, 134]) and Z rapidity (CDF [135, 136]) data, jets (CDF [137, 138], D0 [139, 142]) and photon-jet (D0 [143]) data. The CJ12 parton set is only available at NLO. The parametrization is given using 27 parameters for the fitting of 5 independent PDFs through a polynomial parametrization. The value for α_S is fixed from an external fit at $\alpha_S = 0.118$. The heavy quark effects are not treated, as the implemented scheme is a zero-mass variable flavor number scheme. The masses of the heavy quarks only have a role in defining the thresholds at which the number of active flavors is increased. The errors on PDFs are determined using the Hessian method with $T^2 = \Delta\chi^2 = 100$. The analysis includes target mass and higher twist corrections needed for the description of deep-inelastic scattering data at large-x and low Q^2, and nuclear corrections for deuterium targets.

Three different parton sets are delivered, determined respectively with three different methods for implementing nuclear corrections. Each set has a central value replica and 38 replicas for the computation of errors.

CT10

The CT10 PDF set [76, 144] is based on a global dataset that is almost equivalent to the one on which is based the CJ12 set. The main differences are that CJ12 doesn't include NuTeV data, while CT10 doesn't include SLAC, JLab, and photon-jet data. This parton set is available at three different perturbative orders: LO, NLO, and NNLO. Six

independent PDFs are fitted using 26 parameters that fix a polynomial functional form used to parametrize the parton functions. The heavy quark mass effects are treated through the implementation of the S-ACOT [37] general mass scheme, introduced in Sect. 2.1. The propagation of experimental uncertainties from data to PDFs is performed using the Hessian method with tolerance $T^2 = \Delta\chi^2 = 100$. The value of α_S is fixed from an external fit to the value $\alpha_S = 0.118$. A series of sets is delivered both at NLO and NNLO, including or excluding the W asymmetry data, and varying the value of the strong coupling constant, with $0.112 < \alpha_S < 0.127$ for the NLO analysis and with $0.110 < \alpha_S < 0.130$ for the NNLO analysis and step $\Delta\alpha_S = 0.001$ in both cases. Also, sets with $n_f = 3, 4$ are produced. Each set is fitted using the standard CTEQ PDF evolution but using the HOPPET α_s running solution.

The CT10 sets are delivered to the user as a 52 eigenvector sets (two eigenvectors for each fitted parameter) plus the central value in the standard analysis, as a single member in the varying α_S studies, and as a two-members set in the cases of fixed n_f.

HERAPDF1.5

The HERAPDF1.5 PDF set [145] is based on preliminary HERA I+II [146] combined dataset. Both NLO and NNLO sets are available, as the result of a polynomial parametrization of 5 independent PDFs. The set of parameters fixed to their best-fit values is composed by 14 elements. Uncertainties from experimental data are propagated to parton functions through the Hessian method with a tolerance $T^2 = \Delta\chi^2 = 1$. Heavy quark mass effects are treated using the Thorne Roberts (TR) [69] general mass scheme introduced in Sect. 2.1. The α_S value is fixed from an external fit to the value $\alpha_S = 0.1176$. Also a series of sets with varying α_S is produced: variations are produced over 12 values of α_s, between $\alpha_S = 0.114$ and $\alpha_S = 0.122$ with step $\Delta\alpha_S = 0.001$ and then for $\alpha_S = 0.1156, 0.1176, 0.1196$ both at NLO and NNLO. An analysis is also performed for varying parameters, as for example heavy quarks mass values.

While the standard sets are delivered for NLO and NNLO as eigenvector sets of respectively 20+1 and 28+1 members, for parameter variation sets the members are 12+1 at NLO and 10+1 at NNLO. For varying α_S the eigenvector sets have 12 members, both for NLO and NNLO.

JR09

The JR09 PDF set [147, 148] is based on the HERA electron-proton measurements [149–153], fixed target SLAC data [126], BCDMS [116, 117], E665 [154, 155], NMC [58, 59], and Drell-Yan dimuon data E866/NuSea [124, 156]. The parton set is available both at

NLO and at NNLO. Its parametrization is obtained using 15 parameter to fix a polynomial functional form for the 5 independent PDFs fitted in the analysis. The α_S coupling constant is also fitted. The errors for PDFs are determined using the Hessian method with tolerance $T^2 = \Delta\chi^2 = 1$. While the NLO parton set is obtained only in the FFN scheme with $n_f = 3$, the NNLO parton set is obtained both in the FFN scheme and in the VFN scheme.

The JR09 parton sets are delivered as a central value plus 26 eigenvector sets for error computation both at NLO and at NNLO.

MSTW08

The MSTW08 PDF set [81,82] is based on a global dataset composed by DIS data, Drell-Yan data, W asymmetry and Z rapidity data, and jets data. A description of each component of the dataset is given in Ref. [81]. This parton set is available at LO, NLO and at NNLO. To determine the best fit a polynomial functional form is used for each PDF. The independent PDFs that are fitted are 7, with a total of 20 free parameters to be determined during the fit. The α_S coupling constant is also fitted and sets with α_S variations are delivered. The errors are determined through the Hessian method, imposing a tolerance $T^2 = \Delta\chi^2 \sim 25$. The heavy quark mass effects are introduced using the TR general mass variable flavor number scheme.

The MSTW08 parton sets are available as 40 eigenvector sets (a set for each minimized parameter for each direction of variation) plus a central value set. the α_S variation sets are instead given as a best-fit central set with $\alpha_S = 0.12018$ plus 21 sets with α_S varying between 0.110 and 0.130, with step $\Delta\alpha_S = 0.001$.

	Reference	Dataset	Perturbative Order	Parametrization	HQ Treatment	α_s	Uncertainties
ABM11	[113]	DIS DY	NLO NNLO	6 indep. PDFs Polynomial $N_{par} = 25$	FFN	Fitted Mult. α_s values	Hessian $T^2 = 1$
CJ12	[125]	Global	NLO	5 indep. PDFs Polynomial $N_{par} = 27$	ZM-VFNS	External	Hessian $T^2 = 100$
CT10	[76] [144]	Global	LO NLO NNLO	6 indep. PDFs Polynomial $N_{par} = 26$	S-ACOT	External Mult. α_s values	Hessian $T^2 = 100$
HERAPDF1.5	[145]	DIS (HERA)	NLO NNLO	5 indep. PDFs Polynomial $N_{par} = 14$	TR	External Mult. α_s values	Hessian $T^2 = 1$
JR09	[147] [148]	DIS DY	NLO NNLO	5 indep. PDFs Polynomial $N_{par} = 15$	FFN VFN	Fitted	Hessian $T^2 = 1$
MSTW08	[81] [82]	Global	LO NLO NNLO	7 indep. PDFs Polynomial $N_{par} = 20$	TR	Fitted Mult. α_s values	Hessian $T^2 \sim 25$
NNPDF2.1/2.3	[40] [87] [50]	Global	LO NLO NNLO	7 indep. PDFs Neural Nets $N_{par} = 259$	FONLL	External Mult. α_s values	MC

Chapter 3

NNPDF Methodology

In this Chapter the strategy followed by the NNPDF collaboration is described. While the general strategy has already been discussed in Sect. 2.1, here the discussion will treat the radically different ingredients developed and implemented within the NNPDF framework.

An alternative to standard methods in error determination is Monte Carlo sampling in data space. This method was presented in Ref. [157]. It was successfully combined with an artificial neural network parametrization in Ref. [158]. Neural networks are universal unbiased interpolators and can be trained on each set of data replicas generated through MC sampling. The flexibility and robustness of this approach has been developed and tested in a series of different cases (structure functions [159], spectral functions for τ decays [160], energy spectra of B decays [161], and cosmic ray neutrino fluxes [162]) and is at the basis of all NNPDF parton releases. To go from the data to the parton parametrization first N_{rep} pseudo-data replicas are generated for each data point in the fit. In this way N_{rep} replicas of the original dataset (that is identified with replica $i_{rep} = 0$) are created.

The construction of physical observables to compare to the corresponding data point is a nontrivial step: it implies to parametrize PDFs with neural nets at a reference scale Q_0^2, evolve at the physical scale Q^2 and convolute the result with hard partonic cross-sections. The physical observable obtained is then used to compute the error function to be minimized. Minimization is obtained through a genetic algorithm that acts in the selection of the best set of parameters that determine the neurons of the networks iteration by iteration during the training. The optimal fit is obtained by dynamical stopping using a cross-validation method.

For each PDF and for each replica $i_{rep} = 1, \ldots, N_{rep}$ of the original dataset a neural network is trained, so that at the end are produced N_{rep} sets of PDFs that are a Monte Carlo representation of probability density in the space of parton distributions. From this representation it is straightforward to determine central values, errors, and correlations of

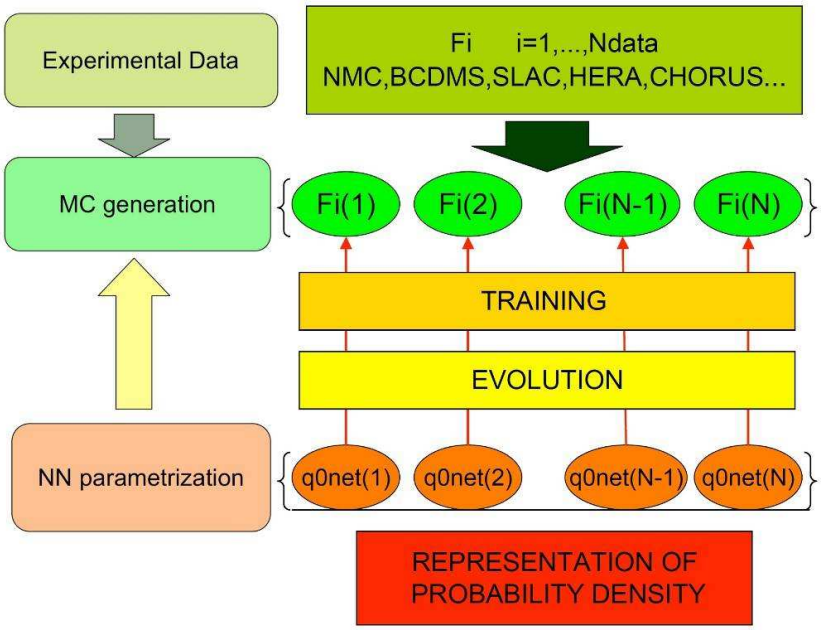

Figure 3.1: Schematic representation of the NNPDF approach.

any function depending on PDFs.

In this Chapter the experimental data and Monte Carlo sampling are at first discussed (Sect. 3.1, 3.2). The central part of the Chapter is dedicated to the minimization process and some related features including the structure and features of the neural networks used for the parametrization, the genetic algorithm and the dynamical stopping (Sect. 3.3). The FastKernel framework is then introduced (Sect. 3.7) and finally an important feature to include new datasets is discussed: the reweighting method (Sect. 3.8).

3.1 Experimental Data

In this Section, a detailed discussion of kinematic cuts, with their respective motivation, and the kinematic coverage of the dataset used for the NNPDF2.1 LO, NLO and NNLO parton distribution function analyses is performed. Particular emphasis is given to the inclusion of new charm structure function data in the global fit, as to correctly fit this observable a general mass scheme is needed. The original datasets are the basis for the Monte Carlo sampling method used to generate pseudo-data replicas. The second part of the Section reports and briefly illustrates the detailed analyses of data performed in Ref. [84, 87]. Finally, also positivity constraints are discussed, paying attention to the different methods used for different perturbative approximations.

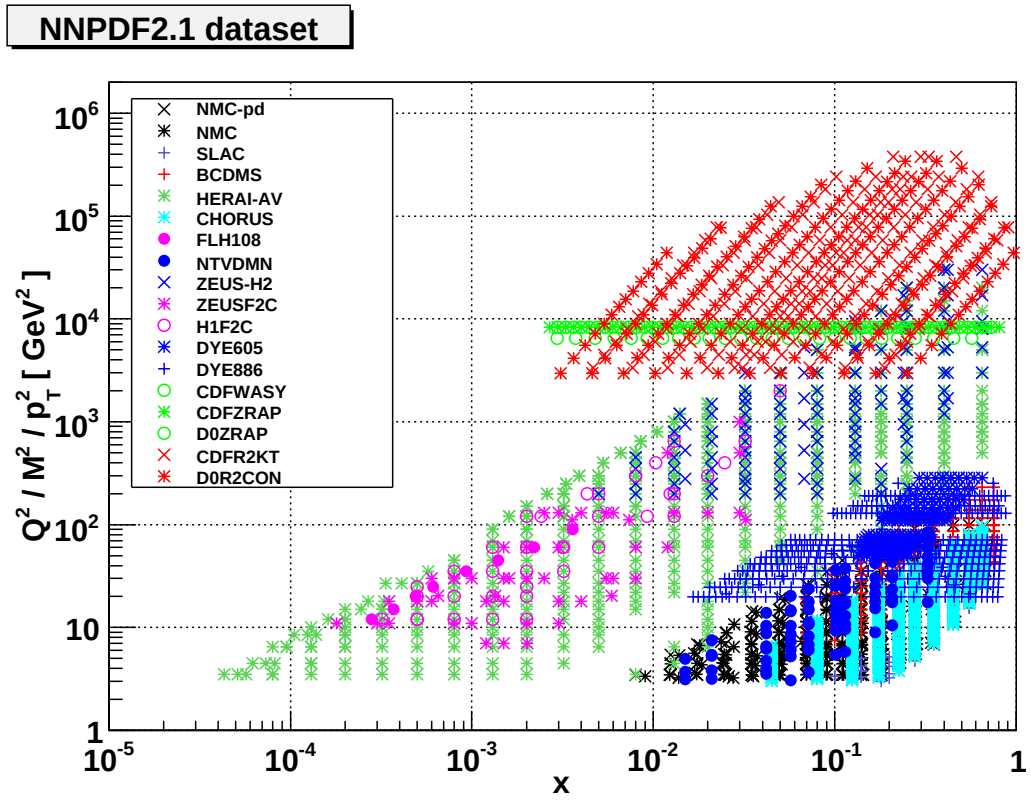

Figure 3.2: Experimental datasets which enter the NNPDF2.1 analysis. The kinematical coverage of each dataset is summarized in Table 3.1.

Kinematic Cuts

First is considered the NNPDF2.1 NLO dataset. The kinematic cuts are performed in variables Q^2 and $W^2 = Q^2(1-x)/x$. While the cut in W^2 is the same as in previous NNPDF analyses and set to $W^2_{\rm min} = 12.5$ GeV^2, the one in Q^2 is now slightly higher, going from $Q^2_{\rm min} = 2$ GeV^2 for DIS data [87] in NNPDF2.0 to $Q^2_{\rm min} = 3$ GeV^2 in the present NNPDF2.1 NLO analysis [40].

The cuts here discussed are motivated by the fact that close to heavy quark threshold F_2^c predictions from the general mass scheme here implemented might be affected by instabilities. The cause can be found in having data points crossing the heavy quark mass threshold while varying its value in different fits. This suggests to use a value of $Q^2_{\rm min}$ at least as large as the maximum value of the charm mass that can be considered acceptable. Then, $Q^2_{\rm min} = 3$ GeV^2 is a reasonable choice since then $m_c^{\rm max} \sim 1.7$ GeV. Another motivation can be found in Ref. [163, 164]. Possible deviations from NLO DGLAP in the small-x and Q^2 HERA data can affect theoretical uncertainty in the PDFs and LHC observables related to their inclusion in the global fit in a moderate way compared to PDF errors and other uncertainties, but removing the HERA points below $Q^2_{\rm min}$ reduces

these theoretical uncertainties even further. The price to pay for this reduced theoretical uncertainty is an increase in statistical uncertainty.

While the kinematic cuts discussed above apply to all DIS experiments included in the global fit, on HERA F_2^c data additional cuts are performed. Concretely, HERA F_2^c data with $Q^2 \leq 4$ GeV^2 and data with $Q^2 \leq 10$ GeV^2 for $x \leq 10^{-3}$ are removed from the fit. The motivation for this is that the FONLL-A general mass scheme for heavy quarks, as discussed in [36], provides a poor description of the data in the smallest x and Q^2 bins due to missing large $\mathcal{O}(\alpha_s^2)$ corrections. This is true for any heavy quark scheme that does not include the $\mathcal{O}(\alpha_s^2)$ corrections, like for example the S-ACOT-χ used in the CTEQ/CT family of PDF sets. Only the FONLL-B scheme can cure this problem since it includes consistently $\mathcal{O}(\alpha_s^2)$ corrections in F_2^c into a NLO fit, as can be seen in [36]. These cuts ensure that all F_2^c experimental data included in the fit are well described by $\mathcal{O}(\alpha_s)$ theory.

Kinematic cuts on the invariant mass W^2 and the scale Q^2 of the DIS final state $W^2_{\min} > 12.5$ GeV^2 and $Q^2 > 3$ GeV^2 are the same for NNPDF2.1 LO, NLO and NNLO datasets, while the F_2^c data are subject to the further cuts $Q^2 > 4$ GeV^2 and $Q^2 > 10$ GeV^2 if $x < 10^{-3}$ only in the LO and NLO analyses, due to the fact that in this region NNLO massive corrections are so large that a NLO approximation is not acceptable (and consequently at LO). These cuts will be removed for the NNLO fit, in which the F_2^c data will only be subject to the cuts which are common to all other DIS data. The charm structure function data included in the NNLO fit are listed in Table 3.2 all other data are the same as in the NLO fit, Table 3.1. The total numbers of datapoints used at LO, NLO and NNLO are also given in Table 3.2.

NNPDF2.1 Dataset

The NNPDF2.1 NLO dataset includes NMC [165,166], BCDMS [116,117] and SLAC [126] deep-inelastic scattering fixed target data; the combined HERA-I DIS dataset [114], HERA F_L [167] and F_2^c structure function data [168–174], ZEUS HERA-II DIS cross-sections [175, 176], CHORUS [177] inclusive neutrino DIS, and NuTeV [115, 178] dimuon production data; fixed-target E605 [123] and E866 [124, 156, 179] Drell-Yan production data; CDF [132] W asymmetry and CDF [135] and D0 [136] Z rapidity distributions; and CDF [138] and D0 [142] Run-II one-jet inclusive cross-sections. A scatter plot of this data in the x, Q^2 plane is displayed in Fig. 3.2, with the values of x determined using LO kinematics.

The dataset is slightly different in the case of LO and NNLO fits. At LO the F_L structure function data are removed, since this observable vanishes at this perturbative order. At NNLO two modifications are present: first, the E866 data, published as x_F distribu-

tions, have been converted into rapidity distributions following the procedure discussed in Ref. [180], since the use of rapidity as kinematic variable makes the inclusion of NNLO corrections simpler. Second, instead of including the NMC proton data as data for structure functions they now are included as data for reduced cross-sections. This decision was taken after that in Ref. [181] was shown that the impact of this different treatment is almost negligible at NLO. The use of cross-section data is in principle preferable, as they are closer to what is actually measured. In Ref. [182] it was claimed that the treatment of this data may have a significant impact on NNLO PDFs, though this claim is not supported by investigations with NNPDF2.1 NNLO [183], or with MSTW08 [184] PDFs.

All the relevant data from H1 and ZEUS experiments at HERA for the charm structure function $F_2^c(x, Q^2)$ is included [168–174]. This inclusion of new data points in NNPDF analyses introduces more information that helps constraining the small-x gluon PDF and moreover the included datasets are sensitive to the value of the charm mass m_c. The kinematic coverage of all the datasets included in NNPDF2.1 is summarized in Table 3.1 and in Fig. 3.2.

Now in turn the features of the various $F_2^c(x, Q^2)$ datasets included in the fit are described:

- The ZEUS 96-97 $D^{*\pm}$ analysis [168].
  In this analysis F_2^c is extracted from the measurement of $D^{*\pm}$ mesons reconstructed via their hadronic decays using data collected in the 1996 and 1997 running periods.

- The ZEUS 98-00 D^* analysis [169].
  As in the previous case, F_2^c is extracted from the measurement of $D^{*\pm}$ mesons reconstructed via their hadronic decays, and uses data collected in the running period between 1998 and 2000.

- The 04-05 ZEUS $D^{\pm}, D^0$ analysis [170].
  In this analysis, based on the HERA-II running period of 2004 and 2005, D mesons are reconstructed via their hadronic decays. An improved precision is obtained reducing the combinatorial background to the D meson signals by using the ZEUS micro-vertex detector to reconstruct displaced secondary vertices.

- The 2005 ZEUS muon analysis [171].
  This dataset is based on the measurement of muons that are generated in charm production from their semileptonic decays. Data was collected during the 2005 HERA-II running period.

- The H1 96-97 $D^{*\pm}$ analysis [172].
  This analysis, based on the 1996-1997 running period, used similar reconstruc-

Deep-inelastic scattering							
Experiment	Set	Ref.	$N_{\rm dat}$	$x_{\rm min}$	$x_{\rm max}$	$Q^2_{\rm min}$ [GeV^2]	$Q^2_{\rm max}$ [GeV^2]
NMC-pd			260 (132)				
	NMC-pd	[165]	260 (132)	0.0015 (0.008)	0.68	0.2 (3.5)	99.0
NMC			288 (221)				
	NMC	[166]	288 (221)	0.0035 (0.009)	0.47	0.8 (3.2)	61.2
SLAC			422 (74)				
	SLACp	[126]	211 (37)	0.07 (0.1)	0.85 (0.55)	0.58 (3.0)	29.2
	SLACd	[126]	211 (37)	0.07 (0.1)	0.85 (0.55)	0.58 (3.2)	29.1
BCDMS			605 (581)				
	BCDMSp	[116]	351 (333)	0.07	0.75	7.5	230.0
	BCDMSd	[117]	254 (248)	0.07	0.75	8.8	230.0
HERAI-AV			741 (592)				
	HERA1-NCep	[114]	528 (379)	$6.2\,10^{-7}$ ($4.3\,10^{-5}$)	0.65	0.045 (3.5)	30000
	HERA1-NCem	[114]	145	$1.3\,10^{-3}$	0.65	90.000	30000
	HERA1-CCep	[114]	34	0.008	0.4	300.0	15000
	HERA1-CCem	[114]	34	0.013	0.4	300.0	30000
CHORUS			1214 (862)				
	CHORUSnu	[177]	607 (431)	0.02 (0.045)	0.65	0.3 (3.0)	95.2
	CHORUSnb	[177]	607 (431)	0.02 (0.045)	0.65	0.3 (3.0)	95.2
FLH108			8				
	FLH108	[167]	8	0.00028	0.0036	12.0	90.000
NTVDMN			90 (79)				
	NTVnuDMN	[115, 178]	45 (41)	0.027	0.36	1.1 (3.1)	116.5
	NTVnbDMN	[115, 178]	45 (38)	0.021	0.25	0.8 (3.1)	68.3
ZEUS-H2			127				
	Z06NC	[175]	90	$5\,10^{-3}$	0.65	200	$3\,10^5$
	Z06CC	[176]	37	0.015	0.65	280	$3\,10^5$
HERA charm structure function data							
ZEUSF2C			69 (50)				
	ZEUSF2C99	[168]	21 (14)	$5\,10^{-5}$ ($3\,10^{-4}$)	0.02	1.8 (7.0)	130
	ZEUSF2C03	[169]	31 (21)	$3\,10^{-5}$ ($1.8\,10^{-5}$)	0.03	2.0 (7.0)	500
	ZEUSF2C08	[170]	9 (7)	$2.2\,10^{-4}$ ($6.5\,10^{-4}$)	0.032	7.0	112
	ZEUSF2C09	[171]	8	$8\,10^{-4}$	0.03	30	1000
H1F2C			47 (38)				
	H1F2C01	[172]	12 (6)	$5\,10^{-4}$	$3.2\,10^{-3}$	1.5 (12)	60
	H1F2C09	[173]	6	$2.4\,10^{-4}$	0.025	120	400
	H1F2C10	[174]	26	$2\,10^{-4}$ ($3.2\,10^{-4}$)	0.05	5.0 (12)	2000
Fixed Target Drell-Yan production							
Experiment	Set	Ref.	$N_{\rm dat}$	$\left[y/x^F_{\rm min}, y/x^F_{\rm max}\right]$	$[x_{\rm min}, x_{\rm max}]$	$M^2_{\rm min}$ [GeV^2]	$M^2_{\rm max}$ [GeV^2]
DYE605			119				
	DYE605	[123]	119	$[-0.20, 0.40]$	$[0.14, 0.65]$	50.5	286
DYE866			390				
	DYE866p	[156, 179]	184	$[0.0, 0.78]$	$[0.017, 0.87]$	19.8	251.2
	DYE866r	[124]	15	$[0.05, 0.53]$	$[0.025, 0.56]$	21.2	166.4
Collider vector boson production							
Experiment	Set	Ref.	$N_{\rm dat}$	$[y_{\rm min}, y_{\rm max}]$	$[x_{\rm min}, x_{\rm max}]$	$M^2_{\rm min}$ [GeV^2]	$M^2_{\rm max}$ [GeV^2]
CDFWASY			13				
	CDFWASY	[132]	13	$[0.10, 2.63]$	$\left[2.9\,10^{-3}, 0.56\right]$	6463	6463
CDFZRAP			29				
	CDFZRAP	[136]	29	$[0.05, 2.85]$	$\left[2.9\,10^{-3}, 0.80\right]$	8315	8315
D0ZRAP			28				
	D0ZRAP	[140]	28	$[0.05, 2.75]$	$\left[2.9\,10^{-3}, 0.72\right]$	8315	8315
Collider inclusive jet production							
Experiment	Set	Ref.	$N_{\rm dat}$	$[y_{\rm min}, y_{\rm max}]$	$[x_{\rm min}, x_{\rm max}]$	$p^2_{T,\rm min}$ [GeV^2]	$p^2_{T,\rm max}$ [GeV^2]
CDFR2KT			76				
	CDFR2KT	[141]	76	$[0.05, 1.85]$	$\left[4.6\,10^{-3}, 0.90\right]$	3364	$3.7\,10^5$
D0R2CON			110				
	D0R2CON	[142]	110	$[0.20, 2.20]$	$\left[3.1\,10^{-3}, 0.97\right]$	3000	$3.4\,10^5$
Total							
Experiment			$N_{\rm dat}$	$x_{\rm min}$	$x_{\rm max}$	$Q^2_{\rm min}$ [GeV^2]	$Q^2_{\rm max}$ [GeV^2]
TOTAL			4520 (3415)	$3.1\,10^{-5}$	0.97	2.0	$3.7\,10^5$

Table 3.1: Experimental datasets included in the NNPDF2.1 global analysis. For DIS experiments in each case the number of data points and the ranges of the kinematical variables are provided before and after (in parenthesis) kinematical cuts. For hadronic data the ranges of parton x covered for each set determined using leading order parton kinematics are shown. Note that hadronic data are unaffected by kinematic cuts. The values of $x_{\rm min}$ and $Q^2_{\rm min}$ for the total dataset hold after imposing kinematic cuts.

Experiment	Set	Ref.	$N_{\rm dat}$	$x_{\rm min}$	$x_{\rm max}$	$Q^2_{\rm min}$	$Q^2_{\rm max}$
ZEUSF2C			69 (62)				
	ZEUSF2C99	[168]	21 (18)	$5\,10^{-5}$ ($1.3\,10^{-4}$)	0.02	1.8 (4)	130
	ZEUSF2C03	[169]	31 (27)	$3\,10^{-5}$ ($7\,10^{-5}$)	0.03	2.0 (4.0)	500
	ZEUSF2C08	[170]	9	$2.2\,10^{-4}$	0.032	7.0	112
	ZEUSF2C09	[171]	8	$8\,10^{-4}$	0.03	30	1000
H1F2C			47 (45)				
	H1F2C01	[172]	12 (10)	$5\,10^{-5}$ ($1.3\,10^{-4}$)	$3.2\,10^{-3}$	1.5 (3.5)	60
	H1F2C09	[173]	6	$2.4\,10^{-4}$	0.025	120	400
	H1F2C10	[174]	29	$2\,10^{-4}$	0.05	5.0	2000
LO Total			3330				
NLO Total			3338				
NNLO Total			3357				

Table 3.2: Charm structure function datasets included in the NNPDF2.1 NNLO analysis. All other data are the same as in the NNPDF2.1 NLO analysis, given in Table 3.1. The number of data points after kinematic cuts are shown in parentheses. In the last three lines is given the total number of datapoints included in the NNPDF2.1 LO, NLO and NNLO fits.

tion strategies as the corresponding ZEUS analysis, namely the reconstruction of $D^{*\pm}\to D^0\pi^+$ using the D^*-D^0 mass difference method.

- The H1 large Q^2 04-07 $D^{*\pm}$ analysis [173].
  This analysis determines F_2^c via identified D mesons produced at large virtualities $Q^2 \geq 100$ GeV^2, and is based on data collected in the HERA-II running period 2004 and 2007.

- The H1 low-Q^2 06-07 $D^{*\pm}$ analysis [174].
  This is analogous to the previous measurement, but now covering the small and medium Q^2 region. It is based on data obtained in the HERA-II 2006-2007 running period. Events containing heavy quarks are distinguished from those containing only light quarks using variables that are sensitive to the longer lifetimes of heavy flavor hadrons, like the transverse displacement of tracks from the primary vertex.

The datasets here above described are not all the published F_2^c HERA datasets. The reason to exclude all the other ones is that in comparison with the ones included they are obsolete and, moreover, the measurements used in the fit are the basis for the combined HERA F_2^c dataset.

In several analyses for the determination of PDF sets, F_2^c data are excluded for the fact that the way F_2^c is usually defined experimentally is affected by mass singularities. Usually F_2^c is defined as the contribution to F_2 with at least one charmed quark in the final state, and the result is not finite in the limit in which $m_c\to 0$. Here a definition of F_2^c as the contribution to F_2^c when only the charm electric charge is nonzero is adopted, which is free of mass singularities. The deviation between this definition and that which is used to define the experimental observable is estimated in Ref. [36] by means of a suitable resummation method, and shown to be negligible in the region of the HERA

$F_2^{\rm NS}(x,Q^2)$			
$N_{\rm rep}$	10	100	1000
$\left\langle PE\left[\left\langle F^{(\rm art)}\right\rangle_{\rm rep}\right]\right\rangle$	20%	6.4%	1.3%
$r\left[F^{(\rm art)}\right]$	0.97	0.99	0.99
$\left\langle V\left[\sigma^{(\rm art)}\right]\right\rangle_{\rm dat}$	$6.1\ 10^{-5}$	$1.9\ 10^{-5}$	$6.7\ 10^{-6}$
$\left\langle PE\left[\sigma^{(\rm art)}\right]\right\rangle_{\rm dat}$	33%	11%	3%
$\left\langle \sigma^{(\rm art)}\right\rangle_{\rm dat}$	0.011	0.011	0.011
$r\left[\sigma^{(\rm art)}\right]$	0.94	0.99	0.99
$\left\langle V\left[\rho^{(\rm art)}\right]\right\rangle_{\rm dat}$	0.10	$9.4\ 10^{-3}$	$1.0\ 10^{-3}$
$\left\langle \rho^{(\rm art)}\right\rangle_{\rm dat}$	0.182	0.097	0.100
$r\left[\rho^{(\rm art)}\right]$	0.47	0.79	0.97
$\left\langle V\left[{\rm cov}^{(\rm art)}\right]\right\rangle_{\rm dat}$	$5.5\ 10^{-9}$	$1.7\ 10^{-10}$	$5.7\ 10^{-11}$
$\left\langle {\rm cov}^{(\rm art)}\right\rangle_{\rm dat}$	$1.3\ 10^{-5}$	$7.6\ 10^{-6}$	$8.1\ 10^{-6}$
$r\left[{\rm cov}^{(\rm art)}\right]$	0.41	0.81	0.98

Table 3.3: Comparison between experimental and Monte Carlo data. The experimental data have $\left\langle \sigma^{(\rm exp)}\right\rangle_{\rm dat} = 0.011$, $\left\langle \rho^{(\rm exp)}\right\rangle_{\rm dat} = 0.107$ and $\left\langle {\rm cov}^{(\rm exp)}\right\rangle_{\rm dat} = 8.6\ 10^{-6}$.
Table taken from Ref. [83].

data. Also, F_2^c is affected by theoretical uncertainties related to the extrapolation from the experimentally accessible region (restricted in p_T and η) to the full phase space. This theoretical uncertainty is estimated using QCD exclusive partonic calculations and added as an extra source of systematic uncertainty in the experimental analysis.

3.2 Monte Carlo Generation

Through Monte Carlo generation, starting from the orginal dataset composed by the $N_{\rm dat}$ central values measured in the various experiments considered, $N_{\rm rep}$ sets of pseudo-data are produced. Considering $F_i^{(\rm exp)}$ as a general individual measurement, it follows:

$$F_i^{(\rm art)(k)} = S_{i,N}^{(k)} F_i^{(\rm exp)} \left(1 + \sum_{p=1}^{N_c} r_{i,l}^{(k)} \sigma_{i,l} + r_i^{(k)} \sigma_{i,s}\right) , \tag{3.1}$$

with

$$S_{i,N}^{(k)} = \prod_{n=1}^{N_a} \left(1 + r_{i,N}^{(k)} \sigma_{i,N}\right) \prod_{n=1}^{N_r} \sqrt{\left(1 + r_{i,N}^{(k)} \sigma_{i,N}\right)} \tag{3.2}$$

$$k = 1, \ldots, N_{\rm rep} \qquad , \qquad i = 1, \ldots, N_{\rm dat}\,.$$

Independent univariate Gaussian random numbers $r^{(k)}$ are used for each independent error source. These sources of error are the same introduced in the definition of covariance matrix in Eq. (2.43). Each Monte Carlo replica of the original experimental data is

generated according to a multi-Gaussian distribution centered in the original point (and so with expectation value equal to the data point value) and error and covariance equal to the corresponding experimental quantities. In this way it is possible to generate a sampling of the probability measure for each experimental data point. As the number of replicas N_{rep} is arbitrary, can be decided with which accuracy the generated pseudo-data would reproduce the original statistical properties of the dataset.

In order to correctly determine a set of replicas, the definition of some statistical estimator is needed: a set of them is defined in the Appendix of Ref. [83]. It is possible to check that averages, variance and covariance of the pseudo-data reproduce central values and covariance matrix elements of the original data. Results for a single observable were determined in Ref. [83] and are presented in Table 3.3. It follows that with $N_{rep} = 1000$ it is possible to ensure average scatter correlations of 99% and accuracies of a few percent on central values, errors and correlations.

3.3 Neural Network Parametrization

The parametrization used in all NNPDF analyses is given in terms of neural networks. This choice allows for an unbiased fit with a very large and redundant set of parameters. The fitting strategy is very important, as the dependence of the observables on PDFs is nontrivial. To minimize the error function for each neural network a genetic algorithm is used. The redundancy of the parametrization implies a subtle problem: surely the underlying *true* distribution given by data can be much better determined, but there is the possibility of minimizing the figure of merit beyond the best fit point. In this case would be fitted not only the physical law, but also statistical fluctuation of data. The solution to this problem is given using a cross-validation method to stop the fit before entering this *overtraining* region. These three ingredients (neural networks, genetic algorithm as fitting tool and cross-validation) are discussed respectively in this Section, in Sect. 3.5, and in Sect. 3.6.

The flexibility of a neural network depends on its size. An infinite size, as a limit case, corresponds to a neural network that can reproduce any continuous function. Of course, it is not possible to deal with such an infinite-sized neural network and a structure needs to be fixed. A neural network is a nonlinear map between input $\xi_i^{(1)}$ and output $\xi_i^{(L)}$ variables. For a correctly chosen structure, they can give a result free of functional biases. The standard methods, by fixing a functional form (for example polynomials of fixed degree) are likely to be affected by this kind of bias.

The structure of the neural networks used in NNPDF's fits is fixed in a 2-5-3-1 architecture. This result is achived through a series of stability tests: as is not known *a*

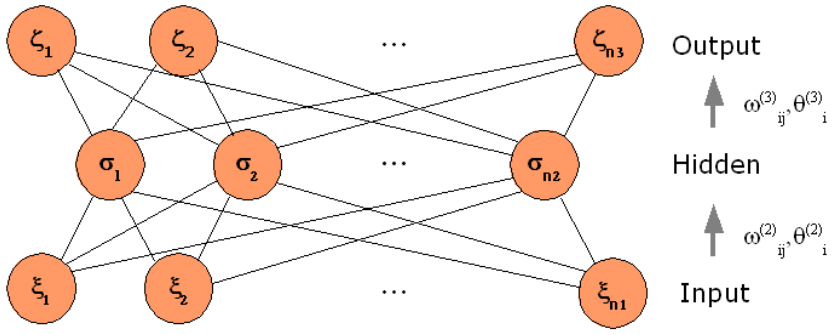

Figure 3.3: Schematic diagram of a feed-forward neural network.

priori which complexity is needed to correctly describe a PDF set, the size of the net is increased until independence of the results upon its variation is reached. Then the architecture is fixed slightly above the critical size that gives stability. A detailed study is performed in [83]. The neural networks used are multi-layer feed-forward neural networks. A schematic representation is given in Fig. 3.3. Neurons are disposed in L layers, each one with n_l neurons. The number of neurons changes with the layers, as is clear looking at the 2-5-3-1 architecture implemented. Each neuron receives input from the neurons of the previous layer and feeds output to the ones in the subsequent layer. These inputs are combined linearly with a weight each and a global threshold, generating the output $\xi_j^{(l)}$ (j-th neuron of the l-th layer) given by a nonlinear activation function $g(x)$

$$\xi_i^{(l)} = g\left(h_i^{(l)}\right) \ , \quad i = 1, \ldots, n_l \ , \quad l = 2, \ldots, L, \tag{3.3}$$

which has as argument the linear combination

$$h_i^{(l)} = \sum_{j=1}^{n_{l-1}} \omega_{ij}^{(l)} \xi_j^{(l-1)} - \theta_i \ , \tag{3.4}$$

where ω_{ij} (weights) and θ_i (thresholds) are free parameters to be determined by the fitting procedure, and $g(x)$ is taken to be a sigmoid in the inner layers,

$$g(x) = \frac{1}{1 + \exp(-x)} \ , \tag{3.5}$$

and linear $g(x) = x$ for the last layer.

In order not to have the neural network defined by weights that span many orders of magnitude, it is a common practice to rescale both the input and the output of the neural network between 0 and 1. Independence with respect to variations of this rescaling have been checked in [83].

The neural network outputs the values $\xi_j^{(L)}$ as a function of the input values $\xi_j^{(1)}$ and the parameters ω_{ij}, θ_i. The training of the neural network consists in the determination of the best-fit values of these parameters given a set of input-output patterns (data). As the same 2-5-3-1 architecture for all the neural networks present in the analysis is given, the number of parameters that need to be determined is of 37 parameters for each PDF, that means a total of 259 parameters. A standard fixed functional form fit has usually less than 30 parameters. In Sect. 3.5 the genetic algorithm needed for the training will be discussed.

Preprocessing

To speed up the fitting procedure, a preprocessing of the data is performed. As already said, a neural network can accomodate any functional form, provided that it has a large enough size and that it is trained for enough time. However, it is common to factorize with preprocessing the asymptotic behaviour that the parton distributions may have for $x \to 0$ and for $x \to 1$. The parametrization of the basis of PDFs implemented in the fit reads

$$\begin{aligned}
\Sigma(x,Q_0^2) &= (1-x)^{m_\Sigma} x^{-n_\Sigma} \mathrm{NN}_\Sigma(x)\,, \\
V(x,Q_0^2) &= A_V (1-x)^{m_V} x^{-n_V} \mathrm{NN}_V(x)\,, \\
T_3(x,Q_0^2) &= (1-x)^{m_{T_3}} x^{-n_{T_3}} \mathrm{NN}_{T_3}(x)\,, \\
\Delta_S(x,Q_0^2) &= A_{\Delta_S} (1-x)^{m_{\Delta_S}} x^{-n_{\Delta_S}} \mathrm{NN}_{\Delta_S}(x)\,, \\
g(x,Q_0^2) &= A_g (1-x)^{m_g} x^{-n_g} \mathrm{NN}_g(x)\,, \\
s^+(x,Q_0^2) &= (1-x)^{m_{s^+}}\, x^{-n_{s^+}} \mathrm{NN}_{s^+}(x)\,, \\
s^-(x,Q_0^2) &= (1-x)^{m_{s^-}}\, x^{-n_{s^-}} \mathrm{NN}_{s^-}(x) - s_{\mathrm{aux}}(x,Q_0^2),
\end{aligned} \tag{3.6}$$

where

$$s_{\mathrm{aux}}(x,Q_0^2) = A_{s^-} \left[x^{r_{s^-}}\, (1-x)^{t_{s^-}} \right]. \tag{3.7}$$

The $\mathrm{NN}(x)$ functions represent the neural networks that are trained during the fit. The preprocessing exponents m, n, are randomly variated on a range given in Table 3.4. The normalization factors A_V, A_{Δ_S}, and A_g are constrained by the sum rules and are discussed in Sect. 3.4. A simple test to check that preprocessing is not introducing a bias is to remove it. For a sufficiently long training the results must be stable. A more sofisticated tool is given in Ref. [87] by the computation of the correlation between a preprocessing coefficient and the χ^2 of the relative PDF computed between the k-th net (trained on

PDF	$[m_{\min}, m_{\max}]$	$[n_{\min}, n_{\max}]$	$r\left[\chi^2, m\right]$	$r\left[\chi^2, n\right]$
$\Sigma(x, Q_0^2)$	$[2.55, 3.45]$	$[1.05, 1.35]$	-0.018	0.131
$g(x, Q_0^2)$	$[1.05, 1.35]$	$[1.05, 1.35]$	-0.002	0.050
$T_3(x, Q_0^2)$	$[2.55, 3.45]$	$[0, 0.5]$	-0.023	-0.130
$V_T(x, Q_0^2)$	$[2.55, 3.45]$	$[0, 0.5]$	0.003	-0.068
$\Delta_S(x, Q_0^2)$	$[12, 14]$	$[-0.95, -0.65]$	0.000	-0.069
$s^+(x, Q_0^2)$	$[2.55, 3.45]$	$[1.05, 1.35]$	0.021	-0.055
$s^-(x, Q_0^2)$	$[2.55, 3.45]$	$[0, 0.5]$	-0.027	-0.015

Table 3.4: The range of random variation of the large-x and small-x preprocessing exponents m and n used in the present analysis (the precise form of these exponents is given in Sect. 3.1 of Ref. [49]). The last two columns give the correlation coefficient Eq. (3.8) between the χ^2 and respectively the large- and small-x preprocessing exponents.

replica k) and experimental data, defined as

$$r\left[\chi^2, m_\Sigma\right] \equiv \frac{\langle \chi^2 m_\Sigma \rangle_{\rm rep} - \langle \chi^2 \rangle_{\rm rep} \langle m_\Sigma \rangle_{\rm rep}}{\sigma^2_{m_\Sigma}} \, . \tag{3.8}$$

The results of this test are given in the last column of Tab. 3.4.

3.4 Sum Rules and Positivity Constraints

While performing a parton fit, it is important to guarantee that only the subspace of acceptable physical solutions is explored by the fitting procedure. It is possible to constrain the space of acceptable solutions by implementing some general features, as for example sum rules and positivity.

Sum rules follow from conservation laws, and read

$$\int_0^1 dx \, x \left[\Sigma(x) + g(x)\right] = 1 \tag{3.9}$$

for the momentum sum rule, and

$$\int_0^1 dx \, \left(u(x) - \bar{u}(x)\right) = 2 \, , \quad \int_0^1 dx \, \left(d(x) - \bar{d}(x)\right) = 1 \tag{3.10}$$

for the valence sum rules. The momentum sum rule Eq. (3.9) fixes the factor A_g as

$$A_g = \frac{1 - \int_0^1 dx \, x \left[(1-x)^{m_\Sigma} {\rm NN}_\Sigma(x)/x^{n_\Sigma}\right]}{\int_0^1 dx \, x \left[(1-x)^{m_g} {\rm NN}_g(x)/x^{n_g}\right]} \, , \tag{3.11}$$

while the valence sum rules Eq. (3.10) fix the factors A_V and A_{Δ_S} as

$$\begin{aligned} A_V &= \frac{3}{\int_0^1 dx\ [(1-x)^{m_V}\mathrm{NN}_V(x)/x^{n_V}]}\,, \\ A_{\Delta_S} &= \frac{1-\int_0^1 dx\ [(1-x)^{m_{T_3}}\mathrm{NN}_{T_3}(x)/x^{n_{T_3}}]}{2\int_0^1 dx\ [(1-x)^{m_{\Delta_S}}\mathrm{NN}_{\Delta_S}(x)/x^{n_{\Delta_S}}]}\,. \end{aligned} \tag{3.12}$$

These rules are satisfied at the initial evolution scale Q_0^2 as an imposed condition. Evolution will preserve this condition at each value of Q^2.

This is almost the same strategy adopted for positivity constraints. Of course all physical cross-sections has to be positive. Positivity needs to be imposed only on hadronic cross-sections and not on partonic quantities (only at LO subsists a probabilistic interpretation for PDFs) [185]. For the NLO case, additional pseudo-datasets are defined for physical cross-sections with extremely small uncertainties in such a way that negative cross-sections would lead to a very large contribution to the χ^2 (Lagrangian multipliers method). All the positivity constraints are implemented at a low scale $Q^2_{\rm pos}$ chosen to be $Q^2_{\rm pos} = 2$ GeV^2, in the range $x \in [10^{-6}, x_{\rm max}]$, where $x_{\rm max}$ is the corresponding kinematical boundary, $x_{\rm max} \sim 0.1$ for NC scattering and $x_{\rm max} \sim 0.5$ for CC scattering. The evolution (DGLAP) then takes care of preserving the positivity properties for higher scales, as for sum rules.

In NNPDF2.1 NLO, positivity of the longitudinal structure function $F_L(x, Q^2)$ is imposed, which constrains the gluon positivity at small-x, of the charm production cross-section in neutrino DIS, ($d^2\sigma^{\nu,c}/dxdy$ [49]), which constrains the strange PDFs both at large and at small-x, beyond the reach of existing data, and of the neutral current DIS charm structure function $F_2^c(x, Q^2)$, useful to impose the positivity of the gluon at very large-x, where it is not constrained by any experimental dataset.

The physical observables for the pseudo-data that implement the positivity constraints are computed consistently at the same perturbative order as all other physical observables, in the present case next–to–leading order perturbative QCD.

The same procedure with Lagrangian multipliers is applied in the case of NNPDF2.1 NNLO. The same could be done at LO, where as already said PDFs have a probabilistic interpretation and it is possible to impose the positivity constraints directly on PDFs. Instead of using the same method with the generation of pseudo-data explained above, it is possible to obtain the same result with a consequent reduction of the fitting time just implementing a minor modification in the PDF parametrization. Within the neural network PDF parametrization which is adopted, this can be done as follows. Recall that

in the architecure that is adopted for neural networks, the response function is a sigmoid

$$\xi_i^{(l)} = g\Big(\sum_j \omega_{ij}^{(l)} \xi_j^{(l-1)} - \theta_i^{(l)}\Big), \qquad g(x) = \frac{1}{1+e^{-x}}\,, \tag{3.13}$$

for hidden layer, but it is linear in the last layer. For the LO fits, instead, for the last layer a quadratic response function

$$\xi_i^{(n_l)} = \Big(\sum_j \omega_{ij}^{(n_l)} \xi_j^{(n_l-1)} - \theta_i^{(n_l)}\Big)^2 \tag{3.14}$$

is adopted. The output of the neural network, and thus the PDF, is then guaranteed to be non-negative.

The basis of PDFs that are parametrized by neural networks in NNPDF fits (Sect. 1.5) includes the gluon, quark singlet, and various other linear combinations of quark PDFs. Of these, only the gluon and singlet must be positive, since all other combinations contain differences of PDFs. However, in practice also the total valence and isospin triplet combination are positive definite. Hence the parametrization Eqs. (3.13)-(3.14) for simplicity is adopted for all PDFs: partonic functions other than singlet, gluon, valence and triplet are allowed to change sign by simply adding to the above form a constant shift. It turns out that with the constraints from the data, this is sufficient in practice to ensure positivity of all PDFs: has been checked a posteriori that for every replica the gluon and all individual quark and antiquark flavors are positive for all values of x and Q^2 for which the NNPDF2.1 LO PDFs are provided.

3.5 Genetic Algorithm Minimization

A reasonable minimum in a very large parameter space and with a figure of merit that is a nonlocal functional of the set of functions that are being determined in the minimization needs to be find. This problem can be efficiently solved using a genetic algorithm.

Starting from a randomly chosen set of parameters, the genetic algorithm generates a pool of possible new sets by mutation of one or more parameters at a time. Each new set that has undergone mutation is a *mutant*. To each set of parameters corresponds a value of the error function that is being minimized, and so those configurations that fall far away from the minimum can be discarded. This procedure is iterated over a sufficiently large number of generations.

The error function to be minimized has the form [84]

$$E^{(k)} = \frac{1}{N_{\rm dat}} \sum_{I,J=1}^{N_{\rm dat}} \left(F_I^{(\rm art)(k)} - F_I^{(\rm net)(k)}\right) \left((\rm cov_{t_0})^{-1}\right)_{IJ} \left(F_J^{(\rm art)(k)} - F_J^{(\rm net)(k)}\right) \ , \quad (3.15)$$

where $F_I^{(\rm art)(k)}$ is the value of the observable F_I at the kinematic point I corresponding to the Monte Carlo replica k, and $F_I^{(\rm net)(k)}$ is the same observable computed from the neural network PDFs, and where the t_0 covariance matrix $\rm cov_{t_0}$ has been defined in Eq. (2.40). For a more detailed explanation on the application of genetic algorithm minimization to PDFs fitting see Ref. [84]. Recent NNPDF analyses have to deal with many different experimental datasets. In the following some more sofisticated features needed in this case with respect to a standard one are discussed.

Targeted Weighted Training

Due to the fact that several different datasets are present in the fit, often with very different chatacteristics, it is likely that during a standard fit the minimization over a specific dataset is much faster than over the other datasets. Of course it may happen also the opposite situation, in which the training of a specific dataset is much slower than all the others. In short, the training of more datasets at a time can well be disomogeneous. This results not only in a very inefficient training, but also in a final uneven figure of merit $E^{(k)}$. A possible solution is to introduce a dynamical weighted fitting technique.

The main idea is to define a weight for each dataset that gets updated during the fit. In a first epoch of the training the error function is modified like this:

$$E_{\rm wt}^{(k)} = \frac{1}{N_{\rm dat}} \sum_{j=1}^{N_{\rm sets}} p_j^{(k)} N_{{\rm dat},j} E_j^{(k)} \ , \quad (3.16)$$

where $E_j^{(k)}$ is the error function in Eq. (3.15) restricted to the dataset j, $N_{{\rm dat},j}$ is the number of points of this dataset and $p_j^{(k)}$ are weights associated to this dataset which are adjusted dynamically. These weights are defined in relation to target values $E_i^{\rm targ}$ for the figure of merit, chosen for each single experiment. Then for each set the weights are defined as

$$p_i^{(k)} = \left(E_i^{(k)}/E_i^{\rm targ}\right)^2 \qquad \text{if} \qquad E_i^{(k)} \geq E_i^{\rm targ} \ , \quad (3.17)$$

while are set to zero ($p_i^{(k)} = 0$) if $E_i^{(k)} < E_i^{\rm targ}$. This procedure has the effect of giving a larger weight to that datasets that are far above their target value, so that the training on them is more intense, and to remove from the training that datasets that already have reached a good minimization. This avoids to waste training resources on datasets that are

already well trained.

The target values $E_i^{\rm targ}$ are determined for all sets with an iterative procedure starting with all $E_i^{\rm targ} = 1$ and going through a first very long fit. The resulting outcome of the fit is used for a first set of $E_i^{\rm targ}$ values. This procedure is iterated until convergence. More details are given in Ref. [87].

The dynamical targeted weighted training only lasts a fixed number of generations $N_{\rm gen}^{\rm wt}$. Only if the error function of a specific replica is above a threshold $E^{(k)} \geq E^{\rm sw}$ then the weighted training is not turned off until $E^{(k)}$ goes below the threshold. The last part of the training then is without weights and the unweighted error function is again Eq. (3.15), computed on experiments (the weighted one was computed on single datasets). This last epoch is important to eliminate any possible residual bias introduced by the $E_i^{\rm targ}$ weighted minimization. Using this procedure the fit is faster and much more uniform.

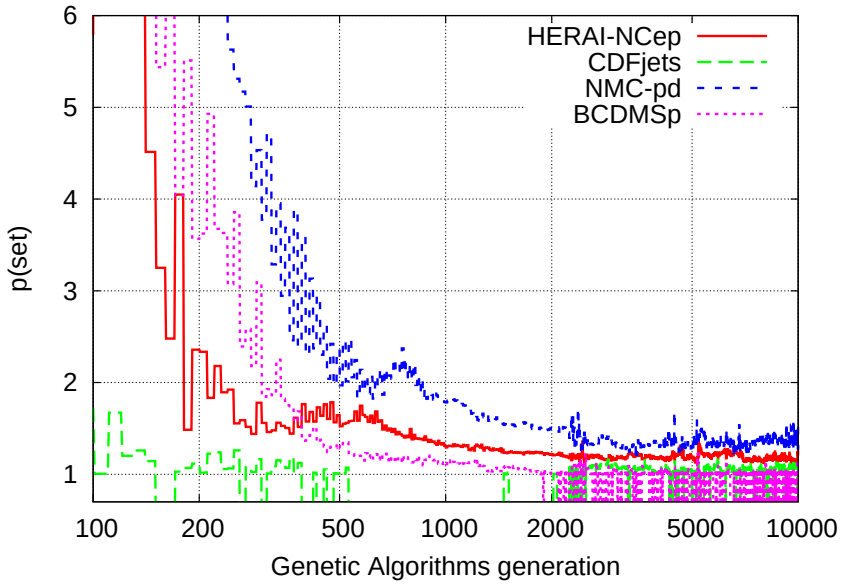

Figure 3.4: Illustration of the weighted training in one particular replica. Individual weights for each dataset converge to a value of p_i which is close to 1 as the training progresses. Only the behaviour of representative datasets is shown.

3.6 Dynamical Stopping

To determine the best fit, the training has to be stopped at a point in which the fit reproduces the underlying physical law contained in the information given by the data with sufficient precision but not the statistical fluctuations of the same data. To this extent, it is necessary to define a dynamical stopping criterion. A series of tests over the conditions for dynamical stopping is turned on only after weighted training has instead turned off,

that is after $N_{\rm gen}^{\rm wt}$ generations. The stopping is applied on the training of each replica, and it is based on the cross-validation method, widely used in the context of neural network training [186]. This method consists in dividing the dataset in two subsets: over one of them the fit is regularly performed (training subset), while the other (validation subset) is used only to compute the error function and validate the result given by the training subset. While the error function computed over the trained points is by definition monotonically non-increasing, the same quantity computed over the validation points would increase when the overlearning regime sets in. The application of the cross-validation method to this case has been described into detail in Refs. [83, 84, 87].

The conditions for dynamical stopping are three:

- the fit must be, as already said, out of the targeted weighted training epoch
- all experiments must have an error function below some reasonable threshold $E_{\rm thres}$
- a moving average for training ($r_{\rm tr}$) and for validation ($r_{\rm val}$), defined below, must satisfy

$$r_{\rm tr} > 1 - \delta_{\rm tr}\,, \quad r_{\rm val} > 1 + \delta_{\rm val}\,. \tag{3.18}$$

In this last condition, the moving averages are defined as

$$r_{\rm tr} \equiv \frac{\langle E_{\rm tr}(i)\rangle}{\langle E_{\rm tr}(i-\Delta_{\rm smear})\rangle}\,, \tag{3.19}$$

$$r_{\rm val} \equiv \frac{\langle E_{\rm val}(i)\rangle}{\langle E_{\rm val}(i-\Delta_{\rm smear})\rangle}\,, \tag{3.20}$$

where the $\Delta_{\rm smear}$ functions are given by

$$\langle E_{\rm tr,val}(i)\rangle \equiv \frac{1}{N_{\rm smear}} \sum_{l=i-N_{\rm smear}+1}^{i} E_{\rm tr,val}(l)\,, \tag{3.21}$$

with $E_{\rm tr,val}(l)$ the value of the error function Eq. (3.15) evaluated for the iteration l of the genetic algorithm and over the training and validation subsets. In practice, for a variation between an iteration and another of the values $r_{\rm tr}$ and $r_{\rm val}$ such that the first is decreased and the second is enhanced by a quantity that exceeds the respective sensitivities $\delta_{\rm tr}$ and $\delta_{\rm val}$, if the first two conditions are satisfied then the fit stops. The two parameters $\delta_{\rm tr}$ and $\delta_{\rm val}$ set the accuracy to which the increase and decrease is required in order to be significant (in order not to confuse a fluctuation with a real increase in validation error function). The tuning of the dynamical stopping parameters is performed into detail in Ref. [87].

$N_{\rm smear}$	$\Delta_{\rm smear}$	$\delta_{\rm tr}$	$\delta_{\rm val}$	$E_{\rm thres}$	$N_{\rm gen}^{\rm max}$
200	200	10^{-4}	$3\,10^{-4}$	6	30000

Table 3.5: Parameter values for the stopping criterion.

The final values of the determined parameters are listed in Table 3.5. In order to avoid unacceptably long fits, when a very large number of iterations $N_{\rm gen}^{\rm max}$ is reached (see Table 3.6) training is stopped anyway. This leads to a small loss of accuracy of the corresponding fits, which is acceptable provided it only happens for a small fraction of replicas.

Genetic Algorithm Parameters

As in the case discussed above for dynamical stopping, also the genetic algorithm is characterized by a certain number of parameters that need to be tuned in order to improve its efficiency and, as a consequence, that of the whole training.

A mutant as a set of parameters associated to the neural network that is being trained has already been introduced. More mutants are obtained from another set of parameters by introducing mutations on them. As can be seen in Table 3.6, each PDF has a fixed number of mutations $N_{\rm mut}$ that can modify its set of parameters at each generation. To each mutation it is associated a mutation rate, that is dynamically adjusted as a function of the iterations $N_{\rm ite}$ in this way:

$$\eta_{i,j} = \eta_{i,j}^{(0)} / N_{\rm ite}^{r_\eta} \,, \tag{3.22}$$

where $\eta_{i,j}^{(0)}$ is the initial mutation rate of the PDF i and the mutation j, $N_{\rm ite}$ is the number of iterations already cycled, and r_η is an exponent randomly initialized at each iteration between 0 and 1 that allows to span the range of all possible beneficial mutations. Looking iteration by iteration which is the value of this exponent for the accepted mutations reveals that the size of the mutation doesn't depend on the stage of the training. Both large and small mutations are accepted with the same frequency independently of $N_{\rm ite}$.

At each iteration, a certain number of mutants is produced to have a pool of parameter configurations from which it is possible to choose the best individual. Before a certain iteration $N_{\rm gen}^{\rm mut}$, the number of generated mutants corresponds to $N_{\rm mut}^a \gg 1$, in order to have a large population in the pool and better explore a parameter space as large as possible. In a more advanced stage of the training it is more useful to reduce this number so that $N_{\rm mut}^b \ll N_{\rm mut}^a$ to help with a reduced population to propagate the beneficial mutations.

The final choices of parameters of the genetic algorithm which have been adopted in the NNPDF2.1 parton determination are summarized in Table 3.6. While at NLO they

	$N_{\rm gen}^{\rm wt}$	$N_{\rm gen}^{\rm mut}$	$N_{\rm gen}^{\rm max}$	$E^{\rm sw}$	$N_{\rm mut}^{a}$	$N_{\rm mut}^{b}$
LO & NLO	10000	2500	30000	2.6	80	10
NNLO	10000	2500	30000	2.3	80	30

	LO & NLO		NNLO	
PDF	$N_{\rm mut}$	$\eta^{\rm k}$	$N_{\rm mut}$	η^{k}
$\Sigma(x)$	2	10,1	2	10,1
$g(x)$	2	10,1	3	10,3,0.4
$T_3(x)$	2	1,0.1	2	1,0.1
$V(x)$	2	1,0.1	3	8,1,0.1
$\Delta_S(x)$	2	1,0.1	3	5,1,0.1
$s^+(x)$	2	5,0.5	2	5,0.5
$s^-(x)$	2	1,0.1	2	1,0.1

Table 3.6: Parameter values for the genetic algorithm for the NNLO fits compared to those of the LO and NLO fits (top). The number of mutations and the values of the mutation rates for the individual PDFs in the NNLO fit as compared to the values of the LO and NLO fits are also given (bottom).

are the same as in Ref. [49, 84, 87], the poorer quality of the LO fit on the one hand, and the greater complexity of NNLO coefficient functions on the other hand, require some retuning of the parameters of the minimization algorithm in these two cases.

At leading order, the best-fit value of the figure of merit $E^{(k)}$ which is being minimized for each replica (which is essentially the χ^2 of the fit of each PDF replica to the given data replica) is on average rather larger than in an NLO fit, because of the poorer accuracy of the LO theory. This is particularly true for the Drell-Yan observables, which have large NLO corrections with a K-factor of order two. As a consequence, the minimum value that $E^{(k)}$ must reach for each experiment in order for the fit to stop has been increased from $E_{\rm th} = 6$ to $E_{\rm th}^{\rm DY} = 12$ for all Drell-Yan experiments. Furthermore, the cross-validation method that is used to determine the optimal fit stops the minimization when the moving average (over iterations of the genetic algorithm) of $E^{(k)}$ increases more than a fixed percentage threshold value r_v, larger than a typical random fluctuation. Because the size of fluctuations of $E^{(k)}$ remains fixed, while its value at best fit has increased, the typical values of r_v are smaller at LO, and thus it turns out to be necessary to reduce the value of r_v required for stopping to $r_v - 1 = 2 \cdot 10^{-4}$, from $r_v - 1 = 3 \cdot 10^{-4}$ used at NLO.

Even with these adjustments, for a sizable fraction of replicas the cross-validation algorithm fails to stop dynamically the minimization even after a large number of generations of the genetic algorithm. This reflects the poor accuracy of LO theory, and it could only be obviated by letting the genetic minimization run much longer. In view of the large theoretical uncertainties inherent to any LO PDF determination, as a practical

compromise replicas that do not stop dynamically after 50000 iterations of the genetic algorithm are discarded, retaining only those replicas for which the stopping criterion was fulfilled. It has been checked that this leads to no significant statistical bias.

At next-to-next-to-leading order the partonic cross-sections have rather more structure than at lower orders, both because of the opening of new partonic channels and because of the appearance of new transcendental functions in the perturbative results (such as higher order harmonic sums). This results in somewhat more complex PDF shapes. As a consequence, it turns out to be necessary to increase the number of mutants and mutations per PDF in the genetic minimization in order to fully explore this more complex space of minima. The NNLO settings for the genetic algorithm used for minimization are summarized in Table 3.6 and compared to those used at LO and NLO.

3.7 The FastKernel Method

In this Section the FastKernel method is briefly described. The main idea is presented in the case of PDFs evolution. A detailed description and benchmark of the method for the construction of the observables is developed in Ref. [87].

The computation of physical observables requires both the computation of PDFs at the initial scale and solving DGLAP evolution equations. In particular, this second step implies the computation of convolutions between PDFs and hard scattering cross-sections. To compute convolutions is a time consuming procedure, and in the case of hadronic observables, where two PDFs are involved - one from each colliding nucleon - a double convolution need to be computed. As developed in Refs. [83,84] and recalled in Sect. 1.5, to solve DGLAP equations it is possible to pre-compute a Green function that can be determined in the N space and that takes a PDF from its initial scale to the scale of physical measurements. This function can be also pre-combined with the hard scattering cross-sections into the definition of an appropriate evolution kernel. This method allows to reduce the computation of any observable to performing a convolution (or two, for hadronic observables).

However, despite this simplification, to compute physical observables is a lengthy process, due to these convolutions. For this reason, in several global PDF fits the computation of Drell-Yan processes is treated using NLO and NNLO K-factor approximation [81, 187]. A method to exactly compute Drell-Yan and collider weak boson production was first proposed in Ref. [87], and is based on analogous ideas as the ones on which are based tools like FastNLO [188] for the case of jet production, APPLGRID [189], or the fast x-space DGLAP evolution code HOPPET [190]. The original idea on which all these tools are based was proposed in Ref. [191].

In Sect. 1.5 has been already introduced a basis in terms of which PDFs can be written:

$$f_j = \{\Sigma, g, V, V_3, V_8, V_{15}, V_{24}, V_{35}, T_3, T_8, T_{15}, T_{24}, T_{35}\}. \tag{3.23}$$

The independent contributions are only seven because heavy intrinsic flavors are not being considered. In this case only the six lightest quark-antiquark contributions and the gluon need to be parametrized.

Equation (1.99) gives

$$f_j(x_I, Q_I^2) = \sum_{k=1}^{N_{\rm pdf}} \int_{x_I}^1 \frac{dy}{y} \Gamma_{jk}\left(\frac{x_I}{y}, Q_0^2, Q_I^2\right) f_k(y, Q_0^2), \tag{3.24}$$

where Γ_{jk} is the matrix of DGLAP evolution kernels and (x_I, Q_I^2) defines the kinematics of a given experimental point. The index I refers both to the kinematic variables which define an experimental point (x, Q^2) and the type of observable. The expression in Eq. (3.24) describes the evolution of a PDF from an initial scale Q_0^2 to the physical scale Q_I^2. The integration in Eq. (3.24) was performed in Ref. [84] by means of Gaussian integration. Here the FastKernel method is first introduced in a simpler (but slower and less accurate) case that relies on triangular basis functions and then is generalized to Hermite cubic functions.

A grid in x that is independent of the experimental point x_I is considered. The grid is defined by a set of points

$$x_{\rm min} \equiv x_1 < x_2 < ... < x_{N_x - 1} < x_{N_x} \equiv 1\ ,$$

that is labeled as x_α with $\alpha = 1, ..., N_x$. It is now possible to define a basis of interpolating functions $\mathcal{I}^{(\alpha)}$ such that

$$\begin{aligned} \mathcal{I}^{(\alpha)}(x_\alpha) &= 1 \\ \mathcal{I}^{(\alpha)}(x_\beta) &= 0\,, \beta \neq \alpha \\ \sum_{\alpha=1}^{N_x} \mathcal{I}^{(\alpha)}(y) &= 1\,, \forall y. \end{aligned} \tag{3.25}$$

An example for such a basis is represented by the triangular functions plotted in Fig. 3.5 where, for any y, only two triangular functions are non zero and their sum is always equal to one. A triangular function $E^{(\alpha)}$ is centered in x_α, where its value is equal to one, and is zero outside the interval $(x_{\alpha-1}, x_{\alpha+1})$.

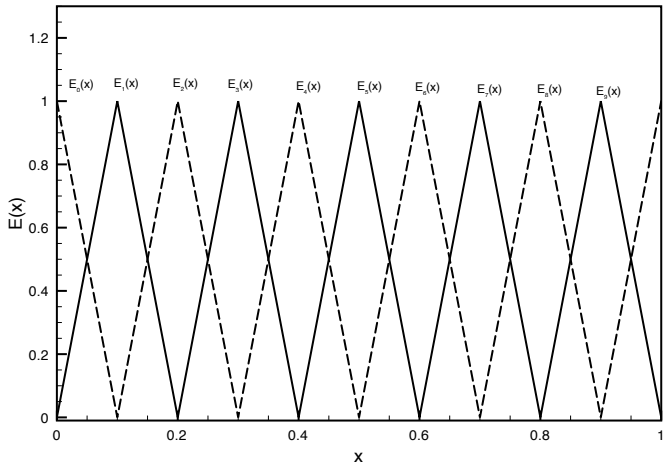

Figure 3.5: Set of interpolating triangular basis functions.

At the initial scale a PDF can be rewritten using the approximation

$$f_k(y,Q_0^2) \equiv f_k^0(y) = \sum_{\alpha=1}^{N_x} f_k^0(x_\alpha)\,\mathcal{I}^{(\alpha)}(y) + \mathcal{O}[(x_{\alpha+1}-x_\alpha)^p]\,, \tag{3.26}$$

given by the general set introduced above. Here p is the lowest order neglected in the interpolation. As triangular functions are linear, in this case $p=2$. It is possible to write in a more explicit way Eq. (3.24) like

$$\begin{aligned} f_j(x_I,Q_I^2) \equiv f_j(x_I) &= \sum_{k=1}^{N_{\rm pdf}}\sum_{\alpha=1}^{N_x} f_k^0(x_\alpha)\int_{x_I}^1 \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\mathcal{I}^{(\alpha)}(y) + \mathcal{O}[(x_{\alpha+1}-x_\alpha)^p] \\ f_j(x_I) &= \sum_{k=1}^{N_{\rm pdf}}\sum_{\alpha=1}^{N_x} \hat{\sigma}_{\alpha k}^{Ij}\, f_k^0(x_\alpha) + \mathcal{O}[(x_{\alpha+1}-x_\alpha)^p], \end{aligned} \tag{3.27}$$

where

$$\hat{\sigma}_{\alpha k}^{j}(x_I,Q_0^2,Q_I^2) \equiv \hat{\sigma}_{\alpha k}^{Ij} = \int_{x_I}^1 \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\mathcal{I}^{(\alpha)}(y). \tag{3.28}$$

The dependence on Q has been dropped on the right hand part of the equations. The indeces j, k, run over the PDFs, α runs over the x-grid points and I refers, as already said, to the experimental point. It is visible from Eq. (3.27) the convenience of this approach: the $\hat{\sigma}_{\alpha k}^{Ij}$ coefficients can be precomputed for each point I and stored. Then the computation of the integral in Eq. (3.24) can be performed by only evaluate N_x times the PDFs, independently of the point at which the evolved PDFs are needed. This reduces the computational effort, but things gets still better by using a more sofisticated basis.

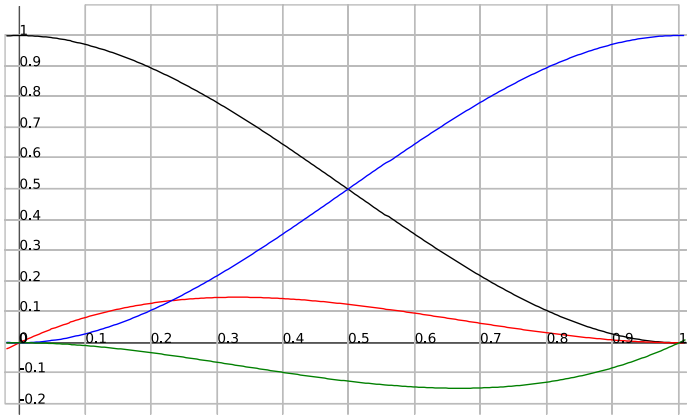

Figure 3.6: Set of interpolating Hermite cubic functions in the [0,1] interval.

In Ref. [87] a basis of cubic Hermite interpolating functions is introduced, defined as

$$\begin{aligned} h_{00}(t) &= 2t^3 - 3t^2 + 1 = (1+2t)(1-t)^2 \\ h_{10}(t) &= t^3 - 2t^2 + t = t(t-1)^2 \\ h_{01}(t) &= -2t^3 + 3t^2 = t^2(3-2t) \\ h_{11}(t) &= t^3 - t^2 = t^2(t-1) \end{aligned} \tag{3.29}$$

and represented in Fig. 3.6.

Using this new basis, it is possible to rewrite a PDF as

$$\begin{aligned} f_k^0(y) &= h_{00}(t) f_k^0(x_\alpha) + h_{10}(t) h_\alpha m_\alpha + h_{01}(t) f_k^0(x_{\alpha+1}) + h_{11}(t) h_\alpha m_{\alpha+1} \\ &\quad + \mathcal{O}[(x_{\alpha+1} - x_\alpha)^4], \end{aligned}$$

where

$$h_\alpha = g(x_{\alpha+1}) - g(x_\alpha), \qquad t = \frac{g(y) - g(x_\alpha)}{h_\alpha}, \tag{3.30}$$

with $g(y)$ a monotonic function in [0,1] which determines the distribution of points in the interval (linear, logarithmic, etc.). Moreover,

$$m_\alpha = \begin{cases} \frac{f_k^0(x_\alpha) - f_k^0(x_{\alpha-1})}{2h_{\alpha-1}} + \frac{f_k^0(x_{\alpha+1}) - f_k^0(x_\alpha)}{2h_\alpha}, & \text{for } 2 \le \alpha \le N_x - 1 \\ \frac{f_k^0(x_{\alpha+1}) - f_k^0(x_\alpha)}{h_\alpha}, & \text{for } \alpha = 1 \\ \frac{f_k^0(x_\alpha) - f_k^0(x_{\alpha-1})}{h_{\alpha-1}}, & \text{for } \alpha = N_x \end{cases} \tag{3.31}$$

are derivatives of the interpolated function, at the right hand side of the interval (m_α) or

at the left hand side ($m_{\alpha+1}$). It is convenient to rewrite the expression in Eq. (3.30) like

$$\begin{aligned} f_k^0(y) &= f_k^0(x_{\alpha-1})\,A^{(\alpha)}(y) + f_k^0(x_\alpha)\,B^{(\alpha)}(y) + f_k^0(x_{\alpha+1})\,C^{(\alpha)}(y) \\ &\quad + f_k^0(x_{\alpha+2})\,D^{(\alpha)}(y) + \mathcal{O}[(x_{\alpha+1}-x_\alpha)^4]\,, \end{aligned} \tag{3.32}$$

where

$$A^{(\alpha)}(y) = \begin{cases} 0, & \text{for } \alpha = 1 \\ -h_{10}(t)\frac{h_\alpha}{h_{\alpha-1}}, & \text{for } \alpha \neq 1 \end{cases} \tag{3.33}$$

$$B^{(\alpha)}(y) = \begin{cases} h_{00}(t) - h_{10}(t) - \frac{h_{11}(t)}{2}, & \text{for } \alpha = 1 \\ h_{00}(t) - \frac{h_{10}(t)}{2}\left(1 - \frac{h_\alpha}{h_{\alpha+1}}\right) - h_{11}(t), & \text{for } \alpha = N_x - 1 \\ h_{00}(t) - \frac{h_{10}(t)}{2}\left(1 - \frac{h_\alpha}{h_{\alpha+1}}\right) - \frac{h_{11}(t)}{2}, & \text{for } \alpha \neq 1, N_x - 1 \end{cases}$$

$$C^{(\alpha)}(y) = \begin{cases} h_{01}(t) + \frac{h_{11}(t)}{2}\left(1 - \frac{h_\alpha}{h_{\alpha+1}}\right) + h_{10}(t), & \text{for } \alpha = 1 \\ h_{01}(t) + h_{11}(t) + \frac{h_{10}(t)}{2}, & \text{for } \alpha = N_x - 1 \\ h_{01}(t) + \frac{h_{11}(t)}{2}\left(1 - \frac{h_\alpha}{h_{\alpha+1}}\right) + \frac{h_{10}(t)}{2}, & \text{for } \alpha \neq 1, N_x - 1 \end{cases}$$

$$D^{(\alpha)}(y) = \begin{cases} 0, & \text{for } \alpha = N_x - 1 \\ h_{11}(t)\frac{h_\alpha}{2h_{\alpha+1}}, & \text{for } \alpha \neq N_x - 1\,. \end{cases}$$

Gathering all the parts defined up to now in Eq. (3.24), the following contributions from the $\hat{\sigma}$ coefficients can be written:

$$\hat{\sigma}_{\alpha k}^{Ij} = \begin{cases} \int_{x_I}^{x_{c+1}} \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\,A^{(c)}(y), & \text{for } \alpha = c, \\ \int_{x_I}^{x_{c+1}} \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\,B^{(c)}(y) & \\ \quad + \theta(N_x - (c+2)) \int_{x_{c+1}}^{x_{c+2}} \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\,A^{(c+1)}(y), & \text{for } \alpha = c+1, \\ \int_{x_I}^{x_{c+1}} \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\,C^{(c)}(y) & \\ \quad + \theta(N_x - (c+2)) \int_{x_{c+1}}^{x_{c+2}} \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\,B^{(c+1)}(y) & \\ \quad + \theta(N_x - (c+3)) \int_{x_{c+2}}^{x_{c+3}} \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\,A^{(c+2)}(y), & \text{for } \alpha = c+2, \\ \theta(N_x - (I-1)) \int_{x_{\alpha-2}}^{x_{\alpha-1}} \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\,D^{(\alpha-1)}(y) & \\ \quad + \theta(N_x - \alpha) \int_{x_{\alpha-1}}^{x_\alpha} \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\,C^{(\alpha-1)}(y) & \\ \quad + \theta(N_x - (\alpha+1)) \int_{x_\alpha}^{x_{\alpha+1}} \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\,B^{(\alpha)}(y) & \\ \quad + \theta(N_x - (\alpha+2)) \int_{x_{\alpha+1}}^{x_{\alpha+2}} \frac{dy}{y}\,\Gamma_{jk}\left(\frac{x_I}{y}\right)\,A^{(\alpha+1)}(y), & \text{for } c+3 \leq \alpha \leq N_x + 1, \\ 0 & \text{for } \alpha < c, \end{cases} \tag{3.34}$$

with c the index such that

$$x_c \leq x_I < x_{c+1}.$$

Even if at first sight these expressions seem more complicated, they are simpler to pre-compute. It can be observed that while in the previous case for each point y the computation of only two contributions from the interpolating basis was needed, in the present case is needed the computation of four contribution. The gain in accuracy (now terms of the fourth order are neglected and not just of the second as before) allows for a much less dense grid, that corresponds to a sensible speed up in computations.

A last issue needs to be addressed: when $y \sim x_I$ the precomputed coefficients need to be regularized, using the same subtraction used in Ref. [84]: considering the case of the first integral of $A^{(\alpha)}(y)$, it can be written

$$\begin{aligned}
&\int_{x_I}^{x_{c+1}} \tfrac{dy}{y}\, \Gamma_{jk}\left(\tfrac{x_I}{y}\right)\, A^{(c)}(y) \\
&\quad = \int_{x_I}^{x_{c+1}} \tfrac{dy}{y}\, \Gamma_{jk}\left(\tfrac{x_I}{y}\right)\, \left(A^{(c)}(y) - \tfrac{x_I}{y} A^{(c)}(x_I)\right) + A^{(c)}(x_I) \int_{x_I}^{x_{c+1}} \tfrac{dy}{y^2}\, \Gamma_{jk}\left(\tfrac{x_I}{y}\right) \\
&\quad = \int_{x_I}^{x_{c+1}} \tfrac{dy}{y}\, \Gamma_{jk}\left(\tfrac{x_I}{y}\right)\, \left(A^{(c)}(y) - \tfrac{x_I}{y} A^{(c)}(x_I)\right) + A^{(c)}(x_I) \int_{x_I/x_{c+1}}^{1} dz\, \Gamma_{jk}(z) \\
&\quad = \int_{x_I}^{x_{c+1}} \tfrac{dy}{y}\, \Gamma_{jk}\left(\tfrac{x_I}{y}\right)\, \left(A^{(c)}(y) - \tfrac{x_I}{y} A^{(c)}(x_I)\right) \\
&\qquad\qquad + A^{(c)}(x_I)\left[\Gamma_{jk}(N)\big|_{N=2} - \int_0^{x_I/x_{c+1}} dz\, \Gamma_{jk}(z)\right].
\end{aligned} \tag{3.35}$$

In this way all the $\hat{\sigma}$ coefficients can be regularized and stored, so that their computation is only needed once for each experimental point, thanks to the fact that they do not depend on PDFs at the initial scale.

The accuracy of a 50 point grid results in a $\mathcal{O}(10^{-5})$ discrepancy in a benchmark against the Les Houches tables given in Ref. [192], that is beyond the accuracy needed for precision phenomenology at LHC. This is the main result of a more detailed study given in Ref. [87].

3.8 New Data Inclusion by Reweighting

The features of the NNPDF approach discussed in this Chapter allow for an interesting exploitation of Bayesian inference to determine the impact of new datasets.

How an existing probability distribution in the space of PDFs may be updated with information from new data shall be discussed. To include the new data, one can of course perform a fit with the new, enlarged dataset. However this is a time consuming task, particularly for observables where no fast code is available. It is therefore desirable to have a faster method of including new data in order to assess it's impact rapidly without

the need for a full refit. NNPDF parton sets are supplied as an ensemble of $N = N_{\rm rep}$ parton distribution replicas $\mathcal{E}$, representing the probability density in PDFs $\mathcal{P}(f)$ based upon the data in the existing fit. It is therefore possible to include new data by weighting each replica f_k in the ensemble by an associated weight w_k. If the replica weights are computed correctly, then reweighting is completely equivalent to a refit.

In order to illustrate the reweighting method, the computation of the expected value of a PDF-dependent observable $\mathcal{F}[f]$ is considered. As the NNPDF Monte Carlo ensemble is a good representation of the probability density $\mathcal{P}(f)$, the expectation value $\langle \mathcal{F}[f] \rangle$ can be calculated as a simple average,

$$\langle \mathcal{F} \rangle = \int \mathcal{F}[f]\, \mathcal{P}(f)\, Df = \tfrac{1}{N} \sum_{k=1}^{N} \mathcal{F}[f_k]\,.$$

New data can be included into the existing ensemble by assigning each replica a unique weight w. This weight assesses the agreement between the replica and new data. The reweighted ensemble now forms a representation of the probability distribution of PDFs $\mathcal{P}_{\rm new}(f)$ conditional on both the existing and new data. The mean value of the observable $\mathcal{F}$ taking account of the new data is then given by the weighted average

$$\langle \mathcal{F} \rangle_{new} = \int \mathcal{F}[f]\, \mathcal{P}_{new}(f)\, Df = \tfrac{1}{N} \sum_{k=1}^{N} w_k \mathcal{F}[f_k],$$

where the weights are given in terms of the individual replica χ^2 to new data by

$$w_k = \frac{(\chi^2_k)^{(n-1)/2} e^{-\frac{1}{2}\chi^2_k}}{\frac{1}{N}\sum_{k=1}^{N} (\chi^2_k)^{(n-1)/2} e^{-\frac{1}{2}\chi^2_k}} = \mathcal{N}_\chi \mathcal{P}(\chi^2 | f_k)\,.$$

Note that after reweighting a given ensemble of N PDF replicas the efficiency in describing the distribution of PDFs is no longer the same. The reweighting procedure will often assign replicas very small weights, therefore these replicas no longer contribute to the ensemble. The efficiency of the representation of the underlying distribution $\mathcal{P}_{new}(f)$ will therefore be less than it would be in a new fit. The loss of information due to reweighting can be quantified using the Shannon entropy to determine the effective number of replicas in the reweighted set:

$$N_{\rm eff} \equiv \exp\{\tfrac{1}{N} \sum_{k=1}^{N} w_k \ln(N/w_k)\}.$$

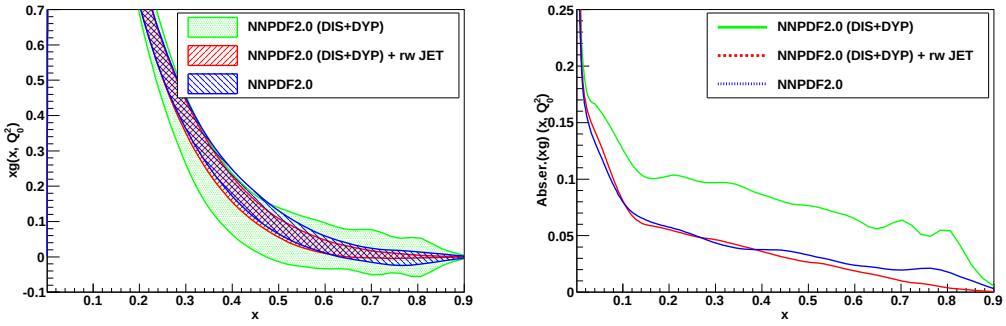

Figure 3.7: The gluon distribution (left) and its uncertainty (right) of the NNPDF2.0(DIS+DY) fit before and after reweighting with the inclusive jet data compared to the refitted gluon from NNPDF2.0.

Unweighting

Once a reweighted PDF set has been determined, it would be interesting to be able to produce a new PDF ensemble with the same probability distribution as a reweighted set, but without the need to include the weight information. A method of *unweighting* has therefore been developed, whereby the new set is constructed by deterministically sampling with replacement the weighted probability distribution. This means that replicas with a very small weight will no longer appear in the final unweighted set while replicas with large weight will occur repeatedly.

If the probability for each replica and the probability cumulants are defined as

$$p_k = \frac{w_k}{N_{\rm rep}} \qquad\qquad P_k \equiv P_{k-1} + p_k = \sum_{j=0}^{k} p_j \,,$$

it is possible to quantitatively describe the unweighting procedure. Starting with $N_{\rm rep}$ replicas with weights w_k, $N_{\rm rep}$ new weights w'_k are determined:

$$w'_k = \sum_{j=1}^{N'_{\rm rep}} \theta\big(\frac{j}{N'_{\rm rep}} - P_{k-1}\big)\theta\big(P_k - \frac{j}{N'_{\rm rep}}\big) \,.$$

These weights are therefore either zero or a positive integer. By construction they satisfy

$$N'_{\rm rep} \equiv \sum_{k=1}^{N_{\rm rep}} w'_k \,,$$

i.e. the new unweighted set consists of $N'_{\rm rep}$ replicas, simply constructed by taking w'_k copies of the k-th replica, for all $k = 1, \ldots, N_{\rm rep}$. This procedure is illustrated graphically in Figure 3.8.

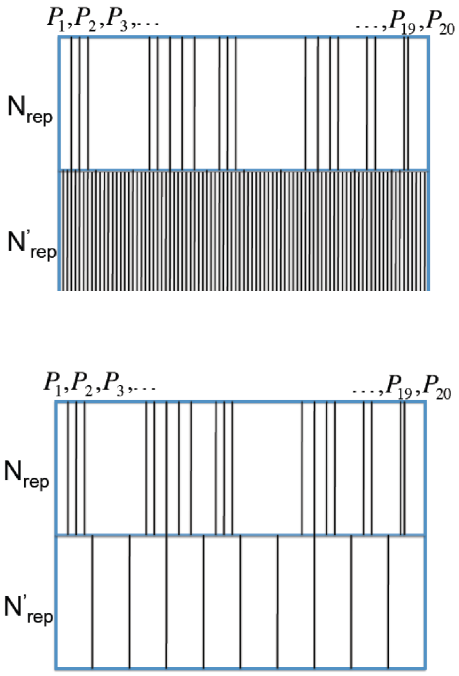

Figure 3.8: Graphical representation of the construction of a set of N'_{rep} unweighted replicas from a set of $N_{\rm rep} = 20$ weighted ones. Each segment is in one-to-one correspondence to a replica, and its length is proportional to the weight of the replica. The cases of $N'_{\rm rep} \gg N_{\rm rep}$ (top) and $N'_{\rm rep} = 10$ (bottom) are shown.

3.9 Closure Test

To verify the effectiveness of the reweighting procedure, here below will be shown that including datasets by reweighting produces an ensemble of PDF replicas statistically equivalent to a full refit. This test is called *closure test*. At first, a new NNPDF2.0 fit including only DIS and Drell-Yan data is produced. The data left out of the fit (Tevatron Run II inclusive jet data) is then reintroduced by reweighting. The resulting reweighted ensemble is then compared to the full NNPDF2.0 fit.

In Figure 3.7 is shown the gluon PDF for the three sets; the prior fit NNPDF2.0(DIS+DY), the reweighted set NNPDF2.0(DIS+DY) with jet data included, and the refitted full set NNPDF2.0. The figure shows excellent agreement between the reweighted set and the full fit. Differences stay well below statistical fluctuations.

To obtain a more precise estimation of the statistical equivalence of the refitted and reweighted parton sets, and also to test the unweighting procedure, it is useful to examine the statistical distances between the new unweighted distributions and the refitted set. The distance formulae are defined in Appendix A of Ref. [87]. If two sets give a

description of the same underlying probability distribution and so are statistically equivalent, the distance between them will fluctuate around a value of one. At $d \sim 7$ the discrepancy between the two sets is at the one-sigma level. In the case of the Tevatron jets reweighting exercise, can be seen in Figure 3.9 that these distances oscillate around one. The reweighted set is therefore equivalent to the refit and there is no significant loss of accuracy in the unweighting procedure.

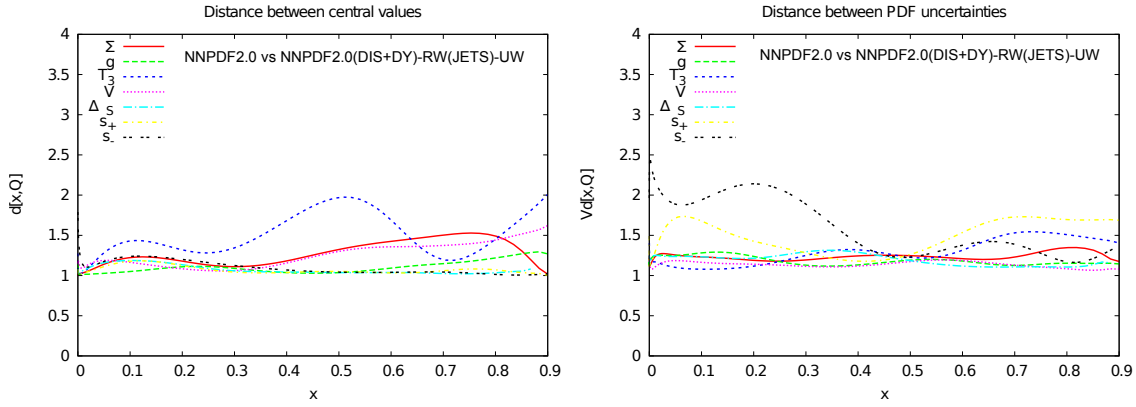

Figure 3.9: Distance between central values (left) and uncertainties (right) of the NNPDF2.0 PDFs and the NNPDF2.0 DIS+DY PDFs reweighted with Tevatron jet data and then unweighted.

Having developed the unweighting procedure, it is possible to perform another check on the consistency of the reweighting method. When adding more than one set of data by reweighting, this method must satisfy combination and commutation properties. Reweighting with both sets must be equivalent to reweighting with one, unweighting then reweighting with the other. Of course switching the order in which the reweighting is performed must produce an equivalent distribution.

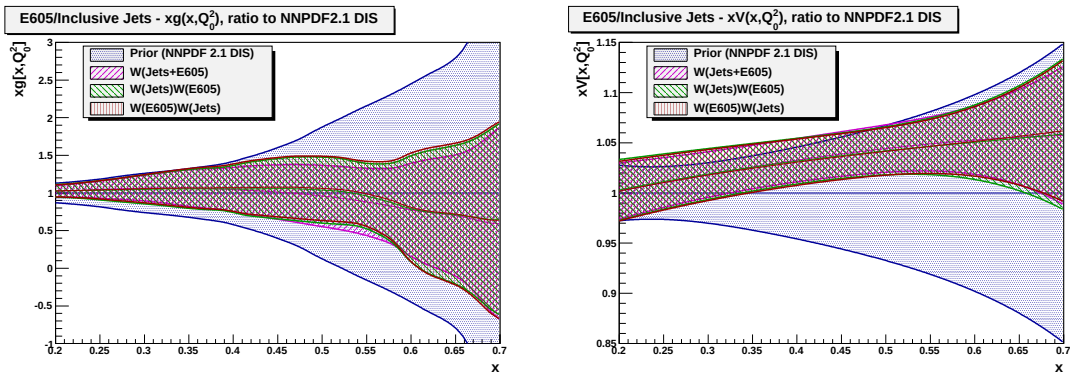

Figure 3.10: Multiple reweighting demonstration. Plots of gluon PDF(left) and valence PDF (right).

To check that the procedure satisfies these properties, a test is performed using the Tevatron jet data as the first dataset and E605 fixed target Drell-Yan data as the second one.

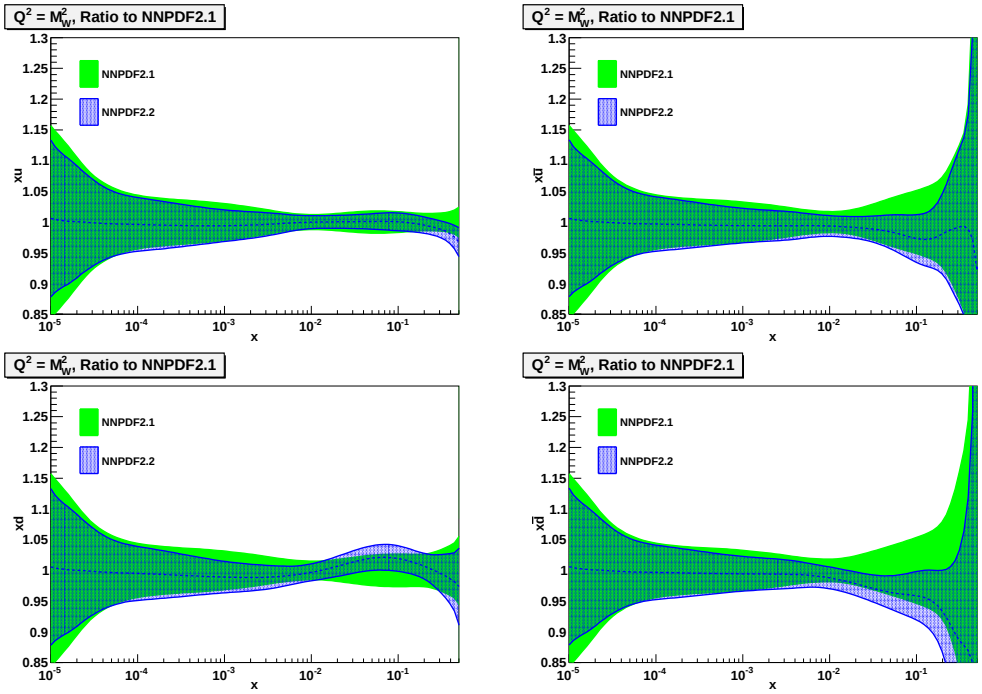

Figure 3.11: Comparison of light quark and antiquark distributions at the scale $Q^2 = M_W^2$ from the global NNPDF2.1 and NNPDF2.2 global fits. Parton densities are plotted normalized to the NNPDF2.1 central value.

In Figure 3.10 are compared the inclusion of the combined set with the inclusion of one set after the other. The result is clearly independent from the order in which the inclusion of single datasets is performed. A distance analysis on the three produced sets confirms that the reweighting method satisfies the combination and commutation requirements.

NNPDF2.2

The Bayesian reweighting method has been used to construct a new NNPDF parton set: NNPDF2.2 [193]. In this set is taken as a prior ensemble the NNPDF2.1 fit and the W-lepton charge asymmetry measurements of the ATLAS, CMS and D0 collaborations are inclueded by reweighting.

The NNPDF2.1 set provides a reasonable description of the new measurements, with $\chi^2_{tot}/N_{dat} = 2.22$. After reweighting with the new data this improves to an excellent level of agreement with $\chi^2_{tot}/N_{dat} = 0.81$. Having reweighted a prior set with $N_{\rm rep} = 1000$ initial replicas, 181 remain, indicating that the data provides a substantial constraint. Using the unweighting procedure outlined above, the new PDF set with $N_{\rm rep} = 100$ has been produced.

Figure 3.11 demonstrates the impact of the new data on the light quark and antiquark PDFs. The uncertainties are significantly constrained by the data in two main regions, is

observed a reduction of around 20% at $x \sim 10^{-3}$ and 30% in the region $x \sim 10^{-2}$ to $x \sim 10^{-1}$. The overall fit quality improves slightly, from a total χ^2_{tot}/N_{dat} of 1.165 with NNPDF2.1 to 1.157. The constraints demonstrated here are the first such constraints upon parton distributions from LHC data.

Such constraints are particularly important given the discrepancies between global parton distribution fits in flavor separation at medium to large x. The W-lepton charge asymmetry data included here may prove useful in resolving some of these discrepancies.

Chapter 4

Impact of Heavy Quark Masses on PDFs

After introducing and discussing the theoretical issues related to parton distribution determination and after discussing the main features of a general NNPDF fit, explaining all the related theoretical and experimental aspects, in this Chapter are presented the results of the NNPDF2.1 analysis. At first the NLO results are presented and compared both to the previous NNPDF2.0 release and to parton sets from other groups. The dependence of this set on the value of heavy quark masses is also discussed. In the central part of the Chapter the LO and NNLO NNPDF2.1 parton sets are presented, and finally perturbative stability of the sets going from LO to NNLO is verified.

4.1 Next-to-Leading Order Results

4.1.1 Statistical Features

The results of the global fit performed to determine the NNPDF2.1 NLO parton set are briefly compared to NNPDF2.0 release and to other PDF sets: CT10 and MSTW08. A first comparison between the two NNPDF sets can be done by computing a set of statistical estimators for NNPDF2.1, that here are shown in Tab. 4.1 for the global fit and in Tab. 4.2 for individual experiments. A direct comparison with NNPDF2.0 χ^2 is given in Tab. 4.2, where the shown values $\chi^2_{2.0}$ has been computed including the normalization uncertainties through the same t_0 prescription as for the NNPDF2.1 case.

The set of estimators shown in Tab. 4.1 are the same as the ones already used in Ref. [87]: $\chi^2_{\rm tot}$ is computed comparing the central (average) NNPDF2.1 fit to the original experimental data, $\left\langle \chi^{2(k)} \right\rangle$ is computed comparing to the data each NNPDF2.1 replica and averaging over replicas, while $\langle E \rangle$ is the quantity which is minimized, i.e. it coincides

$\chi^2_{\rm tot}$	1.16
$\langle E\rangle \pm \sigma_E$	2.24 ± 0.09
$\langle E_{\rm tr}\rangle \pm \sigma_{E_{\rm tr}}$	2.22 ± 0.11
$\langle E_{\rm val}\rangle \pm \sigma_{E_{\rm val}}$	2.28 ± 0.12
$\langle {\rm TL}\rangle \pm \sigma_{\rm TL}$	$(1.6 \pm 0.6)\ 10^4$
$\left\langle \chi^{2(k)}\right\rangle \pm \sigma_{\chi^2}$	1.25 ± 0.09
$\left\langle \sigma^{(\rm exp)}\right\rangle_{\rm dat}$(%)	11.3
$\left\langle \sigma^{(\rm net)}\right\rangle_{\rm dat}$(%)	4.4
$\left\langle \rho^{(\rm exp)}\right\rangle_{\rm dat}$	0.18
$\left\langle \rho^{(\rm net)}\right\rangle_{\rm dat}$	0.56

Table 4.1: Table of statistical estimators for NNPDF2.1 NLO with $N_{\rm rep} = 1000$ replicas. The total average uncertainty is given in percentage. All the χ^2 and E values have been computed using the same t_0 covariance matrix [103] used for minimization.

Experiment	χ^2	$\chi^2_{2.0}$	$\langle E\rangle$	$\left\langle \sigma^{(\rm exp)}\right\rangle_{\rm dat}$(%)	$\left\langle \sigma^{(\rm net)}\right\rangle_{\rm dat}$(%)	$\left\langle \rho^{(\rm exp)}\right\rangle_{\rm dat}$	$\left\langle \rho^{(\rm net)}\right\rangle_{\rm dat}$
NMC-pd	0.97	1.04	2.04	1.9%	0.5%	0.03	0.37
NMC	1.73	1.73	2.79	5.0%	1.5%	0.16	0.71
SLAC	1.27	1.42	2.34	4.4%	1.6%	0.31	0.79
BCDMS	1.28	1.30	2.33	5.7%	2.3%	0.47	0.60
HERAI-AV	1.07	1.15	2.15	2.5%	1.2%	0.06	0.35
CHORUS	1.15	1.24	2.23	15.1%	4.7%	0.08	0.32
FLH108	1.37	1.50	2.36	72.0%	4.0%	0.64	0.67
NTVDMN	0.76	0.73	1.77	21.1%	14.1%	0.04	0.62
ZEUS-H2	1.29	1.33	2.32	13.4%	1.2%	0.27	0.51
ZEUSF2C	0.78	-	1.80	23.3%	3.1%	0.08	0.41
H1F2C	1.50	-	2.52	17.3%	3.0%	0.30	0.40
DYE605	0.84	0.87	1.92	22.3%	7.9%	0.47	0.76
DYE866	1.27	1.29	2.37	20.1%	9.2%	0.20	0.52
CDFWASY	1.86	1.84	3.08	6.0%	4.4%	0.51	0.75
CDFZRAP	1.65	1.85	2.80	11.5%	3.6%	0.82	0.72
D0ZRAP	0.60	0.60	1.62	10.2%	3.1%	0.53	0.76
CDFR2KT	0.97	1.01	2.10	22.2%	4.0%	0.78	0.57
D0R2CON	0.84	0.86	1.92	16.8%	4.5%	0.77	0.59

Table 4.2: Same as Table 4.1 for individual experiments. All estimators have been obtained with $N_{\rm rep} = 1000$ replicas. Note that experimental uncertainties are always given in percentage. In the second and third column the NNPDF2.1 and NNPDF2.0 set [87] χ^2 have been computed with the t_0 prescription.

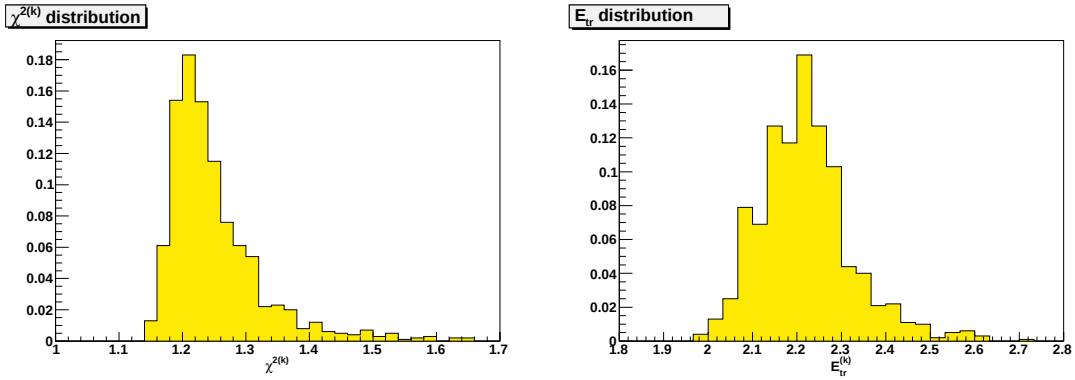

Figure 4.1: Distribution of $\chi^{2(k)}$ (left) and $E_{\rm tr}^{(k)}$ (right), over the sample of $N_{rep} = 1000$ replicas.

with the χ^2 computed comparing each NNPDF2.1 replica to the data replica it is fitted to, with the three values given corresponding to the total, training, and validation datasets.

The value of $\chi^2_{\rm tot}$ for the NNPDF2.0 global fit computed using the t_0 method, to be compared to the NNPDF2.1 value of Tab. 4.1, is $\chi^2_{\rm tot} = 1.23$ (very close to the value $\chi^2_{\rm tot} = 1.21$ of Tab. 9 in Ref. [87], computed with the "standard" covariance matrix). The value $\chi^2_{\rm tot} = 1.16$ for the NNPDF2.1 NLO fit is rather better than for the NNPDF2.0 analysis, where heavy quark mass effects are not included. It is possible to notice, looking at Tab. 4.2, that with respect to that analysis all datasets improves or remain similar: the improvement is particularly evident for the HERA-I average dataset (in consequence of the more tight kinematic cuts along with the implementation of heavy quark mass effects) but also for CHORUS dataset description.

It can happen that a replica never satisfies the conditions that activate dynamical stopping (Sect. 3.6), reaching the maximum number of possible iterations $N_{\rm gen}^{\rm max}$ fixed in the genetic algorithm. This can affect the quality of the fit. Looking at Fig. 4.28 it is clear that while most of the replicas fulfill the stopping criterion, a fraction ($\sim 12\%$) of them stops at $N_{\rm gen} = N_{\rm gen}^{\rm max}$. By performing a series of fits in which the training length is raised more and more, it has been checked that these replicas stop at higher lengths and that the loss of accuracy due to the choice of $N_{\rm gen}^{\rm max}$ is reasonably small, in that the features of the global fit change very little if $N_{\rm gen}^{\rm max}$ is raised.

If now the NNPDF2.1 NLO set and the CT10 analysis are compared, it can be noticed the effect of different kinematic cuts over the respective common datasets by looking at the number of datapoint for each set in Tab. 4.3. Moreover, the χ^2 is defined in a somewhat different way by the CTEQ/CT group, specifically, but not only, in what concerns the treatment of normalization errors (see Ref. [76]): hence this comparison should be taken with care. From this comparison, it is clear that the two sets have a comparable fit quality to fixed target DIS, CT10 being somewhat better for BCDMS proton and NNPDF2.1

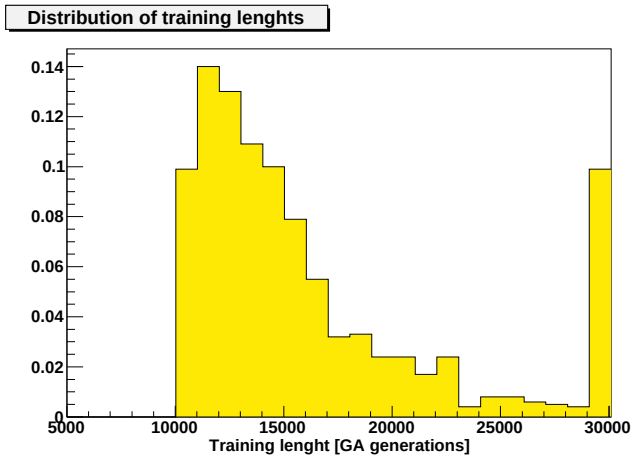

Figure 4.2: Distribution of training lengths over the sample of $N_{rep} = 1000$ replicas.

NLO rather better for NMC deuteron/proton ratio. The fit to HERA-I and Tevatron jet data is rather better in NNPDF2.1 NLO. Comparable fit quality to the Drell-Yan and vector boson production data is obtained in the two cases, with somewhat smaller χ^2 in the CT10 fit. No comparison is attempted for the HERA F_2^c data because of the very different kinematic cuts used in the two fits. A similar comparison to MSTW08 would be less significant because in the MSTW08 fit correlated systematics are not included in the covariance matrix for some datasets.

4.1.2 Parton Distributions

A comparison between NNPDF2.1 NLO and NNPDF2.0 deliveries is shown for the singlet and non-singlet sector respectively in Figs. 4.3-4.4.

The main differences can be found at medium and small-x in the singlet, where for NNPDF2.1 the PDF is slightly larger, at small-x the gluon is also larger than in NNPDF2.0 as its medium- and small-x uncertainty, that again is slightly larger in the 2.1 delivery. Below will be shown that the gluon central value shift with respect to NNPDF2.0 is a consequence of the general mass scheme treatment, while the wider uncertainty is due to the different kinematic cut. The valence PDF is poorly affected by these changes, but is equally subject to minor modifications in consequence of cross-talk induced by sum rules and other constraints. It is possible to better understand this effect due to sum rules by looking at the strange PDF in NNPDF2.1: including heavy quark mass effects, an enhancement with respect to NNPDF2.0 of the strange distribution would be expected. In the NNPDF2.0 set an Improved-ZM scheme is implemented for the dimuon data, which mainly constrain the strangeness. This scheme overestimates heavy quark mass effects, thus generating a charm distribution lower than the one a GM implementation would

Experiment	NNPDF2.1NLO $N_{\rm dat}$	NNPDF2.1NLO χ^2	CT10 $N_{\rm dat}$	CT10 χ^2
NMC-pd	132	0.97	121	1.28
NMC	221	1.73	196	1.71
BCDMSp	333	1.28	337	1.14
BCDMSd	248	1.15	250	1.12
HERAI-AV	592	1.07	579	1.17
NTVnuDMN	41	0.50	38	0.94
NTVnbDMN	38	0.42	33	0.91
DYE605	119	0.85	119	0.81
DYE866p	184	1.31	184	1.21
DYE866r	15	0.77	15	0.64
CDFZRAT	29	1.62	29	1.44
D0ZRAP	28	0.59	28	0.54
CDFR2KT	76	0.97	76	1.55
D0R2CON	110	0.84	110	1.13

Table 4.3: Comparison of χ^2 per data point for experiments which are common to the NNPDF2.1 NLO and CT10 PDF determinations. For each PDF set the number of data points after kinematic cuts is given.

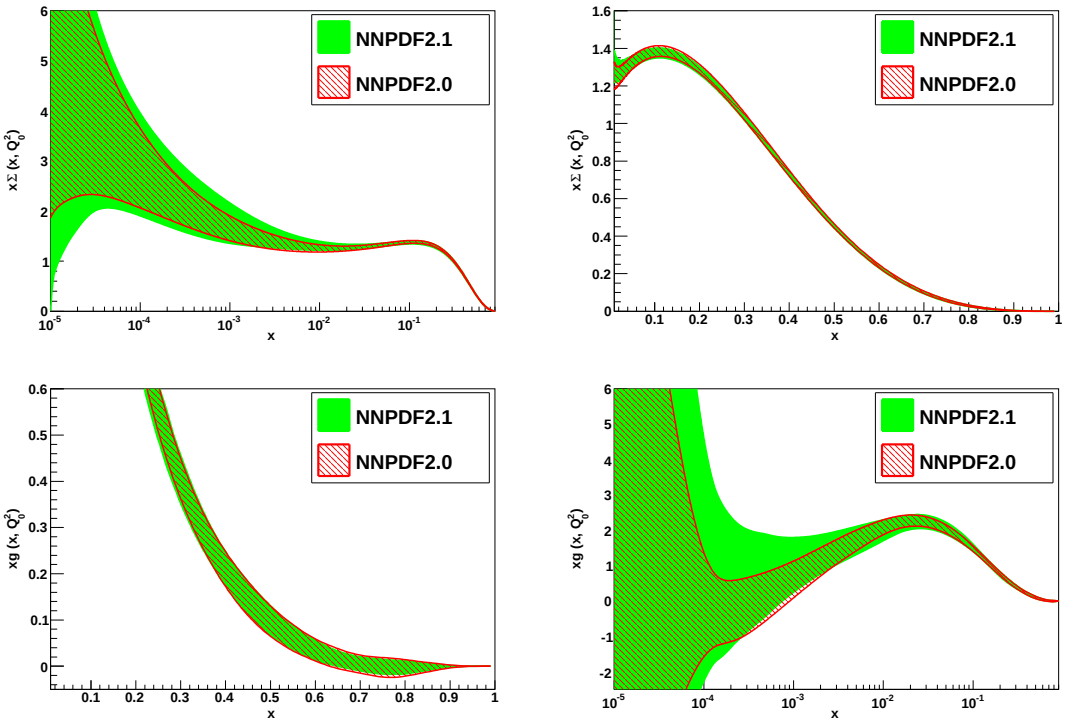

Figure 4.3: Comparison of NNPDF2.1 and NNPDF2.0 singlet sector PDFs, computed using $N_{\rm rep} = 1000$ replicas from both sets. All error bands shown correspond to one sigma.

generate, and as a consequence all other PDFs are enhanced. In the following will be shown that comparing the NNPDF2.1 result for strangeness with a pure ZM fit reveals the expected behaviour: the GM fit for strangeness will be somewhat enhanced with respect to the ZM case.

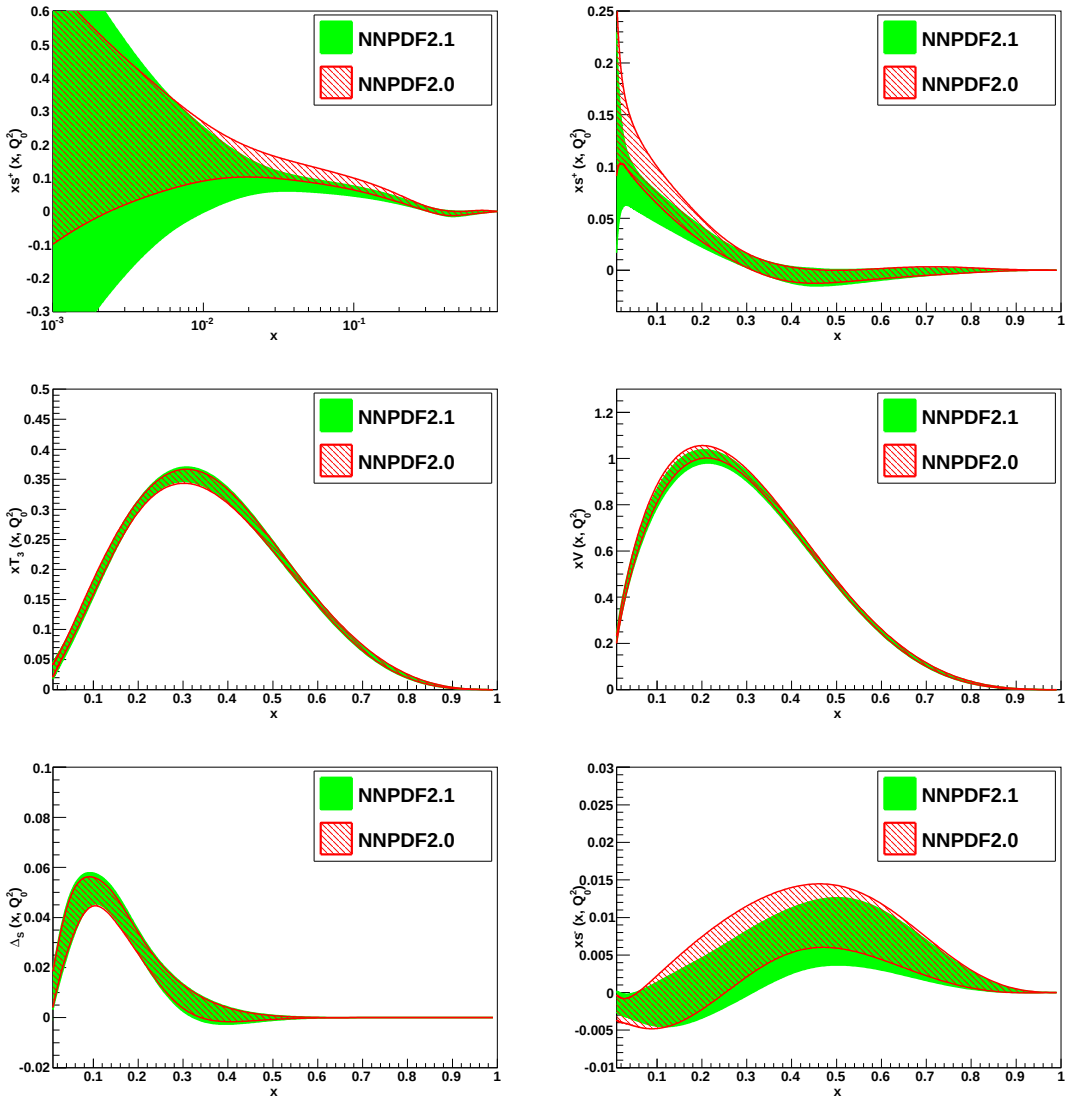

Figure 4.4: Same as Fig. 4.29 for the non-singlet sector PDFs.

As defined in Appendix A of Ref. [87], the distance between two sets of PDFs can be computed in order to quantify the differences. This statistical tool allows to establish whether two sets are a representation of the same underlying probability distribution or not. For $d \sim 1$ the two sets come from the same distribution while for $d \sim 7$, in the case of $N_{\rm rep} = 100$ replicas, a one-sigma difference is observed. The differences between NNPDF2.1 NLO and NNPDF2.0 are mainly present in the medium-x strangeness and to

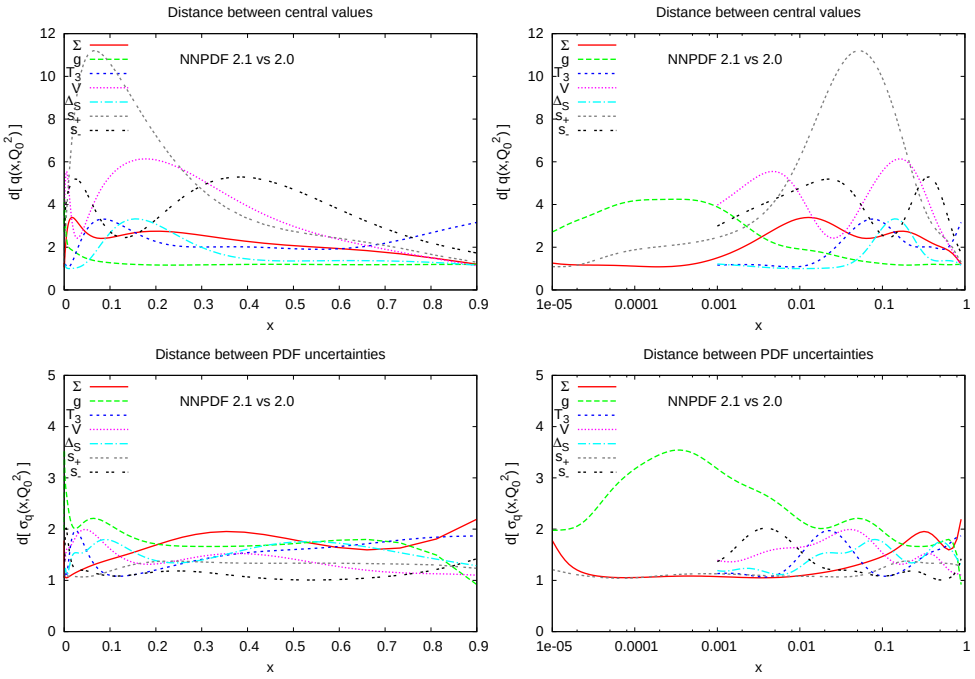

Figure 4.5: Distance between the NNPDF2.0 and NNPDF2.1 parton sets. All distances are computed from sets of $N_{\rm rep} = 100$ replicas.

a lesser extent in the medium and small-x gluon. In general, the two sets are clearly not describing the same underlying distribution but all PDFs are consistent at the one-sigma level. Only the strangeness is consistent at the 90% confidence level.

These differences that have been discussed for the PDF sets at the initial scale are propagated through evolution also to the W and Z scale: at $Q^2 = 10^4$ GeV^2 differences in gluon and light sea quark distributions up to the one-sigma level at small-x are observed.

4.1.3 Comparison to NNPDF2.0 Parton Set

It is an interesting exercise to consider one by one the main differences introduced in the NNPDF2.1 analysis with respect to the NNPDF2.0 fit and check the corresponding impact on PDFs.

With respect to NNPDF2.0, with the inclusion of heavy quark mass effects mainly a new $Q^2_{\rm min}$ kinematic cut is applied, the FONLL-A general mass scheme is implemented, HERA F_2^c data are included, and a threshold prescription in the GM scheme is introduced.

In Fig. 4.6 is shown the distance between NNPDF2.0 PDFs and a fit with the same dataset but with the new cut $Q^2_{\rm cut} = 3$ GeV^2, denoted by NNPDF2.0RED (reduced). Also, in order to ease the subsequent discussion on the impact of heavy quark mass effects, in NNPDF2.0RED a pure ZM scheme is used for all observables, rather than the I-ZM

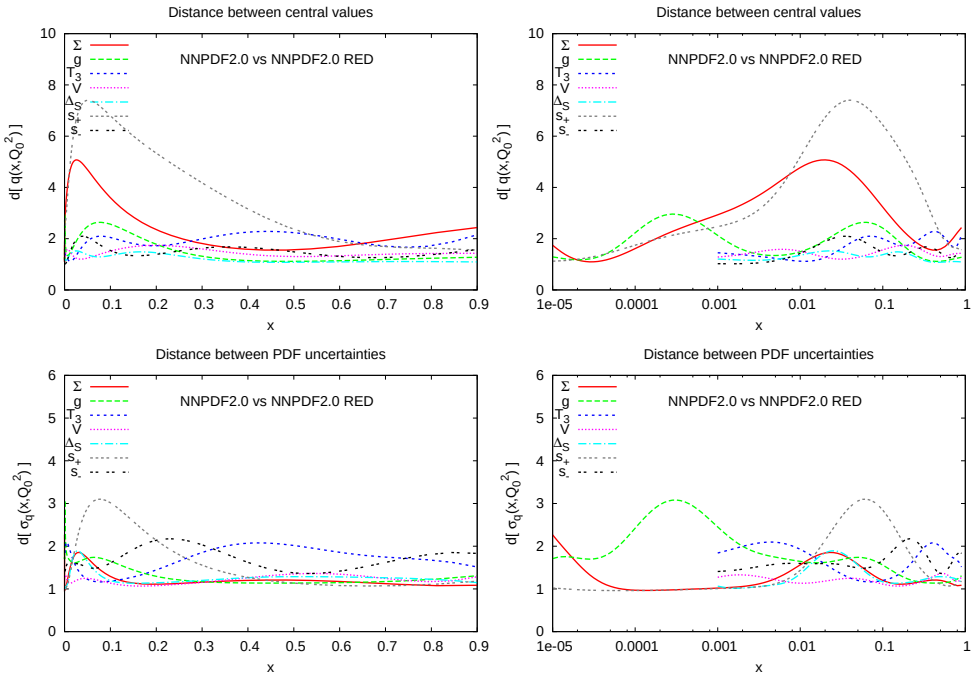

Figure 4.6: Distance between the NNPDF2.0 PDF set and a fit to the same data but with $Q^2_{\rm cut} = 3$ GeV^2 and the ZM-VFN scheme for all observables (NNPDF2.0 RED). All distances are computed from sets of $N_{\rm rep} = 100$ replicas.

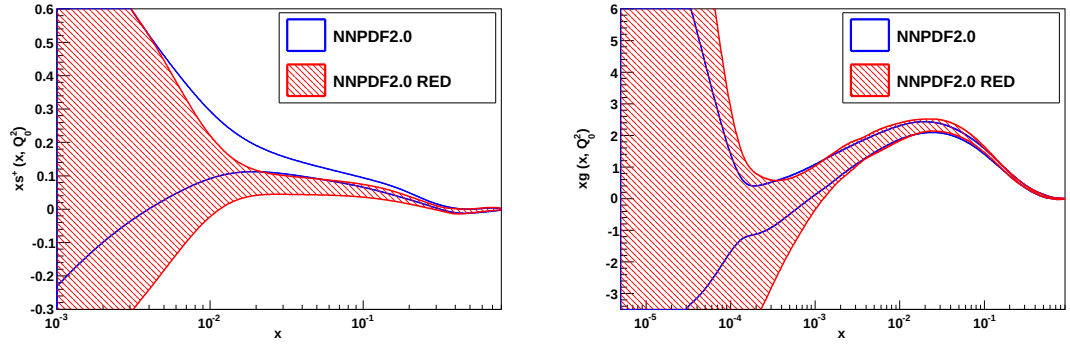

Figure 4.7: Comparison of the small-x total strangeness and gluon in NNPDF2.0 and in NNPDF2.0RED (the distances are shown in Fig. 4.6).

scheme [39] used for dimuon data in Ref. [87].

The two mainly affected PDFs are the medium-x strange and the small-x gluon PDFs, as can be seen by looking at Fig. 4.7 and at distances in Fig. 4.6. The strange is rather smaller in the ZM as compared to the I-ZM scheme, where it was enhanced due to the approximate inclusion of charm suppression. Its modification can be checked to be the responsible for the singlet deviation visible in the distance plot. The strange PDF contributes in fact to the singlet. The gluon is somewhat smaller at small-x and with rather larger uncertainties, due to the reduction in dataset at small-x caused by the new kinematic

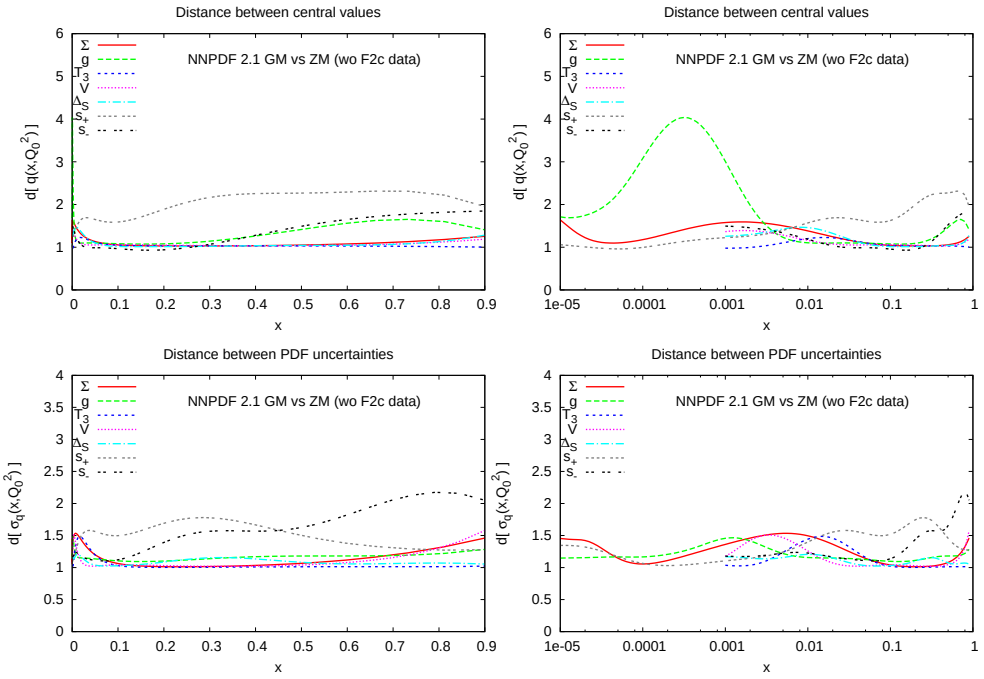

Figure 4.8: Distance between the NNPDF2.1 PDF sets in the GM and in the ZM schemes, in both cases without HERA F_2^c data. All distances are computed from sets of $N_{\rm rep} = 100$ replicas.

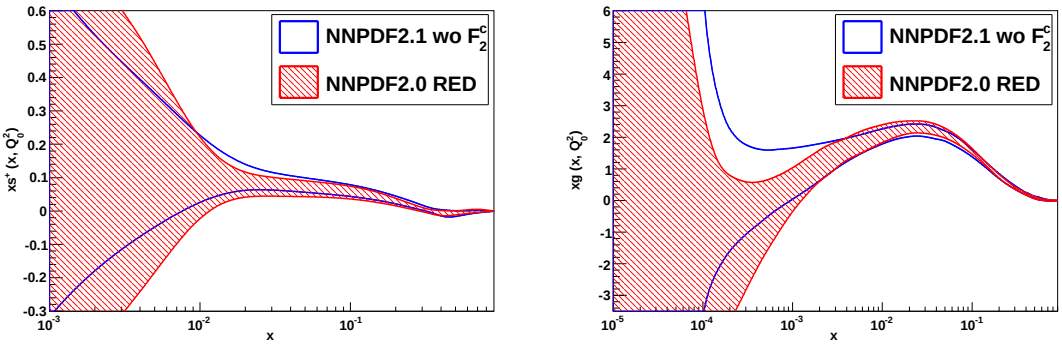

Figure 4.9: Comparison of the small-x total strangeness and gluon in NNPDF2.0RED and NNPDF2.1 without F_2^c data (distances are shown in Fig. 4.8).

cut.

Considering now the NNPDF2.1 fit without F_2^c data, it is compared to the above introduced fit NNPDF2.0RED. In this way exactly the same dataset is used and the only difference is in the implemented scheme: the FONLL-A GM scheme versus a pure ZM scheme.

It is possible to verify that the interpretation previously given of the strange PDF behaviour was correct: as in the previous case, the only PDFs which undergo a change worth mentioning are the strange and gluon PDFs. The singlet is instead unaffected. These PDFs and the corresponding distances are presented in Fig. 4.9 and in Fig. 4.8

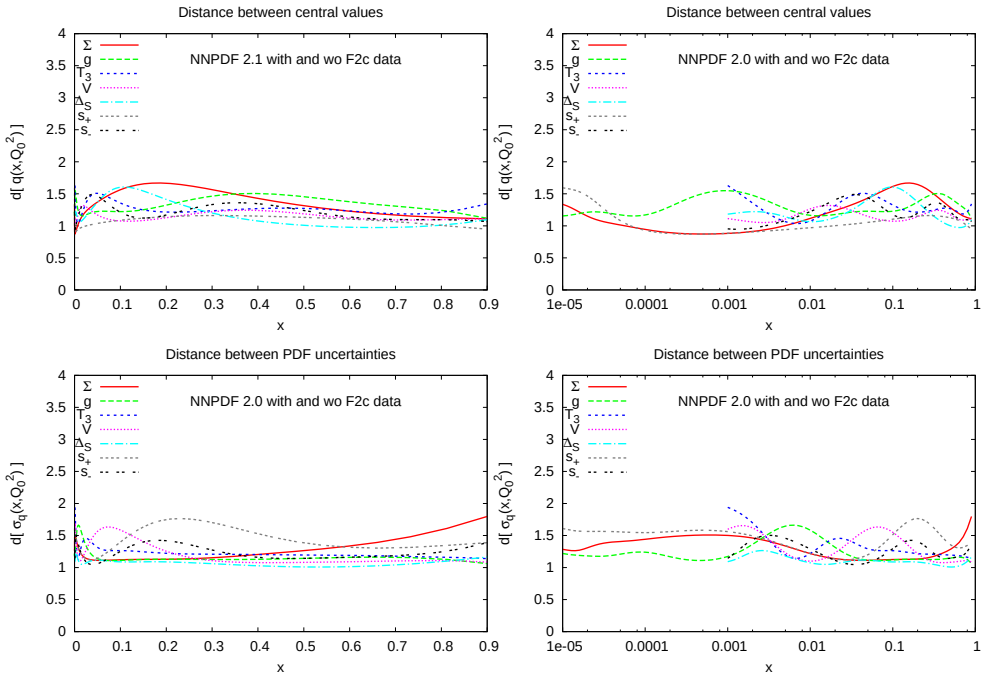

Figure 4.10: Distance between NNPDF2.1 PDF sets with and without HERA $F_2^c(x,Q^2)$ data. Distances have been computed from sets of $N_{\rm rep} = 100$ replicas.

respectively. The GM scheme leads to a larger gluon for $x \lesssim 2 \cdot 10^{-3}$, as well as to a somewhat medium-x larger strangeness. This confirms what previously stated about the I-ZM approximation for dimuon data in NNPDF2.0, which overestimates charm mass effects.

The impact of HERA F_2^c data is quite low. It is possible to see this by looking at distances between NNPDF2.1 with or without the inclusion of this dataset, represented in Fig. 4.10. Almost all distances lay around $d \sim 1$, due in part to the relatively large uncertainties on current F_2^c data and in part to the kinematic cuts. In fact, low x and Q^2 data, which are most sensitive to the gluon PDF, are excluded by cuts. Inclusion of $\mathcal{O}(\alpha_s^2)$ heavy quark mass effects (e.g. by means of the FONLL-B scheme) is necessary in order to take advantage of these data.

Regarding the last issue concerning heavy quark mass effects implementation, a NNPDF2.1 fit with a pure FONLL-A scheme (i.e. without the threshold damping factor) is compared to the standard NNPDF2.1 fit. The difference between these cases should provide a reasonable estimate of the spread of results obtained by including heavy quark masses according to different prescriptions, as suggested in Ref. [38]. Looking at distances in Fig. 4.11 is easy to see that the singlet and gluon PDFs at medium-x, shown in Fig. 4.12, are the most affected. Without damping factor, the F_2^c structure function is closer to the massless result even at moderate Q^2, and this explains why the singlet PDF is somewhat

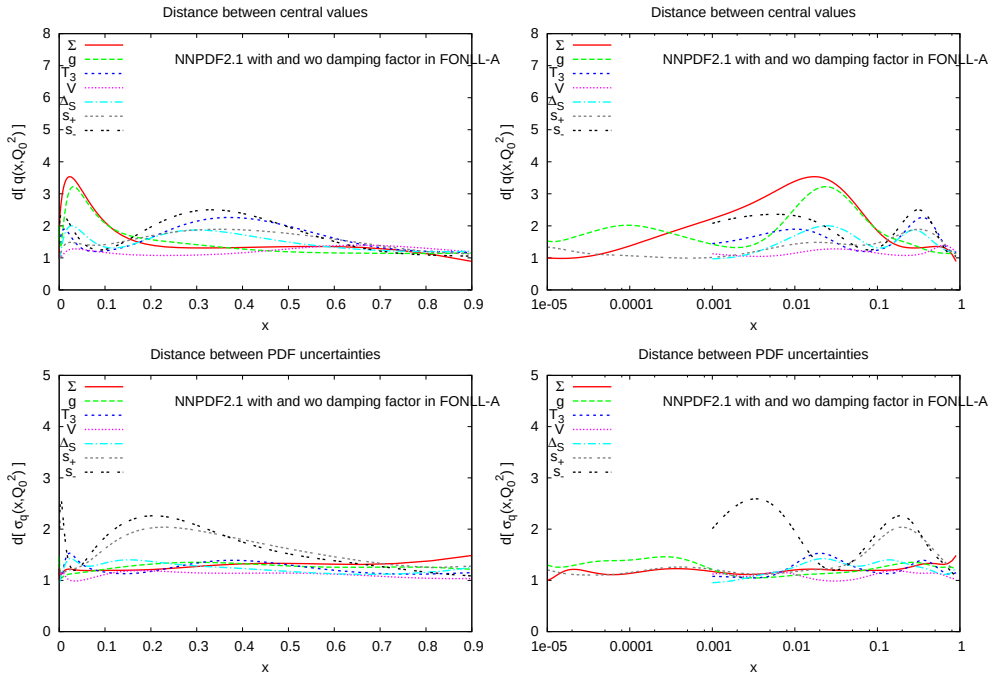

Figure 4.11: Distance between the NNPDF2.1 reference set and the same set obtained without threshold damping factor in the computation of the FONLL-A structure functions. .

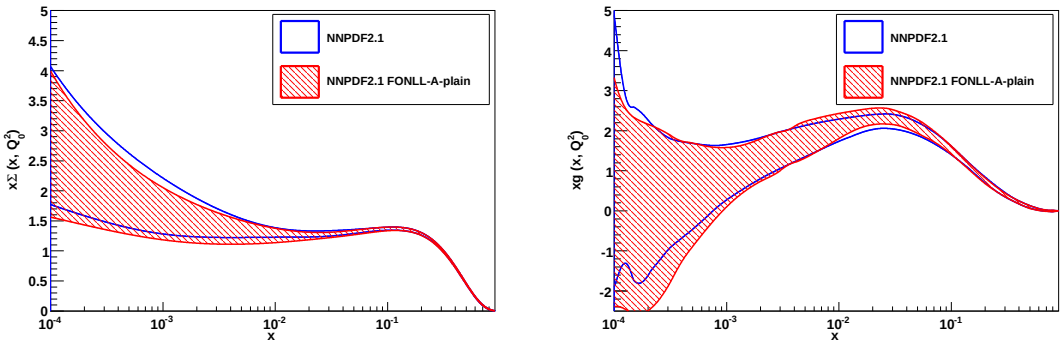

Figure 4.12: The small-x singlet and gluon PDFs, in the reference NNPDF2.1 set and in NNPDF2.1 obtained with FONLL-A without threshold damping factor (distances are shown in Fig. 4.11).

smaller at medium-x.

4.1.4 Comparison to CT10 and MSTW08 Parton Sets

As already done in Ref. [87] among NNPDF2.0, CTEQ6.6 and MSTW08 PDF sets (Figs. 18-19 of that reference), here NNPDF2.1 NLO PDFs are compared to other global PDF sets, CT10 [76] and MSTW08 [81], in Figs. 4.13-4.14. It is interesting here to refer to that previous comparison.

- It is likely that the medium-x singlet marginal agreement is a consequence of the poor agreement in strangeness.

- While the large-x gluon PDF is in marginal agreement with the other sets, in the medium-/small-x region the agreement is improved. The inclusion of heavy quark mass effects brings the NNPDF central value nearer to MSTW08 value. Moreover, the CT10 parametrization is more flexible than the one used for the CTEQ6.6 set, generating a central value and an error band in much better agreement with the wider NNPDF and MSTW uncertainties.

- The small changes in valence and triplet distributions between NNPDF2.0 and NNPDF2.1 go anyway in the direction of improving the agreement with the other global sets.

- The strange PDFs are quite different, presumably due to the fact that a much less flexible parametrization is adopted by CT/CTEQ and MSTW in comparison to NNPDF.

In the next Chapter, the differences analyzed above regarding PDF sets are translated to the observable level, focusing on LHC observables: will be shown that on the whole a reasonable agreement holds between global sets. The main significant differences are mostly related to the rather different large-x gluon in CT10 as shown in Fig. 4.13.

4.2 Heavy Quark Mass Dependence

The MSTW collaboration performed in Ref. [194] a study of the dependence of PDFs on the values of heavy quark masses m_c and m_b, using MSTW08 release as a basis for this study. In this Section a similar analysis is performed using NNPDF2.1 NLO as baseline fit and following the lines of that reference. At first, how different features of the NNPDF2.1 PDFs depend on the values of m_c and m_b is discussed. Then some LHC observables are considered and their dependence on m_c is studied, along with the correct treatment of heavy quark mass uncertainties in the Monte Carlo approach.

In order to quantify the dependence of PDFs on heavy quark masses, several fits are performed, shifting the default value for charm and bottom heavy quarks to various different values. The default masses are shown in Table 4.4, summarized and compared to those of other PDF sets, while the values for the different fits are 1.5, 1.6 and 1.7 GeV for the charm mass m_c and 4.25, 4.5, 5.0 and 5.25 GeV for the bottom mass m_b. All the other settings are left unchanged and equal to the reference fit with standard mass values. It is important to observe that at the order at which the analysis is performed, the perturbative

	m_c [GeV]	m_b [GeV]
NNPDF2.1	1.414	4.75
NNPDF2.0 [87]	$\sqrt{2}$	4.3
CT10 [76]	1.30	4.75
MSTW2008 [194]	1.40	4.75
ABKM09 [94]	1.50	4.50
HERAPDF1.0 [114]	1.40	4.75

Table 4.4: The default values of the heavy quark masses used in NNPDF2.1 and in several recent PDF sets.

definition of the heavy quark mass is immaterial: indeed different definitions (such as, for example, the pole and $\overline{\rm MS}$ mass definitions) differ by terms of $\mathcal{O}(\alpha_s)$. However, heavy quark mass corrections are included up to $\mathcal{O}(\alpha_s)$ only, so the difference is subleading (it becomes relevant once $\mathcal{O}(\alpha_s^2)$ heavy quark corrections are included, for example using the FONLL-B scheme). This implies that the value considered for the quark mass in the NNPDF analysis and in analogous determinations which are based on an NLO ACOT treatment of heavy quarks can be read as a pole mass or an $\overline{\rm MS}$ mass, equivalently. The

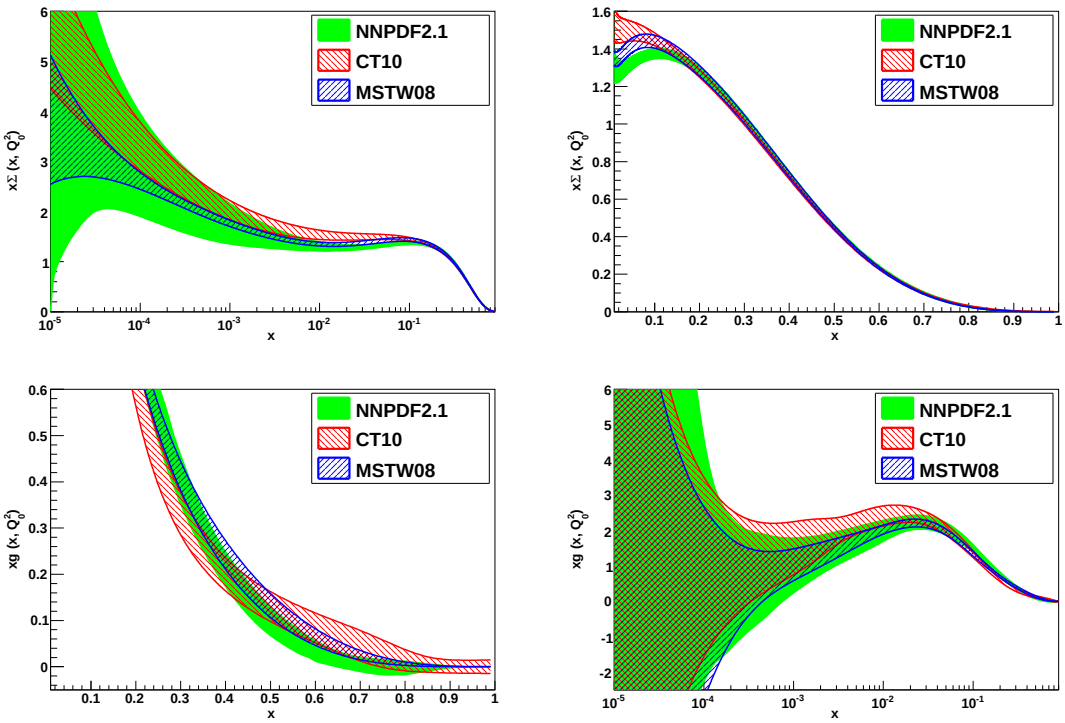

Figure 4.13: The NNPDF2.1 singlet sector PDFs, compared with the CT10 and MSTW08 PDFs. The results for NNPDF2.1 have been obtained with $N_{\rm rep} = 1000$ replicas. All PDF errors are given as one-sigma uncertainties.

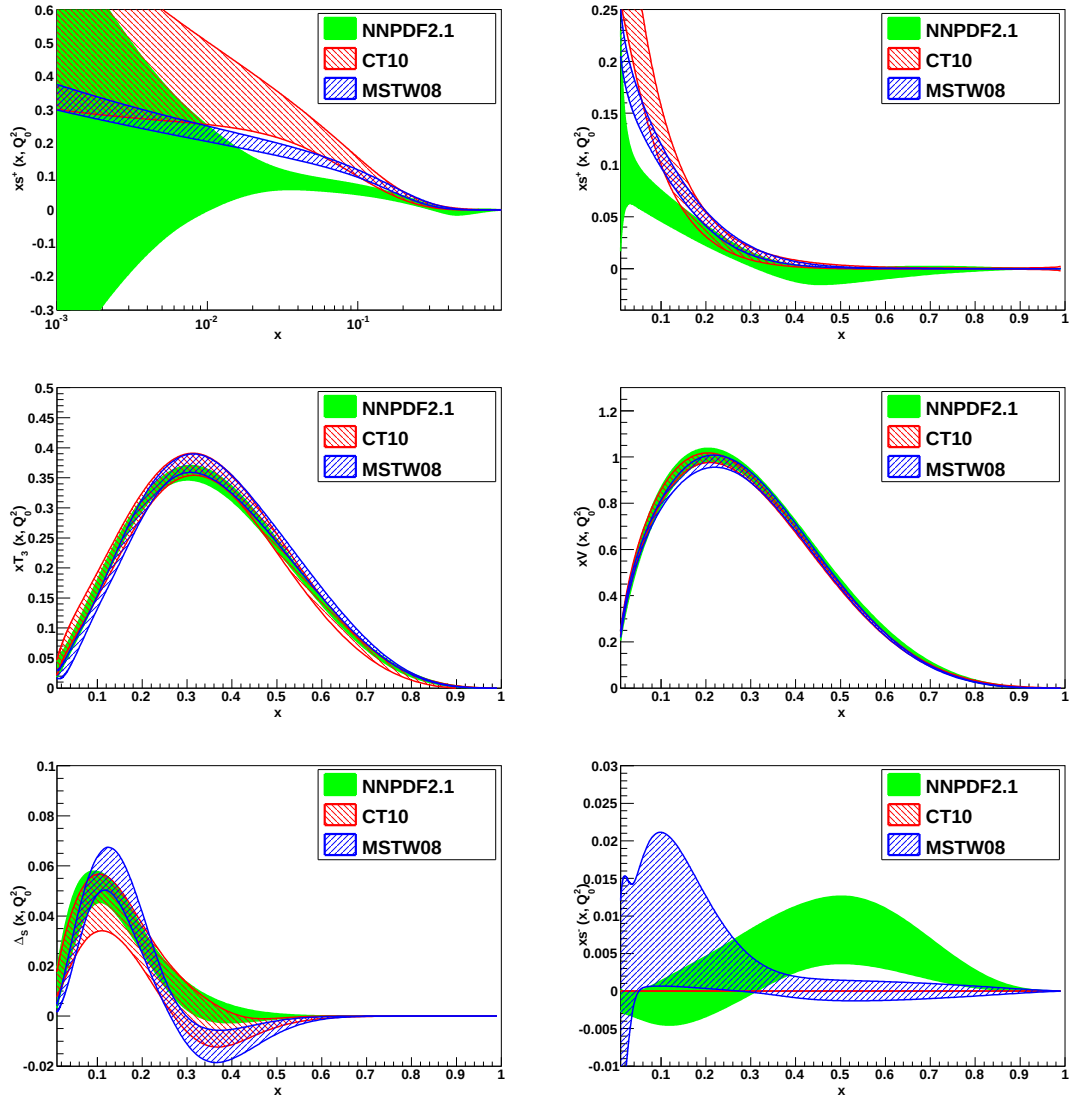

Figure 4.14: Same as Fig. 4.13 for the non-singlet sector PDFs.

$\overline{\rm MS}$ mass is better known, and it has been shown [195] to lead to perturbatively more stable results for deep-inelastic structure functions.

The central value of PDF sets with heavy quark mass variations is shifted accordingly to the value of the heavy quark mass: for larger values of the mass, smaller PDFs are obtained. This can be seen in Figs. 4.15-4.16, where the ratio of PDFs for different values of m_c and m_b to the reference NNPDF2.1 fit is plotted as a function of x for $Q^2 = 10^4$ GeV^2. This effect can be explained recalling that heavy flavors are generated radiatively and their PDFs vanish when the energy scale reaches the value of the heavy quark mass. Hence to a smaller value of the mass corresponds a longer evolution length and so a larger PDF. Because of the momentum sum rule, if the charm PDF becomes larger, other PDFs are ac-

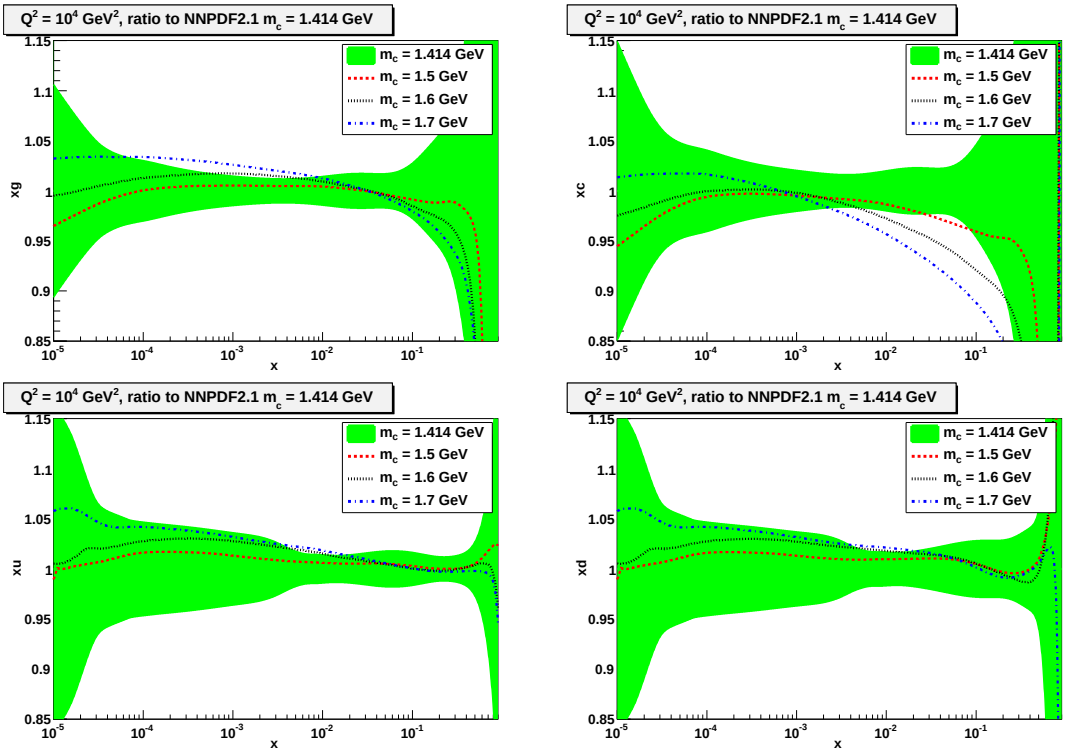

Figure 4.15: Ratio of NNPDF2.1 PDFs obtained for different values of the charm quark mass to the reference NNPDF2.1 set at $Q^2 = 10^4$ GeV^2. Top left: gluon; top right: charm; bottom left: up; bottom right: down.

cordingly smaller (and conversely), as anticipated in Sect. 4.1.2. For bottom in principle the same mechanism is at work, but in practice the effect on all other PDFs is negligible.

4.2.1 Mass Uncertainties and LHC Observables

A more detailed analysis of the phenomenological consequences due to the computation of LHC observables using NNPDF2.1 NLO parton set will be given in the next Chapter. Here the impact on LHC standard candles of charm and bottom mass values is considered. The dependence of light quark distributions and the gluon on the charm mass displayed in Fig. 4.15 is strong enough to affect these observables at the percent level or more, as observed in Ref. [187]. Instead, the bottom mass value affects all PDFs (except the same b distribution) in a much weaker way, below the percent level, so that only the observables that directly depend on the b distribution are significantly affected.

A sample of LHC standard candles is computed using each time a different PDF set, corresponding to the values of the charm mass of Fig. 4.15. The results of this computation are collected in Table 4.5, and are graphically represented in Fig. 4.17. It can be seen that to a variation of the m_c charm mass of the order of 10% corresponds a variation of

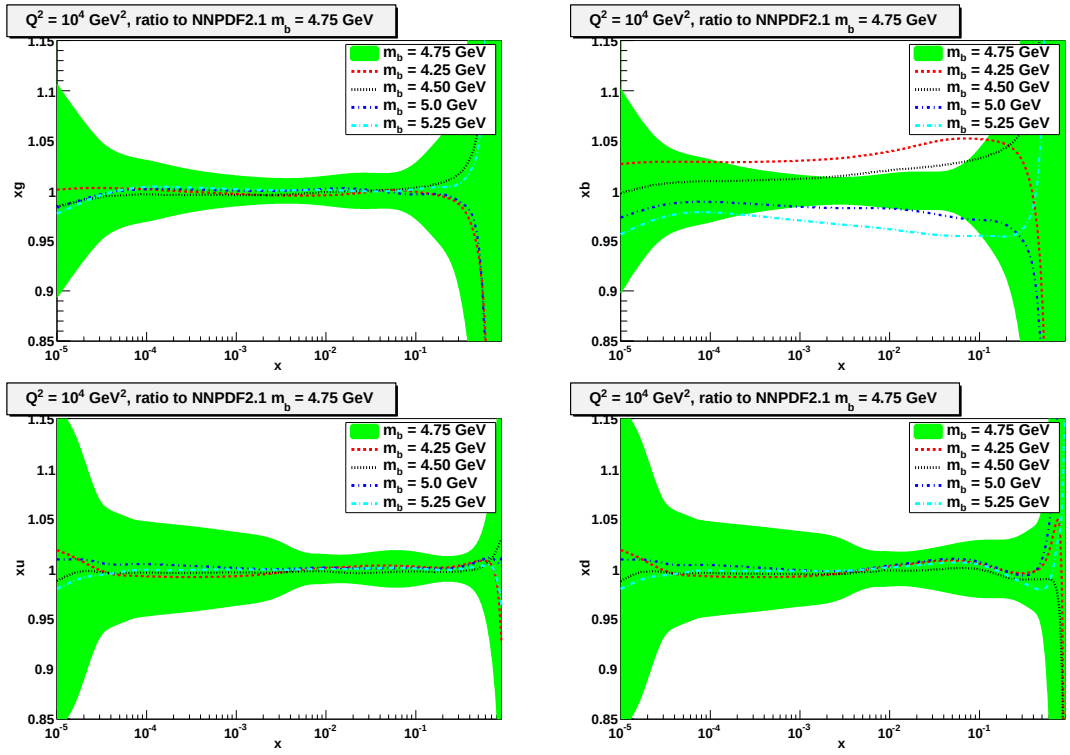

Figure 4.16: Ratio of NNPDF2.1 PDFs obtained for different values of the bottom quark mass to the reference NNPDF2.1 set at $Q^2 = 10^4$ GeV^2. Top left: gluon; top right: bottom; bottom left: up; bottom right: down.

the LHC standard candle values around the percent level.

Combined PDF+m_h Uncertainties and Correlations

The various sets obtained with heavy quark mass variations can be used to combine the uncertainty over PDFs with the heavy quark mass dependence. The Monte Carlo approach developed within the NNPDF collaboration makes the determination of this combined uncertainty much easier. The notation PDF$^{(k_{ij},i,j)}$ is introduced to refer to a precise replica k_{ij} of a precise set with defined $m_c^{(i)}$ and $m_b^{(j)}$ heavy quark mass values. It follows that the mean value of any observable $\mathcal{F}$ can be written as

$$\langle \mathcal{F} \rangle_{\rm rep} = \frac{1}{N_{\rm rep}} \sum_{i=1}^{N_{m_c}} \sum_{j=1}^{N_{m_b}} \sum_{k_{ij}=1}^{N_{\rm rep}^{(i,j)}} \mathcal{F}\left({\rm PDF}^{(k_{ij},i,j)}, m_c^{(i)}, m_b^{(j)} \right) , \tag{4.1}$$

where $N_{\rm rep}$ is the total number of replicas given by

$$N_{\rm rep} = \sum_{i=1}^{N_{m_c}} \sum_{j=1}^{N_{m_b}} N_{\rm rep}^{(i,j)} , \tag{4.2}$$

LHC 7 TeV	$W^+B_{l\nu}$ [nb]	$W^-B_{l\nu}$ [nb]	$Z^0B_{l\bar{l}}$ [nb]	$t\bar{t}$ [pb]	$gg{\to}H$ [pb]
$m_c = 1.414$ GeV	5.99 ± 0.14	4.09 ± 0.09	0.932 ± 0.020	170 ± 5	11.64 ± 0.17
$m_c = 1.5$ GeV	6.06 ± 0.17	4.14 ± 0.12	0.943 ± 0.024	169 ± 6	11.65 ± 0.25
$m_c = 1.6$ GeV	6.11 ± 0.14	4.17 ± 0.10	0.951 ± 0.020	167 ± 6	11.70 ± 0.21
$m_c = 1.7$ GeV	6.14 ± 0.14	4.19 ± 0.09	0.956 ± 0.019	166 ± 5	11.71 ± 0.22
$\rho\left[\sigma, m_c\right]$	0.44	0.41	0.48	-0.31	0.16

LHC 14 TeV	$W^+B_{l\nu}$ [nb]	$W^-B_{l\nu}$ [nb]	$Z^0B_{l\bar{l}}$ [nb]	$t\bar{t}$ [pb]	$gg{\to}H$ [pb]
$m_c = 1.414$ GeV	12.00 ± 0.27	8.84 ± 0.17	1.99 ± 0.036	946 ± 19	37.50 ± 0.40
$m_c = 1.5$ GeV	12.01 ± 0.31	8.94 ± 0.22	2.01 ± 0.04	942 ± 24	37.62 ± 0.62
$m_c = 1.6$ GeV	12.24 ± 0.28	9.02 ± 0.20	2.03 ± 0.04	939 ± 22	37.90 ± 0.55
$m_c = 1.7$ GeV	12.37 ± 0.28	9.10 ± 0.18	2.05 ± 0.04	935 ± 19	38.15 ± 0.58
$\rho\left[\sigma, m_c\right]$	0.48	0.50	0.56	-0.19	0.41

Table 4.5: LHC standard candles at $\sqrt{s} = 7$ TeV (upper table) and 14 TeV (lower table) obtained using NNPDF2.1 fits with different values of the charm mass m_c. The bottom line of each table gives the correlation coefficient between the observable and the mass.

with $N_{\rm rep}^{(i,j)}$ distributed according to a two dimensional Gaussian

$$N_{\rm rep}^{(i,j)} \propto \exp\left(-\frac{\left(m_c^{(i)} - m_c^{(0)}\right)^2}{2\delta_{m_c}^2} - \frac{\left(m_b^{(j)} - m_b^{(0)}\right)^2}{2\delta_{m_b}^2}\right) \tag{4.3}$$

with mean $(m_c^{(0)}, m_b^{(0)})$ and width $(\delta m_c, \delta m_b)$ and making the assumption that the values of charm and bottom masses are uncorrelated. Of course, a different probability distribution (possibly including a correlation between heavy quark mass values) could be assumed. It is consequently possible to compute the combined PDF+m_h uncertainty as the standard deviation of the observable over the replica sample as

$$\delta_{\rm PDF+m_h}\mathcal{F} = \sqrt{\langle\mathcal{F}^2\rangle - \langle\mathcal{F}\rangle^2}\,, \tag{4.4}$$

where to determine averages over replicas Eq. (4.1) must be used.

The correlation between PDFs and heavy flavor masses m_h can be easily computed:

$$\rho\left[m_h, \mathrm{PDF}\left(x,Q^2\right)\right] = \frac{\langle m_h \mathrm{PDF}\left(x,Q^2\right)\rangle_{\rm rep} - \langle m_h\rangle_{\rm rep}\langle \mathrm{PDF}\left(x,Q^2\right)\rangle_{\rm rep}}{\sigma_{m_h}\sigma_{\mathrm{PDF}(x,Q^2)}}\,, \tag{4.5}$$

where averages over replicas are to be understood in the sense of Eq. (4.1). The correlation Eq. (4.5), computed assuming $m_c = 1.55 \pm 0.15$ GeV and $m_b = 4.75 \pm 0.25$ GeV, is displayed in Fig. 4.18, as a function of x for $Q^2 = 10^4$ GeV^2. Again, it can be seen how, as the mass is increased, the corresponding heavy quark PDF is reduced. This implies

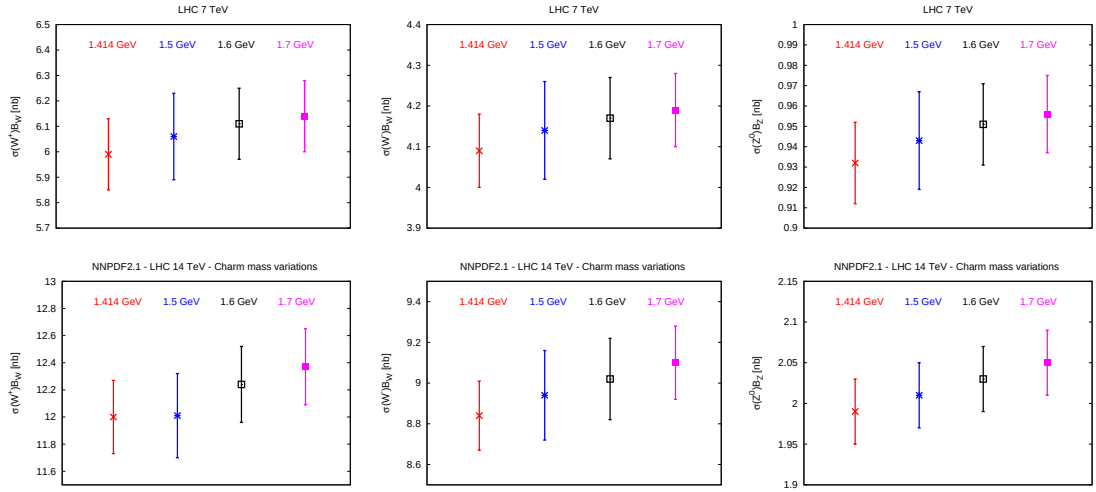

Figure 4.17: Graphical representation of the results of Table 4.5.

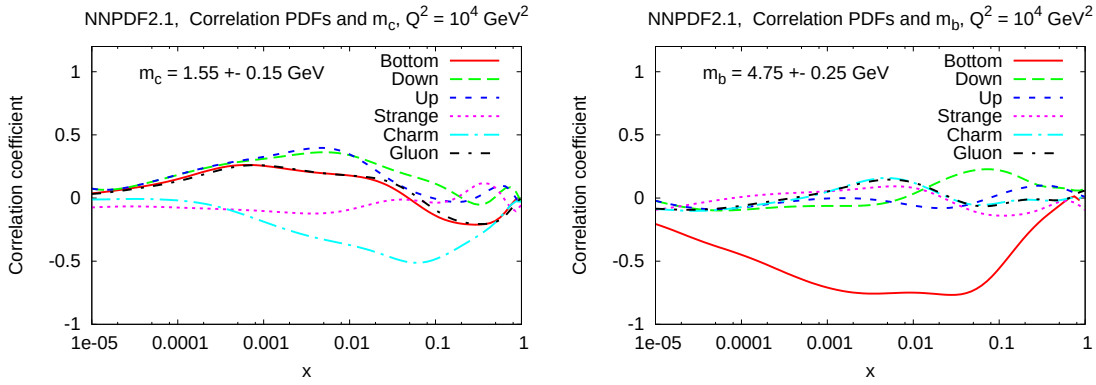

Figure 4.18: Correlation between PDFs and the heavy quark masses at a typical LHC scale $Q^2 = 10^4$ GeV^2: charm mass variations (left plot) and bottom mass variations (right plot). These correlations quantify the qualitative behaviour observed in Figs. 4.15-4.16.

that the PDF is strongly anticorrelated with its mass. As already discussed, because of the sum rules this effect have repercussions on the other PDFs: to the heavy quark suppression corresponds an enhancement of other PDFs, that is translated in a positive correlation in Fig. 4.18. For the bottom quark the effect is negligible.

4.3 Leading Order Results

The main reasons for producing a LO PDFs set are that they are mostly used with leading order Monte Carlo event generators, and are also of interest for comparison of QCD calculations at different perturbative orders. Even if the use of LO PDFs for calculations which has LO accuracy is not mandatory, using for example NLO PDFs, as suggested in

Ref. [196], with LO matrix elements may lead to a poorly behaved perturbative expansion and to bad phenomenology. On the other hand, to use a full LO standard PDF set, even if it is the simplest choice, may not be the best one. The problem is to determine which is the best PDF definition to match with a LO Monte Carlo event generator. The simplest and more consistent choice seems to be the standard full-LO PDF determination, but such a fit over a global dataset may give unacceptably low quality in the description of some sets. Several different "recipes" were proposed as possible modifications to the standard LO determination and are discussed below.

As expected, an optimal LO parton set lies typically at a rather large distance outside the experimental error band of a standard NLO PDF determination, and it is possible to argue from this fact that in a LO analysis theoretical uncertainties are predominant. The solution proposed in [196] to just use NLO PDFs within the LO Monte Carlo can be supported by the fact that the difference between LO and NLO determinations is mostly due to the difference between LO and NLO DIS hard scattering matrix elements used in the fit for the global PDF analyses. In hadronic collider's processes the difference between matrix elements in the two perturbative approximations is much lower and so NLO PDFs combined with LO collider matrix elements within a Monte Carlo generator of events may give a good approximation. This is the case for example of PYTHIA, or equivalently of HERWIG: the main problem here is the substantial retuning of the parameters in the event generator needed for the inclusion of a NLO PDFs set.

Another possible set of PDFs can be defined including some dominant NLO corrections to the LO matrix elements. This is actually what the MSTW collaboration did when producing the MSTW08LO PDFs (see Ref. [81]): a sizable fraction of the large NLO and NNLO K factor for Drell-Yan comes from contributions which have the same kinematics as the LO. This allows for a simple rescaling of the LO cross-section for the Drell-Yan data, giving an intermediate solution between the two analyzed up to now.

In Ref. [197] other ideas on how to modify standard LO determinations were proposed. These ideas arise after better focusing on the main problem with LO PDFs. In the cited reference the observation that the LO fit quality mostly deteriorates because of the faster gluon evolution at small-x and the slower quark density evolution at large-x brings to the determination of a PDFs set, MRST2007lomod: in this set the NLO value for the strong coupling constant together with its two-loop running is used. This leads to smaller values of α_s in the low Q^2 region where the small-x data are concentrated, and thus to slower PDF evolution. Another choice can be made to avoid another consequence of the faster small-x gluon evolution: this leads through the momentum sum rule to depletion of the gluon content at medium-/large-x, which may cause a poor description of large-x fixed-target data. A possible way out is to relax the momentum sum rule, justifying this

	NLO	LO α_s =0.119	LO* α_s =0.119	LO α_s =0.130	LO* α_s =0.130
Total χ^2	1.16	1.74	1.76	1.68	1.74
$\langle\chi^{2\,(k)}\rangle$	1.25 ± 0.07	1.95 ± 0.21	1.89 ± 0.22	1.95 ± 0.19	1.94 ± 0.18
NMC-pd	0.97	1.43	1.13	1.18	1.12
NMC	1.72	2.05	1.68	1.74	1.72
SLAC	1.29	3.77	3.00	2.91	2.70
BCDMS	1.24	1.87	1.82	1.76	1.75
HERAI-AV	1.07	1.70	1.55	1.58	1.59
CHORUS	1.15	1.51	1.67	1.53	1.67
NTVDMN	0.45	0.69	0.71	0.71	0.78
ZEUS-H2	1.29	1.51	1.42	1.43	1.44
ZEUSF2C	0.78	1.75	1.26	1.56	1.34
H1F2C	1.51	1.77	2.00	1.81	2.02
DYE605	0.85	1.86	2.02	1.70	1.83
DYE886	1.26	1.99	2.52	2.59	3.11
CDFWASY	1.83	1.80	2.50	2.16	2.29
CDFZRAP	1.64	2.88	3.89	2.08	2.58
D0ZRAP	0.59	1.07	1.29	0.87	1.02
CDFR2KT	0.96	2.60	3.22	2.45	2.76
D0R2CON	0.83	1.18	1.56	1.17	1.35
$[M]$	1	1	1.16 ± 0.03	1	1.09 ± 0.03

Table 4.6: Fit quality for the global fit and for all experiments included in it for each of the NNPDF2.1 LO PDF sets. The corresponding values for the NNPDF2.1 NLO set of Ref. [103] are given for comparison. The value of the momentum integral $[M]$ Eq. (4.6) is also shown. All the fits have $N_{\rm rep} = 100$ replicas.

choice as an *ad hoc* phenomenological patch.

A last strategy that is worth discussing is depicted in Ref. [198] and is the basis for CT09MC1/MC2 PDF analysis. The main idea is that, as these LO parton sets are conceived to be used as a combination with Monte Carlo event generators, they should be determined by optimizing the agreement with the data of the predictions obtained by using them already in such a combination. This involves considering all the various modifications of the minimal LO framework discussed above, and also introducing suitable pseudodata to optimize the agreement with Monte Carlo generators.

4.3.1 Quality of the Fit

Four PDF sets are produced as a result of the variation between a standard LO fit and a LO* fit (in which the momentum sum rule is not imposed) on one side and of the value of α_S between 0.119 and 0.130 on the other. The combination of these variations produce a set of four different PDF sets. The running of α_S is always computed at LO in each single case.

The χ^2 of the four LO NNPDF2.1 sets, both for the global fit and for individual experiments, are collected in Table 4.6 and compared to the corresponding results of the NNPDF2.1 NLO set. The first line shows the total χ^2 and then also the χ^2 experiment by

experiment computed between the experimental data and the predictions obtained using the central value of the PDF set (replica 0, i.e. the mean over all other replicas) are shown for each case. The mean value over replicas of the total χ^2, computed replica by replica, is shown in the second line, while in the last one is given the value of the momentum integral. All these quantities, as described in Sect. 2.3, are computed including normalization uncertainties.

Comparing the four PDF sets, it is possible to say that varying the α_s value all the χ^2 are poorly affected, so that the case of a NLO running of α_s is not investigated. Relaxing the momentum sum rule instead on one side doesn't affect much the total χ^2, but on the other affects the experiment by experiment result: the description is improved for several DIS experiments, especially for HERA data, but at the same time hadronic data quality lowers significantly. On the whole, the fit quality of the four fits is almost unchanged: the values of $\langle \chi^{2\,(k)} \rangle$ differ from each other by less than a standard deviation.

Contrary to what previously discussed in the case of MSTW08LO (Ref. [81]), here can be seen that Drell-Yan data and DIS data are simultaneously fitted without the need for rescale of DY data. Surely an optimization of the LO determination here performed would bring to a better fit quality of these datasets, also in view of a possible combination with an event generator, but the purpose here is to built a PDFs set based on pure LO theory.

In Table 4.6 the value of the momentum integral

$$[M] \equiv \int_0^1 dx\, x\Sigma\left(x, Q^2\right) + \int_0^1 dx\, xg\left(x, Q^2\right) \ , \tag{4.6}$$

is also given for each LO PDF sets. These are determined at the starting scale $Q_0^2 = 2$ GeV^2, but note that the momentum integral $[M]$ does not depend on scale. A discussion of the behaviour of the momentum integral at LO, NLO and NNLO will be given in the next Chapter.

The conclusion is that analyzing the four sets produced and comparing them with the NLO PDF set previously introduced, a marked worsening in the fit quality is indeed observed with respect to the NLO case. Considering one by one the modifications introduced in the various LO cases no significant improvement is observed. A curious effect is found by relaxing positivity constraint on PDFs: the χ^2 of the LO fit then becomes only about 10% higher than in the NLO case. Of course, as at LO PDFs are subject to the probabilistic interpretation, the imposition of this constraint cannot be avoided. If relaxed, it leads to a negative gluon PDF at large-x, that also may give negative cross-sections.

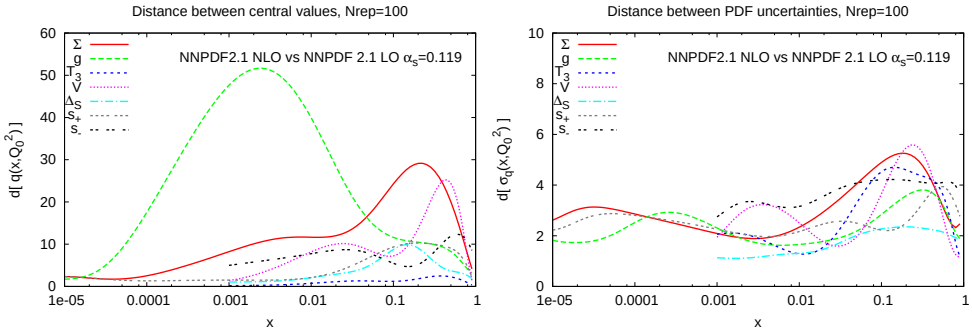

Figure 4.19: Distances between the reference LO and NLO NNPDF2.1 sets. Here and in subsequent figures in this Section, the left plot shows the distance between central values, while the right plot shows the distance between the uncertainties. Both have $\alpha_s(M_z) = 0.119$.

4.3.2 Parton Distributions

It is interesting to make a detailed comparison between PDF sets introduced above, extending the comparison also to parton distributions determined within other groups than the NNPDF collaboration. Again, the distances introduced in Appendix A of Ref. [87], and already used in Sect. 4.1.2 are computed. Descriptions of an identical underlying probability distribution have $d \sim 1$ while for statistically nonequivalent but consistent descriptions at the n-sigma level distances are of the order of $d \sim 7n$.

As a first comparison the standard NNPDF2.1 LO set (with α_s=0.119) is analyzed versus the NNPDF2.1 NLO set introduced and described in Sect. 4.1. This LO fit can be considered as the reference for the LO determination. It can be seen, looking at the distance plot represented in Fig. 4.19, that while the uncertainties are consistent within one sigma, the central values are separated by several sigmas. This tells us that the uncertainties are equivalent in the two cases, reflecting the fact that the two fits have been performed over the same data. The difference of many sigmas documented for central values reflects the fact that theoretical uncertainties are much larger for the LO case than the experimental error bands of either determinations, due to the lack of inclusion of higher order corrections.

Looking now into detail at single parton functions, the gluon is the one that present the largest difference: in the range of medium- small-x ($10^{-4} \leq x \leq 0.05$) is in fact possible to see a difference of more than five sigmas. This is consistent with the fact that the gluon decouples from LO observables. Also, in the same range of x but slightly shifted towards larger x values, the singlet and valence PDFs are lower than the NLO respective parton functions by more than three sigmas as can be seen in Fig. 4.20. In the same set of plots it is clear that the LO gluon is instead larger than the NLO determination. The only x

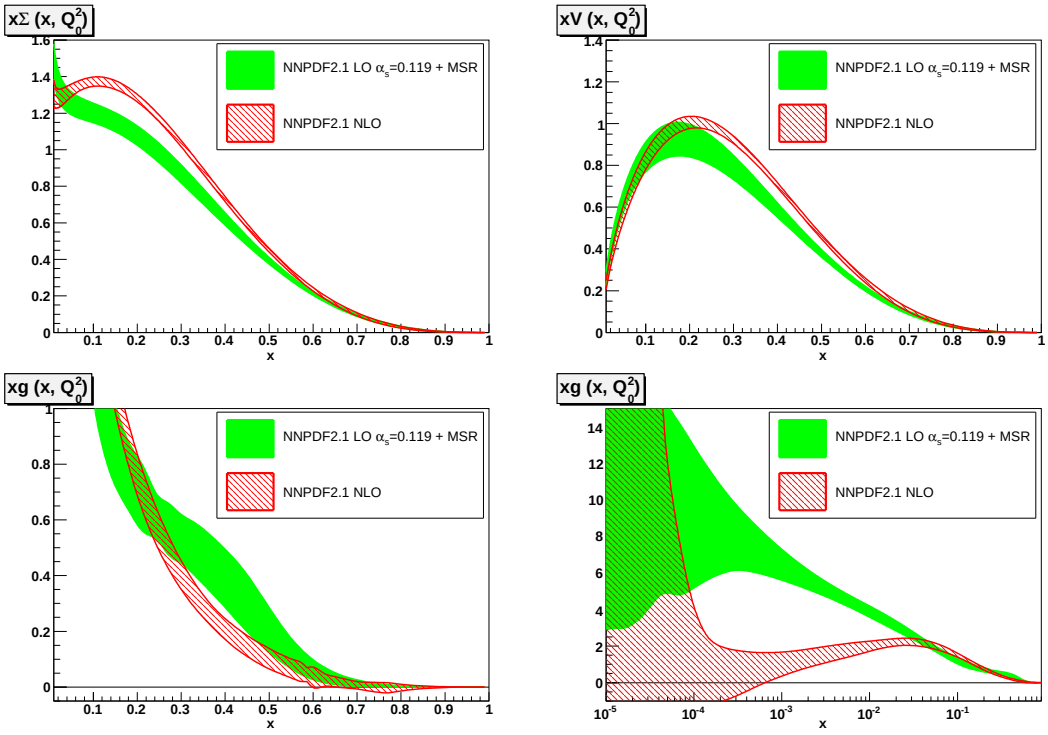

Figure 4.20: Comparison of the quark singlet, valence and gluon distributions for the pair of PDF fits whose distances are plotted in Fig. 4.19.

regions where consistency is found are at small-x, i.e. for $x \leq 10^{-4}$, because of the lack of information to constrain gluon PDFs there with the consequent blow up of error bands, and at large-x. The LO quark is rather smaller (by more than one sigma) than the NLO one for large-x ($x > 0.1$), but it becomes compatible with it at the one-sigma level for smaller x. Finally, the light sea and strangeness asymmetries are minimally affected and quite close at LO and NLO. The fact that the quark LO distributions are smaller or comparable to the NLO ones tells us that the Drell-Yan data actually have relatively little effect on the LO fit, other than through the determination of the $\bar{u} - \bar{d}$ light flavor asymmetry. The missing large NLO K-factors in Drell-Yan data should enhance the LO quark distributions in comparison to the NLO ones.

Looking again at Tab. 4.6, the reference LO set with $\alpha_s(M_Z) = 0.119$ is now compared with the other pure LO set with $\alpha_s(M_Z) = 0.130$. It can be seen the effect of evolving a larger value of α_s down to a scale $Q^2 \sim 10$ GeV^2 using LO evolution: in this way a value for α_s that is preferred by data in this region is reached. The larger value leads to a better description of scaling violations at low scale, and conversely.

It is interesting to look at Fig. 4.21 and Fig. 4.22: it can be seen that again uncertainties are not affected, because mostly driven by experimental uncertainties. Among central value distances, the gluon present the higher shift at medium- small-x, being smaller

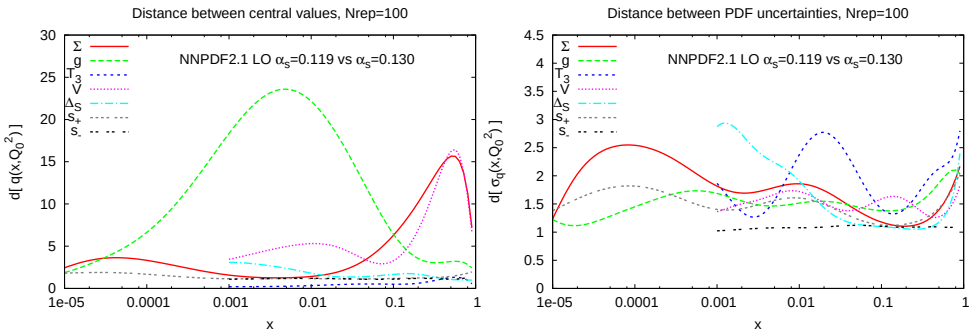

Figure 4.21: Distances between the NNPDF2.1 LO sets with α_s=0.119 and α_s=0.130.

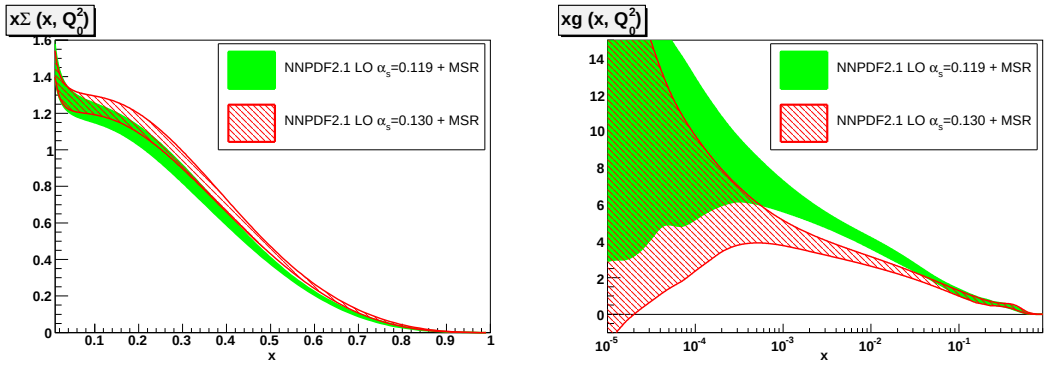

Figure 4.22: Comparison of the quark singlet, and gluon distributions for the pair of PDF fits whose distances are plotted in Fig. 4.21.

for the higher α_S value. The momentum sum rule instead constrains the gluon to larger values at large-x. The difference between the two gluon distributions is however at most of the same order of their uncertainty. Even if the gluon PDF is the only one significantly affected, also has to be noticed that the large-x singlet and valence quark PDFs increase somewhat when α_s is raised, especially at large-x ($x \sim 0.3$), where a shift of about two sigma is observed.

The comparison between the reference LO set and the set with same $\alpha_s(M_Z) = 0.119$ value but relaxed momentum sum rule condition (LO*) is given in Fig. 4.23. The main difference is seen in the medium-x gluon, as shown in Fig. 4.24: the LO* gluon is rather larger than the LO one. However, the central values for all quark PDFs are very close to the standard LO ones.

Next, the NNPDF LO sets are compared with PDF sets from other groups. A first comparison is made among NNPDF2.1 LO with $\alpha_s = 0.130$, MSTW08 LO [81] ($\alpha_s = 0.139$) and CTEQ6L1 [71] ($\alpha_s = 0.130$). In Fig. 4.25 the comparison of the three determinations for the singlet, isospin triplet and gluon distributions is shown. The comparison between

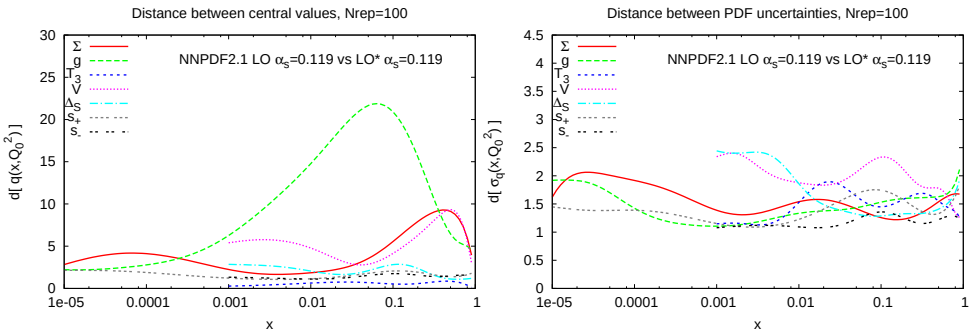

Figure 4.23: Distances between the NNPDF2.1 LO and LO* sets with α_s=0.119.

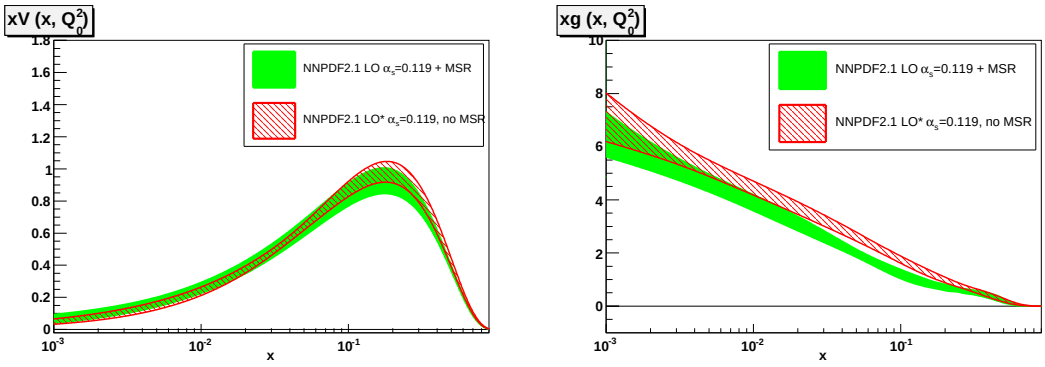

Figure 4.24: Comparison of the valence and gluon distributions for the pair of PDF fits whose distances are plotted in Fig. 4.23.

CTEQ6L1 and the other sets is somehow less precise for the lack of error bands of this delivery. Between the NNPDF analysis and MSTW08 LO again differences are especially large for the gluon distribution, both at small and large-x, and for the isospin triplet at large-x. However NNPDF and MSTW08 LO determinations are mostly compatible within the large error bands.

As a last comparison, are considered the modified LO PDF sets MRST2007lomod [197], obtained relaxing the momentum sum rule and using two-loop running of α_s, with $\alpha_s(M_z) = 0.121$, and the dedicated Monte Carlo sets of the CTEQ/TEA collaboration [198], CT09MC1, CT09MC2 and CT09MCS, based on an LO QCD analysis framework of data which, on top of the standard global dataset used for the NLO PDF determination, also includes a set of LHC pseudo-data generated using NLO PDFs. The CT09MC1 and CT09MC2 are performed without imposing the momentum sum rule and using respectively a one- and two-loop expressions for α_s. The CT09MCS is instead extracted from an analysis in which the two-loop strong coupling is used and the momentum sum rule is imposed during the fit. All these sets are compared to the reference NNPDF2.1 LO ($\alpha_s = 0.119$)

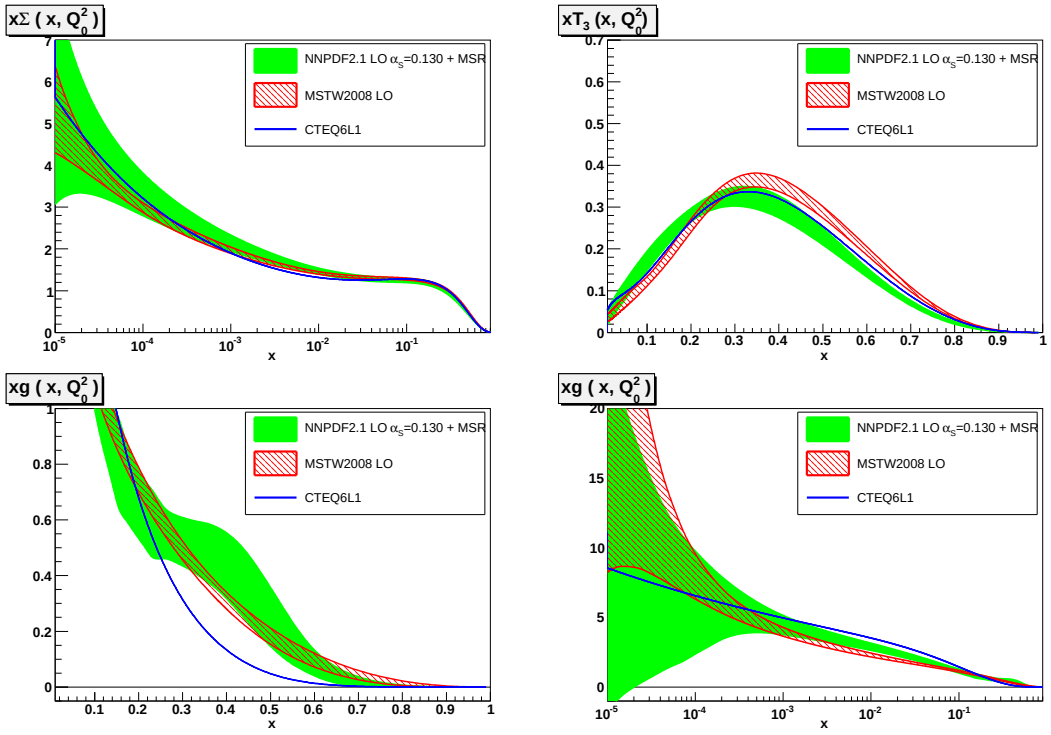

Figure 4.25: Comparison of LO PDFs: the quark singlet, triplet and gluon PDFs are shown for the NNPDF2.1, MSTW08 and CTEQ6L1 sets.

in Fig. 4.26. It is possible to argue that differences among these sets are due to the various different methodologies and assumptions on which is based each of them, being larger than the typical difference between the NNPDF2.1 LO and LO* sets.

4.4 Next-to-Next-to-Leading Order Results

In this Section NNLO PDFs are presented and compared to NNLO results from other groups and also to previous LO and NLO results discussed above. Parton sets at this perturbative order are mainly used for the computation of hadron colliders standard candle processes such as W, Z, top and Higgs production. Phenomenological consequences deriving from the use of the NNPDF2.1 NNLO set for LHC standard candles computation are discussed in the next Chapter.

The statistical features of the NNPDF2.1 NNLO fit are determined through the same statistical estimators used in previous Sections to analyze the NLO and LO cases. Here their values are given in Table 4.7 for the global fit and in Table 4.8 for individual experiments. In this second case, also the χ^2 for the NLO parton set is presented, to ease the comparison between the two fits as was done for the LO set.

It is observed a general similarity of the fit quality between NLO and NNLO results, as

$\chi^2_{\rm tot}$	1.16
$\langle E\rangle \pm \sigma_E$	2.22 ± 0.07
$\langle E_{\rm tr}\rangle \pm \sigma_{E_{\rm tr}}$	2.19 ± 0.09
$\langle E_{\rm val}\rangle \pm \sigma_{E_{\rm val}}$	2.27 ± 0.10
$\langle {\rm TL}\rangle \pm \sigma_{\rm TL}$	$(17 \pm 7)\ 10^3$
$\left\langle \chi^{2(k)}\right\rangle \pm \sigma_{\chi^2}$	1.23 ± 0.05
$\left\langle \sigma^{(\rm exp)}\right\rangle_{\rm dat}$(%)	11.9
$\left\langle \sigma^{(\rm net)}\right\rangle_{\rm dat}$(%)	3.2
$\left\langle \rho^{(\rm exp)}\right\rangle_{\rm dat}$	0.18
$\left\langle \rho^{(\rm net)}\right\rangle_{\rm dat}$	0.53

Table 4.7: Table of statistical estimators for the NNPDF2.1 NNLO fit with $N_{\rm rep} = 1000$ replicas.

Experiment	χ^2	$\chi^2_{\rm nlo}$	$\langle E\rangle$	$\left\langle \sigma^{(\rm exp)}\right\rangle_{\rm dat}$(%)	$\left\langle \sigma^{(\rm net)}\right\rangle_{\rm dat}$(%)	$\left\langle \rho^{(\rm exp)}\right\rangle_{\rm dat}$	$\left\langle \rho^{(\rm net)}\right\rangle_{\rm dat}$
NMC-pd	0.93	0.97	1.98	1.8	0.5	0.03	0.34
NMC	1.63	1.73	2.67	5.0	1.8	0.16	0.75
SLAC	1.01	1.27	2.05	4.4	1.8	0.31	0.78
BCDMS	1.32	1.24	2.38	5.7	2.6	0.47	0.58
HERAI-AV	1.10	1.07	2.16	7.6	1.3	0.06	0.44
CHORUS	1.12	1.15	2.18	15.0	3.5	0.08	0.37
FLH108	1.26	1.37	2.25	72.1	4.8	0.65	0.68
NTVDMN	0.49	0.47	1.74	21.0	14.0	0.04	0.64
ZEUS-H2	1.31	1.29	2.33	14.0	1.3	0.28	0.55
ZEUSF2C	0.88	0.78	1.89	23.0	3.7	0.07	0.40
H1F2C	1.46	1.50	2.48	18.0	3.5	0.27	0.36
DYE605	0.81	0.84	1.88	25.0	7.2	0.55	0.76
DYE866	1.32	1.27	2.40	21.0	8.7	0.23	0.48
CDFWASY	1.65	1.86	2.80	6.0	4.3	0.52	0.61
CDFZRAP	2.12	1.65	3.21	12.0	3.6	0.82	0.67
D0ZRAP	0.67	0.60	1.69	10.0	3.0	0.54	0.70
CDFR2KT	0.74	0.97	1.84	23.0	4.8	0.77	0.61
D0R2CON	0.82	0.84	1.89	17.0	5.5	0.78	0.62

Table 4.8: Same as Table 4.7 for individual experiments. All estimators have been obtained with $N_{\rm rep} = 1000$ replicas. Note that experimental uncertainties are always given in percentage. For reference the NNPDF2.1 NLO χ^2 for the various experiments is also provided.

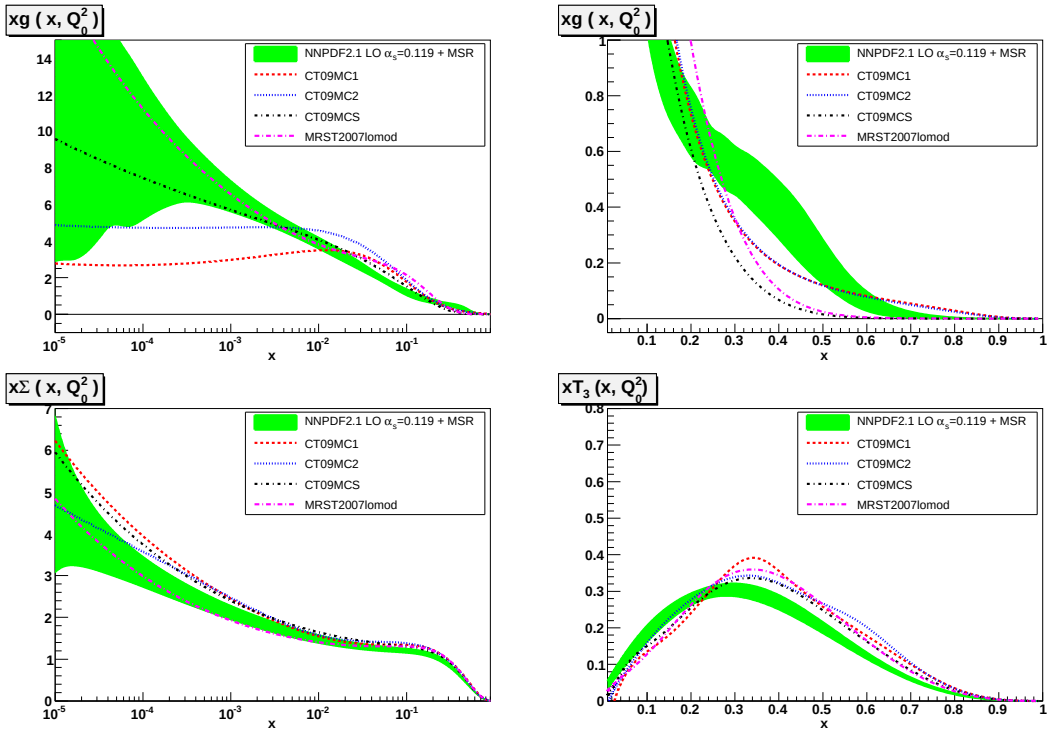

Figure 4.26: Comparison of the NNPDF2.1 LO PDF to modified LO PDF sets: MRST2007lomod, CT09MC1, CT09MC2 and CT09MCS.

also between all the other statistical estimators. Looking into the detail of the experiment by experiment χ^2 of Table 4.8 the difference is less then 10% for each experiment. Only SLAC, W asymmetry and CDF jet data differ more than that (because the NNLO fit gives a better quality), and the Z rapidity distribution (for which the quality instead is worse at NNLO). It is also interesting to notice that the description of H1F2C (the HERA F_2^c data) is almost the same and slightly better than at NLO: while at NLO an optimized cut and a tuning of the treatment of heavy quarks was needed to give a proper description of this set, here the FONLL-C general mass scheme is introduced (described in Sect. 1.6) that permits to release the kinematic cut, obtaining a larger dataset.

The distribution of $\chi^{2(k)}$, $E_{\rm tr}^{(k)}$, and training lengths among the $N_{\rm rep} = 1000$ NNPDF2.1 NNLO replicas are shown in Fig. 4.27 and Fig. 4.28 respectively. In the latter histogram it is visible that not all the replicas stop dynamically and a cospicuous fraction of replicas ($\sim 20\%$) stops at the maximum training length $N_{\rm gen}^{\rm max}$. This fraction is not too higher than the one in the NLO fit. As was done in the NLO analysis, also here is checked that the impact of these replicas is negligible: in fact, discarding all replicas that do not stop dynamically the PDFs change by an amount which is smaller than a statistical fluctuation. Raising the training length reduces also in this case the fraction of unstopped replicas, proving that the problem is only of computational efficiency.

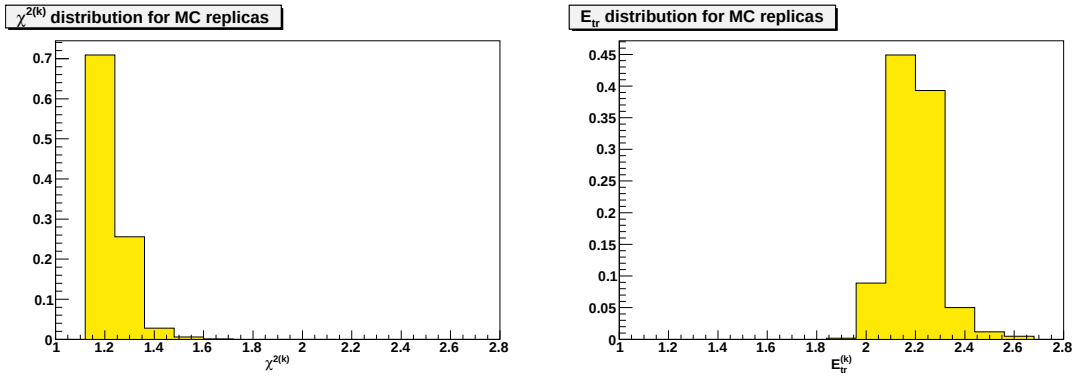

Figure 4.27: Distribution of $\chi^{2(k)}$ (left) and $E_{\rm tr}^{(k)}$ (right), over the sample of $N_{rep} = 1000$ replicas.

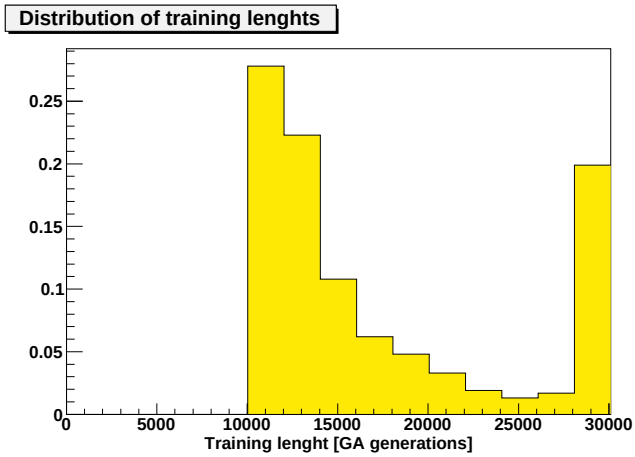

Figure 4.28: Distribution of training lengths over the sample of $N_{rep} = 1000$ replicas.

4.4.1 Parton Distributions

The parton distributions presented here, the NNPDF2.1 NNLO set, is plotted in the basis in which the set is parametrized and directly compared to the NLO fit in Figs. 4.29 and 4.30 at the input scale $Q_0^2 = 2$ GeV^2.

Observing the distances between these two sets plotted in Fig. 4.31, it can be seen that the central values are almost everywhere within one sigma from each other, showing the stability of PDFs while going from NLO to NNLO. The largest variations are observed for quarks at $x \sim 0.1$, while the small-x PDFs (gluon and light quark sea) are very similar to their NLO counterparts. The fact that theoretical uncertainties are not included in PDF error bands and the quite similar quality of the NLO and NNLO fits explain the particularly small distances shown in Fig. 4.31 for PDF uncertainties.

One of the useful exercises that can be done using the NNLO release is to assess the impact of NNLO corrections on physical observables. This is easier if PDFs are

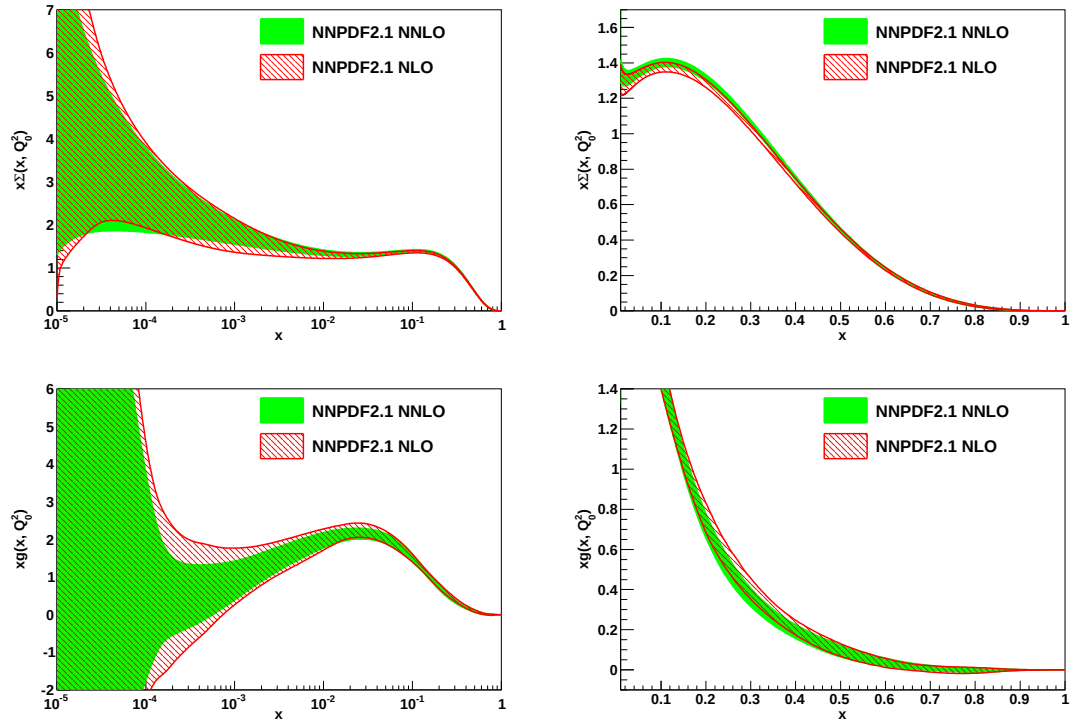

Figure 4.29: Comparison of NNPDF2.1 NLO and NNLO singlet sector PDFs, computed using $N_{\rm rep} = 1000$ replicas from both sets. All error bands shown correspond to one sigma.

represented in the flavor basis at a typical hard scale as in Fig. 4.32, where the ratio between NNLO and NLO parton functions is plotted as a function of x at $Q^2 = 10^4$ GeV^2. The most noticeable difference is at $x \sim 10^{-3}$ for the light quark sea, where the two plotted bands almost don't overlap. Other changes are visible at small-x in the quark distributions, that are larger in that region, and also smaller large-x quarks. As a consequence of evolution is also observed a larger small-x gluon.

Another interesting comparison is performed between NNPDF2.1 NNLO PDFs to those from the MSTW08 NNLO set. A more precise estimate of likeness and difference is made using the same $\alpha_s\,(M_Z)$=0.119 value. The two PDF sets are plotted in Figs. 4.33 and 4.34. The MSTW08 NNLO gluon, unlike its NNPDF2.1 counterpart, is unstable at small-x, where it becomes very negative. Unusually small uncertainty bands are observed for MSTW determination and in general the two sets show reasonable agreement for central values. The only significant difference is found in the strange distribution, probably due to the extremely restrictive parametrization given for the $s + \bar{s}$ and $s - \bar{s}$ MSTW08 PDFs.

A last comparison is made in Figs. 4.35 and 4.36 with the ABKM09 NNLO set (with fixed flavor number $n_f = 3$) [94]. A NNPDF2.1 NNLO PDF set with $\alpha_s = 0.114$ is chosen, as the ABKM set is only provided for $\alpha_s\,(M_Z) = 0.1135 \pm 0.0014$. Further-

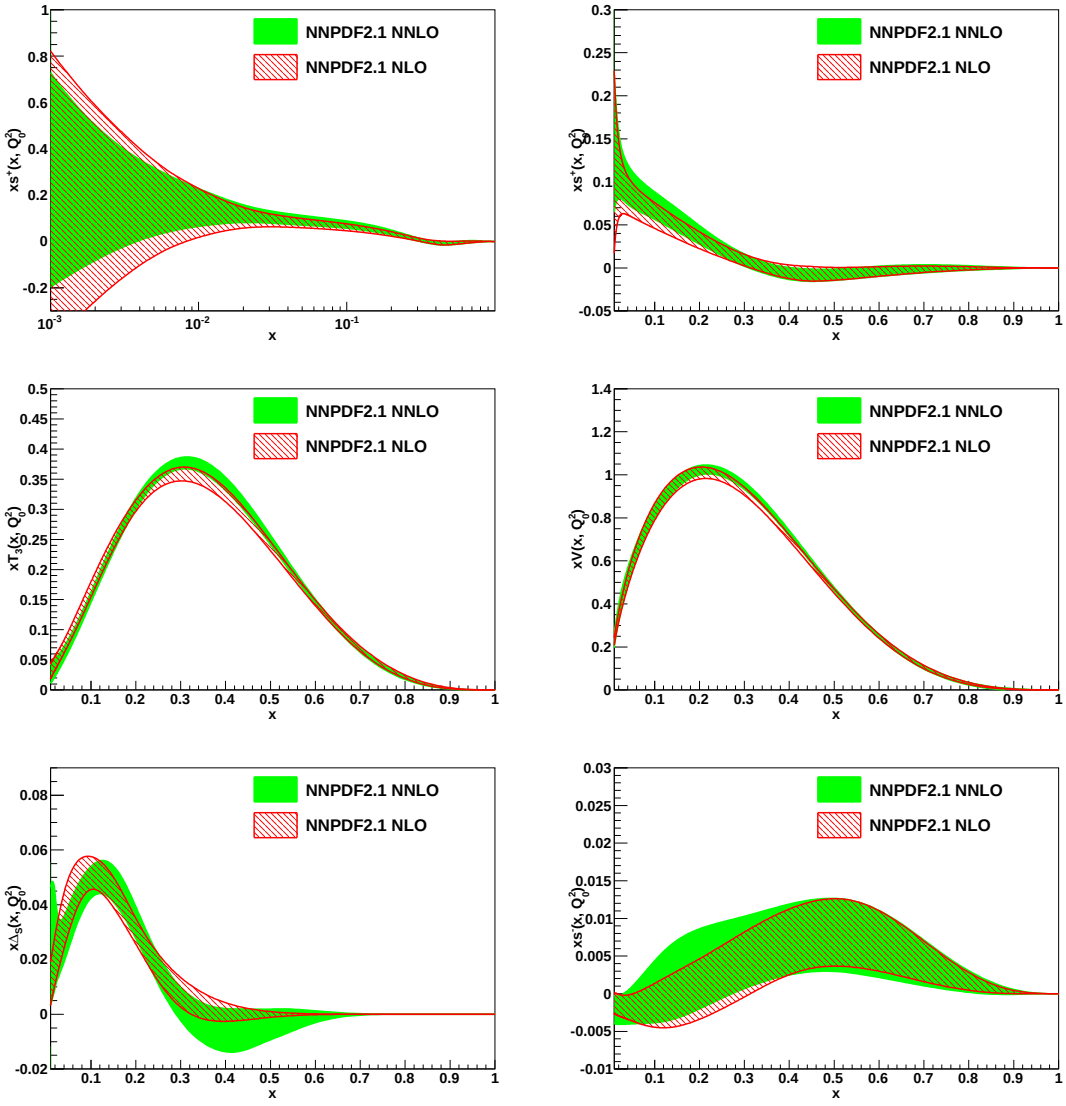

Figure 4.30: Same as Fig. 4.29 for the non-singlet sector PDFs.

more for this set (and its NLO counterpart) only combined PDF+α_s uncertainties can be determined, unlike other sets for which PDF uncertainties with fixed α_s may also be computed. The energy scale is fixed at $Q_0^2 = 2$ GeV^2 and in general is observed a somewhat bad agreement.

4.5 Perturbative Stability

It can be exploited the fact that to determine LO, NLO, and NNLO PDF sets the same data and the same methodology are used in the three cases: because of this can be addressed issues of perturbative stability by comparing individual PDFs but also looking at the behaviour of the total momentum fraction carried by partons.

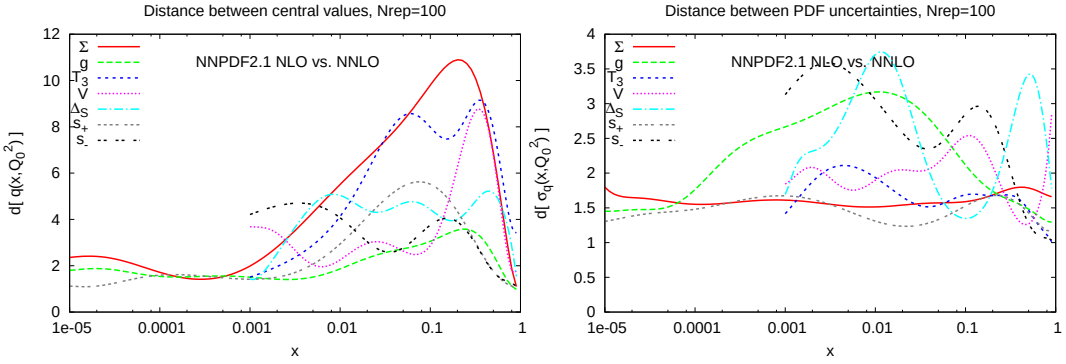

Figure 4.31: Distances between the NNPDF2.1 NLO and NNLO parton sets shown in Figs. 4.29 and 4.30. All distances are computed from sets of $N_{\rm rep} = 100$ replicas.

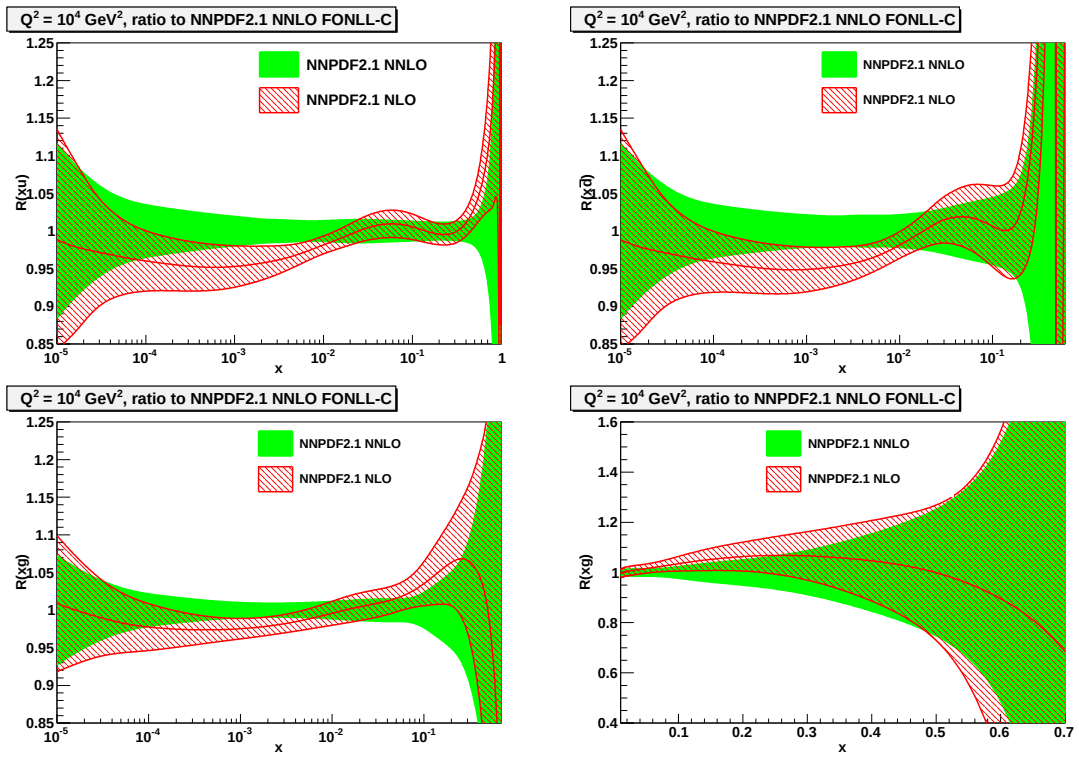

Figure 4.32: Comparison between NNPDF2.1 NLO and NNLO light quark and gluon PDFs at $Q^2 = 10^4$ GeV^2. The results have been obtained with $N_{\rm rep} = 1000$ replicas. All curves are shown as ratios to the central NNPDF2.1 NNLO result.

The LO, NLO, and NNLO parton functions are plotted together to assess the perturbative stability of NNPDF determination, as can be seen in Figs. 4.37 and 4.38 at the starting scale Q_0^2=2 GeV^2 in the basis in which they are independently parametrized by neural networks. In Figs. 4.39 and 4.40 a similar comparison is provided but this time at the scale Q^2=(100 GeV)^2 in the basis of individual flavors. All the error bands take into

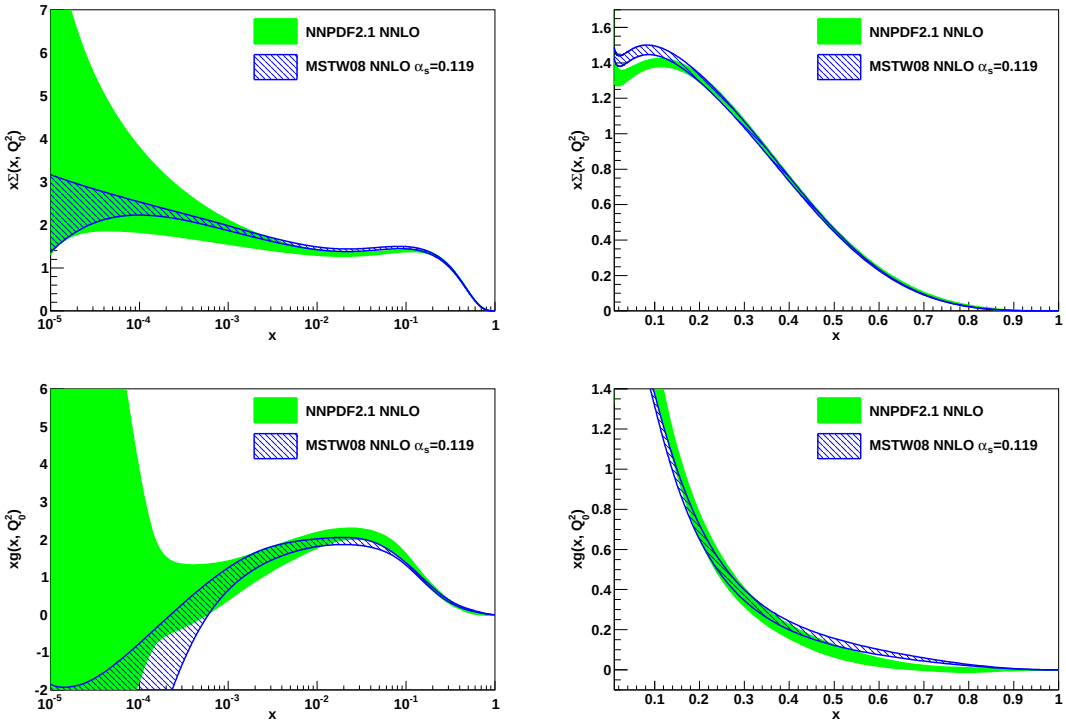

Figure 4.33: The NNPDF2.1 NNLO singlet sector PDFs, compared to MSTW08 PDFs. The results for NNPDF2.1 NNLO have been obtained with $N_{\rm rep} = 1000$ replicas. All PDF errors are given as one sigma uncertainties. In the comparison a common value of $\alpha_s\,(M_Z)$=0.119 has been used.

account also possible deviations from Gaussianity, as they are defined as 68% confidence levels, rather than as standard deviations.

These comparisons manifest the excellent convergence of the perturbative expansion within the kinematic region covered by the experimental data. Also, in the kinematic region where resummation is supposed to affect the result (small- and large-x regions) [199, 200] and the perturbative stability should start to be compromised, no evidence of instability is seen in the PDFs, thus suggesting that resummation corrections are smaller than current PDF uncertainties (at small-x, this is borne out by the dedicated study of Refs. [163, 164]).

Focusing on the differences between NNLO and NLO parton functions, it is possible to see that in almost all cases there is agreement within uncertainties. This is true at the initial scale (Figs. 4.37-4.38), where the NNLO central value is within (or just outside) the NLO uncertainty band, and it is still better at higher scales (Figs. 4.39-4.40) The only noticeable differences are at $Q_0^2 = 2\text{GeV}^2$ for the isospin triplet distribution around the valence peak $x \sim 0.3$, where the NLO and NNLO bands overlap, but the NNLO central value is clearly outside the NLO band, and at higher scale again is observed as in Fig. 4.32

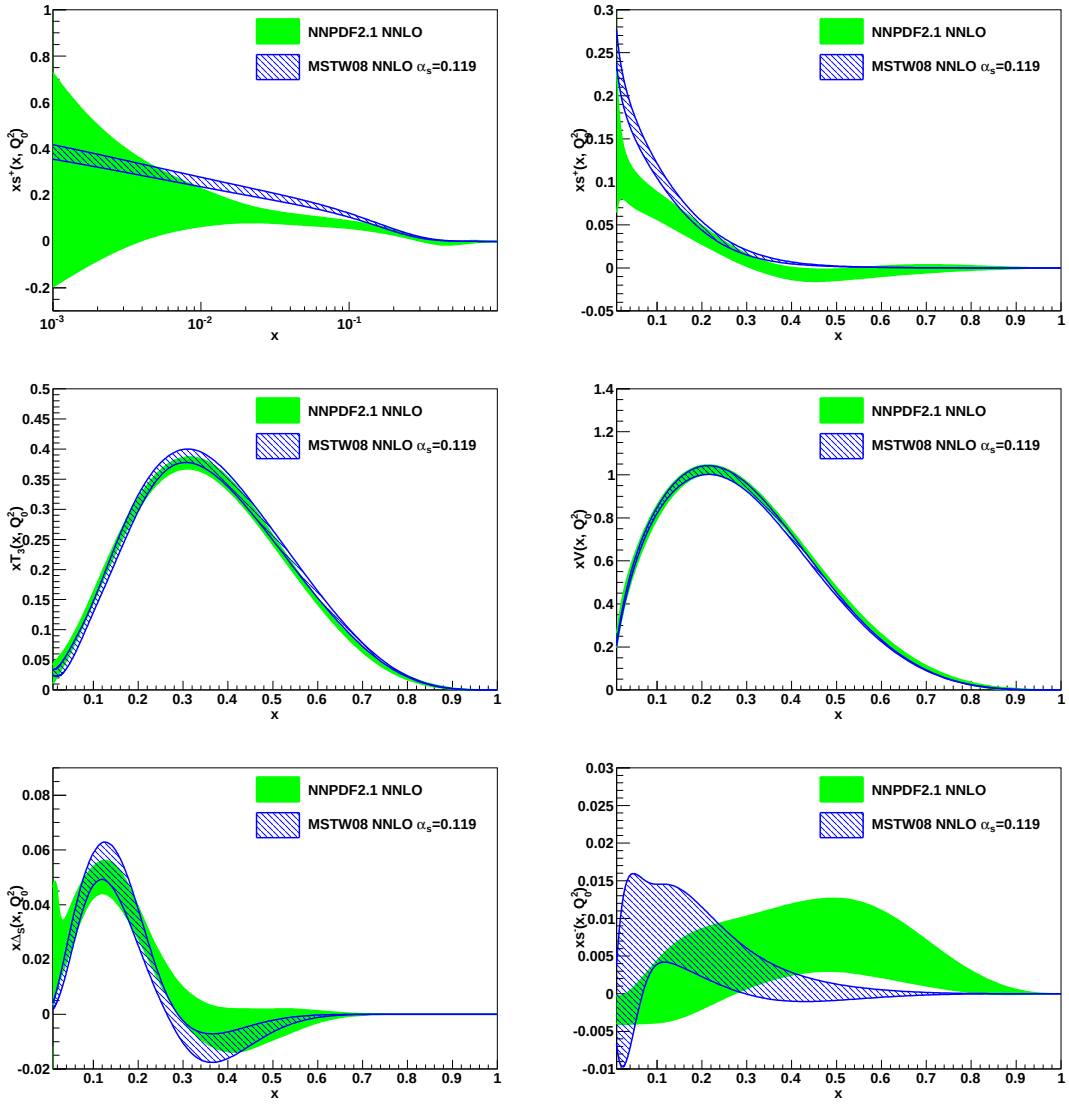

Figure 4.34: Same as Fig. 4.33 for the non-singlet sector PDFs.

a small discrepancy in light quark distributions for $x \gtrsim 10^{-3}$.

It can be concluded that effectively PDF uncertainties only reflect the data uncertainty. Theoretical uncertainties due to higher orders in perturbative QCD excluded in the analysis are not taken into account in the shown error bands. These kind of uncertainty can be estimated by varying the renormalization and factorization scale during the PDF fit, and at NLO by a direct comparison with the NNLO results. Thanks to this comparison it can be safely said that at present the size of the uncertainty due to theoretical inaccuracy is reasonably smaller than the PDF uncertainties, derived exclusively from data.

The same cannot be said for the LO PDFs. Here in fact it is clear that central values differ by several standard deviations from NLO parton functions. The situation improves

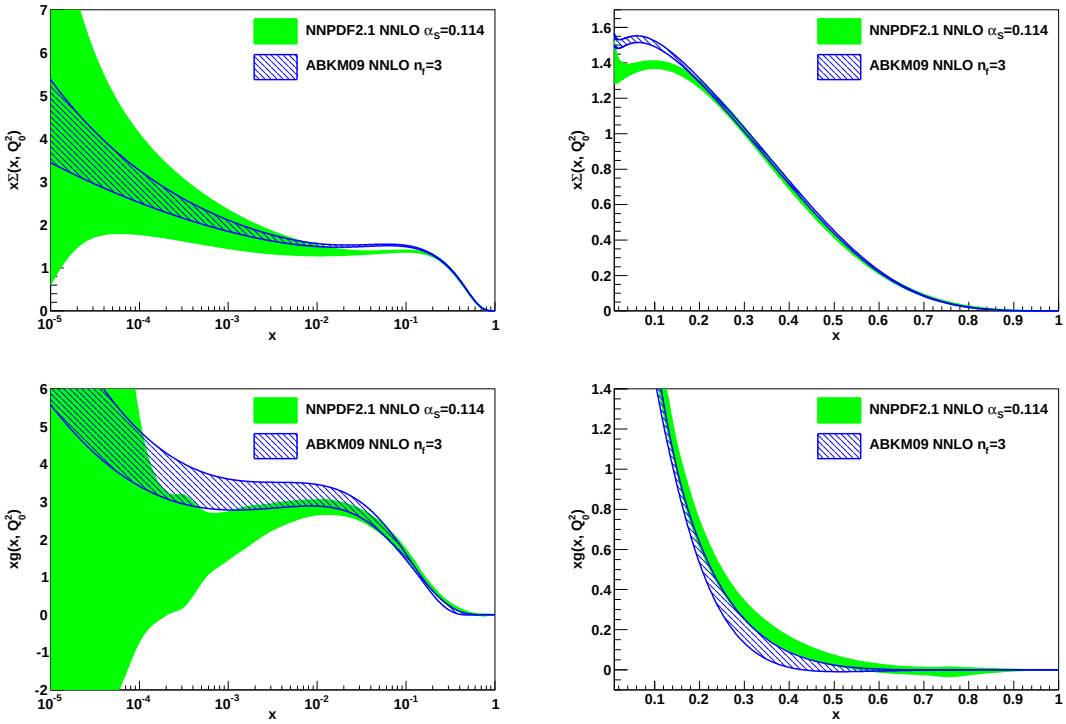

Figure 4.35: The NNPDF2.1 NNLO singlet sector PDFs, compared to the ABKM09 three-flavor set. The results for NNPDF2.1 NNLO have been obtained with $N_{\rm rep} = 100$ replicas. The NNPDF2.1 set with $\alpha_s = 0.114$ is shown because ABKM PDFs are only available for this value of α_s. Note that for ABKM uncertainties also include the uncertainty on α_s while for NNPDF they are pure PDF uncertainties.

somewhat at high scale (Figs. 4.39-4.40), but the difference between LO and NLO remains large for the gluon. Hence, it can be concluded that in this case the theoretical uncertainty is dominant, and the error represented by the plotted bands is only a fraction of the effective total error. These conclusions are true for NNPDF sets as for PDFs from other groups.

The value of the total momentum carried by quarks and gluons and its dependence on the perturbative order provide a strong consistency check of the perturbative QCD framework. The momentum fraction carried by a parton distribution is

$$[q]\left(Q^2\right) \equiv \int_0^1 dx\, xq\left(x, Q^2\right) \ . \tag{4.7}$$

Using the LO*, NLO* and NNLO* PDF sets, where the * indicates that the momentum

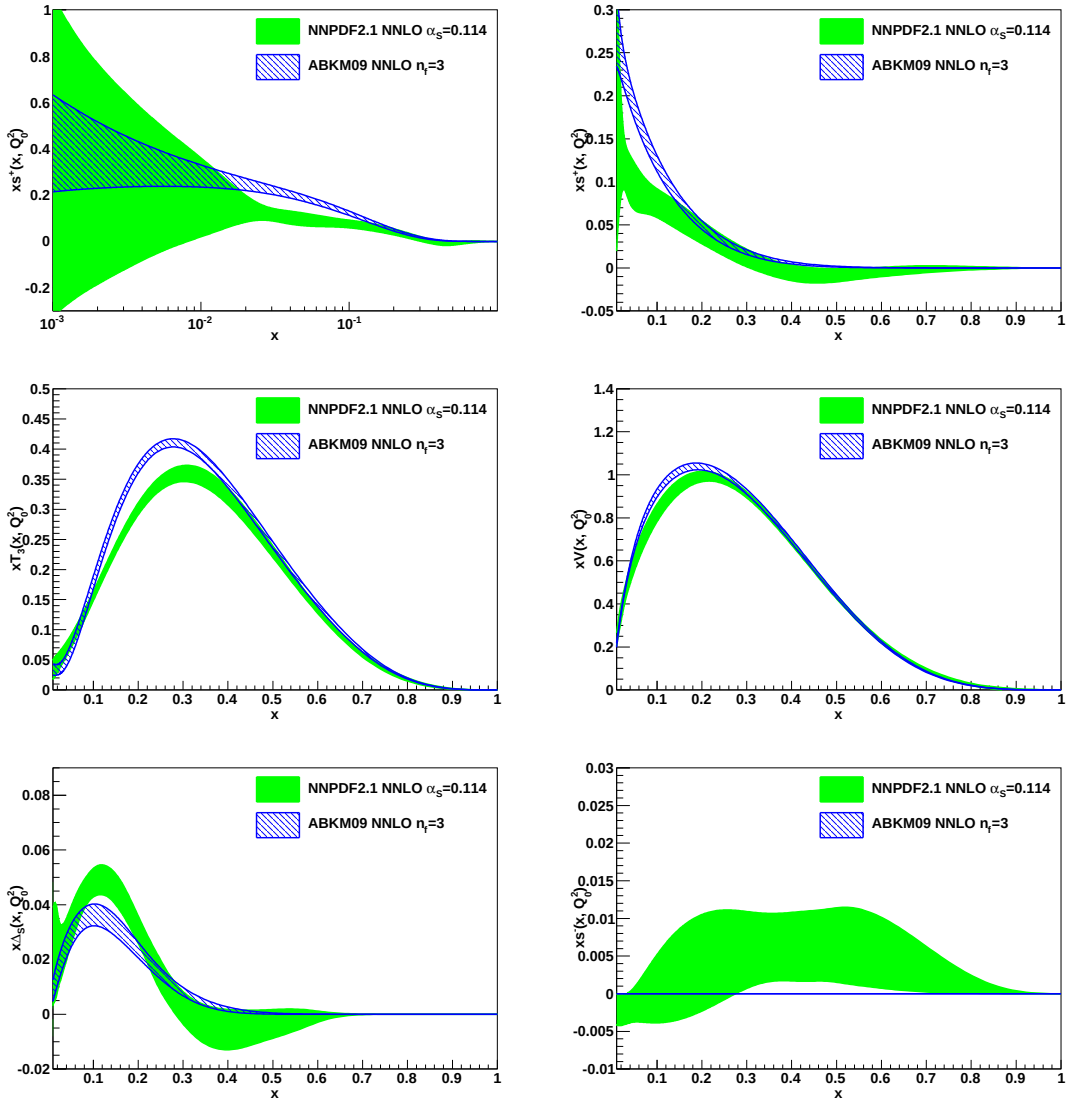

Figure 4.36: Same as Fig. 4.35 for the non-singlet sector PDFs.

sum rule has been relaxed, the total momentum $[M] = [\Sigma] + [g]$ carried by partons is

$$\begin{aligned} [M]_{\rm LO} &= 1.161 \pm 0.032\,, \\ [M]_{\rm NLO} &= 1.011 \pm 0.018\,, \\ [M]_{\rm NNLO} &= 1.002 \pm 0.014\,. \end{aligned} \tag{4.8}$$

where the uncertainty is only from PDFs (and thus does not include any theoretical uncertainty).

Estimating the theoretical uncertainty as the difference between results at two subsequent perturbative orders, is seen that at LO the theoretical uncertainty is dominant, as

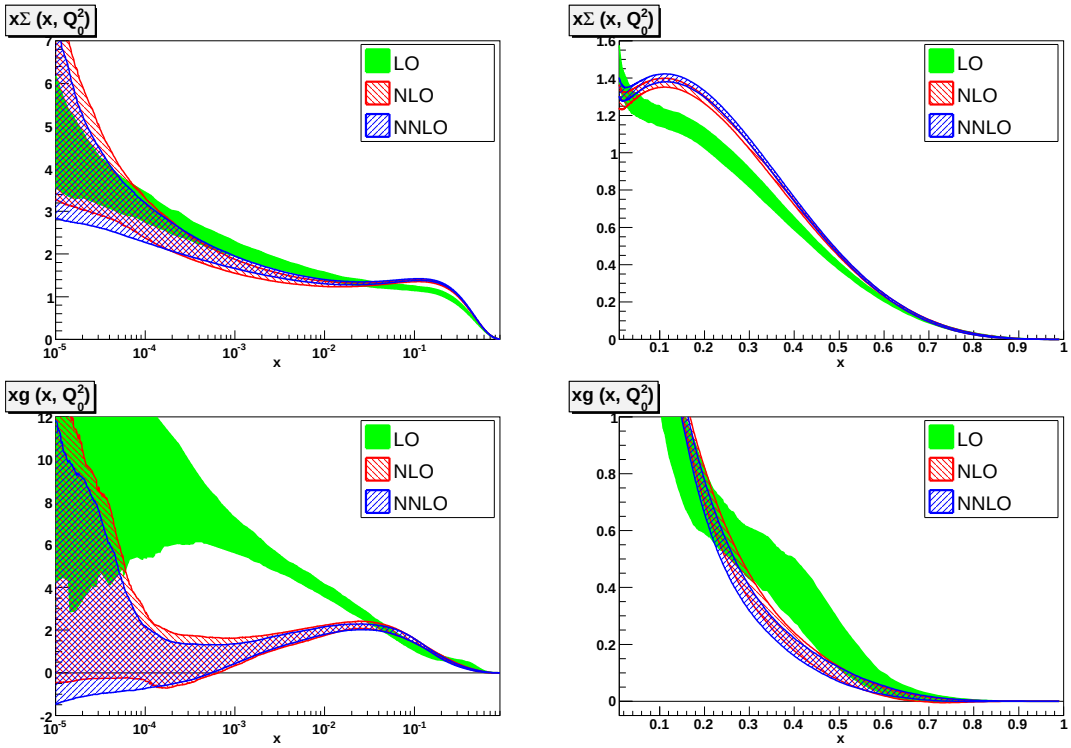

Figure 4.37: Comparison of NNPDF2.1 LO, NLO and NNLO singlet sector PDFs at Q_0^2=2 GeV^2. All uncertainty bands are defined as 68% confidence levels.

already concluded from the PDF plots Figs. 4.37-4.38 above. On the other hand already at NLO the theoretical uncertainty is half of the PDF uncertainty, $\Delta^{\rm th}\left[M\right]_{\rm NLO} = 0.01$, and thus at NNLO the theoretical uncertainty is likely to be negligible.

It is also interesting to determine the momentum fraction carried by individual PDFs. These are tabulated in Tables 4.9-4.10 at a low scale $Q_0^2 = 2$ GeV^2 and at a high scale $Q^2 = 10^4$ GeV^2, both before (Table 4.9, * PDF sets) and after (Table 4.10, standard PDF sets) imposing the momentum sum rule.

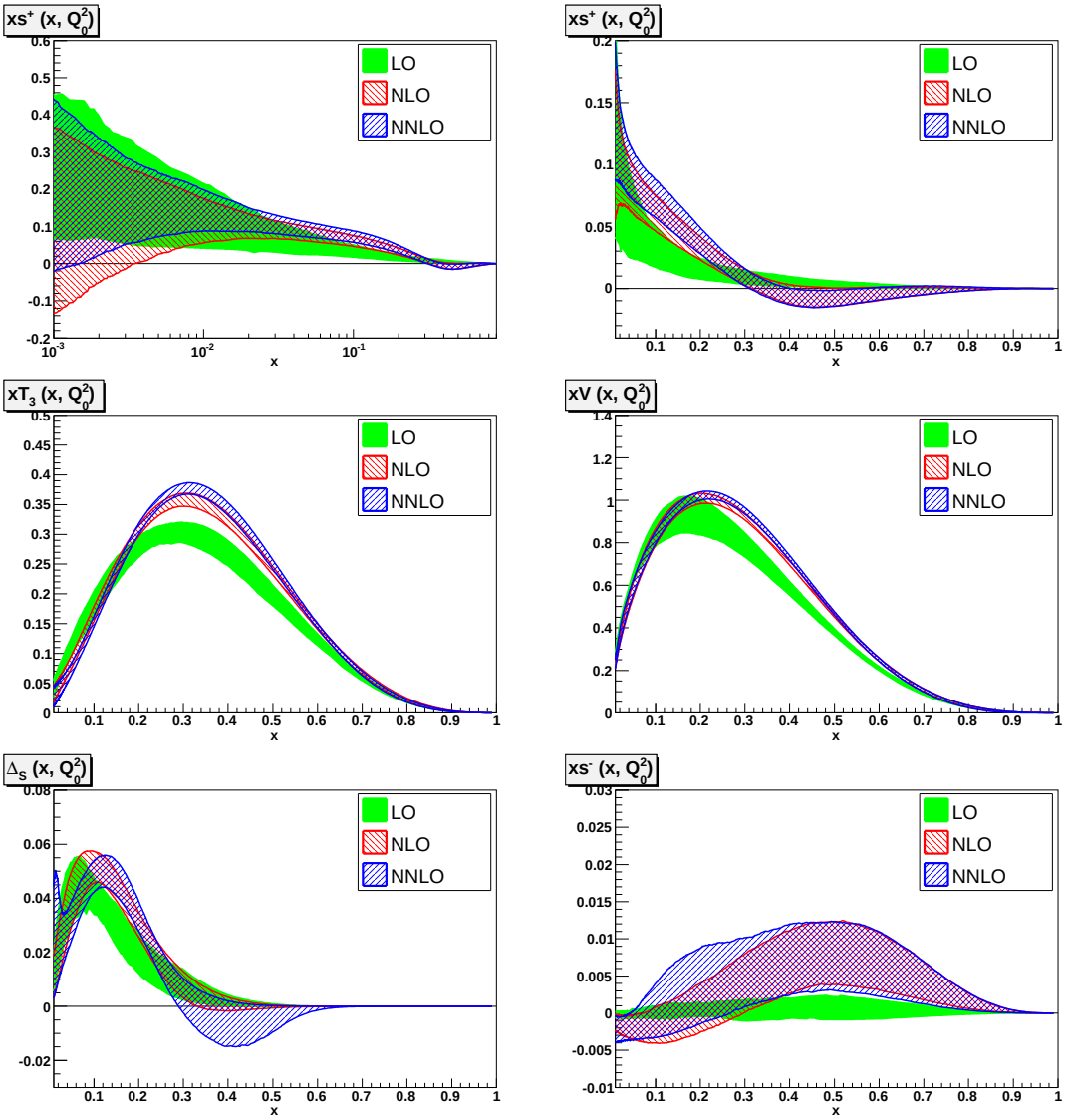

Figure 4.38: Same as Fig. 4.37 for the non-singlet sector PDFs.

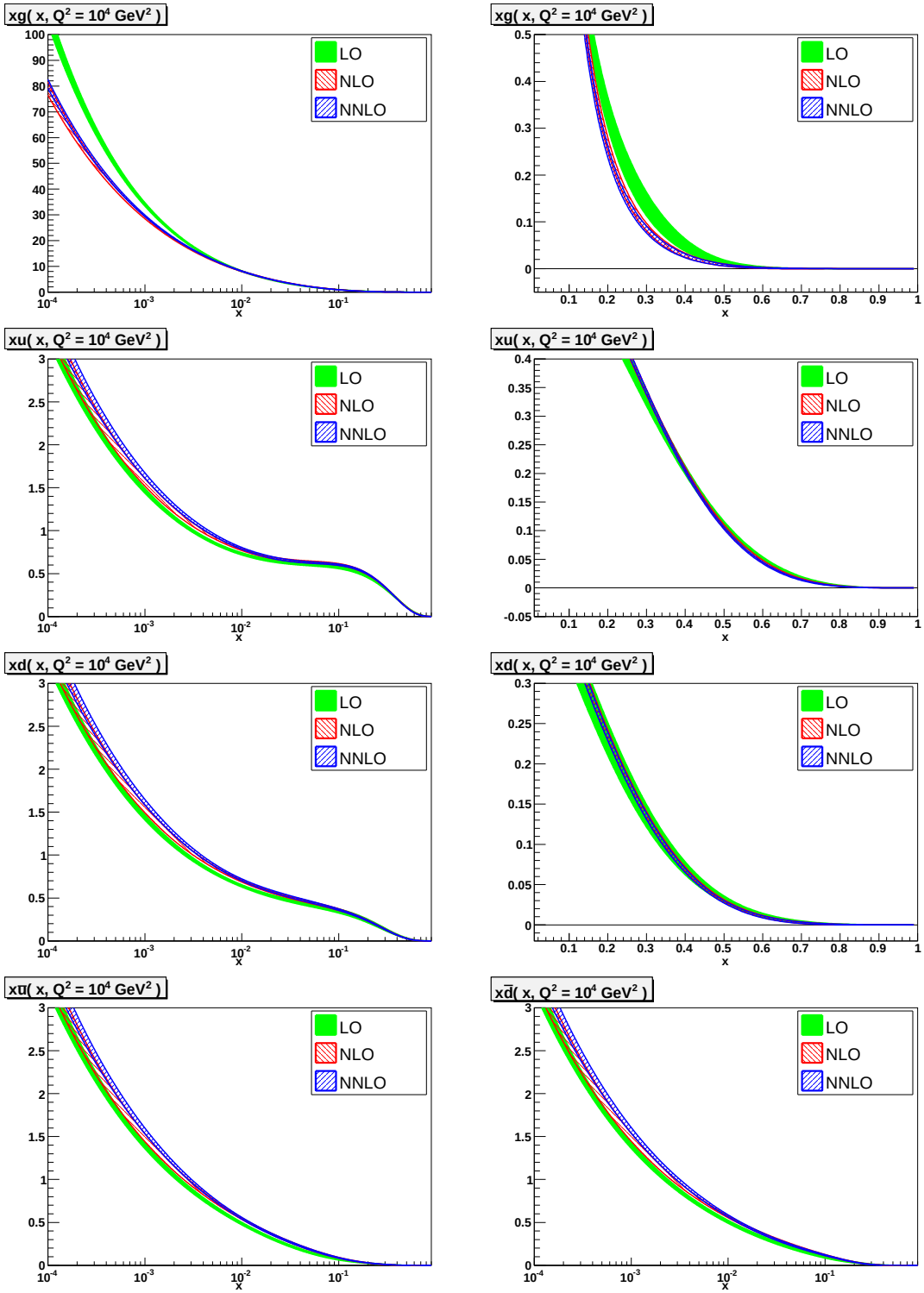

Figure 4.39: Comparison of NNPDF2.1 LO, NLO and NNLO PDFs in the flavor basis at $Q^2 = 10^4$ GeV^2: light quarks and gluon.

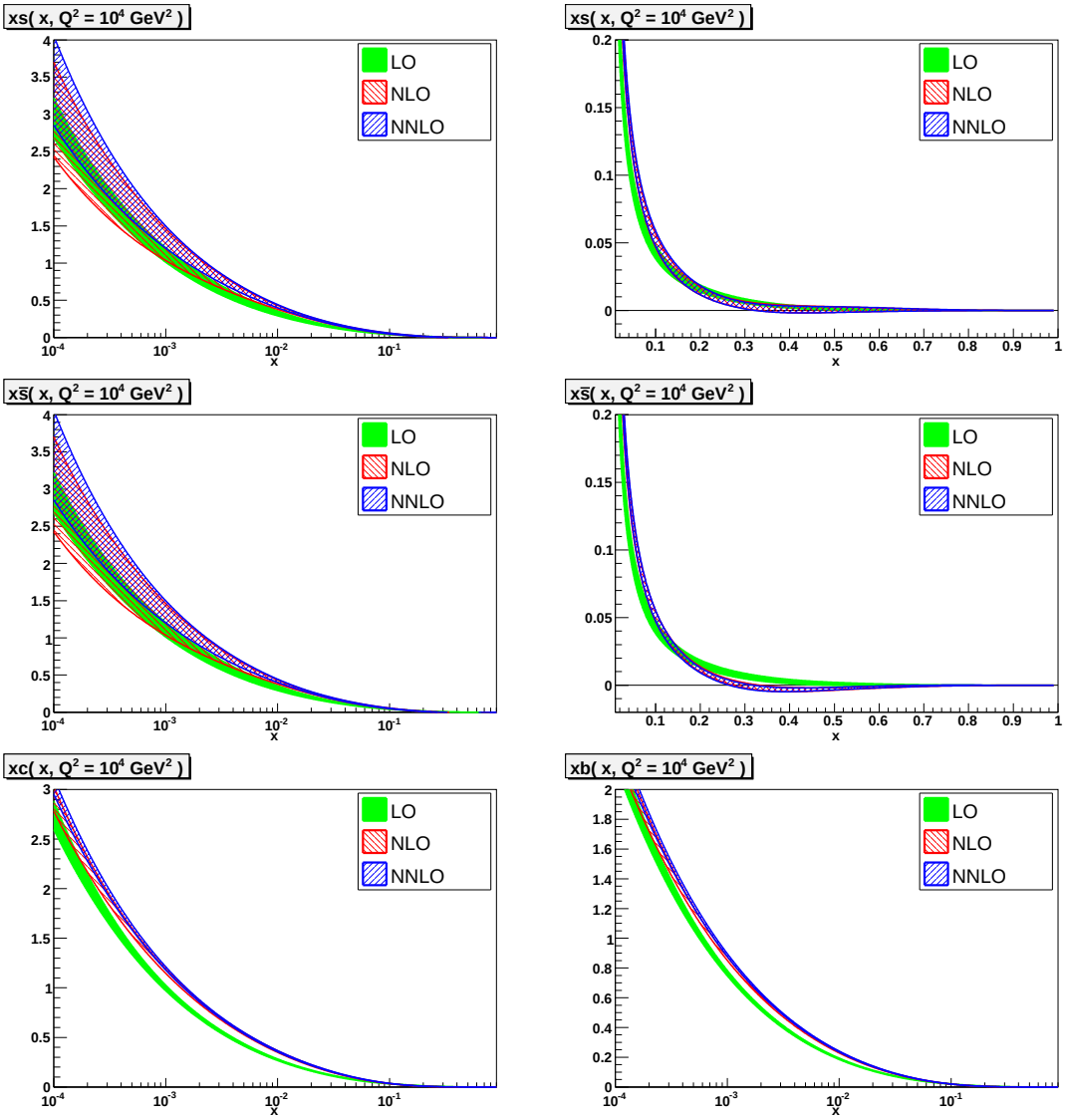

Figure 4.40: Comparison of NNPDF2.1 LO, NLO and NNLO PDFs in the flavor basis at $Q^2 = 10^4$ GeV^2: strange and heavy quarks.

PDF combination	LO*	NLO*	NNLO*
$[\Sigma + g]$	1.161 ± 0.032	1.011 ± 0.018	1.002 ± 0.014
$Q_0^2 = 2$ GeV^2			
$[\Sigma]\,(Q_0^2)$	0.550 ± 0.025	0.591 ± 0.010	0.602 ± 0.010
$[g]\,(Q_0^2)$	0.612 ± 0.028	0.421 ± 0.021	0.400 ± 0.018
$[(u+\bar{u})]\,(Q_0^2)$	0.346 ± 0.015	0.371 ± 0.005	0.376 ± 0.005
$\left[\left(d+\bar{d}\right)\right]\,(Q_0^2)$	0.192 ± 0.011	0.206 ± 0.005	0.209 ± 0.003
$[(s+\bar{s})]\,(Q_0^2)$	0.012 ± 0.004	0.014 ± 0.006	0.017 ± 0.006
$Q^2 = 10^4$ GeV^2			
$[\Sigma]\,(Q^2)$	0.546 ± 0.017	0.528 ± 0.009	0.527 ± 0.005
$[g]\,(Q^2)$	0.615 ± 0.020	0.486 ± 0.018	0.475 ± 0.011
$[(u+\bar{u})]\,(Q^2)$	0.264 ± 0.009	0.256 ± 0.003	0.255 ± 0.002
$\left[\left(d+\bar{d}\right)\right]\,(Q^2)$	0.165 ± 0.007	0.160 ± 0.002	0.159 ± 0.001
$[(s+\bar{s})]\,(Q^2)$	0.048 ± 0.003	0.047 ± 0.004	0.048 ± 0.004
$[(c+\bar{c})]\,(Q^2)$	0.041 ± 0.002	0.039 ± 0.002	0.039 ± 0.001
$\left[\left(b+\bar{b}\right)\right]\,(Q^2)$	0.027 ± 0.001	0.025 ± 0.001	0.025 ± 0.001

Table 4.9: Momentum fractions of various PDF combinations at low scale $Q_0^2 = 2$ GeV^2 and high scale $Q^2 = 10^4$ GeV^2 when the momentum sum rule is not imposed (LO*, NLO* and NNLO* PDF sets). All results are obtained with $N_{\rm rep} = 100$ replicas.

PDF combination	LO	NLO	NNLO
$[\Sigma + g]$	1	1	1
$Q_0^2 = 2$ GeV^2			
$[\Sigma]\,(Q_0^2)$	0.521 ± 0.023	0.590 ± 0.009	0.609 ± 0.013
$[g]\,(Q_0^2)$	0.479 ± 0.022	0.411 ± 0.009	0.391 ± 0.012
$[(u+\bar{u})]\,(Q_0^2)$	0.328 ± 0.012	0.371 ± 0.005	0.381 ± 0.007
$\left[\left(d+\bar{d}\right)\right]\,(Q_0^2)$	0.181 ± 0.010	0.206 ± 0.004	0.211 ± 0.005
$[(s+\bar{s})]\,(Q_0^2)$	0.012 ± 0.005	0.013 ± 0.006	0.017 ± 0.005
$Q^2 = 10^4$ GeV^2			
$[\Sigma]\,(Q^2)$	0.492 ± 0.010	0.523 ± 0.003	0.529 ± 0.004
$[g]\,(Q^2)$	0.509 ± 0.010	0.477 ± 0.003	0.471 ± 0.005
$[(u+\bar{u})]\,(Q^2)$	0.245 ± 0.007	0.255 ± 0.003	0.257 ± 0.004
$\left[\left(d+\bar{d}\right)\right]\,(Q^2)$	0.150 ± 0.006	0.159 ± 0.002	0.159 ± 0.002
$[(s+\bar{s})]\,(Q^2)$	0.041 ± 0.003	0.046 ± 0.003	0.048 ± 0.002
$[(c+\bar{c})]\,(Q^2)$	0.033 ± 0.001	0.0383 ± 0.0004	0.0393 ± 0.0006
$\left[\left(b+\bar{b}\right)\right]\,(Q^2)$	0.021 ± 0.001	0.0245 ± 0.0002	0.0249 ± 0.0003

Table 4.10: Same as Table 4.9, but when the momentum sum rule is imposed (LO, NLO and NNLO PDF sets).

Chapter 5

LHC Phenomenology

This Chapter is dedicated to the study of phenomenology at the LHC. The target of modern parton fits is to provide a reliable description of the various components inside the nucleon. For this reason is of primary importance to compute predictions at LHC, benchmark the results against other groups predictions and physical measurements of the considered observables.

After a brief discussion on parton luminosities (Sect. 5.1), the main topic of the Chapter is treated in Sect. 5.2. Then other minor but not less important issues are discussed: NuTeV anomaly (Sect. 5.3), accuracy of NNLO PDFs (Sect. 5.4) and finally an interesting and surprizingly precise estimate of the α_S strong coupling constant is given in Sect. 5.5.

5.1 Parton Luminosities

Once more the factorization properties can be exploited: considering hadronic observables, it is known that in this case a PDF for each colliding hadron is involved in observables computation. Following Ref. [196], then parton luminosity can be defined as

$$\Phi_{ij}\left(M_X^2\right) = \frac{1}{s}\int_\tau^1 \frac{dx_1}{x_1} f_i\left(x_1, M_X^2\right) f_j\left(\tau/x_1, M_X^2\right) \ , \tag{5.1}$$

where $f_i(x, M^2)$ is a PDF and $\tau \equiv M_X^2/s$. The behaviour of parton luminosity is strictly related to factorized observables behaviour, and the dependence of hadronic observables on PDFs can be mostly read by looking at parton luminosity. As a matter of fact, parton luminosities correspond to individual parton subprocesses. In particular, the gluon-gluon luminosity, the total quark-gluon and quark-antiquark luminosities are considered, defined as

$$\Phi_{qg} \equiv \sum_{i=1}^{n_f} \Phi_{q_i g} \ , \quad \Phi_{q\bar{q}} \equiv \sum_{i=1}^{n_f} \Phi_{q_i \bar{q}_i} \ . \tag{5.2}$$

and the charm and beauty quark-antiquark luminosities.

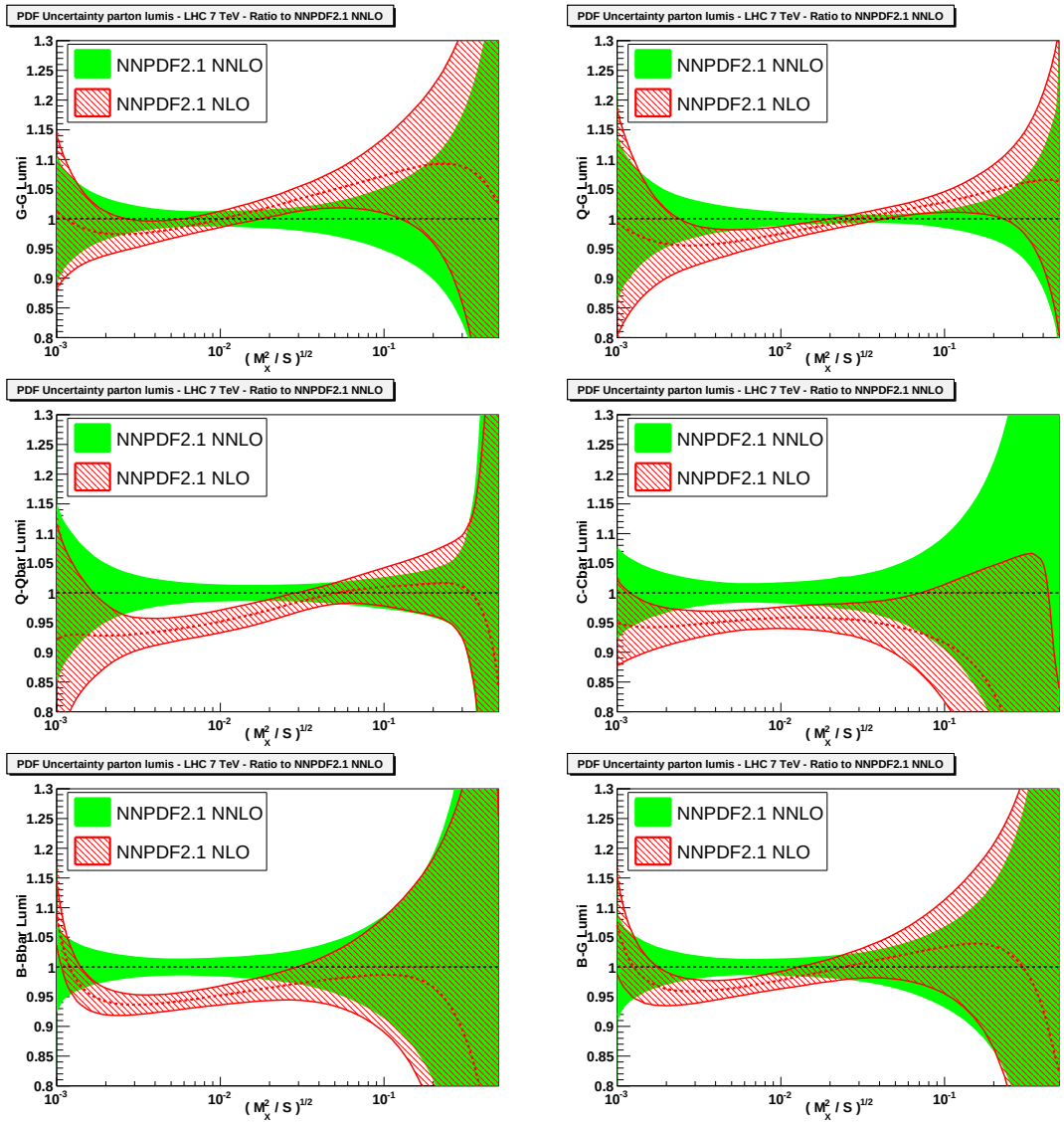

Figure 5.1: Comparison of the parton luminosities Eqs. (5.1)-(5.2) for LHC at 7 TeV, computed using the NNPDF2.1 NLO and NNLO PDFs, using $N_{\rm rep} = 100$ replicas from both sets. From left to right are shown Φ_{gg}, Φ_{qg}, (top) $\Phi_{q\bar{q}}$, $\Phi_{c\bar{c}}$, (middle) $\Phi_{b\bar{b}}$, Φ_{bg} (bottom). All luminosities are plotted as ratios to the NNPDF2.1 NNLO central value. All uncertainties shown are one sigma.

Another test for perturbative stability is to look at the comparison between parton luminosities computed from NNPDF2.1 NLO and NNLO parton sets. This comparison is given in Fig. 5.1 for all luminosities normalized to the NNPDF2.1 NNLO central value. The compatibility, and thus the perturbative stability, is good for all luminosities, as expected from PDFs comparison of Figs. 4.29-4.31. It is interesting to look at the gluon-gluon luminosity, that is the relevant one for Higgs production at the LHC, and observe that the result is quite stable especially in the region where recently a Higgs candidate has been observed at ATLAS [61] and CMS [62] ($M_X = 125 - 126$ GeV, $\sqrt{s} = 7 - 8$

TeV). The value for the NNLO luminosity however gets quite smaller for higher values of the invariant mass. Focusing now on quark-antiquark luminosity and going back to Fig. 4.32, the differences noticed there in light quark distributions are here reflected and enhanced because of the square dependence of parton luminosities on PDFs. As a consequence, non-negligible differences are seen for this luminosity, which is significantly larger at NNLO in the region relevant for W and Z production. Similar but somewhat smaller differences are seen in the quark-gluon channel. The heavy quark PDFs follow the behaviour of the gluon, from which they are generated dynamically via perturbative evolution.

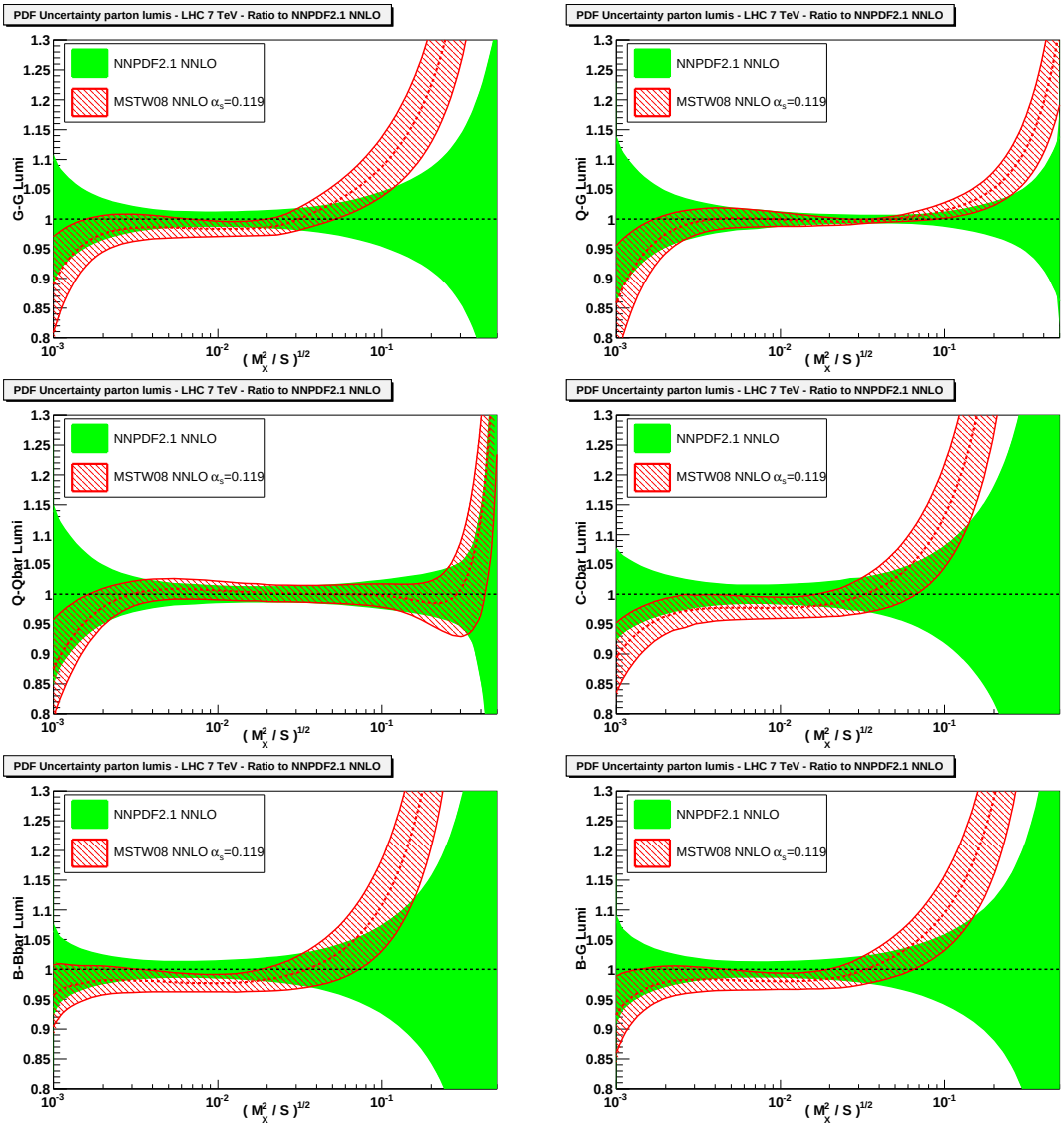

Figure 5.2: Same as Fig. 5.1, but for NNPDF2.1 NNLO and MSTW2008 NNLO PDFs. For both sets PDFs corresponding to the same value $\alpha_s = 0.119$ have been used for consistency.

Indeed a useful exercise is to compare NNPDF results with the ones from other

	$\sigma(W^+)B_{l\nu}$ [nb]	$\sigma(W^-)B_{l\nu}$ [nb]	$\sigma(Z^0)B_{ll}$ [nb]
NNPDF2.0	5.84 ± 0.14	3.97 ± 0.09	0.91 ± 0.02
NNPDF2.1	5.99 ± 0.14	4.09 ± 0.09	0.93 ± 0.02
CT10 - $\alpha_s = 0.118$	6.00 ± 0.13	4.10 ± 0.09	0.94 ± 0.02
CT10 - $\alpha_s = 0.119$	6.04 ± 0.13	4.13 ± 0.09	0.95 ± 0.02
MSTW08 - $\alpha_s = 0.119$	5.91 ± 0.11	4.16 ± 0.08	0.94 ± 0.02
MSTW08 - $\alpha_s = 0.120$	5.95 ± 0.11	4.19 ± 0.08	0.95 ± 0.02

	$\sigma(t\bar{t})$ [pb]	$\sigma(H, m_H = 120\,\mathrm{GeV})$ [pb]
NNPDF2.0	168 ± 7	11.59 ± 0.22
NNPDF2.1	170 ± 5	11.64 ± 0.17
CT10 - $\alpha_s = 0.118$	158 ± 7	10.99 ± 0.21
CT10 - $\alpha_s = 0.119$	161 ± 7	11.17 ± 0.21
MSTW08 - $\alpha_s = 0.119$	164 ± 5	11.48 ± 0.18
MSTW08 - $\alpha_s = 0.120$	168 ± 5	11.69 ± 0.18

Table 5.1: Cross-sections for W, Z, $t\bar{t}$ and Higgs production at the LHC at $\sqrt{s} = 7$ TeV and the associated PDF uncertainties. All quantities have been computed at NLO using MCFM for the NNPDF2.1, NNPDF2.0, CT10 and MSTW08 PDF sets. All uncertainties shown are one sigma.

groups: in Fig. 5.2 the NNPDF2.1 NNLO luminosities are compared to MSTW08 NNLO (at a common value of $\alpha_s = 0.119$), again as a ratio to NNPDF2.1 NNLO central value. A clear worsening of the agreement is observed for low and especially high values of M_X: this is mainly a consequence of the small-x discrepancy seen in Fig. 4.33 for singlet and especially for gluon distribution, related to the unstable behaviour of the MSTW08 NNLO gluon. In general, in the region of M_X^2/s which is relevant for typical electroweak final state masses M_X at the LHC, the agreement is good.

5.2 Predictions and Benchmarks at the LHC

In this Section predictions are computed and benchmarks are produced between both NLO and NNLO observables at LHC and the results obtained from other groups. In the first part of the Section predictions for a set of LHC standard candles are computed using the NNPDF2.1, NNPDF2.0, CT10 and MSTW08 sets. Results are presented and compared for $W^{\pm}$, Z^0, $t\bar{t}$ and Higgs production via gluon fusion with $m_H = 120$ GeV both at $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =14 TeV. These predictions, presented in Tables 5.1 and 5.2, and in the corresponding Figs. 5.3 and 5.5, are all computed using MCFM [201, 202] code at NLO QCD. In this comparison the value of α_s in the case of CT10 and MSTW08 sets is fixed both at the respective default value and at the common value of $\alpha_s(M_Z) = 0.119$, obtained using the PDF sets of Refs. [82, 203].

At first the comparison between NNPDF2.0 and NNPDF2.1 PDF sets is discussed: differences are driven mainly by the heavy quark mass effects but also by different kinematic cuts. For $W^{\pm}$ and Z production at 7 TeV differences are below or equal to one

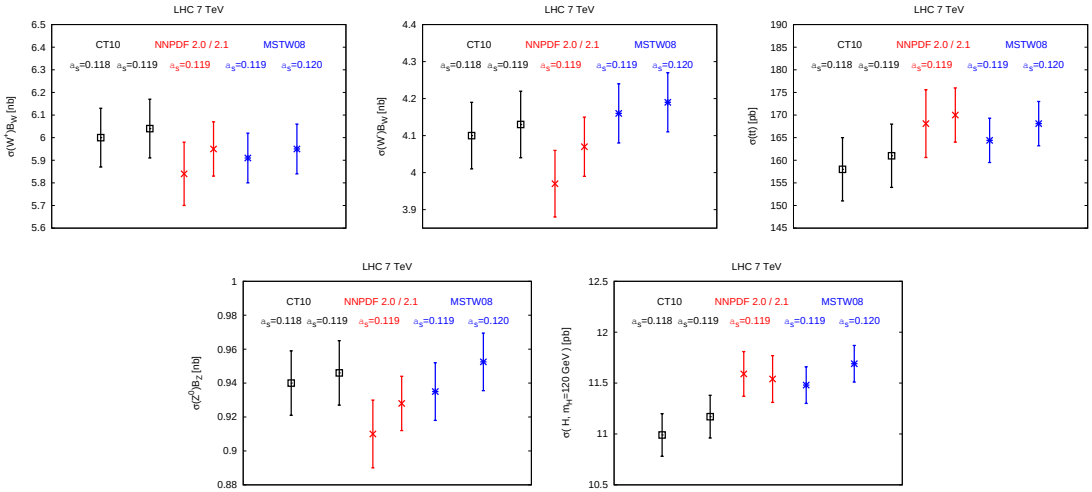

Figure 5.3: Graphical representation of the results of Table 5.1.

sigma and for $t\bar{t}$ and Higgs are essentially unchanged, as can be seen from Fig. 5.3 and Table 5.1. It can be concluded that these LHC observables are only marginally affected by the implementation of heavy quark treatment.

Coming now to the comparison among NNPDF2.1, MSTW08, and CT10 is visible a rather good agreement between NNPDF2.1 and MSTW08 observables and slightly larger differences with CT10, especially for Higgs and $t\bar{t}$ production, observables which are most sensible to the behaviour of gluon distribution already analyzed in Fig. 4.13. Things goes towards better agreement by using a common value of the strong coupling α_s. This effect is stronger for observables as Higgs production in gluon fusion [204], because of the characteristic dependence on α_s already at leading order of this observable. The first measurements released by the ATLAS and CMS experiments [205, 206] for $W^{\pm}$, Z, and $t\bar{t}$ production at 7 TeV are compared to the presented predictions: unfortunately, on that measurements luminosity uncertainty still dominated ($\mathcal{O}(11\%)$), so that their constraining power on PDFs is spoiled out.

If now the energy of the center of mass is increased from $\sqrt{s}$ =7 TeV to 14 TeV, the impact of heavy quark treatment should be larger. By looking at the results collected in Fig. 5.5 and Table 5.2, it can be seen that the $W^{\pm}$ and Z cross-sections, going from NNPDF2.0 to NNPDF2.1, are subject to an enhancement of slightly more than one sigma. The Higgs and $t\bar{t}$ production are instead almost unchanged. The comparison with CT10 and MSTW08 is similar as before, but with the agreement somewhat better for the Higgs and somewhat worse for top production.

A widely used technique to reduce experimental uncertainties is to consider the ratio between cross-sections, as for example the W^+/W^- and W/Z cross-section ratios at the

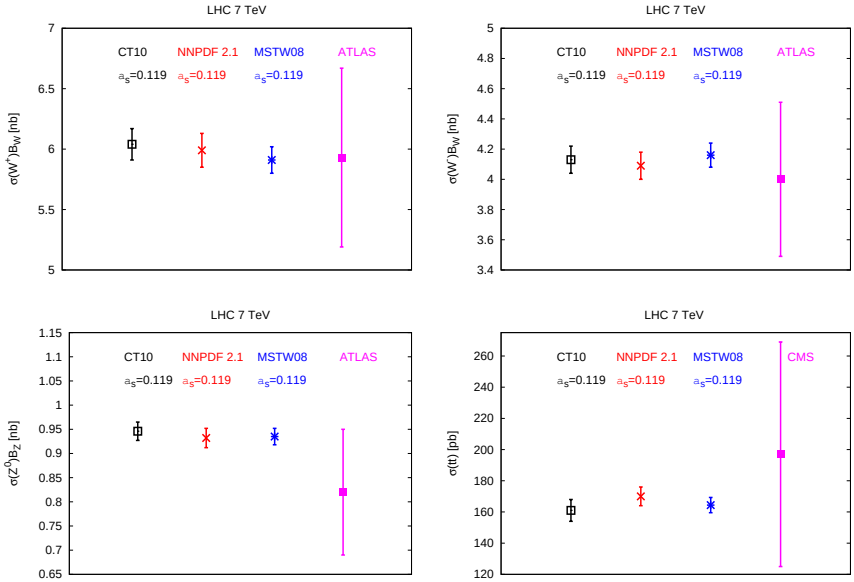

Figure 5.4: LHC measurements of the $W^{\pm}$, Z and $t\bar{t}$ cross-sections at $\sqrt{s}$ =7 TeV from the ATLAS [206] and CMS experiments [205] compared to the predictions of Fig. 5.3.

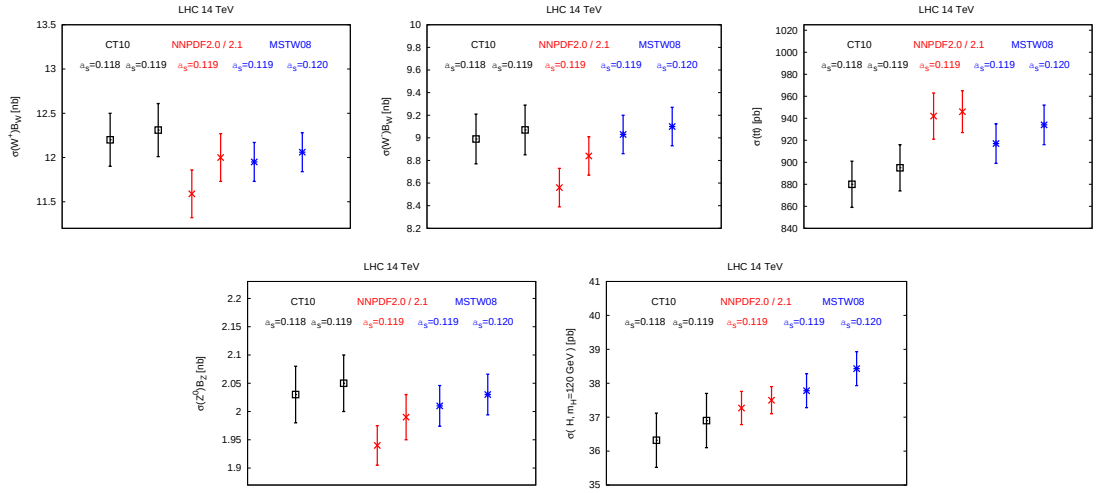

Figure 5.5: Graphical representation of the results of Table 5.2.

LHC. Here normalization uncertainties cancel out. Predictions for these ratios are given in Fig. 5.6 for NNPDF2.1, CT10, and MSTW08 both at 7 and 14 TeV. It can be observed that for W^+/W^- ratio CT10 and NNPDF2.1 are in good agreement but MSTW08 is lower by more than two sigmas, so that in general it seems in better agreement the result for total cross-sections. Also for the W/Z ratio at 7 TeV the agreement is only marginal, while at 14 TeV a better result is found. For these observables the dependence on α_s is negligible.

	$\sigma(W^+)B_{l\nu}$ [nb]	$\sigma(W^-)B_{l\nu}$ [nb]	$\sigma(Z^0)B_{ll}$ [nb]
NNPDF2.0	11.59 ± 0.27	8.56 ± 0.17	1.94 ± 0.04
NNPDF2.1	12.00 ± 0.27	8.84 ± 0.17	1.99 ± 0.04
CT10 - $\alpha_s = 0.118$	12.20 ± 0.30	9.00 ± 0.22	2.03 ± 0.05
CT10 - $\alpha_s = 0.119$	12.31 ± 0.30	9.07 ± 0.22	2.05 ± 0.05
MSTW08 - $\alpha_s = 0.119$	11.95 ± 0.22	9.03 ± 0.17	2.01 ± 0.04
MSTW08 - $\alpha_s = 0.120$	12.06 ± 0.22	9.10 ± 0.17	2.03 ± 0.04

	$\sigma(t\bar{t})$ [pb]	$\sigma(H, m_H = 120\,\mathrm{GeV})$ [pb]
NNPDF2.0	942 ± 21	37.3 ± 0.50
NNPDF2.1	946 ± 19	37.5 ± 0.40
CT10 - $\alpha_s = 0.118$	880 ± 21	36.32 ± 0.80
CT10 - $\alpha_s = 0.119$	895 ± 21	36.90 ± 0.80
MSTW08 - $\alpha_s = 0.119$	917 ± 18	37.78 ± 0.50
MSTW08 - $\alpha_s = 0.120$	934 ± 18	38.43 ± 0.50

Table 5.2: Same as Table 5.1 for the LHC at $\sqrt{s} = 14$ TeV.

Another interesting exercise to understand the relations between PDFs and observables is to compute their correlation. This is a way to quantify the relevance of each PDF for different physical observables [187, 207]. In Fig. 5.7 is shown the correlation coefficient between PDFs and W^+, Z total cross-sections, W^+/W^-, W/Z cross-section ratios at 7 TeV for the LHC. Knowing the dependence of total W and Z cross-sections, a dominant correlation with the u and d sea quarks and an anticorrelation of similar size in modulus with the strange quarks is observed as expected. Instead, the gluon and heavy quarks that are dynamically generated from it are less correlated with these observables. Looking now at the cross-section ratios, it is interesting to notice that, for the W/Z cross-section ratio, correlations are only slightly suppressed in comparison to the previous analyzed cases, while for the W^+/W^- ratio the suppression is much more intense. This observation suggests that the latter ratio should be less sensitive to PDF uncertainties.

In the treatment of heavy quark mass effects theoretical ambiguities due to subleading terms are present. Here is given a partial study of this theoretical uncertainty related to heavy quark inclusion to $\mathcal{O}(\alpha_s)$ by comparing results obtained from the three sets discussed in Sect. 4.1.3: NNPDF2.0 RED (without heavy quark mass terms, but the same kinematic cuts as NNPDF2.1), the default NNPDF2.1, and NNPDF2.1 without damping terms in the FONLL-A method. As can be seen in Fig. 5.8 and in the associated Table 5.3, the results for the NNPDF2.1 “plain" FONLL-A (i.e. without damping terms) are always half way between the ZM-VFN treatment (NNPDF2.0 RED) and the standard NNPDF2.1 global fit at 7 TeV. At the higher energy of 14 TeV the NNPDF2.1 “plain" FONLL-A results sit nearer to the standard NNPDF2.1 set results. As discussed in Sect. 4.1.3, the difference between the NNPDF2.1 results with and without damping terms can be taken as a conservative estimate of the theoretical uncertainty associated to the uncertainty in the inclusion of heavy quark mass effects to $\mathcal{O}(\alpha_s)$.

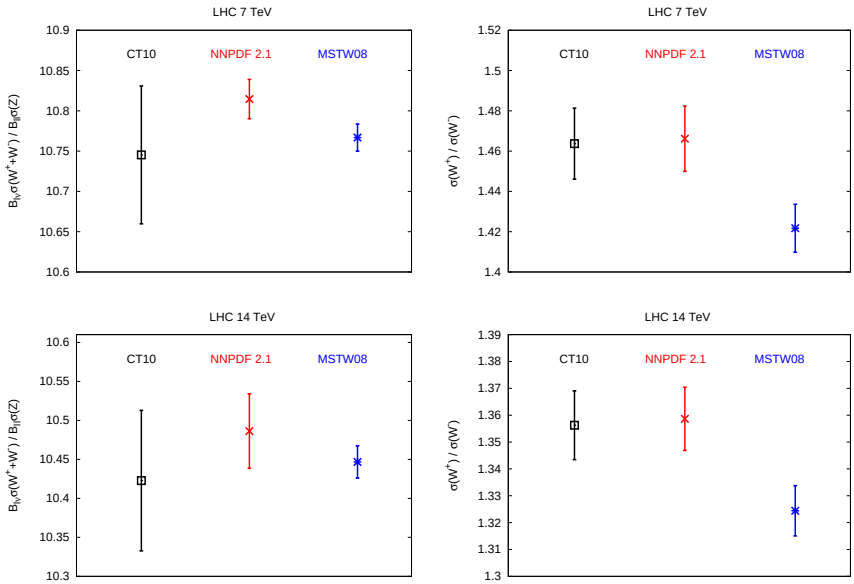

Figure 5.6: Comparison between predictions from different PDF sets for the W^+/W^- and W/Z ratios at the LHC $\sqrt{s}$ =7 TeV (upper plots) and $\sqrt{s}$ =14 TeV (lower plots).

In this second part of the Section a similar set of LHC standard candles are computed using NNPDF2.1 NNLO. The physical observables that here will be considered are the total cross-section for Higgs, weak vector bosons production, approximate NNLO $t\bar{t}$ production and the $(W^+ + W^-)/Z^0$ and W^+/W^- cross-section ratios.

As already seen in the case of NLO observables, their dependence on α_s at hadron colliders is nontrivial: the hard matrix elements depend on the strong coupling and PDFs are correlated with the value of α_s [82, 208, 209], especially in the case of the gluon PDF. A very important issue that needs to be considered when comparing predictions trying to understand their dependence on PDFs and their associated uncertainty is related to the mixing between uncertainties due to PDFs with uncertainties due to the choice of external parameters [210]. This mixing should be avoided. A detailed quantitative comparison is performed between NNPDF2.1 and MSTW08, both at NLO and NNLO QCD. For sake of illustration also the comparison with ABKM09 NNLO will be performed, even if these PDFs are provided for $\alpha_s(M_Z) = 0.1135 \pm 0.0014$, and their uncertainties always include also the contribution due to the variation of α_s in this range.

The first comparison that needs to be discussed is related to Higgs production from gluon-gluon fusion. The total inclusive cross-section for this gluon-gluon channel has attracted considerable attention: in Ref. [211] was claimed that the determination of uncertainties due to PDFs through the so-called PDF4LHC prescription [106] might suffer of a substantial underestimation. More details on this discussion are given in Ref. [181] and especially in Ref. [184]. In Fig. 5.9 the results for the total inclusive cross-section for

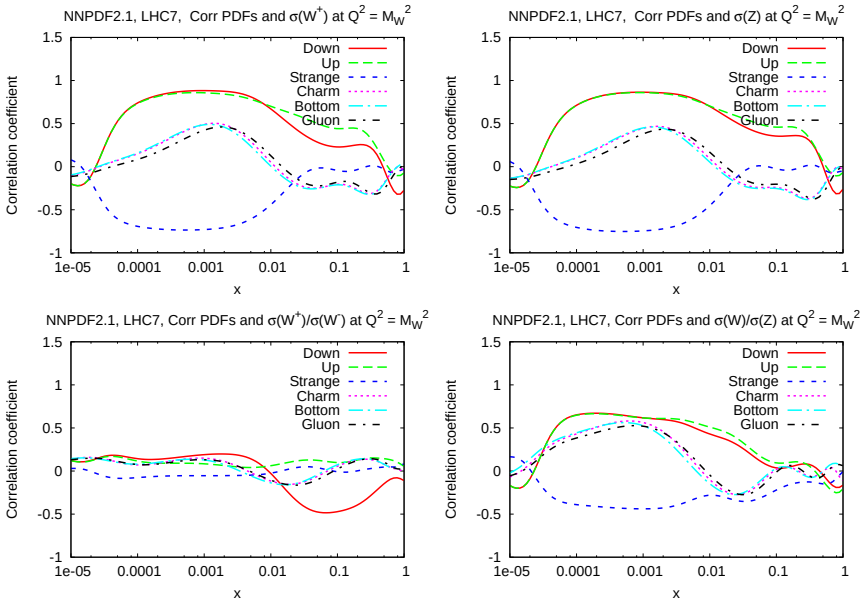

Figure 5.7: The correlation between PDFs and vector boson production total cross-sections (upper plots) and their ratios (lower plots) for LHC 7 TeV. Correlations for W^- (not shown) are very similar.

Higgs production from gluon fusion are compared as a ratio to NNPDF2.1 NNLO with $\alpha_s(M_Z) = 0.119$. The code of Refs. [212, 213] is used to compute this physical observable for NNPDF2.1, MSTW08, and ABKM09 PDF sets at the respective default value of α_s (in the order: $\alpha_s(M_Z) = 0.119$, $\alpha_s(M_Z) = 0.1171$, and $\alpha_s(M_Z) = 0.1135 \pm 0.0014$) as a function of the Higgs mass m_H. All uncertainties are one-sigma error bands. As already said, ABKM determination includes also the uncertainty on α_s. The shown observable is also computed using NNPDF2.1 with $\alpha_s(M_Z) = 0.117$ to have a more precise comparison with MSTW08 prediction. The NNPDF result for this last computation and MSTW default result are in excellent agreement. The same cannot be said for ABKM: surely part of the several sigmas difference between NNPDF2.1 and ABKM is due to the different value of α_S used for the computation, but looking at Fig. 5.16 below it is clear that even adopting the same value for the strong coupling constant the difference would persist.

The results for top production are shown in Fig. 5.10 where the total $t\bar{t}$ cross-section determined at NLO and NNLO at the LHC 7 TeV with m_t =172 GeV (pole mass) is compared with NNPDF2.1, MSTW08, and ABKM. The ABKM prediction is given only at NNLO with $\alpha_s(M_Z) = 0.1135 \pm 0.0014$ as in the previous case, while for NNPDF2.1 NLO and NNLO $\alpha_s(M_Z) = 0.119$ and for MSTW08 $\alpha_s(M_Z) = 0.1202$, 0.1171 respectively for NLO and NNLO predictions. To ease the comparison between NNPDF2.1 and each of the other predictions from MSTW08 and ABKM, the NNPDF results with

7 TeV	$\sigma(W^+)B_{l\nu}$ [nb]	$\sigma(W^-)B_{l\nu}$ [nb]	$\sigma(Z^0)B_{ll}$ [nb]
NNPDF2.1	5.99 ± 0.14	4.09 ± 0.09	0.93 ± 0.02
NNPDF2.0 RED	5.81 ± 0.13	3.98 ± 0.08	0.91 ± 0.02
NNPDF2.1 FONLL-A plain	5.90 ± 0.12	4.03 ± 0.08	0.92 ± 0.02

14 TeV	$\sigma(W^+)B_{l\nu}$ [nb]	$\sigma(W^-)B_{l\nu}$ [nb]	$\sigma(Z^0)B_{ll}$ [nb]
NNPDF2.1	12.00 ± 0.27	8.84 ± 0.17	1.99 ± 0.04
NNPDF2.0 RED	11.57 ± 0.25	8.57 ± 0.17	1.93 ± 0.04
NNPDF2.1 FONLL-A plain	11.82 ± 0.22	8.72 ± 0.15	1.96 ± 0.03

Table 5.3: Cross-sections for W, Z, $t\bar{t}$ and Higgs production at the LHC at $\sqrt{s} = 7$ TeV and the associated PDF uncertainties for the reference NNPDF2.1 set compared to those obtained using sets with different treatment of heavy quarks: NNPDF2.0RED, without heavy quark mass effects, and NNPDF2.1 FONLL-A plain with heavy quark mass effects but without threshold damping terms.

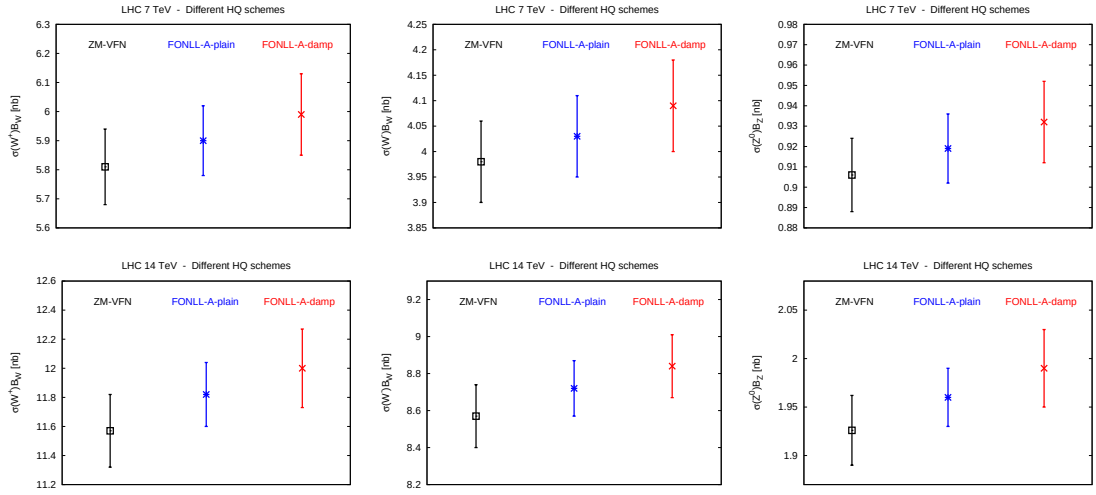

Figure 5.8: Graphical representation of the results of Table 5.3.

$\alpha_s(M_Z) = 0.120$ (NLO) and $\alpha_s(M_Z) = 0.114,\ 0.117$ (NNLO) are also computed. To determine all predictions for this observable, approximate NNLO expressions based on threshold resummation have been constructed [214] and implemented in the public HATHOR code [215]. These theoretical predictions can be compared to the average of the first measurements from CMS [216, 217], $\sigma\left(t\bar{t}\right) = 158 \pm 19$ pb, and ATLAS [218], $\sigma\left(t\bar{t}\right) = 180 \pm 19$ pb. Averaging the most accurate results, which have been obtained with a luminosity of ~ 36 pb^{-1}, and assuming that the two measurements are independent, yields $\sigma\left(t\bar{t}\right) = 169 \pm 13$ pb (shown in Fig. 5.10 as a dashed band).

Figure 5.10 shows that the NNPDF2.1 and MSTW08 predictions are in good agreement both at NLO and NNLO: however, once again, it is important to use a common value of α_s. Also, of course, one should remember that the uncertainties shown in Fig. 5.10 are only PDF uncertainties. In particular theoretical uncertainties, such as may be estimated

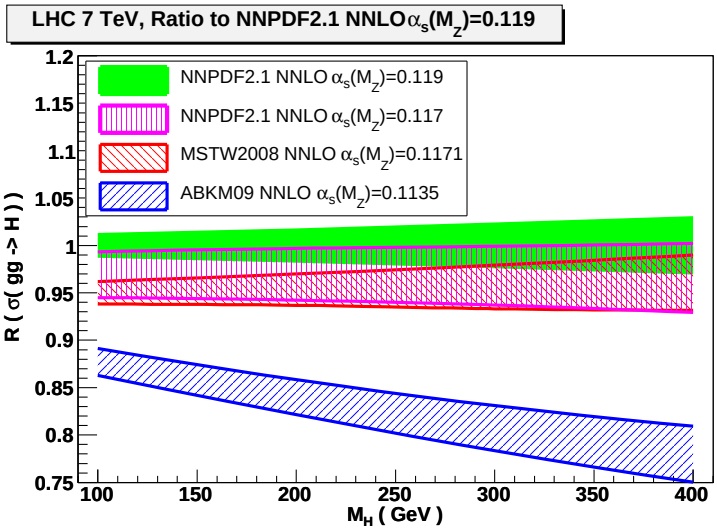

Figure 5.9: The total cross-section for Higgs boson production via gluon fusion at NNLO as a function of m_H. Results are shown for NNPDF2.1 with $\alpha_s(M_Z) = 0.119$ and $\alpha_s(M_Z) = 0.117$, MSTW08 with $\alpha_s(M_Z) = 0.1171$, and ABKM09 with $\alpha_s(M_Z) = 0.1135 \pm 0.0014$, all displayed as ratios to the central NNPDF2.1 curve. The NNPDF result is obtained using $N_{\rm rep} = 100$ replicas. All uncertainties shown are one sigma; for NNPDF and MSTW they are pure PDF uncertainties, while for ABKM they also include the α_S uncertainty corresponding to their given range.

by scale variation, and uncertainties due to the dependence on the top mass, are not shown and might also be significant. Again, the ABKM09 prediction is significantly lower, and the disagreement persists even when a common value of α_s is adopted: already the LHC data are starting to discriminate between PDF sets.

A closely related recent measurement by CMS [216] is the ratio of $t\bar{t}$ and Z cross-sections. Predictions has been computed for this observable using the VRAP code [219] together with HATHOR, for the same PDF sets and settings. At NNLO this ratio is only weakly dependent on the value of α_s. Results are also shown in Fig. 5.10 and compared to the CMS measurement, again shown as a dashed band. The conclusions are similar.

As a last case, the electroweak vector boson production total cross-sections and their ratios at the LHC are considered. In Fig. 5.11, as for the case of top production, is shown NNPDF2.1 with the same values for α_s used there (NNPDF preferred value and the MSTW and ABKM default one). These observables has been computed with the VRAP code [219], within the narrow-width approximation (including the γ^* contribution to gauge boson production). For cross-section ratios error bands are determined as a 68% confidence level instead of a standard deviation, because has been verified that the distribution of results can be markedly non-Gaussian. Again, these theoretical predictions are compared to the first ATLAS [220] and CMS [221] results for this observable, corresponding to an integrated luminosity of 36 pb^{-1}, averaged together. The results are shown as dashed bands on the plots. The single results and their averages are summarized

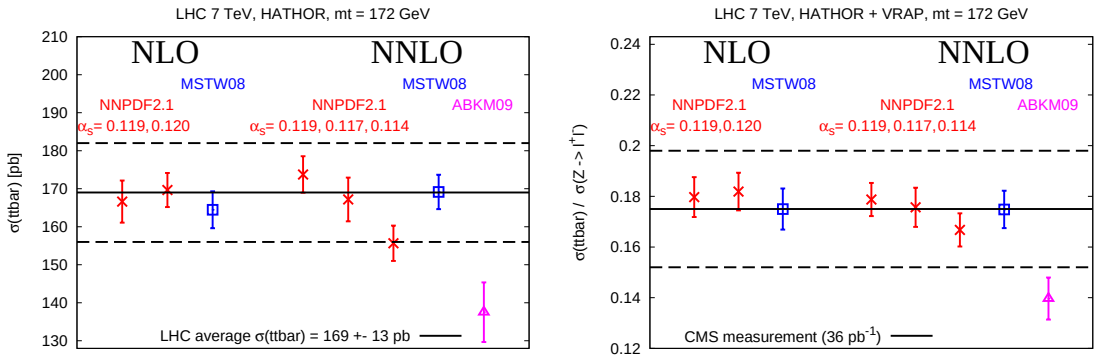

Figure 5.10: The total cross-section for $t\bar{t}$ production at the LHC 7 TeV computed using HATHOR (left) and its ratio to the total Z production cross-section computed using VRAP (right). Results are shown for NNPDF2.1 with $\alpha_s(M_Z) = 0.119$ (NLO and NNLO), $\alpha_s(M_Z) = 0.120$ (NLO) and $\alpha_s(M_Z) = 0.114, 0.117$ (NNLO), MSTW08 with $\alpha_s(M_Z) = 0.1202$ (NLO) and $\alpha_s(M_Z) = 0.1171$ (NNLO), and ABKM09 with $\alpha_s(M_Z) = 0.1135 \pm 0.0014$ (NNLO). The NNPDF results are obtained using $N_{\rm rep} = 100$ replicas. All uncertainties shown are one sigma; for NNPDF and MSTW they are pure PDF uncertainties, while for ABKM they also include the α_S uncertainty corresponding to the given range. The band corresponds to the combination of the most recent CMS and ATLAS measurements (left, see text) and to the CMS measurement [216] (right).

in Table 5.4.

As these processes are quark-dominated, their dependence on α_s is weaker than the ones previously considered. This is also due to the fact that the Born level is independent of the strong coupling and because they are affected by smaller NNLO corrections (though not negligible, as can be seen from the plots in Fig. 5.11). Differences between PDF sets are also less significant, except for the W^+/W^- cross-section ratio which is a very sensitive probe of the quark flavor decomposition. It can be concluded that the LHC data used here, and particularly the W cross-section ratio, can already provide some discrimination between PDFs. Also, from the comparison between LHC data and theoretical predictions for some of the standard candles here analyzed a discrimination between NLO

	ATLAS	CMS	Average
$\sigma\left(W^+\right) B\left(l^+\nu\right)$ (nb)	6.26 ± 0.32	6.02 ± 0.26	6.11 ± 0.20
$\sigma\left(W^-\right) B\left(l^-\nu\right)$ (nb)	4.15 ± 0.21	4.26 ± 0.19	4.21 ± 0.14
$\sigma\left(Z^0\right) B\left(l^+l^-\right)$ (nb)	0.945 ± 0.051	0.975 ± 0.044	0.962 ± 0.033
$\sigma\left(W^+ + W^-\right) B\left(l\nu\right) / \sigma\left(Z^0\right) B\left(l^+l^-\right)$	10.91 ± 0.28	10.54 ± 0.19	10.65 ± 0.16
$\sigma\left(W^+\right) / \sigma\left(W^-\right)$	-	1.421 ± 0.034	1.421 ± 0.034

Table 5.4: Recent results from CMS at ATLAS for the total cross-sections for W^+, W^- and Z^0 production and their ratios, obtained with an integrated luminosity of ~ 36 pb^{-1}, together with their average. The average has been obtained assuming the two measurements to be completely uncorrelated.

and NNLO sets may be drawn.

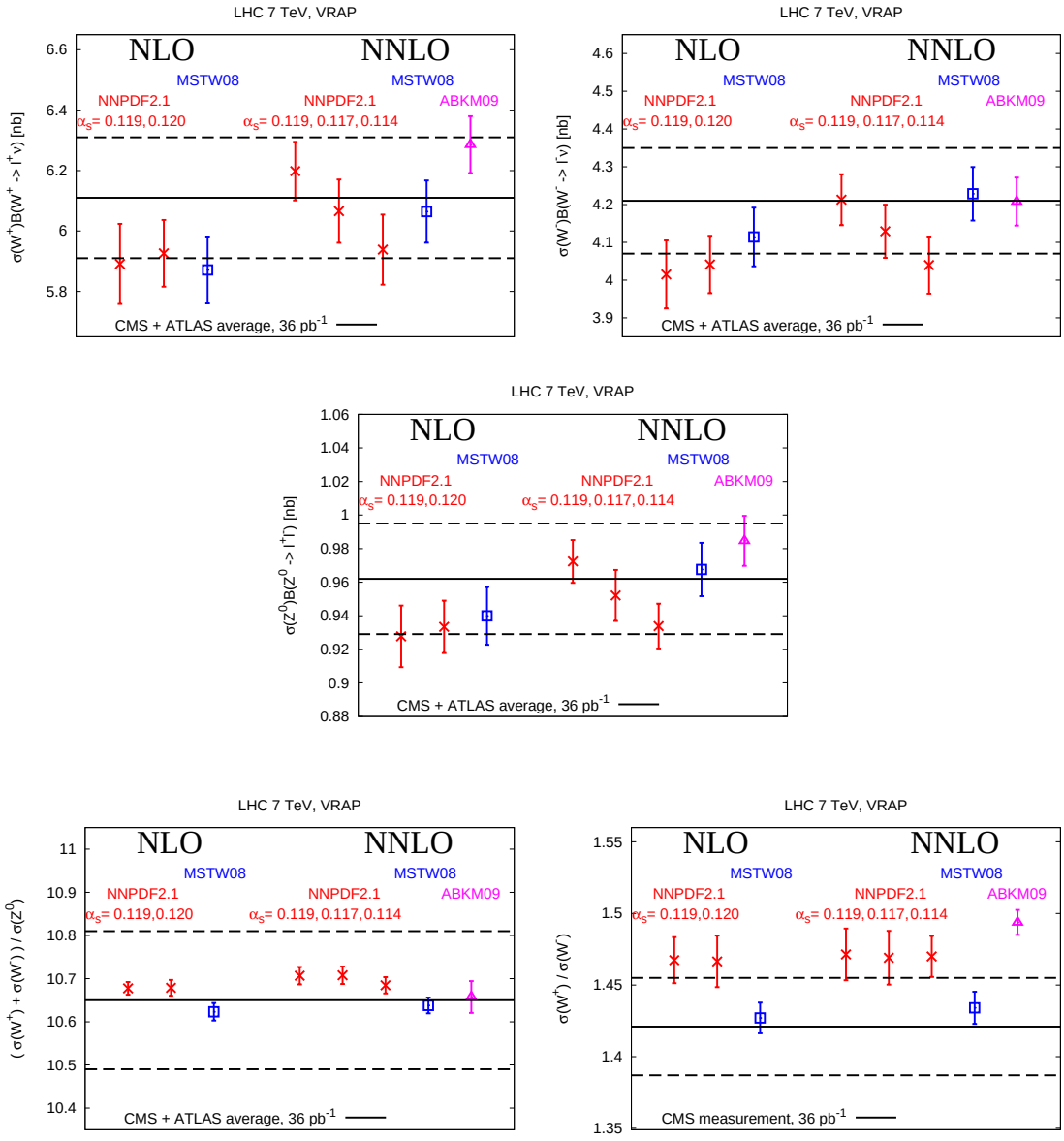

Figure 5.11: The total cross-sections for W^+, W^- and Z^0 production at the LHC 7 TeV and their ratios $\sigma(W^+ + W^-)/\sigma(Z^0)$ and $\sigma(W^+)/\sigma(W^-)$. Predictions are shown for the same PDF sets and values of α_s shown in Fig. 5.10, and compared to the ATLAS and CMS data summarized in Table 5.4, shown as a dashed band. Uncertainties shown are one sigma for absolute cross-sections, and 68% confidence levels for cross-section ratios.

5.3 The NuTeV Anomaly

In Ref. [222] the estimated value for $\sin^2\theta_W$ extracted from the NuTeV data was found to be three standard deviations above the value given by the Standard Model prediction. In previous NNPDF releases [49, 87] the implications that the determination of the strangeness asymmetry $s^-(x,Q^2)$ has on the so-called NuTeV anomaly [223] were studied. By looking at the updated plot in Fig. 5.12, it is clear that the values

$$\sin^2\theta_W\big|_{\rm NuTeV} = 0.2277 \pm 0.0017 \ , \ \ \sin^2\theta_W\big|_{\rm EWfit} = 0.2223 \pm 0.0003 \tag{5.3}$$

are effectively significantly different, but looking at the determinations of the Weinberg angle [48] corrected for the strangeness asymmetry using the values from previous [49,87] and current NNPDF2.1 sets it can be observed how the one-sigma error bar gets bigger using NNPDF1.2, where a more reliable uncertainty determination for strange flavor is given, and in particular for NNPDF2.1

$$R_S(Q^2) \equiv 2\frac{\int_0^1 dx x s^-(x,Q^2)}{\int_0^1 dx x \left(u^-(x,Q^2)+d^-(x,Q^2)\right)} = 2\frac{[S^-]}{[U^-+D^-]} = (1.37 \pm 0.77)\ 10^{-2}. \tag{5.4}$$

This result, that includes heavy quark mass effects, is quite near to the NNPDF2.0 result that does not include them, thereby showing that the impact of heavy quark mass effects on the determination of the strangeness asymmetry is very small, and can also be seen that the three corrected values are in excellent agreement with the electroweak fit and with each other.

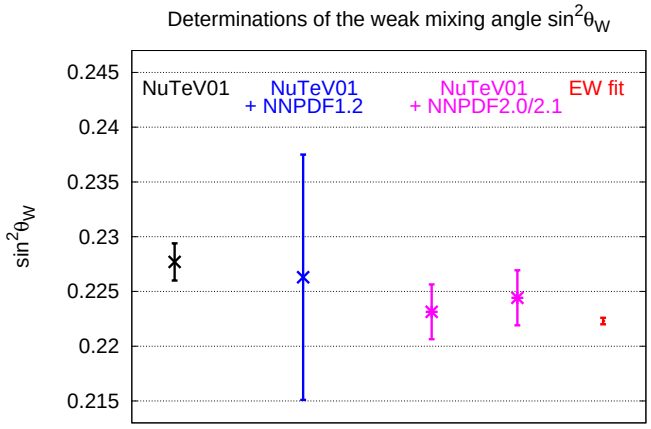

Figure 5.12: Determination of the Weinberg angle from the uncorrected NuTeV data [48], with $[S^-]$ correction determined from NNPDF1.2, NNPDF2.0 and NNPDF2.1. The uncertainty shown on NNPDF2.0 and NNPDF2.1 is the one-sigma PDF uncertainty only.

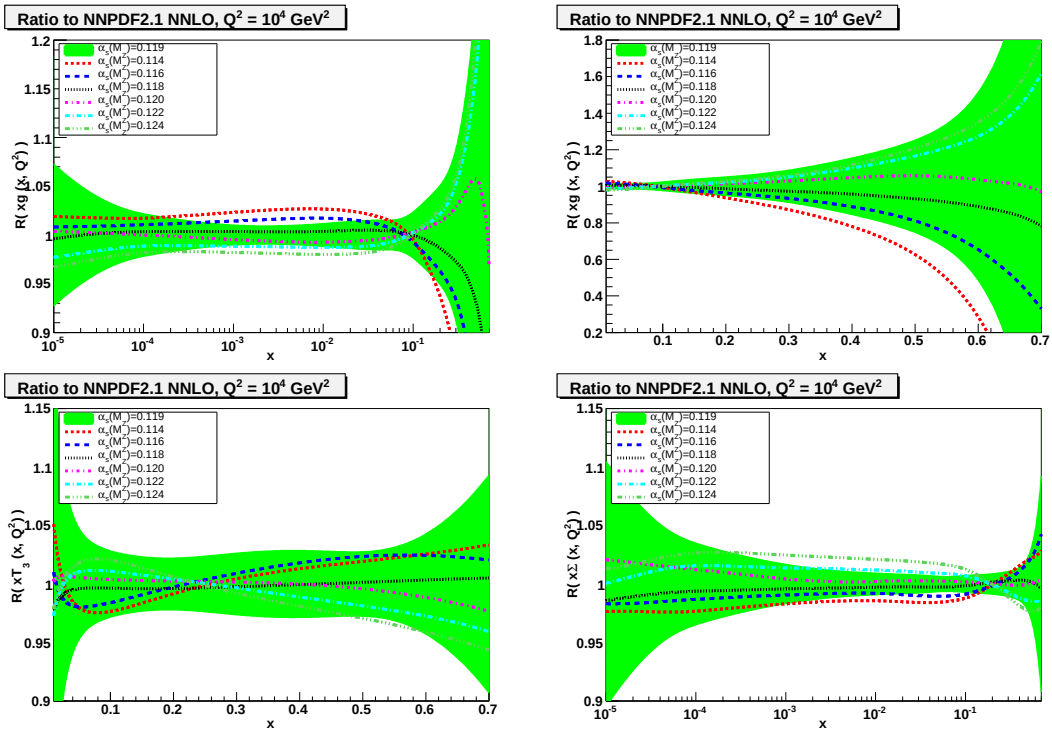

Figure 5.13: Comparison of NNPDF2.1 NNLO sets with different values of the strong coupling, shown as ratios to reference set with $\alpha_s = 0.119$ at $Q^2 = 10^4$ GeV^2: gluon at small and large-x (top), triplet at large-x and singlet at small-x (bottom). To improve readability PDF uncertainties are only shown for the $\alpha_s = 0.119$ set.

5.4 Accuracy of the NNLO PDF Determination

The aim of this Section is to discuss the main sources of uncertainty that affects NNPDF results and their influence on them. The main contribution to PDF uncertainties comes from the underlying dataset (more details are given in Ref. [87, 224]). Of course, another factor that plays an important role is the choice of QCD parameters used in the partonic fit, as the value of the strong coupling α_s, but also of the quark masses. The variation of these parameters allows for an estimation of the related uncertainties on PDFs and can be shown that in some cases their impact is comparable to that of data [38,40,204,209]. In the following are discussed the results obtained generating various fits with different values of α_S both at NLO and NNLO and also by repeating NNPDF2.1 NNLO determination with various subsets of the global dataset. Theoretical uncertainties will not be studied here: as argued in Sect. 4.5, at NNLO the uncertainty related to higher order corrections (as might be estimated by renormalization and factorization scale variation) is usually subdominant, as are those related to the treatment of heavy quarks [38].

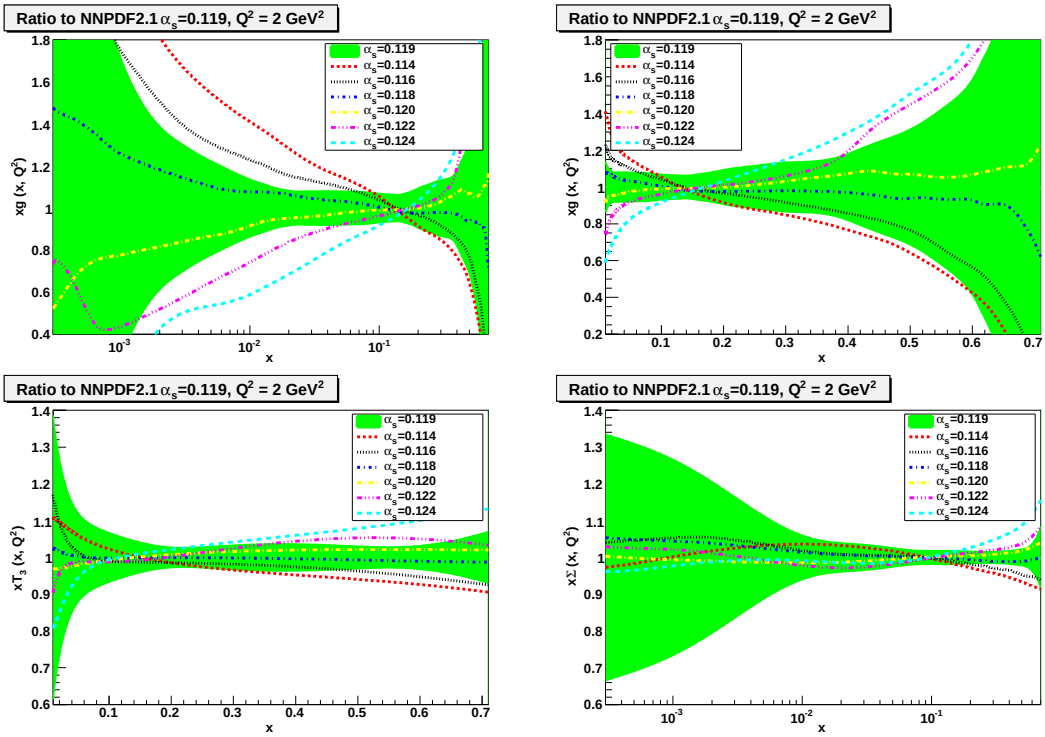

Figure 5.14: Comparison between NNPDF2.1 NLO sets with different values of the strong coupling, shown as ratios with respect the reference fit with $\alpha_s = 0.119$. The PDFs shown are the gluon at small and large-x (upper plots), the triplet at large-x and the singlet at small-x (lower plots).

Dependence on $\alpha_s\,(M_Z)$

First, the NNPDF2.1 NLO PDFs set is analyzed. Several sets with $\alpha_s\,(M_Z)$ in the range from 0.114 to 0.124 in steps of 0.001 are generated and the correlations between the $\alpha_s\,(M_Z)$ values and each PDF are computed as defined in Eq. (82) of Ref. [209]. The results are shown in Fig. 5.15 as a function of x, both at $Q^2 = 2$ GeV^2 and $Q^2 = 10^4$ GeV^2. As expected, the most sensitive PDF is the gluon, and indeed correlations gets weaker increasing the scale because of asymptotic freedom. In this determination the uncertainty on α_s at the 68% confidence level is considered as $\Delta\alpha_s = 0.0012$.

To determine the combined uncertainty on observables, due both to PDF and α_s uncertainties, can be followed the procedure described in Ref. [204], that combines sets with different values of α_s. The method there described suggests to pack together a new set of replicas choosing in a proper way a certain number of replicas from each of the delivered sets with varied α_s values. The suggested proper way is to construct a Gaussian distribution with the assumed central value and uncertainty for α_s and then select according to this distribution $N_{\rm rep}$ replicas from the original generated sets. In this way are produced prepacked sets containing a number of replicas for each value of α_s which corresponds to a Gaussian distribution with given mean and standard deviation. Of course the accuracy

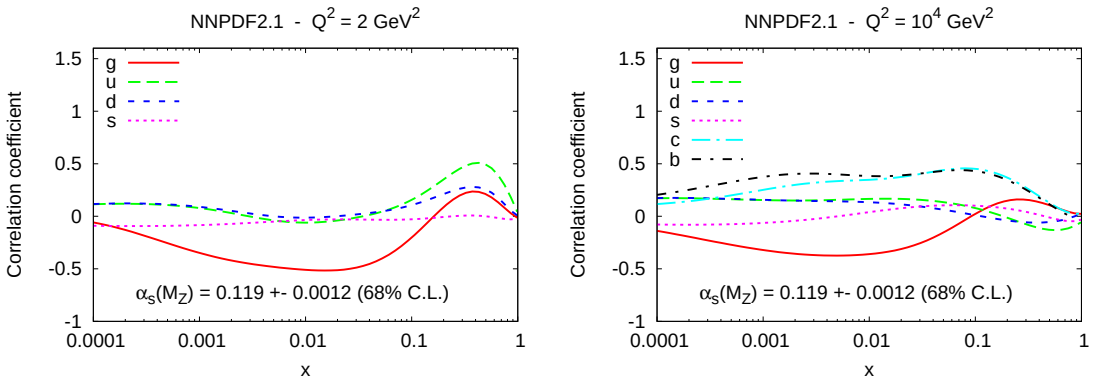

Figure 5.15: Correlation coefficient between PDFs and $\alpha_s(M_Z)$ computed assuming $\alpha_s(M_Z) = 0.119 \pm 0.0012$ at 68% C.L. Results are shown at low scale ($Q^2 = 2$ GeV^2, left) and high scale ($Q^2 = 100$ GeV^2, right).

of the prediction using a combined PDF+α_s set depends on the number of replicas $N_{\rm rep}$.

Following the PDF4LHC recommendation [106] for the combination of PDF+α_s uncertainties, two prepacked PDF+α_s uncertainty sets with $\alpha_s\,(M_z) = 0.119$ and uncertainties $\delta_{\alpha_s} = 0.0012$ and $\Delta\alpha_s = 0.002$ as one-sigma errors are delivered. It is enough to follow the procedure explained above to produce sets with any other values. Below $N_{\rm rep} \sim 100$ is observed a somewhat less accurate result for Higgs production in gluon fusion at LHC 7 TeV, while above this number of replicas in the prepacked set the result looks independent of $N_{\rm rep}$. In conclusion, $N_{\rm rep}$ =100 is recommended.

The same range of α_s values is covered also with the NNPDF2.1 NNLO treatment. Results for the gluon, the quark singlet and isospin triplet are displayed in Fig. 5.13, where the ratio of the central PDFs for each value of α_s to the default $\alpha_s\,(M_Z) = 0.119$ set is shown, and compared to the PDF uncertainty on the central set. Again, the gluon PDF is the most sensitive to α_s variations. In particular, looking at Fig. 5.16 where the Higgs cross-section from gluon-gluon fusion is computed using the NNLO sets with different values of α_s and normalized to the central value, can be seen that its dependence on the value of α_S is remarkable: the factors discussed above that make the gluon distribution dependent on the α_S value affect the result through a quadratic gluon PDF dependence. Moreover, the Higgs cross-section from gluon-gluon fusion undergoes NLO corrections which are as large as the LO contribution, and NNLO corrections which are about half of the LO (see Ref. [184, 204] for detailed studies of the relative size of PDF and α_s uncertainties on this process).

The values of heavy quark masses are varied in the same way as already done for the NNPDF2.1 NLO fit: $m_c = 1.5,\ 1.6$ and 1.7 GeV (in addition to the default $m_c = \sqrt{2}$ GeV), and with $m_b = 4.25,\ 4.5,\ 5.0$ and 5.25 GeV (in addition to the default $m_b =$

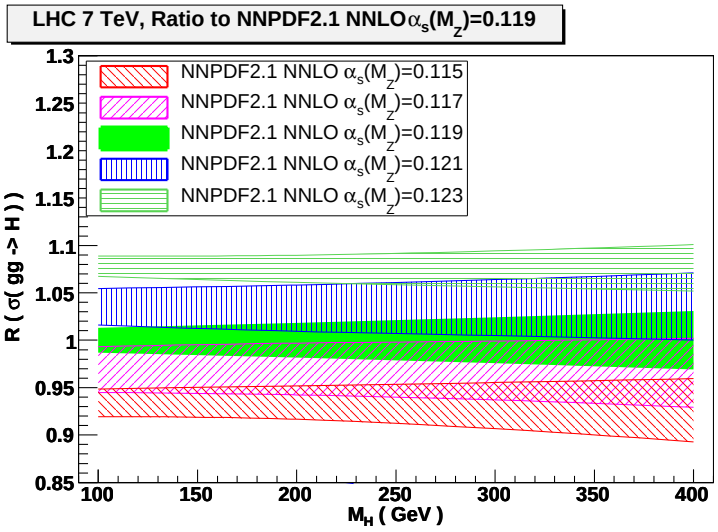

Figure 5.16: Same as Fig. 5.9, but for NNPDF2.1 NNLO sets with different values of $\alpha_s(M_Z)$. Results are shown as ratios to the NNPDF2.1 NNLO reference with $\alpha_s(M_Z) = 0.119$.

4.75 GeV). The dependence of PDFs on these variations is similar to that observed for the NLO case.

Dependence on the Dataset

The focus is now set on the dependence of PDFs and their uncertainties on the dataset they are extracted from. It is important to study into detail this dependence because if on one side it is true that a wider dataset always carries more information, on the other side it is also true that smaller datasets may be more consistent and more reliable.

To better understand which are the consequences of reducing the dataset to gain more consistency of the fitted data, four different parton sets are produced on a smaller dataset, that is a subset of the full NNPDF2.1 NNLO global dataset:

- HERA data only. This set is fitted on the smaller dataset of the four subsets considered, and at the same time is the most consistent. It includes: the combined HERA-I inclusive data, the H1 and ZEUS F_2^c data and the ZEUS HERA-II data. Parton distribution functions based on this dataset have also been determined and published by the HERAPDF group [114].

- Deep-inelastic scattering (DIS) only. From the global dataset are excluded all hadron-hadron data, i.e. DY and jets, which one may perhaps consider theoretically or experimentally less clean than lepton-hadron data.

- Deep-inelastic scattering and Drell-Yan (DIS+DY) only. This determination is in

Experiment	Global	HERA–only	DIS–only	DIS+DY	Collider–only
$N_{\rm dat}$	3357	834	2783	3171	1090
Total	1.16	1.07	1.15	1.18	1.02
NMC-pd	0.93	[13.15]	0.88	0.94	[3.43]
NMC	1.63	[1.91]	1.69	1.69	[2.06]
SLAC	1.01	[3.17]	0.97	1.03	[1.23]
BCDMS	1.32	[2.15]	1.28	1.30	[2.22]
HERAI-AV	1.10	1.05	1.09	1.09	1.06
CHORUS	1.12	[2.63]	1.08	1.13	[1.74]
FLH108	1.26	1.32	1.27	1.26	1.26
NTVDMN	0.49	[60.51]	0.45	0.54	[23.02]
ZEUS-H2	1.31	1.21	1.26	1.28	1.30
ZEUSF2C	0.88	0.77	0.86	0.88	0.75
H1F2C	1.46	1.30	1.47	1.50	1.24
DYE605	0.81	[9.06]	[6.86]	0.82	[1.34]
DYE866	1.32	[12.41]	[2.70]	1.32	[5.76]
CDFWASY	1.65	[7.71]	[13.94]	1.64	1.07
CDFZRAP	2.12	[3.74]	[2.15]	1.91	1.22
D0ZRAP	0.67	[1.11]	[0.67]	0.65	0.61
CDFR2KT	0.74	[1.15]	[0.99]	[1.25]	0.64
D0R2CON	0.82	[1.28]	[0.88]	[1.03]	0.83

Table 5.5: Quality of the fit for NNLO PDF sets based on datasets of varying size. The total number of data points is given in the first row, followed by the χ^2 normalized to the number of data points both for the total fitted set and for each of the individual experiments. The χ^2 values for experiments which are not fitted are also shown in square brackets.

principle the only truly NNLO one, as it excludes jet data, for which only approximate NNLO matrix elements are known.

- Collider data only (lepton and hadron collider). In this case all the fixed-target data are excluded. They are considered less clean because of the low energy range at which they are determined and also because of the nuclear targets used for example in the case of neutrino DIS data. This determination is of greater complexity than DIS+DY, despite having a smaller number of datapoints, because it also includes jet data.

In each case, a set of $N_{\rm rep} = 100$ PDF replicas has been constructed. The NLO counterparts of the DIS and DIS+DY PDF determinations were discussed in Ref. [87] and are available from LHAPDF both for NNPDF2.0 and NNPDF2.1; the HERA-only NLO PDFs were briefly discussed in Ref. [225]; collider-only PDFs are presented here for the first time.

The possibility of obtaining reliable PDFs from datasets of widely varying size (more than a factor three, in this case) without having to modify any aspect of the methodol-

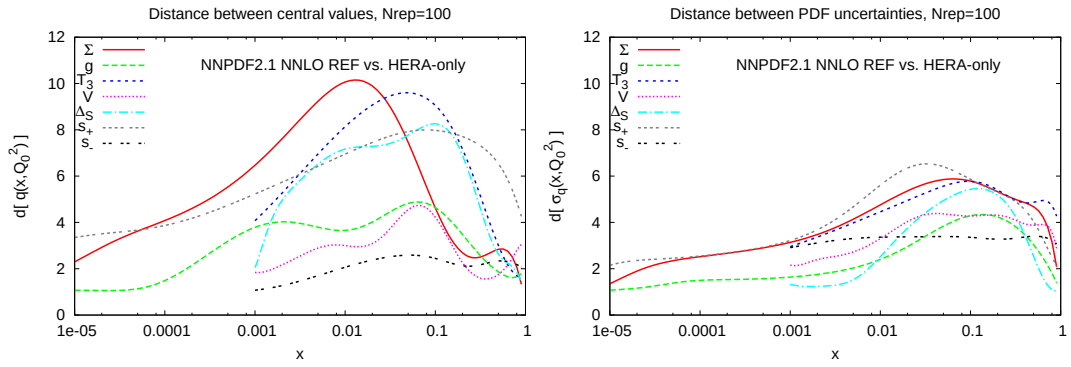

Figure 5.17: Distances between central values (left) and uncertainties (right) for PDFs in the HERA-only and default NNPDF2.1 NNLO fits. All distances are computed from sets of $N_{\rm rep} = 100$ replicas.

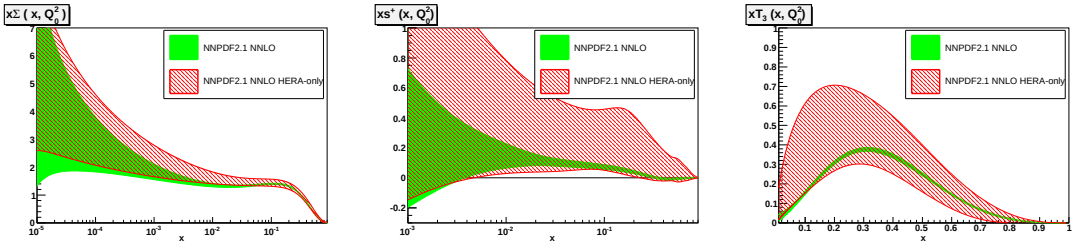

Figure 5.18: Comparison of NNPDF2.1 NNLO singlet, total strangeness and isotriplet PDFs in the global and in the HERA-only fits.

ogy (and in particular without having to change the parametrization, see Ref. [85]) is an advantage of the NNPDF approach, since it allows a meaningful comparison of uncertainties. Each of these PDF sets is made available through the standard LHAPDF interface, and each so far as it goes is as good as the default one, the only difference between them being the smaller amount of experimental information that goes into them.

In Table 5.5 the total χ^2 and that of the individual experiments of the fits presented above are shown, together with the global NNPDF2.1 NNLO fit and the respective number of points included in the fit. The columns of Table 5.5 are sorted by decreasing complexity of the fitted data and the χ^2 values shown in square brackets correspond to the experiments excluded from the corresponding fit.

To quantify the comparison between each of these PDF sets, again are used the distances introduced in Appendix A of Ref. [87] and already used in previous Chapters. In order to assess the impact of individual data each of the reduced-data fits, the default NNPDF2.1, and the fit with immediately greater complexity are compared in turn. Some of the pairs of PDFs with the largest distances will also be compared directly. In this way

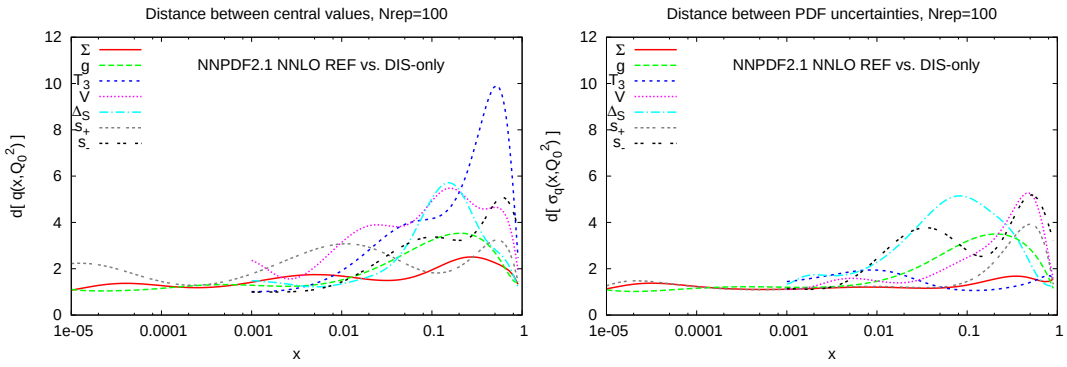

Figure 5.19: Distances between central values (left) and uncertainties (right) for PDFs in the DIS-only and default NNPDF2.1 NNLO fits. All distances are computed from sets of $N_{\rm rep} = 100$ replicas.

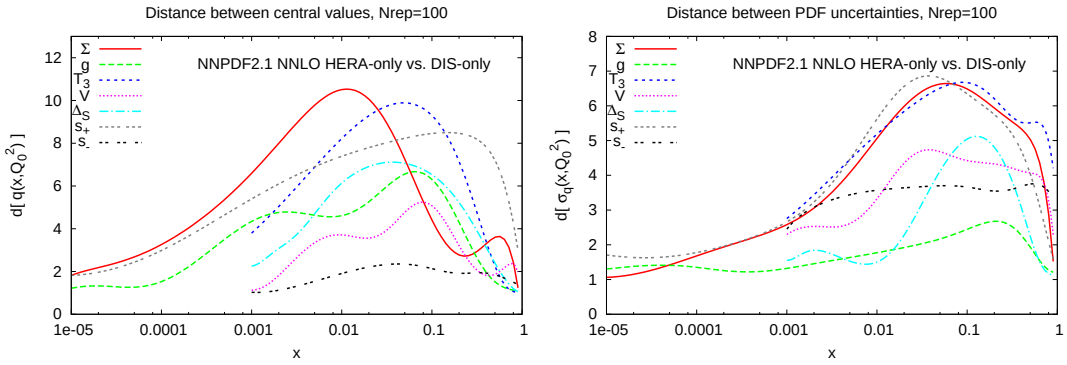

Figure 5.20: Distances between central values (left) and uncertainties (right) for PDFs in the HERA-only and DIS-only NNPDF2.1 NNLO fits. All distances are computed from sets of $N_{\rm rep} = 100$ replicas.

the impact of individual data on PDFs is assessed.

Considering the first, smaller reduced dataset with only HERA data included, rises the problem that charged current DIS data are enough to determine at most four independent linear combinations of quark PDFs (see e.g. Ref. [226]). This implies that strangeness is completely unconstrained: in the HERAPDF [114] set an independent parametrization is only provided for the combinations $d + s$ and $\bar{d} + \bar{s}$, but not separately for strangeness. As can be noticed from Table 5.5, the fit quality to NuTeV data is extremely poor, as also poor is the description for all datasets which are sensitive to the singlet-triplet separation (such as fixed-target DIS and DY data) to the light sea decomposition (such as W production data) and, to a lesser extent, the valence-sea separation (such as neutrino data). Of the excluded datasets, jet data are the best described but still marginal quality is found.

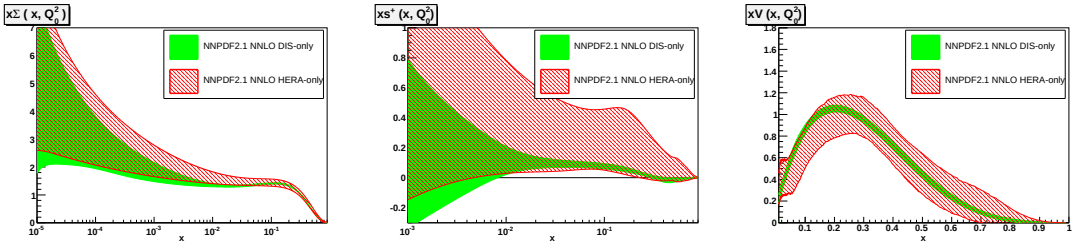

Figure 5.21: Comparison of NNPDF2.1 NNLO singlet, total strangeness and total valence PDFs in the DIS-only and in the HERA-only fits.

The distances for the HERA-only fit compared to the default one are shown in Fig. 5.17. Triplet, strangeness, sea asymmetry, and valence distributions have the largest distances between uncertainties, and triplet and strangeness central values have also large distances. For singlet and gluon are observed even larger distances: the large shifts in strangeness and triplet are accompanied by a corresponding increase in their uncertainty, while the increase in the uncertainty of the singlet and gluon is more moderate, so the change in central value ends up being statistically more significant, as can be clearly seen in the direct PDF comparison of Fig. 5.18. It can be clearly seen that the singlet and gluon have a distance from the NNPDF2.1 NNLO global fit that is more than one sigma, while all other PDFs stay around the 90% C.L. If W and Z production at the LHC is considered, the strange contribution is of order 15-25% of the total cross-section [227]. Using HERA-only PDF sets to compute cross-sections which depends on strangeness gives a theoretical uncertainty of that same order. This implies a very uncertain phenomenology for this kind of predictions. This uncertainty is not due to inaccuracy of HERA data from which this PDFs set is extracted, but to lack of information necessary to disentangle and constrain each flavor contribution.

Flavor decomposition considerably improves by including in the fit data with neutrino beams or deuterium targets, as is the case for DIS-only PDF determination. Gluon and singlet distances (Fig. 5.19) show almost perfect agreement with the global fit, with only a deviation around the half-sigma level in the large-x region of the gluon distribution. Instead, distances between strangeness, valence and light sea asymmetry determined in this fit and those of the global fit, are rather smaller than one sigma (though uncertainties are still significantly larger). A direct comparison of this fit with the HERA-only fit is shown in Fig. 5.21: the most remarkable improvement is observed in singlet, strangeness, and valence PDFs. The triplet distribution also shows a significant decrease in uncertainty, but around the valence peak it only agrees with that of the global fit at the 90% confidence level. This may suggest some tension between deuterium DIS and hadron collider data (as has been discussed elsewhere [76, 228]), though it could also be a statistical fluctuation.

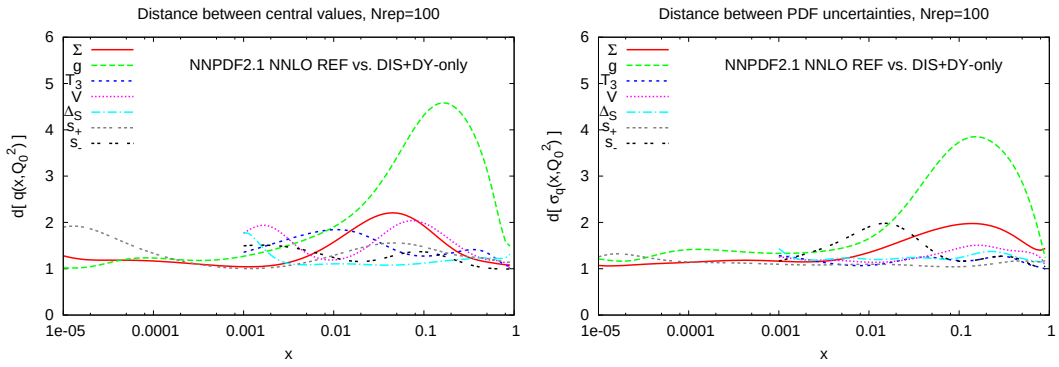

Figure 5.22: Distances between central values (left) and uncertainties (right) for PDFs in the DIS+DY and default NNPDF2.1 NNLO fits. All distances are computed from sets of $N_{\rm rep} = 100$ replicas.

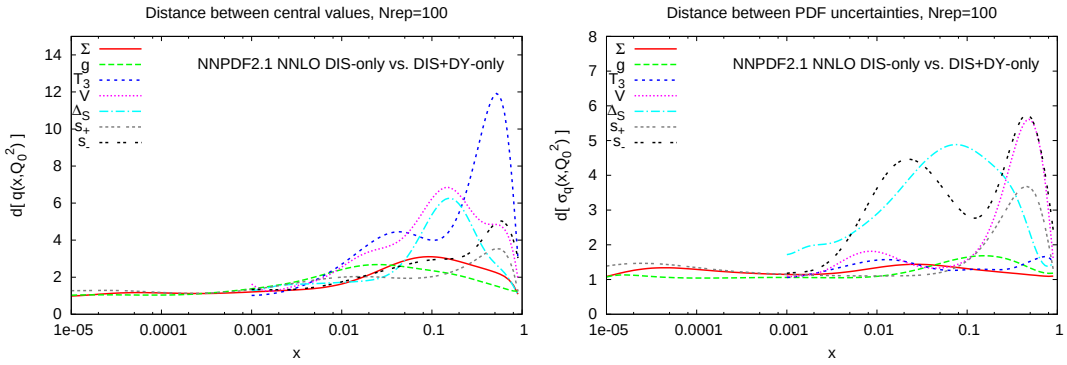

Figure 5.23: Distances between central values (left) and uncertainties (right) for PDFs in the DIS-only and DIS+DY NNPDF2.1 NNLO fits. All distances are computed from sets of $N_{\rm rep} = 100$ replicas.

Another interesting comparison can be made by looking at the drastic improvement in uncertainty distances going from Fig. 5.20 to Fig. 5.19.

Despite all the improvements commented above in comparison to the HERA-only fit, looking at Table 5.5 a poor description of all DY data is given, especially for W and Z production, thus showing that a DIS-only PDF fit is not adequate for precision hadron collider phenomenology. It is anyway remarkable that the quality of jet data description is pretty near to the one given by the global fit. The gluon distribution determined by DIS scaling violations is in good agreement with that of the global fit, even in the large-x region where jet data have an impact on its uncertainty. Other PDF fits based on reduced datasets (such as HERAPDF or ABKM) do not seem [184] to provide an equally good fit to jet data, presumably because of their less flexible PDF parametrization.

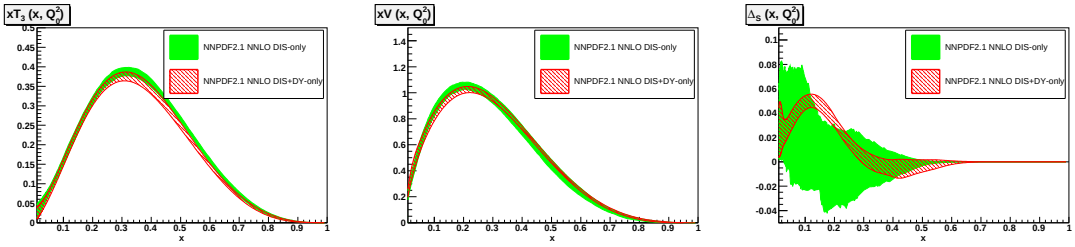

Figure 5.24: Comparison of NNPDF2.1 NNLO isotriplet, total valence and sea asymmetry PDFs in the DIS-only and in the DIS+DY fits.

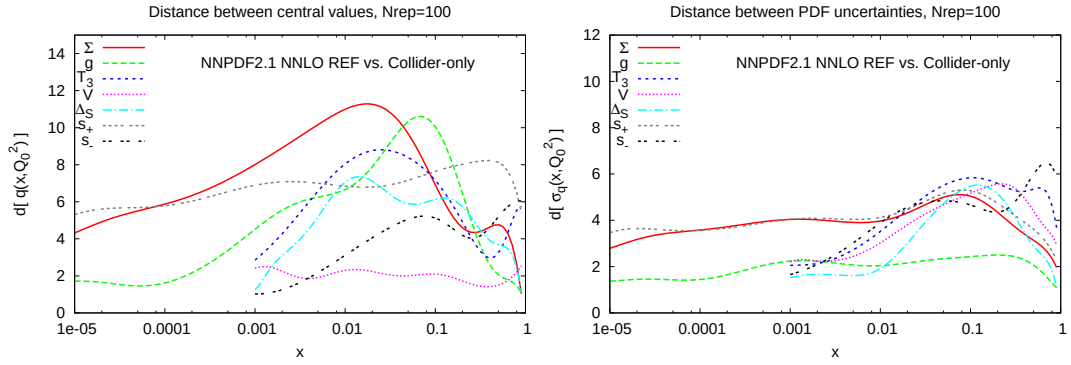

Figure 5.25: Distances between central values (left) and uncertainties (right) for PDFs in the collider-only and default NNPDF2.1 NNLO fits. All distances are computed from sets of $N_{\rm rep} = 100$ replicas.

It is natural then to consider the DIS+DY fit. Observing the plot of distances with respect to the global fit represented in Fig. 5.22, it can be seen that almost all PDFs are statistically equivalent. If the distances with respect to the DIS-only fit are also computed, as shown in Fig. 5.23, it is clear that the inclusion of Drell-Yan data provides information that helps flavor separation: the uncertainties of the flavor decomposition are strongly reduced, as can also be seen by directly comparing PDFs (Fig. 5.24).

The gluon presents a shift of about half a sigma of its central value at large-x. This shows explicitly that jet data has an impact on gluon PDF, constraining it at large-x where no other dataset among the ones used in the global fit include the needed information. Also a related slight shift of the singlet and valence distributions is observed.

Finally, the case of a collider-only fit is analyzed: the same datasets as in the HERA-only fits supplemented by the Tevatron weak-boson production and inclusive jet production data are used. Comparing from Table 5.5 the values of the total χ^2 for this fit and the global one, a significant improvement from χ^2 =1.16 of the latter to χ^2 =1.02 of the former is noticed. This fact supports the idea that a collider-only dataset might be more

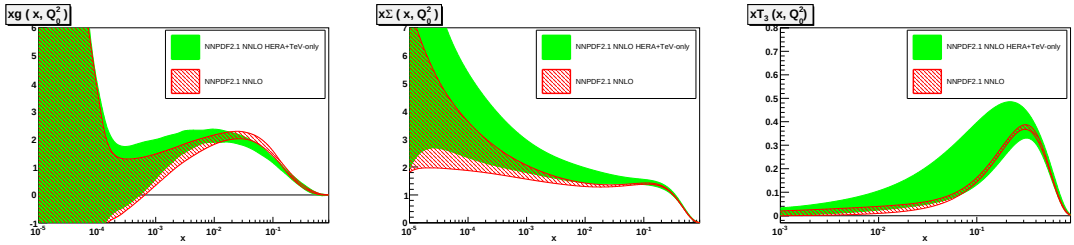

Figure 5.26: Comparison of NNPDF2.1 NNLO gluon, singlet and triplet PDFs in the collider-only and reference fits.

consistent than one which also includes fixed-target data.

In spite of this improvement, checking the distances with respect to the global fit (Fig. 5.25) reveals that almost all PDFs undergo a shift at the one-sigma level, with uncertainties significantly larger in the collider-only fit. The description that this set gives of fixed-target data is generally poor, especially for deuterium data (such as NMCpd or DYE866), which control the up-down separation, and the dimuon data (which control strangeness). These increased uncertainties end up contaminating also the singlet and gluon PDFs, as shown by the distance computation and by direct PDF comparison (Fig. 5.26). The extremely poor determination of strangeness and light quark asymmetry PDFs and the slightly better but still rather poor determination of valence, triplet and even singlet PDFs show that a collider-only fit at present does not provide competitive accuracy for phenomenology.

5.5 Determination of $\alpha_S(M_z)$ Using NNPDF2.1 NLO

The value of the strong coupling constant is a fundamental ingredient for collider phenomenology [229]. If for example the gluon fusion channel for Higgs production is considered, the dominant source of uncertainty is due to the value of $\alpha_s\left(M_Z\right)$ [204]. The PDG [47] value is determined by a combination of results obtained by different processes as the τ decay rate and the total $e^+e^-\rightarrow$ hadrons cross-section, which are independent of PDFs, but also DIS data, which indeed depends on them. The value given in Ref. [229] is

$$\alpha_s\left(M_Z\right) = 0.1184 \pm 0.0007 \tag{5.5}$$

and as the uncertainty in this result may seem too optimistic given the spread of value on which is based and the significant dependence on the perturbative order of some of them, the use of a somewhat more conservative estimate of the uncertainty has been recommended [204, 230], such as $\Delta\alpha_s = 0.0012$ at 68% confidence level for LHC phe-

nomenology.

To exploit the same data used for PDFs determination to determine a value of $\alpha_s(M_Z)$ is interesting in that it is possible at the same time to take advantage of the dependence on the coupling of scaling violations as well as that on individual hard matrix elements of the various processes under consideration. A quite accurate determination is therefore expected. However, in such a determination the value of α_s is necessarily correlated to the best-fit form of the PDFs, and thus subject to potential sources of bias, such as for example an insufficiently flexible PDF parametrization.

An example of the possible pitfalls of a simultaneous determination of PDFs and α_s is highlighted by the analysis of Ref. [231], in which the extraction of α_s from BCDMS and NMC deep-inelastic scattering data was performed using a methodology (scaling violations of truncated moments) which avoids completely the use of parton distributions. The result found, $\alpha_s(M_Z) = 0.124^{+0.005}_{-0.008}$, had rather different central value and uncertainties than those obtained by direct analysis of the same BCDMS ($\alpha_s(M_Z) = 0.113 \pm 0.005$ [232]) and NMC ($\alpha_s(M_Z) = 0.117^{+0.011}_{-0.016}$ [233]) data by the respective collaborations. This suggests that the latter results, obtained using a PDF parametrization, were biased by it.

The fit to α_s that will be discussed in this Section is determined using the previously presented NNPDF2.1 NLO parton set (an analog determination using the NNLO set is given in [234]). The NNPDF methodology, as widely commented, is intended to minimize the parametrization bias through the use of a Monte Carlo approach combined with neural networks as underlying unbiased interpolating functions. Moreover, the use of the NNPDF approach allows for the analysis of different datasets without having to retune the fitting procedure (such as, for instance, the form of parton parametrization) according to the size of the dataset. This enables a direct comparison of values of α_s obtained from different subsets of data which enter the global fit, and also an analysis of the correlation between individual datasets, individual PDFs, and the value of α_s. In this way, it will also be possible to check if DIS data systematically settle on lower values of α_s than hadron or e^+e^- collider data.

The quality of the fit as it is delivered by the NNPDF collaboration, i.e. as a Monte Carlo sample, is a random variable which tends to a constant value for an infinite number of replicas. The size of the fluctuations of the χ^2 is of the order of the square root of the number of data points included in the fit $N_{\rm dat}$, while the size of fluctuations of the average over a sample of $N_{\rm rep}$ replicas decreases as $1/N_{\rm rep}^{1/2}$. The value of the χ^2 for a typical NNPDF fit (with $N_{\rm dat} = 3338$) is shown as a function of $N_{\rm rep}$ in Fig. 5.27. To reduce fluctuations in order to be sensitive to variations of the total χ^2 by a few units one needs for each value of α_s a number of replicas of the same order of magnitude as the number

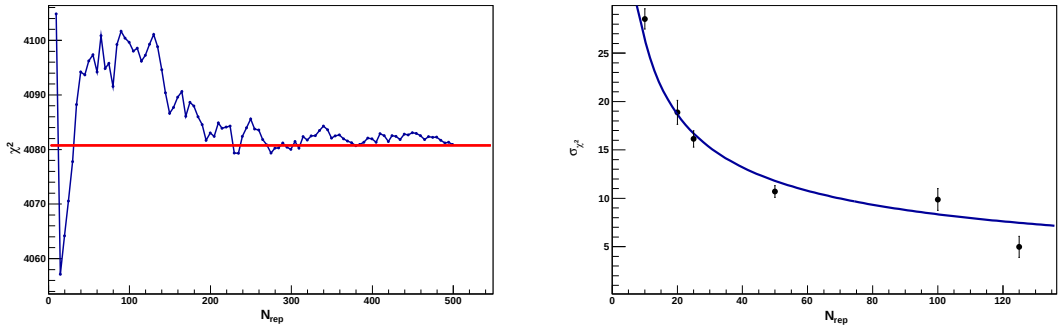

Figure 5.27: Left: The χ^2 as a function of the number of replicas $N_{\rm rep}$ for NNPDF2.0; the horizontal line shows the value for $N_{\rm rep} = 500$. Right: The uncertainty σ_{χ^2} Eq. (5.6) as a function of $N_{\rm rep}$, averaged over all values of α_s. A fit of the form $AN_{\rm rep}^{-1/2}$ is also shown.

of independent data points. This implies a quite large number of replicas to determine α_s and thus a rather computationally intensive effort.

The uncertainty on the value of the χ^2 due to the finite size of the replica sample may be computed using the so-called bootstrap method. Namely, the sample of $N_{\rm rep}$ replicas is divided into $N_{\rm part}$ disjoint partitions with $\widetilde{N}_{\rm rep} = N_{\rm rep}/N_{\rm part}$ replicas each. The variance of the χ^2 for the full $N_{\rm rep}$ replica sample is then found from the variance of the $N_{\rm part}$ values $\widetilde{\chi}^2$ of each replica subsample according to

$$\left(\sigma_{\chi^2}\right)^2 \equiv \frac{1}{N_{\rm part}} \left[\frac{1}{N_{\rm part}} \sum_{k=1}^{N_{\rm part}} \left(\widetilde{\chi}^2_k\right)^2 - \left(\frac{1}{N_{\rm part}} \sum_{k=1}^{N_{\rm part}} \widetilde{\chi}^2_k \right)^2 \right] . \tag{5.6}$$

The value of σ_{χ^2}, averaged over all the (eleven) values of α_s to be considered, is displayed in Fig. 5.27. A fit of the form $AN_{\rm rep}^{-1/2}$, also shown in Fig. 5.27, shows that the expected decrease of the fluctuations with $1/N_{\rm rep}^{1/2}$ is borne out by the data.

The procedure for the α_s determination is quite simple: once a sufficiently large set of PDF replicas has been produced for N_{α_s} fixed values of α_s, for each of these values the corresponding χ^2 and their uncertainty σ_{χ^2} Eq. (5.6) are computed. Then, a parabolic fit is performed on these χ^2 values, seen as a function of α_s. The quality of the parabolic fit is then determined by evaluating the corresponding $\chi^2_{\rm par}/N_{\rm dof}$, with $N_{\rm dof} = N_{\alpha_s} - 3$. A reasonable value of $\chi^2_{\rm par}/N_{\rm dof}$ may be used to confirm that the parabolic approximation to $\chi^2(\alpha_s)$ is adequate in the range of α_s under investigation. The best value for the strong coupling constant is then given by the minimum of the parabola and the $\Delta\chi^2 = 1$ range gives the uncertainty on it at a 68% confidence level. The further uncertainty due to the finite size of the replica sample is determined by error propagation of σ_{χ^2} Eq. (5.6) on the position of the minimum of the parabola.

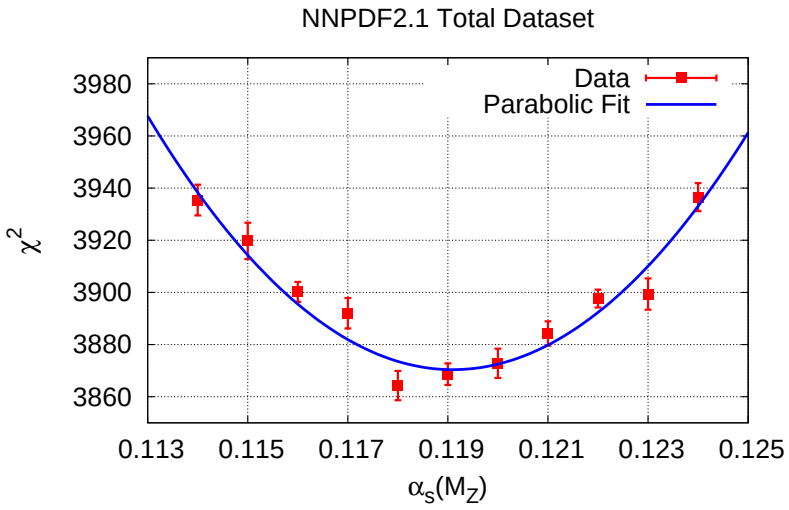

Figure 5.28: The χ^2 as a function of $\alpha_s(M_Z)$ for the NNPDF2.1 global fit. The statistical uncertainties in the χ^2 for each value of α_s have been determined from Eq. (5.6). The solid line is the result of a parabolic fit.

The number of replicas used for each value of α_S is $N_{\rm rep} = 500$, with equally-spaced larger sets with $N_{\rm rep} = 1000$ meant to increase accuracy. Together with the global NNPDF2.1 NLO fit, the two reduced sets over HERA- and DIS-only data are also considered, along with the NNPDF2.0 global and DIS-only sets. Considering the HERA data only fit, the sensitivity to α_s is weaker due to the much smaller size of the data sample, and so the range of α_s values is enlarged as shown in Table 5.6 to ensure that the location of the minimum is more or less at the center of that range. In Table 5.6 are also gathered the α_S values and their respective number of replicas for the NNPDF2.1 global and DIS-only fits.

In Fig. 5.28 the parabolic profile of χ^2 as a function of $\alpha_s(M_Z)$ is shown for the NNPDF2.1 global fit. The results for the DIS-only and HERA-only data fits are presented in Fig. 5.29. The numerical values of each fit are summarized in Table 5.7, where the values and uncertainties of $\alpha_s\,(M_Z)$ are shown, with the *exp* experimental uncertainty determined from the $\Delta\chi^2 = 1$ range and with the *proc* propagated procedural uncertainty, due to the finite size of the replica sample. The quality of the parabolic fit is also shown in each case through the $\chi^2_{\rm par}/N_{\rm dof}$ value. The fits denoted by the *red.* label are simply equal to the respective standard fits but with $N_{\rm rep} = 500$ for each α_S value considered in the procedure.

The fact that a further increase in the number of replicas used is not needed can be seen by looking at the procedural uncertainty: its value is in general very small and with $N_{\rm rep} = 500$ for all α_S values almost no fluctuation at all is observed on the uncertainty,

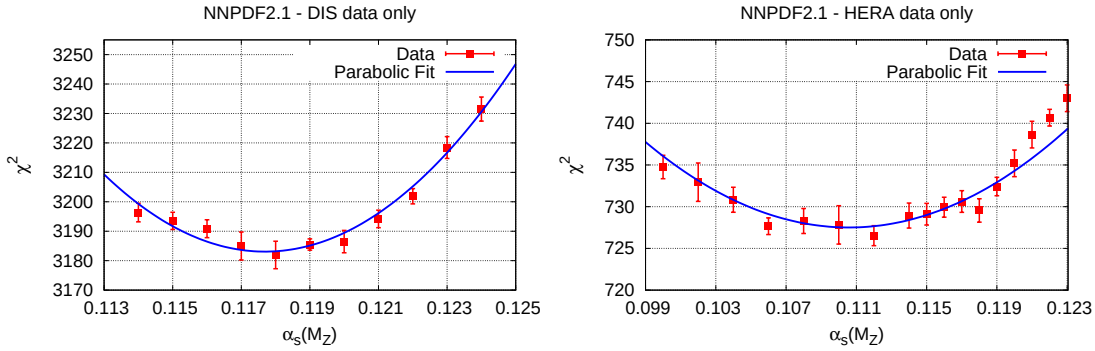

Figure 5.29: Same as Fig. 5.28 but for a fit to DIS data only (left) and to HERA data only (right).

2.1 global	
$\alpha_s\,(M_Z)$	$N_{\rm rep}$
0.114	500
0.115	500
0.116	1000
0.117	500
0.118	500
0.119	1000
0.120	500
0.121	500
0.122	1000
0.123	500
0.124	500

2.1 DIS–only	
$\alpha_s\,(M_Z)$	$N_{\rm rep}$
0.114	500
0.115	500
0.116	1000
0.117	500
0.118	500
0.119	1000
0.120	500
0.121	500
0.122	1000
0.123	500
0.124	500

2.1 HERA–only	
$\alpha_s\,(M_Z)$	$N_{\rm rep}$
0.100	1000
0.102	500
0.104	500
0.106	1000
0.108	500
0.110	500
0.112	1000
0.114	500
0.115	500
0.116	1000
0.117	500
0.118	500
0.119	1000
0.120	500
0.121	500
0.122	1000
0.123	500
0.124	500

Table 5.6: The values of $\alpha_s\,(M_Z)$ and the number of replicas $N_{\rm rep}$ used in each case for various determinations of $\alpha_s\,(M_Z)$.

thus confirming the reliable determination of the finite-size uncertainty. Also, the change in value of α_s is always smaller than the procedural uncertainty as the number of replicas is decreased. The χ^2 values of the parabolic fit are expected to follow a χ^2 distribution with $N_{\rm dof} = 8$ degrees of freedom for the global and DIS fits, and $N_{\rm dof} = 12$ for the HERA only fit. The standard deviation of $\chi^2_{\rm par}/N_{\rm dof}$ is thus expected to be of order 0.5, as indeed observed. Another check that has been performed is to exclude the points at the edge of the fit, adding extra parameters: it follows that results are almost unaffected and no improvement in fit quality is observed.

In Fig. 5.30 the results for various best-fit α_S values are shown. It is interesting to notice the correct behaviour of uncertainties, that increase when reducing the size of the dataset, and in particular the good agreement between the NNPDF2.1 best-fit value and the published PDF value. Moreover, the found experimental uncertainty is surprisingly small. Interestingly, the value found using HERA data only is much smaller, even though, because of the considerable (almost sixfold) increase in statistical uncertainty it is still less

	$\alpha_s(M_Z)$	$\chi^2_{\rm par}/N_{\rm dof}$
NNPDF2.1	$\mathbf{0.1191 \pm 0.0006^{\rm exp} \pm 0.0001^{\rm proc}}$	**1.6**
NNPDF2.1 DIS–only	$0.1178 \pm 0.0009^{\rm exp} \pm 0.0002^{\rm proc}$	0.7
NNPDF2.1 HERA–only	$0.1101 \pm 0.0033^{\rm exp} \pm 0.0003^{\rm proc}$	0.7
NNPDF2.1 *red.*	$0.1191 \pm 0.0006^{\rm exp} \pm 0.0001^{\rm proc}$	1.5
NNPDF2.1 DIS–only *red.*	$0.1177 \pm 0.0009^{\rm exp} \pm 0.0002^{\rm proc}$	0.5
NNPDF2.1 HERA–only *red.*	$0.1103 \pm 0.0032^{\rm exp} \pm 0.0004^{\rm proc}$	1.1
NNPDF2.0	$0.1168 \pm 0.0007^{\rm exp} \pm 0.0001^{\rm proc}$	0.4
NNPDF2.0 DIS–only	$0.1145 \pm 0.0010^{\rm exp} \pm 0.0003^{\rm proc}$	1.4.

Table 5.7: Values of $\alpha_s(M_Z)$ and associated uncertainties. All uncertainties shown are 68% confidence levels, with the experimental uncertainty obtained by requiring $\Delta\chi^2 = 1$ about the minimum, and the procedural uncertainty from propagation of σ_{χ^2} Eq. (5.6) due to finite size of the replica sample. The quality of the parabolic fit as measured by $\chi^2_{\rm par}/N_{\rm dof}$ is also shown in each case. For the global, DIS-only and HERA-only fits (first three rows), the maximum number of replicas, given in Tab. 5.6, has been used. The three reduced replica fits (subsequent three rows) only differ from these because of the use of $N_{\rm rep} = 500$ for all α_s values. The NNPDF2.0 fits of the last two rows also have $N_{\rm rep} = 500$ always.

than three σ from the global fit. The fact that HERA data prefer a lower value of α_s may be related to the deviations between HERA data and the predicted NLO scaling violations which was observed in Refs. [163, 164] for the smallest x and Q^2 HERA data. These may be affected by small-x resummation or saturation effects. As shown there, scaling violations in this region are weaker than predicted from the behaviour observed in other kinematic regions, and thus would tend to bias the value of α_s downwards. A dedicated analysis would be required to prove conclusively that this is the case. Focusing on the DIS-only result, there is no evidence that a lower value of α_s is preferred: the results found in this case are perfectly compatible with the global fit at the one-sigma level.

It is possible to briefly comment on theoretical uncertainties in NNPDF fit, without the intention of giving to this issue an exhaustive and detailed description. A possible source of theoretical uncertainty could be introduced by inefficiencies of the global PDF fit (such as, for example, any residual bias related to parton distributions). This kind of uncertainty should show up in the behaviour of the χ^2 as a function of α_s, either as point-to-point fluctuations or as a systematic deviation from the underlying unbiased quadratic behaviour (if they are correlated to the value of α_s). The good quality of the parabolic fit suggests that these uncertainties are small, and thus that NNPDF uncertainty is an accurate assessment of the total uncertainty due to the statistical and systematic uncertainties in the experimental data. Indeed, the uncertainties due to NNLO and higher orders in perturbative QCD introduced in the computation of the several different processes included in the analysis still need to be assessed and surely affect up to some extent the results here presented. Other possible sources of theoretical uncertainty may be due to the lack of resummation

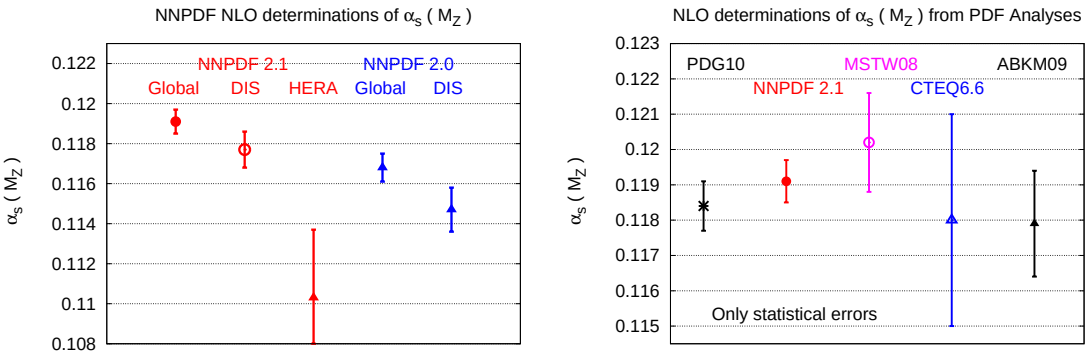

Figure 5.30: Left: Graphical representation of the values of $\alpha_s(M_Z)$ of Table 5.7 (reduced replica fits not included). Uncertainties have been added in quadrature. Right: comparison to other recent determinations of $\alpha_s(M_Z)$ from NLO PDF analysis. The PDG value of Ref. [229] is also shown.

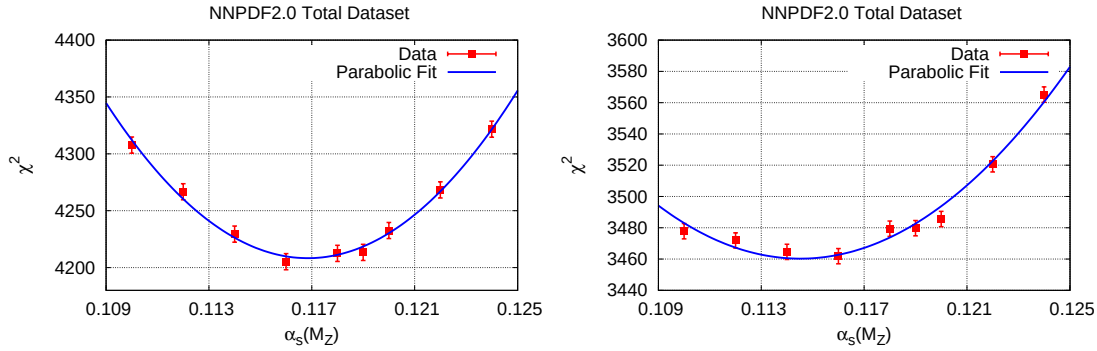

Figure 5.31: Same as Fig. 5.28 but for the NNPDF2.0 global fit (left) and NNPDF2.0 DIS only fit (right).

of higher order QCD corrections in some kinematic regions and to the treatment of heavy quark mass effects. In cases in which both NLO and NNLO determinations are available, such as Refs. [82, 94], a sizable downward shift of the best-fit value, of order of several percentage points, has been observed when going from NLO to NNLO. Looking at the more recent result obtained using NNPDF2.1 NNLO in Ref. [234] good stability is found: the shift is around 1.5%.

In order to give an approximated estimate of the impact of heavy quark masses treatment, the α_S fit has been also performed using the NNPDF2.0 [87] PDF set, both with global and DIS-only dataset, which is based on a zero-mass variable flavor number scheme in which all heavy quark masses are neglected. In these fits, all α_S values considered in the procedure have a number of replicas $N_{\rm rep} = 500$. Considering again Table 5.7 and Fig. 5.30, a remarkable downwards shift is observed for both the NNPDF2.0 determina-

tions here considered: if one were to conservatively estimate the uncertainty due to heavy quark mass effects as the difference between the NNPDF2.1 and NNPDF2.0 results one would get, for the global fit, $\Delta\alpha_s^{\rm hq} \approx 0.002$. In the DIS-only case a still larger shift is observed, because hadronic data is unaffected by the treatment of heavy quark mass effects. The theoretical uncertainty due to neglected perturbative higher order corrections is expected to be of the same order of magnitude, being thus the dominant uncertainty for these fits but also for the ones from other groups.

The result obtained from the global NNPDF2.1 NLO set is compared with the ones determined using PDF sets from other groups: MSTW [82], CTEQ [203] and ABKM [94] PDF sets. The CTEQ collaboration is the only one among the groups here considered that to determine the best fit value and its uncertainty studies the dependence of the fit quality on α_s as is done here. The other two, ABKM and MSTW, fit the α_s value introducing it as a free parameter in the parton fit performed to determine the PDF set. In this way they obtain a correlated Hessian matrix which mixes the PDF parameters with α_s. The equivalence of the α_s uncertainty obtained from either methods is explicitly shown in Ref. [203]. Observing the plot in Fig. 5.30, the agreement among the various different results is clear. It can be seen that CTEQ and MSTW [82, 203] have a quite larger statistical uncertainty. This is due to the tolerance [97] criterion, that is based on a substantial rescaling of the uncertainty ranges in parameter space. In NNPDF procedure this is not needed because the χ^2 fluctuations are kept under control by a suitable choice of the size of the Monte Carlo sample, as discussed above. The dataset used in the ABKM determination is smaller than the other ones because it does not include collider jet and vector boson production data. The CTEQ and MSTW collaborations instead use almost the same dataset as done in the NNPDF analysis (see Sect. 2.5).

As a cross-check, the Gaussianity of the distribution of results obtained performing a fit experiment by experiment is verified, applying a variant of the method proposed in Ref. [235]. The results are displayed in Fig. 5.32 together with the uncertainties due to the finite size of the replica sample, determined as above. For each experiment that is included in the global fit already performed, a dedicated parabolic fit to the χ^2 profile is operated. The results for each of them (among the ones that have a minimum in the investigated range) are shown in Table 5.8. Only NMCratio, SLAC, CHORUS, H1F2C and FLH108 data have no minimum in the fitted range.

The distribution of results can be studied defining the pull

$$P_i \equiv \frac{\alpha_s^i\left(M_Z\right) - \alpha_s^{\rm tot}\left(M_Z\right)}{\sqrt{\sigma_{\alpha_s}^{i,2} + \sigma_{\alpha_s}^{\rm tot,2}}}, \tag{5.7}$$

where $\alpha_s^i\left(M_Z\right)$ is the best fit value for the i-th experiment and $\sigma_{\alpha_s}^i$ the associated statisti-

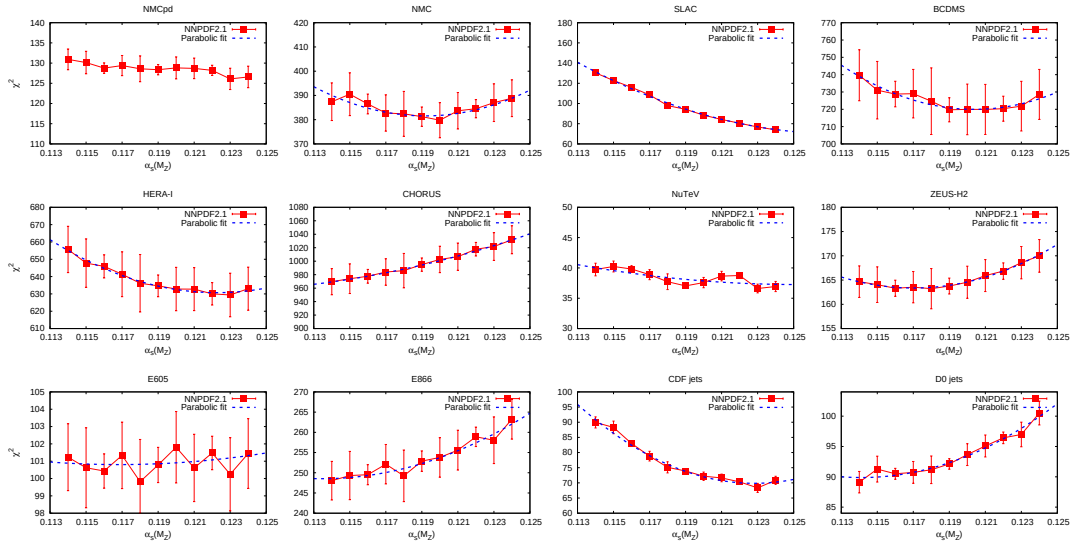

Figure 5.32: The χ^2 profiles for the individual experiments in the NNPDF2.1 global fit together with the results of the corresponding parabolic fits to α_s. The uncertainties due to the finite size of the replica sample are shown on each value.

cal uncertainty, obtained from the $\Delta\chi^2 = 1$ rule. The pulls are summarized in Table 5.8 and displayed graphically in Fig. 5.33. A Gaussian fit to the distribution of pulls is performed and also displayed in Fig. 5.33. The Gaussian fit is in good agreement with the histogram data with mean $\langle P \rangle = 0.04$ and standard deviation $\sigma_P = 1.3$. The standard deviation would be further reduced somewhat if finite-size uncertainties were included. This however would require a lengthy correlation analysis. The conclusion is that the value of the tolerance required to get a perfectly Gaussian distribution of pulls is smaller than 1.3 — a value which is clearly compatible with a statistical fluctuation.

Thanks to the particular procedure used, that is the same for datasets of different size, the issue of whether and why different datasets may prefer different values of α_s can be addressed. In Fig. 5.34 the χ^2 profiles for the global NNPDF2.1 fit, already shown in Fig. 5.32, are compared to the same quantities determined for the fit to DIS data only. The interesting behaviour of the BCDMS dataset is discussed: as already observed in several other studies [82, 232, 236], this dataset seem to prefer a lower value of α_S, in particular lower than Eq. (5.5). There is a clear difference in Fig. 5.34 between the χ^2 profile for global or DIS-only fits. This seems to suggest that, using a lower value of α_S, the parton fit would show a better quality because of this allowed direction in DIS data, but in the global fit the quality that in DIS sector improves would worsen somewhere else.

A clearer picture of the situation can be gained by studying the correlation [40] between parton distributions and the value of the χ^2 for individual experiments. Looking at the best-fit results, if nonvanishing correlations are present means that the fit for that

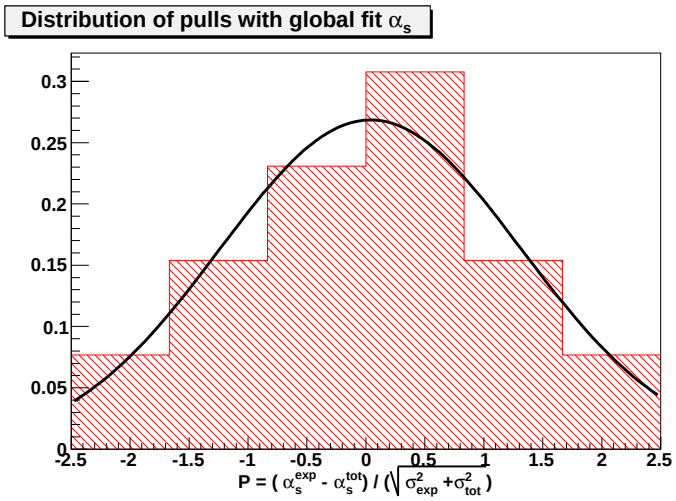

Figure 5.33: Distribution of pulls, Eq. (5.7), for the value of α_s preferred by the individual experiments included in the global fit. These pulls have been summarized in Table 5.32).

specific dataset can be improved by changing the given PDF. Correlations of opposite sign denote a "tension" between different datasets: the PDF is pulled in opposite directions. As an example, here the gluon PDF is considered, and the correlation coefficient for a pair of values of x as a function of α_s is computed and the result is shown in Fig. 5.35 (all from a set of $N_{\rm rep} = 500$ replicas). While for large values of α_S all correlations have the same sign, near the best-fit value of the strong coupling HERA and BCDMS data on one side and jet data on the other pull the gluon in opposite directions. Hence, a determination of α_s including DIS data only can easily be biased. The fact that BCDMS data prefer a lower value of α_s in a DIS-only fit, but not if the gluon is constrained by jet data was also found recently in Ref. [236], in the context of the MSTW08 parton determination. However, in that case the BCDMS data was also found to significantly bias downwards the value of α_s for the DIS fit, perhaps due to the fact that the MSTW gluon parametrization, though more flexible than that of other groups, is still less flexible than that of the present analysis.

The fact that runaway directions for the χ^2 may appear in the joint α_s-gluon space can be understood by noting that in DIS the gluon is determined by scaling violations, hence a smaller value of α_s can be partially compensated by a larger gluon and conversely. However, the jet cross-section pins down the size of the gluon (at the rather larger scale of the jet data) thereby quenching this potential instability. Hence, can be concluded that even though in this fit the DIS-only value of α_s is not significantly smaller than that for the global fit (possibly due to the great flexibility of the functional form of these PDFs), a fit to DIS data, and specifically to BCDMS data, has a potential instability in the direction of lower values of α_s which is only kept under control by the inclusion of jet data.

Experiment	$\alpha_s^i \pm \sigma_{\alpha_s}^i$	P_i
NMCp	0.1192 ± 0.0018	-0.05
BCDMS	0.1204 ± 0.0015	-0.78
HERA-I	0.1223 ± 0.0018	-1.65
ZEUS-H2	0.1170 ± 0.0027	0.75
NuTeV	0.1252 ± 0.0068	-0.89
ZEUSF2C	0.1144 ± 0.0060	0.77
E605	0.1168 ± 0.0100	0.22
E866	0.1135 ± 0.0029	1.87
CDFWASY	0.1181 ± 0.006	0.16
CDFZRAP	0.1150 ± 0.0034	1.18
D0ZRAP	0.1227 ± 0.0067	-0.53
CDFR2KT	0.1228 ± 0.0021	-1.67
D0R2CON	0.1141 ± 0.0031	1.57

Table 5.8: The pulls P_i Eq. (5.7) for individual experiment included in the NNPDF2.1 global fit case, computed for each experiment which has a minimum in the range considered.

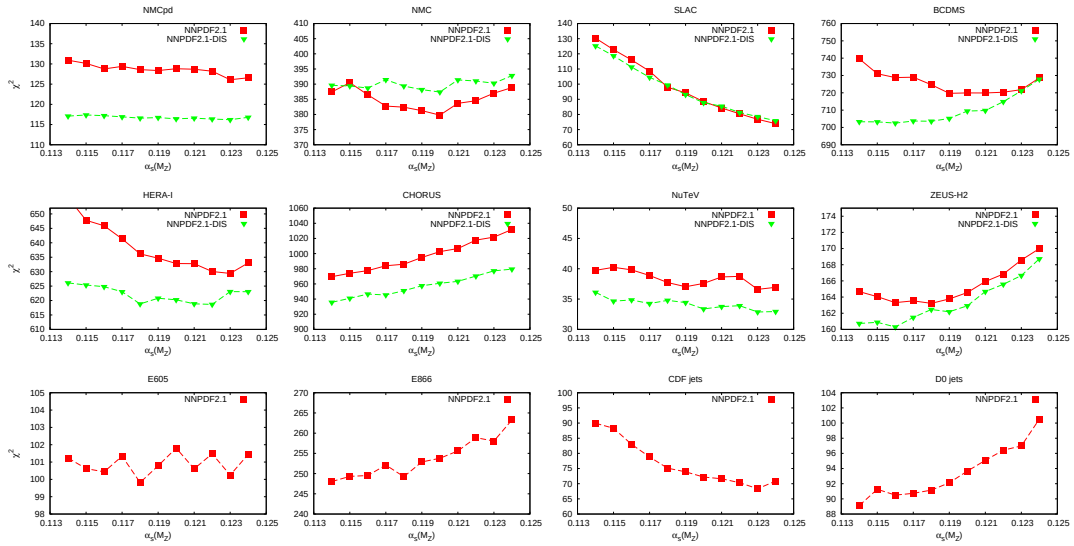

Figure 5.34: Comparison of the χ^2 profiles for the global NNPDF2.1 fit (same as in Fig. 5.32; red, solid curves) to those determined for the DIS-only NNPDF2.1 fit (green, dashed curves).

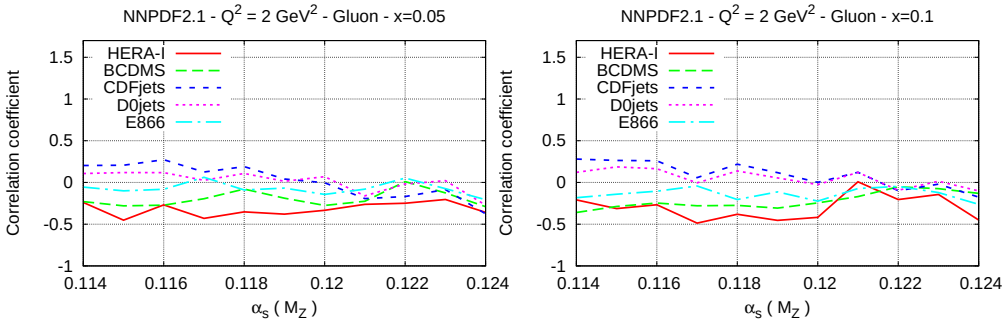

Figure 5.35: Correlation between the χ^2 and the input gluon as a function of $\alpha_s(M_Z)$ for $x = 0.05$ (left) and $x = 0.1$ (right) for the NNPDF2.1 global PDF set.

Summarizing the results discussed in this Section, the strong coupling α_s has been determined from a next-to-leading order analysis of processes used for the NNPDF2.1 global parton determination. A value $\alpha_s(M_Z) = 0.1191 \pm 0.0006^{\rm exp}$ has been found, where the uncertainty includes all statistical and systematic experimental uncertainties, but not purely theoretical uncertainties, which are expected to be rather larger. A detailed study of the dependence of results on the dataset has been performed, analyzing each single experiment included in the global fit. Finally, has been provided evidence that individual data subsets can have runaway directions due to poorly determined PDFs, thus suggesting that a global dataset is necessary for a reliable determination.

Conclusions and Outlook

In this thesis the main problematics related to the determination of a set of parton distribution functions have been presented and, after a brief discussion on the strategies adopted by other groups to address them, the focus has been set on NNPDF methodology.

The fit performed on a global dataset using the combined implementation of Monte Carlo sampling in the space of data, neural network parametrization, and genetic algorithm minimization allows the NNPDF collaboration to deliver PDF sets which behave in a statistically consistent way and minimize parametrization bias.

Three PDF sets are presented and analyzed: NNPDF2.1 LO, NLO, and NNLO. Heavy quark mass effects are taken into account through the implementation of the FONLL-A GM-VFN scheme for the NNPDF2.1 NLO parton set and using FONLL-C for the NNPDF2.1 NNLO delivery. At LO heavy quark mass effects intervene to a much minor extent. The NLO PDF set has been compared to the previous NNPDF2.0 release, which does not include HQ mass effects, and also with parton sets from other groups as CT10 and MSTW08, which instead include them but through different treatments. The leading order determination has been compared both with the NLO one and with the LO PDF sets from other groups as MSTW and CTEQ. The NNPDF2.1 NNLO parton set, after the usual comparison to the NLO release, has been instead compared to MSTW08 NNLO and ABKM09 NNLO. In general, has been verified consistency at the one-sigma level among NNPDF releases (with the only exception of LO determination, where theoretical uncertainties are relevant) while rather significant shifts have been observed in specific cases in the comparison with results from other groups, especially in the NNLO analysis.

Several sets of PDFs have been determined with varying values of α_s and of the heavy quark masses m_c, m_b, to allow the computation of the respective uncertainties.

Theoretical predictions for a set of LHC observables (inclusive vector boson, top and Higgs production) have been computed for benchmarking against results from other groups and against first LHC available measurements. The change in this kind of inclusive cross-sections has been found to be rather small when including heavy quark mass effects. The change is instead substantial when looking at observables that directly probe heavy quark distributions.

Also, has been shown that an estimate of the α_s value is possible exploiting the simultaneous dependence of the global dataset on PDFs and α_s value. It has been determined $\alpha_s(M_Z) = 0.1191 \pm 0.0006^{\text{exp}}$, with a surprisingly low statistical uncertainty.

Having determined LO, NLO, and NNLO PDF sets using the very same methodology and the same data, allows for the assessment of perturbative stability. Excellent convergence of the perturbative expansion within the kinematic region covered by the experimental data has been shown. Moreover, has been shown that the provided uncertainty on PDFs at LO is only a fraction of the theoretical uncertainty (that is not included). Looking at the NLO fit it can be estimated as dominant in this case. The PDF sets produced by the NNPDF collaboration as well as the ones delivered by other groups only include data's experimental uncertainties. A similar estimate of theoretical uncertainty can be given also for the NLO determination by looking at the following perturbative order (NNLO): in this case the size of theoretical uncertainty is smaller than that of experimental one, thus at NLO and beyond it is reasonable to neglect theoretical uncertainty. The theoretical uncertainty could be evaluated in a more precise way for each single perturbative order determination by varying the renormalization and factorization scale during the PDF fit.

The inclusion of the amazing amount of new data that the LHC is delivering will be the main target in the next years. The NNPDF2.2 PDF set, that has been briefly presented in this thesis, already includes a part of LHC data through the reweighting technique. The NNPDF2.3 parton set [50] includes sistematically all the relevant LHC data. The presence of these data will be more and more important in parton fits, leading to much better constrained PDFs. Also, new processes will be available at the energies at which the LHC operates, hopefully allowing for a competitive PDF determination using collider data, for exclusion of the less cleaner fixed-target data and in general of low energy measurements.

Acknowledgments

I’m grateful to Professor José Ignacio Latorre for his patient work and for the fruitful discussions we had during my Ph.D. thesis. I’d like to thank Dr. Juan Rojo for introducing me to my advisor and for the help he gave me especially in the first thesis year. I’d like to thank all and each of the members of the NNPDF collaboration for the diligent restless effort in pursuing improvements and results.

Appendix

Appendix A

Heavy Quark Coefficient Functions to $\mathcal{O}(\alpha_s)$ in Mellin Space

Neutral Current Coefficient Functions

In this first part of Appendix A the analytic Mellin transform of the x-space $\mathcal{O}(\alpha_s)$ heavy quark neutral current coefficient functions is computed and is discussed the implementation and benchmarking of FONLL neutral current structure functions in the FastKernel framework. The corresponding results for charged current structure functions follow below.

The x-space gluon $\mathcal{O}(\alpha_s)$ heavy quark coefficient function is given by Eq. (1.130). Its Mellin transform is defined in the standard way as

$$C_{2,g}^{(n_l),1}\left(N, \frac{Q^2}{m_h^2}\right) = \int_0^{(1+4m_h^2/Q^2)^{-1}} dz z^{N-1} C_{2,g}^{(n_l),1}\left(z, \frac{Q^2}{m_h^2}\right) . \tag{A.1}$$

It is easy to see that the integral Eq. (A.1) can be written in the following way

$$\begin{aligned} C_{2,g}^{(n_l),1}(N,\epsilon) &= T_R a^N \int_0^1 dt\, t^{N-1} \Big\{ \big[1 + 2a(2\epsilon-1)t + 2a^2(1-6\epsilon-4\epsilon^2)t^2\big] \ln\frac{1+v}{1-v} \\ &\qquad - \big[1 + 4a(\epsilon-2)t - 4a^2(\epsilon-2)t^2\big] v \Big\} \qquad \text{(A.2)} \\ &= T_R a^N \int_0^1 dt\, t^{N-1} \Big\{ \big[1 + (1-3a)t - \tfrac{1}{2}(1+4a-9a^2)t^2\big] \ln\frac{1+v}{1-v} \\ &\qquad - \big[1 + (1-9a)t - a(1-9a)t^2\big] v \Big\}, \qquad \text{(A.3)} \end{aligned}$$

where $a(\epsilon) \equiv (1+4\epsilon)^{-1}$ is defined to simplify the coefficients. The needed integrals are

thus

$$J_1(N) \equiv \int_0^1 dt\, t^{N-1} \ln \frac{1+v}{1-v}\,, \qquad J_2(N) \equiv \int_0^1 dt\, t^{N-1} v, \tag{A.4}$$

since extra powers of t can be accommodated by a shift in N by an integer. Here as usual $v = (1-t)^{1/2}/(1-at)^{1/2}$.

The two integrals are related by an integration by parts. To show this, is needed

$$\begin{aligned} \frac{d}{dt} \ln \frac{1+v}{1-v} &= \frac{dv}{dt}\frac{d}{dv} \ln \frac{1+v}{1-v} \\ &= \Big(-\tfrac{1}{2} \frac{1-a}{(1-t)^{1/2}(1-at)^{3/2}} \Big) \Big(\frac{2(1-at)}{(1-a)t} \Big) \\ &= -\frac{1}{t} \frac{1}{(1-t)^{1/2}(1-at)^{3/2}}. \end{aligned} \tag{A.5}$$

Thus

$$J_1(N) = -\frac{1}{N} \int_0^1 dt\, t^N \frac{d}{dt} \ln \frac{1+v}{1-v} = \frac{1}{N} I(N), \tag{A.6}$$

where

$$I(N) \equiv \int_0^1 dt\, t^{N-1} (1-t)^{-1/2} (1-at)^{-1/2}. \tag{A.7}$$

Note that the boundary term in the integration by parts vanishes for all $\mathrm{Re}N > 0$, and thus its analytic continuation vanishes for all N, so it can be safely ignored. Trivially $J_2(N) = I(N) - I(N+1)$. The integral $I(N)$ may be evaluated in the usual way in terms of a standard hypergeometric function:

$$I(N) = \frac{\Gamma(N)\Gamma(\frac{1}{2})}{\Gamma\left(N+\frac{1}{2}\right)}\, {}_2F_1(\tfrac{1}{2}, N, N+\tfrac{1}{2}; a). \tag{A.8}$$

Note that when $a = 0$, this reduces to $B(N, \frac{1}{2})$, as it should.

Thus it is obtained

$$\begin{aligned} &C_{2,g}^{(n_l),1}(N,\epsilon) = T_R a^N \Big\{ \big[\tfrac{1}{N} I(N) + \tfrac{1-3a}{N+1} I(N+1) - \tfrac{1}{2} \tfrac{1+6a-9a^2}{N+2} I(N+2) \big] \\ &- \big[I(N) - I(N+1) + (1-9a)(I(N+1) - I(N+2)) \\ &- a(1-9a)(I(N+2) - I(N+3)) \big] \Big\} \\ &= T_R a^N \Big\{ (\tfrac{1}{N} - 1) I(N) + (\tfrac{1-3a}{N+1} + 9a) I(N+1) \\ &- (\tfrac{1}{2} \tfrac{1+4a-9a^2}{N+2} - (1+a)(1-9a)) I(N+2) - a(1-9a) I(N+3) \Big\}\,. \end{aligned} \tag{A.9}$$

This result is the required ingredient to implement the FONLL-A neutral current structure functions in the FastKernel framework.

A cross-check of the Mellin transform of the massive coefficient function Eq. (A.9)

is provided by the fact that its massless limit coincides with the Mellin transform of the x-space massive asymptotic $(n_l, 0)$ coefficient function, Eq. (1.134). To this purposes, Eq. (A.9) needs to be expanded near $a = 1$. Near $a = 1$, i.e. $\epsilon = 0$, the asymptotic expansion

$$F(\tfrac{1}{2}, N, N+\tfrac{1}{2}; a) = \frac{\Gamma(N+\frac{1}{2})}{\Gamma(\frac{1}{2})^2\Gamma(N)^2} \sum_{n=0}^{\infty} \frac{\Gamma(n+\frac{1}{2})\Gamma(N+n)}{(n!)^2} \times [2\psi(n+1) - \psi(n+\tfrac{1}{2}) - \psi(N+n) - \ln(1-a)](1-a)^n, \quad \text{(A.10)}$$

is needed, so that

$$I(N) = \frac{1}{\Gamma(\frac{1}{2})\Gamma(N)} \sum_{n=0}^{\infty} \frac{\Gamma(n+\frac{1}{2})\Gamma(N+n)}{(n!)^2} \times [2\psi(n+1) - \psi(n+\tfrac{1}{2}) - \psi(N+n) - \ln(1-a)](1-a)^n. \quad \text{(A.11)}$$

The $n = 0$ term then gives the $\ln \epsilon$ collinear divergence, which is subtracted by the massless coefficient function: as $\epsilon \to 0$

$$I(N) = -\ln(4\epsilon) - 2\gamma_E - \psi(\tfrac{1}{2}) - \psi(N) + O(\epsilon). \quad \text{(A.12)}$$

Substituting in Eq. (A.9) the result

$$C_g^{(n_l),1}(N,\epsilon) = T_R \frac{[-N^3 + 3N^2 - N(2+N+N^2)(\ln\epsilon + \gamma_E + \psi(N))]}{N^2(1+N)(2+N)} + \mathcal{O}(\epsilon), \quad \text{(A.13)}$$

is obtained, as expected: the coefficient of the singularity is precisely the LO anomalous dimension $\gamma_{qg}^{(0)}(N)$. Therefore it has been checked that the massless limit is properly reproduced,

$$C_g^{(n_l),1}(N,\epsilon) + \mathcal{O}(\epsilon) = C_g^{(n_l,0),1}(N,\epsilon), \quad \text{(A.14)}$$

with the massive asymptotic coefficient function given by the Mellin transform of Eq. (1.134), as desired.

For completeness, the correponding expressions for the $\mathcal{O}(\alpha_s)$ heavy quark coefficient function for the longitudinal structure function $F_{L,c}$ is also provided, which is implicitly contained in Eq. (A.9) since $F_{2,c} = F_{T,c} + F_{L,c}$. The x-space expression for the longitudinal heavy quark coefficient function is

$$C_{L,g}^{(n_l),1}\left(z, \frac{Q^2}{m^2}\right) = \theta\left(W^2 - 4m^2\right) \times T_R \left[-8\epsilon z^2 \log\frac{1+v}{1-v} + 4vz(1-z)\right]. \quad \text{(A.15)}$$

Its Mellin transform can be computed using the integrals discussed above, with the result

$$C_{L,g}^{(n_l),1}\left(N,\frac{Q^2}{m^2}\right) = T_R a^{N+1}\Big[-8\epsilon a\frac{I(N+2)}{N+2}+4\left(I(N+1)-I(N+2)(1+a)+I(N+2)\right)\Big] .$$

The massless limits of the x- and N-space results are straightforwardly computed and checked to be related by Mellin transformation as they ought to.

Now the implementation and benchmarking of these results in the FastKernel framework is considered. The major improvement in the FastKernel framework as compared to Ref. [87] is the inclusion of heavy quark mass effects in deep-inelastic scattering structure functions, following the FONLL-A general-mass scheme [36]. As discussed in Ref. [87], FastKernel requires to write down all the DIS observables in Mellin space and precomputing all the associated x-space Green's functions. Therefore, to extend FastKernel with FONLL structure functions a formulation of FONLL in Mellin space is needed.

The x-space expression for the FONLL-A heavy quark structure functions, Eq. (1.127), can be easily written down in N-space as follows:

$$\begin{aligned} F_{2,h}^{\rm FONLL}(N,Q^2) &= F_{2,h}^{(n_l)}(N,Q^2) \\ &- \theta\left(Q^2-m^2\right)\left(1-\frac{m^2}{Q^2}\right)^2\left[F_{2,h}^{(n_l,0)}(N,Q^2)-F_{2,h}^{(n_l+1)}(N,Q^2)\right] , \end{aligned} \tag{A.16}$$

with the default damping factor as threshold prescription. In order to implement Eq. (A.16) in the FastKernel framework, the Mellin space expressions of the heavy quark coefficient function in the (n_l), $(n_l,0)$ and (n_l+1) schemes are needed. While the last two are known, the former was not available in a closed form suitable for analytical continuation. The details of the computation have been presented above, and the desired result is Eq. (A.9).

With all the Mellin space heavy quark coefficient functions available, it becomes possible to implement the FONLL-A heavy quark structure functions, Eq. (A.16) into the FastKernel framework. To show that the N-space implementation has the required accuracy, in Table A.1 for $F_{2,c}$ and in Table A.2 for $F_{L,c}$, the results for the Les Houches Heavy Quark benchmarks [38] for FONLL-A obtained with the FONLLdis code [36, 51] and with the FastKernel framework for various relevant values of Q^2 are compared. The benchmark settings for the PDFs and α_s are used for this comparison. What can be seen is that the accuracy on the FONLL heavy quark structure functions is essentially always below the percent level, enough for precision phenomenological studies. For completeness the analogous results for the case of the massive scheme results, where similar accuracies are obtained, are also shown.

	FONLL-A			FFN		
x	FONLLdis	FastKernel	Accuracy	FONLLdis	FastKernel	Accuracy
$Q^2 = 4$ GeV^2						
10^{-5}	0.1507	0.1501	0.4%	0.1088	0.1091	0.3%
10^{-4}	0.0936	0.0931	0.5%	0.0697	0.0698	0.1%
10^{-3}	0.0506	0.0504	0.4%	0.0392	0.391	0.2%
10^{-2}	0.0174	0.0177	1.5%	0.0136	0.0137	0.7%
$Q^2 = 10$ GeV^2						
10^{-5}	0.563	0.561	0.4%	0.3598	0.3602	0.1%
10^{-4}	0.312	0.311	0.3%	0.2007	0.2011	0.2%
10^{-3}	0.1499	0.1495	0.3%	0.0981	0.0982	0.1%
10^{-2}	0.05056	0.05052	0.1%	0.0328	0.0327	0.3%
$Q^2 = 100$ GeV^2						
10^{-5}	2.28636	2.28577	0.02%	1.9779	1.9877	0.5%
10^{-4}	1.12186	1.12082	0.1%	0.9161	0.9184	0.3%
10^{-3}	0.48008	0.47919	0.2%	0.3644	0.3647	0.1%
10^{-2}	0.15207	0.15200	0.04%	0.1037	0.1038	0.1%

Table A.1: Results of the benchmark comparison for the $F_{2c}(x, Q^2)$ structure function in the FONLL-A scheme for the FONLLdis code [36] and for the FastKernel code. Results are provided at the benchmark kinematical points in x, Q^2. Results for the massive (FFN) scheme are also given for completeness.

	FONLL-A			FFN		
x	FONLLdis	FastKernel	Accuracy	FONLLdis	FastKernel	Accuracy
$Q^2 = 4$ GeV^2						
10^{-5}	0.0130174	0.013094	0.6%	0.009077	0.009081	0.04%
10^{-4}	0.008347	0.008316	0.4%	0.005913	0.005910	0.05%
10^{-3}	0.004795	0.004778	0.3%	0.003511	0.003509	0.06%
10^{-2}	0.001910	0.001907	0.2%	0.001403	0.001406	0.2%
$Q^2 = 10$ GeV^2						
10^{-5}	0.073235	0.073022	0.3%	0.049856	0.049982	0.2%
10^{-4}	0.041392	0.041251	0.3%	0.028402	0.028423	0.07%
10^{-3}	0.020754	0.020707	0.2%	0.014463	0.014456	0.05%
10^{-2}	0.007616	0.007595	0.3%	0.005350	0.005346	0.07%
$Q^2 = 100$ GeV^2						
10^{-5}	0.471889	0.4729	0.2%	0.3955	0.397855	0.6%
10^{-4}	0.2236	0.2235	0.1%	0.18656	0.186914	0.2%
10^{-3}	0.0920	0.09188	0.1%	0.0765	0.076393	0.1%
10^{-2}	0.027822	0.02782	0.1%	0.023079	0.023100	0.1%

Table A.2: Same as Table A.1 for the $F_{Lc}(x, Q^2)$ structure function.

Charged Current Coefficient Functions

In this second part of Appendix A the analysis of the previous part is repeated for charged current structure functions.

The F_2^c charm structure functions in charged current DIS is given by Eq. (1.135). The expression for structure functions in neutrino-induced charged current scattering in the FFN scheme is

$$\mathcal{F}_i^c(x,Q^2) = \frac{1}{2}s'(\xi,\mu^2) + \frac{1}{2}\frac{\alpha_s(\mu^2)}{2\pi}\Bigg\{\int\limits_{\xi}^{1}\frac{d\xi'}{\xi'}\Bigg[C_{i,q}^{(n_l)}(\xi',\mu^2,\lambda)s'\left(\frac{\xi}{\xi'},\mu^2\right) + \\ + C_{i,g}^{(n_l)}(\xi',\mu^2,\lambda)g\left(\frac{\xi}{\xi'},\mu^2\right)\Bigg]\Bigg\}, \quad \text{(A.17)}$$

with $i = 1,2,3$. In Eq. (A.17) have used the following definitions:

$$s' = 2|V_{cs}|^2 s + 2|V_{cd}|^2[f\,d + (1-f)u]; \qquad f = \frac{N_p}{N_p+N_n}; \\ \xi = x\left(1+\frac{m_c^2}{Q^2}\right); \qquad \lambda = \frac{Q^2}{Q^2+m_c^2}\,. \quad \text{(A.18)}$$

The explicit x-space expressions of the $\mathcal{O}(\alpha_s)$ contributions $C_{i,q(g)}^{(n_l)}$ to the coefficient functions are given in Refs. [44, 45]. The standard structure functions are related to those defined in Eq. (A.17) through

$$F_1^c \equiv \mathcal{F}_1^c; \qquad F_2^c \equiv 2\xi\mathcal{F}_2^c = x\frac{2}{\lambda}\mathcal{F}_2^c; \qquad F_3^c \equiv 2\mathcal{F}_3^c\,, \quad \text{(A.19)}$$

so that

$$F_L^c \equiv F_2^c - 2xF_1^c = 2\xi\left(\mathcal{F}_2^c - \lambda\mathcal{F}_1^c\right). \quad \text{(A.20)}$$

Before Mellin- transforming the x-space quark coefficient functions of Refs. [44, 45] they are all rewritten in the form

$$C_{i,q}^{(n_l)}(x) = K\delta(1-x) + f(x) + [g(x)]_+,$$

where K is a constant and $f(x)$ is regular function in $x \in [0,1]$ (so in general $g(x)$ is not

regular in $x = 1$). It gives

$$
\begin{aligned}
C_{1,q}^{(n_l)}(z) &= -C_F\left(4+\frac{1}{2\lambda}+\frac{\pi^2}{3}+\frac{1+3\lambda}{2\lambda}K_A\right)\delta(1-z) \\
&+ C_F\left[-\frac{(1+z^2)\ln z}{1-z}-2(1+z)\ln(1-z)+(1+z)\ln(1-\lambda z)+(3-z)+\frac{z-z^2}{1-\lambda z}\right] \\
&+ C_F\left[4\frac{\ln(1-z)}{1-z}-2\frac{\ln(1-\lambda z)}{1-z}-2\frac{1}{1-z}+\frac{1}{2}\frac{1-z}{(1-\lambda z)^2}-\frac{1+z^2}{1-z}\ln\lambda\right]_+ ; \qquad \text{(A.21)} \\
C_{2,q}^{(n_l)}(z) &= -C_F\left(4+\frac{1}{2\lambda}+\frac{\pi^2}{3}+\frac{1+\lambda}{2\lambda}K_A\right)\delta(1-z) \\
&+ C_F\left[-\frac{(1+z^2)\ln z}{1-z}-2(1+z)\ln(1-z)+(1+z)\ln(1-\lambda z)\right. \\
&+ \left.\left(2z+2-\tfrac{2}{z}\right)+\tfrac{\frac{2}{z}-1-z}{1-\lambda z}\right] \\
&+ C_F\left[4\frac{\ln(1-z)}{1-z}-2\frac{\ln(1-\lambda z)}{1-z}-2\frac{1}{1-z}+\frac{1}{2}\frac{1-z}{(1-\lambda z)^2}-\frac{1+z^2}{1-z}\ln\lambda\right]_+ ; \qquad \text{(A.22)} \\
C_{3,q}^{(n_l)}(z) &= -C_F\left(4+\frac{1}{2\lambda}+\frac{\pi^2}{3}+\frac{1+3\lambda}{2\lambda}K_A\right)\delta(1-z) \\
&+ C_F\left[-\frac{(1+z^2)\ln z}{1-z}-2(1+z)\ln(1-z)+(1+z)\ln(1-\lambda z)+(1+z)+\frac{1-z}{1-\lambda z}\right] \\
&+ C_F\left[4\frac{\ln(1-z)}{1-z}-2\frac{\ln(1-\lambda z)}{1-z}-2\frac{1}{1-z}+\frac{1}{2}\frac{1-z}{(1-\lambda z)^2}-\frac{1+z^2}{1-z}\ln\lambda\right]_+ ; \qquad \text{(A.23)}
\end{aligned}
$$

with $K_A = (1-\lambda)\ln(1-\lambda)/\lambda$.

The gluon coefficient functions do not need any further work and are given by

$$
\begin{aligned}
C_{1,g}^{(n_l)}(z) &= T_f(2z^2-2z+1)\left\{2\ln(1-z)-2\ln z-\ln[\lambda(1-\lambda)]\right\}+ \\
&\quad [4-4(1-\lambda)]z(1-z)+(1-\lambda)\frac{z}{1-\lambda z}+ \qquad \text{(A.24)} \\
&\quad 2(1-\lambda)\left[z\ln\frac{1-\lambda z}{(1-\lambda)z}-2\lambda z^2\ln\frac{1-\lambda z}{(1-\lambda)z}\right]-1;
\end{aligned}
$$

$$
\begin{aligned}
C_{2,g}^{(n_l)}(z) &= T_f(2z^2-2z+1)\left\{2\ln(1-z)-2\ln z-\ln[\lambda(1-\lambda)]\right\}+ \\
&\quad [8-18(1-\lambda)+12(1-\lambda)^2]z(1-z)+(1-\lambda)\frac{1}{1-\lambda z}+ \\
&\quad 6\lambda(1-\lambda)\left[z\ln\frac{1-\lambda z}{(1-\lambda)z}-2\lambda z^2\ln\frac{1-\lambda z}{(1-\lambda)z}\right]-1;
\end{aligned}
\tag{A.25}
$$

$$
\begin{aligned}
C_{3,g}^{(n_l)}(z) &= T_f(2z^2-2z+1)\left\{2\ln(1-z)-2\ln(1-\lambda z)+\ln\left(\frac{1-\lambda}{\lambda}\right)\right\}+ \\
&\quad 2(1-\lambda)z(1-z)+2(1-\lambda)\left[(1+\lambda)z^2\ln\frac{1-\lambda z}{(1-\lambda)z}-z\ln\frac{1-\lambda z}{(1-\lambda)z}\right].
\end{aligned}
\tag{A.26}
$$

In order to transform to the N-space the above x-space expressions, in Tables A.3 and A.4 are tabulated the Mellin transforms of all terms involved. In these tables the analytic continuation of the harmonic sum

$$
S_l \equiv S_l(N) = \sum_{k=1}^{N}\frac{1}{k^l} = \zeta(l) - \frac{(-1)^l}{(l-1)!}\psi^{(l-1)}(N+1),
$$

is used, where $\zeta(l)$ is the Riemann ζ-function, with $\zeta(1) = \gamma_{EM}$, $\psi((l-1))$ is the polygamma, and ${}_2F_1(a,b,c;N)$ is the Gauss hypergeometric function .

As an example of use of Tables A.3-A.4, here is presented the complete N-space quark and gluon coefficient functions for F_2^c

$$
\begin{aligned}
C_{2,q}^{(n_l)}(N) &= C_F\Bigg[-\left(4+\frac{1}{2\lambda}+\frac{\pi^2}{3}+\frac{1+\lambda}{2\lambda}K_A\right)-2(S_2-\zeta_2)+\frac{1}{N^2} \\
&- \frac{1}{(N+1)^2}+2\left(\frac{S_1}{N}+\frac{S_1}{N+1}+\frac{1}{(N+1)^2}\right)+\lambda\frac{{}_2F_1(1,N+1,N+2;\lambda)}{N(N+1)} \\
&+ \frac{\ln(1-\lambda)}{N}+\lambda\frac{{}_2F_1(1,N+2,N+3;\lambda)}{(N+1)(N+2)}+\frac{\ln(1-\lambda)}{N+1}+\frac{2}{N+1}+\frac{2}{N} \\
&- \frac{2}{N-1}+2\frac{{}_2F_1(1,N-1,N,\lambda)}{N-1}-\frac{{}_2F_1(1,N,N+1,\lambda)}{N} \\
&- \frac{{}_2F_1(1,N+1,N+2,\lambda)}{N+1}+2\left(S_1^2+S_2-2\frac{S_1}{N}\right)-2J_\lambda(N)-2\left(\frac{1}{N}-S_1\right) \\
&+ \frac{1}{2}\left(\frac{{}_2F_1(2,N,N+2,\lambda)}{N(N+1)}+\frac{\lambda+\ln(1-\lambda)}{\lambda^2}\right)-\left(\frac{1}{N}-\frac{1}{N+1}-2S_1+\frac{3}{2}\right)\ln\lambda\Bigg];
\end{aligned}
\tag{A.27}
$$

$f(z)$	$\mathbf{M}[f](N)$
$\delta(1-z)$	1
z^l	$\dfrac{1}{N+l}$
$\dfrac{(1+z^2)\ln z}{1-z}$	$2(S_2-\zeta_2)-\dfrac{1}{N^2}+\dfrac{1}{(N+1)^2}$
$(1+z)\ln(1-z)$	$-\dfrac{S_1}{N}-\dfrac{S_1}{N+1}-\dfrac{1}{(N+1)^2}$
$(1+z)\ln(1-\lambda z)$	$\lambda\dfrac{{}_2F_1(1,N+1,N+2;\lambda)}{N(N+1)}+\dfrac{\ln(1-\lambda)}{N}+$ $\lambda\dfrac{{}_2F_1(1,N+2,N+3;\lambda)}{(N+1)(N+2)}+\dfrac{\ln(1-\lambda)}{N+1}$
$\dfrac{z-z^2}{1-\lambda z}$	$\dfrac{{}_2F_1(1,N+1,N+2,\lambda)}{N+1}-\dfrac{{}_2F_1(1,N+2,N+3,\lambda)}{N+2}$
$\dfrac{\frac{2}{z}-1-z}{1-\lambda z}$	$2\dfrac{{}_2F_1(1,N-1,N,\lambda)}{N-1}-\dfrac{{}_2F_1(1,N,N+1,\lambda)}{N}$ $-\dfrac{{}_2F_1(1,N+1,N+2,\lambda)}{N+1}$
$\dfrac{1-z}{1-\lambda z}$	$\dfrac{{}_2F_1(1,N,N+1,\lambda)}{N}-\dfrac{{}_2F_1(1,N+1,N+2,\lambda)}{N+1}$
$\left[\dfrac{\ln(1-z)}{1-z}\right]_+$	$\dfrac{1}{2}\left(S_1^2+S_2-2\dfrac{S_1}{N}\right)$
$\left[\dfrac{\ln(1-\lambda z)}{1-z}\right]_+$	$J_\lambda(N)=\displaystyle\sum_{k=1}^{\infty}\dfrac{\lambda^k}{k}\left[S_1(N+k)-S_1(k)-\dfrac{1}{N+k}\right]$
$\left[\dfrac{1}{1-z}\right]_+$	$\dfrac{1}{N}-S_1$
$\left[\dfrac{1-z}{(1-\lambda z)^2}\right]_+$	$\dfrac{{}_2F_1(2,N,N+2,\lambda)}{N(N+1)}+\dfrac{\lambda+\ln(1-\lambda)}{\lambda^2}$
$\left[\dfrac{1+z^2}{1-z}\right]_+$	$\dfrac{1}{N}-\dfrac{1}{N+1}-2S_1+\dfrac{3}{2}$

Table A.3: Mellin transforms of the terms involved in the NLO charged current quark coefficient functions.

$f(z)$	$\mathbf{M}[f](N)$
$[z^2+(1-z)^2]\times$ $\left\{2\ln\left(\frac{1-z}{z}\right)-\ln[\lambda(1-\lambda)]\right\}$	$\frac{4-2N(N-3)-N(N^2+N+2)\{2S_1+\ln[\lambda(1-\lambda)]\}}{N^2(N+1)(N+2)}$
$[z^2+(1-z)^2)]\times$ $\left\{2\ln\left(\frac{1-z}{1-\lambda z}\right)+\ln\left(\frac{1-\lambda}{\lambda}\right)\right\}$	$-\frac{2}{\lambda}\left(\frac{\lambda^2}{N}-\frac{2\lambda}{N+1}+\frac{2}{N+2}\right)\frac{{}_2F_1(1,N+1,N+2,\lambda)}{N+1}$ $-\frac{4(\lambda-1)}{\lambda(N+1)(N+2)}-\frac{(N^2+N+2)\left\{2S_1-\ln\left(\frac{1-\lambda}{\lambda}\right)\right\}}{N(N+1)(N+2)}$
$z\ln\frac{1-\lambda z}{(1-\lambda)z}$	$\frac{{}_2F_1(1,N+1,N+2;\lambda)}{(N+1)^2}$
$z^2\ln\frac{1-\lambda z}{(1-\lambda)z}$	$\frac{{}_2F_1(1,N+1,N+2;\lambda)-1}{\lambda(N+1)(N+2)}$
$\frac{1}{1-\lambda z}$	$\frac{{}_2F_1(1,N,N+1;\lambda)}{N}$
$\frac{z}{1-\lambda z}$	$\frac{{}_2F_1(1,N+1,N+2;\lambda)}{N+1}$

Table A.4: Mellin transforms of the terms involved in the NLO charged current gluon coefficient functions.

$$
\begin{aligned}
C_{2,g}^{(n_l)}(N) &= T_f\frac{4-2N(N-3)-N(N^2+N+2)\{S_1+\ln[\lambda(1-\lambda)]\}}{2N^2(N+1)(N+2)}+\\
&\frac{8-18(1-\lambda)+12(1-\lambda)^2}{(N+1)(N+2)}+\frac{(1-\lambda){}_2F_1(1,N,N+1;\lambda)}{N}+ \qquad \text{(A.28)}\\
&6\lambda(1-\lambda)\left[\frac{{}_2F_1(1,N+1,N+2;\lambda)}{(N+1)^2}-\frac{2{}_2F_1(1,N+1,N+2;\lambda)-1}{(N+1)(N+2)}\right]-\frac{1}{N}.
\end{aligned}
$$

As a cross-check of the Mellin space results, it is possible to compute the asymptotic limit $\lambda\to 1$ of these expressions. The asymptotic expansion of the hypergeometric functions is needed, Eq. (A.10) up to $\mathcal{O}(\lambda-1)$ terms. In particular,

$$
{}_2F_1(1,N+1,N+2;\lambda)=-(1+N)\left(\ln(1-\lambda)+\gamma_E+\psi^{(0)}(N+1)\right)+\mathcal{O}((\lambda-1)), \qquad \text{(A.29)}
$$

$$
{}_2F_1(2,N,N+2;\lambda)=-N(1+N)\left(\ln(1-\lambda)+\gamma_E+\psi^{(0)}(N)\right)+\mathcal{O}((\lambda-1)). \qquad \text{(A.30)}
$$

Substituting in Eq. (A.27), one can see that all collinear heavy quark logarithms and that the massless limit of the massive charged current heavy quark coefficient functions reduces to the usual ZM-VFN result, as known from x-space.

Now the implementation and benchmarking of the above results into the FastKernel framework is discussed. Analogously to the neutral current sector, the FONLL-A charged current structure functions in Mellin space can be written as

$$\begin{aligned} F_{i,h}^{\mathrm{CC,FONLL}}(N,Q^2) &= F_{i,h}^{CC(n_l)}(N,Q^2) \\ &- \theta\left(Q^2-m^2\right)\left(1-\frac{m^2}{Q^2}\right)^2\left[F_{i,h}^{CC(n_l,0)}(N,Q^2)-F_{2,h}^{CC(n_l+1)}(N,Q^2)\right] . \end{aligned} \tag{A.31}$$

with $i = 1, 2, 3$. The Mellin space expressions of the massive heavy quark coefficient functions have been computed above, and the other ingredients of Eq. (A.31) are their massless limits and the standard Mellin transform of the ZM-VFN coefficient functions.

With these results, the FONLL-A charged current structure functions Eq. (A.31) have been implemented into the FastKernel framework. As it has been done in for neutral currents, here the accuracy of this FONLL scheme implementation is benchmarked. Again the same settings of the Les Houches heavy quarks benchmark study are used. The benchmarking of the FONLL-A CC structure function implementation in FastKernel is performed for the charm production cross-section in neutrino induced DIS, defined by Eq. (1.159), that combines all three charged current structure functions. It has been checked that the comparison of individual structure functions has a similar level of accuracy.

Results for the benchmark comparison are shown in Table A.5. As discussed above, the FONLL-A calculation of charged current structure functions has been implemented in a x-space code, FONLLdisCC, that will be used for the benchmarking with the FastKernel implementation. Results are shown for various values of Q^2 relevant for the analysis of experimental data. The accuracy is similar to the one achieved for neutral current structure functions (see Tables A.1-A.2), at the per mil level, suitable for precision PDF determinations.

	FONLL-A			FFN		
x	FONLLdisCC	FastKernel	Accuracy	FONLLdisCC	FastKernel	Accuracy
$Q^2 = 4$ GeV^2						
10^{-5}	163.14	164.06	0.6%	158.70	158.15	0.3%
10^{-4}	109.48	109.55	0.1%	106.81	106.64	0.2%
10^{-3}	69.24	69.35	0.2%	67.86	67.88	0.1%
10^{-2}	37.75	37.87	0.3%	37.27	37.30	0.1%
10^{-1}	13.56	13.57	0.1%	13.53	13.51	0.1%
$Q^2 = 10$ GeV^2						
10^{-5}	279.31	278.71	0.2%	261.49	261.55	0.02%
10^{-4}	167.02	166.85	0.1%	157.27	157.11	0.1%
10^{-3}	92.90	92.87	0.03%	88.33	88.12	0.2%
10^{-2}	44.92	44.93	0.02%	43.36	43.23	0.3%
10^{-1}	14.50	14.48	0.1%	14.26	14.28	0.1%
$Q^2 = 100$ GeV^2						
10^{-5}	674.55	674.53	0.02%	651.21	645.94	0.1%
10^{-4}	345.73	345.81	0.02%	331.17	329.14	0.5%
10^{-3}	161.70	161.78	0.05%	153.94	152.36	0.1%
10^{-2}	64.20	64.26	0.1%	61.11	61.06	0.1%
10^{-1}	15.79	15.83	0.2%	15.33	15.42	0.1%

Table A.5: Results of the benchmark comparison for the dimuon charm production cross-section Eq. (1.159), in the FONLL-A scheme for the FONLLdisCC charged current code and for the FastKernel framework. Results are provided at the benchmark kinematical points in x, Q^2. Results for the massive (FFN) scheme are also given for completeness. The inelasticity variable in the dimuon cross-section for this benchmark table has been taken to be $y = 0.5$. The Les Houches Heavy Quark benchmark settings [38] have been used for the comparison.

Appendix B

Heavy Quark Coefficient Functions to $\mathcal{O}\left(\alpha_s^2\right)$ in Mellin Space

In this Appendix expressions for the Mellin transforms of the $\mathcal{O}\left(\alpha_s^2\right)$ massive heavy quark coefficient functions in the $Q^2 \to \infty$ limit are provided. These asymptotic coefficient functions were first computed long in Ref. [237] in x-space. Their Mellin transforms have been given in Refs. [238, 239], and will be rederived here in a form which is suitable for the purposes of the present analysis.

In order to perform the Mellin transform of the x-space FFN heavy quark coefficient functions of Ref. [237] it is convenient to rewrite them in terms of independent Mellin integrals, which can be then tabulated.

Following the notation introduced in Ref. [237], first the gluon coefficient function for F_L is considered. The corresponding Mellin transform can be written as

$$H_{L,g}^{(2)}\left(N, \frac{Q^2}{m^2}, \frac{\mu^2}{m^2}\right) = 4T_f \left[C_{L,g}^{\rm const}(N) + C_{L,g}^{Q}(N) \ln\frac{Q^2}{m^2} - C_{L,g}^{\mu}(N) \ln\frac{\mu^2}{m^2}\right]. \quad \text{(B.1)}$$

The coefficient function has been separated into three terms: a Q^2-independent term $C^{\rm const}(N)$, a collinear log term $C^{Q}(N)$, and a scale variation term $C^{\mu}(N)$. A similar decomposition will be performed for all coefficient functions.

The individual terms are:

$$\begin{aligned} C_{L,g}^{\rm const} &= C_F \;\Big[\tfrac{16}{15}A_{10}^{(-2)} - \tfrac{16}{3}A_{10}^{(1)} + \tfrac{32}{5}A_{10}^{(3)} + 8A_8^{(1)} + 8A_6^{(1)} - \tfrac{16}{3}A_3^{(1)} - \tfrac{16}{5}A_3^{(3)} \\ &\quad -\zeta(2)\big(\tfrac{16}{3}A_1^{(1)} - \tfrac{32}{5}A_1^{(3)}\big) + 4A_4^{(0)} + 12A_4^{(1)} - 16A_4^{(2)} - \tfrac{16}{15}A_2^{(-1)} \\ &\quad -\tfrac{52}{15}A_2^{(0)} - \tfrac{104}{5}A_2^{(1)} + \tfrac{48}{5}A_2^{(2)} + \tfrac{16}{15}A_1^{(-1)} - \tfrac{64}{15}A_1^{(0)} - \tfrac{152}{5}A_1^{(1)} + \tfrac{168}{5}A_1^{(2)}\Big] \\ &\quad +C_A \;\Big[16A_{10}^{(1)} + 16A_{10}^{(2)} - 32A_8^{(1)} + 16\zeta(2)A_1^{(2)} - 48A_6^{(1)} + 16A_6^{(2)} \\ &\quad +8A_5^{(1)} - 8A_5^{(2)} + 24A_3^{(1)} + \tfrac{8}{3}A_4^{(-1)} - 8A_4^{(0)} - 72A_4^{(1)} + \tfrac{232}{3}A_4^{(2)} \\ &\quad +8A_2^{(0)} + 64A_2^{(1)} - 104A_2^{(2)} - \tfrac{8}{9}A_1^{(-1)} + \tfrac{8}{3}A_1^{(0)} + \tfrac{136}{3}A_1^{(1)} - \tfrac{424}{9}A_1^{(2)}\Big], \end{aligned} \tag{B.2a}$$

$$\begin{aligned} C_{L,g}^{Q} &= C_F[8A_2^{(1)} + 4A_1^{(0)} + 4A_1^{(1)} - 8A_1^{(2)}] \\ &\quad +C_A\big[16A_4^{(1)} - 16A_4^{(2)} - 32A_2^{(1)} + \tfrac{8}{3}A_1^{(-1)} - 8A_1^{(0)} - 40A_1^{(1)} + \tfrac{136}{3}A_1^{(2)}\big], \end{aligned} \tag{B.2b}$$

$$C_{L,g}^{\mu} = C_A\big[16A_4^{(1)} - 16A_4^{(2)} - 32A_2^{(1)} + \tfrac{8}{3}A_1^{(-1)} - 8A_1^{(0)} - 40A_1^{(1)} + \tfrac{136}{3}A_1^{(2)}\big], \tag{B.2c}$$

where the $A_N^{(l)}$ are the independent Mellin integrals

$$A_n^{(l)} \equiv A_n(N+l). \tag{B.3}$$

Those for which closed-form analytic expressions will be used are collected in Tables B.1 and B.2. Some of these Mellin transforms were already computed in Ref. [240] and are repeated here for completeness. The remainder, for which numerical approximations will be used, are evaluated at the end of this Appendix.

The quark coefficient function for F_L can be similarly written as

$$H_{L,q}^{(2)}\left(N, \frac{Q^2}{m^2}, \frac{\mu^2}{m^2}\right) = 4T_f\left[C_{L,q}^{\rm const}(N) + C_{L,q}^{Q}(N)\ln\frac{Q^2}{m^2} - C_{L,q}^{\mu}(N)\ln\frac{\mu^2}{m^2}\right], \tag{B.4}$$

where

$$\begin{aligned} C_{L,q}^{\rm const} &= C_F \;\big[-8A_8^{(1)} - 8A_6^{(1)} + 8A_3^{(1)} + \tfrac{8}{3}A_4^{(-1)} - 8A_4^{(0)} + \tfrac{16}{3}A_4^{(2)} + 8A_2^{(0)} \\ &\quad -8A_2^{(1)} - 16A_2^{(2)} - \tfrac{8}{9}A_1^{(-1)} + \tfrac{8}{3}A_1^{(0)} - \tfrac{32}{3}A_1^{(1)} + \tfrac{80}{9}A_1^{(2)}\big], \end{aligned} \tag{B.5a}$$

$$C_{L,q}^{Q} = C_{L,q}^{\mu} = C_F\big[-8A_2^{(1)} + \tfrac{8}{3}A_1^{(-1)} - 8A_1^{(0)} + \tfrac{16}{3}A_1^{(2)}\big]. \tag{B.5b}$$

Finally, the gluon radiation coefficient function for F_L can be written as

$$H_{L,GR}^{(2)}\left(N, \frac{Q^2}{m^2}, \frac{\mu^2}{m^2}\right) = 4T_f\left[C_{L,GR}^{\rm const}(N) + C_{L,GR}^{Q}(N)\ln\frac{Q^2}{m^2}\right], \tag{B.6}$$

where

$$C_{L,GR}^{\rm const} = C_F\big[\tfrac{4}{3}(A_4^{(1)} - 2A_2^{(1)} + A_1^{(0)}) - \tfrac{50}{9}A_1^{(1)}\big], \tag{B.7a}$$

$$C_{L,GR}^{Q} = C_F\tfrac{4}{3}A_1^{(1)}. \tag{B.7b}$$

Let us now turn to the F_2 heavy quark coefficient functions. In comparison to the longitudinal structure function, there are extra pieces $C^{2Q}(N)$ and $C^{\mu Q}(N)$ arising from the double collinear logarithm. The gluon coefficient function for F_2 can be written as

$$H_{2,g}^{(2)}\left(N,\frac{Q^2}{m^2},\frac{\mu^2}{m^2}\right) = 4T_f\Bigg[C_{2,g}^{\rm const}(N) + C_{2,g}^{2Q}(N)\ln^2\frac{Q^2}{m^2} + C_{2,g}^{Q}(N)\ln\frac{Q^2}{m^2} - C_{2,g}^{\mu}(N)\ln\frac{\mu^2}{m^2} - C_{2,g}^{\mu Q}(N)\ln\frac{\mu^2}{m^2}\ln\frac{Q^2}{m^2}\Bigg], \tag{B.8}$$

where

$$\begin{aligned}
C_{2,g}^{\rm const} &= C_F \;\Big\{4\Big[-2B_{16}^{+}-2\zeta(2)B_{17}^{+}+B_{21}^{+}-\tfrac{3}{2}B_{8}^{+}\Big]\\
&\quad+2\Big[-4B_{15}^{-}-2B_{22}^{-}+12B_{19}^{-}+B_{18}^{-}-\tfrac{5}{2}B_{14}^{-}+2B_{12}^{-}+B_{13}^{-}\\
&\quad-4B_{20}^{-}+4\zeta(2)B_{2}^{-}-\tfrac{1}{3}B_{11}^{-}+3B_{6}^{-}\Big]+2\big(4A_{15}^{(2)}-2A_{22}^{(2)}+A_{18}^{(2)}\\
&\quad-\tfrac{7}{2}A_{14}^{(2)}+4A_{12}^{(2)}-A_{13}^{(2)}-4\zeta(2)A_{4}^{(2)}+4A_{23}^{(2)}\\
&\quad+4\zeta(2)A_{2}^{(2)}-A_{11}^{(2)}+11A_{8}^{(2)}\big)+32A_{19}^{(1)}+\zeta(3)(28A_{1}^{(0)}\\
&\quad-24A_{1}^{(1)}+48A_{1}^{(2)})-28A_{6}^{(1)}+36A_{6}^{(2)}+\tfrac{4}{15}A_{10}^{(-2)}+24A_{10}^{(0)}\\
&\quad+\tfrac{32}{3}A_{10}^{(1)}+\tfrac{48}{5}A_{10}^{(3)}+\zeta(2)\big(12A_{1}^{(0)}-\tfrac{52}{3}A_{1}^{(1)}+26A_{1}^{(2)}+\tfrac{48}{5}A_{1}^{(3)}\big)\\
&\quad-\tfrac{11}{2}A_{5}^{(0)}+22A_{5}^{(1)}-21A_{5}^{(2)}-A_{3}^{(0)}+\tfrac{2}{3}A_{3}^{(1)}-13A_{3}^{(2)}-\tfrac{24}{5}A_{3}^{(3)}\\
&\quad+7A_{4}^{(0)}-33A_{4}^{(1)}+24A_{4}^{(2)}-\tfrac{4}{15}A_{2}^{(-1)}-\tfrac{178}{15}A_{2}^{(0)}+\tfrac{34}{5}A_{2}^{(1)}\\
&\quad-\tfrac{168}{5}A_{2}^{(2)}+\tfrac{4}{15}A_{1}^{(-1)}-\tfrac{226}{15}A_{1}^{(0)}+\tfrac{17}{5}A_{1}^{(1)}+\tfrac{82}{5}A_{1}^{(2)}\Big\}\\
&\quad+C_A \;\Big\{4\Big[C_{24}-C_{25}-C_{26}+\tfrac{3}{4}C_{21}+\tfrac{3}{2}C_{20}\Big]\\
&\quad+2\Big[5(A_{15}^{(0)}+2A_{15}^{(1)})+(A_{16}^{(0)}+2A_{16}^{(1)})-3(A_{19}^{(0)}+2A_{19}^{(1)})+\zeta(2)(A_{17}^{(0)}+2A_{17}^{(1)})\Big]\\
&\quad+8(2A_{15}^{(1)}+A_{23}^{(1)})-16A_{22}^{(1)}+2A_{21}^{(2)}-4A_{20}^{(2)}+12A_{14}^{(1)}-4A_{14}^{(2)}\\
&\quad-2A_{12}^{(0)}-16A_{12}^{(1)}+4A_{12}^{(2)}+4A_{13}^{(0)}+16A_{13}^{(1)}-\zeta(2)(10A_{4}^{(0)}-12A_{4}^{(1)}\\
&\quad+16A_{4}^{(2)})+\tfrac{4}{3}A_{11}^{(0)}+4A_{11}^{(1)}-\zeta(2)(4A_{2}^{(0)}+40A_{2}^{(1)}-8A_{2}^{(2)})\\
&\quad-\zeta(3)(3A_{1}^{(0)}+14A_{1}^{(1)}+2A_{1}^{(2)})-\tfrac{8}{3}A_{10}^{(-1)}-12A_{10}^{(0)}+4A_{10}^{(1)}\\
&\quad+\tfrac{52}{3}A_{10}^{(2)}+\tfrac{16}{3}A_{8}^{(-1)}+5A_{8}^{(0)}-16A_{8}^{(1)}+\tfrac{20}{3}A_{8}^{(2)}\\
&\quad-\zeta(2)\big(8A_{1}^{(-1)}+A_{1}^{(0)}+52A_{1}^{(1)}-\tfrac{199}{3}A_{1}^{(2)}\big)+4A_{6}^{(0)}\\
&\quad-72A_{6}^{(1)}+73A_{6}^{(2)}+\tfrac{4}{3}A_{5}^{(-1)}-\tfrac{3}{2}A_{5}^{(0)}+16A_{5}^{(1)}-\tfrac{107}{6}A_{5}^{(2)}\\
&\quad+46A_{3}^{(1)}-\tfrac{57}{2}A_{3}^{(2)}+\tfrac{52}{9}A_{4}^{(-1)}-\tfrac{28}{3}A_{4}^{(0)}-\tfrac{215}{3}A_{4}^{(1)}+\tfrac{749}{9}A_{4}^{(2)}+\tfrac{73}{3}A_{2}^{(0)}\\
&\quad+83A_{2}^{(1)}-\tfrac{1445}{9}A_{2}^{(2)}+\tfrac{20}{9}A_{1}^{(-1)}+\tfrac{233}{18}A_{1}^{(0)}+\tfrac{65}{9}A_{1}^{(1)}-\tfrac{439}{18}A_{1}^{(2)}\Big\},
\end{aligned}$$

$$\begin{aligned}
C_{2,g}^{2Q} &= C_F \;\Big[2A_{4}^{(0)}-4A_{4}^{(1)}+4A_{4}^{(2)}-A_{2}^{(0)}+2A_{2}^{(1)}-4A_{2}^{(2)}-\tfrac{1}{2}A_{1}^{(0)}+2A_{1}^{(1)}\Big]\\
&\quad+C_A \;\Big[2A_{4}^{(0)}-4A_{4}^{(1)}+4A_{4}^{(2)}+2A_{2}^{(0)}+8A_{2}^{(1)}+\tfrac{4}{3}A_{1}^{(-1)}+A_{1}^{(0)}+8A_{1}^{(1)}\\
&\quad-\tfrac{31}{3}A_{1}^{(2)}\Big],
\end{aligned} \tag{B.9a}$$

$$
\begin{aligned}
C_{2,g}^{Q} &= C_F \quad [2A_8^{(0)} - 4A_8^{(1)} - (8A_1^{(0)} - 16A_1^{(1)} + 16A_1^{(2)})\zeta(2) - 6A_6^{(0)} \\
&\quad +12A_6^{(1)} - 16A_6^{(2)} + 4A_5^{(0)} - 8A_5^{(1)} + 8A_5^{(2)} + 2A_3^{(0)} - 4A_3^{(1)} \\
&\quad +8A_3^{(2)} - 7A_4^{(0)} + 24A_4^{(1)} - 20A_4^{(2)} + 2A_2^{(0)} - 12A_2^{(1)} + 20A_2^{(2)} \\
&\quad +9A_1^{(0)} - 17A_1^{(1)} + 4A_1^{(2)}] \\
&\quad +C_A \quad \big[- 4A_{10}^{(0)} - 8A_{10}^{(1)} - 8A_{10}^{(2)} + 4A_8^{(0)} + 16A_8^{(1)} \\
&\quad -(4A_1^{(0)} + 8A_1^{(2)})\zeta(2) + 24A_6^{(1)} - 8A_6^{(2)} + 2A_5^{(0)} - 4A_5^{(1)} \\
&\quad +4A_5^{(2)} - 4A_3^{(0)} - 12A_3^{(1)} + \tfrac{8}{3}A_4^{(-1)} - 2A_4^{(0)} + 40A_4^{(1)} - \tfrac{134}{3}A_4^{(2)} \\
&\quad -48A_2^{(1)} + 50A_2^{(2)} + \tfrac{52}{9}A_1^{(-1)} - \tfrac{55}{3}A_1^{(0)} - \tfrac{92}{3}A_1^{(1)} + \tfrac{407}{9}A_1^{(2)}\big],
\end{aligned}
\tag{B.9b}
$$

$$
\begin{aligned}
C_{2,g}^{\mu} &= C_A \quad \big\{4A_8^{(0)} + 16A_8^{(1)} - (4A_1^{(0)} - 8A_1^{(1)} + 8A_1^{(2)})\zeta(2) + 24A_6^{(1)} \\
&\quad -8A_6^{(2)} + 4A_5^{(0)} - 8A_5^{(1)} + 8A_5^{(2)} - 2A_3^{(0)} - 8A_3^{(1)} + \tfrac{8}{3}A_4^{(-1)} \\
&\quad -2A_4^{(0)} + 48A_4^{(1)} - \tfrac{158}{3}A_4^{(2)} - 2A_2^{(0)} - 64A_2^{(1)} \\
&\quad +\tfrac{62}{3}A_2^{(2)} + \tfrac{4}{3}A_1^{(-1)} - \tfrac{43}{3}A_1^{(0)} - \tfrac{242}{3}A_1^{(1)} + \tfrac{281}{3}A_1^{(2)}\big\},
\end{aligned}
\tag{B.9c}
$$

$$
\begin{aligned}
C_{2,g}^{\mu Q} &= C_A \quad \big[4A_4^{(0)} - 8A_4^{(1)} + 8A_4^{(2)} + 4A_2^{(0)} + 16A_2^{(1)} + \tfrac{8}{3}A_1^{(-1)} + 2A_1^{(0)} \\
&\quad +16A_1^{(1)} - \tfrac{62}{3}A_1^{(2)}\big],
\end{aligned}
\tag{B.9d}
$$

where

$$
B_n^{\pm} \equiv A_n^{(0)} \pm 2A_n^{(1)} + A_n^{(2)}, \qquad C_n \equiv A_n^{(0)} + 2A_n^{(1)} + 2A_n^{(2)}. \tag{B.10}
$$

The quark coefficient function for F_2 can be written similarly as

$$
\begin{aligned}
H_{2,q}^{(2)}\left(N, \frac{Q^2}{m^2}, \frac{\mu^2}{m^2}\right) &= 4T_f\Bigg[C_{2,q}^{\rm const}(N) + C_{2,q}^{2Q}(N)\ln^2\frac{Q^2}{m^2} + C_{2,q}^{Q}(N)\ln\frac{Q^2}{m^2} \\
&\quad -C_{2,q}^{\mu}(N)\ln\frac{\mu^2}{m^2} - C_{2,q}^{\mu Q}(N)\ln\frac{\mu^2}{m^2}\ln\frac{Q^2}{m^2}\Bigg],
\end{aligned}
\tag{B.11}
$$

where

$$
\begin{aligned}
C_{2,q}^{\rm const} &= C_F \quad \big[8(A_{15}^{(0)} + A_{15}^{(1)}) + 4(A_{13}^{(0)} + A_{13}^{(1)}) + 2(A_{14}^{(0)} + A_{14}^{(1)}) - 4(A_{12}^{(0)} + A_{12}^{(1)}) \\
&\quad -8\zeta(2)(A_2^{(0)} + A_2^{(1)}) + \tfrac{4}{3}(A_{11}^{(0)} + A_{11}^{(1)}) - \tfrac{8}{3}A_{10}^{(-1)} - 8A_{10}^{(0)} - 8A_{10}^{(1)} \\
&\quad -\tfrac{8}{3}A_{10}^{(2)} + \tfrac{16}{3}A_8^{(-1)} + 4A_8^{(0)} - 4A_8^{(1)} + \tfrac{8}{3}A_8^{(2)} - \zeta(2)\big(8A_1^{(-1)} + 4A_1^{(0)} \\
&\quad +4A_1^{(1)} - \tfrac{16}{3}A_1^{(2)}\big) + 8A_6^{(2)} + \tfrac{4}{3}A_5^{(-1)} + A_5^{(0)} - A_5^{(1)} - \tfrac{4}{3}A_5^{(2)} \\
&\quad +10A_3^{(1)} - 4A_3^{(2)} + \tfrac{52}{9}A_4^{(-1)} - \tfrac{52}{3}A_4^{(0)} + \tfrac{40}{3}A_4^{(1)} - \tfrac{16}{9}A_4^{(2)} + \tfrac{70}{3}A_2^{(0)} \\
&\quad -22A_2^{(1)} - \tfrac{176}{9}A_2^{(2)} + \tfrac{20}{9}A_1^{(-1)} + \tfrac{76}{9}A_1^{(0)} - \tfrac{304}{9}A_1^{(1)} + \tfrac{208}{9}A_1^{(2)}\big],
\end{aligned}
\tag{B.12a}
$$

$$
C_{2,q}^{2Q} = C_F \quad \big[2A_2^{(0)} + 2A_2^{(1)} + \tfrac{4}{3}A_1^{(-1)} + A_1^{(0)} - A_1^{(1)} - \tfrac{4}{3}A_1^{(2)}\big], \tag{B.12b}
$$

$$
\begin{aligned}
C_{2,q}^{Q} &= C_F \quad \big[4A_8^{(0)} + 4A_8^{(1)} + 4A_6^{(0)} + 4A_6^{(1)} - 4A_3^{(0)} - 4A_3^{(1)} + \tfrac{8}{3}A_4^{(-1)} \\
&\quad +2A_4^{(0)} - 2A_4^{(1)} - \tfrac{8}{3}A_4^{(2)} + 8A_2^{(2)} + \tfrac{52}{9}A_1^{(-1)} - \tfrac{52}{3}A_1^{(0)} \\
&\quad +\tfrac{40}{3}A_1^{(1)} - \tfrac{16}{9}A_1^{(2)}\big],
\end{aligned}
\tag{B.12c}
$$

$$C_{2,q}^{\mu} = C_F \quad \{4A_8^{(0)} + 4A_8^{(1)} + 4A_6^{(0)} + 4A_6^{(1)} - 2A_3^{(0)} - 2A_3^{(1)} + \tfrac{8}{3}A_4^{(-1)} + 2A_4^{(0)} - 2A_4^{(1)} - \tfrac{8}{3}A_4^{(2)} - 2A_2^{(0)} - 10A_2^{(1)} + \tfrac{8}{3}A_2^{(2)} + \tfrac{4}{3}A_1^{(-1)} - \tfrac{40}{3}A_1^{(0)} + \tfrac{4}{3}A_1^{(1)} + \tfrac{32}{3}A_1^{(2)}\}, \tag{B.12d}$$

$$C_{2,q}^{\mu Q} = C_F \quad \left[4A_2^{(0)} + 4A_2^{(1)} + \tfrac{8}{3}A_1^{(-1)} + 2A_1^{(0)} - 2A_1^{(1)} - \tfrac{8}{3}A_1^{(2)}\right]. \tag{B.12e}$$

The gluon radiation coefficient function for F_2 can be written as

$$H_{2,GR}^{(2)}\left(N, \frac{Q^2}{m^2}, \frac{\mu^2}{m^2}\right) = 4T_f \left[C_{2,GR}^{\rm const}(N) + C_{2,GR}^{2Q}(N) \ln^2 \frac{Q^2}{m^2} + C_{2,GR}^{Q}(N) \ln \frac{Q^2}{m^2}\right], \tag{B.13}$$

where

$$C_{2,GR}^{\rm const} = C_F \quad \left[-\tfrac{2}{3}(A_{33}^{(0)} + A_{33}^{(2)}) - \tfrac{2\zeta(2)}{3}(2A_{27}^{(0)} - A_1^{(0)} - A_1^{(1)}) - \tfrac{4}{3}(A_{32}^{(0)} + A_{32}^{(2)}) + \tfrac{1}{3}(2A_{30}^{(0)} - A_5^{(0)} - A_5^{(1)}) + A_{31}^{(0)} + A_{31}^{(2)} - \tfrac{29}{18}(2A_{28}^{(0)} - A_4^{(0)} - A_4^{(1)}) + \tfrac{67}{18}(A_{29}^{(0)} + A_{29}^{(2)}) + \tfrac{359}{108}(2A_{27}^{(0)} - A_1^{(0)} - A_1^{(1)}) + \tfrac{1}{6}A_4^{(0)} + \tfrac{13}{6}A_4^{(1)} - \tfrac{1}{2}A_2^{(0)} - \tfrac{23}{6}A_2^{(1)} + \tfrac{29}{36}A_1^{(0)} - \tfrac{295}{36}A_1^{(1)} + \tfrac{134\zeta(2)}{18} + \tfrac{265}{36}\right], \tag{B.14a}$$

$$C_{2,GR}^{2Q} = C_F\left[\tfrac{1}{3}(2A_{27}^{(0)} - A_1^{(0)} - A_1^{(1)}) + \tfrac{1}{2}\right], \tag{B.14b}$$

$$C_{2,GR}^{Q} = C_F \quad \left[\tfrac{2}{3}(2A_{28}^{(0)} - A_4^{(0)} - A_4^{(1)}) - \tfrac{29}{18}(2A_{27}^{(0)} - A_1^{(0)} - A_1^{(1)}) - \tfrac{4}{3}(A_{29}^{(0)} + A_{29}^{(2)}) + \tfrac{1}{6}A_1^{(0)} + \tfrac{13}{6}A_1^{(1)} - \tfrac{8\zeta(2)}{3} - \tfrac{19}{6}\right]. \tag{B.14c}$$

In addition to the elementary Mellin transforms listed in Tables B.1-B.2, the coefficient functions contain terms whose Mellin transform is known in closed form, but is expressed in terms of generalized harmonic sums $S_{i_1,\ldots,i_n}(N)$. The Mellin transform of these functions has been evaluated through suitable numerical approximations. The x-space expression whose Mellin transform will be evaluated in this way are the following:

$$f_{16}(z) = 2{\rm Li}_2(-z)\ln(1+z) + \ln(z)\ln^2(1+z) + 2S_{1,2}(-z), \tag{B.15a}$$

$$f_{22}(z) = {\rm Li}_3(1-z), \tag{B.15b}$$

$$f_{24}(z) = {\rm Li}_3\left(\frac{1-z}{1+z}\right) - {\rm Li}_3\left(-\frac{1-z}{1+z}\right), \tag{B.15c}$$

$$f_{25}(z) = \ln(z)\ln(1-z)\ln(1+z), \tag{B.15d}$$

$$f_{26}(z) = \ln(1-z){\rm Li}_2(-z), \tag{B.15e}$$

$$f_{33}(z) = \frac{{\rm Li}_2(1-z)}{1-z}. \tag{B.15f}$$

Each of these functions will now be considered in turn.

In order to determine the Mellin transform of Eq. (B.15a), Eq. (32) of Ref. [240] is

n	$f_n(z)$	$A_n(N) = \mathbf{M}[f_n(z)](N)$
1	1	$\frac{1}{N}$
2	$\ln(z)$	$-\frac{1}{N^2}$
3	$\ln^2(z)$	$\frac{2}{N^3}$
4	$\ln(1-z)$	$-\frac{S_1(N)}{N}$
5	$\ln^2(1-z)$	$\frac{S_1^2(N)+S_2(N)}{N}$
6	$\ln(z)\ln(1-z)$	$\frac{S_1(N)}{N^2}+\frac{S_2(N)-\zeta(2)}{N}$
8	$\mathrm{Li}_2(1-z)$	$-\frac{S_2(N)-\zeta(2)}{N}$
10	$\mathrm{Li}_2(-z)+\ln(z)\ln(1+z)$	$-\frac{\zeta(2)}{2N}-\frac{1}{4N}\left[S_2\left(\frac{N-1}{2}\right)-S_2\left(\frac{N}{2}\right)\right]$
11	$\ln^3(z)$	$-\frac{6}{N^4}$
12	$\ln^2(z)\ln(1-z)$	$\frac{2}{N}\left[\zeta(3)+\frac{\zeta(2)}{N}-\frac{S_1(N)}{N^2}-\frac{S_2(N)}{N}-S_3(N)\right]$
13	$\ln(1-z)\mathrm{Li}_2(1-z)-\mathrm{Li}_3(1-z)$	$\frac{S_1(N)S_2(N)-\zeta(2)S_1(N)+S_3(N)-\zeta(3)}{N}$
14	$\ln(z)\ln^2(1-z)$	$\frac{2}{N}\Big\{\zeta(3)+\zeta(2)S_1(N)-\frac{1}{2N}\left[S_1^2(N)+S_2(N)\right]$ $-S_1(N)S_2(N)-S_3(N)\Big\}$
15	$S_{1,2}(1-z)$	$-\frac{1}{N}\left[S_3(N)-\zeta(3)\right]$
17	$\ln(1+z)$	$\frac{\ln(2)}{N}+\frac{1}{2N}\left[S_1\left(\frac{N-1}{2}\right)-S_1\left(\frac{N}{2}\right)\right]$
18	$\ln^3(1-z)$	$-\frac{S_1^3(N)+3S_1(N)S_2(N)+2S_3(N)}{N}$
19	$\mathrm{Li}_3(-z)$	$-\frac{3\zeta(3)}{4N}+\frac{\zeta(2)}{2N^2}-\frac{\ln(2)}{N^3}-\frac{1}{2N^3}\left[S_1\left(\frac{N-1}{2}\right)-S_1\left(\frac{N}{2}\right)\right]$

Table B.1: Elementary Mellin transforms.

n	$f_n(z)$	$A_n(N) = \mathbf{M}[f(z)](N)$
20	$\ln(z)\mathrm{Li}_2(-z)$	$-\frac{1}{2N^2}\left[\frac{2}{N}S_1\left(\frac{N-1}{2}\right)+\frac{1}{2}S_2\left(\frac{N-1}{2}\right)+\frac{4\ln(2)}{N}\right]$ $+\frac{1}{2N^2}\left[\frac{2}{N}S_1\left(\frac{N}{2}\right)+\frac{1}{2}S_2\left(\frac{N}{2}\right)+\zeta(2)\right]$
21	$\ln^2(z)\ln(1+z)$	$\frac{1}{2N}\left[\frac{2}{N^2}S_1\left(\frac{N-1}{2}\right)+\frac{1}{N}S_2\left(\frac{N-1}{2}\right)+\frac{1}{2}S_3\left(\frac{N-1}{2}\right)+\frac{4\ln(2)}{N^2}\right]$ $-\frac{1}{2N}\left[\frac{2}{N^2}S_1\left(\frac{N}{2}\right)+\frac{1}{N}S_2\left(\frac{N}{2}\right)+\frac{1}{2}S_3\left(\frac{N}{2}\right)\right]$
23	$\ln(z)\mathrm{Li}_2(1-z)$	$\frac{1}{N^2}[S_2(N)-\zeta(2)]+\frac{2}{N}[S_3(N)-\zeta(3)]$
27	$\left(\frac{1}{1-z}\right)_+$	$-S_1(N-1)$
28	$\left(\frac{\ln(1-z)}{1-z}\right)_+$	$\frac{1}{2}S_1^2(N-1)+\frac{1}{2}S_2(N-1)$
29	$\frac{\ln(z)}{1-z}$	$S_2(N-1)-\zeta(2)$
30	$\left(\frac{\ln^2(1-z)}{1-z}\right)_+$	$-\frac{1}{3}S_1^2(N-1)-S_1(N-1)S_2(N-1)-\frac{2}{3}S_3(N-1)$
31	$\frac{\ln^2(z)}{1-z}$	$-2[S_3(N-1)-\zeta(3)]$
32	$\frac{\ln(z)\ln(1-z)}{1-z}$	$\zeta(3)+\zeta(2)S_1(N-1)-S_1(N-1)S_2(N-1)-S_3(N+l-1)$

Table B.2: Continuation of Table B.1.

used, which can be written as

$$A_{16}(N) = \mathbf{M}[f_{16}(z)](N) \quad = \frac{1}{N}\mathbf{M}\left[\frac{z\widetilde{\Phi}(z)}{1+z}\right] - \frac{1}{2N}\mathbf{M}\left[\frac{z\ln^2(z)}{1+z}\right] + \frac{\zeta(2)}{N}\mathbf{M}\left[\frac{z}{1+z}\right] - \frac{\zeta(2)\ln(2)}{N} + \frac{\zeta(3)}{4N}. \tag{B.16}$$

The last two Mellin transforms are respectively[1]:

$$\mathbf{M}\left[\frac{z\ln^2(z)}{1+z}\right] = -\frac{1}{4}\left[S_3\left(\frac{N-1}{2}\right) - S_3\left(\frac{N}{2}\right)\right], \tag{B.18}$$

and

$$\mathbf{M}\left[\frac{z}{1+z}\right] = -\frac{1}{2}\left[S_1\left(\frac{N-1}{2}\right) - S_1\left(\frac{N}{2}\right)\right]. \tag{B.19}$$

The Mellin transform of the term involving the special function $\widetilde{\Phi}(z)$ is given in Ref. [240] in terms of the generalized harmonic sum $S_{1,2}(N)$. To avoid evaluating this directly for complex N, it can be rewritten as

$$\frac{z\widetilde{\Phi}(z)}{1+z} \cong \sum_{k=1}^{10} a_k z^k \widetilde{\Phi}(z), \tag{B.20}$$

where the values of the coefficients a_k are determined by fitting the polynomial $\sum_{k=1}^{10} a_k z^k$ to the function $z/(1+z)$ on the unit interval. One can then use the Mellin transform of the function $z^k\widetilde{\Phi}(z)$, which reads

$$\mathbf{M}[z^k\widetilde{\Phi}(z)] = \frac{1}{(N+k)^3} + \frac{1}{2(N+k)}\left[S_2\left(\frac{N+k-1}{2}\right) - S_2\left(\frac{N+k}{2}\right)\right]. \tag{B.21}$$

The desired Mellin transform $A_{16}(N)$ is immediately found combining Eqs. (B.18), (B.19), (B.20) and (B.21).

[1] In general

$$\mathbf{M}\left[\frac{z\ln^n(z)}{1+z}\right] = \frac{(-1)^{n+1}n!}{2^{n+1}}\left[S_{n+1}\left(\frac{N-1}{2}\right) - S_{n+1}\left(\frac{N}{2}\right)\right]. \tag{B.17}$$

This Mellin transform is superficially different from the Mellin transform number 15 in the Appendix of Ref. [240], but they turn out to be equivalent after suitable simplification. This apparent difference is also responsible for the mismatch between entries 13, 28, 43, and 62 in the Table in the Appendix of Ref. [240] and entries 17, 21, 20 and 19 in Tables B.1 and B.2, respectively. For the same reason entries number 4 and 57 of [240] look different from Eqs. (B.18)-(B.21).

In order to determine the Mellin transform of Eq. (B.15b) the expansion

$$\mathrm{Li}_3(z) = \sum_{k=1}^{\infty} \frac{z^k}{k^3} \tag{B.22}$$

is used. It follows that

$$A_{22}(N) = \mathbf{M}[f_{22}(z)](N) = \sum_{k=1}^{\infty} \frac{1}{k^3} \int_0^1 z^{N-1}(1-z)^k dz, \tag{B.23}$$

but

$$\int_0^1 z^{N-1}(1-z)^k dz = \frac{\Gamma(N)\Gamma(k+1)}{\Gamma(N+k+1)} = \frac{k!}{(N+k)\ldots(N+1)N}, \tag{B.24}$$

so that

$$A_{22}(N) = \sum_{k=1}^{\infty} \frac{k!}{k^3(N+k)\ldots(N+1)N}. \tag{B.25}$$

In the NNPDF implementation this series has been truncated at $k = 30$.

Next, the Mellin transform of Eq. (B.15c) is discussed. In this case the function

$$g(z) = (1-z)^{b-1} \sum_{k=0}^{10} c_k z^k \tag{B.26}$$

is fitted to the function $f_{24}(z)$. However, one can show that

$$f_{24}(0) = \mathrm{Li}_3(1) - \mathrm{Li}_3(-1) = \tfrac{7}{4}\zeta(3) = c_0, \tag{B.27}$$

so that c_0 is fixed, and needs to be fitted only

$$g(z) = (1-z)^{b-1}\big[\tfrac{7}{4}\zeta(3) + \sum_{k=1}^{10} c_k z^k\big]. \tag{B.28}$$

The Mellin transform of Eq. (B.15c) follows immediately, because

$$\begin{aligned} A_{24}(N) = \mathbf{M}[f_{24}(z)](N) &= \textstyle\sum_{k=0}^{10} c_k \int_0^1 z^{N+k-1}(1-z)^{b-1} dz \\ &= \frac{\Gamma(N)\Gamma(b)}{\Gamma(N+b)} \sum_{k=0}^{10} c_k \frac{(N+k)\ldots(N+1)N}{(N+b+l+k)\ldots(N+b+l)}. \end{aligned} \tag{B.29}$$

In order to determine the Mellin transform of Eq. (B.15d), the representation given as Eq. (21) of Ref. [241] is used, in which the function $\ln(1+z)$ is approximated by the

polynomial

$$\ln(1+z) \cong \sum_{k=1}^{8} d_k z^k. \tag{B.30}$$

Using also Table B.1, it then gives

$$A_{25}(N) = \mathbf{M}[f_{25}(z)](N) = \sum_{k=1}^{8} d_k \mathbf{M}[z^k \ln(z)\ln(1-z)] = \sum_{k=1}^{8} d_k \left[\frac{S_1(N+k)}{(N+k)^2} + \frac{S_2(N+k)-\zeta(2)}{N+k}\right]. \tag{B.31}$$

In order to determine the Mellin transform of Eq. (B.15e) the expansion

$$\mathrm{Li}_2(-z) = \sum_{k=1}^{\infty} \frac{(-1)^k z^k}{k^2} \tag{B.32}$$

is used, so

$$A_{26}(N) = \mathbf{M}[f_{26}(z)](N) = \sum_{k=1}^{\infty} \frac{(-1)^k}{k^2} \mathbf{M}[z^k \ln(1-z)] = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^2} \frac{S_1(N+k)}{N+k}. \tag{B.33}$$

In this case has been decided to truncate the series at $k = 100$.

Finally, the Mellin transform of Eq. (B.15f) is determined. In this case the geometric series for $1/(1-z)$ is used so that

$$A_{33}(N)\mathbf{M}\left[f_{33}(z)\right](N) = \sum_{k=0}^{\infty} \mathbf{M}[z^k \mathrm{Li}_2(1-z)] = -\sum_{k=0}^{\infty} \frac{S_2(N+k)-\zeta(2)}{N+k}. \tag{B.34}$$

As in the case above, the series is truncated at $k = 100$.

Note that the Mellin transform of the x-space coefficient functions involve terms of the form $z^l f_n(z)$. The Mellin transform $\mathbf{M}[z^l f_n(z)](N)$ can be obtained from $\mathbf{M}[f_n(z)](N)$ using the identity $\mathbf{M}[z^l f_n(z)](N) = \mathbf{M}[f_n(z)](N+l)$. Thus the Mellin transform of any of the terms in Tables B.1-B.2 and also any of the terms Eqs.(B.15a)-(B.15f) multiplied by a factor z^l can be obtained replacing N with $N+l$.

In conclusion, this calculation has been checked by comparing the inverse Mellin transform of the N-space coefficients with the original x-space results. In Table B.3 the accuracy of this comparison is shown for the various coefficient functions. Excellent accuracy is found for all coefficients and all values of x.

x	$\epsilon_{\rm rel}\left(C_{2,g}^{(n_l,0),2}\right)$	$\epsilon_{\rm rel}\left(C_{2,q}^{(n_l,0),2}\right)$	$\epsilon_{\rm rel}\left(C_{L,g}^{(n_l,0),2}\right)$	$\epsilon_{\rm rel}\left(C_{L,q}^{(n_l,0),2}\right)$
10^{-7}	7×10^{-12}	4×10^{-11}	2×10^{-10}	7×10^{-12}
10^{-6}	3×10^{-11}	1×10^{-11}	4×10^{-11}	3×10^{-12}
10^{-5}	2×10^{-12}	3×10^{-11}	4×10^{-13}	2×10^{-12}
10^{-4}	6×10^{-10}	1×10^{-11}	3×10^{-11}	4×10^{-12}
10^{-3}	6×10^{-9}	2×10^{-12}	1×10^{-10}	2×10^{-11}
10^{-2}	7×10^{-8}	7×10^{-12}	2×10^{-10}	1×10^{-11}
10^{-1}	1×10^{-7}	4×10^{-11}	1×10^{-11}	5×10^{-13}
3×10^{-1}	9×10^{-7}	1×10^{-11}	1×10^{-11}	5×10^{-13}
5×10^{-1}	4×10^{-6}	3×10^{-11}	8×10^{-12}	1×10^{-13}
7×10^{-1}	1×10^{-5}	9×10^{-11}	6×10^{-12}	5×10^{-13}
9×10^{-1}	1×10^{-5}	6×10^{-8}	7×10^{-8}	1×10^{-8}

Table B.3: Comparison of the inverse Mellin transforms of coefficient functions computed here to the original x-space expressions of Ref. [237]: the percentage difference between the original expression and the numerical Mellin inverse is shown in each case.

Appendix C

FastKernel Implementation of FONLL-C

In this Appendix the implementation and benchmarking of the FONLL-C neutral current structure functions in the FastKernel framework is discussed. At the same time has been implemented and benchmarked FONLL-B.

The implementation of FONLL-B and FONLL-C structure functions in FastKernel requires the $\mathcal{O}\left(\alpha_s^2\right)$ massive heavy quark coefficient functions as well as their asymptotic $Q^2 \to \infty$ limit in Mellin space. In Ref. [40] analytic results for the Mellin space $\mathcal{O}\left(\alpha_s\right)$ heavy quark coefficient functions for neutral current and charged current scattering are presented (see also Ref. [242] for the latter). For the $\mathcal{O}\left(\alpha_s^2\right)$ heavy quark coefficient functions in Mellin space the parametrization of Ref. [244] is used. The Mellin transforms in the asymptotic limit as $Q^2 \to \infty$ have been determined as in Appendix B.

The $\mathcal{O}\left(\alpha_s^2\right)$ massless coefficient functions were first computed in Refs. [46, 245–247]. For the implementation in the FastKernel framework the fast Mellin space parametrizations of the exact coefficient functions as given in Refs. [248, 249] has been used.

The gluon radiation terms, namely contributions with heavy quarks in the final state but where the struck quark is light, have to be treated with care. As discussed in Ref. [36], the gluon radiation contribution (which first appears at $\mathcal{O}\left(\alpha_s^2\right)$) is part of the light quark structure functions. It has been checked (see Fig. C.1) that the size of these terms is very small both for the F_2 and F_L structure functions, typically below 1%.

Now the implementation and benchmarking of these results in the FastKernel framework are discussed. Benchmarking has been performed by comparing the FastKernel results with the FONLLdis code [36], an x-space code that implements all FONLL schemes. In Table C.1 are shown the results of the benchmark comparison for the $F_{2c}(x, Q^2)$ and $F_{Lc}(x, Q^2)$ structure functions in the massive fixed flavor number scheme at $\mathcal{O}\left(\alpha_s^2\right)$ for the FONLLdis code and for the FastKernel code. Results are provided at the reference

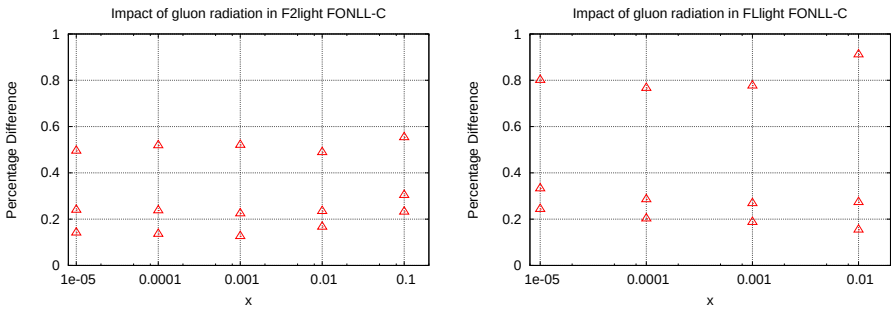

Figure C.1: Percentage difference between the NNLO light quark structure functions F_2 (left) and F_L (right) with and without gluon radiation contributions. From bottom to top the points correspond to scales of Q^2=4, 10 and 100 GeV^2 respectively.

points in the (x, Q^2) plane and with the settings of the Les Houches heavy quark benchmarks [38]. The accuracy is given as the percentage difference between the FastKernel and FONLLdis calculations. The accuracy is never worse than 1%, which is amply sufficient for the purposes of the present analysis. The accuracy of Table C.1 is a little worse than that of its $\mathcal{O}(\alpha_s)$ counterpart, shown as Tables 14-15 of Ref. [40]. This may be due to the fact that the Mellin space parametrizations of the heavy quark coefficient functions that are used [244] are in turn based on a parametrization of the exact x-space coefficient functions, while FONLLdis uses the original exact coefficient functions of Ref [250]. This loss of accuracy is negligible for the needs of current phenomenology, but more detailed studies of this issue may be needed in the future when final combined HERA heavy quark structure function data become available.

The same comparisons for the FONLL-B and FONLL-C are presented in Tables C.2 and C.3 respectively. Comparable accuracy is achieved for these two GM-VFN schemes, again sufficient for the purposes of the present analysis.

	F_{2c} FFNS			F_{Lc} FFNS		
x	FONLLdis	FastKernel	Accuracy (%)	FONLLdis	FastKernel	Accuracy (%)
$Q^2 = 4$ GeV^2						
10^{-5}	0.24591	0.24244	1.43	0.02215	0.02184	1.41
10^{-4}	0.13658	0.13481	1.31	0.01306	0.01283	1.75
10^{-3}	0.06384	0.06308	1.20	0.00662	0.00653	1.41
10^{-2}	0.02025	0.02007	0.92	0.00238	0.00237	0.38
$Q^2 = 10$ GeV^2						
10^{-5}	0.53701	0.53904	0.38	0.08031	0.08105	0.91
10^{-4}	0.29558	0.29550	0.03	0.04611	0.04579	0.71
10^{-3}	0.13909	0.13852	0.24	0.02273	0.02254	0.86
10^{-2}	0.04689	0.04664	0.10	0.00832	0.00826	0.70
$Q^2 = 100$ GeV^2						
10^{-5}	1.99594	2.00744	0.57	0.44200	0.43976	0.51
10^{-4}	1.00912	1.01479	0.56	0.22148	0.21880	1.22
10^{-3}	0.43527	0.43410	0.27	0.09487	0.09380	1.14
10^{-2}	0.13574	0.13492	0.61	0.03019	0.03002	0.57

Table C.1: Benchmark comparisons for the $F_{2c}(x, Q^2)$ and $F_{Lc}(x, Q^2)$ structure functions in the FFN scheme at O(α_s^2) obtained using the FONLLdis code [36] and the FastKernel code. Results are provided at the benchmark kinematic points in x, Q^2 and with the settings of the Les Houches heavy quark benchmarks [38]. The accuracy is given as the percentage difference between the FastKernel and FONLLdis [36] calculations.

	F_{2c} FONLL-B			F_{Lc} FONLL-B		
x	FONLLdis	FastKernel	Accuracy (%)	FONLLdis	FastKernel	Accuracy (%)
$Q^2 = 4$ GeV^2						
10^{-5}	0.24787	0.24858	0.29	0.02519	0.02524	0.21
10^{-4}	0.13556	0.13598	0.31	0.01435	0.01435	0.01
10^{-3}	0.06360	0.06350	0.15	0.00718	0.00715	0.34
10^{-2}	0.02062	0.02051	0.52	0.00258	0.00258	0.08
$Q^2 = 10$ GeV^2						
10^{-5}	0.55100	0.55088	0.02	0.09637	0.09679	0.43
10^{-4}	0.30114	0.30150	0.14	0.05229	0.05222	0.14
10^{-3}	0.14371	0.14375	0.02	0.02507	0.02499	0.32
10^{-2}	0.05012	0.05015	0.01	0.00908	0.00907	0.06
$Q^2 = 100$ GeV^2						
10^{-5}	2.10034	2.08834	0.57	0.48769	0.48716	0.11
10^{-4}	1.04510	1.05096	0.56	0.23569	0.23418	0.65
10^{-3}	0.45879	0.45916	0.08	0.09923	0.09888	0.35
10^{-2}	0.15039	0.15030	0.06	0.03170	0.03174	0.15

Table C.2: Same as Table C.1 for the FONLL-B GM-VFN scheme.

	F_{2c} FONLL-C			F_{Lc} FONLL-C		
x	FONLLdis	FastKernel	Accuracy (%)	FONLLdis	FastKernel	Accuracy (%)
$Q^2 = 4$ GeV^2						
10^{-5}	0.27830	0.28163	1.18	0.02468	0.02500	1.30
10^{-4}	0.14709	0.14858	1.00	0.01423	0.01441	1.23
10^{-3}	0.06556	0.06591	0.52	0.00733	0.00735	0.24
10^{-2}	0.02034	0.02034	0.00	0.00281	0.00283	0.74
$Q^2 = 10$ GeV^2						
10^{-5}	0.69412	0.69873	0.66	0.09909	0.10062	1.52
10^{-4}	0.34662	0.34911	0.88	0.05520	0.05550	0.52
10^{-3}	0.15025	0.15114	0.32	0.02682	0.02699	0.63
10^{-2}	0.04986	0.05022	0.13	0.01002	0.01008	0.58
$Q^2 = 100$ GeV^2						
10^{-5}	2.36920	2.37887	0.41	0.47822	0.47994	0.36
10^{-4}	1.12695	1.13870	1.03	0.23916	0.23914	0.01
10^{-3}	0.47058	0.47317	0.55	0.10262	0.10293	0.30
10^{-2}	0.15175	0.15236	0.40	0.03312	0.03327	0.47

Table C.3: Same as Table C.1 for the FONLL-C GM-VFN scheme.

Appendix D

Distribuciones Partónicas para Fenomenología de Precisión del GCH

Los desarrollos de la física de colisionadores de los últimos cincuenta años han permitido alcanzar resultados fundamentales a nivel experimental y consecuentemente a nivel teórico. El descubrimiento de nuevas partículas, los quarks y los gluones, fue una sólida confirmación de la teoría de la cromodinámica cuántica (QCD), se enmarca en el modelo estándar de las partículas elementales. Los parámetros de este modelo han sido medidos con elevada precisión, confirmando las teorías y confiriendoles gran poder predictivo. Toda la física de los colisionadores de hadrones, en particular la del Gran Colisionador de Hadrones (GCH), activo al CERN de Ginebra, se apoya en la cromodinámica cuántica y en general en el modelo estándar.

Cromodinámica Cuántica Perturbativa

La cromodinámica cuántica describe la física de los hadrones basandose en la dinámica y en las interacciones fundamentales de quarks y gluones. Al ser la QCD una teoría no-Abeliana, los gluones (bosones responsables de la interacción entre quarks) pueden interactuar entre ellos, al contrario de lo que ocurre para los fotones en el caso de la electrodinámica cuántica (QED). Esta característica de auto-interacción de los gluones provoca un comportamiento de la constante de acoplamiento α_S de la teoría diferente respecto a la análoga constante de acoplamiento de la electrodinámica α_{QED}. En el caso de la QCD a bajas energías el valor de α_S no permite el uso de la teoría perturbativa, siendo la interacción fuerte. Esta propiedad se traduce en *confinamiento*: los quarks no pueden ser aislados. Fueron D.J. Gross, F. Wilczek, y H.D. Politzer quienes demonstraron la *libertad asintótica* de la QCD, mereciendo el Nobel en el año 2004. Al subir en la escala de energía la constante α_S se vuelve más pequeña, permitiendo así la aproximación per-

turbativa. Gracias a la libertad asintótica y a las propiedades del *teorema de factorización* la QCD perturbativa se vuelve una teoría de precisión. Los estados iniciales y finales de cualquier proceso que involucre hadrones están constituidos por quarks y gluones fuera del régimen perturbativo. En el cálculo de los observables una parte no-perturbativa puede factorizarse de la parte puramente perturbativa, la cual puede ser calculada con la teoría. Considérese en el caso de un proceso de *deep inelastic scattering* (DIS) el ejemplo de la función de estructura F_2 medida en un punto cinemático (x, Q^2) dado. Es posible escribir este observable como la convolución de una distribución partónica f_i (parte no-perturbativa) con la parte calculable perturbativamente (observable partónica $\hat{F}_2$). Gracias a las ecuaciones de evolución DGLAP es posible calcular perturbativamente la dependencia de las distribuciones partónicas con la escala de factorización. En el marco de la teoría perturbativa los efectos debidos a las masas de los quarks pesados son tenidos en cuenta gracias al uso del esquema FONLL. Los términos masivos son incluidos al orden α_s (FONLL-A) o α_s^2 (FONLL-C) para determinar respectivamente las distribuciones NNPDF2.1 NLO y NNPDF2.1 NNLO.

Determinación de Distribuciones Partónicas

Las distribuciones partónicas (PDFs) son extraidas a partir de datos de diferentes experimentos. Determinar una densidad de probabilidad en el espacio de las funciones a partir de un número finito de puntos experimentales es un problema de dificil solución y además la forma funcional de las distribuciones no es conocida.

Varios grupos se dedican a la determinación de las PDFs. La mayoría de ellos fija una forma funcional (generalmente una polinomial) y reduce así el problema a la determinación de un conjunto finito de parámetros. Los errores experimentales son transferidos a las PDFs utilizando el método del Hessiano. De este modo, es introducida fácilmente una desviación sistemática en la determinación de los resultados debida a la rigidez de la parametrización escogida y una sub-estimación de los errores, parcialmente corregida a traves del uso de tolerancias más grandes que uno.

La colaboración NNPDF utiliza una parametrización de redes neuronales cuyo entrenamiento es efectuado utilizando algoritmos genéticos. De esta manera, la desviación debida a la parametrización es minimizada. También se utiliza el método Monte Carlo para generar $N_{\rm rep}$ réplicas de los datos experimentales y poder así determinar $N_{\rm rep}$ conjuntos de PDFs en lugar de uno solo. De esta forma, es posible determinar bandas de error para las PDFs que mantienen una interpretación estadística rigurosa. El uso de redes neuronales permite una parametrización muy flexible y redundante. A consecuencia de esto es necesario parar el entrenamiento antes de alcanzar el mínimo χ^2 posible que correspondería a un régimen de sobre-entrenamiento. Para no caer en esta situación un

método de *cross-validation* es utilizado para determinar el mejor ajuste.

Resultados

En este trabajo, realizado en el ámbito de la colaboración NNPDF, los conjuntos de distribuciones partónicas determinados incluyen los efectos de masa debidos a los quarks pesados. Se han determinado tres conjuntos: NNPDF2.1 LO, NLO y NNLO. En el caso NLO, las PDFs se han comparado con el conjunto NNPDF2.0, el cual no incluye efectos de masa, y también con distribuciones de otros grupos como CT10 y MSTW08, que los incluyen usando métodos diferentes del FONLL. Las distribuciones LO se han comparado tanto con las NLO como con las distribuciones de las colaboraciones MSTW y CTEQ, mientras que las distribuciones NNLO se han comparado con las MSTW08 NNLO y ABKM09 NNLO (además de compararlas con NNPDF2.1 NLO). Se ha verificado que en general las distribuciones de la colaboración NNPDF son consistentes dentro de una sigma (solo en el caso LO esto no es verdad, debido a la importancia de la incertidumbre teórica en este caso). En comparación con los resultados de los demás grupos se han observado diferencias de alguna forma significativas en ciertos casos específicos, sobretodo en el caso NNLO. Comparando los tres conjuntos LO, NLO y NNLO es posible estudiar la estabilidad estadística, siendo los tres determinados según la misma metodología y usando los mismos datos. Se observa una excelente convergencia de la expansión perturbativa en la región cinemática de los datos. De la comparación de un orden perturbativo con el siguiente es posible dar una estimación de la incertidumbre teórica que afecta al orden más bajo. En el caso de la determinación a LO, se ve claramente que el error experimental solo es una fracción del error teórico (que no está aquí incluido). En cambio, para el conjunto NLO (y plausiblemente NNLO) la incertidumbre teórica resulta despreciable en comparación a la experimental.

Han sido también producidos varios conjuntos de PDFs con diferentes valores tanto de α_s como de m_c y m_b. De esta forma, es posible determinar las incertidumbres combinadas para las PDFs y para el parámetro que varía.

Se han calculado las predicciones para un conjunto de observables del GCH (producción de bosones vectores, top quarks y bosones de Higgs) usando las distribuciones partónicas NNPDF2.1 NLO y NNLO. Estas predicciones han sido comparadas entre ellas y también con predicciones obtenidas usando las PDFs de otros grupos. Ha sido posible comparar también estos resultados con los primeros datos de las colaboraciones ATLAS y CMS. Al incluir los efectos debidos a las masas de los quarks pesados se observa una modificación bastante pequeña en estas observables. En cambio, la diferencia es substancial si se consideran observables que sondean directamente las distribuciones partónicas de los quarks pesados.

La metodología usada por la colaboración NNPDF ha sido aprovechada también para la determinación de la constante de acoplamiento fuerte $\alpha_s(M_Z)$. Esto es posible gracias a que las predicciones teóricas para el conjunto global de datos dependen simultaneamente tanto de las PDFs como de α_s. El valor así determinado ha sido $\alpha_s(M_Z) = 0.1191 \pm 0.0006^{\text{exp}}$. En la determinación de este parámetro la incertidumbre teórica es dominante y significativa, y no está incluida en el error aquí determinado que en cambio solo representa la incertidumbre estadística.

List of Figures

List of Tables

Bibliography

[1] R.K. Ellis, W.J. Stirling, and B.R. Webber, Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol. **8**, 1 (1996).

[2] M.E. Peskin and D.V. Schroeder, Reading, USA: Addison-Wesley (1995) 842 p.

[3] F. Halzen and A.D. Martin, Reading, USA: John Wiley and Sons (1984) 396 p.

[4] M. Gell-Mann, Caltech Synchrotron Laboratory Report No. CTSL-20 (1961).

[5] Y. Ne'eman, Nucl. Phys. **26**, 222 (1961).

[6] M. Gell-Mann, Phys. Lett. **8**, 214 (1964).

[7] G. Zweig, *An SU3 Model for Strong Interaction Symmetry and Its Breaking* CERN Report No. TH 412 (Geneva, 1964); G. Zweig, *Symmetries in Elementary Particles* p.192 (Academic Press, NY, 1965).

[8] R.E. Taylor, Rev. Mod. Phys. **63** 573 (1991); H.W. Kendall, Rev. Mod. Phys. **63** 597 (1991); J.I. Friedman, Rev. Mod. Phys. **63** 615 (1991).

[9] M. Breidenbach *et al.,* Phys. Rev. Lett. **23**, 935 (1969).

[10] J.I. Friedman, Rev. Mod. Phys. **63**, 615 (1991).

[11] E. Rutherford, Phil. Mag. **21**, 669 (1911).

[12] Drell and Walecka, Ann. Phys. **28**(1), 18 (1964).

[13] `http://www.atlas.ch/photos/events-collision-proton.html`.

[14] J.J. Aubert *et al.,* Phys. Rev. Lett. **33**, 1404 (1974).

[15] G. Hanson *et al.,* Phys. Rev. Lett. **35**, 1609 (1975).

[16] G. Hanson *et al.,* Phys. Rev. Lett. **35**, 1609 (1975).

[17] R. Brandelik *et al.,* Phys. Lett. **86B**, 243 (1979); D. P. Barber *et al.,* Phys. Rev. Lett. **43**, 830 (1979); Ch. Berger *et al.,* Phys. Lett. **86B**, 418 (1979).

[18] D.J. Gross and F. Wilczek, Phys. Rev. Lett. **30**, 1343 (1973).

[19] H.D. Politzer, Phys. Rev. Lett. **30**, 1346 (1973).

[20] O.V. Tarasov, A.A. Vladimirov and A.Yu. Zharkov, Phys. Lett. **B93**, 429 (1980); S.A. Larin and J.A.M. Vermaseren, Phys. Lett. **B303**, 334 (1993).

[21] A.J. Buras, Rev. Mod. Phys. **52**, 199 (1980).

[22] J.C. Collins and D.E. Soper, Ann. Rev. Nucl. Part. Sci. **37**, 383 (1987).

[23] F. Bloch and A. Nordsieck, Phys. Rev. **52**, 54 (1937).

[24] T. Kinoshita, J. Math. Phys. **3**, 650 (1962).

[25] T.D. Lee and M. Nauenberg, Phys. Rev. **133B**, 1549 (1964).

[26] C.F. von Weizsäcker, Z. Phys. **88**, 612-625 (1934).

[27] E.J. Williams, K. Dan. Vidensk. Selsk. Mat.-Fys. Medd. **13**, (1935).

[28] H: Georgi and H.D. Politzer, Phys. Rev. **D9**, 416 (1974).

[29] D. Gross and F. Wilczek, Phys. Rev. **D9**, 980 (1974).

[30] S. Moch, J.A.M. Vermaseren and A. Vogt, Nucl. Phys. **B621**, 413 (2002), hep-ph/0110331.

[31] S. Moch, J.A.M. Vermaseren and A. Vogt, Nucl. Phys. **B688**, 101 (2004), hep-ph/0403192.

[32] A. Vogt, S. Moch and J.A.M. Vermaseren, Nucl. Phys. **B691**, 129 (2004), hep-ph/0404111.

[33] A. Vogt, Comput. Phys. Commun. **170**, 65 (2005), hep-ph/0408244.

[34] M. Cacciari, M. Greco and P. Nason, JHEP 9805:007 (1998), hep-ph/9803400.

[35] J.C. Collins, F. Wilczek and A. Zee, Phys. Rev. **D18**, 242 (1978).

[36] S. Forte *et al.,* Nucl. Phys. **B834**, 116 (2010), 1001.2312.

[37] M. Kramer, F.I. Olness and D.E. Soper, Phys. Rev. **D62** 096007 (2000), hep-ph/0003035.

[38] J. Rojo *et al.*, Chapter 22 in: J.R. Andersen *et al.,* "The SM and NLO multileg working group: Summary report", 1003.1241, (2010).

[39] P.M. Nadolsky and W.K. Tung, Phys. Rev. **D79**, 113014 (2009), 0903.2667.

[40] R.D. Ball *et al.,* Nucl. Phys. **B849**, 296 (2011), hep-ph/1101.1300.

[41] E. Witten, Nucl. Phys. **B104**, 445 (1976).

[42] M.A. Shifman, A.I. Vainshtein and V.I. Zakharov, Nucl. Phys. **B136**, 157 (1978).

[43] J.P. Leveille and T.J. Weiler, Nucl. Phys. **B147**, 147 (1979).

[44] M. Gluck, S. Kretzer and E. Reya, Phys. Lett. **B380**, 171 (1996), hep-ph/9603304.

[45] T. Gottschalk, Phys. Rev. **D23**, 56 (1981).

[46] E.B. Zijlstra and W.L. van Neerven, Nucl. Phys. **B383**, 525 (1992).

[47] Particle Data Group, K. Nakamura, J. Phys. **G37** 075021 (2010).

[48] D. Mason *et al.*, Phys. Rev. Lett. 99 192001 (2007).

[49] The NNPDF collaboration, R.D. Ball *et al.*, Nucl. Phys. **B823**, 195 (2009), 0906.1958.

[50] R.D. Ball *et al.*, (2012), hep-ph/1207.1303.

[51] J. Rojo *et al.*, Fonlldis: `http://wwwteor.mi.infn.it/~rojo/fonlldis.html`, (2010).

[52] G. Watt, Private communication, (2010).

[53] NNPDF Collaboration, Nucl. Phys. **B855**, 153 (2012), 1107.2652.

[54] M. Buza and W.L. van Neerven, Nucl. Phys. **B500**, 301 (1997), hep-ph/9702242.

[55] R. McElhaney and S.F. Tuan, Phys. Rev. **D8**, 2267 (1973).

[56] M. Gluck and E. Reya, Phys. Rev. **D14**, 3034 (1976).

[57] A.J. Buras and K.J.F. Gaemers, Nucl. Phys. **B132**, 249 (1978).

[58] New Muon Collaboration, M. Arneodo *et al.*, Nucl. Phys. **B483**, 3 (1997).

[59] New Muon Collaboration, M. Arneodo *et al.*, Nucl. Phys. **B487**, 3 (1997).

[60] M. Klein and R. Yoshida, Prog. Part. Nucl. Phys. 61:343, (2008).

[61] ATLAS Collaboration, Phys. Lett. **B716**, 1 (2012).

[62] CMS Collaboration, Phys. Lett. **B716**, 30 (2012).

[63] D.W. Duke and J.F. Owens, Phys. Rev. **D30**, 49 (1984).

[64] F. Olness and I. Schienbein, Nucl. Phys. Proc. Suppl. **191**, 44 (2009), 0812.3371.

[65] J.C. Collins, Phys. Rev. **D58**, 094002 (1998), hep-ph/9806259.

[66] M.A.G. Aivazis, J.C. Collins, F.I. Olness, and W.K. Tung, Phys. Rev. **D50**, 3102 (1994), hep-ph/9312319.

[67] W.K. Tung, S. Kretzer, and C. Schmidt, J. Phys. **G28**, 983 (2002), hep-ph/0110247.

[68] S. Kretzer, H.L. Lai, F.I. Olness, and W.K. Tung, Phys. Rev. **D69**, 114005 (2004), hep-ph/0307022.

[69] R.S. Thorne and R.G. Roberts, Phys. Rev. **D57**, 6871 (1998), hep-ph/9709442.

[70] R.S. Thorne, Phys. Rev. **D73**, 054019 (2006), hep-ph/0601245.

[71] J. Pumplin *et al.,* JHEP 07, 012 (2002), hep-ph/0201195.

[72] J. Huston, J. Pumplin, D. Stump, and W.K. Tung, JHEP 06, 080 (2005), hep-ph/0502080.

[73] W.K. Tung *et al.,* JHEP 02, 053 (2007), hep-ph/0611254.

[74] J.F. Owens *et al.,* Phys. Rev. **D75**, 054030 (2007), hep-ph/0702159.

[75] H.L. Lai *et al.,* JHEP 04, 089 (2007), hep-ph/0702268.

[76] H.L. Lai *et al.,* Phys. Rev. **D82**, 074024 (2010), hep-ph/1007.2241.

[77] A.D. Martin, R.G. Roberts, W.J. Stirling, and R.S. Thorne, Eur. Phys. J. **C28**, 455 (2003), hep-ph/0211080.

[78] A.D. Martin, R.G. Roberts, W.J. Stirling, and R.S. Thorne, Eur. Phys. J. **C35**, 325 (2004), hep-ph/0308087.

[79] A.D. Martin, R.G. Roberts, W.J. Stirling, and R.S. Thorne, Phys. Lett. **B604**, 61 (2004), hep-ph/0410230.

[80] A.D. Martin, W.J. Stirling, R.S. Thorne, and G. Watt, Phys. Lett. **B652**, 292 (2007), 0706.0459.

[81] A.D. Martin, W.J. Stirling, R.S. Thorne, and G. Watt, Eur. Phys. J. **C63**, 189 (2009), 0901.0002.

[82] A.D. Martin, W.J. Stirling, R.S. Thorne, and G. Watt, Eur. Phys. J. **C64** 653 , (2009), 0905.3531.

[83] NNPDF, L. Del Debbio, S. Forte, J.I. Latorre, A. Piccione, and J. Rojo, JHEP 03, 039 (2007), hep-ph/0701127.

[84] NNPDF, R.D. Ball *et al.,* Nucl. Phys. B809, 1 (2009), 0808.1231.

[85] NNPDF, J. Rojo *et al.,* (2008), 0811.2288.

[86] NNPDF, R.D. Ball *et al.,* Nucl. Phys. B823, 195 (2009), 0906.1958.

[87] R.D. Ball *et al.,* Nucl. Phys. **B838**, 136 (2010), 1002.4407.

[88] R.D. Ball *et al.,* Nucl. Phys. **B855**, 153 (2012).

[89] S. Alekhin, Phys. Rev. **D63**, 094022 (2001), hep-ph/0011002.

[90] S. Alekhin, Phys. Rev. **D68**, 014002 (2003), hep-ph/0211096.

[91] S. Alekhin, JETP Lett. **82**, 628 (2005), hep-ph/0508248.

[92] S. Alekhin, K. Melnikov, and F. Petriello, Phys. Rev. **D74**, 054033 (2006), hep-ph/0606237.

[93] H.L. Lai and W.K. Tung, Z. Phys. **C74**, 463 (1997), hep-ph/9701256; CTEQ, H.L. Lai *et al.,* Eur. Phys. J. **C12**, 375 (2000), hep-ph/9903282; S. Kretzer, H.L. Lai, F.I. Olness, and W.K. Tung, Phys. Rev. **D69**, 114005 (2004).

[94] S. Alekhin, J. Blumlein, S. Klein, and S. Moch, Phys. Rev. D81, 014032 (2010), 0908.2766.

[95] W.T. Giele and S. Keller, Phys. Rev. **D58**, 094023 (1998), hep-ph/9803393.

[96] W.T. Giele, S.A. Keller, and D.A. Kosower, (2001), hep-ph/0104052.

[97] J. Pumplin *et al.,* Phys. Rev. **D65**, 014013 (2001), hep-ph/0101032.

[98] J. Pumplin, D.R. Stump, and W.K. Tung, Submitted to Phys. Rev. D [hep-ph/0008191].

[99] D. Stump *et al.,* Phys. Rev. **D65**, 014012 (2001), hep-ph/0101051.

[100] G. D'Agostini, "Bayesian reasoning in data analysis: A critical introduction", World Scientific, (2003).

[101] H.J. Behrend *et al.,* [CELLO Collaboration], Phys. Lett. **B183**, 400 (1987).

[102] G. D'Agostini, Nucl. Instrum. Meth. **A346**, 306 (1994).

[103] R.D. Ball *et al.,* JHEP 1005:075 (2010), hep-ph/0912.2276.

[104] H. PadÃ©, Ann. Ecole Nor. **3**, 9 (1892).

[105] H. Jung *et al.,* hep-ph/0903.3861.

[106] M. Botje *et al.,* hep-ph/1101.0538.

[107] S. Alekhin *et al.,* hep-ph/1101.0536.

[108] W. Giele *et al.,* hep-ph/0204316.

[109] M.R. Whalley *et al.,* hep-ph/0508110.

[110] D. Bourilkov *et al.*, hep-ph/0605240.

[111] LHAPDF: `http://hepforge.cedar.ac.uk/lhapdf/`.

[112] G. Watt, presented at the PDF4LHC meeting of March 26, 2010 (CERN). Plots from http://projects.hepforge.org/mstwpdf/pdf4lhc/ .

[113] S. Alekhin, J. Blumlein, S. Moch, (2012), hep-ph/1202.2281.

[114] H1 and ZEUS Collaborations, F.D. Aaron *et al.*, JHEP **1001**, 109 (2010), 0911.0884.

[115] NuTeV, M. Goncharov *et al.*, Phys. Rev. **D64**, (2001) 112006, hep-ex/0102049.

[116] BCDMS, A.C. Benvenuti *et al.*, Phys. Lett. **B223**, 485 (1989).

[117] BCDMS, A.C. Benvenuti *et al.*, Phys. Lett. **B237**, (1990) 592.

[118] A. Bodek *et al.*, Phys. Rev. **D20**, 1471 (1979).

[119] W. Atwood *et al.*, Phys. Lett. **B64**, 479 (1976).

[120] M. Mestayer *et al.*, Phys. Rev. **D27**, 285 (1983).

[121] J. Gomez *et al.*, Phys. Rev. **D49**, 4348 (1994).

[122] S. Dasu *et al.*, Phys. Rev. **D49**, 5641 (1994).

[123] G. Moreno *et al.*, Phys. Rev. **D43**, 2815 (1991).

[124] FNAL E866/NuSea, R.S. Towell *et al.*, Phys. Rev. **D64**, 052002 (2001), hep-ex/0103030.

[125] J.F. Owens, A. Accardi, W. Melnitchouk, (2012), hep-ph/1212.1702.

[126] L.W. Whitlow *et al.*, Phys. Lett. **B282**, 475 (1992).

[127] S.P. Malace *et al.*, Phys. Rev. **C80**, 035207 (2009).

[128] F.D. Aaron *et al.*, JHEP **1001**, 109 (2010).

[129] E.A. Hawker *et al.*, Phys. Rev. Lett. **80**, 3715 (1998); J. Webb, Ph.D. Thesis, New Mexico

[130] F. Abe *et al.*, Phys. Rev. Lett. **81**, 5754 (1998).

[131] D. Acosta *et al.*, Phys. Rev. **D71**, 051104(R) (2005).

[132] CDF Collaboration, T. Aaltonen *et al.*, Phys. Rev. Lett. **102**, 181801 (2009), 0901.2169.

[133] V. M. Abazov *et al.*, Phys. Rev. **D77**, 011106(R) (2008).

[134] V. M. Abazov *et al.*, Phys. Rev. Lett. **101**, 211801 (2008).

[135] CDF Collaboration, T.A. Aaltonen *et al.,* Phys. Lett. **B692**, 232 (2010), 0908.3914.

[136] D0 Collaboration, V.M. Abazov *et al.,* Phys. Rev. **D76**, 012003 (2007), hep-ex/0702025.

[137] T. Affolder *et al.,* Phys. Rev. **D64**, 032001 (2001).

[138] CDF Collaboration, T. Aaltonen *et al.,* Phys. Rev. **D78**, 052006 (2008), 0807.2204.

[139] B. Abbott *et al.,* Phys. Rev. Lett. **86**, 1707 (2001).

[140] CDF, T. Aaltonen *et al.,* Phys. Lett. **B692**, 232 (2010), 0908.3914.

[141] CDF - Run II, A. Abulencia *et al.,* Phys. Rev. **D75**, 092006 (2007), hep-ex/0701051.

[142] D0 Collaboration, V.M. Abazov *et al.,* Phys. Rev. Lett. **101**, 062001 (2008), 0802.2400.

[143] V.M. Abazov *et al.,* Phys. Lett. **B666**, 435 (2008).

[144] J. Gao *et al.,* (2013), hep-ph/1302.6246.

[145] V. Radescu, (2011), hep-ex/1107.4193.

[146] ZEUS-prel-10-017, H1prelim-10-141.

[147] P. Jimenez-Delgado, E. Reya, Phys. Rev. **D79**, 074023 (2009), hep-ph/0810.4274.

[148] P. Jimenez-Delgado, E. Reya, Phys. Rev. **D80**, 114011 (2009), hep-ph/0909.1711.

[149] C. Adloff *et al.,* (H1 Collaboration), Eur. Phys. J. **C13**, 609 (2000).

[150] C. Adloff *et al.,* (H1 Collaboration), Eur. Phys. J. **C19**, 269 (2001)

[151] C. Adloff *et al.,* (H1 Collaboration), Eur. Phys. J. **C21**, 33 (2001).

[152] C. Adloff *et al.,* (H1 Collaboration), Eur. Phys. J. **C30**, 1 (2003).

[153] S. Chekanov *et al.,* (ZEUS Collaboration), Eur. Phys. J. **C21**, 443 (2001).

[154] M.R. Adams *et al.,* Phys. Rev. **D54**, 3006 (1996).

[155] M.R. Adams *et al.,* Phys. Rev. Lett. **75**, 1466 (1995).

[156] NuSea, J.C. Webb *et al.,* (2003), hep-ex/0302019.

[157] W.T. Giele, S.A. Keller, and D.A. Kosower, hep-ph/0104052.

[158] S. Forte *et al.,* JHEP **05**, 062 (2002), hep-ph/0204232.

[159] NNPDF Collaboration, JHEP **03**, 080 (2005), hep-ph/0501067.

[160] J. Rojo and J.I. Latorre, JHEP **01**, 055 (2004), hep-ph/0401047.

[161] J. Rojo, JHEP **05**, 040 (2006), hep-ph/0601229.

[162] M.C. Gonzalez-Garcia, M. Maltoni, and J. Rojo, JHEP **10**, 075 (2006), hep-ph/0607324.

[163] F. Caola, S. Forte and J. Rojo, Nucl. Phys. **A854**, 32 (2011), 1007.5405.

[164] F. Caola, S. Forte and J. Rojo, Phys. Lett. **B686**, 127 (2010), 0910.3143.

[165] New Muon Collaboration, M. Arneodo *et al.*, Nucl. Phys. **B487**, 3 (1997), hep-ex/9611022.

[166] New Muon Collaboration, M. Arneodo *et al.*, Nucl. Phys. **B483**, 3 (1997), hep-ph/9610231.

[167] H1 Collaboration, F.D. Aaron *et al.*, Phys. Lett. **B665**, (2008) 139, 0805.2809.

[168] ZEUS Collaboration, J. Breitweg *et al.*, Eur. Phys. J. **C12**, 35 (2000), hep-ex/9908012.

[169] ZEUS Collaboration, S. Chekanov *et al.*, Phys. Rev. **D69**, 012004 (2004), hep-ex/0308068.

[170] ZEUS Collaboration, S. Chekanov *et al.*, Eur. Phys. J. **C63**, 171 (2009), 0812.3775.

[171] ZEUS Collaboration, S. Chekanov *et al.*, Eur. Phys. J. **C65**, 65 (2010), 0904.3487.

[172] H1 Collaboration, C. Adloff *et al.*, Phys. Lett. **B528**, 199 (2002), hep-ex/0108039.

[173] H1 Collaboration, F.D. Aaron *et al.*, Phys. Lett. **B686**, 91 (2010), 0911.3989.

[174] H1 Collaboration, F.D. Aaron *et al.*, Eur. Phys. J. **C65**, 89 (2010), 0907.2643.

[175] ZEUS Collaboration, S. Chekanov *et al.*, Eur. Phys. J. **C62**, (2009) 625, 0901.2385.

[176] ZEUS Collaboration, S. Chekanov *et al.*, Eur. Phys. J. **C61**, (2009) 223, 0812.4620.

[177] CHORUS, G. Onengut *et al.*, Phys. Lett. **B632**, (2006) 65.

[178] D.A. Mason, FERMILAB-THESIS-2006-01.

[179] J.C. Webb, (2003), hep-ex/0301031.

[180] M. Bonvini, S. Forte and G. Ridolfi, Nucl. Phys. **B847**, 93 (2011), 1009.5691.

[181] NNPDF Collaboration, R.D. Ball *et al.*, (2011), 1102.3182.

[182] S. Alekhin, J. Blumlein and S. Moch, Eur. Phys. J. **C71**, 1723 (2011), 1101.5261.

[183] F. Cerutti, (2011), 1107.1095.

[184] R.S. Thorne and G. Watt, JHEP 1108, 100 (2011), 1106.5789.

[185] G. Altarelli, S. Forte and G. Ridolfi, Nucl. Phys. **B534**, 227 (1998), hep-ph/9806345.

[186] C.M. Bishop, Neural Networks for Pattern Recognition (Oxford University Press, 1995).

[187] P.M. Nadolsky *et al.,* Phys. Rev. **D78**, 013004 (2008), 0802.0007.

[188] T. Kluge, K. Rabbertz and M. Wobisch, (2006), hep-ph/0609285.

[189] T. Carli *et al.,* (2009), 0911.2985.

[190] G.P. Salam and J. Rojo, Comput. Phys. Commun. **180**, 120 (2009), 0804.3755.

[191] T. Carli, G.P. Salam and F. Siegert, (2005), hep-ph/0510324.

[192] M. Dittmar *et al.,* (2005), hep-ph/0511119.

[193] R.D. Ball *et al.,* NNPDF Collaboration, Nucl. Phys. **B855**, 608 (2012), hep-ph/1108.1758.

[194] A.D. Martin *et al.,* Eur. Phys. J. **C70**, 51 (2010), 1007.2624.

[195] S. Alekhin and S. Moch, (2010), 1011.5790.

[196] J.M. Campbell, J.W. Huston and W.J. Stirling, Rept. Prog. Phys. **70**, 89 (2007), hep-ph/0611148.

[197] A. Sherstnev and R.S. Thorne, Eur. Phys. J. **C55**, 553 (2008), 0711.2473.

[198] H.L. Lai *et al.,* JHEP **04**, 035 (2010), 0910.4183.

[199] E. Laenen, Pramana **63**, 1225 (2004).

[200] S. Forte, G. Altarelli and R.D. Ball, Nucl.Phys.Proc.Suppl. **191**, 64 (2009), 0901.1294.

[201] J. Campbell and R.K. Ellis, Phys. Rev. **D65**, 113007 (2002), hep-ph/0202176.

[202] MCFM, http://mcfm.fnal.gov.

[203] H.L. Lai *et al.,* Phys. Rev. **D82**, 054021 (2010), 1004.4624.

[204] F. Demartin *et al.,* Phys. Rev. **D82**, 014002 (2010), 1004.0962.

[205] The ATLAS collaboration, JHEP 1012:060 (2010), 1010.2130.

[206] The CMS Collaboration, Phys. Lett. **B695**, 424 (2011), 1010.5994.

[207] A. Guffanti and J. Rojo, (2010), 1008.4671.

[208] J. Pumplin *et al.,* JHEP **02**, 032 (2006), hep-ph/0512167.

[209] R.D. Ball *et al.*, Chapter 21 in: J.R. Andersen *et al.*, "The SM and NLO multileg working group: Summary report", 1003.1241, (2010).

[210] G. Watt, JHEP **09**, 069 (2011), 1106.5788.

[211] J. Baglio, A. Djouadi, R. Godbole, 1107.0281.

[212] R. Bonciani, G. Degrassi and A. Vicini, JHEP **11**, 095 (2007), 0709.4227.

[213] U. Aglietti *et al.*,JHEP **01**, 021 (2007), hep-ph/0611266.

[214] S. Moch and P. Uwer, Phys. Rev. **D78**, 034003 (2008), 0804.1476.

[215] M. Aliev *et al.*, Comput. Phys. Commun. **182**, 1034 (2011), 1007.1327.

[216] The CMS Collaboration, S. Chatrchyan *et al.*, JHEP **07**, 049 (2011), 1105.5661.

[217] The CMS Collaboration, (2011), CMS-PAS-TOP-11-001.

[218] The ATLAS Collaboration, (2011), ATLAS-CONF-2011-040.

[219] C. Anastasiou *et al.*, Phys. Rev. **D69**, 094008 (2004), hep-ph/0312266.

[220] The ATLAS Collaboration, ATLAS-CONF-2011-041.

[221] The CMS Collaboration, CMS-PAS-EWK-10-005.

[222] The NuTeV Collaboration, Phys. Rev. Lett. **88**, 091802 (2002), hep-ex/0110059.

[223] S. Davidson *et al.*, JHEP **02**, 037 (2002), hep-ph/0112302.

[224] The NNPDF Collaboration, R.D. Ball *et al.*, (2011), 1110.1863

[225] S. Lionetti *et al.*, Phys. Lett. **B701**, 346 (2011), 1103.2369.

[226] S. Forte, Acta Phys. Polon. **B41**, 2859 (2010), 1011.5247.

[227] G. Bozzi, J. Rojo and A. Vicini, Phys. Rev. **D83**, 113008 (2011), 1104.2056.

[228] The NNPDF Collaboration, R.D. Ball *et al.*, Nucl. Phys. **B849**, 112 (2011), 1012.0836.

[229] S. Bethke, Eur. Phys. J. **C64**, 689 (2009), 0908.1135.

[230] LHC Higgs Cross Section Working Group, S. Dittmaier *et al.*, (2011), 1101.0593.

[231] S. Forte *et al.*, Nucl. Phys. **B643**, 477 (2002), hep-ph/0205286.

[232] M. Virchaux and A. Milsztajn, Phys. Lett. **B274**, 221 (1992).

[233] M. Arneodo *et al.*, Phys. Lett. **B309**, 222 (1993).

[234] NNPDF Collaboration, R.D. Ball *et al.,* Phys. Lett. **B707**, 66 (2012), 1110.2483.

[235] J. Pumplin, Phys. Rev. **D81**, 074010 (2010), 0909.0268.

[236] A.D. Martin *et al.,* (2011), http://indico.mppmu.mpg.de/indico/getFile.py/access?contribId=27&sessionId=8&resId=0&materialId=slides&confId=973 .

[237] M. Buza *et al.,* Nucl. Phys. **B472**, 611 (1996), hep-ph/9601302.

[238] J. Blumlein, A. De Freitas, W.L. van Neerven, S. Klein, Nucl. Phys. **B755**, (2006) 272, hep-ph/0608024.

[239] I. Bierenbaum, J. Blumlein, S. Klein, Nucl. Phys. **B780**, 40 (2007), hep-ph/0703285.

[240] J. Blumlein and S. Kurth, Phys. Rev. **D60**, 014018 (1999), hep-ph/9810241.

[241] J. Blumlein and S. Kurth, (1997), hep-ph/9708388.

[242] J. Blumlein *et al.,* (2011), 1104.3449.

[243] S. Alekhin and J. Blumlein, Phys. Lett. **B594**, 299 (2004), hep-ph/0404034.

[244] S. Alekhin and J. Blumlein, Phys. Lett. **B594**, 299 (2004), hep-ph/0404034.

[245] W. van Neerven and E. Zijlstra, Phys. Lett. **B272**, 127 (1991).

[246] W. van Neerven and E. Zijlstra, Phys. Lett. **B273**, 476 (1991).

[247] W. van Neerven and E. Zijlstra, Phys. Lett. **B297**, 377 (1992).

[248] S. Moch, J.A.M. Vermaseren and A. Vogt, Phys. Lett. **B606**, 123 (2005), hepph/0411112.

[249] J.A.M. Vermaseren, A. Vogt and S. Moch, Nucl. Phys. **B724**, 3 (2005), hepph/0504242.

[250] E. Laenen *et al.,* Nucl. Phys. **B392**, 162 (1993).