UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Fisica, Astrofisica e Fisica Applicata
Dipartimento di Fisica

Corso di Dottorato in Fisica, Astrofisica e Fisica Applicata
Ciclo XXX

Parton distribution functions with percent level precision

Settore Scientifico Disciplinare FIS/02

Tesi di Dottorato di:
Zahari KASSABOV

Supervisore: Prof. Stefano FORTE

Coordinatore: Prof. Francesco RAGUSA

Anno Accademico 2017-2018

Committee of the final examination:

External Referees:
Prof. Giampiero PASSARINO Prof. Alexander MITOV

External Members:
Prof. Nigel. GLOVER
Dr. Andre David TINOCO

I would like to thank my supervisor Stefano Forte for his always useful advice and for the many intriguing discussions.

Thanks to the theory group in Milan for providing me with such a pleasant environment, and particularly to Giancarlo Ferrera for always helping me with the most varied problems.

I was funded by the HiggsTools network which provided me with great opportunities to further my scientific career. Professors Nigel Glover and Giampiero Passarino were instrumental in making the network a success despite the many challenges, for which I am immensely grateful. I would also like to thank all the members of the network for making all the meetings memorable.

I have worked within the NNPDF collaboration for most of my research. I thank all its members for creating such an exciting environment where so many new ideas can materialize. Particular thanks to Stefano Carrazza, who was an NNPDF student in Milan with me the first year and has been of great help since my initial projects.

Finally I thank my mother Tania and my sister Isabel for supporting me during this time.

Abstract

This thesis is dedicated to the construction and applications of Parton Distribution Functions (PDFs), which are precise enough that are suitable for comparison with high-precision collider data from the LHC. We first review the theoretical framework and explain how PDFs should be used in practice, presenting tools that allow to seemingly convert between different representations of PDF uncertainties. We then discuss how these tools are used to construct the PDF4LHC15 combined sets, which implement the recommendation of the PDF4LHC group, and provide detailed benchmarks of each of the combined sets. We describe NNPDF 3.1, the first global PDF analysis to provide percent level uncertainties for many relevant observables while being validated by a closure test, with particular emphasis to the issues that appeared during its preparation and the resources used solve them. Finally we present a determination of the strong coupling constant based on the global NNPDF3.1 fit. We use a new methodology that correctly propagates all the uncertainties from the PDFs to the result.

Contents

Chapter 1

Introduction

After the discovery of the Higgs Boson [1, 2] the main focus of the Large Hadron Collider (LHC) and indeed of the field of Particle Physics Phenomenology is to determine the properties of the Standard Model with enough precision that small deviations from it can be discovered in the experimental data. Since the most stringent tests of the Standard Model currently from experiments involving proton-proton collisions, it is of vital importance to attain a precise and accurate description of the structure of the proton, which in turn cannot be determined from first principles with the current understanding of the underlying theory, Quantum Chromodynamics (QCD). The structure of the proton is described in terms of Parton Distribution Functions (PDFs), which at the lowest order in perturbation theory constitute probability densities to find a given constituent of the proton (called parton, for example a quark or a gluon) carrying a given momentum fraction.

The techniques used to determine Parton Distribution Functions have evolved greatly since the early eighties, when PDFs were based on simple ad-hoc models with significative differences among themselves and no way to estimate uncertainties [3]. Nowadays the standard is analyses based on the fit to all physical processes where experimental data exists and the corresponding theoretical description to elucidate their implant on the PDFs is available. These analyses feature advanced theoretical treatments required to account for the effect of the heavy quark masses, and uncertainty estimates that take into account both the uncertainties of the input experimental data and those related to the selection of the model used to fit them (which is not given from first principles). The PDF sets constructed that fulfil all these criteria agree well within uncertainties [4], which are furthermore small enough to allow for quantitative tests of allow for rigorous quantitative tests of the Standard Model in collider experiments.

An important contribution to the progress made is due to the NNPDF collaboration [7]. NNPDF analyses parametrize the PDFs in terms of neural networks, and employ a methodology that propagates the uncertainties in a conceptually simple way while minimizing the assumptions on the parametrization. The methodology is tuned to satisfy a closure tests that ensures its self consistency under very general assumptions [8]. The software used in the fits allows to perform the convolution operation in a quick way [9], allowing to tackle problems that would otherwise be computationally intractable. The heavy quark mass scheme used to implement the partonic evolution, so

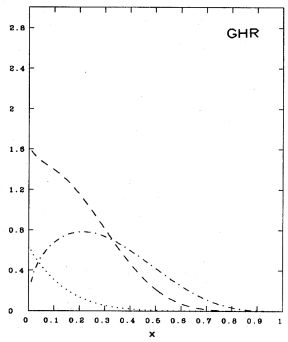


FIG. 25. Parton distributions of Glück, Hoffmann, and Reya (1982), at $Q^2=5$ GeV^2: valence quark distribution $x[u_v(x)+d_v(x)]$ (dotted-dashed line), $xG(x)$ (dashed line), and q_v (dotted line).

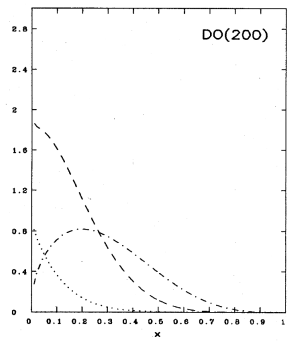
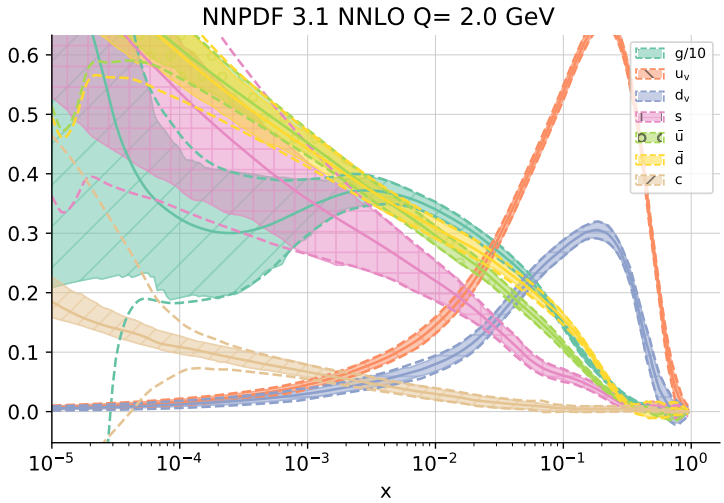


FIG. 27. "Soft-gluon" ($\Lambda=200$ MeV) parton distributions of Duke and Owens (1984) at $Q^2=5$ GeV^2: valence quark distribution $x[u_v(x)+d_v(x)]$ (dotted-dashed line), $xG(x)$ (dashed line), and $q_v(x)$ (dotted line).



Comparison between some early characterizations of parton densities [5, 6], from Ref. [3] (up) and the latest results from the NNPDF determination (down)

called FONLL [10], can be extended to account for charm initiated contributions and fit them explicitly [11]. The latest iteration of the NNPDF global analysis, NNPDF 3.1 [12], combines these features with the addition of new high precision data from the LHC experiments, resulting in significative improvements in precision and accuracy.

Some open problems in PDF determinations include:

- The consistent treatment of *theoretical uncertainties* (related to the fact that the theory used to analyse the data is only known in a perturbative approximation), and that are currently not included in the determinations.

- Characterization of experimental systematics in precise measurements: When the systematic uncertainties dominate over the statistical ones, the experimental covariance matrices can become near singular, so that a small change in the

treatment of the systematics (consistent with the precision at which they are determined) has a great impact on the result.

- Search for better fitting procedures: The recent development in Machine Learning to deal with problems that have similarities with PDF fitting suggests that more advances procedures can be developed allowing for faster fits and better control of the procedure.

Overview

Chapter 2: Review of the theoretical framework

We discuss shortly the main theoretical results that underpin a PDF determination. While the treatment is superficial, it aims to provide a more formal, yet intuitive, understanding of *what* is a PDF and how is it constructed from theory.

Chapter 3: PDFs for practical uses

The main motivation for developing precise PDFs is their usage in the context of experimental analyses to measure properties of the Standard Model, possibly finding deviations from it, that may indicate New Physics. To that end, the PDFs must be packaged in a way that facilitates the computation of PDF depended quantities and their uncertainties. In particular, it is preferable that PDFs provide the so called *Hessian Uncertainties* in many experimental settings. In Sec. 3.3, we discuss a method [13, 14] to transform the *Monte Carlo PDFs* that are obtained natively in the NNPDF fitting procedure into equivalent Hessian PDF sets. A further challenge is that frequently, the same computation has to be repeated for all of the *error sets* that characterize the uncertainties of the PDFs. The Hessian conversion algorithm has the side effect of reducing the number of error sets of the resulting PDF, compared to the starting Monte Carlo one for an equivalent statistical accuracy (at least when some assumptions on the Gaussianity of the starting error set hold). However the prospective reduction of error sets is interesting enough to be developed further: It may be a worthwhile trade-off to construct some specialized PDF sets that make some assumptions on which use cases they will be employed for in exchange for a speedup in the computations. This idea is implemented by the SM-PDF algorithm [14], discussed in Sec 3.4.

Chapter 4: The PDF4LHC recommendation

Apart from NNPDF, several other groups regularly perform PDF determinations. In practice, it is often recommended that the predictions from several PDFs from different collaborations are used when assessing the uncertainty of a results such as the Higgs Cross section. A natural question is then which PDF sets to consider and how to combine their predictions. The PDF4LHC working group produces guidelines on how to treat PDFs. The 2015 recommendation [4] recognized the improved understanding of the determinations of PDFs and the consequent improvement in the agreement between the results from several independent collaborations. As a result, it provided a less conservative prescription for the uncertainty computation based on the delivery of combined PDF sets.

The Monte Carlo conversion method introduced in the previous subsection and described in Sec. 3.3 was used in the elaboration of these combined sets, which were benchmarked in detail in Ref. [15].

The improved agreement between collaborations and the altered prescription for the combination had the combined effect of bringing down the PDF uncertainties (e.g. by a factor of two for the Higgs production in gluon fusion [16]).

Chapter 5: NNPDF 3.1

The NNPDF 3.1 PDF set [12] builds upon the methodology constructed for the NNPDF 3.0 sets [8], and specifically on the fact that it is optimized based on *closure tests*. The new features in NNPDF 3.1 include the addition of new high precision collider data, that collectively constrain significatively the results (thereby increasing the precision), includes a consistent theoretical treatment of the charm PDF (thereby increasing the theoretical accuracy). As a consequence of the more stringent precision targets, the numerics have also been tested more carefully and verified to not distort the results.

All these improvements make NNPDF 3.1 able to reliably predict phenomenologically relevant physical observables like the W and Z total cross sections with PDF uncertainties below 1%. As the NNPDF project grows more complex the software tools that underpin it need to by upgraded to handle it appropriately. The required improvements include enhanced flexibility of the code so that the different parts of the methodology can be verified to function correctly and to the required precision, better performance or easier deployment in diverse High Performance Computing facilities. In 5.5, we discuss the development from the ground of a programming framework that is suited to the particular characteristics of scientific computing, together with an NNPDF-specific application based on it. This framework was used extensively in NNPDF 3.1 and the determination of α_s.

Chapter 6: A determination of the strong coupling constant

The improvements leading to the NNPDF 3.1 set make it unprecedented in terms of the precision and accuracy it achieves. This progress can then be translated to other quantities that depend on the PDFs. A prime example is the strong coupling constant, α_s. A determination of α_s based on NNPDF 3.1 and featuring a new methodology is presented here.

Original research in this thesis

Much of the research has been done in collaboration with other colleagues, particularly from the NNPDF collaboration. This makes it hare. In general, wherever results are presented, I have reported here in more detail the research where I feel I have made a significative contribution, and discussed parts of our research where I contributed less as necessary for context. In particular I believe I made a large contribution to the design, implementation and benchmarking of the PDF transformation methods discussed in Sec. 3.3 and Sec. 3.4, the benchmarks presented Sec. 4.4, the analyses in Sec. 5.4, the development of coding framework in Sec 5.5, as well as several other parts

of the NNPDF code which were helpful to obtain the results in Chapter 5, and the determination of α_s presented in 6. On the other hand, several more technical results in already published work (particularly in Refs. [13, 4, 14, 15, 12]) were not included here if they would have implied merely transcribing the, without providing additional useful context to the discussion.

I have produced all the figures where the source is not explicitly stated in the caption.

Chapter 2

Review of the theoretical framework

We review some fundamental aspects of Quantum Chromodynamics (QCD), particularly those related with the determination of PDFs.

2.1 Fundamentals of QCD

QCD is a gauge field theory where the gauge group is $SU(3)$. The gauge bosons of the theory are called *gluons* and are massless. The fermions are called *quarks* and have fractional electric charge (either $2/3$ or $-1/3$ for the quarks, and the opposite sign for the antiquarks). There are three *families* each containing a pair of quarks with their corresponding antiquarks. The corresponding classical Lagrangian, which is invariant under $SU(3)$ transformations is given in terms Yang-Mill Lagrangian density

$$\mathcal{L}_{\text{classical}} = \sum_{\text{flavours}} \bar{\psi}_a (i\gamma_\mu D^\mu - m)_{ab} \psi_b - \frac{1}{4} G^A_{\mu\nu} G^{\mu\nu}_A , \tag{2.1.1}$$

where ψ_a are the quark fields, D^μ is the covariant derivative

$$D^\mu = \partial^\mu + igt^a A^\mu_a \ , \tag{2.1.2}$$

where in turn we have introduced the gluon field, A^μ_a and the matrices t^a, corresponding to the eight generators of the $SU(3)$ color group in the fundamental representation. Arbitrarily, we define their norm to satisfy $\text{tr}(t^a t^b) = \frac{1}{2}\delta^{ab}$. $G^a_{\mu\nu}$ is the field strength tensor, which can be defined in terms of the gluon fields and the *structure constants* of the $SU(3)$ group, f^{abc} as

$$G^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g f^{abc} A^b_\mu A^c \nu \ . \tag{2.1.3}$$

The parameter g is the bare coupling of the theory.

There exist multiple pieces of experimental evidence that such theory can in fact explain the strong interactions. Starting from 1963 Gell-Mann, Ne'eman and Zweig showed that a model exhibiting the symmetries of QCD could be used to explain compactly the then puzzling proliferation of new particles that were being found in nuclear experiments [17, 18, 19, 20]. In the following sections, we shall recognize the quantitative features of perturbative QCD that have been tested experimentally to a high degree of precision, particularly regarding the description of Deep Inelastic Scattering (DIS).

2.2 The QCD coupling constant

We define the strong coupling constant in terms of the coupling from Eq. 2.1.1

$$\alpha_s = \frac{g^2}{4\pi} \tag{2.2.1}$$

As we depart from the classical Lagrangian and consider the quantum version of the theory, the couplings must be altered so that they absorb the unphysical dependence of the renormalization scale; thus, the renormalization of implies that strength of the coupling *runs*; that is, it depends on the energy scale of the process in which it enters, given by the parameter μ^2, with units of squared energy. The dependence is determined by the renormalization group equation (Callan-Symanzik):

$$\mu^2 \frac{d}{d\mu^2}\alpha_s(\mu^2) = \beta(\alpha_s(\mu^2)) \ , \tag{2.2.2}$$

where the β function admits a perturbative expansion in α_s, of the form

$$\beta(\alpha_s) = -\alpha_s^2(\beta_0 + \beta_1\alpha_s + \beta_2\alpha_s^2 + ...) \tag{2.2.3}$$

At leading order, the running is determined by the β_0 coefficient, which is

$$\beta_0(\alpha_s) = \frac{33 - 2n_f}{12\pi} \ , \tag{2.2.4}$$

where n_f is the number of flavours that are light at the scale μ^2. Since $n_f < 17$ at any scale in QCD, β function is negative, implying that the strength of the coupling increases as the scale of the interaction decreases. This property of QCD is known as *asymptotic freedom* [21, 22].

At leading order, the solutions to the Renormalization Group Equation are given in terms of one parameter, Λ, known as the *QCD scale*.

$$\alpha_s = \frac{1}{b_0 \log \frac{\mu^2}{\Lambda^2}} \tag{2.2.5}$$

The value of Λ is not given by the theory and must therefore be determined experimentally. Since it marks the energy scale at which Eq. 2.2.5 becomes infinity (note that leading order approximation is entirely inadequate as a consequence), it can be used as a rough estimate of the scale at which the perturbative description of the theory breaks down. Equivalently, it is possible to parametrize the running in Eq. 2.2.5 in terms of the value of the coupling constant at some arbitrary fixed scale Q_0^2. We have then:

$$\alpha_s(\mu^2) = \alpha_s(Q_0^2)\left(1 - \alpha_s(Q_0^2)\beta_0 \log\frac{\mu^2}{Q_0^2} + \mathcal{O}(\alpha_s^2)\right) \tag{2.2.6}$$

thus, in principle it is enough to measure the strong coupling at one scale, and it can then be related to any other trough renormalization scaling. The values of the strong coupling are usually tabulated at the scale of the mass of the Z boson, $M_Z^2 \approx 91.2GeV^2$. This parametrization is more useful than the one in Eq. 2.2.5 beyond leading order, since Λ becomes dependent on the choice of renormalization scheme.

The scaling of the strong coupling constant has been measured in the range of around $1 - 200GeV$ and found to agree with the experimental data. A review of the determinations of α_s and evidence can be found in Ref. [23]. I present an updated determination based on the NNPDF framework in Chapter 6.

2.3 Characterization Deep Inelastic Scattering

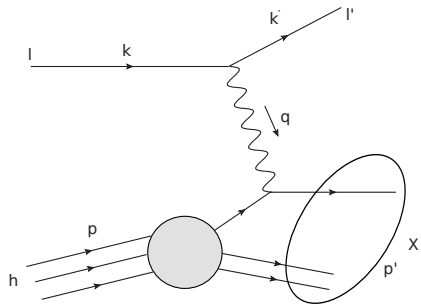


Figure 2.1: The DIS process. An incoming lepton l and four momentum k scatters off an hadron h, that is broken into the arbitrary final state X. The measured leptonic final state is l'.

The history of the quantitative predictions of QCD begins in the 1970s with the establishment of the approximate scaling behaviour in Deep Inelastic Scattering (DIS). Given its particular importance in the determination of PDFs, as well as in establishing the concept in the first place, we revise it here in enough detail , but making simplifying assumptions where they won't affect the argument.

A deep inelastic scattering experiment consists on the measurement of the scattering of a beam of leptons off an hadronic target (that is, the process $l + h \longrightarrow l' + X$). A lepton interacts with an hadron though the exchange of a virtual vector boson. In what follows, we assume that the interaction is mediated by a photon, for simplicity (in particular, we assume that the interaction conserves parity since it is purely electromagnetic). The target hadron absorbs the virtual particle and produces a final state X. If the target hadron remains intact after the interaction (that is $X = h$), the process is said to be elastic. Instead, the deep inelastic regime occurs when the hadron is fragmented into many particles. Usually the final state of the lepton l' is measured, but not that of the hadron.

Assuming we are considering the rest frame of the target hadron and neglecting the masses of the incoming and outgoing leptons, compared to the other energy scales in the process, the kinematics, depicted in Figure. 2.1 are described in terms of the following variables:

M The mass of the target hadron.

E The energy of the incoming lepton.

k The momentum of the incoming lepton.

E' The energy of the outgoing lepton l'.

k' The momentum of the outgoing lepton.

p The momentum of the target hadron.

$q = k - k'$, the momentum transfer.

$\nu = E - E'$, the energy loss of the lepton.

$y = \nu/E$, the fractional energy loss of the lepton.

$Q^2 = -q^2 = -(k - k')^2$

$M_X^2 = (p + q)^2$, the invariant mass of the final hadronic state.

$x = Q^2/2M\nu = Q^2/2pq = Q^2/2ME$

The kinematic characterization of the process can be found in e.g. Ref. [24]. The variable x was introduced by Bjorken [25] and is central to the understanding of the DIS process in QCD, as we shall see. In particular, QCD predicts that at leading order, the cross section is a function of x and independent of Q^2, in the large Q^2 limit. This property is known as *scaling*. We note that we can write x as

$$x = 1 - \frac{M_X^2 - M^2}{2pq} \ , \tag{2.3.1}$$

thus, $x = 1$ implies $M_X^2 = M^2$, and therefore elastic scattering. Since the baryon number is conserved, we must have that $M_X^2 > M$, and consequently $x \leq 1$. But since both Q^2 and ν are positive, x must be positive as well, so $x \geq 0$. Similar arguments apply to y (i.e. noting that $E' \leq E$). Therefore, we have

$$0 \leq x \leq 1; \ 0 \leq y \leq 1 \ . \tag{2.3.2}$$

Since most experiments are performed with no sensitivity to the polarization of the incoming and outgoing leptons, we will restrict ourselves to the discussion of the spin averaged case. Thus, neglecting spin labels, and in the Feynaman gauge, the amplitude of the process is given by

$$\mathcal{M} = ie^2 \bar{u}(k')\gamma^\mu u(k) \left(i\frac{g_{\mu\nu}}{Q^2} \right) \langle X|J_h^\nu|P\rangle \ , \tag{2.3.3}$$

where J_h^ν is the hadronic current. The fundamental difficulty in the computation of the process is the fact that we ignore the wavefunctions for the hadronic initial and final states ($|X\rangle$ $|P\rangle$), since they are non computable in perturbation theory, due to the large value of the strong coupling constant at energy scales comparable to the mass of the hadrons. We can study the problem better by factorizing it into a part that is computable in perturbation theory ($L_{\mu\nu}$) and a part that depends on the hadronic structure ($W^{\mu\nu}$); so that we can write

$$\left|\bar{\mathcal{M}}\right|^2 = \frac{1}{Q^2} L_{\mu\nu} W^{\mu\nu} \ . \tag{2.3.4}$$

It is easy to obtain the leptonic tensor from Eq.2.3.3 under the assumptions we are using (including all the spin contributions and neglecting the masses of the leptons):

$$L_{\mu\nu} = e^2 \sum_{\text{spin}} \bar{u}(k')\gamma_\mu u(k)\bar{u}(k)\gamma_\nu u(k') \, , \tag{2.3.5}$$

$$= e^2 \mathrm{tr}[\not{k}'\gamma_\mu \not{k}\gamma_\nu] \tag{2.3.6}$$

$$= 4e^2[k_\mu k'_\nu + k_\nu k'_\mu - g_{\mu\nu}(kk')] \ . \tag{2.3.7}$$

The hadronic part is formally given by summing over all possible final states:

$$W^{\mu\nu} \sim \sum_X \left\langle P(p)|J_h^{\mu\dagger}|X\right\rangle \langle X|J_h^{\nu}|P(p)\rangle \ , \tag{2.3.8}$$

$$\sim \left\langle P(p)|J_h^{\mu\dagger}J_h^{\nu}|P(p)\right\rangle \tag{2.3.9}$$

While $W^{\mu\nu}$ it is not computable from first principles, its tensor structure is constrained by the symmetries of the theory. In particular, by requiring that the tensor is symmetric under parity transformations and imposing the conservation of the hadronic current, $q_\mu W^{\mu\nu} = q_\nu W^{\mu\nu} = 0$, it is possible to find that the most general tensor structure is:

$$W_{\mu\nu} = F_1\left(-g_{\mu\nu} + \frac{q_\mu q_\nu}{q^2}\right) + \frac{F_2}{pq}\left(p_\mu - q_\mu\frac{p.q}{q^2}\right)\left(p_\nu - q_\nu\frac{p.q}{q^2}\right) \tag{2.3.10}$$

where the (Lorentz scale) coefficients F_1 and F_2 are called *structure functions*. More general structures are possible if we consider a parity violating interaction (mediated by a W boson), giving rise to a third F_3 structure function. Also when the spin of the incoming and outgoing leptons is determined, spin dependent structure functions contribute to the polarized cross section [24].

We can isolate the different components of the hadronic tensor by making suitable projections that yield a specific dependence on the structure functions. For example we may make the projection $p^\mu p^\nu W_{\mu\nu}$ (equivalent to assuming that the absorbed virtual photon is longitudinally polarized). We use the result to define the longitudinal structure function,

$$F_L = F_2 - 2xF_1 = \frac{Q^4}{(pq)^3}p^\mu p^\nu W_{\mu\nu} \tag{2.3.11}$$

As we shall see explicitly, at leading order we have that $F_L = 0$. This is the Callan-Gross relation [26] and provides evidence that the quarks are spin $1/2$ particles.

We can similarly extract F_2 by projecting in the direction of a vector n satisfying $p.n = 1$ $n.q = 0$, $n^2 = 0$. Neglecting the mass of the proton, we also have $p^2 = 0$. Then

$$F_2 = (p.q)n^\mu n^\nu W_{\mu\nu} \tag{2.3.12}$$

2.4 DIS in the parton model

The argumentation in this section is based on Ref. [27].

So far we have only made simplifying assumptions on the form of the hadronic tensor, $W^{\mu\nu}$. We now introduce some *model assumptions* based on QCD, which will allow us to establish some elemental properties. In particular, one of the elemental components in the establishment of the QCD as a theory for the strong interactions was the proposal of the *parton model* [28]. Fundamentally, it states that the hadron is made of individual components, called *"partons"*, and that an interaction at sufficiently high energy probes directly an interaction with the partons, which can be considered approximately free and on-shell (indeed a more precise characterization of the phrase "sufficiently high energy" is "high enough that the binding energy that holds the partons inside the hadrons can be neglected").

The parton model suggests that the hadronic tensor admits a probabilistic interpretation. It is given in terms of Parton Distribution Functions (PDFs) which encode the probability of the hard boson interacting with a parton. That is, instead of interacting with the "whole" hadron with momentum p we are interacting with a parton of type i and momentum ξp with infinitesimal probability $f_i(\xi, Q^2)d\xi$. Here ξ is a momentum fraction (thus $0 < \xi < 1$), and we are neglecting the transverse components of the momentum of the parton on the grounds that we are studying an interaction in the limit where the transferred energy is much higher than the binding transverse momentum. By definition, the total longitudinal momentum of the hadron is the sum of the momenta of the individual partons (since we assume they are non interacting)

$$\sum_{i}^{\text{partons}} \int_0^1 d\xi \xi f_i(\xi, Q) = 1 \ . \tag{2.4.1}$$

PDFs must also yield the quantum numbers that characterize the hadron. For a proton, we have:

$$\int_0^1 d\xi (u(\xi, Q) - \bar{u}(\xi, Q)) = 2 \ , \tag{2.4.2}$$

$$\int_0^1 d\xi (d(\xi, Q) - \bar{d}(\xi, Q)) = 1 \ , \tag{2.4.3}$$

$$\int_0^1 d\xi (q(\xi, Q) - \bar{q}(\xi, Q)) = 0, \ q = s, c, b, t \ . \tag{2.4.4}$$

Relations 2.4.1-2.4.4 are known as *sum rules* and hold to all orders in perturbation theory.

Now we can view the hadronic interaction as a probabilistic sum of the possible parton level interactions with the total proton momentum p is replaced by ξp:

$$W_{\mu\nu} = \int_0^1 \frac{d\xi}{\xi} \sum_{i}^{\text{partons}} f_i(\xi, Q^2) \widetilde{W}^i_{\mu\nu}(\xi, Q^2) \ , \tag{2.4.5}$$

where we have introduced the parton tensors $\widetilde{W}^i_{\mu\nu}$. The factor $1/\xi$ is necessary because the proton states are conventionally normalized to $2p^0$ while the hadron states are normalized to $2\xi p^0$. Thus the factor $1/\xi$ converts the parton flux to the correct normalization for the proton flux.

The parton level tensor must obey the same symmetries as the hadron level tensor $W^{\mu\nu}$. Therefore, under the same assumptions that were made in Eq. 2.3.10, the form of $\widetilde{W}^i_{\mu\nu}$ is

$$\widetilde{W}^i_{\mu\nu} = \widetilde{F}^i_1 \left(-g_{\mu_\nu} + \frac{q_\mu q_\nu}{q^2} \right) + \xi^2 \frac{\widetilde{F}^i_2}{pq} \left(p_\mu - \frac{p.qq_\mu}{q^2} \right) \left(p_\nu - \frac{p.qq_\nu}{q^2} \right) \ . \tag{2.4.6}$$

We have now introduced the parton level structure functions $\widetilde{F}^i_1(x, Q^2)$ and $\widetilde{F}^i_2(x, Q^2)$. They are also related to the hadron level structure functions F_1 and F_2 trough the PDFs,

$$F_J(x, Q^2) = \int_0^1 \frac{d\xi}{xi} \sum_{i}^{\text{partons}} \widetilde{F}^J_1(x, Q^2), \ J = 1, 2, L \ . \tag{2.4.7}$$

The crucial advantage in this characterization is that the parton level structure functions are now computable in perturbation theory. The parton level tensor $\widetilde{W}^i_{\mu\nu}$ is the spin-averaged squared amplitude for the partonic subprocess in Fig. 2.1. The corresponding matrix element is

$$\mathcal{M}_\mu = -ie_{q_i}\bar{u}(l)\gamma^\mu u(\xi p) \ . \tag{2.4.8}$$

We note that this is now analogous to the leptonic tensor from Eq. 2.3.5, with the replacements $k \to \xi p$, $k' \to l'$, $q \to -q$, and $e^2 \to e^2_{q_i}$. Then, by comparing the tensor structures in Equations 2.3.5 and 2.4.6, neglecting again the quark masses and applying the projections Eqs. 2.3.12 and 2.3.11, it is possible to arrive at

$$\widetilde{F}^i_2 = 2e^2_{q_i}\delta(l^2) \ . \tag{2.4.9}$$

The above is only non-zero when

$$l^2 = (\xi p + q)^2 = 2\xi pq - Q^2 = 0 \ , \tag{2.4.10}$$

and therefore

$$\xi = Q^2/2pq = x \ . \tag{2.4.11}$$

We have found that at leading order, we can identify the Bjorken variable x as the momentum fraction of an incoming parton.

Similarly, we find

$$\widetilde{F}^i_L = 0 \ , \tag{2.4.12}$$

and

$$\widetilde{F}^i_1 = 2e^2_{q_i}\frac{\xi^2}{x}\delta(\xi - x) \ . \tag{2.4.13}$$

We have now studied all the parts needed to write the electromagnetic structure functions of a proton at leading order: Combining Eq. 2.4.7 with Eqs. 2.4.13 and 2.4.9, and inserting the correct phase space factor, we arrive at

$$F_2(x, Q^2) = 2x \sum_i^{\text{partons}} f_i(x)e^2_{q_i} \ , \tag{2.4.14}$$

$$F_1(x, Q^2) = \sum_i^{\text{partons}} f_i(x)e^2_{q_i} \ . \tag{2.4.15}$$

As we anticipated, we find the Callan-Gross relation at leading order, $F_2 = 2xF_1$ since $F_L = 0$. The result that $F_L = 0$ encodes the fact that a spin $1/2$ particle cannot absorb a longitudinally polarized vector boson.

2.5 Higher order corrections and factorization

The results in Equations. 2.4.13, 2.4.9 and 2.4.12 can be interpreted as the contributions to the coefficient functions $C_{ij}(x, Q^2)$ at leading order. The coefficient functions encode the partonic cross sections and admit expansions in perturbative QCD,

$$C_{ij}(x, Q^2) = \widetilde{F}_i(x, Q^2)\delta_{ij} + \alpha_s C^{(1)}_{ij}(x, Q^2) + \alpha^2_s C^{(2)}_{ij}(x, Q^2) + \dots \tag{2.5.1}$$

When accounting for higher order corrections in QCD, we have to consider both loop ("virtual") corrections and additional emissions of new partons ("real corrections"). The QCD corrections have both ultraviolet (UV) and infrared (IR) divergences. In particular, the loop diagrams are affected by both; the ultraviolet divergences are dealt with following some renormalization procedure. The IR loop divergences exist because of the assumption that the partons are massless. These IR singularities cancel out when combined with real emission diagrams. This result holds for all the Standard Model, and is known as the Kinoshita-Lee-Nauenberg theorem [29, 30]. However real emission diagrams also introduce collinear singularities (that is, the transverse momentum of the new emitted particle tends to zero), that do not cancel out trivially. We can postulate that it is possible to treat these infinites in a similar way as the renormalization cures the UV divergences: by reabsorbing them into a redefinition of some *bare* quantity. In this case, the bare quantities are the PDFs, and they are modified in such a way that they correspond to finite measurable quantities. The PDFs now depend on an new energy scale μ (like we had already written but not described) and we can define the structure functions to all orders in perturbation theory as finite quantities given by

$$F^{(i)}(x,Q^2) = \sum_{j}^{\text{partons}} \int_x^1 \frac{dy}{y} C_{ij}(\frac{x}{y}, \alpha_s(Q^2), \mu^2) f_j(y,\mu^2) \ . \tag{2.5.2}$$

We can express this more succinctly introducing the convolution operator $\otimes$, which shall be defined by

$$F^{(i)}(x,Q^2) = \sum_{j}^{\text{partons}} C_i(x,Q^2) \otimes f_i(x,\mu^2) \ , \tag{2.5.3}$$

that is,

$$f(x) \otimes g(x) \equiv \int_x^1 \frac{dy}{y} f\left(\frac{x}{y}\right) g(y) \ . \tag{2.5.4}$$

The fact that Eq. 2.5.2 holds is by no means trivial. We have taken a number of items for granted along the way: The first one is the fact that the divergent part factorizes in a way that is independent of the observable. The second is that the collinear corrections do not apply to the interference terms between the partonic and the non perturbative hadronic part of the calculation.

It turns out that both of these features are true for a large class of processes. This result is known as the *Collinear factorization theorem*, and is described in detail in Ref. [31]. The reason why we can neglect the interference terms for sufficiently hard interactions is that they are suppressed by powers of the hard scale Q^2. The terms are commonly called *higher twist*.

Notice that in Eq.2.5.3 the right hand side of the has a term that depends on the arbitrary scale μ^2, while the left hand side does not depend on it. This is as expected, since observable quantities must not depend on this arbitrary choice. In the following section we will explicitly construct PDFs that have this property. Here we remark that we can choose how to split the finite contributions of μ dependence between the PDF and the coefficient function. This choice does not affect the results at any order in perturbation theory, and is called *factorization scheme*. The most common choice is the $\overline{\text{MS}}$ scheme, where the finite counterterms that are kept in the coefficient

functions are process independent. For example, at next-to-leading order (NLO), the finite counterterms are $\log 4\pi - \gamma_E$, (where γ_E is the Euler-Mascheroni constant).

2.6 DGLAP evolution

We obtain the scale dependence of the PDFs by requiring that the structure functions in Eq. 2.5.3 are independent of the arbitrary factorization scale choice μ^2.

$$\mu \frac{d}{d\mu} F(x, Q^2) = 0 \ . \tag{2.6.1}$$

This condition leads to renormalization group equations for the PDFs and coefficient functions that are solved in terms of the Altarelli-Parisi splitting functions P_{ij}:

$$\mu \frac{d}{d\mu} f_i(x, \mu^2) = \sum_j P_{ij}(x, \alpha_s(\mu)) \otimes f_j(x, \mu^2) \tag{2.6.2}$$

$$\mu \frac{d}{d\mu} C_i(x, Q^2/\mu^2, \alpha_s) = -\sum_j P_{ij}(x, \alpha_s(\mu)) \otimes C_j(x, Q^2/\mu^2, \alpha_s) \tag{2.6.3}$$

These relations are known as Altarelli-Parisi or DGLAP equations [32, 33, 34], and the results can be proven using the operator product expansion formalism. Note that the coupling constant is evaluated at the factorization scale. The splitting functions P_{ij} are known up to NNLO [35, 36].

Since the rank of the evolution matrix P_{ij} is not maximal, there are several subspaces of flavour combinations that are preserved by the evolution. This is a consequence of the flavour symmetry that QCD exhibits in the limit where we neglect the quark masses; For example, a gluon only splits into a quark and an antiquark of the same flavour. The thirteen partons (6 quarks, 6 antiquarks and the gluon). We give a basis that transforms the flavours in the PDFs so as to make the evolution operator P_{ij} as diagonal as possible. We first define:

$$q_i^{\pm} = q_i \pm \bar{q}_i \ . \tag{2.6.4}$$

because of the baryon number conservation, the combinations with a negative sign, called *valences* are preserved, and thus decouple. Similarly we can define *triplet* combinations:

$$T_3 = u^+ - d^+ \,, \tag{2.6.5}$$
$$T_8 = u^+ + d^+ - 2s^+ \,, \tag{2.6.6}$$
$$T_{15} = u^+ + d^+ + s^+ - 3c^+ \,, \tag{2.6.7}$$
$$T_{24} = u^+ + d^+ + s^+ + c^+ - 4b^+ \,, \tag{2.6.8}$$
$$T_{35} = u^+ + d^+ + s^+ + c^+ + b^+ - 5t^+ \,. \tag{2.6.9}$$

Instead, the singlet distribution,

$$\Sigma = \sum_i^{\text{quarks}} q^+ \tag{2.6.10}$$

can couple with the gluon. Therefore, for the non-singlet sector composed by valences and triplets, we have

$$\mu\frac{d}{d\mu}f_i^{\rm NS}(x,\mu^2) = P_i^{\rm NS}(x,\alpha_s)\otimes f(x,\mu^2) \tag{2.6.11}$$

while for the gluon and the singlet,

$$\mu\frac{d}{d\mu}\begin{pmatrix} g \\ \Sigma \end{pmatrix}(x,\mu^2) = \begin{pmatrix} P_{gg} & P_{g\Sigma} \\ P_{\Sigma g} & P_{\Sigma\Sigma} \end{pmatrix}(x,\alpha_s)\otimes\begin{pmatrix} g \\ \Sigma \end{pmatrix}(x,\mu^2) \tag{2.6.12}$$

We have now established that the scale dependence on the PDFs can be computed in perturbative QCD given an initial condition. In practice, the solution to the DGLAP equations are implemented in numerical codes trough x space integrations. Examples of such code include HOPPET [37], QCDNUM [38] or APFEL [39], which is currently employed in the NNPDF fits. It is also possible to solve the DGLAP equations in *Mellin* space, where they are trivial, since a Mellin transform of Equation 2.6.2 transforms the convolution integrals into products. The trade-off is then recovering the PDFs in x space, which requires a numerically involved inversion procedure. The QCD-PEGASUS [40] implements this approach.

2.7 Treatment of heavy flavours

So far we have been consistently making the approximation that the quarks participating in the PDF evolution are massless. Since the masses of the three lightest quarks, u, d, s are far below Λ, this approximation is entirely reasonable for them. The treatment of the rest of the quarks is more delicate, particularly since we are interesting in studying the regimes where they masses are either smaller or bigger than the characteristic scale of the interaction, and therefore procedures that are able to interpolate all the regimes are needed.

There exit several *flavour number schemes* that incorporate the heavy quark effects under different assumptions. Depending on the relation between the mass of a heavy quark m_q and the scale at which we probe the PDF, Q, we can identify two limiting cases:

$m_q \ll Q$ In this case we can simply treat the heavy quark as another massless proton, so that it is perturbatively generated by the DGLAP evolution.

$m_q \gtrsim Q$ The heavy quark can be considered as a purely final state that does not participate in the evolution (since there is no energy to produce it). This then allows to consider fully the mass effects in the matrix elements of the final state.

The first of these limits is well realized in the so called Zero Mass Variable Flavour Number Scheme (ZM-VFNS): The heavy quark is treated as a massless parton above its mass threshold (typically chosen to coincide with the mass of the quark, defined in e.g. the $\overline{\rm MS}$ subtraction scheme), and a corresponding heavy quark PDF is introduced (which is set to zero below the threshold). The approximation becomes problematic near the mass threshold because coefficients of order $\log(m_q^2/Q^2)$ appear in the coefficient functions and constitute potentially large corrections [10].

Instead the *Fixed Flavour Number Schemes* (FFNS) are accurate in the second limit: The heavy quarks are treated as a purely final state particle and only the lighter partons are considered in the evolution. Instead the heavy quark enters into the coefficient functions; at lowest order, they incorporate the splitting of a gluon into pair $q\bar{q}$ of heavy quarks, considering the mass effects. At scales much greater than the mass, this approximation becomes unreliable because collinear logarithms of Q^2/m_q^2 are left unresummed by the DGLAP evolution.

The General Mass Variable Flavour Number Schemes (GM-VFNS) are procedures that attempt to interpolate between these two limits, so that the effects of the heavy quarks are accounted for at all scales.

The essential idea [41] is to switch from FFNS PDFs with n_f flavours considered in the evolution to PDFs with n_f+1 flavours, at the matching point $\mu = m_q$. The PDFs above and below the threshold are related order by order in α_s by the $(n_f+1)\times n_f$ transition matrix $A_{jk}(\mu/m_q)$,

$$f_j^{(nf+1)}(\mu \to m_q^+) = A_{jk} \otimes f_k^{(n_f)}(\mu \to m_q^-) \ . \tag{2.7.1}$$

Here the superindexes $+$ and $-$ indicate the direction of the limit.

The elements of the A_{jk} are known up to NNLO [42]. Requiring that the theoretical expressions of the structure functions are continuous at the matching threshold, we find

$$F(x,Q^2) = C_k^-(m_q/Q) \otimes f_k^-(Q) = C_j^+(m_q/Q) \otimes f_j(Q) \tag{2.7.2}$$

$$\equiv C_j^+ \otimes A_{jk}(m_q/Q) \otimes f_k^-(Q) \ . \tag{2.7.3}$$

Since the PDFs are continuous at the threshold, the coefficient functions must separately satisfy

$$C_k^-(m_q/Q) = C_j^+ \otimes A_{jk}(m_q/Q) \ . \tag{2.7.4}$$

This condition defines the *minimal GM-VFNS*: Since the matrix is not square, the equation is underspecified and one can redefine the scheme by swapping terms of $\mathcal{O}(m_q/Q)$ between the two sides.

A number of GM-VFNS variants exist: They include ACOT [43, 44, 45] TR [46] (which require the continuity of the derivatives of the structure functions) and FONLL [10]. The FONLL method provided a theoretical framework that improved the understanding of the differences and equivalences in the earlier methods (and in particular to establish its equivalence with several ACOT variants) and allows for extensions such as the description of charm initiated contributions [47].

2.8 Hadroproduction

The result of the collinear factorization in Sec 2.5 also holds for processes where to hadrons interact to yield a given final state [31]. This is a basic property of QCD and allows to relate experimental measurements obtained in an hadron collider such as the LHC to theory predictions obtained in Perturbation Theory. The experimental results are typically obtained for hadronic cross sections ($\sigma_{pp\to X}$) for a given final state X (typically taken to be inclusive in all hadronic particles accompanying the desired

event), while the theory predictions are usually computed for hard (partonic) quantities, $\hat{\sigma}_{ab\to X}$. Using the notation from Ref. [48], the basic structure of hadroproduction processes is:

$$\sigma_{pp\to X}(s, M_X^2) = \sum_{a,b} \int_{x_{\min}} dx_1 dx_2 f_a(x_1, M_X^2) f_b(x_2, M_X^2) \hat{\sigma}_{ab\to X}(x_1 x_2 s, M_X^2) \ . \tag{2.8.1}$$

Here the sum is taken over all possible constituents of the proton, including quarks, antiquarks and gluon, but also photon and leptons. The underlying assumption of the factorization theorem is that similarly to the DIS case, discussed in Sec. 2.5, the singularities associated with the partons in the initial state are universal for all processes, so that it is possible to absorb these singularities into the bare parton densities [49]. We may recast Eq. 2.8.1 in terms of the convolution (see Sec. 2.5) between a *partonic luminosity* dependent on the PDFs, and a hard *coefficient function*,

$$\sigma_{pp\to X}(s, M_X^2) = \int_\tau^1 \frac{dx}{x} \mathcal{L}(x) C\left(\frac{\tau}{x}, \alpha_s(M_x^2)\right) \ , \tag{2.8.2}$$

where τ is the minimum value of x_1 or x_2 that allows kinematically to produce a final state with invariant mass M_X,

$$\tau = \frac{M_X^2}{s} \tag{2.8.3}$$

The partonic luminosity $\mathcal{L}$ is defined by

$$\mathcal{L}(x, M_X^2) \equiv \sum_{a,b} \int_x^1 \frac{dz}{z} f_a(z, M_X^2) f_b(x/z, M_X^2) \ . \tag{2.8.4}$$

We can define the luminosity of a given *partonic channel* by restricting the sum in Eq. 2.8.4 to a particular subset of partons. For example if a and b are restricted to be quarks we talk about *quark luminosity*, $\mathcal{L}_{qq}$. Note that, since a and b are indistinguishable in a pp collider like the LHC, it only makes sense to define partonic channels that are symmetric under the exchange $a \leftrightarrow b$. For example $\mathcal{L}_{u\bar{d}}$ is defined as

$$\mathcal{L}_{u\bar{d}}(x, M_X^2) \equiv \int_x^1 \frac{dz}{z} f_u(z, M_X^2) f_{\bar{d}}(x/z, M_X^2) + \int_x^1 \frac{dz}{z} f_{\bar{d}}(z, M_X^2) f_u(x/z, M_X^2) \ . \tag{2.8.5}$$

The coefficient function is defined analogously to the DIS case Eq 2.5.1, as dimensionless quantities that encode the partonic cross section,

$$\hat{\sigma}_{ab\to X} = \sigma_0 C_{ab}\left(\tau, \alpha_s(M_X^2)\right) \tag{2.8.6}$$

where

$$C_{ab}\left(\tau, \alpha_s(M_X^2)\right) = c_{ab}\delta(1-x) + \mathcal{O}(\alpha_s) \tag{2.8.7}$$

with the numbers c_{ab} nonzero only if the given partons couple to the final state X at leading order.

Chapter 3

PDFs for practical usage

In this chapter we consider several transformations of PDF sets that make them more useful in certain practical situations.

3.1 PDFs from the point of view of users

A user of PDFs is typically someone who wishes to compute a PDF dependent quantity (such as cross sections for hadronic processes). Users generally possess the means to compute the corresponding parton level quantity (e.g. given in terms of a parton level simulation [50, 51, 52, 53, 54, 55, 56] or as a precomputed grid [57, 9]), and need to convolve it with a PDF to obtain the corresponding observable result.

From the point of view of an user, PDF sets are generally given as a set of computer files in the LHAPDF [58] format. An LHAPDF set consists of a set of *members*, each consisting of the value of a PDF sampled at a grid of points in (x, Q) and for each flavour. The LHAPDF software then interpolates to obtain the value of the PDF at arbitrary points in (x, Q) (it can also extrapolate, but that is generally not advised since the extrapolation does not preserve basic properties of the PDF like the sum rules in Eqs. 2.4.1-2.4.4). The *members* of the PDF sets are used to compute PDF dependent quantities together with their uncertainties. The following types of PDF members are often employed:

Central values Representing the "best fit" PDF.

Error members These represent variations around the best result. They are used to compute "PDF uncertainties" in a way prescribed by each set. The most common error types are *Monte Carlo* and *Hessian*. We will describe them next in Sec. 3.2.

Parameter fluctuations Most commonly α_S. These members are used to correlated the change in the PDFs with the variations of the parameter, and account for it when for example considering the uncertainties of that parameter.

To obtain an estimate of the PDF uncertainty of a given quantity, one must compute it for each of the PDF members. This is often computationally prohibitive as it requires to repeatedly compute expensive convolutions Eq 2.8.1. We will dedicate

much of the remaining of this chapter to the problem of *compressing* the information contained in the error members in a way that they reproduce the features of the starting set of error members as faithfully as possible.

In the context of experimental analyses at the LHC, PDF uncertainties need to be combined with those from other theory and experimental parameters. This is often done employing maximum likelihood estimation method (see e.g. Ref [59]) which assume that the error parameters can be continuously varied. The Hessian error representation is suited for this task, while the Monte Carlo is not. Therefore, it is advantageous for PDF sets that are obtained as Monte Carlo samples like NNPDF (see Sec 5.1) to be convertible to error sets of the Hessian type. In Sec. 3.3 we describe a method that implements this transformation.

Finally it may be desirable to somewhat modify the assumptions made in the PDF fit, for example to impose some asymptotic behaviour at small x consistent with the expected behaviour at leading order.

3.2 Representation of uncertainties in PDFs

We note that the way PDF uncertainties are represented in the grid seen by the user (See Sec 3.1) is not necessarily related to the way PDFs were determined. Indeed in the following sections we discuss methods to convert between the different representations of the uncertainties a posteriori.

3.2.1 Monte Carlo errors

For Monte Carlo PDF sets, the error members describe samples from some distribution of a set of functions (that is, one PDF for each flavour) that is constructed propagating the uncertainties in the fitting procedure. These are typically the uncertainties of the data used to constrain the PDFs, the random state of the fitting algorithms that selects potentially different minima (since the PDFs are not fully constrained by the data) and in some instances, additional theory errors. We discuss a Monte Carlo fitting procedure in more detail in Sec. 5.1. We call each error member, a *Monte Carlo replica*.

Any quantity $\mathcal{O}$ that depends on the PDFs adopts a different value for each replica, $\mathcal{O}_r$. The PDF dependence can then be characterized by computing statistical estimators over the set of values $\{\mathcal{O}_r\}$, $r = 1 \dots N_{\rm rep}$. We can define its central value as the mean of the values that it adopts for each of the PDF replicas:

$$\langle \mathcal{O} \rangle = \langle \mathcal{O} \rangle_r = \frac{1}{N_{\rm rep}} \sum_r^{N_{\rm rep}} O_r \tag{3.2.1}$$

PDF uncertainty on any quantity $\mathcal{O}$ as the standard deviation of the ensemble of values that $\mathcal{O}$ has for each replica:

$$\Delta_{\mathcal{O}} = \left(\frac{1}{N_{\rm rep} - 1} \sum_r^{N_{\rm rep}} (\mathcal{O}_r - \langle \mathcal{O} \rangle_r)^2 \right)^{\frac{1}{2}} \tag{3.2.2}$$

Other statistical estimators, such as the median (instead of the mean) and the interquartile range (instead of the standard deviation) may be advisable in that they are more resilient to outliers.

The central value PDF written to the LHAPDF grids is similarly the mean of the replicas. Note that using the value of the observable computed with the central value PDF is different from Eq. 3.2.1 if the observable is not linear in the PDFs. The difference can be significative compared with the PDF uncertainty, for example when computing the best fit value of $\alpha_s\left(M_Z^2\right)$, as we will discuss in Chapter 6.

3.2.2 Hessian errors

The Hessian error formalism is a natural way to represent PDF uncertainties when a given PDF parametrization is assumed. The PDFs are then fitted by maximizing the agreement with the data and the uncertainties are assumed to be Gaussian fluctuations around the best fit values for each parameter [60, 61]. That is, at some fixed scale Q_0^2, the PDF is defined by a set of parameters. For example, the functional form assumed in older PDF fits by the CTEQ collaboration was [62]

$$xf(x,Q_0^2) = a_0 x^{a_1}(a-x)^{a_2}\exp(a_3 x + a_4 x^2 + a_5\sqrt{x} + a_6 x^{-a_7}) \ . \tag{3.2.3}$$

Current Hessian determinations employ more complicated functional forms containing a polynomial basis to some high order [63, 64, 65]. The parameters $\vec{a}$ ($\vec{a} = \{a_0 \dots a_7\}$ in the example above) are fixed by minimizing some error function like

$$\chi^2(\vec{a}) = \sum_{ij}(d_i - t_i(\vec{a}))C_{ij}^{-1}(d_j - t_j(\vec{a})) \ , \tag{3.2.4}$$

where d_i is the experimentally measured value for the data point i, t_i is the corresponding theoretical prediction obtained from the PDF and C_{ij} measures the experimental covariance between the points i and j. The parametrization, and thus the PDF is then obtained as

$$\vec{a_0} = \arg\min_{\vec{a}} \chi^2(\vec{a}) \ , \tag{3.2.5}$$

where the arg min notation stands for the values in the domain of the function $\chi^2(\vec{a})$ where it is minimized. Assuming the minimum is unique and that the error function is quadratic around the minimum, a small deviation from the minimum when we change the parameters, is given by

$$\Delta\chi^2 = \chi^2(\vec{a}) - \chi^2(\vec{a}_0) = (\vec{a} - \vec{a}_0)H(\vec{a} - \vec{a}_0) \ , \tag{3.2.6}$$

where H is the Hessian matrix of the error function χ^2 (with an added factor 1/2 for convenience as will be clear from Eq. 3.2.9), evaluated at the minimum of the parameters,

$$H_{ij} = \frac{1}{2}\left.\frac{\partial^2\chi^2(\vec{a})}{\partial a_i \partial a_j}\right|_{\vec{a}=\vec{a}_0} \ . \tag{3.2.7}$$

In terms of the shifts,

$$\vec{\delta} = \vec{a} - \vec{a}_0 \ , \tag{3.2.8}$$

the variation is

$$\Delta\chi^2 = \vec{\delta}H\vec{\delta} \tag{3.2.9}$$

Since H is symmetric, it can be diagonalized in terms of a complete orthogonal basis. We define a set of rescaled eigenvectors $\vec{e}_i$ as the normalized eigenvectors $\vec{v}_i$ over the square root of the eigenvalues λ_i; that is $\vec{e}_i = \vec{v}_i/\sqrt{\lambda_i}$. We can now project $\vec{\delta}$ in the basis of $\vec{e_i}$,

$$\vec{\delta} = \sum_i z_i \vec{e}_i \ , \tag{3.2.10}$$

with $z_i = \vec{\delta}.\vec{e_i}$. Replacing in Eq. 3.2.9, we find:

$$\Delta\chi^2 = \sum_i z_i^2 \tag{3.2.11}$$

This defines a sphere in the rotated parameter space defined by the bases $\{e_i\}$, centered in $\vec{a_0}$ and corresponding to increases in the figure of merit of up to $\Delta\chi^2$, and thus the expected value of the parameters.

Given any observable O that depends on the parameters trough the PDF, we can now find the maximum deviation from its best fit value (when $\delta = 0$) that is consistent with a fixed increase in the error function $\Delta\chi^2$. In the linear approximation we are assuming, we have

$$O(\delta) - O(0) = \sum_i \left.\frac{dO}{d\delta_i}\right|_{\delta_i=0} \delta_i = \sum_i \left.\frac{dO}{dz_i}\right|_{z_i=0} z_i \ . \tag{3.2.12}$$

The vector of maximum deviation is parallel to the gradient, that is:

$$z_i = \frac{dO}{dz_i} \sqrt{\Delta\chi^2 / \sum_j \left(\left.\frac{dO}{dz_j}\right|_{z_j=0} \right)^2} \tag{3.2.13}$$

Then the square of the maximum deviation is, from Eqs 3.2.12 and 3.2.13,

$$\Delta_O^2 = \Delta\chi^2 \sum_i \left.\frac{dO}{dz_i}\right|_{z_i=0}^2 \tag{3.2.14}$$

Assuming again linear error propagation, we can construct a general set of PDFs that allow to compute the uncertainty in Eq. 3.2.14. Each error member is constructed by shifting the best fit parameters in the direction of the corresponding eigenvectors of the covariance matrix,

$$\vec{a}^{(k)} = \vec{a_0} + t\vec{e_k} \ , \tag{3.2.15}$$

where, $t^2 = \Delta\chi^2$.

The final recipe for estimating uncertainties is then similar to what we find in the Monte Carlo Method, Eq. 3.2.2:

$$\Delta_{\mathcal{O}} = \left(\sum_k^{N_{\rm eig}} (\mathcal{O}(\vec{a}_k) - \mathcal{O}(\vec{a}_0))^2 \right)^{\frac{1}{2}} \tag{3.2.16}$$

Some caveats of the fitting Hessian methodology include:

- The $\Delta\chi^2$ parameter needs to be chosen in a way that accounts for both the possible inconsistencies in the input data and limitations in the parametrization that may limit the maximum achievable agreement.

- It is complicated to account for parameters that are too sensitive to the fluctuations in the experimental data *overfitting*. This is accomplished with relatively involved procedures employed to fix the parametrization (e.g. [66, 64]), or directly not accounted for. This also makes it difficult to estimate the uncertainty associated to making a particular choice of parametrization.

- The approximation of linear error propagation that is the basis to the Hessian approximation is not necessarily precise enough. It is possible to partially correct for that by setting the parameter t in Eq. 3.2.15. Indeed to improve the robustness of the Hessian representation, ofen positive and negative fluctuations are construncted with respect to each eigenvector. That is, for each eigenvector, in addition to an error member constrcuted with the parametrization in Eq. 3.2.15, we also have,

  $$\vec{a}^{(k-)} = \vec{a_0} - t\vec{e_k} \tag{3.2.17}$$

  Note that this comes at the cost of duplicating the total number of convolutions needed to compute the PDF uncertainty for a fixed number of eigenvectors.

Note that only the last point applies to the Hessian error representation, as opposed to fits performed with the Hessian methodology: The uncertainty on the predictions will be accurate only up to a linear approximation in the dependence of a given observable on the PDF. A related limitation of the Hessian representation is that the shifts can only correspond to Gaussian fluctuationas in parameter space; as we show in Sec 4.4, the PDF uncertainties in kinematical regions were PDFs are not so well constrained (that is, small and large x) are best characterized by non Gaussian distributions. This is because the uncertainty is mainly determined by the methodology and non Gaussian constrains like positivity (see Sec 5.1.3) rather than by experimental data (which are often assumed Gaussian).

3.3 Monte Carlo to Hessian transformation

3.3.1 Introduction

One crucial advantage is the Hessian error representation of PDFs described in Sec. 3.2.2 over the Monte Carlo error representation (Sec. 3.2.1) is that it allows to interpret the PDF fluctuations in terms of continuous parameter variations, and the orthogonal Hessian eigenvectors can be used as nuisance parameters in experimental analyses. In this way PDF uncertainties are more easily combined with other experimental or theoretical sources of uncertainty, at least as long as the theoretical uncertainties do not depend on the PDFs, since otherwise coherent variations of the PDFs also the corresponding variations would be needed (see Sec. 6.4.1 for an example).

NNPDF however uses a Monte Carlo fitting procedure, that instead has the advantage that it allows to construct PDFs with an arbitrary functional form (neural network) and characterized by a very large number of parameters (and with a built in

procedure to eliminate possible over-learning). It also does not need to assume that the distributions around the best fit are Gaussian, or linear error propagation.

Even though deviations from Gaussianity may be important in specific kinematic regions, especially when limited experimental measurements are available and PDF uncertainties are driven by theoretical constraints (such as for example the large-x region, relevant for new physics searches), in most cases, and specifically when PDF uncertainties are small and driven by abundant experimental data, the Gaussian approximation is reasonably accurate. This then raises the question of whether in such case, in which everything is Gaussian and the Hessian approximation is adequate, one could have the best of possible worlds: a Hessian representation with the associate advantages, but without having to give up the use of a general-purpose flexible functional form. We develop a methodology for the construction of a Hessian representation of Monte Carlo PDFs that achieves this goal in a straight forward manner. Starting from an initial Monte Carlo PDF set, the Hessian error sets are directly obtained as the eigenvectors of the covariance matrix in PDF space. The method offers a direct way to achieve *"PDF compression"*: That is, to represent the PDF uncertainties of the initial set with a much smaller number of error sets and minimal loss of accuracy. Furthermore, since the inverse problem of obtaining a Monte Carlo representation of a Hessian set was already solved [67], the method allows to convert between the two main representations of PDF errors. This in turn enables the construction of combined PDF sets [48, 67, 68, 69] in either representation. In this context, the problem of obtaining a common Hessian PDF representation has also been tackled in the so-called "Meta-PDF" approach [69], which is based on first parametrizing the Monte Carlo members in terms of a common functional form thus induced the problems associated with a particular choice of parametrization, and in particular the accuracy loss due to the finite flexibility of the assumed functional form. In Ref. [13], we proposed a method that minimizes the bias by using the Monte Carlo replicas themselves as a basis for the parametrisation, and employs a genetic algorithm to find the linear combination that best describes each replica. In the method described here (described for the first time in the Appendix of Ref. [13] and explained in more detail in Ref. [14]), we skip the parametrization step altogether and employ directly the eigenvectors of the covariance matrix as Hessian parameters.

The benchmarks done in the context of the PDF4LHC recommendation [4, 15] and discussed in Chapter. 4 proved that this method is superior to either of the two at a fixed number of error sets, in that it provides a description of the PDF uncertainty of the most relevant PDF dependent observables of the LHC that is closer to the prior. The method is also more conceptually simple and straight forward to implement efficiently (the topical runtime is of order one minute on a laptop while the GA based method required a day on a computing cluster).

In the following, we dub the Monte Carlo to Hessian transformation SVD+PCA.

3.3.2 Methodology

Here we will assume the central value to be the same as in the prior PDF set, from which, if the prior is given as a Monte Carlo, it is typically determined as a mean (note that this is not uniquely the best choice, as pointed out in Sec. 3.2.1).

Since we are interested in the construction of a multigaussian representation in

PDF space, the only information we need is the corresponding covariance matrix. This is constructed starting with a matrix X which samples over a grid of points the difference between each PDF replica, $f_\alpha^{(k)}(x_i, Q)$, and the central set, $f_\alpha^{(0)}(x_i, Q)$, namely

$$X_{lk}(Q) \equiv f_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q)\,, \tag{3.3.1}$$

where α runs over the N_f independent PDF flavors at the factorization scale $\mu_F = Q$, i runs over the N_x points in the x grid where the PDFs are sampled, $l = N_x(\alpha - 1) + i$ runs over all $N_x N_f$ grid points, and k runs over the $N_{\rm rep}$ replicas. The sampling is chosen to be fine-grained enough that results will not depend on it.

The desired covariance matrix in PDF space is then constructed as

$$\mathrm{cov}(Q) = \frac{1}{N_{\rm rep} - 1} X X^t\ . \tag{3.3.2}$$

The key idea which underlies the SVD method is to represent the $(N_x N_f) \times (N_x N_f)$ covariance matrix Eq. (3.3.2) over the $N_{\rm rep}$ dimensional linear space spanned by the replicas (assuming $N_{\rm rep} > N_x N_f$), by viewing its $N_x N_f$ eigenvectors as orthonormal basis vectors in this space, which can thus be represented as linear combinations of replicas. The subsequent PCA optimization then simply consists of picking the subspace spanned by the dominant eigenvectors, *i.e.*, those with largest eigenvalues.

The first step is the SVD of the sampling matrix X, namely

$$X = USV^t\ , \tag{3.3.3}$$

where U and V^t are orthogonal matrices, with dimensions respectively $N_x N_f \times N_{\rm eig}^{(0)}$ and $N_{\rm rep} \times N_{\rm rep}$, S is a diagonal $N_{\rm eig}^{(0)} \times N_{\rm rep}$ positive semi-definite matrix, whose elements are the so-called singular values of X, and the initial number of singular values is given by $N_{\rm eig}^{(0)} = N_x N_f$.

The matrix $Z = US$ then has the important property that

$$ZZ^t = XX^t, \tag{3.3.4}$$

but also that it can be expressed as

$$Z = XV, \tag{3.3.5}$$

and thus it provides the sought-for representation of the multigaussian covariance matrix in terms of the original PDF replicas: specifically, V_{kj} is the expansion coefficient of the j-th eigenvector over the k-th replica. We assume henceforth that the singular values are ordered, so that the first diagonal entry of S correspond to the largest value, the second to the second-largest and so forth.

The PCA optimization then consists of only retaining the principal components, *i.e.* the largest singular values. In this case, U and S are replaced by their sub-matrices, denoted by u and s respectively, with dimension $N_x N_f \times N_{\rm eig}$ and $N_{\rm eig} \times N_{\rm rep}$, with $N_{\rm eig} < N_{\rm eig}^{(0)}$ the number of eigenvectors which have been retained. Due to the ordering, these are the upper left sub-matrices. Because s has only $N_{\rm eig}$ non-vanishing diagonal entries, only the $N_{\rm rep} \times N_{\rm eig}$ submatrix of V contributes. We call this the principal submatrix P of V:

$$P_{kj} = V_{kj} \qquad k = 1\,, \ldots, N_{\rm rep}\,, \qquad j = 1, \ldots, N_{\rm eig}\,. \tag{3.3.6}$$

The optimized representation of the original covariance matrix, Eq. (3.3.2), is then found by replacing V with its principal submatrix P in Eq. (3.3.5). This principal matrix P is thus the output of the SVD+PCA method: it contains the coefficients of the linear combination of the original replicas or error sets which correspond to the principal components, which can be used to compute PDF uncertainties using the Hessian method.

Indeed, given a certain observable σ_i (which could be a cross-section, the value of a structure function, a bin of a differential distribution, etc.) its PDF uncertainty can be computed in terms of the original Monte Carlo replicas by

$$s_{\sigma_i} = \left(\frac{1}{N_{\rm rep}-1} \sum_{k=1}^{N_{\rm rep}} \left(\sigma_i^{(k)} - \sigma_i^{(0)} \right)^2 \right)^{\frac{1}{2}} = \frac{1}{\sqrt{N_{\rm rep}-1}} \left\| d(\sigma_i) \right\| , \tag{3.3.7}$$

where $\sigma_i^{(k)}$ is the prediction obtained using the k-th Monte Carlo PDF replica, $\sigma_i^{(0)}$ is the central prediction, and in the last step we have defined the vector of differences

$$d_k(\sigma_i) \equiv \sigma_i^{(k)} - \sigma_i^{(0)} , \qquad k = 1, \ldots, N_{\rm rep} , \tag{3.3.8}$$

with norm

$$\left\| d(\sigma_i) \right\| \equiv \left(\sum_{k=1}^{N_{\rm rep}} d_k^2(\sigma_i) \right)^{\frac{1}{2}} . \tag{3.3.9}$$

Note that this is another way of writting Eq.3.2.2.

Assuming linear error propagation and using Eq. (3.3.5), the norm of the vector $\{d_k(\sigma_i)\}$, Eq. (3.3.8), can be represented on the eigenvector basis:

$$\left\| d(\sigma_1) \right\| = \left\| d^V(\sigma_1) \right\| \tag{3.3.10}$$

where the rotated vector

$$d^V{}_j(\sigma_i) = \sum_{k=1}^{N_{\rm rep}} d_k(\sigma_i) V_{kj}, \qquad j = 1, \ldots, N_{\rm eig}^{(0)} , \tag{3.3.11}$$

has the same norm as the original one because of Eq. (3.3.4).

Replacing V by the principal matrix P in Eq. (3.3.11), *i.e.*, letting j only run up to $N_{\rm eig} < N_{\rm eig}^{(0)}$ we get

$$\tilde{s}_{\sigma_i} = \frac{1}{\sqrt{N_{\rm rep}-1}} \left\| d^P(\sigma_i) \right\| , \tag{3.3.12}$$

where now the vector is both rotated and projected

$$d^P{}_j(\sigma_i) = \sum_{k=1}^{N_{\rm rep}} d_k(\sigma_i) P_{kj}, \qquad j = 1, \ldots, N_{\rm eig} . \tag{3.3.13}$$

The norm of d^P is only approximately equal to that of the starting vector of differences d: $\left\| d^P(\sigma_1) \right\| \approx \left\| d(\sigma_1) \right\|$. However, it is easy to see that this provides the linear

combination of replicas which minimizes the difference in absolute value between the prior and final covariance matrix for given number of eigenvectors. As the difference decreases monotonically as $N_{\rm eig}$ increases, the value of $N_{\rm eig}$ can be tuned to any desired accuracy goal, with the exact equality Eq. (3.3.10) achieved when $N_{\rm eig} = N_{\rm eig}^{(0)}$. Note that, of course, the optimization step can be performed also starting with a symmetric Hessian, rather than Monte Carlo, prior. In such case, the index k runs over Hessian eigenvectors, Eq. (3.3.2) is replaced by $\mathrm{cov}(Q) = XX^t$, and the rest of the procedure is unchanged.

An interesting feature of this SVD+PCA method is that the matrix V (and thus also the principal matrix P) in Eq. (3.3.11) simply represents the coefficients of a linear combination of replicas, and this does not depend on the value of the PDF factorization scale Q (note that the evolution operator, viewed as a matrix would act on U instead): the scale dependence is thus entirely given by the DGLAP evolution equation satisfied by the original Monte Carlo replicas. Of course, the subsequent PCA projection may depend on scale if there are level crossings, but this is clearly a minor effect if a large enough number of principal components is retained. Because of this property, the SVD+PCA methodology can be used for the efficient construction [4] of a Hessian representation of combined PDF sets, even when the sets which enter the combination satisfy somewhat different evolution equations, *e.g.*, because of different choices in parameters such as the heavy quark masses, or in the specific solution of the DGLAP equations.

3.3.3 Number of error sets

The SVD+PCA method can represent Monte Carlo PDF sets with enough replicas that the statistical error is negligible, such as the 900 replicas of the prior PDF4LHC 2015 combined set [4], the 1000 replicas of the NNPDF 3.1 prior sets [12] or the 1000 replicas of the NNPDF 3.0 set [15] with around 100 Hessian error sets in such a way that the information loss due to the compression is negligible compared to the error due to the Hessian approximation.

While a smaller number of error sets can still yield reasonable accuracy (certainly comparable to other proposed solutions such as Meta PDF), one cannot control very accurately where the information loss is happening: smaller eigenvectors in the covariance matrix correspond to large-x (where the value of the PDF us numerically smaller) and small-x (where the correlation length is small and the covariance has many small contributions). Therefore observables that depend on the behaviour at the most extreme x values, like notably jet distributions would be the first to deteriorate as the number of error sets goes down.

We therefore develop a compression procedure where one can choose which behaviour of the PDF uncertainty is more important to reproduce. This is the idea behind the SMPDF method discussed next in Sec. 3.4.

3.4 The SMPDF algorithm

3.4.1 Introduction

The SMPDF method aims at producing PDFs with a minimal number of error sets, designed to provide accurate representations of PDF uncertainties for specific processes or classes of processes. The SMPDFs (*Specialized minimal PDFs*) are constructed in such a way that sets corresponding to different input processes can be combined together without losing information on their correlations, and therefore an existing set can also be enlarged to describe a new process.

While other compression methods such as MCH [13, 14], META PDFs [69] and CMC [70] aim at providing an optimized representation in all kinematic regions, here we exploit the well known fact [71] that if one is interested only in a specific set of cross sections, the number of PDF error members can be greatly reduced without significant accuracy loss. This allows us to reduce the number of error sets required from around a 100 required by each of the methods to achieve a reasonable description of the PDF uncertainty to as little as a dozen or less for sufficiently inclusive processes (see the detailed comparison in Ref [14]).

Our methodology is based on the SVD-PCA [13, 14], method discussed in Sec.3.3. Starting from either a Hessian or a Monte Carlo prior set, and a list of collider processes, the SM-PDF algorithm leads to a set of eigenvectors optimized for the description of the input processes within some given tolerance.

In comparison to existing methods, such as data set diagonalization [72], our methodology has the advantage that no information is lost in the process of the construction of the specialized set. This is because the specialized set is constructed through a suitable linear transformation, whereby the starting space is separated into a subspace spanned by the optimized SM-PDF set, and its orthogonal subspace. This then implies that any given SM-PDF set can be iteratively expanded in order to maintain a given accuracy for an increasingly large set of processes, and also, that SM-PDF sets optimized for different sets of processes can be combined into a single set, either *a priori*, at the level of PDFs, or *a posteriori*, at the level of cross-sections. This, for example, enables the a-posteriori combination of previous independent studies for a signal process and its corresponding backgrounds, with all correlations properly accounted for.

We describe the method en detail next. An example is shown in Sec. 4.4.

3.4.2 Methodology

In the SM-PDF method, this same SVD+PCA optimization is performed, but now with the goal of achieving a given accuracy goal not for the full prior PDF set in the complete range of x and Q^2, but rather for the aspects of it which are relevant for the determination of a given input set of cross-sections, and in such a way that all the information which is not immediately used is stored and can be *a posteriori* recovered either in part or fully, *e.g.* if one wishes to add further observables to the input list. The method allows the user to choose the desired accuracy of the representation, and has only one additional free parameter that is fixed by optimizing to data.

The algorithm is constructed by supplementing the SVD+PCA methodology of Sec. 3.3 with three additional features: a measure of the accuracy goal with which

the uncertainties are to be reproduced; a way of singling out the relevant part of the covariance matrix; and a way of keeping the information on the rest of the covariance matrix in such a way that the full covariance matrix can be recovered at a later stage, to improce the description of the next observable. The main input to the algorithm is the set of N_σ observables which we want to reproduce, $\{\sigma_i\}$, with $i = 1, \dots N_\sigma$. Theoretical predictions for the cross-sections $\{\sigma_i\}$ are computed using a prior PDF set, which we assume for definiteness to be given as a Monte Carlo, though the method works with obvious modifications also if the starting PDFs are given in Hessian form (one just needs to convert between Eqs. 3.2.2 and 3.2.16). The goal of the SM-PDF methodology is to evaluate the PDF uncertainties s_{σ_i}, Eq. (3.2.2), in terms of a reduced number of Hessian eigenvectors,

$$\tilde{s}_{\sigma_i} = \left(\sum_{n=1}^{N_{\rm eig}} \left(\widetilde{\sigma}_i^{(n)} - \widetilde{\sigma}_i^{(0)} \right)^2 \right)^{\frac{1}{2}} , \tag{3.4.1}$$

with the number $N_{\rm eig}$ being as small as possible within a given accuracy goal. We thus define a measure T_R of the accuracy goal (tolerance) by the condition

$$T < T_R; \qquad T \equiv \max_{i \in (1, N_\sigma)} \left| 1 - \frac{\tilde{s}_{\sigma_i}}{s_{\sigma_i}} \right| \tag{3.4.2}$$

in other words, T_R is the maximum relative difference which is allowed between the original and reduced PDF uncertainties, $\tilde{s}_{\sigma_i}$ and s_{σ_i} respectively, for all the observables $\{\sigma_i\}$.

In order to determine the part of the covariance matrix relevant for the description of the input observables $\{\sigma_i\}$, we define the correlation function

$$\rho\left(x_i, Q, \alpha, \sigma_i\right) \equiv \frac{N_{\rm rep}}{N_{\rm rep}-1} \left(\frac{\langle X(Q)_{lk} d_k(\sigma_i) \rangle_{\rm rep} - \langle X(Q_{\sigma_i})_{lk} \rangle_{\rm rep} \langle d_k(\sigma_i) \rangle_{\rm rep}}{s_\alpha^{\rm PDF}(x_i, Q) s_{\sigma_i}} \right) , \tag{3.4.3}$$

where the matrix of PDF differences $X(Q)$ and the grid index $l = N_x(\alpha - 1) + i$ have been defined in Sec. 3.3 Eq. (3.3.1); $s_\alpha^{\rm PDF}(x_i, Q)$ is the standard deviation of the PDFs in the prior Monte Carlo representation, given by the usual expression, analogously to Eq. 3.3.7

$$s_\alpha^{\rm PDF}(x_i, Q) = \left(\frac{1}{N_{\rm rep}-1} \sum_{k=1}^{N_{\rm rep}} \left[f_\alpha^{(k)}(x_i, Q) - \langle f_\alpha(x_i, Q) \rangle \right] \right)^{\frac{1}{2}} , \tag{3.4.4}$$

and s_{σ_i}, the standard deviation of the i-th observable σ_i, is given by Eq. (3.3.7). The function in Eq. (3.4.3) measures the correlation between the observables σ_i and the l-th PDF value (*i.e.* $f_\alpha(x_i, Q)$, with $l = N_x(\alpha - 1) + i$).

The basic idea of the SM-PDF construction is to apply the SVD to the subset of the covariance matrix which is most correlated to the specific observables that one wishes to reproduce $\{\sigma_i\}$, one at a time, through an iterative procedure schematically represented in Fig. 3.1.

The iteration loop (contained in the dashed box in Figure 3.1) is labeled by an iteration index j, such that at each iteration an extra eigenvector is added, thereby

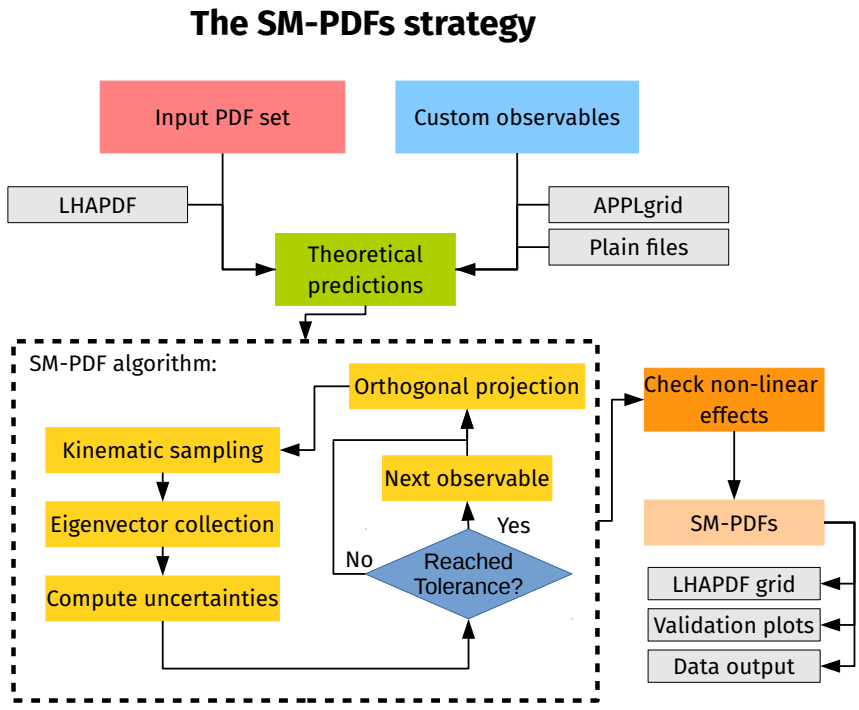


Figure 3.1: Schematic representation of the SM-PDF strategy.

increasing the accuracy. If the accuracy goal is achieved for all observables after j iterations, then the final reduced Hessian set contains $N_{\rm eig} = j$ eigenvectors as error sets. These are delivered as a new principal matrix P, which provides the expansion coefficients of the eigenvectors over the replica basis: namely, P_{kj} is the component of the j-th eigenvector in terms of the k-th replica. They thus replace the principal matrix of the previous PCA procedure as a final output of the procedure, and can be used in exactly the same way.

To set off the iterative procedure, we select one of the observables we wish to reproduce from the list, σ_1, and compute the correlation coefficient $\rho\left(x_i, Q, \alpha, \sigma_1\right)$ for all grid points (x_i, α) and for a suitable choice of scale Q. We then identify the subset Ξ of grid points for which ρ exceeds some threshold value:

$$\Xi = \left\{(x_i, \alpha) : \rho\left(x_i, Q_{\sigma_1}, \alpha, \sigma_1\right) \geq t\rho_{\rm max}\right\} . \tag{3.4.5}$$

The threshold value is expressed as a fraction $0 < t < 1$ times the maximum value $\rho_{\rm max}$ that the correlation coefficient takes over the whole grid, thereby making the criterion independent of the absolute scale of the correlation.

We then construct a reduced sampling matrix X_Ξ, defined as in Eq. (3.3.1), but now only including points in the $\{x_i, \alpha\}$ space which are in the subset Ξ. We perform the SVD of the reduced matrix

$$X_\Xi = USV^t , \tag{3.4.6}$$

and we only keep the largest principal component, *i.e.* one single largest eigenvector, which is specified by the coefficients of its expansion over the replica basis, namely,

assuming that the singular values are ordered, by the first row of the V matrix. We thus start filling our output principal matrix P by letting

$$P_{kj} = V_{k1}^{(j)}\,, \quad j = 1\,, \quad k = 1, \ldots, N_{\rm rep}\,. \tag{3.4.7}$$

Note that j on the left-hand side labels the eigenvector (P_{kj} provides expansion coefficients for the j-the eigenvector) and on the right-hand side it labels the iteration ($V_{k1}^{(j)}$ is the first row of the V-matrix at the j-th iteration), which we can identify because, as mentioned, at each iteration we will add an eigenvector. The remaining eigenvectors of the principal matrix span the linear subspace orthogonal to P, and we assign them to a residual matrix R:

$$R_{km}^{(j)} = V_{k(m+1)}^{(j)} \qquad j = 1\,, \qquad m = 1, \ldots, N_{\rm rep} - 1\,, \qquad k = 1, \ldots, N_{\rm rep}\,. \tag{3.4.8}$$

At the first iteration, when there is only one eigenvector, the principal matrix P has just one row, and it coincides with the principal component of V. So far, the procedure is identical to that of the SVD+PCA method, and we can thus use again Eq.(3.3.12) to compute uncertainties on observables, check whether the condition Eq. (3.4.2) is met, and if it is not add more eigenvectors. The procedure works in such a way that each time a new eigenvector is selected, exactly the same steps are repeated in the subspace orthogonal to that of the previously selected eigenvectors, thereby ensuring that information is never discarded. This is achieved by a projection method.

Specifically, we project the matrix X and the vector of observable differences $\{d_k(\sigma_i)\}$ on the orthogonal subspace of P, namely, the space orthogonal to that spanned by the eigenvectors which have already been selected (as many as the number of previous iterations). The projections are performed by respectively replacing $d(\sigma_i)$ and X by

$$d^R(\sigma_i) = d(\sigma_i) R^{(j-1)}\,, \tag{3.4.9}$$

$$X^R = X R^{(j-1)}\,, \tag{3.4.10}$$

where the first iteration of the residual matrix $R^{(1)}$ has been defined in Eq. (3.4.8).

After the projection, we proceed as in the first iteration. We first determine again the subset Ξ, Eq. (3.4.5), of the projected sampling matrix X^R, thereby obtaining a new sampling matrix X^R_Ξ: this is possible because everything is expressed as a linear combination of replicas anyway. Once the new matrix X^R_Ξ has been constructed, the procedure is restarted from Eq. (3.4.6), leading to a new matrix V^R. The number of columns of the projected matrix X^R_Ξ (and therefore of V^R) is $N_{\rm rep} - (j-1)$, which is the dimension of the subspace of the linear combinations not yet selected by the algorithm (that is, $N_{\rm rep} - 1$ for $j = 2$, and so on). We can now go back to Eq. (3.4.7) and proceed as in the previous case, but with the projected matrices: we add another row to the matrix of coefficients to the principal matrix by picking the largest eigenvector of the projected matrix, and determining again the orthogonal subspace.

At the j-th iteration, this procedure gives

$$P^{R\,(j)}_{k} = V^{R\,(j)}_{k1}\,, \quad k = 1, \ldots, N_{\rm rep} - (j-1)\,, \tag{3.4.11}$$

$$R^{R\,(j)}_{km} = V^{R\,(j)}_{k(m+1)}\,, \quad m = 1, \ldots, N_{\rm rep} - j\,, \quad k = 1, \ldots, N_{\rm rep} - (j-1)\,. \tag{3.4.12}$$

which respectively generalize Eqs. (3.4.7) and (3.4.8) for $j \geq 1$. The projected row of coefficients P^R Eq. (3.4.11) can be used to determine the corresponding unprojected row of coefficients of the principal matrix and of the residual matrix by using the projection R matrix in reverse, *i.e.*, at the j-th iteration

$$P^{(j)}_{kh} = \sum_{k'} R^{(j-1)}_{kk'} P^{R^{(j)}}_{k'h}\,, \tag{3.4.13}$$

$$R^{(j)}_{kh} = \sum_{k'} R^{(j-1)}_{kk'} R^{R^{(j)}}_{k'h}\,. \tag{3.4.14}$$

We thus end up with a principal matrix which has been filled with a further eigenvector, and a new residual matrix and thus a new projection.

In summary, at each iteration we first project onto the residual subspace, Eq. (3.4.9), then pick the largest eigenvector in the subspace, Eq. (3.4.11), then re-express results in the starting space of replicas, Eq. (3.4.13), so P is always the first row of V in each subspace, and Eqs. (3.3.13-3.3.12) remain valid as the P matrix is gradually filled. Determining the correlation and thus Ξ after projection ensures that only the correlations with previously unselected linear combinations are kept. The fact that we are always working in the orthogonal subspace implies that the agreement for the observables σ_i which had already been included can only be improved and not deteriorated by subsequent iterations. It follows that we can always just check the tolerance condition on one observable at a time. The procedure is thus unchanged regardless of whether we are adding a new observable or not. In any case, the subset Ξ Eq. (3.4.5) is always determined by only one observable, namely, the one that failed to satisfy the tolerance condition at the previous iteration. The procedure is iterated until the condition is satisfied for all observables $\{\sigma_i\}$ in the input list. The number of iterations j until convergence defines the final number of eigenvectors $N_{\rm eig}$.

The output of the algorithm is the final $N_{\rm rep} \times N_{\rm eig}$ principal matrix P, which can be used to compute uncertainties on observables using Eqs. (3.3.12-3.3.13). However, for the final delivery we wish to obtain a set of Hessian eigenvectors. These can be obtained by performing the linear transformation given by P (a rotation and a projection) in the space of PDFs. The X matrix Eq. (3.3.1) then becomes

$$\widetilde{X} \equiv \sqrt{\frac{1}{N_{\rm rep}-1}} XP\,, \tag{3.4.15}$$

so, substituting in Eq. (3.3.1), the final $N_{\rm eig}$ eigenvectors are found to be given by

$$\widetilde{f}^{(k)}_\alpha(x_i, Q) = f^{(0)}_\alpha(x_i, Q) + \widetilde{X}_{lk}(Q)\,, \quad k = 1, \ldots, N_{\rm eig}\,. \tag{3.4.16}$$

This is the same result as with the SVD+PCA algorithm of Sect. 3.3, but now generally with a smaller number of eigenvectors, namely, those which are necessary to describe the subset of the covariance matrix which is correlated to the input set of observables.

Chapter 4

The PDF4LHC recommendation

4.1 Introduction

There exist several collaborations that produce PDF sets [65, 64, 73, 74, 75, 12] which are advertised as adequate "For High precision collider data", or "for the LHC era". These sets can however lead to significatively different predictions (and importantly, also different sizes of the corresponding PDF uncertainties), and therefore leave non-expert users confused as to how to interpret the results. This is particularly relevant in the context of precise experimental analyses at the LHC [48, 68]: The level of precision achievable by the experiments is such that differences in the PDFs may well be the main reason of potential disagreements between theory and data. Also in this context, it is particularly convenient to have an agreed upon prescription that different analysis groups can use to compare their results.

The PDF4LHC working group has the task to elucidate the differences between the PDF sets used at the LHC, and to provide a protocol for both experimentalists and theorists to calculate PDF dependent quantities, with the necessarily corresponding estimate of their PDF uncertainty. This is then translated into recommendations [76, 4] for the broader community, with guidelines on how to compute PDF+α_s uncertainties. While alternative sets of recommendations exist (e.g. [77]), the PDF4LHC guidelines are mostly employed by the community (as measured by citation count or by e.g. their adoption by the Higgs Cross Section Working Group [78, 16]). The PDF4LHC recommendations specify that PDF dependent quantities should be estimated taking into account the results from different PDF fits.

Th evolution of the PDF4LHC guidelines reflects the progress in the field. In 2010 [76], at the time of the first recommendation, PDFs did not incorporate LHC data, the importance of the parametrization uncertainties wasn't sufficiently recognized and the GM-VFNS schemes (see Sec. 2.7) were only in the process of being implemented. The difference among PDF sets was much bigger than the PDF uncertainties they quoted, and the origin of the disagreements was unclear [79, 80]. These facts required that the recommendation for estimating PDF uncertainties was conservative and largely independent of the PDF uncertainties provided by each group since they were too small to account for the poorly understood differences.

By 2015, the situation was much improved. The updated versions of the PDF sets entering the previous recommendation agreed within their respective uncertainties,

and provided predictions for LHC observables that were broadly in agreement. This suggested that now the uncertainties from each group could be assumed to have a proper statistical meaning and that the recommendation should take them into account in a statistically sound way. Consequently, the need for an updated recommendation was identified [81].

Naturally, an essential task of the recommendation is to define a set of requirements for PDF sets to be part of it. These criteria are presented in Sec. 4.2. In Sec. 4.3 we present the combination method that was adopted in the 2015 recommendation and compare some of its features with the older 2010 prescription. Next, in Sec. 4.4 we present the methods that were used to obtain the final PDF sets together with detailed benchmarks.

4.2 Criteria for sets entering the combination

The guidelines in the PDF4LHC recommendation are based on the most up to date understanding on the subject of PDF determination. The 2015 recommendation steams from the combined result of the progress made by individual groups and common benchmark studies, particularly Refs. [82, 79, 80, 83] and references therein. The first task of the recommendation is to identify a set of criteria that specify which PDF sets are suitable to be included. These criteria are:

Based on global dataset It has been recognized [64, 75, 12] that including including diverse types of processes from both fixed target and collider experiments contributes to reduce the experimental uncertainties by providing more stringent constraints on the PDFs. This increases the requirements on the PDF fitting methodologies themselves: A successful PDF determination should be able to accommodate all available experimental data within its experimental and theoretical uncertainties, as well as to detect problems in the experimental and theoretical inputs (see e.g. the approach taken by NNPDF in Sec. 5.4). While a generalized procedure to treat theoretical uncertainties in PDF fits does not exist currently, one can reasonably expect that different physical process that constraint the similar partonic channels can provide increased robustness against missing higher order corrections. For example top quark pair production, jet production, and the Z transverse momentum distribution all constrain the gluon PDF at medium x. Since there is no reason to think that the higher order corrections of these processes are correlated, including them all is likely partly average out the undue pull that the missing higher order of each process has in the PDFs. Similarly including data from different experiments allows to assess possible underestimated experimental uncertainty by comparing the constraints provided by them all (i.e. we would find poor agreement between the dataset with underestimated uncertainties and the rest of the data).

NNLO theory with a GM-VFNS All the process included modern PDF analyses can be computed at two loops in α_s: It is possible to include NNLO corrections at the level of total cross sections and for exclusive distributions for Deep Inelastic Scattering [84, 85, 86], Drell Yan[87, 88, 89], and recently top quark pair production [90, 91], Z transverse distribution[92, 93, 94, 95] and dijet production [96].

Furthermore it has been established that the effect of the masses of the heavy quarks needs to be taken into account properly, thought the use of the General-Mass Variable Flavour Number Schemes (GM-VFNS). Indeed, in the detailed benchmarks leading to the 2015 recommendations, it was found that using a 3-flavour fixed flavour number scheme, with otherwise the same methodology as the one used in MSTW [97] and NNPDF [98], leads to a markedly worse agreement with the data on average, and a significantly lower preferred value of α_s. Ref. [99] concludes that some of the most significant differences between PDF sets are due to the choice of flavour scheme, specifically in the choice of a GM-VFNS versus a FFNS, with the first favoured on theoretical grounds: For example, the PDF evolution in a 3-flavour FFNS, where the charm quark is not generated perturbatively, yields logarithmic terms of the form $(\alpha_s \log(Q^2/m_c^2))^n)$ that are neglected. GM-VFNS have also been found to be advantageous by direct comparison of the fit quality to experimental data. On the other hand, the differences between the specific GM-VFNS have been shown to be subdominant: The variations are both formally higher order and numerically small compared to the experimental uncertainties; the differences between GM-VFNS were studied in detail in Chapter 22 of Ref. [82].

Usage of the world average of $\alpha_s\left(M_Z^2\right)$ The uncertainty on $\alpha_s\left(M_Z^2\right)$ has a strong impact both on the determination of PDFs and the prediction on PDF-dependent quantities [100, 101, 102]. While it is possible to determine α_s from the best fit PDF (see Chapter 6), it is considered advantageous to deliver default results at a commonly agreed value that is consistent with the PDG *World Average* [23]. Two main reasons motivate this requirement. Firstly, PDF based determinations of α_s miss independent constrains coming from Lattice QCD, τ decays or electroweak global fits, that are included in the PDG average (in Chapter. 6 we show that the determination based on top quark pair production included in the World Average is in fact not an independent constraint). The second reason is that in practice a common value of α_s is simpler for practical uses, particularly for the preparation of combined PDF sets. In fact, it was chosen to use the same central value of α_s at NLO and NNLO, namely $\alpha_s\left(M_Z^2\right) = 0.118$, which is consistent with the PDG average (based on determinations at, at least, NNLO accuracy). While a selecting different value at NLO could lead to an overall better agreement to the data with NLO PDFs, the differences in NLO predictions are due to finite higher order (that is NNLO) terms in α_s, and therefore consistent with NLO theory.

Similarly to α_s, the rest of the Standard Model parameters entering the PDF fit should be consistent with the PDG values.

Self validation methods The PDF methodologies entering the combination should be based on methodologies that contain an assessment of the uncertainties induced not only by the experimental data entering the fit but also due to the fitting procedure (particularly due to the choice of parametrization). NNPDF implements this requirement by tuning the methodology to closure tests that ensure that the uncertainties are consistent (See Sec. 5.1.9). CTEQ and MMHT implement dynamic tolerances [64, 75] (see Sec. 3.2.2).

At the time of producing the recommendation, CT14, MMHT20014 and NNPDF3.0, were the sets identified as satisfying the above requirements.

4.3 The PDF4LHC combination

The improvements in PDF determination we just summarized in Sec 4.1 call for a more statistically meaningful prescription for combining the results from multiple PDF than that adopted in 2011 (taking the *envelope* of the uncertainty bands). This is easily achievable when PDFs are in the Monte Carlo representation [48, 67, 68, 69]: One can simply concatenate together sets of Monte Carlo replicas from different groups. The meaning of such combined sets is that, when computing Monte Carlo errors following Sec 3.2.1 one has a given probability to sample a replica from any of the groups and therefore the uncertainties take into account both the individual estimate of PDF errors and the dispersion between different determinations. Note than in the case where all the individual PDF set agree perfectly both in terms of central values and uncertainties, the combined set is statistically equivalent to each of the combined sets rather than one with reduced uncertainties. This is the correct behaviour considering all the PDFs entering the combination use a similar input dataset. The Hessian PDF sets entering the combination, CT14, MMHT20014, can be converted to Monte Carlo following the method in Ref [67]. It was found that taking 300 replicas from each of the three PDF sets is stable upon statistical fluctuations. In this way we arrive at a 900 replica prior set.

We compare the application of the 2011 and 2015 prescriptions to Higgs production in gluon fusion, in Fig. 4.1. The agreement between the newer PDF sets improves significantly with respect to the old ones (in this particular case mainly due to the improvements in the NNPDF methodology driven by the closure test validation). The statistical combination also presents somewhat smaller uncertainties compared to the envelope procedure.

In Fig 4.2 we have verified the agreement at the PDF level. We find that the dispersion between the individual PDFs entering the combination is comparable to each of the individual PDF uncertainties, and thus generally compatible with statistical fluctuations. Additional variations can be attributed to the differences in experimental inputs, fitting methodology, and theory settings.

4.4 The final PDF4LHC deliverables

4.4.1 The PDF4LHC PDF sets

The PDF4LHC prior set uses 900 replicas, which is too many to be practical in most applications. Additionally, as discussed in Chapter 3, the Hessian error representations are preferable for many contexts. Therefore it was found that providing more compact PDF sets was a necessity. Three methods were proposed to achieve a more practical representation:

Meta PDFs [69] It is a Monte Carlo to Hessian transformation which consist on first fitting the Monte Carlo replicas to a given functional form and then applying the standard Hessian procedure, described in Sec. 3.2.2 to obtain the uncertainties.

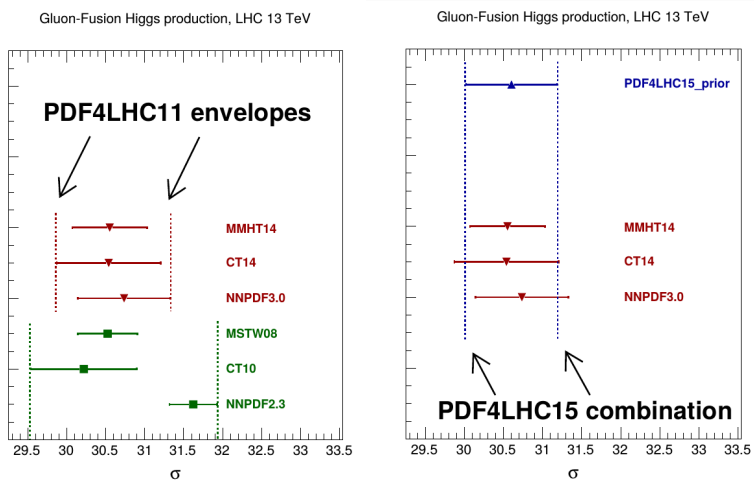


Figure 4.1: Comparison of the PDF uncertainties in the Higgs production in gluon fusion employing the 2011 PDF4LHC prescription based on envelopes (left) and the 2015 based on a combined dataset. The three PDF predictions for MMHT14, CT14 and NNPDF3.0 enter the 2015 prescription, and the older 2011 prescription was based on MSTW08, CT10, and NNPDF2.3. The central values and uncertainties using the combined PDF4LHC15 set, labeled as PDF4LHC15_prior, are displayed on the top of the figure on the right. The figures have been taken from Ref. [103].

Compressed Monte Carlo (CMC) [70] It is a Monte Carlo compression technique that minimizes the number of replicas while preserving a number of statistical estimator in the original sample.

MCH [13] The Monte Carlo to Hessian transformation described in Sec 3.3.

The three methods are used to implement different PDF4LHC combined sets: The PDF4LHC15_30 sets are based on Meta PDF, the PDF4LHC15_mc are based on the compressed Monte Carlo approach and PDF4LHC15_100 is based on MCH. The PDF4LHC15_100 and PDF4LHC15_mc contain 100 error members, while PDF4LHC15_30 contains 30.

The properties of these sets were thoroughly benchmarked during the preparation of the 2015 recommendation (a large number of comparison plots is archived in Ref. [104]), and also subsequently in Ref [15].

4.4.2 Comparison of Hessian reductions

For the two Hessian transformations it is trivial to reproduce the central values of the prior set. The only other relevant quantity in the Hessian approximation is the covariance matrix in PDF space, Eq. 3.3.2. Since the MCH method directly optimizes the agreement with the covariance matrix, it is expected that it will perform better than Meta PDF in this regard. As shown in Fig 4.1 this is indeed the case. We see that the correlation matrix (which is displayed instead of the covariance so that

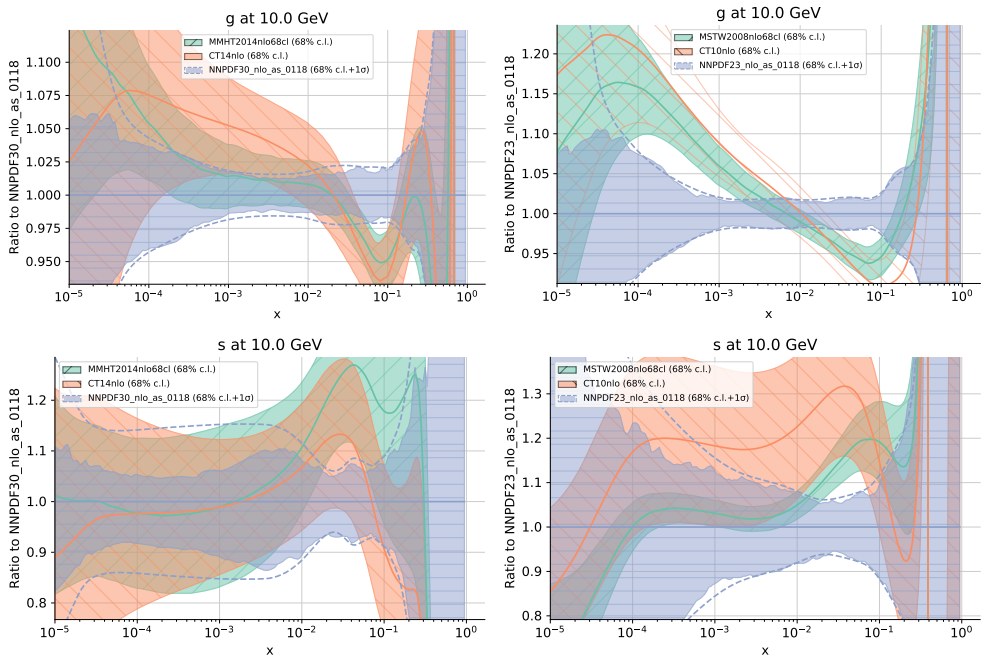


Figure 4.2: Comparison between newer (left) and older (right) versions of the PDF set entering the PDF4LHC recommendation. The upper plots display the gluon PDF and the lower the strange quark.

the total uncertainties are normalized away) is more closely reproduced in the MCH method when the number of error members is large enough. When a small number of eigenvectors is desired (around 30), the two methods perform similarly (MCH fails to reproduce the regions at large x where the covariance matrix is numerically small while the differences for Meta-PDF is somewhat more spread in all kinematical regions).

We now compare the performance of the three reduced sets at NLO for all the hadronic cross sections included in the NNPDF3.0 analysis [8].

The predictions have been computed at $\sqrt{s} = 7$ TeV using NLO theory with `MCFM` [105], `NLOjet++` [106] and `aMC@NLO` [107, 108] interfaced to `APPLgrid` [57]. The dataset we are considering contains $N_\sigma \simeq 600$ data points for electroweak gauge boson, jet production and top quark pair production. We display the results on the (x, Q) plane, associating leading order kinematics to each process (see Sec. 5.2.2). We assess the compressed methods on the relative difference between the standard deviation, $s_i^{(\mathrm{red})}$, of the cross-section σ_i computed with the reduced sets, and that of the prior, $s_i^{(\mathrm{prior})}$:

$$\Delta_i \equiv \frac{\left| s_i^{(\mathrm{prior})} - s_i^{(\mathrm{red})} \right|}{s_i^{(\mathrm{prior})}}, \quad i = 1, \ldots, N_\sigma \,. \tag{4.4.1}$$

Here $s_i^{(\mathrm{prior})}$ have been compute employing all the 900 replicas of the prior (using Eq. 3.2.2) while the $s_i^{(\mathrm{red})}$ have been computed with each of the reduced sets (using

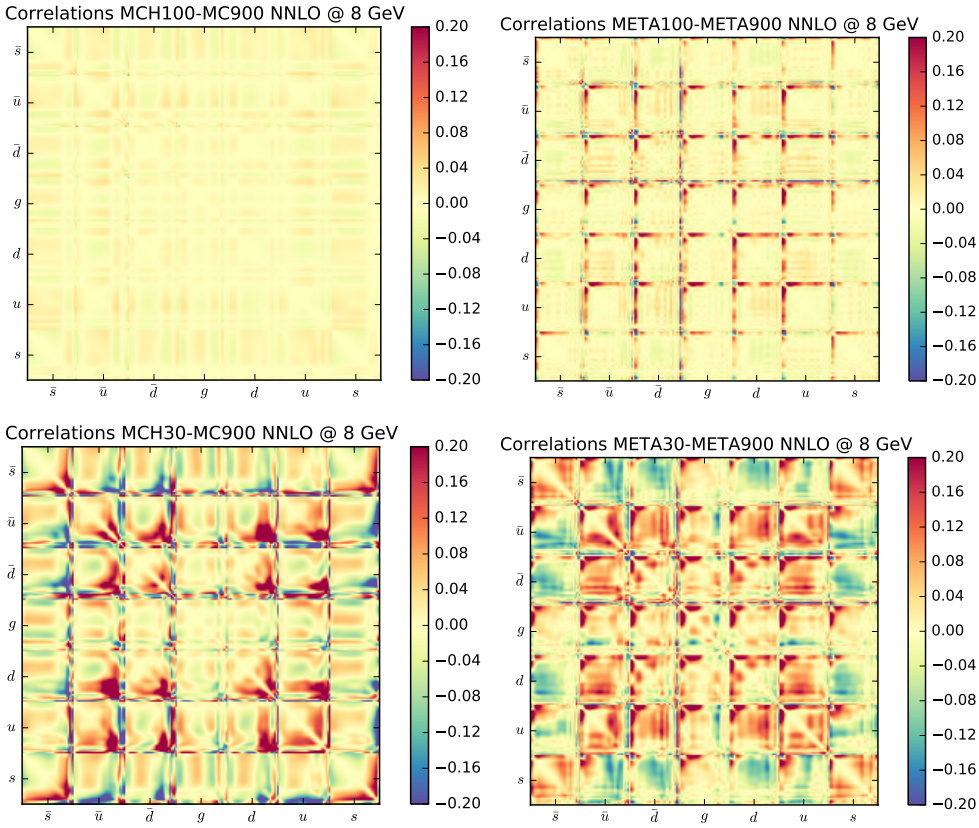


Figure 4.1: Differences in the correlation matrix between the Hessian representations of the PDF4LHC prior set and the prior itself. The MCH representation is on the left and the META PDF on the right. We show the results for 100 eigenvectors (up) and 30 (down).

again Eq. 3.2.2 for PDF4LHC_nlo_mc and Eq. 3.2.16 for PDF4LHC_nlo_30 and PDF4LHC_nlo_100).

We have presented the results in Fig. 4.2. We find that the PDF4LHC_nlo_mc and PDF4LHC_nlo_100 reproduce the uncertainties to better than around 30% in all the cases, while PDF4LHC_nlo_30 can cause deviations bigger than 50% for certain outliers. This may warrant some caution when using these sets in a context where PDF uncertainties are important.

The results we have obtained suggest that it may be interesting to test the performance of the SM-PDF based sets (see Sec. 3.4) as general purpose reduced sets. While they would, by construction, fail to reproduce observables they were not optimized for, it may be the case that a sufficiently inclusive input data set (see Sec. 3.4.2) leads to a similar performance as the Meta PDF sets, but with a more explicit control on which observables should be expected to work (and obviously the possibility to alter the input to suit a particular application while retaining the possibility to combine uncertainties

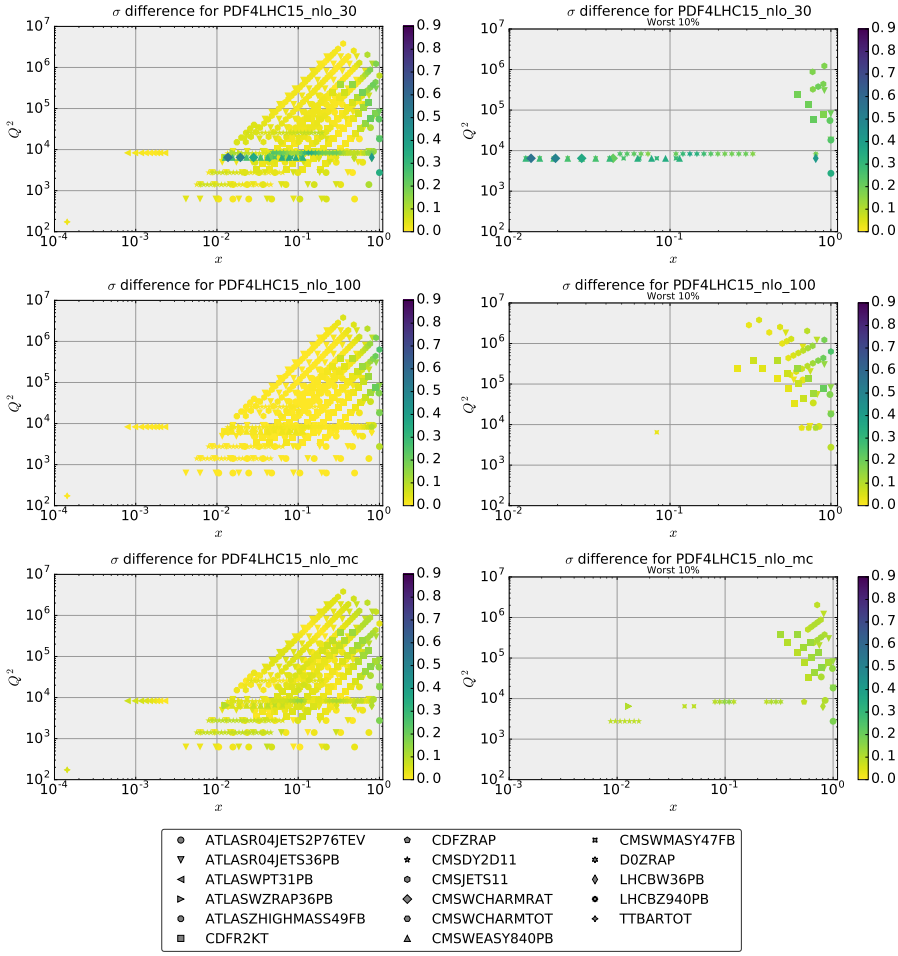


Figure 4.2: Relative difference Eq. (4.4.1), between the PDF uncertainties computed using the reduced set and the prior computed for all hadronic observables included in the NNPDF3.0 fit, shown as a scatter plot in the (x, Q^2) at the corresponding point, determined using leading-order kinematics. From top to bottom results for PDF4LHC_nlo_30, PDF4LHC_nlo_100 and PDF4LHC_nlo_mc are shown. In the left, all the points are shown, while on the right, we display only the 10% of points with maximal deviation.

later). We have constructed the so called SM-PDF-Ladder PDF set at NLO using the theoretical predictions in Table. 4.1, calculated with the `aMC@NLO` code. We set the tolerance parameter T_R, Eq. 3.4.2 to 5% and obtain a set with 17 eigenvectors. These are the same settings as in Sec 3.3 of Ref. [14]. In Fig. 4.3 we compare the Δ_i ratio Eq 4.4.1 for the NNPDF3.0 dataset, like we did for the PDF4LHC sets (in Fig. 4.2). The results are almost equivalent to those for PDF4LHC_nlo_30, even though the SM-PDF-Ladder sets has half the number of error sets. This shows that, even in its most unspecialised form, the SM-PDF methodology can provide a competitive relation

process	distribution	$N_{\rm bins}$	range
$gg \to h$	$d\sigma/dp_t^h$	10	[0,200] GeV
	$d\sigma/dy^h$	10	[-2.5,2.5]
VBF hjj	$d\sigma/dp_t^h$	5	[0,200] GeV
	$d\sigma/dy^h$	5	[-2.5,2.5]
hW	$d\sigma/dp_t^h$	10	[0,200] GeV
	$d\sigma/dy^h$	10	[-2.5,2.5]
hZ	$d\sigma/dp_t^h$	10	[0,200] GeV
	$d\sigma/dy^h$	10	[-2.5,2.5]
$ht\bar{t}$	$d\sigma/dp_t^h$	10	[0,200] GeV
	$d\sigma/dy^h$	10	[-2.5,2.5]

process	distribution	$N_{\rm bins}$	range
Z	$d\sigma/dp_t^{l^-}$	10	[0,200] GeV
	$d\sigma/dy^{l^-}$	10	[-2.5,2.5]
	$d\sigma/dp_t^{l^+}$	10	[0,200] GeV
	$d\sigma/dy^{l^-}$	10	[-2.5,2.5]
	$d\sigma/dp_t^Z$	10	[0,200] GeV
	$d\sigma/dy^Z$	5	[-4,4]
	$d\sigma/dm^{ll}$	10	[50,130] GeV
	$d\sigma/dp_t^{ll}$	10	[0,200] GeV

process	distribution	$N_{\rm bins}$	range
$t\bar{t}$	$d\sigma/dp_t^{\bar{t}}$	10	[40,400] GeV
	$d\sigma/dy^{\bar{t}}$	10	[-2.5,2.5]
	$d\sigma/dp_t^t$	10	[40,400] GeV
	$d\sigma/dy^t$	10	[-2.5,2.5]
	$d\sigma/dm^{t\bar{t}}$	10	[300,1000]
	$d\sigma/dp_t^{t\bar{t}}$	10	[20,200]
	$d\sigma/dy^{t\bar{t}}$	12	[-3,3]

process	distribution	$N_{\rm bins}$	range
W	$d\sigma/d\phi$	10	[0,200] GeV
	$d\sigma/dE_t^{\rm miss}$	10	[-2.5,2.5]
	$d\sigma/dp_t^l$	10	[0,200] GeV
	$d\sigma/dy^l$	10	[-2.5,2.5]
	$d\sigma/dm_t$	10	[0,200] GeV
	$d\sigma/dp_T^W$	5	[-4,4]
	$d\sigma/y^W$	10	[50,130] GeV

Table 4.1: LHC processes and the corresponding differential distributions used as input in the construction of the SM-PDF-Ladder set. In each case we indicate the range spanned by each distribution and the number of bins $N_{\rm bins}$. All processes have been computed for $\sqrt{s} = 13$ TeV. Higgs bosons and top quarks are stable, while weak gauge bosons are assumed to decay leptonically. No acceptance cuts are imposed with the exception of the leptons from the gauge boson decay, for which we require $p_T^l \geq 10$ GeV and $|\eta^l| \leq 2.5$.

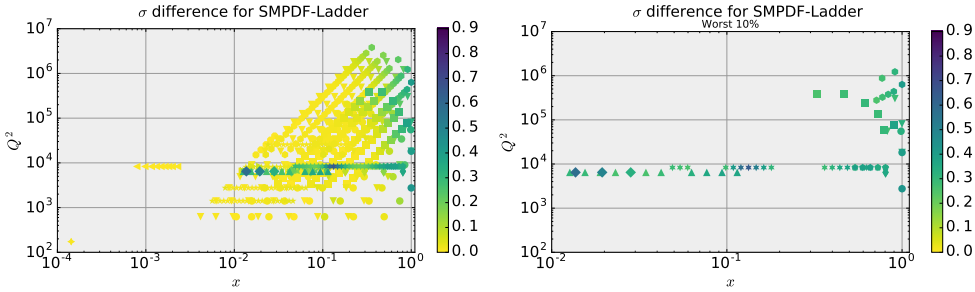


Figure 4.3: Same as Fig. 4.2 but testing the SM-PDF-Ladder PDFs.

between computational efficiency and accuracy in reproducing the uncertainties.

4.4.3 Gaussianity of the PDF4LHC predictions

While the CMC method is expected to perform worse than the Hessian based reductions when reproducing purely Gaussian properties, such as correlations and standard deviations (when compared with a fixed number of error members), it may be advantageous to employ the CMC compressed sets for observables where the distribution of the predictions of the replicas of the prior PDF4LHC set is non Gaussian. Here

we reproduce some of the results from Ref [15] that shed light on the situations when using the PDF4LHC15_mc set is advantageous.

In order to quantify the degree of Gaussianity of the predictions, we first transform the Monte Carlo sample (one value of the observable for each of the 900 replicas) into a continuous probability distribution. We then compare that probability distribution with a Gaussian with the same mean and standard deviation as the sample as well as the distributions obtained for the MCH and CMC compressed sets. The first step is accomplished using the Kernel Density Estimate (KDE) method. The second, using the KullbackLeibler (KL) divergence as a measure of the difference between two probability distributions (for a brief review of both methods see e.g. Ref. [109]).

The KDE method consists of constructing the probability distribution corresponding to a sample as the average of kernel functions K centered at each point in the sample. In our case, given $k = 1, \ldots, N_{\rm rep}$ replicas of the i-th cross-section $\{\sigma_i^{(k)}\}$, the probability distribution is

$$P(\sigma_i) = \frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} K\left(\sigma_i - \sigma_i^{(k)}\right) , \quad i = 1, \ldots, N_\sigma \, . \tag{4.4.2}$$

We specifically choose

$$K(\sigma - \sigma_i) \equiv \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{(\sigma - \sigma_i)^2}{h}\right) , \tag{4.4.3}$$

where the parameter h, known as bandwidth, is set to

$$h = \hat{s}_i \left(\frac{4}{3N_{\rm rep}}\right)^{\frac{1}{5}} , \tag{4.4.4}$$

where $\hat{s}_i$ is the standard deviation of the given sample of replicas. This choice is known as *Silverman rule*, and, if the underlying probability distribution is Gaussian, it minimizes the integral of the square difference between the ensuing distribution and this underlying Gaussian [110].

The KullbackLeibler divergence measures the information loss when using a probability distribution $Q(x)$ to approximate a prior $P(x)$, and is given by

$$D_{\rm KL}^{(i)}(P|Q) = \int_{-\infty}^{+\infty} \left(P(x) \cdot \frac{\log P(x)}{\log Q(x)}\right) dx. \, , \tag{4.4.5}$$

We have studied the Gaussianity of the observables listed in Table 4.1. We have proceeded as follows: For each cross section we have obtained a KDE representation of the prior distribution (i.e. the predictions for the 900 replicas in the prior) and we have compared (by computing the KL divergence Eq. 4.4.5) them to the KDE estimate of the distribution of predictions obtained with the PDF4LHC15_mc set, the Gaussian distribution obtained directly PDF4LHC15_100 (see Sec. 3.2.2) and a Gaussian that has the same mean and standard deviation as the prior sample. We present the results for all the cross sections in Fig. 4.4. We see that the MCH method reproduces the prior distribution essentially as well as possible within a Gaussian approximation, since the values of the KL divergence between PDF4LHC15_100 and

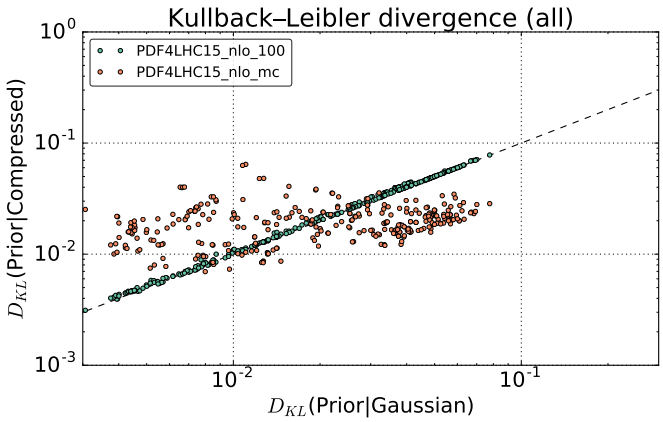


Figure 4.4: The KL divergence, Eq. (4.4.5) between the prior and each of its two reduced representations PDF4LHC15_nlo_prior (Monte Carlo) and PDF4LHC15_nlo_mc (Hessian) vs. the divergence between the prior and its Gaussian approximation, computed for all observables listed in Table 4.1.

the prior and practically equal to those of the Gaussian approximation of the prediction. In Ref. [15] we provided an intuitive way to understand the absolute values of the KL divergences and found them to correspond to generally good agreement with the Gaussian approximation (corresponding to reproducing the PDF uncertainties to about 20%). We find that the performance of the CMC compression (vertical axis of Fig. 4.4) is largely uncorrelated with the degree of Gaussianity (horizontal axis), and indeed using the CMC compression proves advantageous for some observables (for which we have $D_{\rm KL}$(Prior|PDF4LHC_nlo_mc) $< D_{\rm KL}$(Prior|Gaussian)).

In order to find out which observables are better described using the CMC compression, we break down the results in Fig. 4.4 by process. Figure 4.5 that the Monte Carlo compressing is advantageous for a significant fraction of the W and Z production data, but not for top and Higgs production. This is consistent with the expectation that non-Gaussian behaviour is mostly to be found in large x PDFs, which are probed by gauge boson production at high rapidity, but not by Higgs and top production which are mostly sensitive to the gluon PDF at medium and small x.

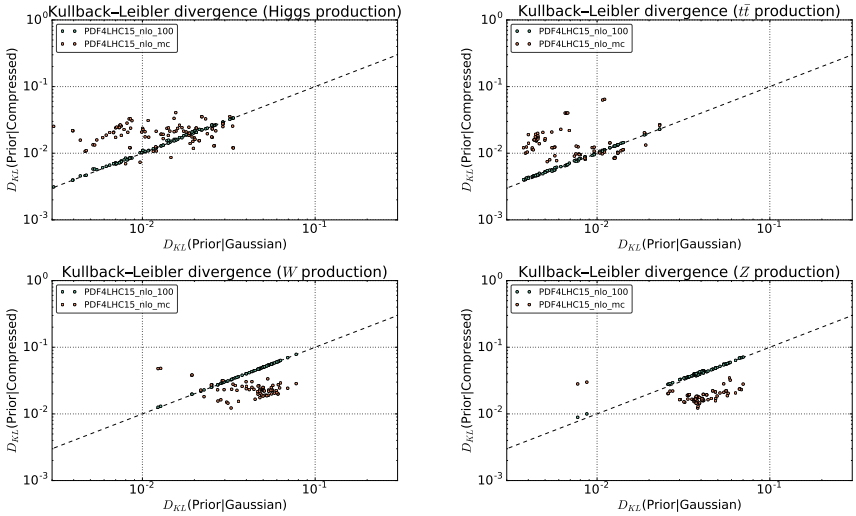


Figure 4.5: Same as Fig. 4.4, now separating the contributions of the different classes of processes of Table 4.1: Higgs production (top left), top quark pair production (top right), W production (bottom left) and Z production (bottom right).

Chapter 5

NNPDF 3.1

NNPDF3.1 [8] is the latest global set of the NNPDF collaboration. Two main developments motivate its release. Firstly the inclusion of new collider data that provides important constraints on the PDFs (see Sec. 5.2.2) thereby increasing the precision of the PDF determination. Secondly, a formalism to fit the charm PDF in the same way as the light quark PDFs within the FONLL scheme, was developed in Refs [111, 47] and implemented in a PDF fit for the first time in Ref [11]. Allowing the charm to be freely parametrized leads to improvements in the fit quality and stabilizes the dependence of the PDFs on the charm mass (see Sec. 5.3.2).

NNPDF3.1 uses the same fitting algorithm as NNPDF3.0, which is briefly described in Sec. 5.1. The adequacy of the methodology was established by means of closure test, described in Sec. 5.1.9. The main features of the resulting PDFs are described in Sec. 5.3. The main challenge in the development of NNPDF3.1 was the increased precision targets driven by the new high precision collider data and the corresponding state of art theory predictions at NNLO. Two remarkable examples of the issues that appeared are presented in Sec. 5.4. The necessity to challenge every aspect of the methodology and the input experimental and theoretical predictions demanded the development of a new brand of analysis tools, described in Sec. 5.5 that extend and complement the existing fitting code, described in Ref. [112], and explore concepts in Computer Science such as Functional Programming, Contract Programming and compiler technology.

NNPDF3.1 takes advantage of the tools presented in Chap 3 to provide compressed sets with higher statistics. Specifically, the default sets are compressed Monte Carlo and Hessian sets of 100 error members obtained from fits with 1000 replicas.

5.1 The NNPDF fitting methodology

The NNPDF fitting methodology for NNPDF3.0 is described in detail Ref.[8]. The fitting algorithm is essentially unchanged in NNPDF3.1. Here we only briefly refer to some aspects of the methodology that are more relevant for this work.

5.1.1 PDF parametrization

As we discussed in Sec.2.6, it is enough to parametrize the PDFs at a fixed scale, since they can be related to any other via DGLAP evolution. In NNPDF3.1, we selected $Q = 1.65$GeV so that the PDFs run always above the charm mass when it is independently parametrized.

Compared to NNPDF3.0, the change is that the charm PDF is added to the basis of independently parametrized flavour combinations. The fitted basis is, in terms of quarks and gluons,

$$g\,, \tag{5.1.1}$$
$$\Sigma = \sum_{u,d,s} q_i + \overline{q}_i\,, \tag{5.1.2}$$
$$T_3 = u^+ - d^+\,, \tag{5.1.3}$$
$$T_8 = u^+ + d^+ - 2s^+\,, \tag{5.1.4}$$
$$V = \sum_{u,d,s} q_i - \overline{q}_i\,, \tag{5.1.5}$$
$$V_3 = u^- - d^-\,, \tag{5.1.6}$$
$$V_8 = u^- + d^- - 2s^-\,, \tag{5.1.7}$$
$$c\,. \tag{5.1.8}$$

Since each PDF combination is characterized by in a very flexible functional form, the choice of basis has little effect on the resulting PDFs, but however can accelerate the convergence of the procure in practice, as was explicit shown in Ref. [8].

Each of the combinations is parametrized in terms of a neural network times a preprocessing term:

$$f_i(x) = A_i x^{-\alpha_i}(1-x)^{\beta_i}\mathrm{NN}_i(x) \tag{5.1.9}$$

A_i is a normalization term that is used to fix the value of the sum rules: Four of these values are fixed at each iteration of the fit as follows:

$$A_g = \frac{1 - \int_0^1 dx x\Sigma(x)}{\int_0^1 dx x\hat{g}(x)}\,, \tag{5.1.10}$$

$$A_V = \frac{3}{\int_0^1 dx \hat{V}(x)}\,, \tag{5.1.11}$$

$$A_{V_3} = \frac{1}{\int_0^1 \hat{V}_3(x)}\,, \tag{5.1.12}$$

$$A_{V_8} = \frac{1}{\int_0^1 \hat{V}_8(x)}\,, \tag{5.1.13}$$

where the notation $\hat{f}(x)$ stands for the unnormalized PDFs. The rest of the normalization constants are set to 1. The preprocessing factor $x^{-\alpha_i}(1-x)^{\beta_i}$ represents the dominant behaviour of the PDF.

The preprocessing exponents α_i and β_i are determined following an iterative procedure where the exponents of a new fit are set from the *preferred exponents* of a previous one. Specifically, we choose points at sufficiently low x for α (at 10^{-6} to 10^{-3}, except for the gluon and singlet where we only use 10^{-6}) and sufficiently high x for β (at 0.6 to 0.95) and compute

$$\alpha_{\text{eff},i}(x) = -\frac{\log f_i(x)}{\log x} \tag{5.1.14}$$

$$\beta_{\text{eff},i}(x) = \frac{\log f_i(x)}{\log(1-x)} \tag{5.1.15}$$

We use these values to set the α_i and β_i parameters from Eq. 5.1.9 for the next iteration: specifically we draw the values from an uniform distribution centered at the mean of $\alpha_{i,\text{eff}}$ and $\beta_{i,\text{eff}}$ over the set of replicas of the previous fit and spanning twice the standard deviation in each direction. We impose limits that guarantee the finiteness of the momentum fraction when the neural network part is constant (that is α_i>2). The resulting values are used to fix the preprocessing exponents of each replica for the next iteration of the fit. The procedure is repeated until the distributions of the preprocessing exponents do not change significantly. Usually convergence is attained after the first or the second iteration, depending on how big the changes in the respective fits were. While it would be desirable to arrive at a procedure that does not require iterating the preprocessing, currently it has a number of advantages: The PDF behaviour at low and large x cannot be described by experimental data and thus some extra knowledge is required to fix it. Additionally, the convergence of the algorithm improves significantly when the dominant part of the PDF behaviour is determined by preprocessing.

Let us briefly introduce some common nomenclature related to neural networks: A neural network can be viewed as a directed graph where each node is either an *input* or an *activation* node. The activation nodes have each associated an *activation function* and the edges indicate that the result of one node is to be used as output for another, as indicated by the direction of the edge. The input nodes represent the external inputs (e.g. the data to be fitted). The *output nodes* characterize the value of the network. Each node (labeled by I) is additionally associated to a threshold θ^I, and each edge connecting the output of the node J to the input of the node I is associated to a *weight*, w_{JI}. Thresholds and weights constitute the *parameters* of the neural network. The activation function of a given node, often takes a single parameter, a, constructed by as the weighted sum of the inputs of the node,

$$a_I = \sum_J^{\text{inputs of } I} w_{JI}\xi_J + \theta_I \tag{5.1.16}$$

where ξ_J is the value of the activation function $g_J(a_J)$ for the node J,

$$\xi_J = g_J(a_J)\ . \tag{5.1.17}$$

A *feed forward* neural network restricts the graph to be acyclic. A multilayer perceptron restricts each node to belong to one element of an ordered list of *layers*. Then, edges can only exists between consecutive layers. The first layer is the *input layer* and

the last is the *output layer*. Any layer in between is called *hidden layer*. In terms of the layer index l and the indexes of the node within the current (i) and previous (j) layer, the value of the activation nodes is

$$\xi_i^{(l)} = g\left(\sum_j^{\text{inputs}} w_{ij}^{(l)} \xi_j^{(l-1)} + \theta_i^l\right) \ . \tag{5.1.18}$$

The index l runs over layers and the indexes i and j over the nodes of the l-th and $(l-1)$-th layer respectively, starting from the first hidden layer. A multilayer perceptron is *fully connected* when all possible edges given the aforementioned restrictions are present.

The neural network we use is a feed forward, fully connected, multilayer perceptron. The architecture, represented in Fig 5.1, is by default 2-5-3-1. The first layer contains

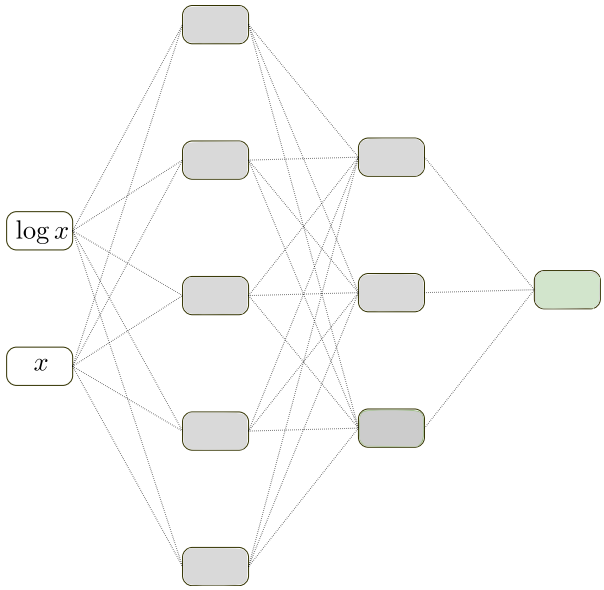


Figure 5.1: A representation of the default neural network used in the fit. The inputs are x and $\log x$. The two hidden layers have a logistic activation function, and the final visible layer, a linear one.

2 input nodes which take x and $\log x$ respectively. The two hidden layers have a logistic activation function:

$$g(a) = \frac{1}{1+\exp^{-a}} \ , \tag{5.1.19}$$

while the last (output) layer has a linear activation function, allowing the final result of the neural network to acquire values outside $(0, 1)$.

$$g(a) = a \ . \tag{5.1.20}$$

The weights $w_{ij}^{(l)}$ and thresholds θ_i^l are the parameters to be determined by the minimization algorithm.

5.1.2 Experimental uncertainties

We now wish to construct the error function to be minimized when fitting PDFs. All the experimental data used in NNPDF comes in the form of a central value for each point, a statistical uncertainty (that is always uncorrelated) and a list of possibly correlated systematic uncertainties. The systematic uncertainties can be *additive* or *multiplicative*. Multiplicative systematic uncertainties are proportional to the value of the observable, while additive ones do not depend on it. We can construct a covariance matrix between two data point i and j as

$$\mathrm{cov}_{ij} = \delta_{ij}\sigma_i^{\mathrm{uncorr}}\sigma_j^{\mathrm{uncorr}} + \sum_s^{\mathrm{additive}} \sigma_{i,s}\sigma j, s + \left(\sum_p^{\mathrm{multiplicative}} \sigma_{i,p}\sigma_{j,p}\right) T_i T_j \ . \quad (5.1.21)$$

The uncorrelated uncertainties contain the statistical ones as well as other correlated systematics. The multiplicative uncertainties are proportional to the value of the corresponding data, T_i. Since the experimental data are assumed to have a multigaussian distribution, a natural candidate for error function is the χ^2 statistic defined as

$$\chi^2 = \sum_{ij}^{N_{\mathrm{data}}} (D_i - T_i)\mathrm{cov}_{ij}^{-1}(D_j - T_j) \ , \quad (5.1.22)$$

where T_i are the theory predictions that are varied to attain the minimum. This however yields inconsistent results. It is well known [113] that the minimum of Eq.5.1.22 does not correspond to an unbiased estimator for the location of the underlying multigaussian distribution of the data, and in practice the difference is large. The problem can be eliminated using the $t0$ prescription [114] in the fit instead. The procedure consists on making the normalization uncertainties depend on the result of a previous fit, and keeping the covariance matrix fixed during the minimization. The result is a covariance matrix identical to Eq. 5.1.21 with the exception that the normalization systematics are not normalized by the current theoretical predictions (that change as the fit progresses) but are fixed to the predictions of a previous fit. Replacing it in Eq. 5.1.22 leads to an objective function that, after the convergence of the procedure, yields unbiased best fit values, as illustrated in Ref. [114].

In the NNPDF fits, the t_0 covariance is iterated at the same time as preprocessing, and the convergence is similarly achieve after less than 3 iterations in most cases (depending on how close the original t_0 PDF we started from is to the current fit configuration).

5.1.3 Positivity constraints

Positivity constraints are a further addition to the target error function. While PDFs at leading order represent probability densities and are therefore positive definite, this is not the case beyond LO [115]. This constitutes an important difficulty in PDF fits: We must allow PDFs to go below zero while at the same time ensuring the positivity of any observable cross section. The experimental data that is fitted obviously provides some positivity constraints, but this however is not enough to guarantee that all possible observables are positive definite, particularly those sensitive to the high or low x regions such as high energy jets or heavy partners in New Physics

searches. We therefore consider quantities that can be predicted by the theory and must be positive, but for which we do not have experimental data. These include the longitudinal structure functions $F_L(x, Q^2)$ on a grid in x, the dimuon cross section from Ref [116], and *tagged* deep-inelastic structure functions and Drell-Yan rapidity distributions, defined by setting to zero the electric charge of all quarks but one [8]. Additionally we include the gluon fusion process for a Higgs-like observable with a mass of 5GeV.

For each positivity observable we add a contribution to the error function that is activated when the observable becomes negative:

$$\Delta_{\rm pos} = -\lambda_{\rm pos} \sum_{i}^{N_{\rm data}} \Theta(-O_{\rm pos,i}) O_{\rm pos,i} \ , \tag{5.1.23}$$

where Θ is the indicator function. The value of $\lambda_{\rm pos}$ is determined as 0.25 times the prediction with the PDF chosen as t0 (see Sec.5.1.2). We take the absolute value if the prediction is negative.

The positivity predictions for the structure functions are computed using APFEL [39], while the Higgs based constraints are obtained from `aMCfast` [117].

This strategy is rather expensive computationally for its effectiveness, since the convolutions with the positivity observables account for around 25% of the time in the fit and require a larger number of iterations of the minimization algorithm to each convergence, with convergence worsening rather quickly as new positivity observables are added which in turn limits the amount we can add. Yet the penalization terms in Eq. 5.1.23 do not by themselves guarantee in practice that all the replicas will predict a positive value for the observables that are included (or indeed any other observable that an user might be interested in). Positivity is hard to guarantee within the NNPDF framework when the PDF uncertainty of the observable is comparable in magnitude with its value. From the user's perspective, the most consistent way of treating the results is manually setting to zero the PDF prediction for any replica that turns out to be negative. Note that the fact that many replicas are negative means that the observable is *consistent with zero within uncertainties* which is in fact an adequate characterization.

Investigations on possible improvements of the positivity minimization strategy (for example using a log-barrier method [118]) suggested that an improved treatment of the positivity requires an upgrade of the current minimization strategy based on a genetic algorithm, described next in Sec. 5.1.7. Such improved minimization strategies are currently under investigation.

5.1.4 Cross validation

The NNPDF parametrization is chosen to be flexible enough to adapt to any reasonable functional form for the PDF as extensively tested in Ref. [114]. This allows a conceptually simpler and more rigorous treatment compared to approaches bases on obtaining the parametrizations based on some physical consideration, since one can allow the fit to find the best functional form without any prejudice. There is however the risk that the fit becomes sensitive to noise in the experimental data as opposed to genuine features of it. This effect is called *overlearning*. To address the problem introducing as little bias as possible, NNPDF fits utilize a *cross validation* procedure.

The idea is to split the experimental data in two sets roughly equal in size called *training* and *validation* sets. The fit algorithm only optimizes the training set and it never tests any of the validation data. However the best parametrization is selected at the point where the unseen validation set has the smallest error function, thereby ensuring that optimizing noise in the training set does not bias the final result of the fit.

The current strategy is to perform the minimization for a fixed number of iterations and *look back* at the parameter value where the validation error function was smallest. This is in general not the latest iteration since at some point improving the training score has a negative effect on the validation. In NNPDF3.1 the number of minimization iterations (see Sec. 5.1.7 next) is 50000.

At the end of the fit, the training and validation scores of the error function are very similar, as we would expect if the fit procedure is working properly. This is shown in Fig. 5.2.

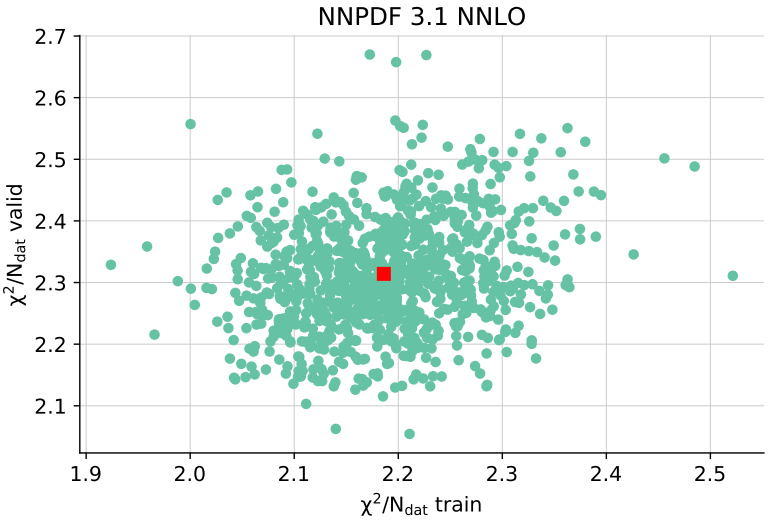


Figure 5.2: Distribution of training and validation error function for the 1000 replica NNPDF 3.1 NNLO default set. Each circle corresponds to one replica and the square is the average value. The training and validation scores are comparable.

5.1.5 Pseudodata generation

One of the most salient features of the NNPDF methodology is the ability to propagate the experimental uncertainties to the PDF fit making a minimal number of assumptions: To perform a fit, we first sample N_{rep} realizations from the probability distribution that we construct from the experimental inputs. We call these samples *pseudodata replicas*. Each of the pseudodata replicas is then fitted to a set of functional forms for each fitted combination as described in Sec 5.1.1. We call each of these sets of functions a *PDF replica*, or simply *replica*. For each of the $N_{\mathcal{D}}$ points entering the

fit, the pseudodata replica samples are obtained as from

$$D_I = D_I^0 + \sum_J^{N_\mathcal{D}} C_{IJ}^{\frac{1}{2}} \delta_J \ , \tag{5.1.24}$$

where D_I is the data point indexed by I, D_I^0 is the corresponding experimentally measured central value, $C^{\frac{1}{2}}$ is the transpose of the Cholesky decomposition of the t0-covariance matrix C (see Sec.5.1.2) and δ_J is a random number sampled from a standard normal distribution. We have $I = 1 \dots N_\mathcal{D}$. For positive observables, D_I is restricted to be positive.

5.1.6 Target error function

We have now described all the elements of the target error function to be optimized: Fitting a given PDF replica consists on performing the minimization of the error function χ^2 as a function of the set of parameters that characterize the PDF functional form, $\{\theta\}$, namely the thresholds and weights of the neural networks (see Sec.5.1.1):

$$\chi^2 \left[\mathcal{T}\left[f(\{\theta\})\right], \mathcal{D}\right] = \frac{1}{N_\mathcal{D}} \sum_{I,J} \left(T_I[f(\{\theta\})] - D_I\right) C_{IJ}^{-1} \left(T_J[f(\{\theta\})] - D_J\right) + (\text{positivity}) \ , \tag{5.1.25}$$

where $T_I[f]$ is the theoretical prediction for the experimental point I using the PDF replica f and D_I is the fluctuated pseudodata experimental measurement (constructed following Eq. 5.1.24). The positivity terms were described in Sec.5.1.3. The cross validation procedure described in Sec. 5.1.4 is employed to avoid overfitting during the minimization. Labeling it cv min, fitting a replica means performing the operation

$$\chi^{2\,\text{min}} = \text{cv} \min_{\{\theta\}} \chi^2 \left[\mathcal{T}\left[f(\{\theta\})\right], \mathcal{D}, \xi\right] \ , \tag{5.1.26}$$

and retrieving the PDF replica from the parametrization that minimizes the error. We have introduced a parameter ξ characterizing the random state of the algorithm (i.e. it is the seed of the random number generator). It affects the results because of the selection of the cross validation split, the finite efficiency of the genetic minimization algorithm and the existence of equivalent local minima. That is, while the error function Eq 5.1.25 does not depend on the random state, the minimum we find in practice does.

5.1.7 Minimization algorithm

NNPDF3.1 uses the same *nodal mutation* genetic algorithm of NNPDF3.0. The algorithm minimizes the target training error function Eq. 5.1.25 as a function of the parameter of the neural network. First all the parameters are initialized by sampling from a standard normal distribution.

Then N_{mut} *mutants* are generated by assigning each node of the neural network a probability of being mutated. If a node is selected, its parameters (both weights and thresholds) are changing following the formula

$$\theta \to \theta + \frac{\eta r_\delta}{N_{\text{ite}}^{r_{\text{ite}}}} \ , \tag{5.1.27}$$

where η is a constant baseline size of the mutation, r_δ is a random uniform number between -1 and 1, different for each parameter, $N_{\rm ite}$ is index of the current iteration. It was found that setting $\eta = 15$ and the mutation probability for a node being selected is 5%.

Once the mutants are generated, the goodness of fit on the training set is computed for each of them. The best mutant with the best store is selected and is used as baseline for the next iteration. The procedures is repeated for a fixed number of iterations.

As discussed in Sec.5.1.4, in the end, the parametrization with best overall validation score is selected from the set of best mutants obtained in each iteration.

5.1.8 Post selection of replicas

The genetic algorithm described in Sec. 5.1.7 does not always find a replica of acceptable quality. We implement a set of post selection criteria that ensure that replicas that failed to converge are discarded. These criteria are:

Error function We discard replicas for which the value of the total error function is bigger that the mean one plus 4 standard deviations. These are unlikely to correspond to pseudodata outliers in a 1000 replica fit, and instead are likely to reflect a poorly converged minimization. Indeed assuming the error function is distributed following a Gaussian distribution, the probability of seeing one or more replicas that is more that 4 sigma away from the mean value in a 1000 replica fit is

$$P = 1 - (1 - 2\text{CDF}(4))^{1000} \approx 0.06 \ , \tag{5.1.28}$$

where with CDF we have indicated the cumulative density function of a standard normal distribution,

$$\text{CDF}(x) = \frac{1}{2}\left(1 + \frac{2}{\sqrt{\pi}}\int_0^{\frac{x}{\sqrt{2}}} e^{-t}dt\right) \ . \tag{5.1.29}$$

We have computed one minus the probability of not seeing any outlier, which implies that it is not outside the $\pm 4\sigma$ band (the factor two accounts for the symmetric range) 1000 times. Therefore the outliers are far more likely to be a consequence of the bad performance of the fitting procedure than they are to be a statistical fluctuation.

Arc-Length Similarly we discard replicas where the arc-length fluctuates more than 4 standard deviation w.r.t the mean value. In this way we discard replicas that are too wiggly and unlikely to correspond to physical PDFs.

Positivity As we discussed in Sec. 5.1.3, our positivity setting strategy is not guaranteed to converge. We therefore discard replicas where any positivity bin is more negative than minus its corresponding $\lambda_{\rm pos}$ from Eq. 5.1.23.

The post fit selection criteria discard between around 20 and 30% of the replicas in a typical fit. Developing more reliable minimization strategies that hold these properties automatically and do not require a separate post processing step is an ongoing project.

5.1.9 Closure tests

Closure tests are a powerful tool to verify the correctness of a PDF fitting procedure. The idea is based on assuming the complete knowledge of the PDFs and testing whether the fitting procedure reproduced them. Specifically one creates *fake* experimental data that is consistent with the assumed functional forms of the PDFs and then follows the usual fitting procedure. The results of the fit are then compared to the functional form that was assumed and consistency is assessed. We define 3 *levels* of closure testing:

Level 0 The predictions on the assumed functional form are fitted directly without applying any extra fluctuation. The goal is to obtain the same functional form with negligible uncertainty, in the region where data is abundant.

Level 1 Experimental fluctuations are added on top of the theory predictions, following Eq. 5.1.24.

Level 2 Pseudodata replicas are sampled from the fluctuated data obtained in Level 1, thus reproducing the NNPDF fitting procedure described in Sec. 5.1.7. The goal is to obtain a distribution in PDF space such that the initial functional form is compatible everywhere within uncertainties.

Thus Level 0 tests the efficiency of the minimization algorithm, and Level 2 the consistency of the complete procedure, and particularly of the size of PDF uncertainties. The difference between the size of Level 2 and Level 1 uncertainties can give a rough idea of the relative size of the *functional* and *experimental* uncertainties: The fits at Level 1 always see the same data and the replicas differ only in the random seed used to initialize the algorithm. As discussed in Sec. 5.1.7, this change the cross validation splitting and is thus a way of finding equivalent minima within the efficiency of the fitting algorithm which is also included. Level 2 uncertainties also include the propagation of experimental errors.

Ref.[114] produced multiple pieces of evidence that the NNPDF methodology is successfully validated by closure tests when the starting distribution is MSTW2008 [119]. Here we only reproduce the most clear and direct evidence of the success of the procedure when it comes to reproduce the original data input, in Fig. 5.3: The χ^2 of the fluctuated data to the fitted result is comparable to the original χ^2 it was sampled from. Further evidence proved that the distances between input and fitted PDFs are compatible with statistical fluctuations thereby proving the consistence of the methodology.

5.2 Experimental and theoretical input to NNPDF 3.1

5.2.1 Overview

NNPDF3.1 features a wealth of high precision experimental data from the LHC included in a PDF determinations for the first time, including data corresponding to two original process: top quark differential distributions and the Z transverse momentum distribution.

We briefly review the complete dataset, referring the reader to Ref. [12] for more details. For each dataset, we list in Tables 5.1, 5.2, and 5.3, the published reference,

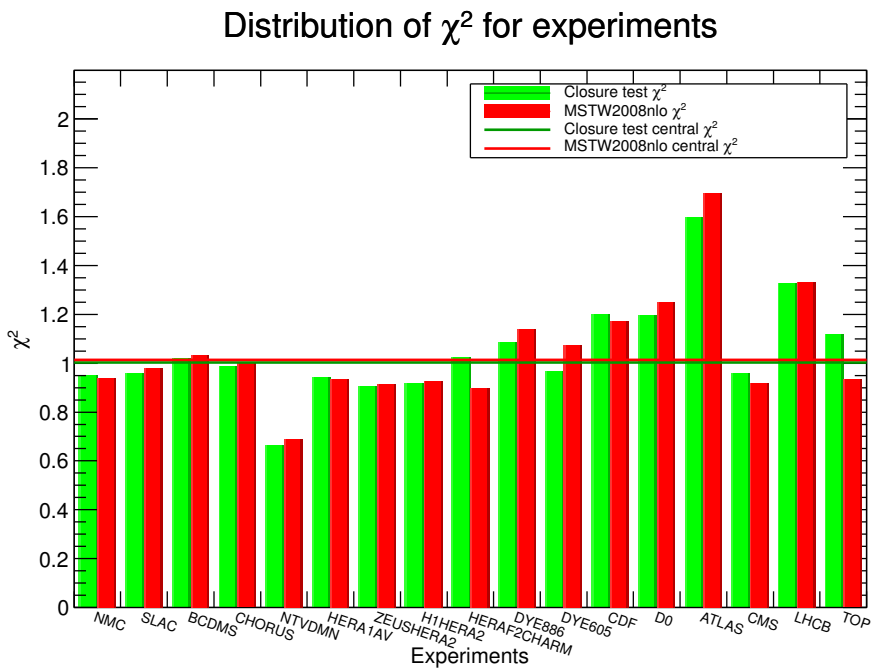


Figure 5.3: Comparison between the error function of the original input data to a Level 2 closure two fit to the original PDF set it was sample from (MSTW2008) and the corresponding NNPDF fit. The figure shows that the NNPDF methodology can attain a result that is statistically equivalent to the true value, in the data region. The figure is taken from Ref. [114].

the number of data points in the NLO/NNLO PDF determinations before and after (in parenthesis) kinematic cuts, the kinematic range covered in the relevant variables after cuts, and the code used to compute the NLO and NNLO results.

The kinamtic coverage of the NNPDF3.1 dataset in the (x, Q^2) plane is shown in Fig. 5.1 where leading order kinematics have been used for illustration purposes. The central rapidity is used when rapidity is integrated over, and we set Q^2 equal to the factorization scale.

5.2.2 New data in NNPDF3.1

We now list the new data included in NNPDF 3.1.

We have included the datasets from experiments which have now finished and have provided the final analysis of their results:

- The final HERA combined data [74] is now available and has been included in NNPDF3.1 replacing previous partial analyses. The effect of this replacement was studied in [168] and found to be small.

- We also include the H1 and ZEUS measurements of the bottoms structure function [129, 130].

Experiment	Obs.	Ref.	$N_{\mathbf{dat}}$	x range	Q range (GeV)	Theory
NMC	F_2^d/F_2^p	[120]	260 (121/121)	$0.012 \leq x \leq 0.68$	$2.1 \leq Q \leq 10$	APFEL
	$\sigma^{\mathrm{NC,p}}$	[121]	292 (204/204)	$0.012 \leq x \leq 0.50$	$1.8 \leq Q \leq 7.9$	
SLAC	F_2^p	[122]	211 (33/33)	$0.14 \leq x \leq 0.55$	$1.9 \leq Q \leq 4.4$	APFEL
	F_2^d	[122]	211 (34/34)	$0.14 \leq x \leq 0.55$	$1.9 \leq Q \leq 4.4$	
BCDMS	F_2^p	[123]	351 (333/333)	$0.07 \leq x \leq 0.75$	$2.7 \leq Q \leq 15.1$	APFEL
	F_2^d	[124]	254 (248/248)	$0.07 \leq x \leq 0.75$	$3.0 \leq Q \leq 15.1$	
CHORUS	$\sigma^{\mathrm{CC},\nu}$	[125]	607 (416/416)	$0.045 \leq x \leq 0.65$	$1.9 \leq Q \leq 9.8$	APFEL
	$\sigma^{\mathrm{CC},\bar{\nu}}$	[125]	607 (416/416)	$0.045 \leq x \leq 0.65$	$1.9 \leq Q \leq 9.8$	
NuTeV	σ_ν^{cc}	[126, 127]	45 (39/39)	$0.02 \leq x \leq 0.33$	$2.0 \leq Q \leq 10.8$	APFEL
	$\sigma_{\bar{\nu}}^{cc}$	[126, 127]	45 (37/37)	$0.02 \leq x \leq 0.21$	$1.9 \leq Q \leq 8.3$	
HERA	$\sigma^p_{\mathrm{NC,CC}}$ **(*)**	[74]	1306 (1145/1145)	$4 \cdot 10^{-5} \leq x \leq 0.65$	$1.87 \leq Q \leq 223$	APFEL
	σ^c_{NC}	[128]	52 (47/37)	$7 \cdot 10^{-5} \leq x \leq 0.05$	$2.2 \leq Q \leq 45$	
	F_2^b **(*)**	[129, 130]	29 (29/29)	$2 \cdot 10^{-4} \leq x \leq 0.5$	$2.2 \leq Q \leq 45$	

Table 5.1: Deep-inelastic scattering data included in NNPDF3.1. New datasets, not included in NNPDF3.0, are denoted **(*)**. The kinematic range covered in each variable is given after cuts are applied. The total number of DIS data points after cuts is 3102/3092 for the NLO/NNLO PDF determinations.

Exp.	Obs.	Ref.	$N_{\mathbf{dat}}$	Kin_1	Kin_2 (GeV)	Theory
E866	$\sigma^d_{\mathrm{DY}}/\sigma^p_{\mathrm{DY}}$	[131]	15 (15/15)	$0.07 \leq y_{ll} \leq 1.53$	$4.6 \leq M_{ll} \leq 12.9$	APFEL+Vrap
	σ^p_{DY}	[132, 133]	184 (89/89)	$0 \leq y_{ll} \leq 1.36$	$4.5 \leq M_{ll} \leq 8.5$	APFEL+Vrap
E605	σ^p_{DY}	[134]	119 (85/85)	$-0.2 \leq y_{ll} \leq 0.4$	$7.1 \leq M_{ll} \leq 10.9$	APFEL+Vrap
CDF	$d\sigma_Z/dy_Z$	[135]	29 (29/29)	$0 \leq y_{ll} \leq 2.9$	$66 \leq M_{ll} \leq 116$	Sherpa+Vrap
	k_t incl jets	[136]	76 (76/76)	$0 \leq y_{\mathrm{jet}} \leq 1.9$	$58 \leq p_T^{\mathrm{jet}} \leq 613$	NLOjet++
D0	$d\sigma_Z/dy_Z$	[137]	28 (28/28)	$0 \leq y_{ll} \leq 2.8$	$66 \leq M_{ll} \leq 116$	Sherpa+Vrap
	W electron asy **(*)**	[138]	13 (13/8)	$0 \leq y_e \leq 2.9$	$Q = M_W$	MCFM+FEWZ
	W muon asy **(*)**	[139]	10 (10/9)	$0 \leq y_\mu \leq 1.9$	$Q = M_W$	MCFM+FEWZ

Table 5.2: Same as Table 5.1 for the Tevatron fixed-target Drell-Yan and W, Z and jet collider data. The total number of Tevatron data points after cuts is 345/339 for NLO/NNLO fits.

- We have included the legacy W lepton asymmetries from D0 using the complete Tevatron luminosity, both in the electron [138] and in the muon [139] channels. These datasets provide important information on quark flavour separation [169], which is currently the less well understood feature of PDFs in the data region.

We have also included some new dataset from the LHC experiments

- From the ATLAS experiment we include:
  - The Z boson (p_T^Z, y_Z) and (p_T^Z, M_{ll}) double differential distributions measured at 8 TeV [145].
  - The inclusive W^+, W^- and Z rapidity distributions at 7 TeV from the 2011 dataset [141],
  - The top-quark pair production normalized y_t distribution at 8 TeV [151];

Exp.	Obs.	Ref.	$N_{\rm dat}$	Kin$_1$	Kin$_2$ (GeV)	Theory
ATLAS	W,Z 2010	[140]	30 (30/30)	$0 \leq \lvert\eta_l\rvert \leq 3.2$	$Q = M_W, M_Z$	MCFM+FEWZ
	W,Z 2011 (*)	[141]	34 (34/34)	$0 \leq \lvert\eta_l\rvert \leq 2.3$	$Q = M_W, M_Z$	MCFM+FEWZ
	high-mass DY 2011	[142]	11 (5/5)	$0 \leq \lvert\eta_l\rvert \leq 2.1$	$116 \leq M_{ll} \leq 1500$	MCFM+FEWZ
	low-mass DY 2011 (*)	[143]	6 (4/6)	$0 \leq \lvert\eta_l\rvert \leq 2.1$	$14 \leq M_{ll} \leq 56$	MCFM+FEWZ
	$[Z\ p_T$ 7 TeV $\left(p_T^Z, y_Z\right)]$ (*)	[144]	64 (39/39)	$0 \leq \lvert y_Z\rvert \leq 2.5$	$30 \leq p_T^Z \leq 300$	MCFM+NNLO
	$Z\ p_T$ 8 TeV $\left(p_T^Z, M_{ll}\right)$ (*)	[145]	64 (44/44)	$12 \leq M_{ll} \leq 150$ GeV	$30 \leq p_T^Z \leq 900$	MCFM+NNLO
	$Z\ p_T$ 8 TeV $\left(p_T^Z, y_Z\right)$ (*)	[145]	120 (48/48)	$0.0 \leq \lvert y_Z\rvert \leq 2.4$	$30 \leq p_T^Z \leq 150$	MCFM+NNLO
	7 TeV jets 2010	[146]	90 (90/90)	$0 \leq \lvert y^{\rm jet}\rvert \leq 4.4$	$25 \leq p_T^{\rm jet} \leq 1350$	NLOjet++
	2.76 TeV jets	[147]	59 (59/59)	$0 \leq \lvert y^{\rm jet}\rvert \leq 4.4$	$20 \leq p_T^{\rm jet} \leq 200$	NLOjet++
	7 TeV jets 2011 (*)	[148]	140 (31/31)	$0 \leq \lvert y^{\rm jet}\rvert \leq 0.5$	$108 \leq p_T^{\rm jet} \leq 1760$	NLOjet++
	$\sigma_{\rm tot}(t\bar{t})$	[149, 150]	3 (3/3)	-	$Q = m_t$	top++
	$(1/\sigma_{t\bar{t}})d\sigma(t\bar{t})/y_t$ (*)	[151]	10 (10/10)	$0 < \lvert y_t\rvert < 2.5$	$Q = m_t$	Sherpa+NNLO
CMS	W electron asy	[152]	11 (11/11)	$0 \leq \lvert\eta_e\rvert \leq 2.4$	$Q = M_W$	MCFM+FEWZ
	W muon asy	[153]	11 (11/11)	$0 \leq \lvert\eta_\mu\rvert \leq 2.4$	$Q = M_W$	MCFM+FEWZ
	$W + c$ total	[154]	5 (5/0)	$0 \leq \lvert\eta_l\rvert \leq 2.1$	$Q = M_W$	MCFM
	$W + c$ ratio	[154]	5 (5/0)	$0 \leq \lvert\eta_l\rvert \leq 2.1$	$Q = M_W$	MCFM
	2D DY 2011 7 TeV	[155]	124 (88/110)	$0 \leq \lvert\eta_{ll}\rvert \leq 2.2$	$20 \leq M_{ll} \leq 200$	MCFM+FEWZ
	[2D DY 2012 8 TeV]	[156]	124 (108/108)	$0 \leq \lvert\eta_{ll}\rvert \leq 2.4$	$20 \leq M_{ll} \leq 1200$	MCFM+FEWZ
	$W^\pm$ rap 8 TeV (*)	[157]	22 (22/22)	$0 \leq \lvert\eta_l\rvert \leq 2.3$	$Q = M_W$	MCFM+FEWZ
	$Z\ p_T$ 8 TeV (*)	[158]	50 (28/28)	$0.0 \leq \lvert y_Z\rvert \leq 1.6$	$30 \leq p_T^Z \leq 170$	MCFM+NNLO
	7 TeV jets 2011	[159]	133 (133/133)	$0 \leq \lvert y^{\rm jet}\rvert \leq 2.5$	$114 \leq p_T^{\rm jet} \leq 2116$	NLOjet++
	2.76 TeV jets (*)	[160]	81 (81/81)	$0 \leq \lvert y_{\rm jet}\rvert \leq 2.8$	$80 \leq p_T^{\rm jet} \leq 570$	NLOjet++
	$\sigma_{\rm tot}(t\bar{t})$	[161, 162]	3 (3/3)	-	$Q = m_t$	top++
	$(1/\sigma_{t\bar{t}})d\sigma(t\bar{t})/y_{t\bar{t}}$ (*)	[163]	10 (10/10)	$-2.1 < y_{t\bar{t}} < 2.1$	$Q = m_t$	Sherpa+NNLO
LHCb	Z rapidity 940 pb	[164]	9 (9/9)	$2.0 \leq \eta_l \leq 4.5$	$Q = M_Z$	MCFM+FEWZ
	$Z \to ee$ rapidity 2 fb	[165]	17 (17/17)	$2.0 \leq \eta_l \leq 4.5$	$Q = M_Z$	MCFM+FEWZ
	$W,Z \to \mu$ 7 TeV (*)	[166]	33 (33/29)	$2.0 \leq \eta_l \leq 4.5$	$Q = M_W, M_Z$	MCFM+FEWZ
	$W,Z \to \mu$ 8 TeV (*)	[167]	34 (34/30)	$2.0 \leq \eta_l \leq 4.5$	$Q = M_W, M_Z$	MCFM+FEWZ

Table 5.3: Same as Table 5.1, for ATLAS, CMS and LHCb data from the LHC Run I at $\sqrt{s} = 2.76$ TeV, $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV. The ATLAS 7 TeV Z p_T and CMS 2D DY 2012 are in brackets because they are only included in a dedicated study but not in the default PDF set. The total number of LHC data points after cuts is 848/854 for NLO/NNLO fits (not including ATLAS 7 TeV Z p_T and CMS 2D DY 2012).

  - The total cross-sections for top quark pair production at 7, 8 and 13 TeV [149, 150]. The top cross section is the only data we include at 13 TeV; we reserve the total W and Z cross sections [170] for benchmarks (some of which are shown in Sec. 5.3.4).
  - The inclusive jet cross-sections at 7 TeV from the 2011 dataset [148];
  - The Low mass Drell-Yan M_{ll} distributions at 7 TeV from the 2010 run [143].

- From CMS we have added:
  - The W^+ and W^- rapidity distributions at 8 TeV [157]
  - The inclusive jet production cross-sections at 2.76 TeV [160];
  - Top-quark pair production normalized $y_{t\bar{t}}$ distributions at 8 TeV [163]
  - Total inclusive $t\bar{t}$ cross-sections at 7, 8 and 13 TeV [161];
  - The distribution of the Z boson double differentially in (p_T, y_Z) at 8 TeV [158].

- Finally, we include the complete LHCb 7 and 8 TeV measurements of the W and Z inclusive production in the muon channel [166, 167].

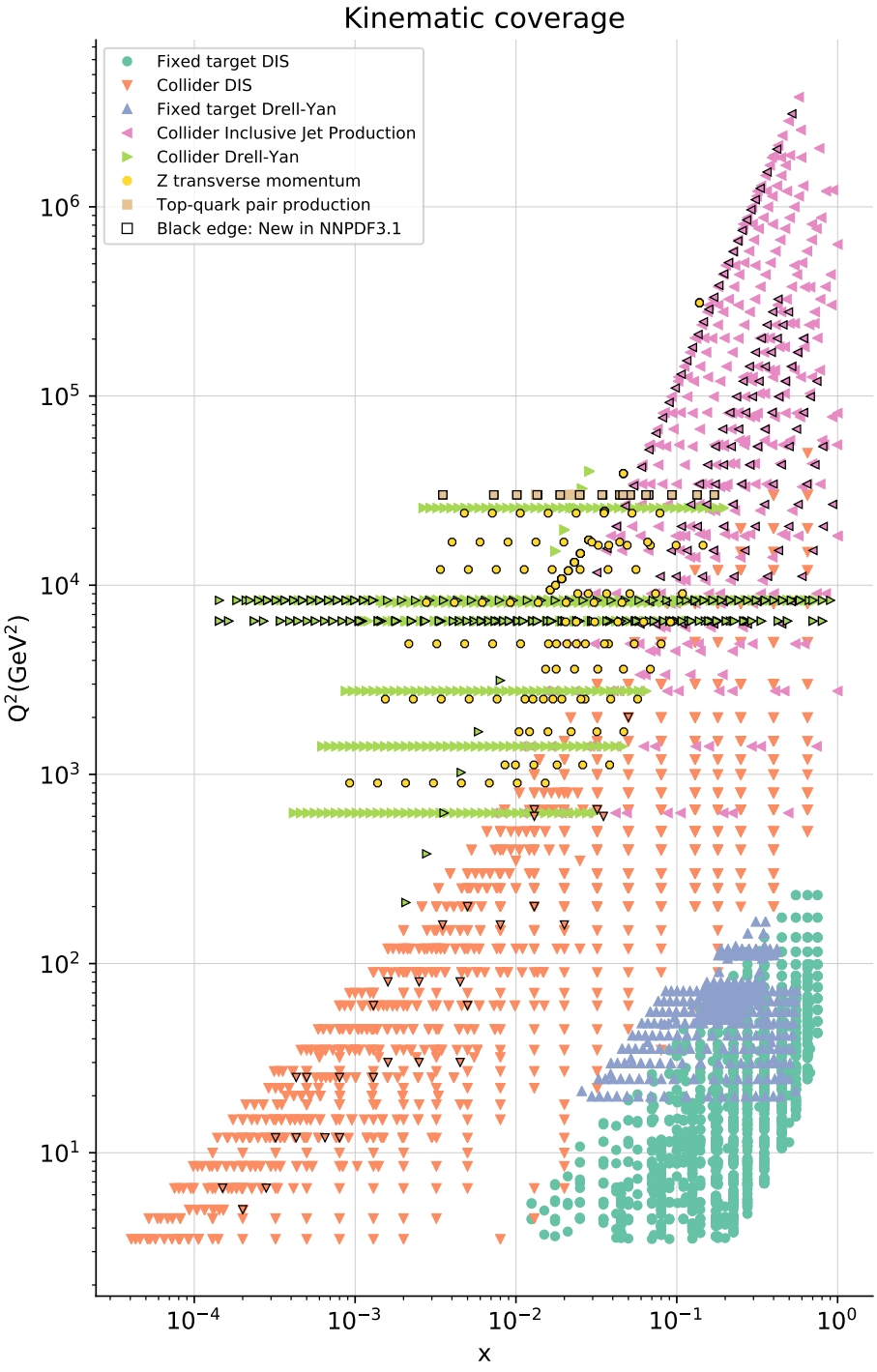


Figure 5.1: Kinematic reach of the NNPDF 3.1 data

5.3 Main characteristics of NNPDF 3.1

We present here a small selection of the results shown in Ref. [12] and the accompanying online gallery [171]; for example, while here we only plot some selected PDF flavours, all of the fitted PDFs for the corresponding comparison can be found in the gallery.

5.3.1 Impact of new data

The change that adding each individual dataset listed in Sec. 5.2.2 causes in the PDFs is relatively small and comparable with statistical fluctuations. To illustrate this, we compare in Fig 5.1 the baseline gluon PDF from the default NNPDF3.1 global fit to fits where the Zp_T and the top data have been excluded (note that both of these processes couple to the gluon PDF at leading order). The uncertainties are similarly increased only slightly, by a few per-mill of the value of the PDF. However the addition of new data taken as a whole results in a change in the central value of around 1σ and a reduction in PDF uncertainties by up to 30%. We show this in Fig. 5.2. This result is satisfactory in that it shows that the new data contributes to constrain the PDFs, but in a way that is consistent with the quoted PDF uncertainties for a given experimental input, thereby supporting the overall consistency of the methodology.

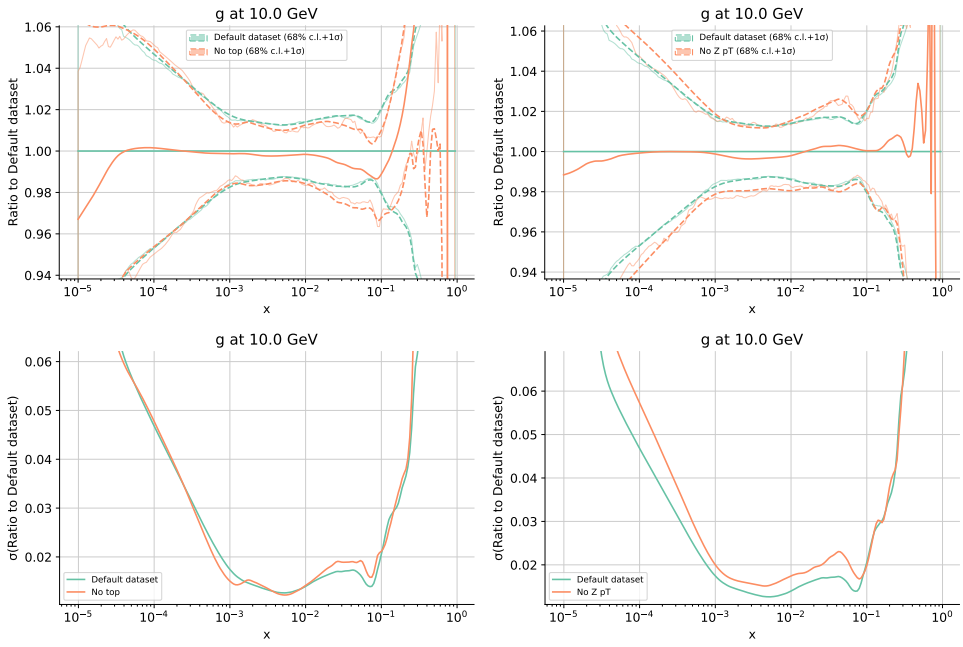


Figure 5.1: Comparison of the gluon PDF (up) and the gluon PDF uncertainty (down) of the default NNPDF 3.1 global fit at NNLO with a fit that does not contain any top data (left) and $Z\,p_T$ data (right).

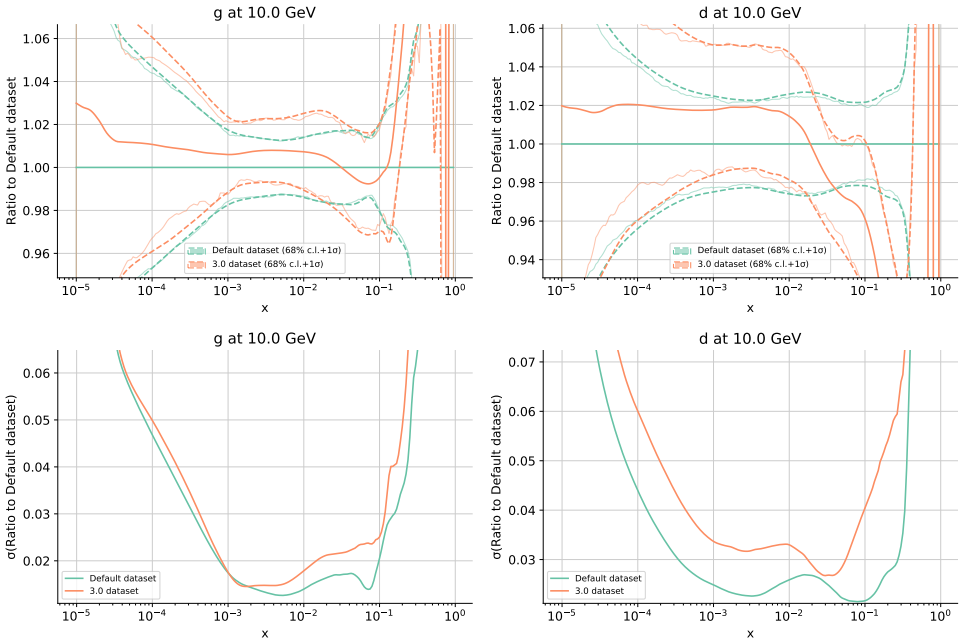


Figure 5.2: Comparison of the default NNPDF3.1 default fit at NNLO with a fit containing a similar dataset to NNPDF3.0, for the gluon (left) and d (right) .

5.3.2 Impact of fitted charm

Fitting the charm PDF independently, in an equivalent way to the strange and light quark distributions proves to be advantageous when determining PDFs, particularly at NNLO accuracy. For example the fit quality of the combined HERA data deteriorates notably at NNLO compared to the fits at NLO, when the charm is generated only perturbatively. Instead the difference in fit quality is significatively smaller when the charm is fitted. The underlying reason is that the constraints on the total quark content from HERA only become compatible with the constrains on strangeness from the ATLAS W, Z 2011 rapidity distributions with the extra degrees of freedom provided by the parametrized charm PDF, as shown in Table 5.1.

	ATLAS W, Z 7 TeV 2011	Hera Combined
NLO fitted charm	3.70	1.14
NLO perturbative charm	4.29	1.15
NNLO fitted charm	2.15	1.16
NNLO perturbative charm	2.75	1.21

Table 5.1: χ^2 per degree of freedom for the ATLAS W,Z rapidity distributions for fits at NLO, NNLO with fitted or perturbative charm

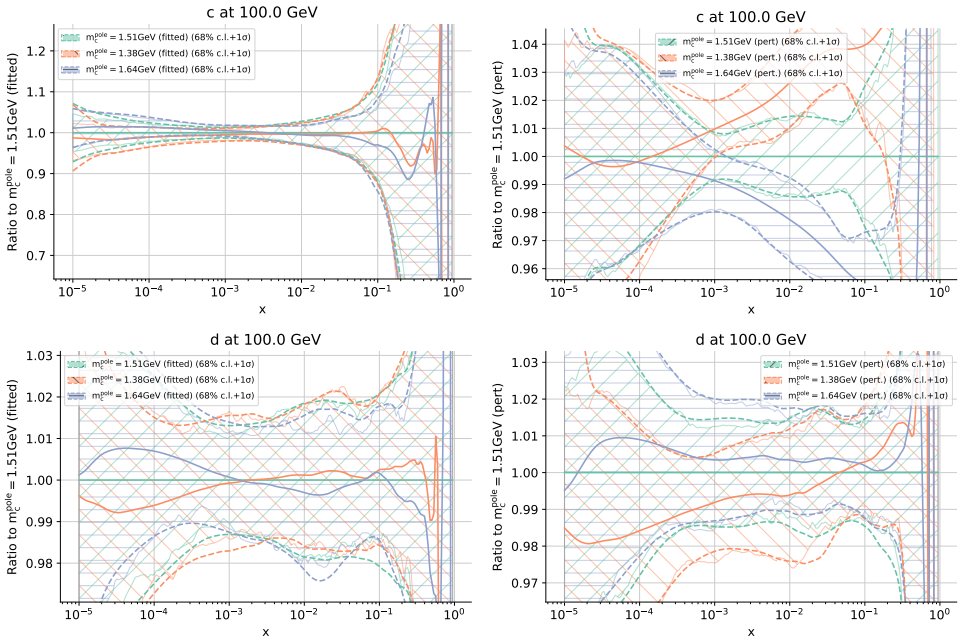


Figure 5.3: Charm (up) and d PDFs fitted with several different values of the charm mass, with a fitted (left) and a purely perturbatively generated (right) charm PDF, at $Q = 100$ GeV

A PDF fit where the charm PDF is purely perturbative is affected by significative uncertainties related to the value of the charm mass. As discussed in Ref. [11], this is mainly due to the position of the perturbative threshold for charm production in the DGLAP evolution (see Sec. 2.7 and references therein): A lower threshold (associated with a lower charm pole mass) can be related to an increase in the charm momentum fraction (which is completely correlated to that of the gluon since no indented parametrization of the charm exists when it is generated perturbatively) at any scale above the threshold, since the range of scales in which charm is produced increases. Therefore the differences is propagated at higher scales and affects high energy cross sections (see the detailed discussion in Ref. [11]).

As we show in Fig. 5.3, the charm mass dependence is much reduced when the charm PDF is fitted: The charm PDF is fundamentally determined by the data (rather than the DGLAP evolution) since extra degree of freedom exist to compensate threshold effects, which are arranged by the PDF fit in such a way that the agreement with the data is optimized; and a change in the charm production threshold only effects a small fluctuation in the charm PDF at the initial scale to accommodate it. Notably, the light quark PDFs are also more stable upon a variation in the charm mass.

5.3.3 Improved uncertainties compared to NNPDF 3.0

To demonstrate the improvement in uncertainties with respect to NNPDF3.0, we plot in Fig. 5.5 the PDF uncertainty on the differential luminosity (see Sec 2.8),

$$\tilde{L}(M_X, y, s) = \sum_{ij}^{\text{channel}} \frac{1}{s} f_i\left(\frac{M_x e^y}{\sqrt{x}}, M_x\right) f_j\left(\frac{M_x e^{-y}}{\sqrt{x}}, M_x\right) \tag{5.3.1}$$

for NNPDF 3.0 and NNPDF 3.1. The figure shows a considerable reduction in uncertainty, which is now below the percent level in parts of the phases space relevant for LHC phenomenology. In particular, the gluon PDF shown in Fig 5.4 is improved due to the combination of many mutually consistent constraints on the gluon from DIS (especially at HERA), Z transverse momentum distributions, jet production, and top pair production, which taken together cover a very wide kinematic range. The uncertainty reduction on the gluon then propagates to the singlet component of the quark distributions via DGLAP evolution, partially explaining the overall improvement in Fig. 5.5.

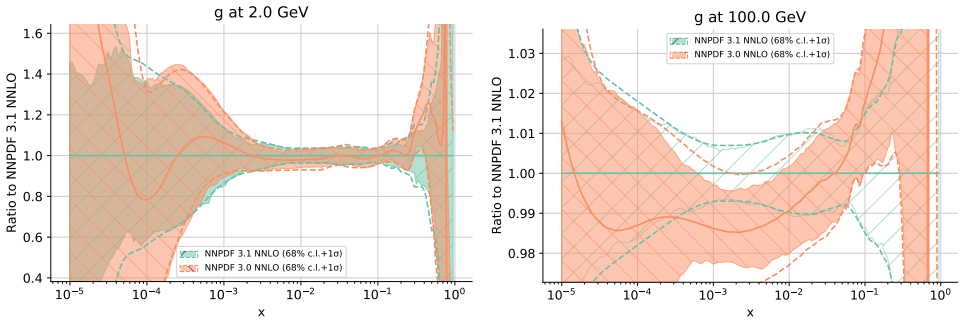


Figure 5.4: Gluon PDF of NNPDF 3.1 and NNPDF3.1, at $Q = 2$GeV (left) and 100 GeV (right)

5.3.4 LHC cross sections

Theoretical predictions based on the NNPDF3.1 to W and Z production data at $\sqrt{s} = 13$ TeV from ATLAS [170] are compared to the data and the results from other PDFs in Fig. 5.6. As mentioned in Sec. 5.2.2, none of these cross sections were included in the fits (neither in NNPDF 3.1 nor in any PDFs used for comparison), and therefore constitute genuine predictions of the theory.

We compute fiducial cross-sections using `FEWZ` [172] at NNLO QCD accuracy, using NNPDF3.1, NNPDF3.0, CT14, MMHT14 and ABMP16 PDFs, together with the corresponding PDF uncertainty band. All calculations are performed with $\alpha_s = 0.118$ (including ABMP16 which uses a different value by default). Electroweak NLO corrections are computed with `FEWZ` for Z production, and with `HORACE3.2` [173] for W production. The fiducial phase space is matched to the ATLAS measurement in Ref. [170].

In Fig 5.6 we display the ratios of total cross sections W^+/W^- and W/Z. Note that in the figures we only show the PDF uncertainties of the theoretical predictions, but not other effects such as α_s or m_c uncertainties or scale variations. Taking this into account, all the results agree reasonably well with the data. The electroweak corrections are sizeable compared to the PDF uncertainty and contribute to improve the agreement with the ATLAS measurements.

The corresponding absolute W^+, W^- and Z cross-sections are shown in Fig. 5.7, normalized in each case to the experimental central value. Theory predictions are generally in agreement with the data, with the exception of ABMP16 for Z production.

These results show that both the precision and accuracy of NNPDF3.1 in describing these high precision observables have significantly improved with respect to NNPDF3.0. This improved agreement is particularly marked for Z production.

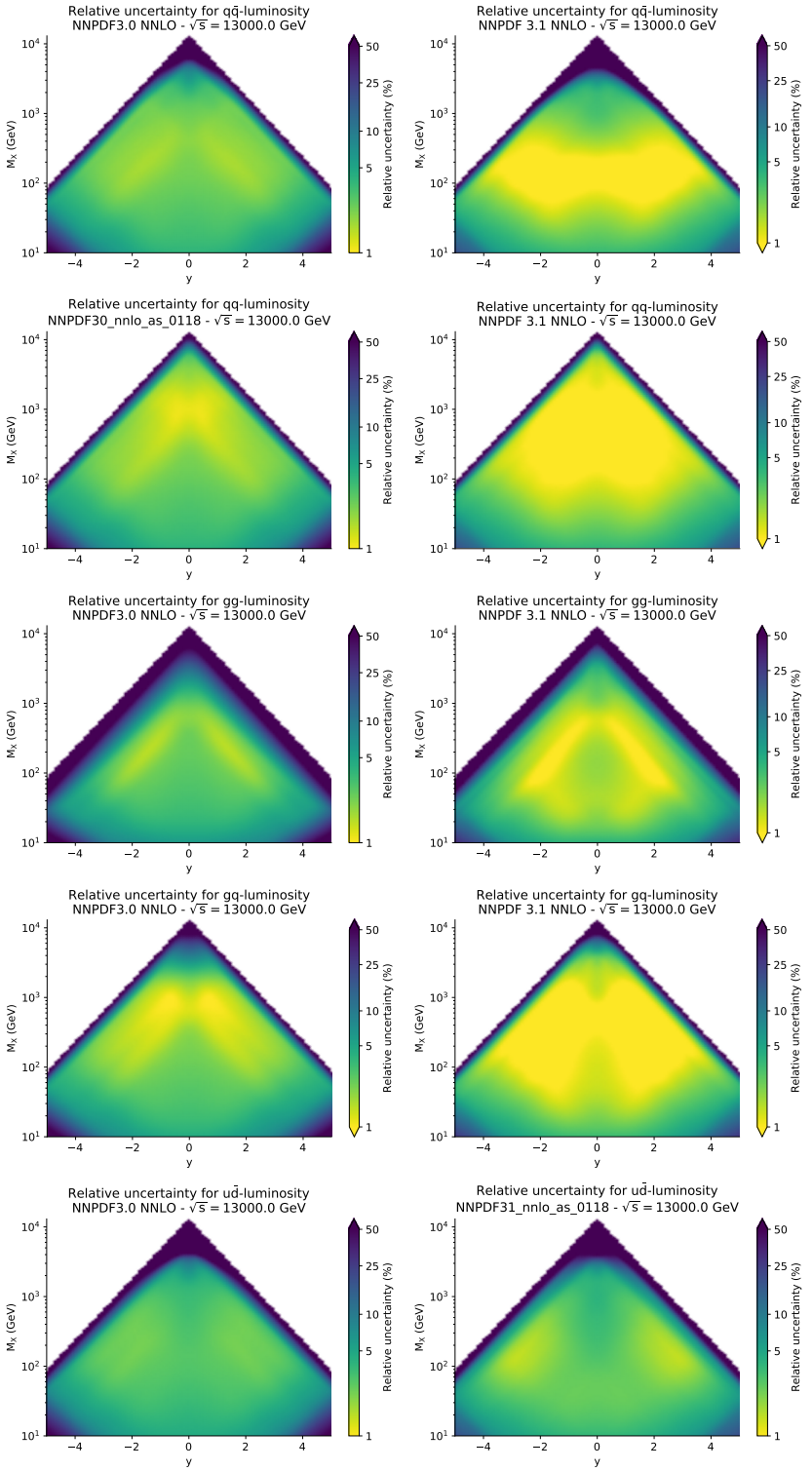


Figure 5.5: The relative uncertainty on the luminosities of plotted as a function of the invariant mass M_X and the rapidity y of the final state; the left plots show results for NNPDF3.0 and the right plots for NNPDF3.1 (upper four rows).

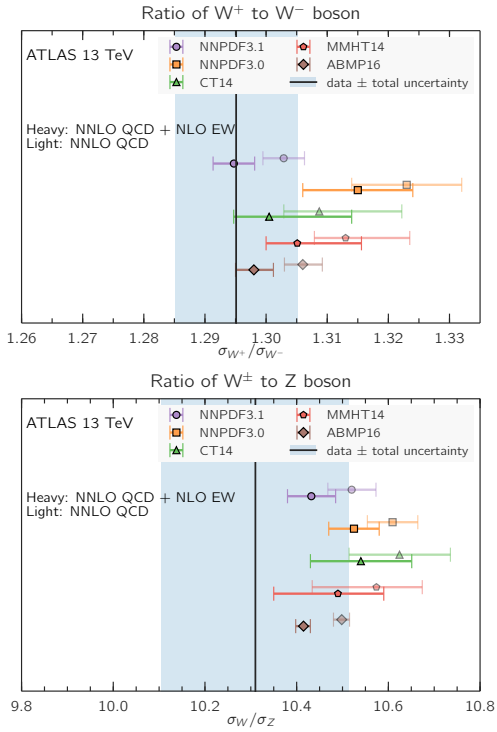


Figure 5.6: Comparison of the ATLAS measurements of the W^+/W^- ratio (left) and the W/Z ratio (right) at $\sqrt{s} = 13$ TeV with theoretical predictions computed with different NNLO PDF sets. Predictions are shown with (heavy) and without (light) NLO EW corrections computed with `FEWZ` and `HORACE`, as described in the text. The figure is taken from Ref. [12].

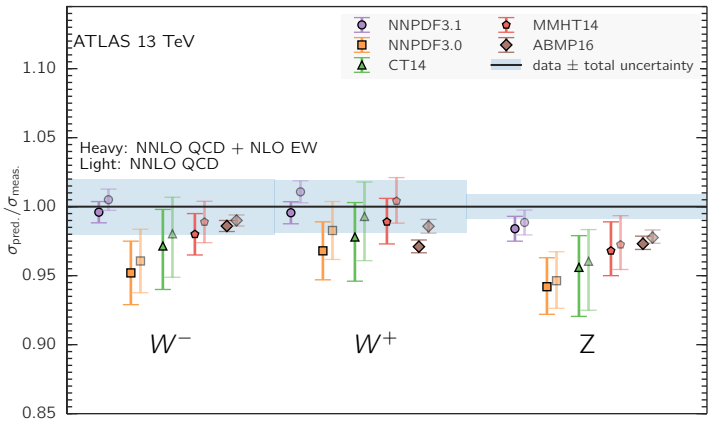


Figure 5.7: Same as Fig. 5.6, now for the absolute W^+, W^- and Z cross-sections. All predictions are normalized to the experimental central value. The figure is taken from Ref. [12].

5.4 Issues with high precision data

As repeatedly mentioned throughout this work, the inclusion of high precision data in the PDF fits requires a constant reassessment of the underlying assumptions and methodological choices. The same holds for the input to the PDF determinations: the experimental data and the corresponding parton level theory predictions. When the nominal uncertainties decrease effects that were previously overlooked become important and comparable to the main uncertainties. We illustrate this with two examples where the complications arising from such overlooked effects required a substantial effort in order to be clarified.

The first one involves an issue with the theoretical computation of the Z transverse momentum distribution: It was found that the fluctuations in the provided results were substantially higher than the quoted Monte Carlo uncertainties. The second one shows that the correlation model for the CMS double differential Drell-Yan distribution from 2012. It was found the provided correlation model leads to results that are largely incompatible with the rest of the fitted data.

5.4.1 Monte Carlo uncertainties in the Zp_T distributions

The transverse momentum distribution of the Z boson was recently computed at NNLO [92, 93, 94, 95]. This accomplishment allowed to include the predictions corresponding predictions from ATLAS [144, 145] and CMS [158], for the first time. In NNPDF3.1 the theoretical predictions have been obtained from Ref. [174], based on the computation of Refs. [94, 95]. Factorization and renormalization scales are chosen as

$$\mu_R = \mu_F = \sqrt{p_T^2 + M_{\ell\ell}^2}\,, \tag{5.4.1}$$

where $M_{\ell\ell}$ is the invariant mass of the final-state lepton pair.

The Zp_T datasets are measured very precisely, with experimental uncertainties few-per-mille for the most precise bins. This in turn requires to place the theory predictions under further scrutiny. In particular it was found that the C-factors, defined as the ratios of the NNLO to the NLO predictions for each bins presented fluctuations that

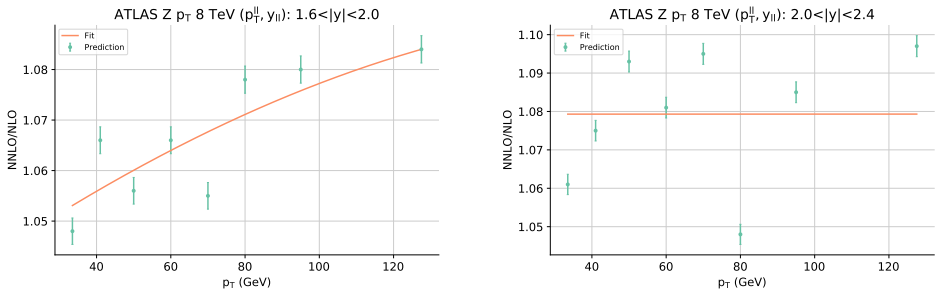


Figure 5.1: C-factors of two bins of the ATLAS Z transverse momentum data. The error bars show the Monte Carlo integration errors as obtained from the theory predictions based on Refs. [94, 95]. The solid line shows the prediction from the Gaussian Process based model described here.

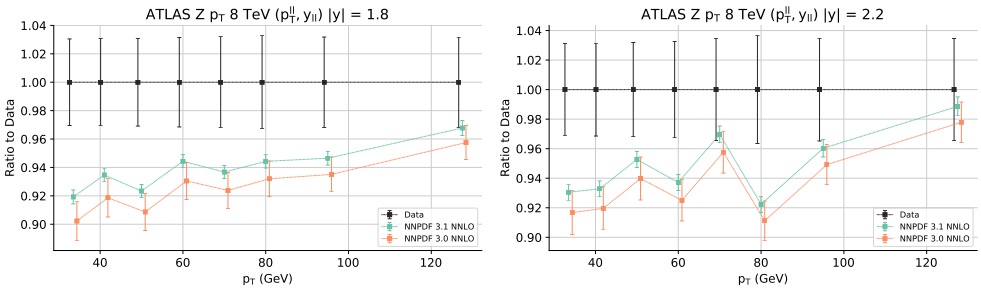


Figure 5.2: Theory predictions for the ATLAS Zp_T 8 TeV data for two rapidity bins, compared and normalized to the experimental data. The comparison with Fig. 5.1 constitutes evidence that the observed fluctuations in the theoretical predictions are predominately due to the fluctuations in the C-factors.

were big compared to the quoted Monte Carlo uncertainties, and were unlikely to be explained by them. The C-factors for two rapidity bins are displayed in Fig 5.1; since it is expected that the C-factors are smooth as a function of p_T, the fluctuations are unlikely to be a genuine theoretical effect. To further support this conclusion, we show in Fig. 5.2 the theoretical predictions for one PDF that does not include the data (NNPDF3.0) and one that does (NNPDF3.1). The fluctuations observed in Fig. 5.2 are clearly correlated with those in Fig 5.1. Furthermore, note that the scale of the fluctuations is bigger than all the other magnitudes in Fig. 5.2, proving that they constitute an impediment to obtain predictions with theoretical uncertainties smaller than the experimental ones.

We can try to further study the size of the fluctuations by fitting the C-factors to a suitable statistical model. While we do not know the amount of smoothness (i.e. the correlation length of the C-factors taken as a function of p_T), we can *learn* it using a similar methodology as in the NNPDF fits. Specifically, we choose a Gaussian Process [175] as implemented by the `scikit-learn` library [176].

We construct a Gaussian Process kernel taking two parameters: the scale of the correlation length in units of p_T, and a noise level η. We now assume a translation-invariant two point covariance function defined by:

$$k(p_{Ti}, p_{Tj}) = \exp\left(-\frac{1}{2}(|p_{Ti} - p_{Tj}|/l)^2\right) + \delta(p_{Ti} - p_{Tj})\eta \tag{5.4.2}$$

The prior of our Gaussian Process model is then a distribution of functions $C(p_T)$ that fulfills:

$$\mathbb{E}(C(p_T)) = m(p_T) = 1 \tag{5.4.3}$$

$$\mathbb{E}\left((C(p_{Ti}) - m(p_{Ti})(C(p_{Tj}) - m(p_{Tj}))\right) = k(p_{Ti}, p_{Tj})\ , \tag{5.4.4}$$

where $\mathbb{E}$ is the expected value over ensembles of functions. In this model the higher moments of the probability density vanish. The result of the procedure is a posterior distribution that is constructed by considering the input C-factors. Assuming an uncorrelated a Gaussian likelihood for the input data, with the mean centered at each data point and the scale set to η, it is possible to arrive at a closed form solution that is also a Gaussian Process [175].

y	(0.4, 0.8)	(0.8, 1.2)	(1.2, 1.6)	(1.6,2.0)	(2.0,2.4)
std	0.0033	0.0038	0.0032	0.0068	0.0162
68%	0.0038	0.0038	0.0035	0.0065	0.0172
90%	0.0051	0.0053	0.0050	0.010	0.0222

Table 5.2: Fluctuations of the differences between the original C-factors and the model predictions for the ATLAS Zp_T 8 TeV data. The columns indicate bind in rapidity and the rows are the standard deviation, the 68% inter quantile range and the 98% inter quantile range respectively, computed over the distribution of absolute value of differences in each bin. The results validate the choice of adding 1% extra uncertainty to the data.

Note that we do not use the Monte Carlo uncertainties in the whole procedure since we wish precisely to estimate them. For this application we can for example compute the differences between the mean value of each of the provided C-factors and the mean of the posterior distribution, and estimate the dispersion for each bin in rapidity.

We fix the hyper-parameters η and l employing a cross validation procedure. Specifically we maximize the mean of the validation score (using a quadratic error function) of a leave-one-out cross validation fit over all the possible combinations. We further require that the training and validation score are not too far apart (specifically that they are within a factor 2) so that an accidentally finding an overfitted minimum for the only validation point does not bias the fit. With this procedure, we finally obtain the model predictions in Fig 5.2. Note that in the second bin, the model is unable to find evidence for $l < \text{inf}$ since it does no combination of parameters with correlation length of the size of the range in p_T improves the error improve the error score. We can then proceed to estimate the size of the fluctuations .

In view of the results, we conclude that there is likely an underestimated source of uncertainty in the preparation of the theory prediction, that is phenomenologically relevant when compared with the magnitude of the experimental errors.

A similar approach based on neural network fits [177] was presented in Ref. [12]. The conclusions were consistent with the ones presented here.

After the decision was to include the data adding an additional 1% uncorrelated uncertainty to all the points in the distribution, which is consistent with the difference . Unfortunately this implies that we cannot take full advantage of the experimental precision of the measurement currently. These findings were further corroborated in Ref. [174].

We note that the approach presented here could be used to improve the accuracy of theory predictions where the Monte Carlo errors are consistent with the size of the fluctuations but it is desired to make them smaller. One could *learn* a correlation model for the predictions based on similar assumption on smoothness and replace the noisy outputs of the computations with smoothed model predictions.

5.4.2 The CMS 8 TeV double-differential Drell-Yan distributions

The NNPDF3.0 PDF determination already included double-differential (in rapidity and invariant mass) Drell-Yan data at 7 TeV from the CMS 2011 dataset [155]. An updated version of the same measurement at 8 TeV based on 2012 data was presented in Ref. [156], including both the absolute cross-sections and the ratio of 8 TeV and 7 TeV measurements. The data is characterized by extremely small uncorrelated uncertainties, and therefore the total uncertainty is dominated by correlated systematics. After extensive checks, it was determined that the data is incompatible with the rest of the constraints in the fit. While it is not possible to determine it with certainty since no breakdown of systematics was provided, we obtained evidence that the incompatibility may be related with the choice of correlation model used to analyze the experimental uncertainties.

Including the 2012 data in the fit causes a global increase in the total error function by $\Delta\chi^2 = 11.5$. The most marked deterioration occurs in the inclusive HERA data. Yet including the data in the fit does not suffice to describe it properly: We obtain a χ^2 per degree of freedom of 2.88. Furthermore, the inclusion of the CMS double-differential Drell-Yan data causes a noticeable change in the gluon PDF, which is unexpected considering that the Drell-Yan process only provides an indirect handle for the gluon PDF. We show some comparisons in Fig. 5.3.

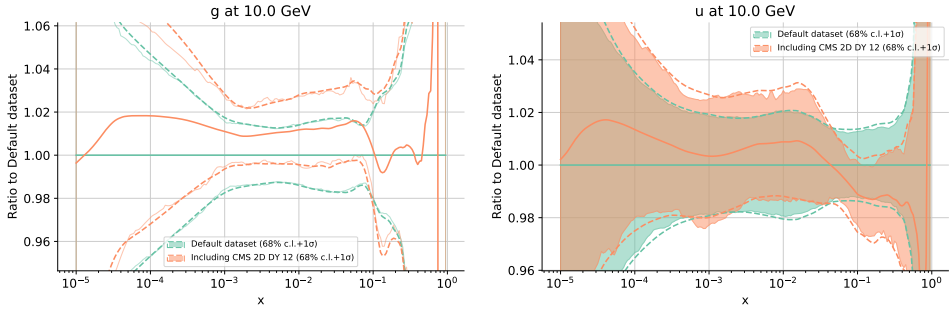


Figure 5.3: Ratios for the gluon (left) and up (right) PDFs, where the impact of including the CMS double differential Drell-Yan data at 8 TeV is more marked.

Even though the exercise is necessarily limited in scope considering the lack of information of the systematics, it is instructive to try to understand the origin of the problem, in particular to discard any possible problems with the global fit. We begin by illustrating the effect of the correlation model, by comparison with the 2011 dataset. In Fig. 5.4 we compare the predictions for two equivalent bins in invariant mass of the 2011 and 2012 datasets, using a variation of 3.1 that includes the updated CMS data. Even though the diagonal uncertainties are in better agreement for the new data, the total χ^2 is markedly worse, showing how these comparison plots can be misleading when the effect of the correlated systematics dominates. We next assess the kinematical region that each of these two bins probe. To this end, we plot in Fig. 5.5 the correlation function used for the SMPDF, Eq. 3.4.3. These plots answer the question "In which region would a reduction of the PDF uncertainty result on a biggest reduction in uncertainty in the corresponding prediction?" and they can be

used as a proxy to estimate the parts of the PDF (in flavour and x space) that are most sensitive to a given prediction. Unsurprisingly we find extremely similar patterns, providing evidence for the fact that the two datasets might be incompatible with each other, since we are unable to fit them both at the same time.

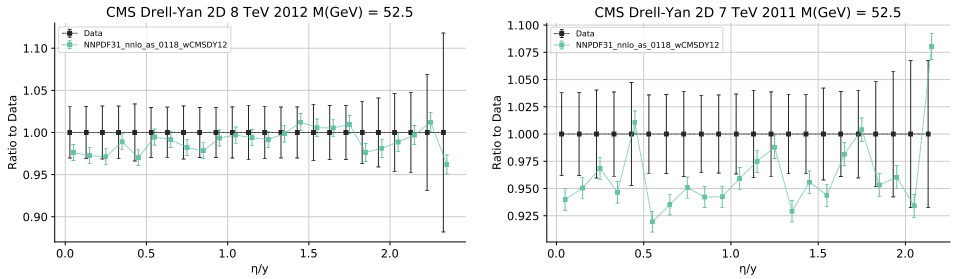


Figure 5.4: Data-theory comparison for two equivalent bins of the CMS double differential Drell-Yan 2012 (left) and (2011) data. The predictions are computed with a fit that includes both datasets. The agreement with the 2012 is much worse despite each point of the predictions being closer to the data in units of the diagonal uncertainty: The χ^2 per degree of freedom is 2.88 for the 2012 data and 1.24 for the 2011 data.

At this point evidence has accumulated to hypothesise that the problem may reside in the stability of the covariance matrix. If uncorrelated uncertainties do not dominate, the diagonal entries may not be big enough to stabilize the results, and we may end up with eigenvectors of very different magnitude. In this situation, a small modification in the covariance matrix, that rotates one big eigenvector into a small one might make a large difference when sampling pseudodata and fitting. In the NNPDF3.1 code, we solve Eq. 5.1.25 using a Cholesky decomposition rather than computing the inverse explicitly. We have checked that the numerical stability of this improved procedure is enough to use it reliably. However the main concern is not the floating point precision, but rather the precision with which the experimental systematics are determined in the first place, which currently is largely uncontrolled. For example it may be the case that assuming that a particular systematic is completely correlated across all datapoints would not be an adequate approximation at this level of precision. In fact, there is evidence that different correlation models for inclusive jet double differential observables have a very large impact on the fit quality [178]. To elucidate further, we compare the cumulative increase in the χ^2 that each eigenvector of the covariance matrix effects when considered. The results in Fig. 5.6 show that only a small number of eigenvectors is responsible for the large increase in χ^2. These eigenvectors may well correspond to incompatible linear combinations of data points that should vanish due to elemental smoothness constraints in the PDFs.

All this evidence, together with private communication with colleagues from the MMHT collaboration led us to judge that it is not advantageous to include the CMS 2012 double differential Drell-Yan data at 8 TeV in the default NNPDF fit.

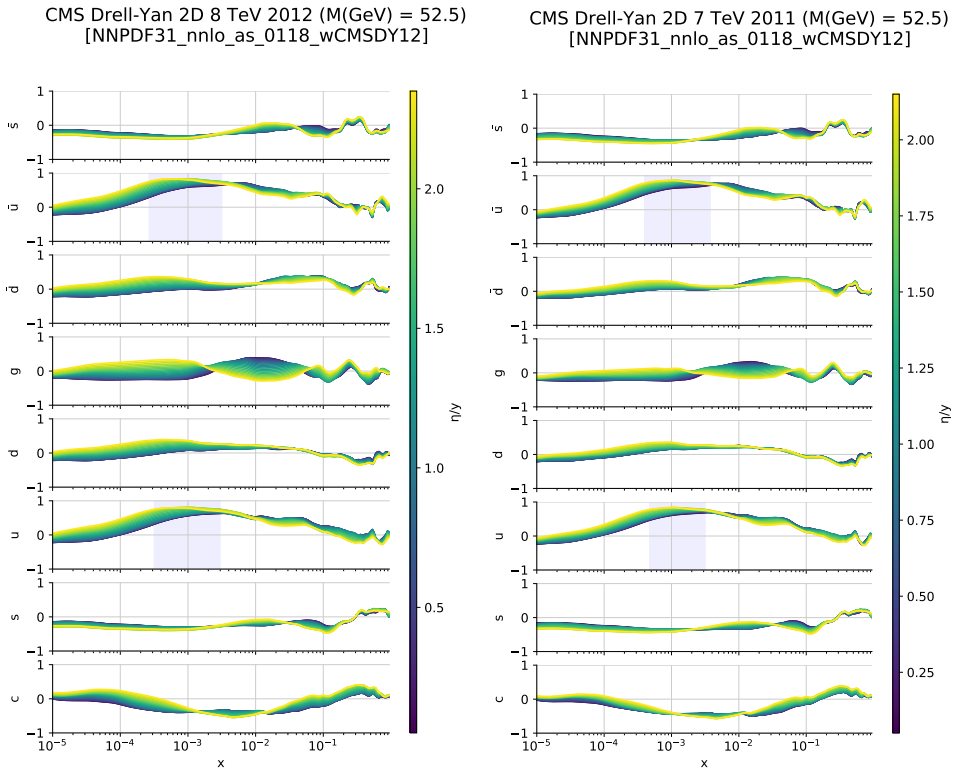


Figure 5.5: Data-PDF corrleations for two equivalent bins of the CMS double differential Drell-Yan 2012 (left) and (2011) data. For each point in rapidity (distinguished by the color code), we plot the result of Eq. 3.4.3 for each flavour. The similarity of the patterns provides evidence of the incompatibility between the two datasets. The test PDF is a variation of NNPDF3.1 fits that include the 2012 data.

5.5 Advanced code tools for NNPDF 3.1

While NNPDF 3.1 did not change much with respect to NNPDF 3.0 [8] in terms of fitting methodology, as we discussed in the previous sections, there was a notable increase in the precision that the whole procedure required, driven mainly by the addition of new high precision LHC data with the corresponding novel theory prediction.

Meeting these precision requirements demanded a careful evaluation of all aspects of the NNPDF methodology described in Sec. 5.1. In fact it was found that multiple aspects of the procedure that were previous overlooked since they constituted completely negligible corrections in the previous generation of PDF sets, were now causing deviations above the PDF uncertainty, or contributing to a unacceptable decrease in the quality of the fit. This in turn required a new generation of analysis and plotting tools, that could process all data used as output and input to the fits to produce the relevant diagnostics.

While the NNPDF code already had a reporting tool, called `validphys` that pro-

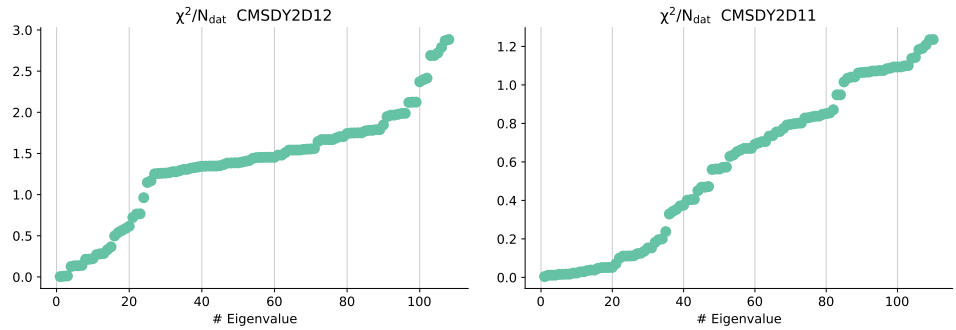


Figure 5.6: Cumulative contribution to the error function per eigenvector for the CMS 2012 double differential Drell-Yan data (left) and the 2011 data (right) , obtained by setting the subsequent eignvectors to zero after the inversion of the covariance matrix. The error function is computed w.r.t. the variation of NNPDF3.1 that does include the 2012 data.

duces a very detailed comparison between two fits (outputting a LaTeX project that is then compiled to PDF), one cannot easily customize its output to perform more general comparisons. On the other hand, tools such as APFEL Web [179] or SM-PDF Web [180] are adequate for casual usages, but not when one wants to assemble large numbers of comparisons in a systematic way. A related problem to solve is the general lack of metadata in the NNPDF internal formats, making it complicated to associate code outputs to interesting quantities (for example to perform a data-theory comparison as a function of the relevant kinematical variables of a given experiment).

It was therefore clear that the best of course of action would be to develop a new set of tools that could be used to analyze the data in a more general and systematic way. A stringent constraint on the development was to not introduce big modifications in the NNPDF code given the originally short timescale of the project, which also imposed the requirement to be able to produce interesting analysis incrementally as the progress in the code development occurs. This was also an opportunity to change the development language from C++ to Python. While the original goal of the project (named `validphys2`) was to design a more general *plotting tool*, eventually its scope extended far beyond that. Currently the project consist on an NNPDF-specific part which is built on top of a open sourced framework to perform scientific computation in the Python programming language, (named `reportengine`). Its main purpose is to contribute to alleviate the problem of reproducibility in software based research while maximizing the usability for practitioners. While it is currently used within the NNPDF collaboration, the code can be employed in other research. Thus the NNPDF specific code `validphys2` is an example of an application that uses the `reportengine` framework.

The basic usage of `validphys2` consists on producing a runcard such as

```
pdfs:
    - NNPDF31_nnlo_as_0118
    - NNPDF30_nnlo_as_0118
```

```
lumi_channel: gg

sqrts: 13000

template_text: |
   # Luminosity uncertainties
   To illustrate the improvement in preciosion compared to
'NNPDF 3.0', we plot the
   uncertainty of the PDF luminosity
   for the gluon-gluon channel ($i=j=g$)

   {@with pdfs@}
   {@plot_lumi2d_uncertainty@}
   {@endwith@}


actions_:
    - - report:
          main: True
```

and obtain essentially Sect. 5.3.3 of this document in HTML format (also with the resolved Markdown source that can be converted to LaTeX easily).

While given a complete description of the `reportengine` and `validphys` codes is outside the scope of this work, and is furthermore likely to become outdated soon, we will describe in some detail which problems they solve and how.

The framework was used to study several aspects of the NNPDF fit, to produce the gallery of plots, and was used extensively in the α_S determination project, presented in Chapter 6.

5.5.1 The `reportengine` framework

Introduction

`reportengine` is a scientific computation framework for the Python programming language. Its aim is to help programmers develop analysis code that is easily accessible to users, reproducible and based on small and side effect free units that are composed automatically to perform a given task.

Usually *reusable* code is associated with an architecture where many clearly separated units (here by *units* we mean for examples functions in the Python language) perform small and specialized tasks and are then composed together to achieve the desired effect. There is however a tension in how small the units can be reasonably made: The simpler they are the more work to put them together is needed, and then the complexity shifts from the core logic to the orchestration. One problem that `reportengine` solves is automating the composition of the small units for many patterns that are typical in scientific computing. This way, the minimum complexity for a given task shifts towards smaller and more reusable logical units that are easy to reason about and compose.

Facing the user, the crucial advantage is that the user's analysis that ultimately determines how the pieces are to be combined is given in terms of a declarative input card. It allows the users of applications based on `reportengine` to write a YAML file instead of a script or a Jupyter Notebook [181] which are instead used extensively for prototyping. As we discuss in Sect. 5.5.1 this has several advantages.

When processing the input of the user, the code acts as a compiler: It processes the runcard to generate a dependency graph, Directed Acyclic Graph (DAG), representing executable code. The nodes correspond to *actions* and the edges to dependency relations between them (i.e. which inputs are requires to execute a given action). Indeed compilers like GCC[182] or LLVM[183] use DAGs as the abstraction to represent the program. After the user input is read, DAG is then *checked* for *"compile time"* errors, and if none is preset, each node (action) is executed in the topological order of the graph (that is, in such an order that each action is executed after all of its dependencies), providing inputs for subsequent actions. Compared to a true compiler, `reportengine` currently does little to optimize the code, and it is mostly focused on the automatic generation of the graph. In fact, it does currently add some overhead compared two writing an equivalent script in Python by hand, which is small for applications like `validphys2` where the runtime is dominated by numeric computation in the user code. One difference with compiled languages like C or C++ is that one can use arbitrary logic in the same language (Python) at compile- and run-time, making the distinction somewhat arbitrary. This allows for much more powerful checks at compile time than what can be achieved with a reasonable amount of effort with the preprocessing or templating facilities of the most popular compiled languages (the D languages allows making use of many characteristics of the language at compile time but not arbitrary I/O). One can for example not just check that the type of a given input variable is a string but also check that a given file is present in the user's system and if not, download it. This way, the framework addresses the need of scientific codes to have convenient checks of the input at initialization time, beyond what a typical complied language can provide when compiling the input and what is easily doable by hand in scripted languages. The convention is therefore that everything that might fail should do so at compile-time and as fast as possible, to allow the user to fix it. Past the checking phase, everything that is long-running can reasonably safely be assumed to succeed. One advantage of having an explicit graph is that the parallelization becomes trivial, as we discuss 5.5.1.

To automatically deduce the graph from the minimal amount of user input, `reportengine` makes extensive use of the introspection features of the Python language and a few conventions: The nodes are obtained directly from Python functions (that we call *provider* functions) defined by the client application that uses `reportengine`. The edges are deduced from the names of the parameters in the signature of the function: It is assumed the functions depends on *resources* that have the same name of the parameters, and are computed either from other provider functions or from the user configuration. For example, the `validphys2` code contains a provider function `sum_rules`, defined as follows:

```
import numbers
from reportengine.checks import check_positive
from validphys.core import PDF
```

```
def sum_rules(pdf:PDF, Q:numbers.Real):
    """Compute the sum rules for each member (as defined by libnnpdf), at the
    energy scale ``Q``. Return a SumRulesGrid object with the list of values
    for each sum rule.
    The integration is performed with absolute and relative tolerance of 1e-4."""
    #Code that computes the sum rules
    ...
```

Then, if `reportengine` determines that the function `sum_rules` needs to be computed, it will search try to find the *resources* `PDF` and `Q`. The resources could either be provided in the user's configuration file or be given in terms of other provider functions. In fact users would not be interested in using the `sum_rules` function directly, but rather on a function that displays nicely the result

```
from reportengine.table import table

@table
def sum_rules_table(sum_rules):
    """Return a table with the descriptive statistics of the sum
    rules, over members of the PDF."""
    # Implementation
    ...
```

Thus a runcard such as

```
#sumrulestable.yaml

Q: 10
pdf: NNPDF31_nnlo_as_0118

actions_:
  - - sum_rules_table
```

would generate the graph in Fig. 5.1, by analyzing the application source code and the runcard without further developer intervention to assemble it. Once executed, by the `validphys` executable (which is a `reportengine` based application), the runcard above saves the table with the descriptive statistics of the sum rules over replicas. An important assumption is that the functions defining the actions (such as `sum_rules_table` above) are *pure in practice*: That is, it is assumed that executing two times the same function with the same outputs will yield the same results for all practical purposes and have no other side effects. The side effects that are needed, such as writing figures and tables to disk are managed by the framework itself. For example the `@table` decorator above takes care of writing the returned Pandas DataFrame[184] structure and write it to a correct path. Similarly, there is a `@figure` decorator that handles Matplotlib[185] based plots.

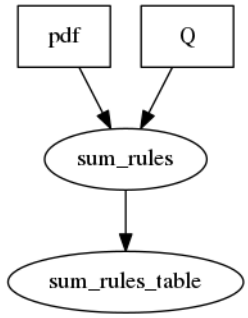


Figure 5.1: DAG representing the computation required to compute a table with sum rules. The graph is generated automatically when the user requests the final action.

All inputs are ultimately derived from the user provided runcard. The `reportehgine` framework contains a `Config` class that is meant to be extended by the client applications. It allows to write code like:

```
from reportengine.config import Config
from validphys.core import PDF
class ValidphysConfig(Config):
    def parse_pdf(self, user_input:str):
        #error out if not found
        ...
        return PDF(user_input)
```

A method starting with `parse_` automatically binds to the corresponding value in the user input card. The type of the parameter (that we declared as string in the example above) is also checked automatically before entering the parsing function. Optional default values are similarly supported. Similarly several inputs can be combined at compile time using *production rules*. These are methods starting with the prefix `produce_` that take zero or more already parsed inputs and return a resource that can be requested by the user or by other provider functions.

The approach to building the graphs as presented so far lacks flexibility because there is a bound fixed value for the parameters of each function (in this case `pdf` and `Q`). The problem is resolved by making extensive use of namespaces, as we describe in Sec.5.5.1.

We now describe how this architecture is designed to solve the problems that typically appear in scientific code.

Main features of the framework

Early, decoupled error checking The framework addresses three particularities of scientific software:

1. Computations can take a long time to run.
2. Outputs are frequently unexpected even when everything is working correctly, for reasons that are not obvious at first.

3. Computations depend on an complicated set of parameters that not always can be encapsulated efficiently.

It is therefore necessary to check the inputs preemptively in order to assure that a large amount of computation time isn't wasted, possibly followed by tedious debugging session because of a trivial mistake in the input parameters.

The further requirement to allow great dynamism in how the nodes are composed and executed implies that the error checking of the parameters cannot be made inside the functions themselves. This is clear considering a provider function like:

```
def check_parameter_is_correct(parameter):
    #Raise error if parameter is not correct
    ...

@figure
def final_plot(complex_calulation, parameter):
   check_parameter_is_correct(parameter)
   #do the plot
   ...
```

This pattern is unpalatable if `complex_calculation` takes a large amount of time and `parameter` is given by the user and known immediately after the program starts. In that case we would like to know if the parameter satisfies the constraints required by the action before any computationally expensive work takes place. `reportengine` offers functionality to do precisely that. We would write instead:

```
from reportengine.checks import make_argcheck

@make_argcheck
def check_parameter_is_correct(parameter):
    #Raise error if parameter is not correct
    ...

@figure
@check_parameter_is_correct
def final_plot(complex_calulation, parameter:int):
   #do the plot
   ...
```

This construct causes the checking function `check_parameter_is_correct` to be called at the time at which the graph is constructed (*compile time*), passing the user-supplied `parameter` as argument. In this way the user can know if there are avoidable problems with the input before committing large amounts of time such as when sending jobs to a computing cluster. The type specification is tested at compile time (in this example, we require that the parameter is an integer), before the check function, allowing to omit the type checks in this contents.

In principle, the checking facilities allows to impose on the client applications the specification that any uncaught exception inside the provider functions, after the input passes all the checks, is a programmer error in the application. Since `reportengine` is specialized in the situation where the output is a deterministic function of the user provided inputs, it is theoretically possible to always check that the input to the provider function is valid. In practice the assumption can be broken, since it is possible that the environment changes (for example the existence of a given file was corroborated at some point but the file was removed later when it is actually needed) or it is possible that some of the input files are corrupt (for example a PDF grid that was extracted only partially from a tarball missing some of the replica files). In such cases it is considered that the programs is in an invalid state and it is therefore adequate to abort it.

The per-function checks as described here complement those defined in the `Config` class: The checks in the functions can be used for parameters that are only going to be used for a particular provider or that restrict further a generally valid resource (for example several actions in `validphys2` require that the error type (see Sec. 3.2) of the PDF is based on replicas).

Declarative input It is convenient to be able to specify the *what* the program should only without any regard or knowledge of *how* that is achieved by the underlying implementation. The primary input of `reportengine` applications are YAML run cards. A valid input card for the `validphys2` application looks like this:

```
pdfs:
    - NNPDF31_nlo_as_0118
    - NNPDF31_nnlo_as_0118

First:
    Q: 1
    flavours: [up, down, gluon]

Second:
    Q: 100
    xgrid: linear

actions_:
    - First:
        - plot_pdfreplicas
        - plot_pdfs
    - Second:
        - plot_pdfreplicas
```

This example illustrates several advantages of the framework

Correct by definition A declarative input specifies what user wants. It is up to the underlying code to try to provide it (or fail with an informative message).

Obvious meaning It is easy for a human to verify that the input is indeed what it was intended. Even without any explanation it should be easy enough to guess what the runcard above does: We declare a list of PDFs that we want to compare in two different ways, determined by the *namespaces* `First` and `Second`. In the `First` comparison, we want to plot the PDF replicas and error bands at $Q = 1$GeV, and only for a subset of flavours. In the `Second` comparison we want to plot the PDF replicas at $Q = 100$GeV on a linear scale.

Implementation independent The input is very loosely coupled with the underlying implementation, and therefore it is likely to remain valid even after big changes in the code are made. For example, in the runcard above, we didn't have to concern ourselves with how LHAPDF grids are loaded, and how the values of the PDFs are reused to produce the different plots. Therefore the underlying mechanism could change easily without breaking the runcard. Thus saving the runcard (which is done automatically when running a `reportengine` application) greatly aids the reproducibility of the result.

Namespaces and loops As we saw in the previous example, one can perform actions with different parameters by wrapping each set of parameters in a *namespace*. Also, arguably the main pattern in high level scientific analysis is testing something scanning for multiple values of a given parameter. The most frequent data structure is a list and the most useful control flow construct is a loop. To implement this, lists of namespaces are created and expanded automatically. Therefore these both have first class support in `reportengine`. The following rules apply to construct namespaces: - Every unrecognized mapping in a configuration file can be interpreted as a namespace. - Every list of mappings can be interpreted as a list of namespaces. - Actions can be performed assuming a particular namespace or list of namespaces. In the later case, all the parameters are resolved within each namespace in the list, and the action is executed with each set of parameters as input. - Namespaces can be stacked together. A given input is searched first in the innermost namespace of the stack, and if it cannot be resolved, it is searched in the outer ones, until reaching the global one. When lists of namespaces are staked, the result is the Cartesian product of the lists.

Furthermore one can trivially construct a parser for a list from a parser of the element of the list, by adding the `element_of` decorator. For example, the following:

```
@element_of('pdfs')
def parse_pdf(self, user_input:str):
    #error out if not found
    ...
    return PDF(user_input)
```

Causes the key `pdfs` in the configuration file to be parsed as a list of PDF objects. Additionally, when used as a namespace specification, the lists will act as a list of namespaces, each containing a single key with the corresponding element (in the example above, `pdf`). For example, the following `validphys2` runcard computes the sum rules (that require a PDF and a value of Q as input) a total of 12 times for the for each of the two PDFs and for each value of Q:

```
pdfs:
    - mcpdf_pdf4lhc_test_replicas
    - mcpdf_test_replicas

Qs:
 - Q: 1
 - Q: 1.2
 - Q: 1.65
 - Q: 2
 - Q: 10
 - Q: 100

actions_:
  -  pdfs:
     - Qs:
        - sum_rules_table
```

The report function offers a more convenient syntax than the nested YAML based `actions_` specification.

The report provider function Reports are implemented as an action of `reportengine`. The action takes a `template_text` argument, which corresponds to a text following the `pandoc` flavour of the Markdown syntax, additionally containing special markers that will be interpreted as actions and namespace specifications. The actions will be resolve as if they where directly specified in the configuration file and when all of them are completed, their value will be substituted in the template. The markers are strings between `{@` and `@}`. There are currently **target** and **with/endwith** tags:

Target tags Specify an action to be executed. The possible syntax is:

```
{@[spec] <action_name>@}
```

where `[]` stands for optional syntax. A few conforming examples are:

```
{@ sum_rules_table@}
```

```
{@Qs::pdfs sum_rules_table@}
```

The optional namespace specification works as described in SOMEWHERE. The different parts of the specification, which can be mappings, or lists of mappings (or special tags implementing that behaviour) are separated with the `::` operator (resembling the C++ scope resolution operator). Actions will be repeated if the specification results in multiple namespaces.

`with/endwith` tags Repeat the content between the tags for each namespace in the specifications. Targets inside the block are repeated and searched within each namespace. The syntax of the `with` tag is:

```
{@with <spec>@}
```

where `spec` is the same as for the target tag. It must be closed by an `endwith` tag

```
{@endwith@}
```

Like in the **target** tag, the spec is separated by `::`.

A version of the example above that generates a report in addition to a set of table files is:

```
pdfs:
    - mcpdf_pdf4lhc_test_replicas
    - mcpdf_test_replicas

Qs:
 - Q: 1
 - Q: 1.2
 - Q: 1.65
 - Q: 2
 - Q: 10
 - Q: 100

meta:
    title: Sum rules for Monte Carlo PDFs
    author: Zahari Kassabov
    keywords: [mcpdfs, test]

template_text: |

    {@with pdfs@}

    # {@pdf@}
    {@with Qs@}
    ## Q = {@Q@} GeV
    {@sum_rules_table@}

    {@endwith@}
    {@endwith@}

actions_:
  - - report:
        main: True
```

The result is shown in Fig 5.2.

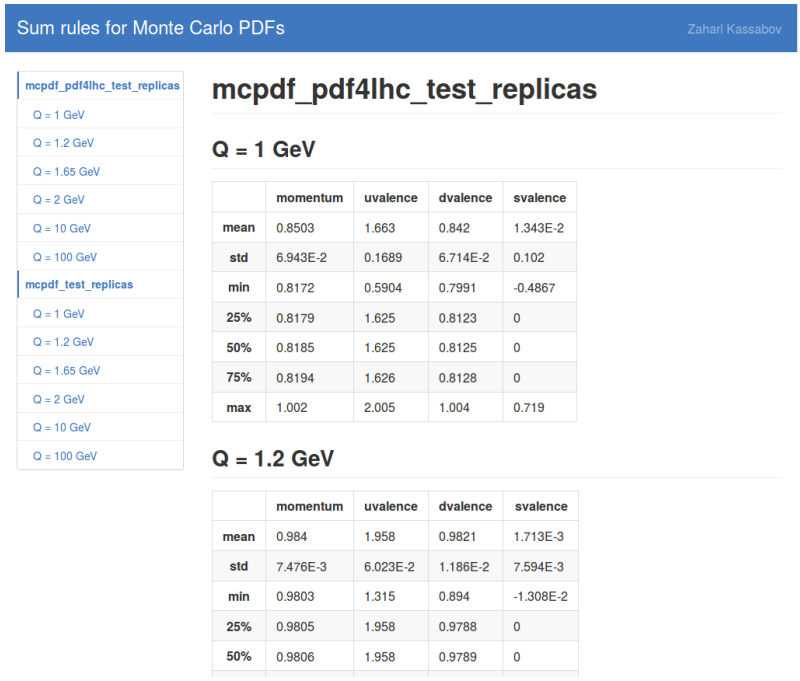


Figure 5.2: Example output of the validphys action that computes sum rules. The result of a scan for multiple energy values is displayed neatly in an HTML page. The results at low Q show that PDF values computed at low scales using the LHAPDF extrapolation feature do not reproduce the sum rules properly.

Parallel processing Since the computation is represented in terms of a DAG where the only allowed interactions between the nodes are the edges, the dependency structure is made explicit. This allows to trivially execute the graph in parallel. The algorithm consists of transvasing the graph in topological order and adding to a queue of nodes that contain no incoming edges (that is, the functions have no inputs). The nodes in the queue are executed in parallel by a pool of process workers and as tasks are completed, the outgoing edges of the corresponding node are removed, thereby allowing new tasks to be added to the queue.

While this approach allows to effectively use the available number of computing cores in some situations, The effective use of this feature is somewhat hampered by two limitations in the CPython implementation: lack of truly immutable data structures and incapacity to take advantage of multiple threads in the same process. Instead one has to resort to creating one operating system process per worker, and the communication between tasks requires that the data is serialized and in the origin and deserialized in the destination, which can be expensive compares to executing a task. Furthermore, some resources (like those backed by C/C++ wrappers like in those coming from the NNPDF code) are not serializable and must be initialized in each process, which often negates completely the multiprocessing advantage. Therefore the approach only works optimally when the nodes perform expensive computation while

exchanging little data.

More advanced approaches based on compacting the graph, so that the amount of I/O is minimized are under investigation.

Help system The `reportengine` framework provides an automated command line based help system for each defined action. The documentation is generated by parsing the *docstring* of the corresponding provider function, as well as by figuring out the dependencies. For example, the output of `validphys --help sum_rules_table` is:

```
sum_rules_table

Defined in: validphys.pdfgrids

Generates: table

sum_rules_table(sum_rules)

Return a table with the descriptive statistics of the sum rules, over
members of the PDF.

The following resources are read from the configuration:

    pdf: A PDF set installed in LHAPDF. Either just an id (str), or a
  mapping with 'id' and 'label'.
  [Used by sum_rules]

The following additionl arguments can be used to control the
behaviour. They are set by default to sensible values:

  Q(Real) [Used by sum_rules]
```

The collect function and arbitrary comparisons The `reportengine` framework provides an API function, called `collect` ,that allows to collect results produced in different namespaces. It takes a function and a namespace specification and returns the result of the function when the inputs are resolved in each of the namespaces spanned by the specifications. The most straight forward use is to map some actions over a list of equivalent items and then collect them into a plot or table. For example:

```
from reportengine import collect
from reportengine.table import table

# def abs_chi2_data_experiment(...) compute the chiš for a given
#                                   experiment

experiments_chi2 = collect(abs_chi2_data_experiment, ('experiments',))

@table
```

```
def experiments_chi2_table(experiments, pdf, experiments_chi2,
                           each_dataset_chi2):
    """Return a table with the chiš to the experiments and each dataset on
    the experiments."""
```

The `collect` function can also solve a more complicated problem in this framework; namely how to perform a comparison where arbitrary differences in the inputs can be introduced. For example, a simple data theory-comparison plot in the NNPDF code depends on the following parameters:

- The name of the dataset.
- The specification of the experimental systematics.
- The theory settings used.
- The cuts of the data.
- The PDF(s) entering the comparison.

Ideally we want to be able to compare arbitrary aggregates of variations of these options. To achieve that we choose an arbitrary name for a list of namespaces (for example, `dataspecs`), and use the standard `reportengine` functionality to resolve the dataset differently within each namespace. For example we might use the data-theory comparison tool `plot_fancy_dataspecs`:

```
fit: NNPDF31_nlo_as_0118_1000
use_cuts: True

normalize_to: data

dataset_input:
    dataset: NMC

dataspecs:
    - theoryid: 52
      pdf: NNPDF31_nlo_as_0118_hessian

    - theoryid: 53
      pdf: NNPDF31_nnlo_as_0118_hessian

template_text: |
    % NLO vs NNLO comparison for {@dataset_input@}
    {@plot_fancy_dataspecs@}

actions_:
    - - report:
            main: True
```

In this example, we are comparing the value of the NMC SOMETHING data, with the theory option 52 (corresponding to a NLO theory) convolved with the NNPDF

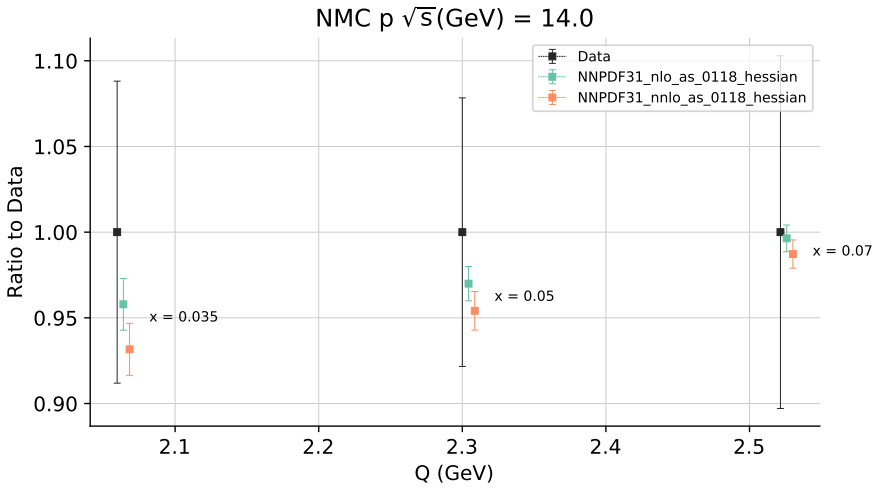


Figure 5.3: An example of an arbitrary comparison performed with the `validphys` code. The user can specify declaratively the parameters of the comparison and have a series of checks ensuring that the input is valid. At the same time, the code does not need to be adapted specifically for each possible variation.

3.1 NLO set and the NNLO theory (`theoryid` 53) with the corresponding NNLO set. The result (for one bin) is displayed in Fig. 5.3. The signature of the plotting function is:

```
@_check_same_dataset_name
@_check_dataspec_normalize_to
@figuregen
def plot_fancy_dataspecs(dataspecs_results, dataspecs_commondata,
                         dataspecs_cuts, dataspecs_speclabel,
                         normalize_to:(str, int, type(None))=None):
```

Where the parameters starting with `dataspecs_` are computed in terms of the `collect` function over the `dataspecs` list of namespaces. For example:

```
dataspecs_commondata = collect('commondata', ('dataspecs',))
```

Note that the plotting function has checks that ensure that the inputs fulfill the relevant constraints (in this case, that we are comparing the same experimental data). This is the most powerful feature of `reportengine`: It allows arbitrary comparisons that are expressed without code in a declarative way and checked for correctness before any expensive computation is executed.

An example: A simple application to debug interpolation problems

During the development of NNPDF 3.1 it was found that some of the interpolations that were used in the fit were causing discrepancies in the theory prediction, when comparing the result produced from the source APPLgrids and the FKTables that were produced from them. The discrepancy was of few percent, which is very notable compared to the PDF uncertainty of precise LHC observables such as the Zp_T data, which are of order 0.2%. The difference in the predicted PDF uncertainties for those observables could reach 50%. Debugging the issue was non trivial because there are several interpolations involved in the FKTable production. These include:

- The interpolated grid in which the DGLAP equations are solved by APFEL.
- The interpolation of the fitted neural networks that is written to LHAPDF.
- The finite grid in which the FKTables are interpolated where the convolution is performed.

Understanding the problem required to methodically disassemble all the effects of these interpolations (and the different ways they commute in different parts of the procedure). The issue was clarified by producing predictions obtained with variations of the interpolation settings. The results were analyzed with a small `reportengine` based application totaling 114 source lines of Python code (SLOC), including all the boilerplate required to set up the framework. The application allowed to quickly process and visualize the results and helped to finally pinpoint the issue (see an example output in Fig. 5.4).

It was found that the settings for the APFEL interpolation grids that were used in the fits and in the preparation of FKTables were the biggest contributors to the interpolation inaccuracy. A change consisting on solving the DGLAP equation on a custom, more dense, grid in x was shown to resolve the issue satisfactorily.

5.5.2 The validphys2 project

As we have illustrated in the previous sections, `validphys2` is a reportengine based application that provides tools to analyze data that is relevant to the NNPDF project. It links to `libnnpdf` trough a wrapper for the Python language constructed with the SWIG program. `validphys2` can also be used as a Python library, allowing more convenient access to several `libnnpdf` features (including elemental ones, such as loading a dataset) that are tedious to attain directly in the C++ code.

The `validphy2` contains over 80 table generating and plotting tools, many of which have been showcases thouought this document. Some are generally useful for all the projects in the collaboration, such as PDF comparisn plots or data-theory predictions while others are specifically useful for a particular project, such as the determination of the strong coupling constant. It is convenient to systematize the tools within the same code because in this way all of them can take advantages of the features of the core code, such as the automatic localization of resources (see Sec. 5.5.2), extensive error checking, and the possiblity to share the result as an HTML page (see Sec. 5.5.2). Rather than decribing the ever-icreasing collection of tools, we proceed to list some more notable features of the code that were implemented on top of the existing C++ code and `reportengine` (See Sec.5.5.1).

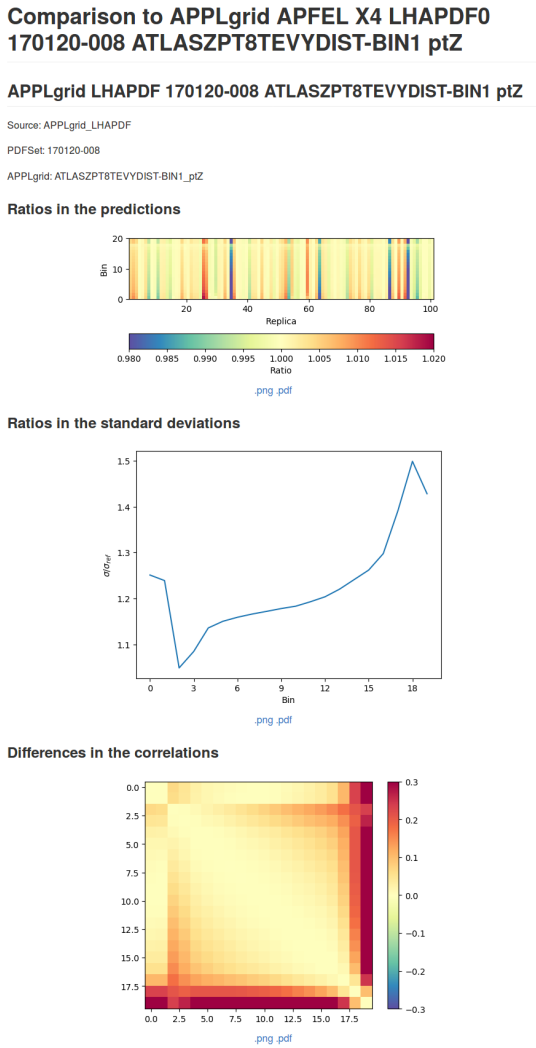


Figure 5.4: Example of a part of the output of the simple program used to debug the interpolation issues. The figure shows the output of a comparison with a particular configuration of various grid settings (labeled "X4"). From top to bottom, the plots show the ratios between the predictions used the internal NNPDF FKTables and the APPLgrids that were used to generated them, the ratios in the PDF standard deviation of the predictions and the difference in correlation coefficients between bins.

Sharing tools

Validphys allows members of the NNPDF collaboration to easily and automatically upload the output of an analysis to a server that is accessible to all of them. This merely requires appending `--upload` flag to the invocation of the `validphys` executable. At the end of the run, the output folder will be stored online and the user

will see an URL where the results can be viewed. The metadata block defined in the user input card is also used to index the result so it can be fount later, as illustrated in Fig. 5.5. While designed with reports in mind, the infrastructure can be used to share and index any type of file. The repositiry is currently hosted at the University of Milan and regulary backed up in two physical locations.

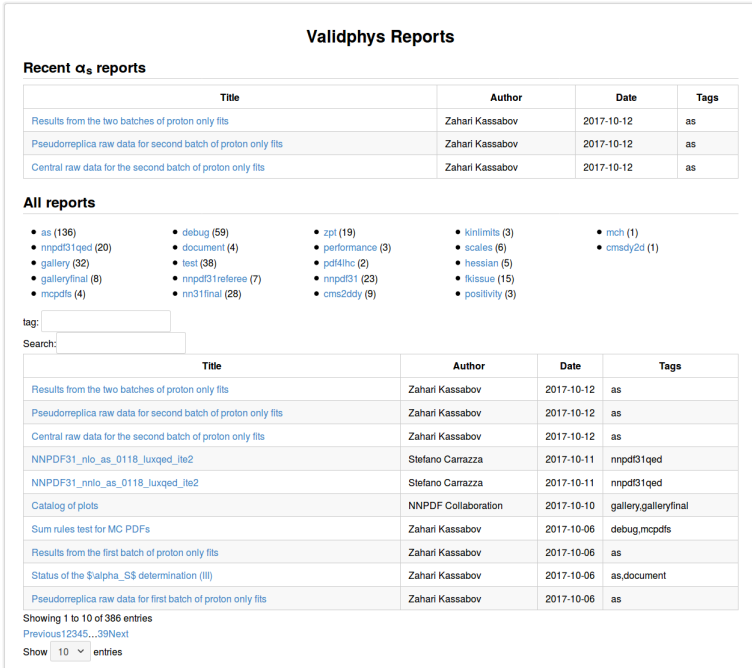


Figure 5.5: The private validphys reports page. The page collects and allows to find all the produced reports in the NNPDF collaboration.

Automatic downloading of resources

When some `validphys` action requires a resource such a completed fit, an LHAPDF set, or set of FKTables, and the files cannot be found locally, they will be automatically downloaded from the relevant NNPDF or LHAPDF repositories. This makes it easy to reproduce an existing `validphys` output given only the runcard to generate it: It is considered a bug if running `validphys <existing runcard>.yaml` does not give a comparable result to the original runcard on a correctly configured system, and without further need to install any grid.

Plotting format specification

A pressing problem of the NNPDF framework is the dissociation between data and metadata. For example, it used to not be easy to automatically plot some prediction as a function of a kinematic variable because that information was not recorded. This

information is now encoded in a declarative way for each included dataset, allowing figures like SOMETHING and SOMETHING to be produced automatically.

Binary packaging

All the NNPDF projects required to run `validphys2` are automatically packaged in the `conda` format, when a change is pushed to the relevant continuous integration services. Currently versions of the packages for both Linux and Mac are produced.

This allows users with access to the NNPDF repositories to set up all the required codes by simply typing

```
conda install validphys
```

which will automatically pull all the dependencies, rather than requiring having them installed by hand. While some basics checks are performed implementing a full testing system is still a pending task.

Chapter 6

A determination of the strong coupling constant

6.1 Introduction

The strong coupling constant α_s (see Sec. 2.2) is an essential input to any calculation in perturbative Quantum Chromodynamics (QCD), since it is the only free parameter in the theory beside the quark masses. It enters the theoretical predictions of hard-scattering processes both though the explicit dependence of the partonic cross section and implicitly though its dependence on other quantities, particularly the PDFs. The precise determination of α_s is therefore relevant for phenomenology at the LHC. Currently α_s uncertainties contribute notably to the uncertainty of relevant Standard Model processes such as the Higgs Cross Section [16]. Furthermore, recent improvements improvements in fixed order computation [186] and determination of PDFs [4] highlight the importance of reducing the α_s uncertainty.

Many different observables can be used to determine α_s; several independent determinations are combined by the Particle Data Group (PDG) to obtain a *World Average* [23]. The determinations entering the PDG average are based on hadronic τ decays, Lattice computations, PDFs, e^+e^- hadronic annihilations, electroweak precision fits and $t\bar{t}$ production data. These observables all require making additional assumptions on the underlying Physics, other than perturbative QCD. This includes reliance on non perturbative Lattice or Monte Carlo models (τ decays and e^+e- processes), or the strict validity of the Standard Model (for EW precision fits).

The determinations of α_s based on PDFs require both a robust fitting methodology, with the suitable adaptations to extract α_s, and a precise understanding of the theoretical and experimental inputs that enter PDF determination. This presents several significative challenges: For examples biases arising from the choice of parametrization of the PDFs have been identified [187] as an important problem and has necessitated relatively recent developments in the methodologies such as closure tests [8] (see Sec. 5.1.9) and dynamical tolerances [64, 75] in order to be satisfactorily resolved.

Moreover, the results of PDF determinations will be affected by the inadequacies of the theoretical calculations of the processes entering the fit such as missing higher order corrections, the values of the heavy quark masses, deuteron and nuclear corrections, or higher twist effects.

In spite of the these complications, a determination of α_s based on a global PDF fit also presents important advantages. Because the fit of α_s is performed on multiple experimental measurements of diverse physical processes, the possible defects in the description of the data (both of theoretical and experimental origin) can be reasonably expected to be predominately uncorrelated, which in turn implies that defects should average out to some extent in the final result. Therefore, despite the difficulty of realizing a precise description of each of the inputs, a small defect in one of them (e.g. due to the underestimation of the experimental uncertainties or large missing higher order corrections) has a smaller impact on the result than it if the determination was based solely on the problematic dataset, or indeed on any less global subset of the data inputs entering the PDF fit, provided that no particular reason for excluding some data from the determination has been identified.

In addition, while several recent determinations of α_s based on hadronic data have been presented [188, 189, 190, 191, 192, 193] (see also [194] for earlier results), the ones based on the simultaneous fitting of the PDFs and $\alpha_s(m_Z)$ take consistently into account the dependence of the result on the whole dataset entering the PDF fit. We discuss this in further detail in Sec 6.4.

In this chapter, we present an update of the previous NNPDF determination [195, 196] based on NNPDF 2.1 [197, 198], employing the NNPDF 3.1 global analysis [12] described in Chapter 5 as an input.

This is motivated by the development of a new methodology to extract α_s, as well as by the many subsequent improvements made to the NNPDF fits, which have culminated in NNPDF3.1 global analysis [12]. The input dataset has been significatively enhanced with new measurements (see Sec 5.2.2), and we use NNLO QCD theory for all experimental predictions.

The main advantage of the new method (which we dub *correlated replica method*) is to propagate directly all the components of uncertainty on the PDFs into the result for α_s, including the experimental uncertainty on the data, and the uncertainty induced by the PDF fitting methodology. We call these uncertainties *PDF uncertainties on* α_s. The previous method employed in Refs [195, 196] only accounted for effects in the central PDF, but not on the error members, and might be outright inadequate *e.g* in the presence of inconsistent experiments. In order to improve both the precision and the accuracy of our results we have employed for the first time a *batch minimization* strategy: We determine α_s from the best of several runs of the PDF minimization algorithm. The effect is to minimize the dependence on outliers that may pull the final value significatively and reduce the dispersion of the results due to the finite efficiency of the minimization algorithm.

The combinations of the improvements in input dataset, theoretical calculations, and fitting methodology, lead to the following result for the strong coupling constant at NNLO

$$\alpha_s(m_Z) = 0.11845 \pm 0.00052^{\rm pdf}\ (0.4\%) \pm 0.000027^{\rm stat}\ (0.02\%)\,, \tag{6.1.1}$$

which is compatible with the PDG world average and with high precision (of 0.4%) when considering only the uncertainties that can be reliably estimated in our framework, which include the experimental uncertainty of the data and that related to the PDF fitting methodology, but notably do not directly provide an estimate for missing higher order uncertainties (MHOU). As a consequence our results need to be evaluated

carefully: The results require a systematic scrutiny of all the elements of the methodology which could contribute significantly to the uncertainty at this level of precision. To that end, we have systematically confirmed that all the procedural choices (such as the number of fitted replicas) induce fluctuations on the value of α_s that are much smaller than the PDF uncertainty and are therefore irrelevant to our determination. We also require at least a rough estimate of the theoretical uncertainty in order to interpret the result properly: We estimate an upper bound on the MHOUs are by halving the difference between the NLO and NNLO determinations of α_s. As we will show, these turn out to be the overall dominant source of uncertainties in the present α_s determination.

6.2 Fitting methodology

The main methodological development in this new determination is the introduction of the *correlated replica method* used to obtain the determination of α_s. Its main purpose is to take into account all sources of uncertainty that enter a standard PDF fit, particularly the experimental uncertainty in the input data, and propagate this uncertainty on the determination of α_s. By contrast, the method described in Refs. [195, 196] (henceforth called the $\Delta\chi^2 = 1$ method) is based on determining α_s from the best fit PDF only and does not take into account the variations in the fitted PDFs at each value of α_s.

6.2.1 The correlated MC replica method

The methodology

To describe the α_s fitting methodology it will be convenient to first repeat several aspects of the NNPDF fitting procedure introduced in Sec. 5.1 and described in detail in Ref. [8]: To perform a fit, we first sample $N_{\rm rep}$ realizations from the probability distribution that describes the experimental inputs. We call these samples *pseudodata replicas* (See Sec. 5.1.5). Each pseudodata replica is then fitted to a set of functional forms (one for each fitted parton distribution) parametrized by neural networks multiplied by a preprocessing function (See Sec. 5.1.1). The set of functions obtained by this procedure is called a *PDF replica*, or simply *replica*. For each of the $N_{\mathcal{D}}$ experimental points entering the fit, the pseudodata replica samples are generated according to the Eq 5.1.24, which we repeat here, now explicitly indicating the replica dependence:

$$D_I^{(r)} = D_I^0 + \sum_{J=1}^{N_{\mathcal{D}}} C_{IJ}^{\frac{1}{2}} d_J^{(r)} \ , \tag{6.2.1}$$

where the index r identifies the replica, $D_I^{(r)}$ is the pseudodata point indexed by $I = 1, \ldots, N_{\mathcal{D}}$, D_I^0 is the corresponding experimentally measured central value, $C^{\frac{1}{2}}$ is the transpose of the Cholesky decomposition of the covariance matrix C and $d_J^{(r)}$ is a random number sampled from a standard normal distribution.

As discussed in Sec 3.2.1, any quantity $\mathcal{O}$ that depends on the PDFs adopts a different value, denoted $\mathcal{O}^{(r)}$, for each replica labeled by the index r.

Fitting a given PDF replica consists of performing the minimization of the error function χ^2 (Eq. 5.1.22) as a function of the set of parameters, $\{\theta\}$ (see Sec 5.1.7), that characterize the PDF functional form. Writing again the replica dependence explicitly, we have

$$\chi^{2(r)}\left[\{\theta\},\alpha_s,\mathcal{D}\right] = \sum_{I,J=1}^{N_{\mathcal{D}}} \left(T_I[\{\theta\},\alpha_s] - D_I^{(r)}\right) C_{IJ}^{-1} \left(T_J[\{\theta\},\alpha_s] - D_J^{(r)}\right) \ , \tag{6.2.2}$$

As explained in Sec. 5.1.4, a cross validation procedure is employed to avoid overfitting during the minimization. The fitting procedure then consists on finding a minimum for Eq 5.1.26,

$$\chi^{2(r)}_{\min}\left[\alpha_s,\mathcal{D},\xi\right] = \mathrm{cv}\min_{\{f\}} \chi^{2(r)}\left[\{f\},\alpha_s,\mathcal{D}\right]\Big|_{\alpha_s,\mathcal{D},\xi} \ , \tag{6.2.3}$$

and retrieving the PDF replica from the parametrization that minimizes the error function. The value of α_s, and the dataset $\mathcal{D}$ are held fixed during the minimization.

Simultaneous minimization of PDFs and α_s

The central idea behind the correlated replica method consists of selecting the best value of α_s by minimizing the error function in Eq. 6.2.3:

$$\alpha_s^{\min(r)}\left[\mathcal{D},\xi\right] = \arg\min_{\alpha_s} \chi^2\left[\alpha_s,\mathcal{D},\xi\right] \ , \tag{6.2.4}$$

where the notation arg min indicates that the right hand side corresponds to the argument of the function that minimises the χ^2.

The crucial point is that the *same* set of pseudodata replicas is used as we scan the values of α_s in search of a minimum, hence the name of *correlated replicas*. It is clear from the notation used that the value of α_s obtained by this procedure depends on the choice of the dataset, and potentially on the details of the minimization. The correlated replica method performs first the minimization in the space of PDFs at fixed α_s, and then the minimization over α_s, using the same set of data replicas for the entire procedure. Note that in this procedure α_s is on the same footing as the parameters $\{\theta\}$ that characterize the functional form of the PDFs, and the minimization procedure spans this enlarged space of parameters. The only difference stems from the cross validation procedure, which is only applied when minimizing over the set $\{\theta\}$, and is expected to have no effect on the best value of α_s. Therefore α_s could in principle be treated as another parameter in the fit. However this cannot be directly implemented in the `FastKernel` framework [9], which is used in the NNPDF fits, and achieves a considerable speedup by fixing all the parameters of the DGLAP evolution, including α_s. Instead we perform multiple fits scanning over a range of values of α_s: we fit a set of PDF replicas $f^{(\alpha,r)}$, organized as a $N_{\alpha_s} \times N_{\mathrm{rep}}$ table, where the index α runs over the discrete set of values of α_s, and the index r characterizes the pseudodata replica. All the replicas with the same index r are fitted using different values of α_s, but the same set of random numbers $\{d_J\}$ when sampling pseudodata in Eq. 6.2.1.

6.2.2 Minimization strategy

We now discuss how to implement the minimization Eq 6.2.4 in practice: We first describe in Sec 6.2.2 how do we associate a value of $\alpha_s\left(M_Z^2\right)$ to a given pseudodata

replica. Next, in Sec 6.2.2, we describe how do we obtain a finite size uncertainty, associates to the finite number of pseudodata replicas we fit. Finally we address some technical difficulties associated to our methodology: Since not all the PDF replicas in satisfy the convergence criteria at the end of the PDF fit, described in Sec. 5.1.8, some entries in $f^{(\alpha,r)}$ will be undefined. Therefore it may not be possible to obtain a value of α_s from all pseudodata replicas. The set of selection criteria that determine which values we consider are explained in Sec. 6.2.2.

Parabolic fitting

We associate to each of the remaining replicas the corresponding value of the minimum of the error function Eq. 5.1.26 attained during the fit,

$$\chi^{2(r,\alpha)} = \chi^2_{\min}[\alpha, \mathcal{D}_r, \xi_r] \ . \tag{6.2.5}$$

In order to increase the resolution that can be achieved with the finite grid in α_s, we take advantage of the fact that the profile of the error function $\chi^{2(r)}(\alpha_s) \equiv \chi^2_{\min}[\alpha_s, \mathcal{D}_r, \xi_r]$ can be expanded as a quadratic polynomial around $\alpha_s^{\min}$; therefore we determine $\chi^{2(r)}(\alpha_s)$ by performing a quadratic fit to the discrete set of N_{α_s} values that we determine by minimization. In Sec 6.3.5 we have checked that the effects of the Taylor approximation are negligible compared to the relevant uncertainties. We end up with a set of *curves*,

$$\chi^{2(r)}(\alpha_s) = m^{(r)} \left[\alpha_s - \alpha_s^{\min(r)}\right]^2 + c^{(r)} \ , \tag{6.2.6}$$

labeled by the index r, each of them obtained from the corresponding replica r, if it satisfied the selection criteria that we will describe in Sec. 6.2.2. The number of curves that we obtain following this prescription is denoted $N_{\rm curves}$, where clearly $N_{\rm curves} \leq N_{\rm rep}$. Given the set of correlated replicas fitted for a fixed pseudodata sample and different values of α_s, $f^{(\alpha,r)}$, we obtain the coefficients $m^{(r)}$, $\alpha_s^{\min(r)}$, $c^{(r)}$ from the fit to a quadratic polynomial, performed by least squares. Finally we select $\alpha_s^{\min(r)}$, the minimum of the parabola, as the value predicted from the curve corresponding to the pseudodata replica indexed by r.

The result of this procedure is a set of values $\{\alpha_s^{\min(r)}\}$, which describes the fluctuations on α_s induced by the fluctuations in the data, and hence effectively propagates the PDF uncertainty, obtained from the standard PDF fits, to the α_s determination.

Bootstrapping resampling

A remaining possible source of uncertainty are the finite size effects that appear due to selecting a limited number of curves that pass the fitting criteria. In principle we can always obtain more replicas and improving the finite size uncertainty is bounded only by computational cost. We produce a large enough number of curves so that this reducible finite size uncertainty is negligible compared with the PDF uncertainty. To estimate this uncertainty, we implement a *case resampling bootstrapping procedure*: In our case the sample is the set of minima of each of the selected *curves*, $\{\alpha_s^{\min(r)}\}_r$, with $r = 1 \ldots N_{\rm curves}$. The method consists on constructing a large number $N_{\rm resamples}$ of resamples of the original sample. Each resample is a new set of $N_{\rm curves}$ values drawn

with replacement from the original sample. We then compute the statistical estimator of interest for each of the resamples, and assess its variation over the ensemble. $N_{\rm resamples}$ is chosen large enough so that the results do not depend on it within the required precision.

We now apply these steps explicitly to estimate the finite size uncertainty of the α_s determination. We estimate the central value of α_s as the mean of the minima of each curve,

$$\alpha_s^{(\rm central)} = \left\langle \{\alpha_s^{\min(r)}\} \right\rangle_r , \ r = 1 \dots N_{\rm curves} \ . \tag{6.2.7}$$

The corresponding bootstrapping uncertainty is then obtained as

$$\Delta_{\alpha_s}^{\rm stat} = {\rm std} \left(\left\langle \{\alpha_s^{\min(\rho),s}\} \right\rangle_\rho \right)_s , \ \rho = 1 \dots N_{\rm curves}, \ s = 1 \dots N_{\rm resamples} \tag{6.2.8}$$

where each $\alpha_s^{\min(\rho),s}$ is sampled with equal probability from the original sample $\{\alpha_s^{\min(r)}\}_r$.

We may similarly estimate the finite size uncertainty on the PDF uncertainty as

$$\Delta_{\alpha_s}^{\rm stat} = {\rm std} \left({\rm std} \left(\{\alpha_s^{\min(\rho),s}\} \right)_\rho \right)_s , \ \rho = 1 \dots N_{\rm curves}, \ s = 1 \dots N_{\rm resamples} \tag{6.2.9}$$

We find that the results are independent of the random seed used to generate the boostrapping resamples (up to the two first significative figures in the uncertainties) when $N_{\rm resamples} = 10000$.

We use the finite size uncertainties on the mean Eq. 6.2.8 as a criterion to decide how many replicas we require to obtain a determination of α_s where the PDF uncertainties dominate over the finite size ones. The uncertainty on the standard deviation Eq. 6.2.9 is used as a criterion for discarding outliers resulting from undersampled curves, as we discuss next in Sec. 6.2.2.

Selection criteria

The approach described so far needs to be refined to take into account two limitations of the NNPDF methodology.

The first one is that not all the replica fits converge, as a consequence of the quality criteria (described in Sec 5.1.8) that are imposed on them. The second limitation is that the random state with which the genetic algorithm is initialized has a small but measurable effect on the value of the error function. As discussed in Sec. 5.1.6, this dependence is due to the limited efficiency of the genetic algorithm and the variations in the cross validation/training splitting. The effect is to decrease the precision of the α_s determination since the parabolic fits of the error function profile $\chi^2(\alpha_s)$ rely crucially on the profile being smooth. We shall now discuss these issues in turn.

Firstly, in the correlated replica method we are using the *same* set of pseudodata to perform fits at *all* the values of α_s that we consider (up to small corrections due to the implementation of normalization uncertainties). It may happen for particular pseudodata configurations that some of the α_s values yield non-convergent fits. As a result there is a reduced number of points available to perform the fit of the parameters in Eq. 6.2.6. A given curve is kept in the procedure only if the number of α_s values for

which the replica has passed all our fit quality criteria is larger than some threshold that we denote by N_{minpts}. Hence the set of selected pseudodata is

$$\left\{ r \mid \sum_{\alpha} \left(\begin{cases} 1 & \text{if } f^{(r,\alpha)} \text{ converged} \\ 0 & \text{otherwise} \end{cases} \right) \geq N_{\text{minpts}} \right\} . \tag{6.2.10}$$

The number of curves N_{curves} is then the number of the selected replicas, correlated across all α_s values.

The threshold N_{minpts} is chosen to ensure the stability of the distribution of minima. Curves with too few points to obtain a sensible parabolic fit, and thus a value of $\alpha_s^{\min(r)}$ will lead to spurious outliers in the distribution of minima over replicas. However, once we have enough points to reliably fit the parabolas, the variations in the distribution will be the result of the differences in the best fit value for different pseudoreplica samples, and will not depend on the number of curves beyond finite size effects. To assess the influence of outliers, we employ the bootstrapping procedure described in Sec. 6.2.2 above. Specifically, we find the value of N_{minpts} that minimizes the bootstrapping estimate of the finite size effects of the uncertainty, Eq. 6.2.9. The value is expected to decrease when increasing the number of curves, if they were sampled independently from the same distribution. However the parameter N_{minpts} controls the trade-off between outliers allowed in the distribution and size of the sample, and thus the underlying distribution depends on N_{minpts}. In particular, an increase in $\Delta_{\alpha_s}^{\text{stat}}$ when N_{curves} increases by allowing curves with less points, is a clear indication of a contamination of the sample due to poorly fitted parabolas, thus warranting a tighter selection implemented as an increase in N_{minpts}. To account for the fact that a too tight criteria yields a small number of samples that are then affected by large statistical uncertainties, we multiply the $\Delta_{\alpha_s}^{\text{stat}}$ by a penalty factor that depends on the number of points: This is the 99% confidence level factor from a two sided Student-t distribution. In this way we minimize the effect of the outliers in the determination of the central value. Indeed, assuming the distribution of $\{\alpha_s^{\min(r)}\}$ is Gaussian, the difference between the sampled and true central value follows a Student-t distribution with $N_{\text{curves}} - 1$ degrees of freedom, zero mean and scale parameter $\Delta_{\alpha_s}^{\text{stat}}/\sqrt{N_{\text{curves}}}$. A given confidence level around the mean is the proportional to the standard deviation $\Delta_{\alpha_s}^{\text{stat}}$, where the proportionality term is a factor that depends on the number of curves and the desired confidence interval. This coefficient is the quantile function of the standardized Student-t distribution evaluated at $1-(1-\text{CL})/2$, where CL is the desired confidence interval. We choose CL = 0.99. Thus, we minimize:

$$N_{\text{minpts}} = \arg \min_{N_{\text{minpts}}} \Delta_{\alpha_s}^{\text{stat}} T_{0.99, N_{\text{curves}}-1} . \tag{6.2.11}$$

where $T_{0.99,N_{\text{curves}}-1}$ is the percentile of the two-sided confidence factor obtained from a Student-t distribution with $N_{\text{curves}} - 1$ degrees of freedom.

As we discuss Sec. 6.3.5, the precise settings of the selection affect the central value of the determination by an amount that is negligible compared to the PDF uncertainties on α_s. The uncertainties themselves are affected by around 10%.

To address the second problem, namely improving the smoothness of the $\chi^2(\alpha_s)$ profiles, we construct several sets of fits (that we dub *batches*), $f_{(1)}^{(\alpha,r)}$ and $f_{(2)}^{(\alpha,r)}$, which differ by their respective random states $\xi_{(1)}$, and $\xi_{(2)}$, but correspond to the same

pseudodata replica and α_s (each also uses a different t_0 PDF, which is a negligible effect in the final result, as we discuss in detail in Sec. 6.3.5). Both batches are constructed after we have achieved the convergence of the t_0 and preprocessing settings on each value of α_s, as required by our standard fitting procedure (which in turn required other batches of fits). We finally combine the batches by selecting the replica that gives the minimum error function of each pair:

$$f^{(\alpha,r)} = \arg \min_{\{f_{(1)}^{(\alpha,r)}, f_{(2)}^{(\alpha,r)}\}} \left\{ \chi^{2(r,\alpha)} \left[f_{(1)}^{(\alpha,r)} \right], \chi^{2(r,\alpha)} \left[f_{(2)}^{(\alpha,r)} \right] \right\} , \tag{6.2.12}$$

and we impose the further condition that the two replicas $f_{(1)}^{(\alpha,r)}$ and $f_{(2)}^{(\alpha,r)}$ have converged. In this way we mitigate the influence of outliers that narrowly pass the post-selection fit criteria. Note that the approach can be extended to arbitrarily many batches which then will be advantageous to include in the combination, provided the t_0 and preprocessing parameters have converged. When we combine more than two batches we still require to have at least of the replicas for trom the batches hace converged in order to select a point. As we explain in detail in Sec. 6.3.3, we use some additional batches both to prove that they do not significantly affect the result and to refine it nevertheless. A further benefit of this technique is that it minimizes the dependence on the t_0 procedure, as we will explain in Sec. 6.3.5.

6.2.3 Final formulas for the α_s determination

Let us summarize here the above discussion by presenting the final formulas used in our determination of α_s.

The central value and uncertainty are computed substituting the minimum of the profile of $\chi^{2(r)}(\alpha_s)$, Eq. 6.2.6, in Eqs.3.2.1 and 3.2.2 respectively. The profiles are computed by fitting quadratic polynomials to the minimum of the error function Eq. 6.2.5 of sets of replicas fitted at discrete values of α_s, where the pseudodata is kept appropriately correlated. The replicas used in the determination are selected according the criteria explained in Sec. 6.2.2. For each profile, we obtain the minimum α_s directly from Eq.6.2.6. Finally, we determine our central value as

$$\alpha_s = \frac{1}{N_{\text{curves}}} \sum_{r}^{N_{\text{curves}}} \alpha_s^{(r)\,\min} , \tag{6.2.13}$$

and the PDF uncertainty as

$$\Delta_{\alpha_s} = \left(\frac{1}{N_{\text{curves}} - 1} \sum_{r}^{N_{\text{curves}}} (\alpha_s^{(r)\,\min} - \alpha_s)^2 \right)^{\frac{1}{2}} \tag{6.2.14}$$

6.3 The strong coupling constant from NNPDF3.1

In this section we present the main results for the strong coupling constant $\alpha_s\left(M_Z^2\right)$ from the NNPDF3.1 global NNLO analysis. We first discuss the details of the PDF fit settings as well as the $\alpha_s\left(M_Z^2\right)$ fits and the range in which we vary $\alpha_s\left(M_Z^2\right)$. We then present the best-fit result, and estimate the experimental uncertainties by means

of the correlated MC replica method, we perform multiple tests of the validity of the approach, and provide some rough estimate on the Missing Higher Order Uncertainty in the computation.

6.3.1 Fit settings

The present PDF fits are a close variant of the recent NNPDF3.1 global analysis [12]. The data inputs are thus described in Sec. 5.2.2, with only one difference: In the NNPDF3.1 NNLO fit, inclusive jets where treated using NNLO evolution but NLO matrix elements, and including the NLO scale variations as additional source of theoretical uncertainties. The reason for this choice was that the NNLO K-factors of Ref. [199] were not available for all the jet datasets included in the NNPDF3.1 fit. Here, in order to ensure that exact NNLO theory is used for all collider experiments included in the $\alpha_s\left(M_Z^2\right)$ determination, we have kept only those datasets for which the exact NNLO calculation is available, specifically the ATLAS and CMS inclusive jet measurements at 7 TeV based on the 2011 dataset. In the former cases, as motivated in [12], only the central rapidity bin has been kept in the fit. Other jet experiments, in particular the CMS and ATLAS data at $\sqrt{s} = 2.76$ TeV and the CDF measurements, are excluded from the present analysis, since the corresponding NNLO matrix elements are still not available presently.

The theory settings for each fit are identical to the ones described in Ref. [12] with the only difference that the value of α_s changes in the PDF evolution and partonic cross sections. Indeed we store enough data for all the processes to change the value of α_s: DIS predictions are explicitly recomputed using the `APFEL` [39] code, hadronic processes are stored as APPLGIDS [57] up to NLO allowing to compute predictions at different values of α_s. The NNLO corrections are stored as ratios of the NNLO over the NLO result for each bin (which we call C-factors, as mentioned in Sec. 5.4). It is simple to rescale the C-factors knowing the value of $\alpha_s\left(M_Z^2\right)$ at which they were computed, $\alpha_s^{\rm old}$ and the new desired value $\alpha_s^{\rm new}$:

$$C(\alpha_s^{\rm new}) = 1 + \left(C(\alpha_s^{\rm old}) - 1\right)\left(\frac{\alpha_s^{\rm new}}{\alpha_s^{\rm old}}\right)^2 \tag{6.3.1}$$

We use values of $\alpha_s\left(M_Z^2\right)$ between 0.106 and 0.130. We have produced fits in steps of $\Delta_{\alpha_s} = 0.002$ between 0.106 and 0.112 and 0.128 to 0.1130, and in steps of $\Delta_{\alpha_s} = 0.001$ between 0.111 and 0.128.

6.3.2 Results

Here we show our main results, together with the corresponding PDF and finite size uncertainties. We show the global average and the results for experiments corresponding to different physical processes. The main result is the value of $\alpha_s\left(M_Z^2\right)$, determined from the minimum of 3 batches of fits at NNLO, as described in Sec. 6.2.2,

$$\alpha_s^{\rm NNLO}(M_Z) = 0.11845 \pm 0.00052^{\rm pdf}\ (0.4\%) \pm 0.000027^{\rm stat}\ (0.02\%)\,, \tag{6.3.2}$$

where the residual finite size uncertainties are clearly negligible with respect to the PDF uncertainty. The sample is based on 379 curves selected by our procedure. Using

the same methodology we have also obtained a result at NLO based on the minimum of two batches,

$$\alpha_s^{\rm NLO}(M_Z) = 0.12067 \pm 0.00064^{\rm pdf}\ (0.5\%) \pm 0.000061^{\rm stat}\ (0.05\%)\,. \tag{6.3.3}$$

In this case the sample size is 108 curves.

Fig 6.1 illustrates the methodology described in Sec. 6.2: Each curve shows the error function as a function of α_s attained by each of the curves produced by combining three batches of fits at NNLO. The color scale shows the fitted value $\alpha_s^{\min(r)}$, Eq. 6.2.6 from the fit to each the correlated pseudodata replica.

The distribution of the minimum values are shown in Fig. 6.2.

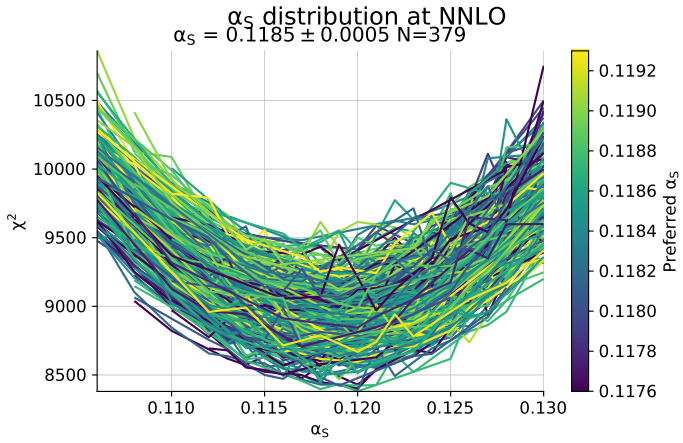


Figure 6.1: Selected curves at NNLO. The lines show the error function Eq. 6.2.5 as a function of α_s for each curve. The color scale shows the minimum α_s value from the parabolic fit to each curve.

6.3.3 Effect of the batch minimization

We have employed for the first time the batch minimization strategy described in Sec 6.2.2. The motivations for using batch minimization here are reducing the dependence of the result from outliers that that barely pass the post selection criteria, and to increase the smoothness of the $\chi^{2(r)}(\alpha_s)$, Eq. 6.2.6. This in turn results in a more precise estimation of the minimum, and correspondingly reduced PDF uncertainties.

In Table 6.1 we have shown the results of the α_s determination at NNLO for each of the individual batches (that is without applying the batch selection), combining two out of three batches using Eq. 6.2.12, and finally combining the three batches (which corresponds to our final NNLO result Eq. 6.3.2). We find that the batch minimization causes a moderate increase of less that half a sigma in the final result $\alpha_s\left(M_Z^2\right)$ (which can be attributed to better control of the outliers), as well as a decrease in the PDF uncertainty of about 20%. We also find that a third batch does not significatively improve the result, since variations are compatible with statistical fluctuations both

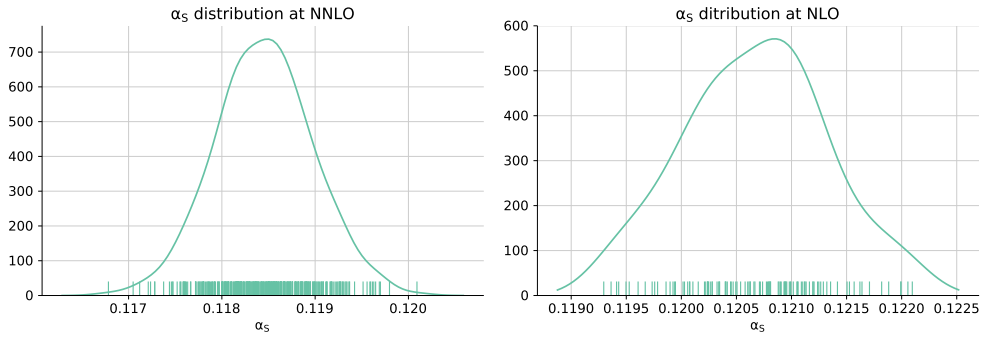


Figure 6.2: Distribution of the minimum of the error function for each selected curve as a function of $\alpha_s\left(M_Z^2\right)$, at NNLO (left) and NLO (right). The markers represent the position of each minimum. The curve is the probability density estimated using Kernel Density Estimate where the bandwidth parameter of the kernel has been computed using the Silverman method.

		NNLO Result
First batch	mean	0.11831
	error	0.00065 (0.55%)
Second batch	mean	0.11828
	error	0.00062 (0.52%)
Third batch	mean	0.11822
	error	0.00072 (0.61%)
First and second	mean	0.11844
	error	0.00054 (0.46%)
First and third	mean	0.11841
	error	0.00058 (0.49%)
Second and third	mean	0.11841
	error	0.00060 (0.51%)
All three	mean	0.11845
	error	0.00052 (0.44%)

Table 6.1: Results from the batch minimization. In the first block we have shown the results we obtain from each of the three indivudual batches, without the batch minimization procedure described in Sec. 6.2.2. In the second block we show the result of combining two of the three batches (in all three possible ways). In the third block, we display our final result, the combination of the three batches. The rows "mean" and "error" show the central value and standard deviation (PDF uncertainty) of the minima of the selected curves.

at the level of central values and PDF uncertainties (see Sec. 6.3.5). This justifies using only two batches for the NLO result Eq. 6.3.3.

6.3.4 Impact of individual datasets and PDF uncertainties

We can try to guess the approximate impact of LHC data upon the determination of α_s, as well as the behaviour of individual processes included in the global fit, which we discuss in the next section. In this way we can gain a very rough qualitative understanding of how the new data has impacted the global best-fits at NLO and NNLO since the previous NNPDF determinations [196, 195]. However, as we will explain in Sec 6.4, a more quantitative understanding of the impact of the different datasets requires more sophisticated analysis.

We can however provide rough estimates on whether a given process prefers a larger or smaller value based on the individual contribution to the error function Eq. 6.2.2 of each individual dataset. Neglecting correlations between different processes, the error function is additive and we may define the partial contribution of the process $\mathcal{P}$ as

$$\chi^2_p\left[\{\theta\},\alpha_s,\mathcal{D},\mathcal{P}\right] = \frac{1}{N_{\mathcal{P}}}\sum_{I,J}^{\mathcal{D}_{\mathcal{P}}}\left(T_I[\{\theta\},\alpha_s]-D_I\right)C^{-1}_{IJ}\left(T_J[\{\theta\},\alpha_s]-D_J\right)\ , \qquad (6.3.4)$$

which is identical to Eq. 6.2.2, with the only difference that we restrict the sum to the set of data points belonging to the process we are interested in, $\mathcal{P}$. It is crucial to emphasize that formula Eq. 6.3.4 depends strongly on the *whole* input dataset $\mathcal{D}$ and therefore in no way gives the α_s determination from that process only. We can however use Eq. 6.3.4 to provide rough qualitative estimates of the *pull* of individual datasets and in particular whether assigning them a bigger weight in the fit would result in smaller or bigger total values of $\alpha_s\left(M_Z^2\right)$. The minima of the partial χ^2 values as a function of α_s are shown in Fig. 6.3, with the experiments grouped by physical process. The uncertainties are computed as standard deviations over the minima from each curve, as explained in 6.2.2. We note that Fig. 6.3 must be interpreted carefully: The minimum values do not correspond neither to the result of the α_s determination with only that process (which would result in much larger uncertainties) nor to the preferred value of α_s taking into account the rest of the data (which would exclude values far from the global minimum value). The minimum values represent merely the points where optimizing the partial error score for the process is most advantageous in order to optimize the global error function Eq. 6.2.2. We will return to this point in Sec 6.4. The number of experimental data points corresponding to each of the physical processes is shown in Table 6.2.

A more insightful way of interpreting the per-process results is though the contribution of the partial χ^2_p per process to the total error function χ^2: In Fig. 6.4 we show the cumulative difference between the χ^2_p of each individual processes and its value computed at the global best fit $\alpha_s\left(M_Z^2\right)$, always using the central PDF from one of the batches, and neglecting the effect of the uncertainties that are correlated between processes. We observe both at NLO and NNLO that the LHC data significantly contributes to constraining α_s. In particular, it is interesting to note that the 13 data points from top pair production data make a large contribution to the total χ^2 outside the best fit region, even though the dataset is composed by almost 4000 points (see Table 6.2). The interpretation of this fact is that there is a small interval of possible values of α_s where the top data is consistent with the rest of the data entering the fit.

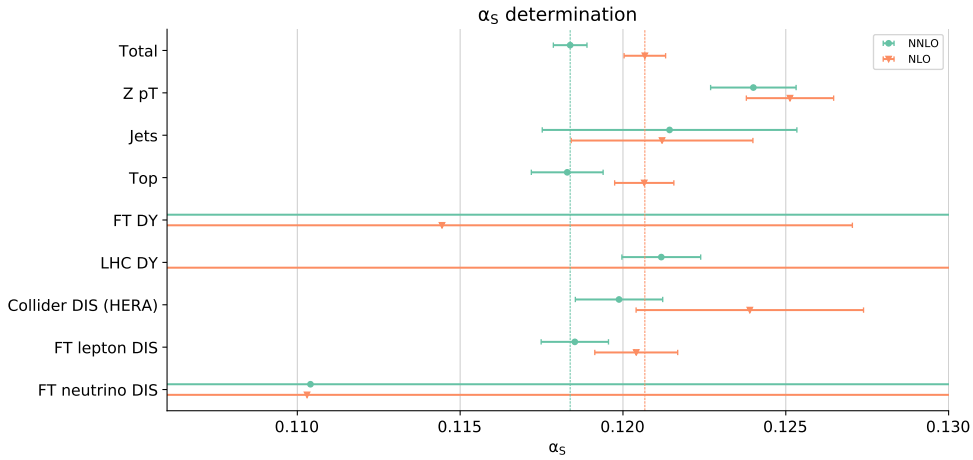


Figure 6.3: minima of the partial χ^2 (see text) at NLO and NNLO for each family of experiments determined with the MC replica method. For the total dataset, this corresponds to the global best fit $\alpha_s\left(M_Z^2\right)$.

	NLO	NNLO
Z pT	120	120
Jets	164	164
Top	13	13
FT DY	189	189
LHC DY	253	273
Collider DIS (HERA)	1221	1211
FT lepton DIS	973	973
FT neutrino DIS	492	492
Total	3950	3865

Table 6.2: Number of points entering the fits at NLO and NNLO, grouped per physical process.

6.3.5 Tests of the methodology

Our determination of α_s results in a remarkably small PDF uncertainty of about 0.4%, which is comparable with the uncertainties of the most precise determinations that enter the PDG average [23]. This clearly indicates that the theory uncertainties that are not presently estimated in our procedure may be significative. We provide rough estimates of the theory uncertainties in Sec. 6.3.6. However, in view of the level of precision we find, we need to analyze in detail the procedural aspects of the correlated replica method described in Sec. 6.2 that could significatively increase the uncertainty in our result. On top of the finite size uncertainties described in Sec. 6.2.2, we have estimated the uncertainties induced by the t_0 method to treat normalization

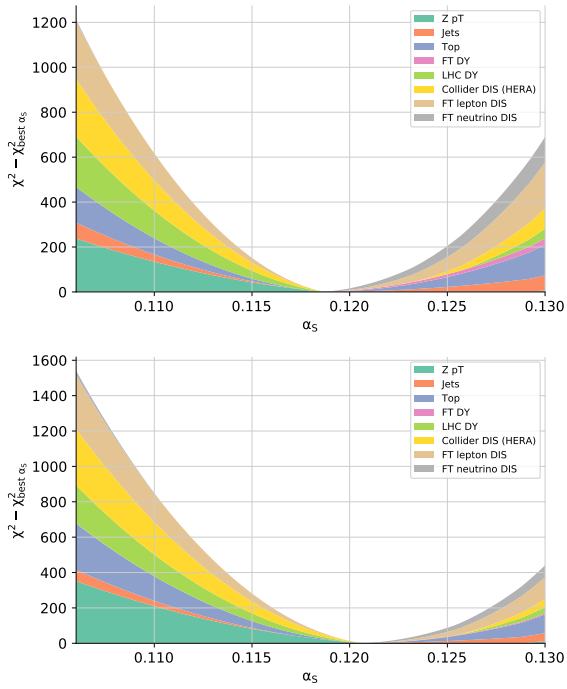


Figure 6.4: Differences between the χ^2_p at $\alpha_s\left(M_Z^2\right)$ and its value at the central best-fit α_s at NNLO (left) and NLO (right). Negative differences are ignored.

uncertainties [114], the fact that we fit quadratic polynomials, as described in Sec 6.2.2, which do not necessarily model the error function $\chi^2(\alpha_s)$ away from the minimum and the criteria used to select the curves, explained in Sec.6.2.2. We find that all these effects induce changes on the final value of $\alpha_s\left(M_Z^2\right)$ that are much smaller than our estimate of the PDF uncertainty on $\alpha_s\left(M_Z^2\right)$ and therefore do not change our final determination.

Effect of the curve selection settings

The selection algorithm in Sect. 6.2.2 contains some arbitrariness: Specifically in the range of fitted values of $\alpha_s\left(M_Z^2\right)$ and the in the criteria to discard a curve with too few points. We now vary these settings within reasonable ranges to probe their effect. We use as input the combined three batches of fits at NNLO from where we obtained our main result, Eq. 6.3.2. To probe the stability of the curve selection, instead of the criterion in Eq 6.2.11, we simply fix the minimum required number of points that must have converged in each curve, N_{minpts} out of the 21 values of α_s we fit, and repeat the determination of α_s for different fixed values of N_{minpts}. The results are presented in Table 6.3. The results show that the dependence of the central result with the number of selected curves is always much smaller than the PDF uncertainty, while the PDF uncertainty itself changes by around 15% in reasonable ranges where we select

a sufficient number of curves. These variations are an estimate of the magnitude of the uncertainty on the PDF uncertainty itself. The PDF uncertainty decreases significatively when we tighten the selection (e.g. by requiring that each curve contains 18 out of 21 points), but at the cost of match reduced statistics. We conclude that our results is stable upon changes in the arbitrary choices in the selection.

		NNLO result
$N_{\text{minpts}} \geq 18$	mean	0.11842
	error	0.00031 (0.26%)
	n	12
$N_{\text{minpts}} \geq 15$	mean	0.11844
	error	0.00044 (0.37%)
	n	92
$N_{\text{minpts}} \geq 6$	mean	0.11845
	error	0.00052 (0.44%)
	n	379
All selected	mean	0.11844
	error	0.00056 (0.47%)
	n	400

Table 6.3: Variation of the results with the minimum number of converged PDF replicas required to select a curve, out of the 21 that were attempted, one for each value of α_s. The rows "mean" and "error" show the central value and standard deviation (PDF uncertainty) of the minima of the selected curves. The third row "$N_{\text{minpts}} \geq 6$" is the final selection Eq. 6.3.2. The row n shows the total number of selected curves entering the determination.

We also check the dependence on the values of α_s that are farthest from the best fit value. Our best fit should be independent of the PDF fits for α_s values that are far from it, as long as the parabolic behaviour of $\chi^2(\alpha_s)$ can be resolved above the statistical fluctuations. To test that this is in fact the case, we repeat the α_s determination removing the several α_s value that are farthest from the best fit from consideration. That is, instead of fitting α_s from all the 21 different values in the range $\alpha_s\left(M_Z^2\right) \in [0.106, 0.130]$, we select a smaller range by trimming the values where the absolute difference with our central result Eq 6.3.2 $|\alpha_s\left(M_Z^2\right) - \alpha_s^{\text{NNLO}}(M_Z)|$ is greatest. We present the results in Table 6.4. We confirm that the central values are consistent within uncertainties even if we only fit half of the points. The resulting PDF uncertainties are also within ten percent. The uncertainties only start growing we trim the 15 most distant values of α_s and fit the 6 central points only. In this case, the selection criterion Eq. 6.2.11 (which is applied after trimming the distant values in α_s) only selects 10 curves.

Effect of the t_0 procedure

The t_0 procedure (see Sec 5.1.2) introduced in Ref [114] implements an unbiased estimate of the best fit PDF when the input experimental data contains normalization

		Total
Default	mean	0.11845
	error	0.00052 (0.44%)
	n	379
Trim one	mean	0.11847
	error	0.00049 (0.41%)
	n	290
Trim two	mean	0.11846
	error	0.00045 (0.38%)
	n	218
Trim 5 farthest	mean	0.11852
	error	0.00051 (0.43%)
	n	290
Trim 10 farthest	mean	0.11869
	error	0.00046 (0.39%)
	n	32
Trim 15 farthest	mean	0.11822
	error	0.00079 (0.67%)
	n	10

Table 6.4: Variation of the result at NNLO as we discard the values of α_s in the fitted range that are furthest from the best fit. The rows "mean", "error" and n have the same meaning as in Table 6.3.

uncertainties. The procedure makes the covariance matrix entering Eq. 6.2.2 dependent on the so called t_0 *PDF set*, which is simply the central PDF from a previous fit. All *multiplicative* uncertainties, i.e. proportional to the central value of the measured data, are made instead made proportional to the theory predictions using the t_0. The fits are iterated fit is statistically equivalent to the previous iteration (that is, the differences are much smaller than the PDF uncertainties).

We first demonstrate that the t_0 procedure produces a both necessary and large correction on the α_s determination. In Fig. 6.5 we have used the PDFs from one of the NNLO batches (which were fitted with the t_0 methods which is standard in our fits) and minimized the error function using instead the unmodified experimental covariance matrix. That is, we have replaced the experimental covariance matrix (where the normalizations) with the t_0 covariance matrix in Eq. 6.2.4 but not in Eq. 5.1.26. We see that all minima are shifted towards smaller values of $\alpha_s\left(M_Z^2\right)$ and the fit we obtain, $\alpha_s\left(M_Z^2\right)^{(\exp\chi^2)} = 0.114 \pm 0.001$ is very far from the result obtained when the normalization uncertainties are treated correctly Eq.6.3.2 in units of the PDF uncertainties. While this is inconsistent since we are optimizing α_s for a different quantity than the PDF fit (redoing the excise optimizing the uncorrected definition of χ^2 has a prohibitive computational cost), it illustrates that a correct treatment of the normalization uncertainties is necessary to achieve a correct determination.

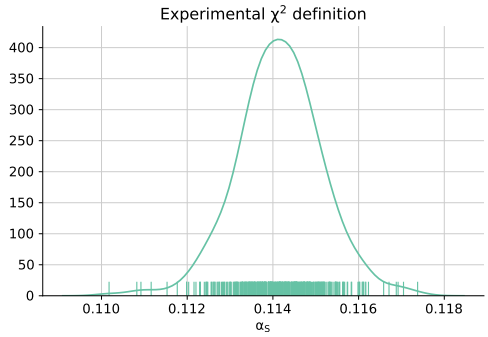


Figure 6.5: Like 6.2 but for one batch of NNLO fits, using the experimental covariance matrix in in the minimization of α_s instead of the t_0 one.

The large effect of the t_0 procedure on the resulting value of α_s suggests that there may also be a dependence on the precise choice of the t_0 PDF that is used in the fits: Even though the fits are iterated until the changes in the t_0 PDF are small compared to the PDF uncertainties, this doesn't necessarily ensure that the finite precision with which we determine the t_0 PDF (i.e. the precision on the central PDF) does not induce a measurable systematic change on α_s. Therefore we need test to test explicitly that the best fit value of α_s, which depends on the t_0 PDF trough the covariance matrix in Eq. 6.2.2 is also not significatively affected. For example changes in the less well constrained jet normalizations might change the global preferred value of α_s though their pull on the gluon PDF.

We have performed the following test: Using the three batches of NNLO fits $f_{(1)}^{(\alpha,r)}$, $f_{(2)}^{(\alpha,r)}$ and $f_{(3)}^{(\alpha,r)}$ we have performed three times combination procedure described in Sec.6.2.2, where each time we have used a different t_0 set when finding the best α_s in Eq. 6.2.4, namely the central value of each of the three batches for each value of α_s. Note that using the same t_0 set for all batches is in fact more consistent that using a different one since the values of the error function for each point can be compared more meaningfully when selecting points in Eq. 6.2.12. We present the results in Table 6.5. Since we observe no changes in the results, we may safely conclude that the t_0 procedure does not add significative procedural uncertainties to our result, while however it is crucial to apply it to correct for normalization uncertainties.

	mean	error	n
First t_0	0.11844	0.00052 (0.44%)	379
Second t_0	0.11845	0.00052 (0.44%)	379
Third t_0	0.11841	0.00051 (0.43%)	356

Table 6.5: Results from the combination of three batches of fits at NNLO using different t_0 variations. The mean and error are the results of the determination, as explained in Sec. 6.2. n is the number of curves selected following Sec 6.2.2.

Effect of the parabolic fit

Our procedure is based on obtaining smooth curves of the dependence of the error function on α_s for each for each of the fitted replicas. These curves have a minimum in the fitted range, and consequently we can expand in Taylor polynomials of second degree around the minimum. This is likely a good approximation as shown in Fig. 6.1. We take advantage of the smoothness of the profiles by reducing the number of α_s variations we need to produce (and that are costly in the NNPDF framework). Indeed if we follow the procedure as described, but simply take the sample minimum over α_s of the error function samples Eq. 6.2.5 instead of producing the quadratic fit Eq. 6.2.6.

In Sec. 6.3.5 we already demonstrated that our result does not depend strongly on the fitted range.

Here we demonstrate quantitatively that the quadratic approximation does not introduce procedural uncertainties that are of importance in comparison to the PDF uncertainties. We first perform a simple qualitative test: Any transformation $\chi^2(\alpha_s) \to \chi^2(f(\alpha_s))$ where f is sufficiently smooth and monotonic in the relevant range should also have the same minimum when replacing $\alpha_s \to f(\alpha_s)$ in Eq. 6.2.4. We consequently expect to also get the same minimum value $\alpha_s^{\min(r)}$ when making the replacement in the parabolic fit Eq. 6.2.6. We have made this test for $f = \exp$ and $f = \log$, and show the results in Table 6.6. In both cases the results are within the PDF uncertainties. This is especially notable in the case of log since its Taylor expansion has a small convergence radius at $\log(\alpha_s) \sim \log(0.1)$.

		Total
log	mean	0.11813
	error	0.00053 (0.45%)
	n	379
exp	mean	0.11849
	error	0.00051 (0.43%)
	n	345
default	mean	0.11845
	error	0.00052 (0.44%)
	n	379

Table 6.6: α_s determination with transformed input to the error function. For the first row we fit $\chi^2(\exp(\alpha_s))$, for the second $\chi^2(\log(\alpha_s))$. In the third our standard determination is presented for comparison. The rows "mean", "error" and n have the same meaning as in Table 6.3.

We also check that the quadratic approximation is in fact sufficient and no higher order terms are needed to describe the error function $\chi^2(\alpha_s)$ in the fitted range. In particular, we wish confirm that fitting a cubic polynomial does not improve the model. For this test, we employ Akaike Information Criterion (AIC)[200]. The AIC provides an estimate of the expected relative distance between the fitted model and the unknown true mechanism [201]. The AIC score balances goodness of fit against simplicity of the model. A lower score corresponds to a lower expected distance (more

precisely, lower KullbackLeibler divergence). We may use it to test whether it is advantageous to include extra orders in the Taylor expansion by comparing the AIC score from our fit with quadratic polynomials Eq. 6.2.6 to the score obtain by fitting cubic polynomials instead. The AIC score is given by

$$\mathrm{AIC} = 2k - 2\log L + \frac{2k(k+1)}{n-k-1} \tag{6.3.5}$$

where k is the number of degrees of freedom, n is the number of fitted points and $\log(L)$ is the log-likelihood associated to the model. For a least squares fit, it is given, up to constant terms, by the sum in quadrature of the residues:

$$\log(L)^{(r)} = \sum_{\alpha}^{n} (\chi^{2(r)}(\alpha_s) - \chi^{2(r,\alpha)})^2 + \text{const.} \tag{6.3.6}$$

where we have used the notation from Eqs. 6.2.5 and 6.2.6: We fit by least squares a polynomial model $\chi^{2(r)}(\alpha_s)$ to the values of the error function obtained from the PDF fit $\chi^{2(r,\alpha)}$ for each curve indexed by r. We have compared in Table 6.7 the mean and standard deviations over curves. Since the AIC score is comparable and lower for , we conclude that there is no evidence that a more complex model is required.

	AIC
Quadratic polynomial	169 ± 37
Cubic polynomial	173 ± 35

Table 6.7: AIC score comparing a quadratic and cubic fit to $\chi^2(\alpha_s)$. The means and standard deviations are taken over curves.

6.3.6 Estimation of theoretical uncertainties

The theory uncertainties on our α_s determination include Missing Higher Order Uncertainties (MHOU), electroweak effects, higher-twist, nuclear corrections and missing mass corrections, amongst others. Here we only estimate a bound on the MHOU, and leave a more complete analysis of the issue to further studies.

We may obtain a bound on the MHOU by comparing the results at NLO and NNLO. Our results differ by

$$\Delta\alpha_s^{\mathrm{pert}} \equiv |\alpha_s^{\mathrm{NNLO}} - \alpha_s^{\mathrm{NLO}}| = 0.0022\,. \tag{6.3.7}$$

By taking half the difference between the NLO and NNLO results, we obtain a conservative estimate of the NNLO uncertainty of

$$\Delta^{\mathrm{NNLO}} = 0.0011\,. \tag{6.3.8}$$

This estimate is significatively lager than the PDF uncertainty. Therefore it indicates that the theoretical uncertainties are at the very least comparable in size to the PDF uncertainties we can estimate reliably as demonstrated in the previous section.

We conclude that we have reached the limit to the current framework for extracting α_s from hadronic data, and that further progress must necessarily account for the Missing Higher Orders quantitatively. In Sec 6.5 we speculatively outline a possible method to do so.

6.4 α_s determination from a partial dataset

We now return our attention to the results from the partial χ^2, Fig 6.3 and interpret more carefully the results from the *partial* χ^2 fits. While the following discussion applies to any fit of a theoretical parameter from hadronic data, we restrict ourselves to determinations of α_s for concreteness.

Two *categories* of determinations based on hadronic measurements enter the World Average: Those based on PDFs [202, 203, 204, 205], which are essentially obtained by optimizing Eq 6.2.2 as a function of α_s, and the $t\bar{t}$ production, currently including only the CMS measurement at 7 TeV [193], which is instead based on minimizing over a χ^2 function that considers explicitly the $t\bar{t}$ data only, Eq 6.3.4 (we call this the *partial* χ^2 *method*). We shall discuss the relation between these categories, and also try to elucidate the noticeable fact that determinations of α_s based on an hadronic dataset, such as Ref [193] as well as more recent ones like Ref [189], give significatively different results from the determinations based on the PDFs that they use as input to compute the predictions in Eq. 2.8.1.

6.4.1 The *Partial* χ^2 *method*

Several recent determinations of α_s based on hadronic data [193, 192, 191, 190, 189, 188] implement the following procedure, which we shall dub *Partial* χ^2:

1. Consider some experimental measurement of hadronic data, $\mathcal{P}$. For example, $t\bar{t}$ production [193, 190], Prompt photon events [192] jet production [191, 189], and Z+jet production [188].

2. Compute theory predictions at discrete values of α_s, following Eq. 2.8.1 and suitably interpolating the results from PDF sets fitted with different values of α_s (i.e. where $\alpha_s(M_Z)$ is a fixed parameter in Eq. 6.2.2).

3. Construct a profile $\chi^2_{\mathcal{P}}(\alpha_s)$ characterizing the agreement between data and theory, Eq 6.3.4.

We point out that the recommendation [4] for estimating α_s uncertainties on the PDFs, of estimating the final result with an upper and a lower PDF variation of $\alpha_s(M_Z)$ does not apply when fitting α_s itself. In this case the value of α_s should be kept matched with the rest of the calculation. Note that this does not imply that theory parameters cannot be fixed in PDF fits by default: For example the value of α_s itself is fixed in the PDF4LHC recommendation [4] to a value consistent with the PDG average [23] on the grounds that it takes into account more information than that provided by hadronic data; we may trade some internal consistency of the input $\mathcal{D}$ within the PDF fitting framework with potentially more reliable external constrains on the theory parameters. On the other hand, theoretical parameters that are to be

fitted do certainly have to be varied consistently in the PDFs. This is a required condition, but, as we argue next, not sufficient.

We now discuss the relation between the partial χ^2 method we just described and the dataset used to fit the PDF by optimizing the *global* χ^2, Eq. 6.2.2. In particular it is pertinent to examine why does the partial χ^2 appear to constrain α_s in all the examples above. That is, why is the value of $\chi^2_p[\alpha_S, \mathcal{P}]$ different at different values of α_s?

6.4.2 Simultaneous PDF and α_s determination from a partial dataset

We note that if the only data used to fit the PDFs was any of the partial datasets above (such as e.g. $t\bar{t}$ production), so that $\mathcal{D} = \mathcal{P}$ then we would certainly not have enough constraints to determine the PDFs and α_s simultaneously: In fact, we would be able to obtain an adequate fit, characterized by $\chi^2/(N_{\mathcal{D}} - 1) \approx 1$ for any reasonable value of α_s. We would however have big PDF uncertainties, associated to the kinematic regions that are not constrained by $\mathcal{P}$. For example if we fitted PDFs to $t\bar{t}$ production data only, we could obtain a good fit at a higher value of $\alpha_s(M_Z)$ by compensating it with a reduced gluon momentum fraction large x as we will show next in a more general situation. Therefore for $\mathcal{D} = \mathcal{P}$, the partial χ^2 in Eq. 6.3.4 is flat and does not allow to determine α_s (in this case, $\chi^2_{\mathcal{P}}$ is also the global χ^2, Eq. 6.2.2).

It follows that for these relatively small datasets, the $\chi^2_p[\alpha_S, \mathcal{P}]$ profile fundamentally measures the disagreement between the partial data set $\mathcal{P}$ and the dataset included in the PDF fit, $\mathcal{D}$, as a function of α_s.

6.4.3 Inconsistency of the partial χ^2 method

The partial χ^2 method neglects the fact that the dataset used in the PDF fits, $\mathcal{D}$, constrains α_s itself, i.e. that the minimum of Eq. 6.2.2 adopts significantly different values for different values of α_s. That is, given the measurement $\mathcal{P}$, if one makes enough assumptions on the input data of the PDFs to be able to extract $\alpha_s(M_Z)$ with competitive uncertainties, then the prior over α_s is not uniform. One cannot simply disregard the constrains from $\mathcal{D}$ on the theory parameters $\{\alpha\}$ while utilizing them for the PDF parameters $\{\theta\}$. In particular, this can lead to evident inconsistencies such as the value selected by the partial χ^2 method being excluded by the PDF on which the theoretical prediction Eq 2.8.1 is based. This is then a logical contradiction, because the result, which, as we have shown in Sec. 6.4.2, is based on the agreement with $\mathcal{D}$, is grounded on a prior that is internally inconsistent to begin with. Moreover, the best fit PDFs away from the global minimum in $(\{\alpha\}, \{\theta\})$ are subject to a large degree of arbitrariness: In an ideal PDF fit where all theory and data are correct, every dataset has a χ^2 per degree of freedom, $\chi^2/d.o.f \approx 1$. However, when not all constraints can be satisfied simultaneously (e.g. because the *wrong* value of α_s has been given as input) the result of the fit depends on the number of points belonging to each particular dataset: The smaller a dataset is (in comparison to others which cannot be fitted simultaneously), the less advantageous is it for the global figure of merit Eq. 6.2.2 to bend the PDF in order to accommodate it. This is clear in the case of the $t\bar{t}$ data in the NNPDF 3.1 [12] fits. The default dataset includes a total of 26 $t\bar{t}$ production

data points corresponding to the ATLAS [149, 150, 151] and CMS [161, 162, 163] measurements of the total cross sections and differential distributions, computed at NNLO [90, 91](see Ref [12] for details). The $t\bar{t}$ data has a large sensitivity to α_s but a low statistical weight in the fit (26 points to be compared to 3979 in total). Therefore its description (i.e. the partial χ^2, Eq. 6.3.4) deteriorates rapidly as we move move α_s away from the best fit value. However, we can modify the assumptions on $\mathcal{D}$ insisting that the $t\bar{t}$ data is described at any value of α_s. For example we set $\alpha_s(M_Z) = 0.121$ where the top data is not so well described in a default NNPDF fit that optimizes Eq 6.2.2 on a large dataset (we have $\chi^2_{t\bar{t}}/d.o.f. = 1.42$) and increase the statistical weight of the top data by fitting 15 identical copies of it. The effect of the reweighting is to greatly improve the description of $t\bar{t}$ (the partial χ^2 becomes $\chi^2_{t\bar{t}}/d.o.f. = 1.02$) while slightly deteriorating the global χ^2. The most significative change between the default fit and that with increased weight happens in the gluon PDF, which is nevertheless compatible within PDF uncertainties, as we show in Fig. 6.1. Indeed, because of the high degeneracy in the space of PDF parameters, $\{\theta\}$, important variations in the input assumptions (that e.g. change drastically the partial χ^2) can be reabsorbed into relatively small changes in the PDFs (both in terms of deterioration of the global χ^2 and distances in PDF space). In this way we have demonstrated that the partial χ^2 does not measure significant physical properties of the hard cross section, but rather properties of the PDF minimization.

In summary, we propose that the most statistically rigorous way to produce an α_s determination from the measurement $\mathcal{P}$ is to include it in a PDF fit and determine simultaneously α_s and the PDFs based on the global χ^2 that now includes $\mathcal{P}$ as well as the rest of the data $\mathcal{D}$. Therefore if $\mathcal{P}$ was already included in the $\mathcal{D}$ the result from optimizing the global χ^2 profile Eq 6.2.2 would be unchanged. Since there is no way to disentangle the $\mathcal{D}$ dependence from Eq. 2.8.1, this method is no more PDF dependent than the partial χ^2 minimization, but it solves the shortcomings that we have described. The correction on the value of α_s when $\mathcal{P}$ is included will either be a small or point to a flaw in the theory, experiment description or fitting methodology. An important advantage is that the process will then be treated using the full fledged PDF fitting machinery (as opposed to a naive minimization of Eq. 6.2.2). In particular, this takes care of implementing the correct treatment of the normalization uncertainties, which has been observed to make a significant difference in an α_s determination (see Sec. 6.3.5). We conclude that it is questionable to consider hadronic results as independent constrains on α_S in World averages, rather than as corrections to the results from the prior PDFs.

6.5 Preferred values

While we have concluded that the quantities suitable for inclusion in global averages are those based on minimizing the global χ^2 profile, Eq. 6.2.2, it is nevertheless interesting to define a *preferred* α_s value from a given dataset $\mathcal{P}$. Some possible usages include the assessment of the constraints provided by the measurement, and possibly the study of the higher order corrections (e.g. one could take the dispersion over ensemble of preferred values of a suitable set of processes as an estimate of Missing Higher Order Uncertainty). We first list some desirable properties that such definition should have.

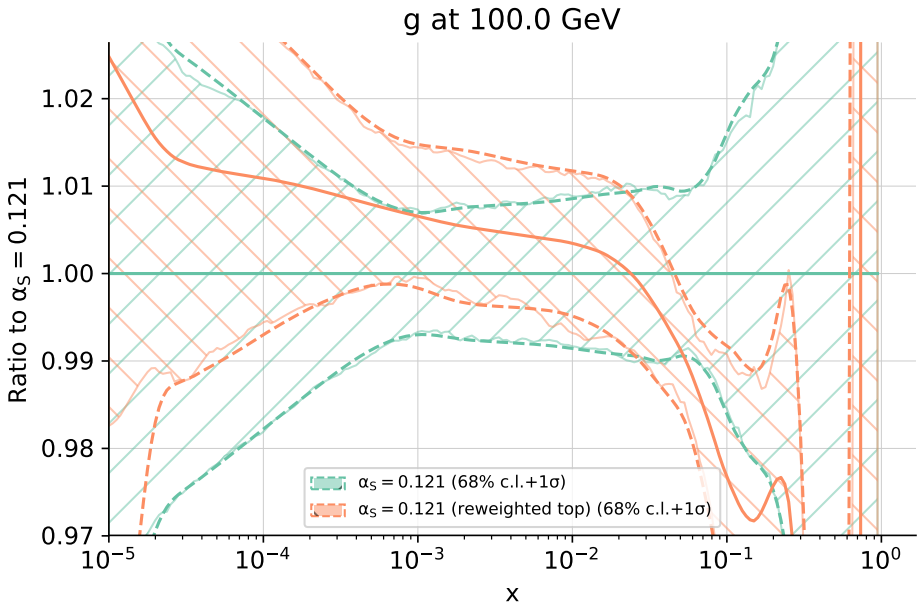


Figure 6.1: Comparison of gluon PDF between an NNLO-like global fit at NNLO where we have set $\alpha_s(M_Z) = 0.121$ and a fit with the only difference that the weight with which the $t\bar{t}$ production data enters the fit has been multiplied by 15. The reweighting causes noticeable decrease in the gluon at large x (but yet roughly within uncertaities) to accommodate the $t\bar{t}$ data which is, which is then described optimally, with $\chi^2_{t\bar{t}}/d.o.f. = 1.02$, to be compared to $\chi^2_{t\bar{t}}/d.o.f. = 1.42$ before the reweighing. The improvement of the description of the $t\bar{t}$ data comes at the cost of a deterioration in the global χ^2 ($\chi^2/d.o.f = 1.215$ before the reweighing and $\chi^2/d.o.f = 1.229$ afterwards).

- Independent on the relation between the number of points in the dataset of interest, $N_{\mathcal{P}}$ and those in the global dataset, $N_{\mathcal{D}}$. Clearly, if we are interested in intrinsic physical properties, the number of points in the dataset should not change the result.

- Explicitly depend on the global dataset used in the PDF fit $\mathcal{D}$. Since, as discussed in the previous section, in general we cannot get rid of the dependence on $\mathcal{D}$, it needs to be clearly acknowledged.

- Converge to the determination from $\mathcal{P}$ alone, in the sense described in Sec. 6.4.2 when it determines α_s by itself. While this definition is likely more interesting for smaller, experimentally cleaner, datasets, this is a logical asymptotic property.

The partial χ^2 method discussed in Sec. 6.4.1 has none of these properties and therefore it is not a particularly good definition of preferred value (it may however approximate the third property reasonably well in practice). On the other hand, the exercise illustrated in Fig 6.1 points at a definition that satisfies them:

Preferred value of α_s for the data $\mathcal{P}$ The value of α_s that corresponds to the minimum of the global χ^2 over values of α_s and PDF parameters $\{\theta\}$, when the PDF parameters are restricted to result in a *good fit* for $\mathcal{P}$ within its experimental uncertainties, for all values of α_s.

The value is preferred in the sense that the constraints from $\mathcal{P}$ take precedence over those from $\mathcal{D}$, in particular regardless of the number of points, thereby satisfying the first requirement. Once the constrains from $\mathcal{P}$ are enforced, a global χ^2 which includes $\mathcal{D}$ is minimized, thus satisfying the second condition.

The main difficulty is to algorithmically specify what a *good fit* means: Intuitively if the dataset if self consistent at a given value of α_s, then we require that $\chi^2_{\mathcal{P}}/d.o.f \approx 1$. If this is the case at every relevant value of α_s then the partial $\chi^2_{\mathcal{P}}$ of this reweighed fit is flat and α_s is determined based on the agreement with $\mathcal{D}$ (but based on PDFs that have been modified to accommodate $\mathcal{P}$ at all values of α_s). If $\mathcal{P}$ determines α_s by itself (in the sense of Sec 6.4.2) then the partial χ^2 will not be flat and will be used to obtain α_s. A suitable interpolating procedure between these two situations could be obtained in the NNPDF framework by minimizing as a function of $\{\theta\}$ and α_s

$$\text{ERF} = \chi^2\left[\{\theta\}, \alpha_s, \mathcal{D}\right] + w\chi^2\left[\{\theta\}, \alpha_s, \mathcal{P}\right] \ , \tag{6.5.1}$$

where w is a large number. Because of the cross validation based regularization, the effect of w will saturate either when we reach $\chi^2_{\mathcal{P}}/d.o.f \approx 1$, so that only the first therm varies as a function of α_s, or else, if $\mathcal{P}$ determines α_s, the curvature of profile will exclusively depend on the second term.

It remains to be studied whether these preferred values can be computed in practice.

Chapter 7

Conclusions and outlook

In this thesis we have presented several results that contribute to the improved understanding of the structure of the proton and related phenomenology. In Sec. 3.3 have presented MCH, a Monte Carlo to Hessian method that extends the range of applicability of the NNPDF global analyses to situations where Hessian errors are required. It was also used to produce some of the PDF4LHC combined sets discussed in Chapter 4. We have also presented a method, SM-PDF, that significatively improves the computational efficiency of PDF error estimates. These developments open some directions. MCH could be used to elucidate the relation between the Monte Carlo uncertainties and dynamic tolerances (see Sec. 3.2.2) used by Hessian determinations of PDFs. This in turn could prove useful to develop a new generation of closure tests (see Sec 5.1.9) and related validation strategies. Methods similar to SM-PDF could be used to obtain compact representations of cross section perturbative coefficients that can be subsequentially combined with arbitrary PDFs.

The PDF4LHC15 sets represented a baseline for the state of art in PDF determinations in 2015. This standard has been suppressed by the NNPDF3.1 analysis presented in Chapter 5, which is currently unparalleled in the extent of the included dataset, the sophistication of the theoretical treatment, and the level of validation of the fitting methodology.

NNPDF3.1 has been used as a foundation of a precise determination of a strong coupling constant, described in Chapter 6, which has also served to understand the influence of the PDFs in the extraction of theory parameters.

Both the NNPDF3.1 analysis and the α_s determination probe the limits of the current NNPDF framework and evidence the necessity of extensive new developments, both within the collaboration and elsewhere. For example, the issues described in Sec. 5.4 required comprehensive and time consuming tests in order to be understood and resolved. This highlights the need to develop better diagnostics in together with the colleagues that provide us with the experimental and theoretical inputs.

The improvements in the PDF analysis have led to reductions in PDF uncertainties that paradoxically has put into question the usefulness of the concept of *PDF uncertainty* as it is currently utilized: There is increasing evidence (see e.g. sec 6.3.6) that the theoretical uncertainties, which are currently not included in PDF determinations, may in fact be dominant. Future generations of PDF fits should certainly include quantitative estimates of theory uncertainties. The procedure to attain such

estimates is not clear at the moment, however. The capability of performing scale variations at NLO was recently implemented in the NNPDF code. Now a significant amount of testing is required to understand the optimal range of variation of the scales as well as the correlation model. The extension to NNLO is not easily achievable with the current technology. We have outlined a possible alternative approach in Sec. 6.5.

The fitting methodology may seem less important when theory uncertainties predominate. This is however not the case. The ability to cleanly map a variation in the input parameters, be it of scales or dataset weights, onto a variation of the PDF that is not affected by defects in the minimization procedure will be crucial to be in position to formulate a recipe to estimate the theory uncertainties. Currently we have some evidence (see e.g. Sec 6.3.3) that a markedly better minimization might be achievable. Additionally a more robust minimization may allow as to dispose of some unpalatable aspects of our methodology such as preprocessing iterations (see Sec. 5.1.1), the difficulty to impose positivity constraints (see Sec. 5.1.3), or the large amount of replicas that are discarded (see Sec. 5.1.8 and a practical consequence in Sec. 6.2.2).

In conclusion, reaching a percent level precision in the determination of PDFs has bred a new set of challenges.

Bibliography

[1] **ATLAS** Collaboration, G. Aad et al., *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett.* **B716** (2012) 1–29, [arXiv:1207.7214].

[2] **CMS** Collaboration, S. Chatrchyan et al., *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Phys. Lett.* **B716** (2012) 30–61, [arXiv:1207.7235].

[3] E. Eichten, I. Hinchliffe, K. D. Lane, and C. Quigg, *Super Collider Physics*, *Rev. Mod. Phys.* **56** (1984) 579–707.

[4] J. Butterworth et al., *PDF4LHC recommendations for LHC Run II*, *J. Phys.* **G43** (2016) 023001, [arXiv:1510.03865].

[5] M. Glück, E. Hoffmann, and E. Reya, *Scaling violations and the gluon distribution of the nucleon*, *Zeitschrift für Physik C Particles and Fields* **13** (1982), no. 2 119–130.

[6] D. Duke and J. Owens, *Qš-dependent parametrizations of parton distribution functions*, *Physical Review D* **30** (1984), no. 1 49.

[7] NNPDF Collaboration, "NPDF website." http://nnpdf.mi.infn.it/, 2017.

[8] **NNPDF** Collaboration, R. D. Ball et al., *Parton distributions for the LHC Run II*, *JHEP* **04** (2015) 040, [arXiv:1410.8849].

[9] V. Bertone, S. Carrazza, and N. P. Hartland, *APFELgrid: a high performance tool for parton density determinations*, *Comput. Phys. Commun.* **212** (2017) 205–209, [arXiv:1605.02070].

[10] S. Forte, E. Laenen, P. Nason, and J. Rojo, *Heavy quarks in deep-inelastic scattering*, *Nucl. Phys.* **B834** (2010) 116–162, [arXiv:1001.2312].

[11] **NNPDF** Collaboration, R. D. Ball, V. Bertone, M. Bonvini, S. Carrazza, S. Forte, A. Guffanti, N. P. Hartland, J. Rojo, and L. Rottoli, *A Determination of the Charm Content of the Proton*, *Eur. Phys. J.* **C76** (2016), no. 11 647, [arXiv:1605.06515].

[12] **NNPDF** Collaboration, R. D. Ball et al., *Parton distributions from high-precision collider data*, *Eur. Phys. J.* **C77** (2017), no. 10 663, [arXiv:1706.00428].

[13] S. Carrazza, S. Forte, Z. Kassabov, J. I. Latorre, and J. Rojo, *An Unbiased Hessian Representation for Monte Carlo PDFs*, *Eur. Phys. J.* **C75** (2015), no. 8 369, [arXiv:1505.06736].

[14] S. Carrazza, S. Forte, Z. Kassabov, and J. Rojo, *Specialized minimal PDFs for optimized LHC calculations*, *Eur. Phys. J.* **C76** (2016), no. 4 205, [arXiv:1602.00005].

[15] J. R. Andersen et al., *Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report*, in *9th Les Houches Workshop on Physics at TeV Colliders (PhysTeV 2015) Les Houches, France, June 1-19, 2015*, 2016. arXiv:1605.04692.

[16] **LHC Higgs Cross Section Working Group** Collaboration, D. de Florian et al., *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, arXiv:1610.07922.

[17] M. Gell-Mann, *Symmetries of baryons and mesons*, *Phys. Rev.* **125** (Feb, 1962) 1067–1084.

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

[19] Y. Dothan, M. Gell-Mann, and Y. Ne'eman, *Series of hadron energy levels as representations of non-compact groups*, *Physics Letters* **17** (1965), no. 2 148–151.

[20] G. Zweig, *An SU(3) model for strong interaction symmetry and its breaking. Version 2*, in *Developments in the quark theory of hadrons. Vol. 1. 1964 - 1978* (D. Lichtenberg and S. P. Rosen, eds.), pp. 22–101. 1964.

[21] D. J. Gross and F. Wilczek, *Ultraviolet behavior of non-abelian gauge theories*, *Physical Review Letters* **30** (1973), no. 26 1343.

[22] H. D. Politzer, *Reliable perturbative results for strong interactions?*, *Physical Review Letters* **30** (1973), no. 26 1346.

[23] **Particle Data Group** Collaboration, C. Patrignani et al., *Review of Particle Physics*, *Chin. Phys.* **C40** (2016), no. 10 100001.

[24] A. V. Manohar, *An Introduction to spin dependent deep inelastic scattering*, in *Lake Louise Winter Institute: Symmetry and Spin in the Standard Model Lake Louise, Alberta, Canada, February 23-29, 1992*, pp. 1–46, 1992. hep-ph/9204208.

[25] J. D. Bjorken, *Asymptotic sum rules at infinite momentum*, *Physical Review* **179** (1969), no. 5 1547.

[26] C. G. Callan Jr and D. J. Gross, *High-energy electroproduction and the constitution of the electric current*, *Physical Review Letters* **22** (1969), no. 4 156.

[27] N. Hartland, *Proton structure at the LHC.* PhD thesis, Edinburgh U., 2014. arXiv:1411.0259.

[28] R. P. Feynman, *Very high-energy collisions of hadrons*, *Physical Review Letters* **23** (1969), no. 24 1415.

[29] T. Kinoshita, *Mass singularities of feynman amplitudes*, *Journal of Mathematical Physics* **3** (1962), no. 4 650–677.

[30] T.-D. Lee and M. Nauenberg, *Degenerate systems and mass singularities*, *Physical Review* **133** (1964), no. 6B B1549.

[31] R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and collider physics.* Cambridge University Press, 1996.

[32] G. Altarelli and G. Parisi, *Asymptotic freedom in parton language*, *Nuclear Physics B* **126** (1977), no. 2 298–318.

[33] Y. L. Dokshitzer, *Calculation of the structure functions for deep inelastic scattering and e+ e- annihilation by perturbation theory in quantum chromodynamics*, *Zh. Eksp. Teor. Fiz* **73** (1977) 1216.

[34] V. N. Gribov and L. N. Lipatov, *Deep inelastic ep-scattering in a perturbation theory.*, tech. rep., Inst. of Nuclear Physics, Leningrad, 1972.

[35] E. Zijlstra and W. Van Neerven, *Order-αs2 qcd corrections to the deep inelastic proton structure functions f2 and fl*, *Nuclear Physics B* **383** (1992), no. 3 525–574.

[36] A. Vogt, S. Moch, and J. A. Vermaseren, *The three-loop splitting functions in qcd: the singlet case*, *Nuclear Physics B* **691** (2004), no. 1 129–181.

[37] G. P. Salam and J. Rojo, *A Higher Order Perturbative Parton Evolution Toolkit (HOPPET)*, *Comput. Phys. Commun.* **180** (2009) 120–156, [arXiv:0804.3755].

[38] M. Botje, *QCDNUM: Fast QCD Evolution and Convolution*, *Comput.Phys.Commun.* **182** (2011) 490–532, [arXiv:1005.1481].

[39] V. Bertone, S. Carrazza, and J. Rojo, *APFEL: A PDF Evolution Library with QED corrections*, *Comput.Phys.Commun.* **185** (2014) 1647, [arXiv:1310.1394].

[40] A. Vogt, *Efficient evolution of unpolarized and polarized parton distributions with qcd-pegasus*, *Comput. Phys. Commun.* **170** (2005) 65–92, [hep-ph/0408244].

[41] R. S. Thorne and W. K. Tung, *PQCD Formulations with Heavy Quark Masses and Global Analysis*, arXiv:0809.0714.

[42] M. Buza, Y. Matiounine, J. Smith, and W. L. van Neerven, *Charm electroproduction viewed in the variable-flavour number scheme versus fixed-order perturbation theory*, *Eur. Phys. J.* **C1** (1998) 301–320, [hep-ph/9612398].

[43] J. C. Collins, *Hard scattering factorization with heavy quarks: A General treatment*, *Phys. Rev.* **D58** (1998) 094002, [hep-ph/9806259].

[44] M. A. G. Aivazis, J. C. Collins, F. I. Olness, and W.-K. Tung, *Leptoproduction of heavy quarks. 2. A Unified QCD formulation of charged and neutral current processes from fixed target to collider energies*, *Phys. Rev.* **D50** (1994) 3102–3118, [hep-ph/9312319].

[45] W.-K. Tung, S. Kretzer, and C. Schmidt, *Open heavy flavour production: conceptual framework and implementation issues*, *Journal of Physics G: Nuclear and Particle Physics* **28** (2002), no. 5 983.

[46] R. S. Thorne, *Variable-flavor number scheme for next-to-next-to-leading order*, *Physical Review D* **73** (2006), no. 5 054019.

[47] R. D. Ball, M. Bonvini, and L. Rottoli, *Charm in Deep-Inelastic Scattering*, *JHEP* **11** (2015) 122, [arXiv:1510.02491].

[48] S. Forte, *Parton distributions at the dawn of the LHC*, *Acta Phys.Polon.* **B41** (2010) 2859, [arXiv:1011.5247].

[49] G. Altarelli and J. Wells, *QCD: The Theory of Strong Interactions*, pp. 27–96. Springer International Publishing, Cham, 2017.

[50] T. Sjostrand, S. Mrenna, and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, *Comput. Phys. Commun.* **178** (2008) 852–867, [arXiv:0710.3820].

[51] *MCFM*, http://mcfm.fnal.gov.

[52] S. Catani and M. Grazzini, *An NNLO subtraction formalism in hadron collisions and its application to Higgs boson production at the LHC*, *Phys.Rev.Lett.* **98** (2007) 222002, [hep-ph/0703012].

[53] R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, R. Pittau, et al., *Four-lepton production at hadron colliders: aMC@NLO predictions with theoretical uncertainties*, *JHEP* **1202** (2012) 099, [arXiv:1110.4738].

[54] J. Currie, E. W. N. Glover, A. Gehrmann-De Ridder, T. Gehrmann, A. Huss, and J. Pires, *Single jet inclusive production for the individual jet p_T scale choice at the LHC*, in *23rd Cracow Epiphany Conference on Particle Theory Meets the First Data from LHC Run 2 Cracow, Poland, January 9-12, 2017*, 2017. arXiv:1704.00923.

[55] S. Gieseke et al., *Herwig++ 2.0 Release Note*, hep-ph/0609306.

[56] T. Gleisberg et al., *Event generation with SHERPA 1.1*, *JHEP* **02** (2009) 007, [arXiv:0811.4622].

[57] T. Carli et al., *A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project*, *Eur.Phys.J.* **C66** (2010) 503, [arXiv:0911.2985].

[58] A. Buckley, J. Ferrando, S. Lloyd, K. Nordström, B. Page, et al., *LHAPDF6: parton density access in the LHC precision era*, *Eur.Phys.J.* **C75** (2015) 132, [arXiv:1412.7420].

[59] G. Chiribella, G. M. D'Ariano, P. Perinotti, and M. F. Sacchi, *Maximum likelihood estimation for a group of physical transformations*, *Int. J. Quantum Inform.* **4** (Jul, 2005) 453. 20 p.

[60] J. Pumplin, D. R. Stump, and W. K. Tung, *Multivariate fitting and the error matrix in global analysis of data*, *Phys. Rev.* **D65** (2001) 014011, [hep-ph/0008191].

[61] D. Stump, J. Pumplin, R. Brock, D. Casey, J. Huston, J. Kalk, H. L. Lai, and W. K. Tung, *Uncertainties of predictions from parton distribution functions. 1. The Lagrange multiplier method*, *Phys. Rev.* **D65** (2001) 014012, [hep-ph/0101051].

[62] P. M. Nadolsky et al., *Implications of CTEQ global analysis for collider observables*, *Phys. Rev.* **D78** (2008) 013004, [arXiv:0802.0007].

[63] R. S. Thorne, L. A. Harland-Lang, A. D. Martin, and P. Motylinski, *The Effect of Final HERA inclusive Cross Section Data MMHT2014 PDFs*, in *Proceedings, 2015 European Physical Society Conference on High Energy Physics (EPS-HEP 2015)*, 2015. arXiv:1508.06621.

[64] S. Dulat, T.-J. Hou, J. Gao, M. Guzzi, J. Huston, P. Nadolsky, J. Pumplin, C. Schmidt, D. Stump, and C. P. Yuan, *New parton distribution functions from a global analysis of quantum chromodynamics*, *Phys. Rev.* **D93** (2016), no. 3 033006, [arXiv:1506.07443].

[65] S. Alekhin, J. Blumlein, S. Moch, and R. Placakyte, *Parton Distribution Functions, α_s and Heavy-Quark Masses for LHC Run II*, arXiv:1701.05838.

[66] A. Martin, A. T. Mathijssen, W. Stirling, R. Thorne, B. Watt, et al., *Extended Parameterisations for MSTW PDFs and their effect on Lepton Charge Asymmetry from W Decays*, *Eur.Phys.J.* **C73** (2013), no. 2 2318, [arXiv:1211.1215].

[67] G. Watt and R. S. Thorne, *Study of Monte Carlo approach to experimental uncertainty propagation with MSTW 2008 PDFs*, *JHEP* **1208** (2012) 052, [arXiv:1205.4024].

[68] S. Forte and G. Watt, *Progress in the Determination of the Partonic Structure of the Proton*, *Ann.Rev.Nucl.Part.Sci.* **63** (2013) 291, [arXiv:1301.6754].

[69] J. Gao and P. Nadolsky, *A meta-analysis of parton distribution functions*, *JHEP* **1407** (2014) 035, [arXiv:1401.0013].

[70] S. Carrazza, J. I. Latorre, J. Rojo, and G. Watt, *A compression algorithm for the combination of PDF sets*, *Eur. Phys. J.* **C75** (2015) 474, [arXiv:1504.06469].

[71] J. Pumplin, *Data set diagonalization in a global fit*, *Phys. Rev.* **D80** (2009) 034002, [arXiv:0904.2425].

[72] J. Pumplin, *Parametrization dependence and $\Delta\chi^2$ in parton distribution fitting*, *Phys.Rev.* **D82** (2010) 114020, [arXiv:0909.5176].

[73] J. F. Owens, A. Accardi, and W. Melnitchouk, *Global parton distributions with nuclear and finite-Q^2 corrections*, *Phys. Rev.* **D87** (2013), no. 9 094012, [arXiv:1212.1702].

[74] **ZEUS, H1** Collaboration, H. Abramowicz et al., *Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data*, *Eur. Phys. J.* **C75** (2015), no. 12 580, [arXiv:1506.06042].

[75] L. A. Harland-Lang, A. D. Martin, P. Motylinski, and R. S. Thorne, *Parton distributions in the LHC era: MMHT 2014 PDFs*, *Eur. Phys. J.* **C75** (2015) 204, [arXiv:1412.3989].

[76] M. Botje et al., *The PDF4LHC Working Group Interim Recommendations*, arXiv:1101.0538.

[77] A. Accardi et al., *A Critical Appraisal and Evaluation of Modern PDFs*, *Eur. Phys. J.* **C76** (2016), no. 8 471, [arXiv:1603.08906].

[78] LHC Higgs Cross Section Working Group, S. Heinemeyer, C. Mariotti, G. Passarino, and R. Tanaka (Eds.), *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties*, *CERN-2013-004* (CERN, Geneva, 2013) [arXiv:1307.1347].

[79] S. Alekhin et al., *The PDF4LHC Working Group Interim Report*, arXiv:1101.0536.

[80] R. D. Ball, S. Carrazza, L. Del Debbio, S. Forte, J. Gao, et al., *Parton Distribution Benchmarking with LHC Data*, *JHEP* **1304** (2013) 125, [arXiv:1211.5142].

[81] J. Rojo et al., *The PDF4LHC report on PDFs and LHC data: Results from Run I and preparation for Run II*, *J. Phys.* **G42** (2015) 103103, [arXiv:1507.00556].

[82] **SM and NLO Multileg Working Group** Collaboration, T. Binoth et al., *The SM and NLO Multileg Working Group: Summary report*, in *Physics at TeV colliders. Proceedings, 6th Workshop, dedicated to Thomas Binoth, Les Houches, France, June 8-26, 2009*, pp. 21–189, 2010. arXiv:1003.1241.

[83] J. R. Andersen et al., *Les Houches 2013: Physics at TeV Colliders: Standard Model Working Group Report*, arXiv:1405.1067.

[84] S. Moch, J. A. M. Vermaseren, and A. Vogt, *The Three loop splitting functions in QCD: The Nonsinglet case*, *Nucl. Phys.* **B688** (2004) 101–134, [hep-ph/0403192].

[85] E. B. Zijlstra and W. L. van Neerven, *Order alpha-s**2 QCD corrections to the deep inelastic proton structure functions F2 and F(L)*, *Nucl. Phys.* **B383** (1992) 525–574.

[86] A. Vogt, S. Moch, and J. A. M. Vermaseren, *The Three-loop splitting functions in QCD: The Singlet case*, *Nucl. Phys.* **B691** (2004) 129–181, [hep-ph/0404111].

[87] R. Hamberg, W. L. van Neerven, and T. Matsuura, *A complete calculation of the order $\alpha - s^2$ correction to the Drell-Yan K factor*, *Nucl. Phys.* **B359** (1991) 343–405. [Erratum: Nucl. Phys.B644,403(2002)].

[88] K. Melnikov and F. Petriello, *Electroweak gauge boson production at hadron colliders through O(alpha(s)**2)*, *Phys. Rev.* **D74** (2006) 114017, [hep-ph/0609070].

[89] S. Catani, L. Cieri, G. Ferrera, D. de Florian, and M. Grazzini, *Vector boson production at hadron colliders: a fully exclusive QCD calculation at NNLO*, *Phys. Rev. Lett.* **103** (2009) 082001, [arXiv:0903.2120].

[90] M. Czakon, D. Heymes, and A. Mitov, *High-precision differential predictions for top-quark pairs at the LHC*, *Phys. Rev. Lett.* **116** (2016), no. 8 082003, [arXiv:1511.00549].

[91] M. Czakon, D. Heymes, and A. Mitov, *Dynamical scales for multi-TeV top-pair production at the LHC*, *JHEP* **04** (2017) 071, [arXiv:1606.03350].

[92] A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, A. Huss, and T. A. Morgan, *Precise QCD predictions for the production of a Z boson in association with a hadronic jet*, *Phys. Rev. Lett.* **117** (2016), no. 2 022001, [arXiv:1507.02850].

[93] A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, A. Huss, and T. A. Morgan, *The NNLO QCD corrections to Z boson production at large transverse momentum*, *JHEP* **07** (2016) 133, [arXiv:1605.04295].

[94] R. Boughezal, J. M. Campbell, R. K. Ellis, C. Focke, W. T. Giele, X. Liu, and F. Petriello, *Z-boson production in association with a jet at next-to-next-to-leading order in perturbative QCD*, *Phys. Rev. Lett.* **116** (2016), no. 15 152001, [arXiv:1512.01291].

[95] R. Boughezal, X. Liu, and F. Petriello, *Phenomenology of the Z-boson plus jet process at NNLO*, *Phys. Rev.* **D94** (2016), no. 7 074015, [arXiv:1602.08140].

[96] J. Currie, A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, A. Huss, and J. Pires, *Precise predictions for dijet production at the LHC*, *Phys. Rev. Lett.* **119** (2017), no. 15 152001, [arXiv:1705.10271].

[97] R. Thorne, *Effect of changes of variable flavor number scheme on parton distribution functions and predicted cross sections*, *Phys.Rev.* **D86** (2012) 074017, [arXiv:1201.6180].

[98] **The NNPDF** Collaboration, R. D. Ball et al., *Theoretical issues in PDF determination and associated uncertainties*, *Phys.Lett.* **B723** (2013) 330, [arXiv:1303.1189].

[99] R. Thorne, *The effect on PDFs and $\alpha_S(M_Z^2)$ due to changes in flavour scheme and higher twist contributions*, *Eur.Phys.J.* **C74** (2014), no. 7 2958, [arXiv:1402.3536].

[100] F. Demartin, S. Forte, E. Mariani, J. Rojo, and A. Vicini, *The impact of PDF and α_s uncertainties on Higgs Production in gluon fusion at hadron colliders*, *Phys. Rev.* **D82** (2010) 014002, [arXiv:1004.0962].

[101] H.-L. Lai et al., *Uncertainty induced by QCD coupling in the CTEQ global analysis of parton distributions*, *Phys. Rev.* **D82** (2010) 054021, [arXiv:1004.4624].

[102] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, *Uncertainties on α_S in global PDF analyses*, *Eur. Phys. J.* **C64** (2009) 653–680, [arXiv:0905.3531].

[103] S. Carrazza, "PDF tools for LHC Run II." Rencontres de Moriond, 2016.

[104] "http://www.hep.ucl.ac.uk/pdf4lhc/mc2h-gallery/website/.".

[105] J. Campbell, R. K. Ellis, and F. Tramontano, *Single top production and decay at next-to-leading order*, *Phys. Rev.* **D70** (2004) 094012, [hep-ph/0408158].

[106] Z. Nagy, *Three jet cross-sections in hadron hadron collisions at next-to-leading order*, *Phys.Rev.Lett.* **88** (2002) 122003, [hep-ph/0110315].

[107] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, *JHEP* **1407** (2014) 079, [arXiv:1405.0301].

[108] V. Bertone, R. Frederix, S. Frixione, J. Rojo, and M. Sutton, *aMCfast: automation of fast NLO computations for PDF fits*, *JHEP* **08** (2014) 166, [arXiv:1406.7693].

[109] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, November, 1995.

[110] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, November, 1986.

[111] R. D. Ball, V. Bertone, M. Bonvini, S. Forte, P. Groth Merrild, J. Rojo, and L. Rottoli, *Intrinsic charm in a matched general-mass scheme*, *Phys. Lett.* **B754** (2016) 49–58, [arXiv:1510.00009].

[112] S. Carrazza, *Parton distribution functions with QED correctionsa*. PhD thesis, Milan U., 2015. arXiv:1509.00209.

[113] G. D'Agostini, *On the use of the covariance matrix to fit correlated data*, *Nucl.Instrum.Meth.* **A346** (1994) 306–311.

[114] **The NNPDF** Collaboration, R. D. Ball et al., *Fitting Parton Distribution Data with Multiplicative Normalization Uncertainties*, *JHEP* **05** (2010) 075, [arXiv:0912.2276].

[115] G. Altarelli, S. Forte, and G. Ridolfi, *On positivity of parton distributions*, *Nucl. Phys.* **B534** (1998) 277–296, [hep-ph/9806345].

[116] **The NNPDF** Collaboration, R. D. Ball et al., *Precision determination of electroweak parameters and the strange content of the proton from neutrino deep-inelastic scattering*, *Nucl. Phys.* **B823** (2009) 195–233, [arXiv:0906.1958].

[117] V. Bertone, R. Frederix, S. Frixione, J. Rojo, and M. Sutton, *aMCfast: automation of fast NLO computations for PDF fits*, *JHEP* **1408** (2014) 166, [arXiv:1406.7693].

[118] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[119] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, *Parton distributions for the LHC*, *Eur. Phys. J.* **C63** (2009) 189, [arXiv:0901.0002].

[120] **New Muon** Collaboration, M. Arneodo et al., *Accurate measurement of F_2^d/F_2^p and $R_d - R_p$*, *Nucl. Phys.* **B487** (1997) 3–26, [hep-ex/9611022].

[121] **New Muon** Collaboration, M. Arneodo et al., *Measurement of the proton and deuteron structure functions, F_2^p and F_2^d, and of the ratio σ_L/σ_T*, *Nucl. Phys.* **B483** (1997) 3–43, [hep-ph/9610231].

[122] L. W. Whitlow, E. M. Riordan, S. Dasu, S. Rock, and A. Bodek, *Precise measurements of the proton and deuteron structure functions from a global analysis of the SLAC deep inelastic electron scattering cross-sections*, *Phys. Lett.* **B282** (1992) 475–482.

[123] **BCDMS** Collaboration, A. C. Benvenuti et al., *A high statistics measurement of the proton structure functions $f_2(x, q^2)$ and r from deep inelastic muon scattering at high q^2*, *Phys. Lett.* **B223** (1989) 485.

[124] **BCDMS** Collaboration, A. C. Benvenuti et al., *A high statistics measurement of the deuteron structure functions $f_2(x, q^2)$ and r from deep inelastic muon scattering at high q^2*, *Phys. Lett.* **B237** (1990) 592.

[125] **CHORUS** Collaboration, G. Onengut et al., *Measurement of nucleon structure functions in neutrino scattering*, *Phys. Lett.* **B632** (2006) 65–75.

[126] **NuTeV** Collaboration, M. Goncharov et al., *Precise measurement of dimuon production cross-sections in ν_μFe and $\bar{\nu}_\mu$Fe deep inelastic scattering at the Tevatron*, *Phys. Rev.* **D64** (2001) 112006, [hep-ex/0102049].

[127] D. A. Mason, *Measurement of the strange - antistrange asymmetry at NLO in QCD from NuTeV dimuon data*, . FERMILAB-THESIS-2006-01.

[128] **H1 , ZEUS** Collaboration, H. Abramowicz et al., *Combination and QCD Analysis of Charm Production Cross Section Measurements in Deep-Inelastic ep Scattering at HERA*, *Eur.Phys.J.* **C73** (2013) 2311, [arXiv:1211.1182].

[129] **H1** Collaboration, F. D. Aaron et al., *Measurement of the Charm and Beauty Structure Functions using the H1 Vertex Detector at HERA*, *Eur. Phys. J.* **C65** (2010) 89–109, [arXiv:0907.2643].

[130] **ZEUS** Collaboration, H. Abramowicz et al., *Measurement of beauty and charm production in deep inelastic scattering at HERA and measurement of the beauty-quark mass*, *JHEP* **09** (2014) 127, [arXiv:1405.6915].

[131] **FNAL E866/NuSea** Collaboration, R. S. Towell et al., *Improved measurement of the anti-d/anti-u asymmetry in the nucleon sea*, *Phys. Rev.* **D64** (2001) 052002, [hep-ex/0103030].

[132] **NuSea** Collaboration, J. C. Webb et al., *Absolute Drell-Yan dimuon cross sections in 800-GeV/c p p and p d collisions*, hep-ex/0302019.

[133] J. C. Webb, *Measurement of continuum dimuon production in 800-GeV/c proton nucleon collisions*, hep-ex/0301031.

[134] G. Moreno et al., *Dimuon production in proton - copper collisions at* $\sqrt{s}$ *= 38.8-GeV*, *Phys. Rev.* **D43** (1991) 2815–2836.

[135] **CDF** Collaboration, T. A. Aaltonen et al., *Measurement of* $d\sigma/dy$ *of Drell-Yan* e^+e^- *pairs in the Z Mass Region from* $p\bar{p}$ *Collisions at* $\sqrt{s} = 1.96$ *TeV*, *Phys. Lett.* **B692** (2010) 232–239, [arXiv:0908.3914].

[136] **CDF - Run II** Collaboration, A. Abulencia et al., *Measurement of the Inclusive Jet Cross Section using the* k_{T} *algorithm in* $p\bar{p}$ *Collisions at* $\sqrt{s}$*=1.96 TeV with the CDF II Detector*, *Phys. Rev.* **D75** (2007) 092006, [hep-ex/0701051].

[137] **D0** Collaboration, V. M. Abazov et al., *Measurement of the shape of the boson rapidity distribution for* $p\bar{p} \to Z/\gamma^* \to e^+e^- + X$ *events produced at* $\sqrt{s}$*=1.96-TeV*, *Phys. Rev.* **D76** (2007) 012003, [hep-ex/0702025].

[138] **D0** Collaboration, V. M. Abazov et al., *Measurement of the electron charge asymmetry in* $\boldsymbol{p\bar{p} \to W + X \to e\nu + X}$ *decays in* $\boldsymbol{p\bar{p}}$ *collisions at* $\boldsymbol{\sqrt{s} = 1.96}$ *TeV*, *Phys. Rev.* **D91** (2015), no. 3 032007, [arXiv:1412.2862]. [Erratum: Phys. Rev.D91,no.7,079901(2015)].

[139] **D0** Collaboration, V. M. Abazov et al., *Measurement of the muon charge asymmetry in* $p\bar{p} \to W{+}X \to \mu\nu + X$ *events at* $\sqrt{s}$*=1.96 TeV*, *Phys.Rev.* **D88** (2013) 091102, [arXiv:1309.2591].

[140] **ATLAS** Collaboration, G. Aad et al., *Measurement of the inclusive $W^{\pm}$ and Z/γ^{*} cross sections in the electron and muon decay channels in pp collisions at $\sqrt{s}$= 7 TeV with the ATLAS detector*, *Phys.Rev.* **D85** (2012) 072004, [arXiv:1109.5141].

[141] **ATLAS** Collaboration, M. Aaboud et al., *Precision measurement and interpretation of inclusive W^{+}, W^{-} and Z/γ^{*} production cross sections with the ATLAS detector*, arXiv:1612.03016.

[142] **ATLAS** Collaboration, G. Aad et al., *Measurement of the high-mass Drell–Yan differential cross-section in pp collisions at $\sqrt{s}$=7 TeV with the ATLAS detector*, *Phys.Lett.* **B725** (2013) 223, [arXiv:1305.4192].

[143] **ATLAS** Collaboration, G. Aad et al., *Measurement of the low-mass Drell-Yan differential cross section at $\sqrt{s}$ = 7 TeV using the ATLAS detector*, *JHEP* **06** (2014) 112, [arXiv:1404.1212].

[144] **ATLAS** Collaboration, G. Aad et al., *Measurement of the Z/γ^{*} boson transverse momentum distribution in pp collisions at $\sqrt{s}$ = 7 TeV with the ATLAS detector*, *JHEP* **09** (2014) 145, [arXiv:1406.3660].

[145] **ATLAS** Collaboration, G. Aad et al., *Measurement of the transverse momentum and ϕ^{*}_{η} distributions of DrellYan lepton pairs in protonproton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *Eur. Phys. J.* **C76** (2016), no. 5 291, [arXiv:1512.02192].

[146] **ATLAS** Collaboration, G. Aad et al., *Measurement of inclusive jet and dijet production in pp collisions at $\sqrt{s}$ = 7 TeV using the ATLAS detector*, *Phys. Rev.* **D86** (2012) 014022, [arXiv:1112.6297].

[147] **ATLAS** Collaboration, G. Aad et al., *Measurement of the inclusive jet cross section in pp collisions at $\sqrt{s}$=2.76 TeV and comparison to the inclusive jet cross section at $\sqrt{s}$=7 TeV using the ATLAS detector*, *Eur.Phys.J.* **C73** (2013) 2509, [arXiv:1304.4739].

[148] **ATLAS** Collaboration, G. Aad et al., *Measurement of the inclusive jet cross-section in proton-proton collisions at $\sqrt{s} = 7$ TeV using 4.5 fb^{1} of data with the ATLAS detector*, *JHEP* **02** (2015) 153, [arXiv:1410.8857]. [Erratum: JHEP09,141(2015)].

[149] **ATLAS** Collaboration, G. Aad et al., *Measurement of the $t\bar{t}$ production cross-section using $e\mu$ events with b-tagged jets in pp collisions at $\sqrt{s}$ = 7 and 8* TeV *with the ATLAS detector*, *Eur. Phys. J.* **C74** (2014), no. 10 3109, [arXiv:1406.5375]. [Addendum: Eur. Phys. J.C76,no.11,642(2016)].

[150] **ATLAS** Collaboration, M. Aaboud et al., *Measurement of the $t\bar{t}$ production cross-section using $e\mu$ events with b-tagged jets in pp collisions at $\sqrt{s}$=13 TeV with the ATLAS detector*, *Phys. Lett.* **B761** (2016) 136–157, [arXiv:1606.02699].

[151] **ATLAS** Collaboration, G. Aad et al., *Measurements of top-quark pair differential cross-sections in the lepton+jets channel in pp collisions at* $\sqrt{s} = 8$ *TeV using the ATLAS detector*, *Eur. Phys. J.* **C76** (2016), no. 10 538, [arXiv:1511.04716].

[152] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the electron charge asymmetry in inclusive W production in pp collisions at* $\sqrt{s} = 7$ *TeV*, *Phys.Rev.Lett.* **109** (2012) 111806, [arXiv:1206.2598].

[153] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the muon charge asymmetry in inclusive pp to WX production at* $\sqrt{s} = 7$ *TeV and an improved determination of light parton distribution functions*, *Phys.Rev.* **D90** (2014) 032004, [arXiv:1312.6283].

[154] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of associated W + charm production in pp collisions at* $\sqrt{s} = 7$ *TeV*, *JHEP* **02** (2014) 013, [arXiv:1310.1138].

[155] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the differential and double-differential Drell-Yan cross sections in proton-proton collisions at* $\sqrt{s} = 7$ *TeV*, *JHEP* **1312** (2013) 030, [arXiv:1310.7291].

[156] **CMS** Collaboration, V. Khachatryan et al., *Measurements of differential and double-differential Drell-Yan cross sections in proton-proton collisions at 8 TeV*, *Eur. Phys. J.* **C75** (2015), no. 4 147, [arXiv:1412.1115].

[157] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the differential cross section and charge asymmetry for inclusive* $\mathrm{pp} \to \mathrm{W}^{\pm} + X$ *production at* $\sqrt{s} = 8$ *TeV*, *Eur. Phys. J.* **C76** (2016), no. 8 469, [arXiv:1603.01803].

[158] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the Z boson differential cross section in transverse momentum and rapidity in protonproton collisions at 8 TeV*, *Phys. Lett.* **B749** (2015) 187–209, [arXiv:1504.03511].

[159] **CMS** Collaboration, S. Chatrchyan et al., *Measurements of differential jet cross sections in proton-proton collisions at* $\sqrt{s} = 7$ *TeV with the CMS detector*, *Phys.Rev.* **D87** (2013) 112002, [arXiv:1212.6660].

[160] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the inclusive jet cross section in pp collisions at* $\sqrt{s} = 2.76$ *TeV*, *Eur. Phys. J.* **C76** (2016), no. 5 265, [arXiv:1512.06212].

[161] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the t-tbar production cross section in the e-mu channel in proton-proton collisions at sqrt(s) = 7 and 8 TeV*, *JHEP* **08** (2016) 029, [arXiv:1603.02303].

[162] **CMS** Collaboration, C. Collaboration, *Measurement of the top quark pair production cross section using eμ events in proton-proton collisions at* $\sqrt{s} = 13$ TeV *with the CMS detector*, .

[163] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the differential cross section for top quark pair production in pp collisions at* $\sqrt{s} = 8$ *TeV*, *Eur. Phys. J.* **C75** (2015), no. 11 542, [arXiv:1505.04480].

[164] **LHCb** Collaboration, R. Aaij et al., *Inclusive W and Z production in the forward region at* $\sqrt{s} = 7$ *TeV*, *JHEP* **1206** (2012) 058, [arXiv:1204.1620].

[165] **LHCb** Collaboration, R. Aaij et al., *Measurement of the cross-section for* $Z \to e^+e^-$ *production in pp collisions at* $\sqrt{s} = 7$ *TeV*, *JHEP* **1302** (2013) 106, [arXiv:1212.4620].

[166] **LHCb** Collaboration, R. Aaij et al., *Measurement of the forward Z boson production cross-section in pp collisions at* $\sqrt{s} = 7$ *TeV*, *JHEP* **08** (2015) 039, [arXiv:1505.07024].

[167] **LHCb** Collaboration, R. Aaij et al., *Measurement of forward W and Z boson production in pp collisions at* $\sqrt{s} = 8$ *TeV*, *JHEP* **01** (2016) 155, [arXiv:1511.08039].

[168] J. Rojo, *Progress in the NNPDF global analysis and the impact of the legacy HERA combination*, in *Proceedings, 2015 European Physical Society Conference on High Energy Physics (EPS-HEP 2015)*, 2015. arXiv:1508.07731.

[169] **HERAFitter developers' Team** Collaboration, S. Camarda et al., *QCD analysis of W- and Z-boson production at Tevatron*, *Eur. Phys. J.* **C75** (2015), no. 9 458, [arXiv:1503.05221].

[170] **ATLAS** Collaboration, G. Aad et al., *Measurement of* $W^{\pm}$ *and Z-boson production cross sections in pp collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, *Phys. Lett.* **B759** (2016) 601–621, [arXiv:1603.09222].

[171] N. Collaboration, "Catalog of plots for NNPDF 3.1." http://pcteserver.mi.infn.it/~nnpdf/nnpdf31-gallery/, 2017.

[172] R. Gavin, Y. Li, F. Petriello, and S. Quackenbush, *W Physics at the LHC with FEWZ 2.1*, *Comput.Phys.Commun.* **184** (2013) 208–214, [arXiv:1201.5896].

[173] C. Carloni Calame, G. Montagna, O. Nicrosini, and A. Vicini, *Precision electroweak calculation of the production of a high transverse-momentum lepton pair at hadron colliders*, *JHEP* **0710** (2007) 109, [arXiv:0710.1722].

[174] R. Boughezal, A. Guffanti, F. Petriello, and M. Ubiali, *The impact of the LHC Z-boson transverse momentum data on PDF determinations*, arXiv:1705.00343.

[175] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, vol. 1. MIT press Cambridge, 2006.

[176] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* **12** (2011) 2825–2830.

[177] S. Carrazza, *Modeling NNLO jet corrections with neural networks*, in *23rd Cracow Epiphany Conference on Particle Theory Meets the First Data from LHC Run 2 Cracow, Poland, January 9-12, 2017*, 2017. arXiv:1704.00471.

[178] J. Bossio, "Atlas results on jets and top." https://indico.cern.ch/event/647565/timetable, 2017. PDF4LHC meeting.

[179] S. Carrazza, A. Ferrara, D. Palazzo, and J. Rojo, *APFEL Web: a web-based application for the graphical visualization of parton distribution functions*, *J.Phys.* **G42** (2015) 057001, [arXiv:1410.5456].

[180] S. Carrazza and Z. Kassabov, *Parton Distribution Functions at LHC and the SMPDF web-based application*, *PoS* **PP@LHC2016** (2016) 020, [arXiv:1606.09248].

[181] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing, "Jupyter Notebooks a publishing format for reproducible computational workflows." 2016.

[182] R. M. Stallman and G. DeveloperCommunity, *Using The Gnu Compiler Collection: A Gnu Manual For Gcc Version 4.3. 3*. CreateSpace, 2009.

[183] C. Lattner and V. Adve, *Llvm: A compilation framework for lifelong program analysis & transformation*, in *Proceedings of the international symposium on Code generation and optimization: feedback-directed and runtime optimization*, p. 75, IEEE Computer Society, 2004.

[184] W. McKinney, *Data structures for statistical computing in python*, in *Proceedings of the 9th Python in Science Conference* (S. van der Walt and J. Millman, eds.), pp. 51 – 56, 2010.

[185] J. D. Hunter, *Matplotlib: A 2d graphics environment*, *Computing In Science & Engineering* **9** (2007), no. 3 90–95.

[186] G. Heinrich, *QCD calculations for the LHC: status and prospects*, in *5th Large Hadron Collider Physics Conference (LHCP 2017) Shanghai, China, May 15-20, 2017*, 2017. arXiv:1710.04998.

[187] G. Altarelli, *The QCD Running Coupling and its Measurement*, *PoS* **Corfu2012** (2013) 002, [arXiv:1303.6065].

[188] M. Johnson and D. Maître, *Strong coupling constant extraction from high-multiplicity Z+jets observables*, arXiv:1711.01408.

[189] **H1** Collaboration, V. Andreev et al., *Determination of the strong coupling constant $\alpha_s(m_Z)$ in next-to-next-to-leading order QCD using H1 jet cross section measurements*, *Eur. Phys. J.* **C77** (2017), no. 11 791, [arXiv:1709.07251].

[190] T. Klijnsma, S. Bethke, G. Dissertori, and G. P. Salam, *Determination of the strong coupling constant $\alpha_s(m_Z)$ from measurements of the total cross section for top-antitop quark production*, *Eur. Phys. J.* **C77** (2017), no. 11 778, [arXiv:1708.07495].

[191] **ATLAS** Collaboration, M. Aaboud et al., *Determination of the strong coupling constant α_s from transverse energy-energy correlations in multijet events at $\sqrt{s} = 8$ TeV using the ATLAS detector*, arXiv:1707.02562.

[192] B. Bouzid, F. Iddir, and L. Semlala, *Determination of the strong coupling constant from ATLAS measurements of the inclusive isolated prompt photon cross section at 7 TeV*, arXiv:1703.03959.

[193] **CMS** Collaboration, S. Chatrchyan et al., *Determination of the top-quark pole mass and strong coupling constant from the t t-bar production cross section in pp collisions at $\sqrt{s} = 7$ TeV*, *Phys.Lett.* **B728** (2014) 496, [arXiv:1307.1907].

[194] J. Rojo, *Constraints on parton distributions and the strong coupling from LHC jet data*, *Int. J. Mod. Phys.* **A30** (2015) 1546005, [arXiv:1410.7728].

[195] S. Lionetti et al., *Precision determination of α_s using an unbiased global NLO parton set*, *Phys. Lett.* **B701** (2011) 346–352, [arXiv:1103.2369].

[196] R. D. Ball, V. Bertone, L. Del Debbio, S. Forte, A. Guffanti, et al., *Precision NNLO determination of $\alpha_s(M_Z)$ using an unbiased global parton set*, *Phys.Lett.* **B707** (2012) 66–71, [arXiv:1110.2483].

[197] **The NNPDF** Collaboration, R. D. Ball et al., *Impact of Heavy Quark Masses on Parton Distributions and LHC Phenomenology*, *Nucl. Phys.* **B849** (2011) 296, [arXiv:1101.1300].

[198] **The NNPDF** Collaboration, R. D. Ball et al., *Unbiased global determination of parton distributions and their uncertainties at NNLO and at LO*, *Nucl.Phys.* **B855** (2012) 153, [arXiv:1107.2652].

[199] J. Currie, E. W. N. Glover, and J. Pires, *NNLO QCD predictions for single jet inclusive production at the LHC*, *Phys. Rev. Lett.* **118** (2017), no. 7 072002, [arXiv:1611.01460].

[200] H. Akaike, *A new look at the statistical model identification*, *IEEE transactions on automatic control* **19** (1974), no. 6 716–723.

[201] K. P. Burnham and D. R. Anderson, *Multimodel inference: understanding aic and bic in model selection*, *Sociological methods & research* **33** (2004), no. 2 261–304.

[202] S. Alekhin, J. Blümlein, and S. Moch, *Parton distribution functions and benchmark cross sections at next-to-next-to-leading order*, *Physical Review D* **86** (2012), no. 5 054009.

[203] P. Jimenez-Delgado and E. Reya, *Dynamical next-to-next-to-leading order parton distributions*, *Physical Review D* **79** (2009), no. 7 074023.

[204] L. Harland-Lang, A. Martin, P. Motylinski, and R. Thorne, *Uncertainties on α_S in the MMHT2014 global PDF analysis and implications for SM predictions*, *The European Physical Journal C* **75** (2015), no. 9 435.

[205] R. Ball, V. Bertone, L. del Debbio, S. Forte, A. Guffanti, J. Latorre, S. Lionetti, J. Rojo, and M. Ubiali, *Precision NNLO determination of α_S using an unbiased global parton set*, *physics letters b* **707** (2012), no. 1 66–71.