UNIVERSITÀ DEGLI STUDI DI MILANO

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

Ciclo XXXV

Theory Predictions for PDF fitting

Settore Scientifico Disciplinare FIS/02

Supervisor:
Prof. Stefano Forte

Candidate:
Alessandro Candido

Anno Accademico 2021/2022

This thesis is written using *ArsClassica* by Lorenzo Pantieri, a style based on the original LaTeX package *classicthesis* of André Miede, inspired by a work of Robert Bringhurst, *"The Elements of Typographic Style"*.

CONTACTS
☞ info@annibale.dev · Write to Alessandro Candido

CONTENTS

INTRODUCTION

The common thread of this thesis are Parton Distribution Functions (PDF) and their ecosystem, half-way between theoretical and experimental High Energy Physics (HEP): being strongly data-driven, they greatly depend on experiments precision and results availability. But theory is also crucial for the extraction, since PDFs are determined to best fit data with theoretical predictions.

As it is possible to infer from the thesis structure, the main subject consists in constructing suitable predictions, in order to allow some new studies in the NNPDF Collaboration, and creating a consistent and improved framework, by designing extensible tools, that will simplify the inclusion of new physical observables and improved theory calculations. This will be discussed extensively in the *first part* of the thesis, and it will include a package dedicated to the solution of DGLAP equations chapter 2, another providing Deep Inelastic Scattering (DIS) predictions and grids chapter 1, and the discussion of the full framework cited chapter 3, where the two packages are integrated in. Finally, a discussion about the impact of Missing Higher Order Uncertainties in PDF is presented in chapter 4, together with present techniques to consume their estimate in a NNPDF fit. Indeed, they will be one of the first important applications of the framework as a whole, even if .

While this has been the main focus for a long time, when me and my collaborators started approaching the completion of the first prototype framework, and the integration in the main NNPDF workflow, some new options for dedicated studies became immediately available. These applications are collected and discussed in the *second part* of the thesis. Most of them are connected not only to the availability of the tools we created, but also to the many achievements and past works of NNPDF, that created a unique foundation for many studies, given the flexibility of its novel methodology and the unparalleled precision reached with its extensive datasets, including experiments spanning multiple decades, and several different processes, from lepton-hadron and hadron-hadron colliders and fixed-target experiments.

Moreover, extracting PDFs is deeply connected to the statistical methods applied. This is more or less true for any analysis based on quantitative experimental data, but it is particularly relevant for the case of PDFs, because of the functional nature of the objected extracted, that exacerbates the dependency of the result on the methodology. From this perspective, the NNPDF methodology is already a novelty, since it required a suitable extension of the usual statistical treatment, based on a given parametrization, in order to access more complex model developed by the Machine Learning community, where direct control of parameter space is difficult, and not very useful. The whole procedure has been recently innovated, even if I have not taken part directly to this process, that

finally led to the release of NNPDF4.0 Ball et al. 2022a, I investigated some limitations of the current approach, especially considering the perspective of the full distribution, that in the NNPDF methodology is only arising at the end of the whole fit, but not used in the individual optimization steps. This led to the proposal of a new candidate methodology, described in chapter 8, that would replace the current usage of a Neural Network (NN) with different techniques. Nevertheless, we currently conjecture the final result to be mostly compatible, but at the time of writing the proposal is still at an early stage, so no full check has yet been performed.

This topic is collected in the *third part*, together with an investigation about the positivity of certain PDFs, described in chapter 7. The reason why the two things are bundled together is that they both affect the final extraction methodology, even if in two very different ways. In NNPDF itself, three main lines of development have always been identified: data implementation, theory computation and extension, and methodological improvements. If the first part is mainly devoted to the theory, this one is instead connected to the methodology.

A final remark is required: software development has been a big share of the main effort. While in the first part this is manifest, it is actually underlying any work described, even though not always to the same extent. The whole NNPDF architecture, and collaboration's main results, are deeply connected to the development of increasingly more reliable tools. Eventually, the main code has been published Ball et al. 2021a, in order to support full transparency, and to make it available for more studies. Potentially, even by authors external to the collaboration. Following this philosophy, all the projects I took part in have been developed open immediately, and they are available in the NNPDF GitHub organization (a few minor ones still in the N3PDF organization), with special care for usability and maintainability, to the best of our abilities.

DECLARATION

This thesis is a report of the research activity conducted during my PhD. Since part of this has already been published in papers, proceedings, or even the documentation of the software developed, some of the material is already appeared in those works.

Following, the description of the sources for each chapter.

CHAPTER 1 content is adapted from the documentation of the `yadism` package, Candido, Hekhorn, and Magni 2022b, available online at `https://yadism.readthedocs.io/`, and section 1.4 in particular has been initially written for a yet unpublished work on low-energy neutrino structure functions

CHAPTER 2 mirrors the `EKO` paper, Candido, Hekhorn, and Magni 2022a, which in turn contains material from the `EKO` documentation, `https://eko.readthedocs.io/`, but some material in the docs has also been (and will be) backported from the paper itself

CHAPTER 3 is based on a proceeding appeared slightly before the thesis itself, Barontini et al. 2022, that is an early presentation of a work that will be discussed in a dedicated publication

CHAPTER 4 has no public source, because the work is based on the toolchain exposed in the previous chapters, and the study itself has not yet reached its final stage; still, part of the material contained was originally authored as an internal note, for the NNPDF collaboration's members, and adapted here for a (possibly) more generic audience

CHAPTER 5 reviews the content of a collaboration's result, Ball, Candido, Cruz-Martinez, et al. 2022, based on `NNPDF4.0` release and `EKO`'s features

CHAPTER 6 is based on an NNPDF work, Ball, Candido, Forte, et al. 2022, based on `NNPDF4.0`, prompted by interaction with experimental users

CHAPTER 7 collects the material appeared in Candido, Forte, et al. 2020

CHAPTER 8 presents a new work-in-progress methodology candidate, thus has no public reference at the moment, even though a likewise work-in-progress report exists online, Petrillo 2022, but it is rather orthogonal to the content of the chapter, since focused on the technical implementation, while chapter 8 introduces just the gist of the idea, and sets the context

0 QUANTUM CHROMODYNAMICS & PARTON DISTRIBUTIONS

Quantum Chromodynamics (QCD) is the theory of strong interactions among colored particles. It is a fundamental constituent of the Standard Model (SM) of particle physics, together with the Electroweak (EW) interaction.

The fundamental feature of QCD is its asymptotic freedom, that makes the coupling perturbative at high enough energies, but non-perturbative at low energies, where the relevant scale to compare with is the intrinsic Λ_{QCD}, whose order of magnitude is roughly the same of the mass of the proton (M_p). This generates a large amount of composite particles, named *hadrons*, whose constituents are colored particles, i.e. quarks, who are determining main quantum numbers on the hadrons, and gluons, the interaction-carrying bosons, arising in the corresponding Yang–Mills (YM) theory, and acting as the binding glue between quarks.

The discovery of hadrons beyond proton and neutron has driven particle physics experiments advancements and theoretical progress in the second half of the twentieth century, resulting in the successful framework of QCD as a Quantum Field Theory following the YM pattern. But other interesting theories has been investigated during the quest for a theory of hadrons, and some of them still have relevant consequences, extending beyond hadrons (e.g. string theory).

Despite the success of the framework, the nature of hadrons is still being intensively investigated, since it would require the solution of non-perturbative QCD dynamics. Different tools are available for this investigation, one of them being the extremely powerful formulation of QCD on a discretized lattice Wilson 1974, but it has not yet been possible to describe the nature of hadrons from first principles, with a sufficiently accurate lattice determination. However, hadrons are ubiquitous in HEP experiments, both as products of high-energy collisions, and as scattering particles at hadronic and semihadronic machines. Therefore a better understanding of the hadronic structure is necessary to formulate precise enough theoretical predictions for collision events, confirming in this way the SM theory, and investigating possible signals of Beyond the Standard Model (BSM) physics.

To circumvent the current limitations about non-perturbative QFT, a different approach has been pursued, whose origin dates even before the discovery of QCD. In order to analyze collisions of composite particles, it was proposed to describe them as packets of collinear point-like constituents, collectively named

partons, each one carrying a fraction of the measured momentum of the scattering particle, Feynman 1969. This partonic picture was successfully applied to predict the high-energy electron-proton collisions, a process called Deep Inelastic Scattering (DIS), resulting in the discovery of Bjorken scaling Bjorken 1967, i.e. the statement that DIS structure functions (cf. section 0.1) do not depend on the exchanged photon virtuality, setting the energy scale of the process.

Bjorken scaling is then violated by QCD corrections, but it is still possible to remain in the framework established by the parton model, because of a fundamental QCD property: factorization, J. C. Collins, Soper, and Sterman 1989. This feature ensures that, up to highly suppressed corrections in the scale ratio, the hadronic cross-sections factorize in a perturbative hard partonic cross-sections, that can be computed in perturbation theory with pQFT calculations, and a universal matrix element, describing the probability that a certain parton constituent from the original hadron enters the hard event.

It is in this framework that Parton Distribution Functions (PDF) are defined, as the non-perturbative matrix element, completing the hadronic cross-section. PDFs universality, granted by factorization (i.e. PDFs being independent from the process considered), makes possible to have a unique set of functions describing the hadronic structure, bridging the gap from the partonic to the hadronic cross-sections. In this context, it is possible to avoid the complexity of a non-perturbative calculation, resorting on a determination of the PDFs from experimental data. Hence, a set of data used to determine the PDFs can constrain the predictions on other events, including different processes.

While the procedure is completely analogue to the determination of other SM model parameters, in the case of PDFs there are two critical differences:

1. the object determined is not an actual parameter of the theory, but we need it because of the complexity of a first principles determination - thus theory first principles theory calculation, able to unfold the non-perturbative dynamics, like lattice, can contribute to the PDFs determination;

2. the unknown parameter is not scalar, but a full function (actually a finite set of them) over the $(0, 1)$ domain, resulting in an infinite amount of degrees of freedom to be determined.

The second point will be further discussed in the context of chapter 8.

PDFs are not the only hadronic object arising from factorization, since also final products observed in semi-inclusive measurements can be described by close analogue, called Fragmentation Functions (FF). More complex object can describe the transverse (non-collinear) dynamics of partons, like Generalized Parton Distributions (GPD) and Transverse Momentum Distributions (TMD). Including more degrees of freedom, and requiring the measurement of more differential observables, the state of this objects is still very raw, compared to collinear PDFs. The research is actively ongoing in this field, but they will not be further described here.

Moreover, other non-perturbative processes are involved in the predictions for observables resulting from high-energy collisions. In particular, the colored scat-

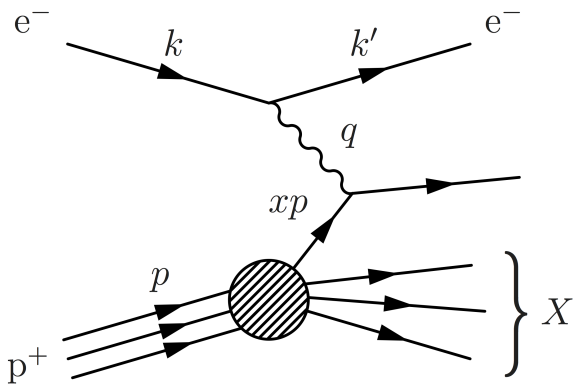

Figure 0.1: The LO Feynman diagram associated to the scattering of a lepton (electron in picture) against an hadron component, mediated by an EW boson.

tering products have to be grouped together in hadrons, since no colored particle is eventually detected, because of QCD *confinement*, another key property of the theory. This mechanism is known as *hadronization*, and it has a close interplay with Parton Shower (PS) and QCD *jets* observables. Also this material is not treated in this thesis, but can be found in the documentation of modern Parton Showers Bellm et al. 2016; Bierlich et al. 2022.

PDFs extraction is therefore essential for HEP research, especially for hadronic machines like LHC. The methodology has been greatly improved in the past few years, and many more processes are now contributing to their determination, together with increasingly accurate theoretical predictions. In the rest of this chapter, some of the main elements related to the PDF theoretical environment will be described. QCD Lagrangian and other fundamental properties are textbook material now, consequently they will not be explicitly described, and instead the interested reader should refer to well-known resources Campbell et al. 2017; J. Collins 2013; R. K. Ellis et al. 2011; Peskin and Schroeder 1995.

0.1 Deep Inelastic Scattering

The Deep Inelastic Scattering process is the scattering of a lepton over an hadron component, mediated by an EW boson fig. 0.1[1]. Since the scattering happens directly on a constituent of the incoming hadron, isolating it from the composite particle (and thus destroying the latter), it is called *deep inelastic*. The leptonic part does not couple directly to QCD, thus the α_s corrections do apply only to the hadronic side (at LO EW), and the EW boson can be seen as emitted from the incoming lepton and absorbed into the hadron. In this picture the pro-

[1] This and the other Feynman diagrams in this chapter are taken from the related CMS wiki page of the Zurich University.

cess can be interpreted as the scattering of an off-shell EW boson over an hadron, probing the hadron composition.

As discussed in the former section, the history of the DIS process is deeply connected to the parton model before, and PDFs determination afterwards, since data from several DIS experiments has been used to provide the main constraints on the PDFs. Multiple old and more recent DIS experiments are still giving a relevant contribution to modern datasets used to extract PDFs, including: Stanford Linear Accelerator Center (SLAC), Bologna-Cern-Dubna-Munich-Saclay (BCDMS), CERN Hybrid Oscillation Research ApparatUS (CHORUS), New Muon Collaboration (NMC), and the more recent NuTev. But the DIS data that mostly constrained PDFs have been the results from H1 and ZEUS experiments at HERA.

The entire chapter 1 will be dedicated to the review of the theory predictions for this process, and to present a software package, `yadism`, dedicated to the calculation of them, with all relevant variants and options.

0.2 DOUBLE HADRONIC & OTHER PROCESSES

Until very recently, DIS has been the one process mostly determining PDFs on its own, even though a non-negligible fraction of Fixed Target Drell–Yan (FTDY) data was already included in the main fits. With the advent of the Large Hadron Collider (LHC), this started changing, since the incredible amount of data generated (and those that will be produced in the future) are compensating the indirectness of the probe. The `NNPDF4.0` release shown for the first time how the LHC data are not only giving sizeable contribution to the PDFs determination, but also able to constraint the PDFs shape on their own Ball et al. 2021b, to a remarkable degree of accuracy.

For this reason, double hadronic initiated processes, like pp at LHC, or $p\bar{p}$ at Tevatron, are now a relevant part of the global QCD dataset used in PDFs extraction. But while for the DIS process analytical calculations are available, most double hadronic processes require the usage of Monte Carlo (MC) integrators, since the resulting integrals are not known analytically, and they quickly extend to many dimensions. Then, in order to obtain the theory predictions required in PDF fits, many different codes are required, since no one implements all possible processes at the state-of-the-art perturbative order (usually NNLO by now, but only for more common processes), and no one is optimal for all of them.

This wide landscape of theoretical predictions, involving increasingly more demanding software tools, is a challenge for a global QCD fits, like PDF ones, since they will need to interface with a variety of codes constantly evolving, and find an effective way to decouple the fit itself from the computational costs involved. This is why interpolation grids have been introduced, to store the results of MC computations and offer a unique and fast interface for their consumption. Grids will be described in more details in chapter 3, where a new grid layout, PineAPPL S. Carrazza et al. 2020a, initially motivated by the extension of existing formats

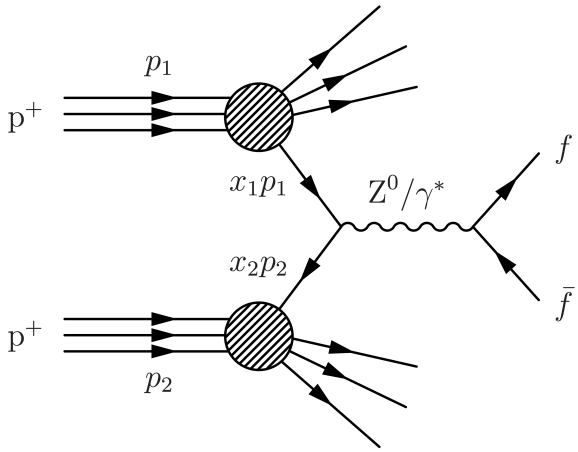

Figure 0.2: The LO Feynman diagram associated to the scattering of two quark components of the proton in the s-channel, generating a virtual EW boson, eventually decaying leptonically.

to allow EW corrections, has been used to construct a full pipeline to streamline the process of producing predictions for PDF and similar fits.

New processes and observables are being observed at the LHC, possibly including new particles and new physics, stressing the need for a more precise determination of the proton structure. The Higgs production, whose main channel is represented in fig. 0.3, that has been first detected by the LHC collaborations ATLAS Aad et al. 2012 and CMS Chatrchyan et al. 2012, is an extremely well-known example of new particle discovery, which heavily involved a proton initiated process, thus depending critically on the PDFs knowledge.

0.3 DOKSHITZER–GRIBOV–LIPATOV–ALTARELLI–PARISI EVOLUTION EQUATIONS

PDFs are a set of functions of two variables (i.e. a function of three variables, of which one is discrete):

$$f_i(z, \mu_F^2) \tag{0.1}$$

The three variables are:

- the *flavor* i of the chosen parton (usually a PID in practice)
- the *momentum fraction* $z \in (0, 1]$ carried by the parton
- the *factorization scale* $\mu_F^2 \in \mathbb{R}^+$

where the last one is required, since PDFs are defined through the factorization theorem, and the factorization scheme used usually involves an unphysical scale,

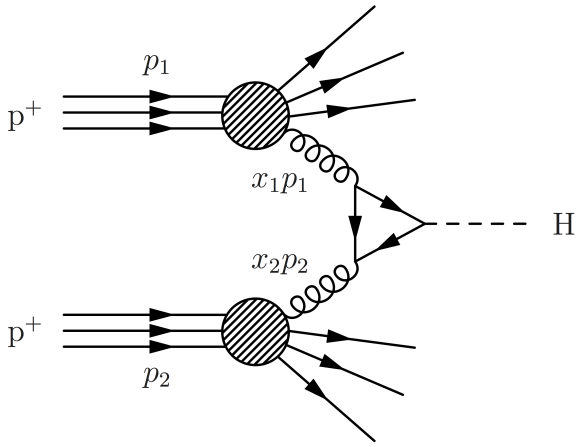

Figure 0.3: The LO Feynman diagram associated to the scattering of two gluon components of the proton, coupling to a virtual quark loop, that finally generates an Higgs boson. This is the Higgs production via gluon fusion, the main channel for Higgs production at LHC.

μ_F^2, in a very similar way to what happens for renormalization schemes R. K. Ellis et al. 2011.

The role of the PDF fits is to determine a border condition at a given value for the unphysical scale, let it be $\mu_{F,0}^2$, since the dependence on the scale is fully encoded in perturbative QCD. Indeed, the various schemes, corresponding to different choices of the unphysical scale μ_F^2, are related by the analogue of Callan-Symanzik equations for factorization, obtained asserting that physical observables must not depend on the choice of the unphysical scale. In such manner, for choices of the scale in the perturbative regime for QCD, some terms are factorized either in the PDF, or in the hard partonic cross section. Moving this scale, the terms are swapped, but thus the PDF values in two different schemes, corresponding to two different values of the scale, should compensate for the difference in the partonic cross sections, both obtained by a perturbative calculation, thus the difference is also determined by perturbative physics.

This relation between PDFs defined at different scales takes the shape of a set of integro-differential equations, called the Dokshitzer–Gribov–Lipatov–Altarelli–Parisi evolution equations (DGLAP) Altarelli and Parisi 1977; Dokshitzer 1977; Gribov and Lipatov 1972:

$$\mu_F^2 \frac{d\mathbf{f}}{d\mu_F^2}(x, \mu_F^2) = \mathbf{P}(a_s(\mu^2), \mu_F^2) \otimes \mathbf{f}(\mu_F^2) \tag{0.2}$$

The equations establish the anomalous scaling of the PDFs, and the kernels $\mathbf{P}$ are called *Altarelli–Parisi splitting functions*.

The equation and its solution is discussed further in chapter 2, where another software package is presented, `EKO`, automating the solution of the associated operator equation. Indeed, any linear equation (as DGLAP) can be solved by

a linear operator, that is actually producing the solution given any boundary condition, thus independently of the boundary condition itself:

$$f_i(z, \mu_F^2) = E_{ij}(\mu_F^2 \leftarrow \mu_{F,0}^2) \otimes f_j(\mu_{F,0}^2) \tag{0.3}$$

We call such an operator, for DGLAP, Evolution Kernel Operators (EKO).

This is another central ingredient in PDF fits, since the data set span different scales, over multiple order of magnitudes, so the PDF determined by the fit has to be evolved first to the suitable scale, to be folded with the partonic calculation at that scale, resulting in the hadronic predictions for the measured observable.

Part I

THEORY

1 DEEP INELASTIC SCATTERING

The Deep Inelastic Scattering (DIS) process has been introduced in section 0.1, briefly outlining its relevance in the PDF determination, and enumerating the various experimental source of DIS data for PDF fits.

In this chapter, the theory of DIS will be described in more details in section 1.1, defining the related kinematics and physical observables, in a wide variety of variants. Then, in section 1.2, it will follow a brief review of the analytic ingredients that is possible to compute in perturbation theory, the so-called *coefficients functions*, with a focus on their analytic properties, since they are crucial to standardize the presentation, in order to allow fully automated numerical integration. It follows a summary of Flavor Number Scheme in section 1.3, since the treatment of quark mass effects is crucial to predict the outcome of DIS experiments, traditionally operating at not-so-high energies (for which charm and bottom mass effects are most relevant).

Finally, `yadism` will be presented, a new program to compute DIS grids and predictions, that implements all the elements discussed in the preceding sections, and a range of further features, including extremely relevant ones, like scale variations (that will be described in more details in chapter 4, but not specifically in the context of DIS), and many other, e.g. TMC and alternative quark mass definitions schemes.

This chapter is not a full review of `yadism` development, since it is partly still work in progress, despite the feature parity with predecessors has already been reached. The online documentation, intended to be a living document, it is already publicly available at:

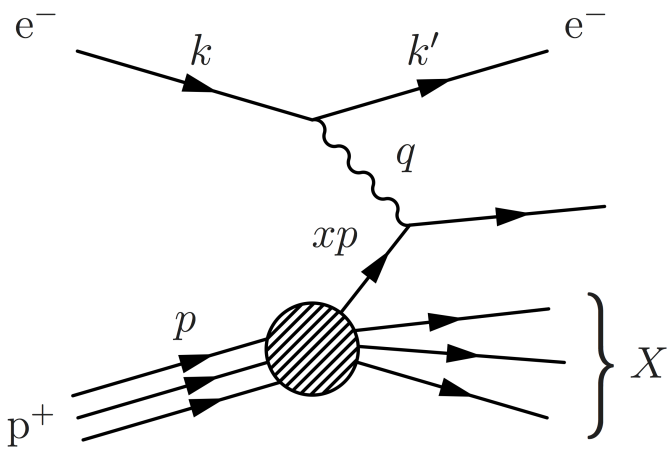

Figure 1.1: The LO Feynman diagram of the DIS process, including the original hadron. Kinematic variables indicated. Notice that, contrary to what is shown in the figure, in the text x will be reserved for the hadronic Bjorken-x, while the partonic momentum fraction will usually be represented by z (the two of them coincide at this perturbative order).

https://yadism.readthedocs.io/

and contains a superset of (almost) all the material in this chapter. Together with the other authors, we hope to present the full set of features in a future publications, Candido, Hekhorn, et al. n.d.

1.1 DEFINITIONS

1.1.1 Kinematics

The following variables are widely used to describe the kinematic of the DIS process (cf. fig. 1.1 for momenta definition):

- $Q^2 = -q^2$ is the EW boson virtuality (photon in the EM process)
- $M_h^2 = p^2$ the mass of the scattered hadron
- $\nu = q \cdot p$, mainly used in the definition of the following
- $x = \frac{Q^2}{2\nu}$, $y = \frac{q\cdot p}{k\cdot p}$ the Bjorken variables

HADRONIC VS PARTONIC Notice that the variables listed here are all **hadronic**, so x is not the partonic momentum fraction (it is only at LO, because the coefficient function is a Dirac δ). In order to avoid confusion the coefficient function variable will be called z, and thus the partonic momentum fraction will be x/z.

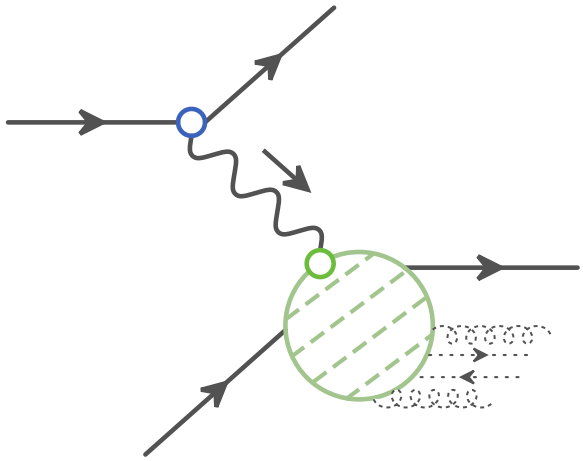

Figure 1.2: In blue the leptonic coupling, the corresponding green one, close to the blob, is instead the hadronic coupling. The blob itself is the hadronic contribution.

It is possible to cut the DIS diagram on the exchanged boson, and compute separately the two sides (cf. fig. 1.2). Since the boson is a Lorentz vector, it carries a space-time index, so the two sides, once squared, will have two indices each, and for this reason they are called the *leptonic* and the *hadronic tensors*. At LO in EW corrections, the leptonic side does not couple to QCD, so the higher order corrections in the strong coupling are all on the side of the hadronic tensor. The fully inclusive DIS cross section σ is thus given by

$$\frac{d\sigma^i}{dxdy} = \frac{2\pi y\alpha^2}{Q^4}\sum_b \eta_b L_b^{\mu\nu} W_{\mu\nu}^b \tag{1.1}$$

where $i \in \{NC, CC\}$ corresponds to the NC or CC processes, respectively. For NC processes, the summation is over $b \in \{\gamma\gamma, \gamma Z, ZZ\}$, whereas for CC interactions there is only W exchange $b = \{W\}$. The normalization factors η_b denote the ratios of the corresponding propagators and couplings to the photon propagator and coupling squared:

$$\eta_{\gamma\gamma} = 1 \tag{1.2}$$

$$\eta_{\gamma Z} = \frac{4\sin^2(\theta_w)}{1-\sin^2(\theta_w)} \cdot \frac{Q^2}{Q^2 + M_Z^2} \tag{1.3}$$

$$\eta_{ZZ} = \eta_{\gamma Z}^2 \tag{1.4}$$

$$\eta_W = \left(\frac{\eta_{\gamma Z}}{2}\frac{1 + Q^2/M_Z^2}{1 + Q^2/M_W^2}\right)^2 \tag{1.5}$$

Based on symmetries and momenta involved, it is possible to characterize the hadronic tensor with three scalar functions[1]:

$$W_{\mu\nu} = \left(-g_{\mu\nu} + \frac{q_\mu q_\nu}{q^2}\right) F_1(x, Q^2) + \frac{\hat{P}_\mu \hat{P}_\nu}{P \cdot q} F_2(x, Q^2) - i\varepsilon_{\mu\nu\alpha\beta} \frac{q^\alpha P^\beta}{2P \cdot q} F_3(x, Q^2) \tag{1.6}$$

with $\hat{P}_\mu = P_\mu - (P \cdot q/q^2) q_\mu$, P the 4-momentum of the hadron and q the 4-momentum of the scattered boson.

These scalar functions are known as DIS *structure functions*.

Instead, the leptonic tensors $L_b^{\mu\nu}$ can all be written in terms of the photonic lepton tensor, because the lepton is assumed massless:

$$L_{\mu\nu}^{\gamma\gamma} = 2\left(k_\mu k'_\nu + k_\nu k'_\mu - (k \cdot k')g_{\mu\nu} - i\lambda\epsilon_{\mu\nu\alpha\beta} k^\alpha k'^\beta\right) \tag{1.7}$$

$$L_{\mu\nu}^{b} = \kappa_b \, L_{\mu\nu}^{\gamma} \tag{1.8}$$

$$\kappa_{\gamma Z} = (g_V^e + e\lambda g_A^e) \tag{1.9}$$

$$\kappa_{ZZ} = (g_V^e + e\lambda g_A^e)^2 \tag{1.10}$$

$$\kappa_W = (1 + e\lambda)^2 \tag{1.11}$$

with $g_V^e = -\frac{1}{2} + 2\sin^2(\theta_w)$ and $g_A^e = -\frac{1}{2}$ the vectorial and axial-vectorial coupling between the Z boson and the lepton with charge $e = \pm 1$ and helicity $\lambda = \pm 1$.

Inserting the leptonic and the hadronic tensors into the cross section we obtain

$$\frac{d\sigma^i}{dxdy} = \frac{4\pi\alpha^2}{xyQ^2}\eta^i \left\{\left(1 - y - \frac{x^2y^2M^2}{Q^2}\right) F_2^i + y^2 x F_1^i \pm \left(y - \frac{y^2}{2}\right) x F_3^i\right\} \tag{1.12}$$

where the $-$ sign in front of F_3 is taken for an incoming e^+ or $\bar{\nu}$ and the $+$ sign for an incoming e^- or ν. The normalization factor η^i are given by $\eta^{NC} = 1$ and $\eta^{CC} = \kappa_W \eta_W$. So unlike in the NC process, in the CC process the leptonic couplings and the propagator corrections are not inside the structure functions but enter only on a cross section level. This is possible because in CC there are no interferences between different bosons. The structure functions are given by

$$F_k^{CC} = F_k^W \tag{1.13}$$

$$F_k^{NC} = F_k^{\gamma\gamma} - (g_V^e \pm \lambda g_A^e)\eta_{\gamma Z} F_k^{\gamma Z} + \left((g_V^e)^2 + (g_A^e)^2 \pm 2\lambda g_V^e g_A^e\right)\eta_{ZZ} F_k^{ZZ} \qquad k \in \{1, 2, L\} \tag{1.14}$$

$$x F_3^{NC} = -(g_A^e \pm g_V^e)\eta_{\gamma Z} x F_3^{\gamma Z} + \left(2 g_V^e g_A^e \pm \lambda((g_V^e)^2 + (g_A^e)^2)\right) x F_3^{ZZ} \tag{1.15}$$

[1] This is true in the case of unpolarized DIS, while a few more functions have to be taken into account in the polarized case.

1.1.2 Structure Functions

As noted above, there are three different structure functions, which we refer to as different **kinds**. Usually, we actually use a different choice for the basis of independent kinds, with respect to what is shown above in eq. (1.6):

$$F_2,\ F_L = F_2 - 2xF_1,\ xF_3 \tag{1.16}$$

Indeed, computing F_L instead of F_1 is advantageous due to the Callan-Gross relation Callan and Gross 1969 $F_L = 0$ in the naive parton model. Notice that the F_L definition it is not always the one above, but it may be corrected, since the actual F_L is the object involved in Callan-Gross relation. Even with this basis, F_1 is still available, but it is treated as a derived quantity, as well as the total cross sections, that is coming from the combination of all the structure functions in the hadronic tensor.

Moreover, the value of xF_3 is often preferred to the bare F_3 structure function, to better represent the way it appears in the full cross section (its "native scaling").

Many experiments either quote the values of the computed structure functions, as resolved by the reconstructed kinematics, or they prefer to report the value of a *reduced cross-section*, but there is not a unique definition for it. Several reduced cross-sections definitions, as defined by experimental collaborations, are implemented and documented in `yadism`.

1.1.3 Process / Currents

DIS is thought as a single process, but it can be drastically different according to which EW boson is actually exchanges. It is actually possible to consistently define three different types of possible **processes**, which correspond to a given *set* of scattering bosons:

- Electromagnetic Current (EC): the only boson allowed to be exchanged is the photon
- Neutral Current (NC): in addition to the photon, the Z boson is also included, so this is a superset of EC. Since now two bosons are allowed also interference terms appear. The Z boson has an axial coupling to the leptons and thus it introduces the problems related to γ_5 Gnendiger et al. 2017. It is relevant to note that there are no Flavor Changing Neutral Currents (FCNC) in the SM, thus NC will always conserve the incoming flavor
- Charged Current (CC): only the W^+ *or* W^- are allowed to be exchanged. The actual boson is determined by the incoming scattering lepton and charge conservation. As the $W^\pm$ are flavor changing additional care is needed in the calculation.

The qualitative behavior of the three processes is drastically different, especially CC concerning flavor structure. But at high energy also NC predictions might be

significantly different from EC, while at low scale the Z contribution is mostly irrelevant.

1.1.4 Heavyness

Another level at which it is possible to split the DIS cross-section is the flavor content of the diagrams involved. By excluding different set of flavors it is possible to obtain what we call different **heavynesses** for structure functions:

total this is the name we give, intuitively, to structure functions in which all possible contributions are taken into account (according to the chosen Flavor Number Scheme)

light for these observables we deny all contributions by heavy quarks (exactly which quarks have to be considered massive depends once more on the FNS)

<*heavy*> e.g. *charm*, contains the contributions in which the heavy quark of selected flavor couples directly to the EW boson (as if only the charge of the given flavor is non-zero, while all the other couplings are set to zero)

All the heavynesses are defined tuning parameters at Lagrangian level, e.g. by imposing that the only the charm quark couples to EW boson, setting to 0 all the other couplings. Because of this all the observables are potential physical observables, since they are well-defined and free of divergences.

Notice that, as a consequence, the contributions in which the heavy quark is present, but does not couple to the EW boson, are not included nor in *light* neither in <*heavy*>, but they are of course present in *total*, thus:

$$O_{\text{total}} \neq O_{\text{light}} + \sum_{h \in \text{heavy}} O_h \tag{1.17}$$

1.2 COEFFICIENT FUNCTIONS

Using the collinear factorization theorem of DIS, J. C. Collins, Soper, and Sterman 1989, we can write any hadronic structure function F_k in terms of PDF $f_j(\xi, \mu_F^2)$ and the coefficient functions $c_{j,k}(z, Q^2, \mu_F^2, \mu_R^2)$ (acting as *partonic structure functions*) using a convolution over the first argument:

$$F_k^{bb'}(x, Q^2, \mu_F^2, \mu_R^2) = \sum_p f_p(\mu_F^2) \otimes_x c_{k,p}^{bb'}(Q^2, \mu_F^2, \mu_R^2) \tag{1.18}$$

where the sum runs over all contributing partons $p \in \{g, q, \bar{q}\}$. In the following we will assume that a quark $\hat{q}$ is hit by the boson. Note that this is *independent* of the incoming parton p. The dependency on the renormalization and factorization scales has to be propagated consistently, in order to be able to use them as an

estimate for MHOU (cf. chapter 4). For those cases in which scale variations are not relevant, it is safe to consider $\mu_R^2 = \mu_F^2 = Q^2$.

Using pQCD, it is possible to expand the coefficient functions in powers of the strong coupling $a_s(\mu_R^2) = \frac{\alpha_s(\mu_R^2)}{4\pi}$:

$$c_{k,p}^{bb'}(z, Q^2, \mu_F^2, \mu_R^2) = \sum_{l=0} a_s^l(\mu_R^2) c_{k,p}^{bb',(l)}(z, Q^2, \mu_F^2, \mu_R^2) \tag{1.19}$$

In practice, different normalization might be used.

Similar to the splitting on the leptonic side we have to split on the partonic side again:

$$c_{k,p}^{bb'} = g_{\hat{q},b}^V g_{\hat{q},b'}^V c_{k,p}^{VV} + g_{\hat{q},b}^A g_{\hat{q},b'}^A c_{k,p}^{AA} \qquad k \in \{1, 2, L\} \tag{1.20}$$

$$c_{3,p}^{bb'} = g_{\hat{q},b}^V g_{\hat{q},b'}^A c_{3,p}^{VA} \tag{1.21}$$

The main categories for coefficients the same of Structure Functions, i.e.:

- the **process** (EM/ NC/ CC)
- the **kind** (F2/L/3)
- the **heavyness** involved; it slightly differ from that of structure functions, since it is referred to individual contributions
  - *light*, when no mass is involved
  - *heavy*, when mass effects are accounted for, with a single quark mass (two mass effects are rather negligible, and very complex to include), since (it is the same for every flavor, just depending on the numerical value of the mass as a parameter)
  - *asymptotic*, that are the limit of heavy contributions, to subtract the double counting in GM-VFNS schemes, like FONLL
  - *intrinsic*, in which the incoming parton is a massive one (can also combine with asymptotic)
- but there is a new one: the **channel** (ns/ps/g), and it is related to the incoming parton:
  - if the EW boson it is coupling to a *quark* line connected to the incoming one, than each PDF it's contributing proportionally to his charge (e.g.: electric charge for the photon); this is called *non-singlet (ns)*
  - otherwise the line to which the EW boson is coupling it will be detached from the incoming by *gluonic* lines, and the gluon is flavor blind, so all the charges are summed and all the PDF are contributing the same way; this is called *pure singlet (ps)*
  - eventually: if a *gluon* is entering all the quarks will couple to the EW boson (if no further restrictions are imposed by the observable, e.g. F2charm), as in the singlet case, and so the charges are summed over; this is called the *gluon (g)* (because *it is* the gluon...)

NLO	light	heavy	intrinsic	asymptotic
NC	✓	✓	✓	✓
CC	✓	✓	✓	✓
NNLO				
NC	✓	✓	✗	✓
CC	✓	tabulated*	✗	✓
N3LO				
NC	✓	✗†	✗	✗‡
CC	✓	✗†	✗	✗

* Already available as K-factors **Gao:2017kkx**, now being integrated in the grid format.

† Full calculation not available but an approximated expression can be constructed from partial results.

‡ Calculation available, to be implemented.

Table 1.1: Overview of the different types and accuracy of the DIS coefficient functions currently implemented in `yadism`. For each perturbative order (NLO, NLO, and N3LO) we indicate the light-to-light ("light"), light-to-heavy ("heavy"), heavy-to-light and heavy-to-heavy ("intrinsic") and "asymptotic" ($Q^2 \gg m_h^2$ limit) coefficients functions which have been implemented and benchmarked. The NNLO heavy quark coefficient functions for CC scattering are available in K-factor format and are being implemented into the `yadism` grid formalism.

  - the **parity structure** (vectorial-vectorial/axial-axial/vectorial-axial), it is relevant only for the NC, and should be taken into account

Of course, there the coefficient functions also depends on the **perturbative order** they are computed at.

A recap of the status of coefficient functions as implemented in `yadism` (cf. section 1.4), is contained in table 1.1.

Notice that the concept of *heavyness* in coefficient functions loosely corresponds to the same one in structure functions. In particular, a light structure functions is only expressed in terms of light coefficient functions. About the massive contributions instead, the idea would be similar, but here there are a few more complications: the contributions to the observable computed in a massive scheme are just divided in two categories, i.e. the massive coefficients for massless PDFs and the intrinsic contributions. When instead a GM-VFNS is being constructed, the need for asymptotic limit of massive observables arises, and they would come with their own asymptotic coefficient functions. However, according to the specific scheme some additional terms related to the matching conditions might be included in the coefficient functions. Moreover, GM-VFNS might construct composite observables, assembling other "elementary" observables. In this cases, the

connection between the heavyness of coefficient functions and observables is completely lost, but it still exists at the level of the single components.

1.2.1 Treatment of distributions

Coefficient functions are not always pure functions of the partonic variables, since in absence of masses, regulating all divergences, the regularization itself can generate distributions. All distributions disappears once convoluted with the PDF (or a suitable interpolation basis, as a placeholder for a generic PDF):

$$\sigma = \sum_j f_j \otimes c_j = \sum_j \int_x^1 \frac{dz}{z} f_j(x/z) c_j(z) \tag{1.22}$$

so they never survive in physical observables.

A generic coefficient function will allow for three ingredients:

- *Regular* functions $r(z)$ that are well behaving, i.e. integrable, for all $z \in (0, 1]$; these typically contain polynomials, logarithms and dilogarithms
- *Dirac*-δ distributions: $\delta(1-z)$
- *Plus* distributions: $[g(z)]_+$ which have a regulated singularity at $z \to 1$ and are defined by

$$\int_0^1 dz\, f(z)\, [g(z)]_+ = \int_0^1 dz\, \left(f(z) - f(1)\right) g(z) \tag{1.23}$$

The "plused" function can be a generic function, but in practice will almost always be $\log^k(1-z)/(1-z)$. The "plused" function has to be regular at $z = 0$. These contributions are related to soft and/or collinear singularities in the physical process.

In order to do the convolution in a generic way we adopt the Regular-Singular-Local (RSL) scheme: i.e. we categorize them by their behavior under the convolution internal. This is needed because of the mismatch in the definitions of the convolution and the plus prescription. Any coefficient function $c(z)$ can be written in the following way:

$$f \otimes c = \int_x^1 \frac{dz}{z} f(x/z) c^R(z) + \int_x^1 dz \left(\frac{f(x/z)}{z} - f(x)\right) c^S(z) + f(x) c^L(x) \tag{1.24}$$

The remapping of the coefficient function ingredients on to the RSL elements is done in the following way:

- Regular functions $c(z) = r(z)$ contribute only to the regular bit:

$$c^R(z) = r(z)\,, \quad c^S(z) = 0 = c^L(x) \tag{1.25}$$

- Dirac delta distributions $c(z) = \delta(1-z)$ only contribute to the local bit:

$$c^R(z) = 0 = c^S(z)\,, \quad c^L(x) = 1 \tag{1.26}$$

- "Raw" plus distributions $c(z) = [g(z)]_+$ contribute to both the singular and the local bit:

$$c^R(z) = 0\,, \quad c^S(z) = g(z)\,, \quad c^L(x) = -\int_0^x dz\, g(z) \tag{1.27}$$

derivation

$$f \otimes [g]_+ = \int_x^1 \frac{dz}{z} f(x/z) \cdot [g(z)]_+ \tag{1.28}$$

$$= \int_0^1 \frac{dz}{z} f(x/z) \cdot [g(z)]_+ - \int_0^x \frac{dz}{z} f(x/z) \cdot [g(z)]_+ \tag{1.29}$$

$$= \int_0^1 dz \left(\frac{f(x/z)}{z} - f(x)\right) \cdot g(z) - \int_0^x dz\, \frac{f(x/z)}{z} \cdot g(z) \tag{1.30}$$

$$= \int_x^1 dz \left(\frac{f(x/z)}{z} - f(x)\right) \cdot g(z) - f(x) \int_0^x dz\, g(z) \tag{1.31}$$

$$\Rightarrow \begin{cases} c^R(z) & = 0 \\ c^S(z) & = g(z) \\ c^L(x) & = -\int_0^x dz\, g(z) \end{cases} \tag{1.32}$$

- A product of a regular function and a plus distribution $c(z) = r(z)\,[g(z)]_+$ contributes to all three bits:

$$c^R(z) = (r(z) - r(1))g(z)\,, \quad c^S(z) = r(1)g(z)\,, \quad c^L(x) = -r(1) \int_0^x dz\, g(z) \tag{1.33}$$

derivation

$$f \otimes c = \int_x^1 \frac{dz}{z} f(x/z) r(z) \cdot [g(z)]_+ \tag{1.34}$$

$$= \int_0^1 \frac{dz}{z} f(x/z) r(z) \cdot [g(z)]_+ - \int_0^x \frac{dz}{z} f(x/z) r(z) \cdot [g(z)]_+ \tag{1.35}$$

$$= \int_0^1 dz \left(\frac{f(x/z)r(z)}{z} - f(x)r(1) \right) \cdot g(z) - \int_0^x dz \, \frac{f(x/z)r(z)}{z} \cdot g(z) \tag{1.36}$$

$$= \int_x^1 dz \left(\frac{f(x/z)r(z)}{z} - f(x)r(1) \right) \cdot g(z) - f(x)r(1) \int_0^x dz \; g(z) \tag{1.37}$$

$$= \int_x^1 dz \left(\frac{f(x/z)(r(z) + r(1) - r(1))}{z} - f(x)r(1) \right) \cdot g(z) - f(x)r(1) \int_0^x dz \; g(z) \tag{1.38}$$

$$= \int_x^1 dz \left(\frac{f(x/z)}{z} - f(x) \right) r(1) \cdot g(z) + \int_x^1 dz \frac{f(x/z)(r(z) - r(1)))}{z} g(z) - f(x)r(1) \int_0^x dz \; g(z) \tag{1.39}$$

$$= \int_x^1 \frac{dz}{z} f(x/z)r(1) \cdot [g(z)]_+ + \int_x^1 dz \frac{f(x/z)(r(z) - r(1)))}{z} g(z) \tag{1.40}$$

$$\Rightarrow \begin{cases} c^S(z) & = r(1)g(z) \\ c^R(z) & = (r(z) - r(1))g(z) \\ c^L(x) & = -r(1) \int_0^x dz \, g(z) \end{cases} \tag{1.41}$$

Also consider that a plus distribution that contains a regular and a singular function $c(z) = [r(z)g(z)]_+$ can be simplified by

$$[r(z)g(z)]_+ = r(z) \, [g(z)]_+ - \delta(1-z) \int_0^1 dy \; r(y) \, [g(y)]_+ \tag{1.42}$$

derivation;

$$\int_0^1 dz \; f(z) \, [r(z)g(z)]_+ = \int_0^1 dz \, (f(z) - f(1)) \, r(z)g(z) \tag{1.43}$$

$$= \int_0^1 (f(z)r(z) - f(1)r(1)) \, g(z) \; dz - f(1) \int_0^1 dz (r(z) - r(1))g(z) \tag{1.44}$$

$$= \int_0^1 dz \; f(z) \left(r(z) \, [g(z)]_+ \right) - f(z) \left(\delta(1-z) \int_0^1 dy \; r(y) \, [g(y)]_+ \right) \tag{1.45}$$

1.3 FLAVOR NUMBER SCHEMES

FNS or Heavy Quark Matching Schemes are dealing with the ambiguity of including massive quark contributions to physical cross sections. In general, it is possible to consider two different kinematic regimes that require a different handling of the massive contributions: for $Q^2 \lesssim m^2$ the heavy quark should be treated with the full mass dependence. $Q^2 \gg m^2$ however the quark should be considered massless, because otherwise a resummation of the occurring terms $\log\left(m^2/Q^2\right)$ would be required.

1.3.1 Fixed Flavor Number Scheme

As the name FFNS suggests, we are considering a fixed number of flavors $n_f = n_l + 1$ with n_l light flavors and *one* (and *only one*) heavy flavor with a finite mass m. The number of light quarks n_l is arbitrary but fixed and can range between 3 and 6. Except for intrinsic contributions we are *not* allowing the heavy PDF to contribute (and those corresponding to flavors not in the scheme as well). This scheme is adequate for $Q^2 \sim m^2$.

- the **light** structure functions corresponds to the interaction of the purely light partons, i.e. the coefficient functions may only be a function of z, Q^2 (and unphysical scales); in particular they can *not* depend on any quark mass. This may be consistently obtained computing contributions for a Lagrangian with all masses set to 0.
  - this definition is consistent with Moch, Rogal, et al. 2008; Moch and Vermaseren 2000; Moch, Vermaseren, and Vogt 2005, 2009; Vermaseren, Vogt, et al. 2005, and `QCDNUM`, M. Botje 2011
  - but is not consistent with `APFEL`, Bertone, Stefano Carrazza, and Rojo 2014, which instead is calling *light* the sum of contributions in which a light quark is coupled to the EW boson
- as noted in section 1.1, the **total** structure functions are *not* the sum of **light** and the single **heavy** ones, but contains additional terms *Fmissing* such as the Compton diagrams in Hekhorn 2019; this is the proper physical object, accounting for all contributions coming from the full Lagrangian.
- the **heavy** structure functions are defined by having in the Lagrangian *only* the EW charges that are associated to the specific *coupling* quark (the only massive one). In NC this corresponds to the electric and weak charges of the quarks but in CC the situation is bit more involved: we divide the CKM-matrix into several parts:

$$V_{CKM} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \tag{1.46}$$

and associate the blue couplings to the charm structure functions, green to bottom and purple to top. For $F_{2,c}^{\nu,p}$ this in effect amounts to

$$F_{2,c}^{\nu,p} = 2x\Big\{ C_{2,q} \otimes \Big[{V_{cd}}^2 (d + \bar{c}) + {V_{cs}}^2 (s + \bar{c}) \Big] + 2 \left({V_{cd}}^2 + {V_{cs}}^2 \right) C_{2,g} \otimes g \Big\} \tag{1.47}$$

Note that even heavier contributions are *not* available. E.g.:

- there is no contributions coming from either *bottom* or *top* to $F_{2,c}$
- while *charm* would contribute to $F_{2,b}$, but only as a massless flavor.

1.3.2 Zero-Mass Variable Flavor Number Scheme

As the name ZM-VFNS suggests, this scheme takes into account a variable number of light flavors n_f with $n_f = n_f(Q^2)$. There is an *activation* scale $Q^2_{thr,i}$ associated to each potentially "heavy" quark (i.e. charm, bottom, and top) and whenever $Q^2 \geqslant Q^2_{thr,i}$ this quark is considered massless, otherwise infinitely massive.

This scheme is adequate for $Q^2 \gg m^2$.

- the **heavy** structure functions are *not* defined, as quark masses are either 0 or ∞ (so no massive correction is available at all)
- **total** ones thus are equal to *light*
- **light** structure functions corresponds to the interaction of the purely light partons, i.e. the coefficient functions may only be a function of z, Q^2 and eventually unphysical scales; so they can *not* depend on any quark mass

ZM-VFNS dependence on thresholds is simple, they just define the Q^2 patches in which n_f is constant (and they are of course different from the quark masses, that are always considered to be zero or infinite). Also note that $Q^2_{thr,i}$ are not necessarily, but usually chosen to be, the quarks' masses.

1.3.3 Fixed Order Next-to-Leading Log (FONLL)

FONLL Forte, Laenen, et al. 2010 is a GM-VFNS that includes parts of the DGLAP equations into the matching conditions. That is: two different schemes are considered, and they are matched at a given scale, accounting for the resummation of collinear logarithms, but also for power suppressed terms from massive corrections at the same time. In the original paper the prescription is only presented for the charm contributions, but just as a placeholder of an arbitrary massive quark. The prescription defines two separate regimes, below and above the *next* heavy quark threshold: Q_{thr,n_f+2}. As in the case of ZM-VFNS, these matching thresholds are not necessarily, but usually chosen to be, the quarks' masses.

- for $Q^2 < Q^2_{thr,n_f+2}$: the general expression, eqs. (14-15) of Forte, Laenen, et al. 2010, is:

$$F^{FONLL}(x,Q^2) = F^{(d)}(x,Q^2) + F^{(n_f)}(x,Q^2) \tag{1.48}$$

$$F^{(d)}(x,Q^2) = F^{(n_f+1)}(x,Q^2) - F^{(n_f,0)}(x,Q^2) \tag{1.49}$$

  Here the scheme change between the schemes with n_f (i.e. the FFNS scheme in which the coupling flavor is the only one considered to be massive) and (n_f+1) flavors (i.e. the FFNS scheme with only massless quarks, including the formerly massive one) is explicitly included.

  This scheme change is related to the DGLAP matching conditions: in particular the massive corrections are only coming from the n_f scheme, but the collinear contribution is present in both:

  - the n_f scheme includes the logarithms of the involved mass, while the PDF of the massive object are scale-independent by definition (since the factorization terms are kept in the matrix element)
  - the (n_f+1) scheme does not account for them in them in the coefficient function, but instead they are resummed in the PDF evolution through the DGLAP equation

  By matching the two schemes a GM-VFNS is obtained, accounting for both the massive corrections and the resummation of collinear logarithms. The matching is obtained subtracting the asymptotic massless limit of the massive expression, namely $F^{(n_f,0)}(x,Q^2)$, while adding the (n_f+1) expression, such that for large Q^2 the massive n_f contribution cancels with the asymptotic one, and only the truly light contribution survives. Actually below the former threshold, so $Q^2 < Q^2_{thr,n_f+1}$, FNS with n_f flavors is employed, i.e. a $\theta(Q^2 - Q^2_{thr,n_f+1})$ is prepended to $F^{(d)}$.

- above this threshold: the ZM-VFNS is employed and this leads to an inconsistency at this Q_{thr,n_f+2} threshold, but a good approximation nevertheless. This amounts to simply make an hard cut to the original smooth decay of massive contributions, and to add the subsequent thresholds for the following massive quarks.

Damping

Up to NLO the scheme change (from $n_f - 1$ flavors to n_f) is continuous, but in general it is not. In order to recover the continuous transition a damping procedure may be adopted, turning the scheme in the so called *damp FONLL*.

Continuity on its own is not an issue, but it is one symptom of a feature of $F^{(d)}$ eq. (1.49): while it improves the behavior at large Q^2 it is unreliable for $Q^2 \sim Q^2_{thr,n_f+1}$. For this reason might be a good idea to suppress $F^{(d)}$ near

threshold, and then this restore continuity. The generic shape of this suppression is written in eq. (17) of Forte, Laenen, et al. 2010, and it is:

$$F^{(d,th)}(x, Q^2) = f_{thr}(x, Q^2) F^{(d)}(x, Q^2) \tag{1.50}$$

In particular the following conditions are needed for $f_{thr}(x, Q^2)$ to fit the task:

- be such that $F^{(d,th)}(x, Q^2)$ and $F^{(d)}(x, Q^2)$ is power suppressed for large Q^2
- enforce the vanishing of $F^{(d,th)}(x, Q^2)$ at and below threshold

A common shape for $f_{thr}(x, Q^2)$ is then:

$$f_{thr}(x, Q^2) = \theta(Q^2 - m^2) \left(1 - \frac{Q^2}{m^2}\right)^2 \tag{1.51}$$

The power used here is 2, but in general this is arbitrary, and thus it is a user choice in `yadism`.

Threshold different from heavy quark mass

The threshold in FONLL plays a relevant role, since it is deciding where (in Q^2) the match should happen. A typical choice is to put the threshold on top of the relevant quark mass (also in ZM-VFNS, mimicking the opening of a new channel). This is not mandatory, as the threshold is just an FNS parameter it can be freely chosen. If the threshold is then chosen *different* from the quark mass, a new scale ratio appears, and the expressions might depend also on this one. Notice that the threshold is only a parameter of FONLL, so it can not affect the FFNS ingredients of the scheme (which can only depend on the real quark masses, through massive propagators). Then only the massless limit (the double counting preventing bit) might include a threshold dependency, and in practice it will only change the relevant logarithm, that:

- instead of being the logarithm of the ratio between the process scale and *the mass*
- is a logarithm of the ratio with *the threshold*

as it is discussed in Forte, Napoletano, et al. 2018.

1.4 yadism: DIS GRIDS PROVIDER

Deep-inelastic structure functions can be evaluated with several public codes such as `APFEL` Bertone, Stefano Carrazza, and Rojo 2014 and QCDNUM M. Botje 2011. These various available DIS codes differ in the accuracy with which structure functions can be computed, whether they are based on the x-space or the

N-space formalism, the treatment of heavy quark mass effects and of target mass corrections, the availability of polarised and time-like coefficient functions, and the inclusion of QED corrections among other considerations.

`yadism` is a new actor on this scene, being created as a framework for the evaluation of DIS structure functions from the same family as the EKO Candido, Hekhorn, and Magni 2022a DGLAP evolution code (cf. chapter 2). The open source `yadism` code can be obtained from its GitHub repository

`NNPDF/yadism`

together with a detailed documentation, tutorials, and user-friendly examples

`https://yadism.readthedocs.io/`

One of the main advantages of `yadism` is that it is integrated with the fast interpolation grid toolbox PineAPPL S. Carrazza et al. 2020a, and hence DIS structure functions can be treated on the same footing as hadronic observables from the point of view of PDF fitting and related applications. PineAPPL provides a unique grid format, with application programming interfaces (APIs) for different programming languages and a user-friendly command-line interface to manage the grid files (cf. chapter 3). Furthermore, `yadism` implements the available N^3LO DIS coefficient functions, which combined with (approximate) N^3LO evolution and heavy quark matching conditions available in `EKO` provide theoretical calculations required to carry out a N^3LO PDF determination. `yadism` will be described in detail in an upcoming publication Candido, Hekhorn, et al. n.d., and here we summarise its main features, in particular those relevant to the present study, and highlight benchmarking studies carried out.

GRID FORMALISM. As already indicated in eq. (1.18), in the perturbative regime DIS structure functions are given by the factorised convolution of process-dependent partonic scattering cross-sections and of process-independent parton distribution functions,

$$F_i(x, Q^2) = \sum_j \int_x^1 \frac{dz}{z} C_{i,j}(z, \alpha_s(Q^2)) f_j\left(\frac{x}{z}, Q^2\right) \equiv C_{j;i} \otimes f_j \tag{1.52}$$

where j is an index that runs over all possible partonic initial states and $C_{i,j}$ is the process-dependent, but target-independent, coefficient function, given by an expansion in the QCD coupling $\alpha_s(Q^2)$. In the third term of eq. (1.52) and in the following, sum over repeated indices is implicit.

As standard for fast interpolation techniques developed in the context of PDF fits Bertone, Rikkert Frederix, et al. 2014; Carli et al. 2010; S. Carrazza et al. 2020a; Wobisch et al. 2011, the PDFs can be expanded over an interpolation basis

$$f_j(\xi) = \sum_\alpha p_\alpha(\xi) f(\xi_\alpha) \equiv p_\alpha(\xi) f_\alpha\,, \qquad \xi = \frac{x}{z}\,, \tag{1.53}$$

with $p_\alpha(x)$ some suitable polynomial basis. This way the convolution in eq. (1.52) can be replaced by a simple contraction

$$F_i = C_{j;i} \otimes f_j = C_{j\alpha;i} \cdot f_\alpha\,, \qquad C_{j\alpha;i} = C_{j;i} \otimes p_\alpha\,, \tag{1.54}$$

in terms of PDFs evaluated at fixed grid points ξ_α and precomputed coefficients $C_{j\alpha;i}$. In `yadism` the polynomial interpolation basis is provided by EKO. The same grid structure can be generalised to accomodate extensions of the basic structure function calculation in eq. (1.52) such as heavy quark mass effects, renormalization and factorization scale variations Abdul Khalek et al. 2019a,b, and target mass corrections, among other effects. Isospin modifications, required to evaluate the neutron, deuteron, or heavy nuclear structure functions, can be accounted for either at the coefficient function level or at the input PDF level.

The grid formalism summarised schematically in eq. (1.54) requires as input the corresponding DIS coefficient functions (cf. section 1.2). table 1.1 provides an overview of the different types and accuracy of the DIS coefficient functions currently implemented in `yadism`. For each perturbative order (NLO, NLO, and N3LO) we indicate the neutral-current and charged-current light-to-light (*light*), light-to-heavy (*heavy*), heavy-to-light and heavy-to-heavy (*intrinsic*) and *asymptotic* ($Q^2 \gg m_h^2$ limit) coefficients functions which have been implemented and benchmarked. The NNLO heavy quark coefficient functions for CC scattering are available in K-factor format and are being implemented into the `yadism` grid formalism. We note that the full calculation of the N3LO NC massive coefficient functions is not available but an approximated expression can be constructed from partial results. Heavy quark structure functions can be evaluated in the FONLL GM-VFNS Forte, Laenen, et al. 2010, as well as in the FFNS and ZM-VFNS. We point out that the list in table 1.1 is going to be updated as new features are added, and therefore the interested user is encouraged to consult the online documentation for an up-to-date states of available coefficient functions.

SCALE VARIATIONS. As done by other public DIS tools, `yadism` also provides the option of varying the renormalization and factorization scales in the calculation. The code follows the definitions of scale variations from W. L. van Neerven and Vogt 2000, 2001, which are consistent with the broader picture of scale variations relevant for PDF fits from Abdul Khalek et al. 2019b where they also affect the DGLAP evolution. There are two kinds of scale variations: renormalization scale μ_R dependence, related to the ultraviolet renormalization scheme, and factorization scale μ_F dependence, related to the subtraction of collinear logarithms in the adopted factorization scheme. The factorization scale μ_F sets the boundary between the coefficient functions and the DGLAP-evolved PDFs. Scale variations at a given perturbative order can be constructed from combining ingredients already present at the previous perturbative order, and hence for this reason they represent a suitable predictor of potentially unknown missing higher orders. Within `yadism`, the scale variation contributions to the DIS structure functions are stored in separate grids such that the values of the scale ratios $\xi_F^2 = \mu_F^2/Q^2$ and $\xi_R^2 = \mu_R^2/Q^2$ can be evaluated a posteriori.

The calculation of scale variations provided by `yadism` and the subsequent determination of the MHOU theory covariance matrix has been benchmarked with the results of Abdul Khalek et al. 2019b, finding good agreement.

BENCHMARKING. The DIS structure functions predictions provided by `yadism` have been thoroughly benchmarked with those from `APFEL` and `QCDNUM`. Specifically, we have verified that we can reproduce the `APFEL` predictions for those coefficient functions listed in table 1.1 which are available in `APFEL`. Excellent agreement is found in all cases considered, with some residual differences well understood as will be discussed in more detail in Candido, Hekhorn, et al. n.d. To illustrate this good agreement, fig. 1.3 displays the comparison of the `yadism` predictions for DIS structure functions at NNLO with the corresponding ones from `APFEL`. The same input theory settings are used in both calculations, in particular the PDFs (in this case `NNPDF4.0` NNLO), strong coupling constant ($\alpha_s(m_Z) = 0.118$), and GM-VFNS (FONLL-C). We display results for four bins of representative DIS datasets included in the `NNPDF4.0` global analysis: fixed-target Neutral Current DIS on a deuteron target from BCDMS, fixed-target Charged Current DIS on a lead target from CHORUS, collider neutral-current positron-proton DIS from HERA, and collider Charged Current electron-proton DIS from HERA. A similar level of agreement is obtained for other bins of DIS datasets.

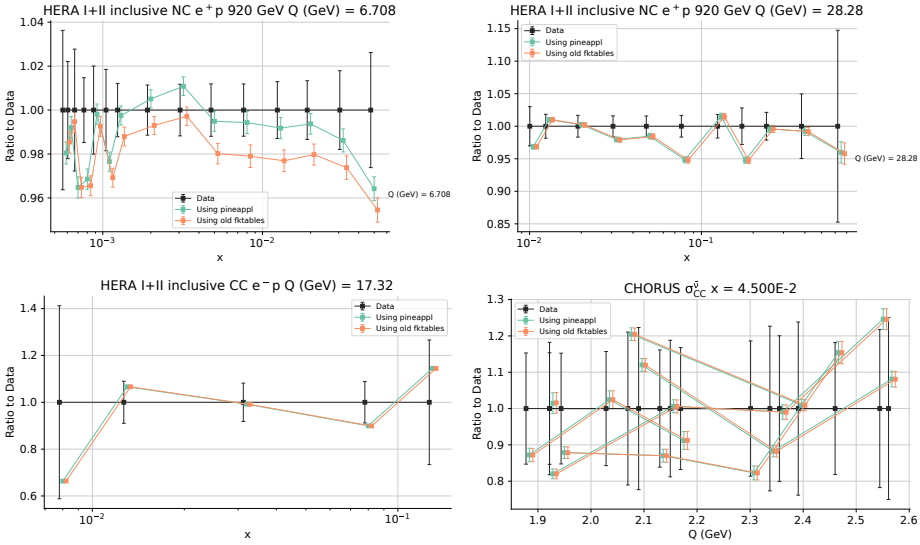

Figure 1.3: Comparison of the `yadism` predictions for DIS structure functions and reduced cross-sections at NNLO with the corresponding ones from `APFEL` for the same choice of input settings. We display predictions for four x bins of representative DIS datasets included in the NNPDF4.0 global analysis: fixed-target neutral-current DIS on a deuteron target from BCDMS, fixed-target charged-current DIS on a lead target from CHORUS, collider neutral-current positron-proton DIS from HERA, and collider charged-current electron-proton DIS from HERA.

2 EVOLUTION OPERATORS

As briefly introduced in section 0.3, a central element of a PDF fit consists in DGLAP evolution. During a fit, this will be used in two contexts:

i. upgrading grids to the so called FK tables (cf. chapter 3), to obtain theory predictions at many different scales (matching the experimental results), starting from the PDF candidate being minimized at a single fitting scale[1]

ii. evolving the result of the fit, to distribute PDF grids for the users already at all the potentially useful scales (interpolating among a finite sets of values)

[1] Or, equivalently, evolving online (i.e. during the fit) at all those scales.

While working on new PDF fits and related issues, a certain set of limitations of the existing DGLAP evolution libraries has been collected, and for this reason we decided to write a new QCD evolution library that matches the requirements of a PDF fitting collaboration, and similar fits as well.

We called it Evolution Kernel Operators (EKO), because the main focus is to compute such operators, which are independent of the initial boundary conditions, and only depend on the selected theory parameters. In such manner, the operators can be computed only once, stored on disk and then reused in the actual application. This method can lead to a significant speed-up when PDFs are repeatedly being evolved, as it is customary in PDF fits. This approach has been introduced by `FastKernel` Ball, Del Debbio, Forte, Guffanti, Latorre, Piccione, et al. 2009; Ball, Del Debbio, Forte, Guffanti, Latorre, Rojo, et al. 2010; Del Debbio, Forte, et al. 2007 and it is further developed here (with FK tables cited above being the output of `FastKernel`). Furthermore, we decided to solve the evolution equations in Mellin space (cf. section 2.1.1) to allow for simpler solution algorithms (both iterative and approximated). Yet, results are provided in momentum fraction space (cf. section 2.1.2) to allow an easy interface with existing codes.

`EKO` currently implements the leading order (LO), next-to-leading order (NLO) and next-to-next-to-leading order (NNLO) solutions J. Blümlein et al. 2021; Moch, Vermaseren, and Vogt 2004; Vogt, Moch, and Vermaseren 2004. However, it is organized in such a way that the addition of higher order corrections, such as the approximate next-to-next-to-next-to-leading order (N^3LO) Moch, Ruijl, et al. 2022, can be achieved with relative ease. This accuracy is needed to match the precision in the determination of the short-distance cross sections for several processes at LHC (see e.g. Duhr and Mistlberger 2022 and references therein). We also expose the associated variations of the various scales (their role in PDFs is exposed in chapter 4, and specifically in section 4.2).

The correct treatment of intrinsic heavy quark distributions is also properly taken into account. While the role of these distributions in the evolution equations is mathematically simple, as they decouple in a specific basis, their integration into the full solution, including matching conditions, is non-trivial. We implement backward evolution too, again including the non-trivial handling of all matching conditions. Both have been already used for studies of the heavy quark content of the nucleon Ball, Candido, Cruz-Martinez, et al. 2022, that is exposed in details in chapter 5.

It is also relevant to remark that `EKO` is another corner stone in a suite of tools that aims to provide new handles to the theory predictions in the PDF fitting procedure. This consists exactly in the calculation of those FK tables mentioned above, and described in chapter 3. But it is worth explicitly mentioning that `EKO` is also powering `yadism` Candido, Hekhorn, et al. n.d., the DIS coefficient function library described in chapter 1.

We adopted Python as the main development language, opting for a high-level one, which is easy to understand and learn for newcomers. In particular, with the advent of Data Science and Machine Learning, Python has become the lan-

guage of choice for many scientific applications, mainly driven by the large availability of packages and frameworks, and in particular some high-quality and widespread ones. We decided to write a code that can be used by everyone who needs QCD evolution, and to make it possible for applications that are not supported yet to be built on top of the shipped tools and ingredients. For this reason the code is developed mainly as a library, that contains physics, math, and algorithmic tools, such as those needed for managing or storing the computed operators. As an example we also expose a runner, making use of the built library to deliver an actual evolution application.

EKO is open-source, allowing easy interaction with users and developers. The project comes with a clean, modular, and maintainable codebase that guarantees easy inspection and ensures it is future-proof. The repository is publicly available, and located at:

https://github.com/N3PDF/eko

The EKO documentation contains the full API reference, tutorials to get started with EKO calculations, and far more complete discussion about the underlying theory and numerical techniques adopted. It can be accessed at:

https://eko.readthedocs.io/en/latest/

This document is also regularly updated and extended upon the implementation of new features.

2.1 THEORY OVERVIEW

The central equations that EKO is solving are the Dokshitzer–Gribov–Lipatov–Altarelli–Parisi evolution equations (DGLAP) evolution equations, whose full expression is:

$$\mu_F^2 \frac{d\mathbf{f}}{d\mu_F^2}(x, \mu_F^2) = \mathbf{P}\left(a_s(\mu^2), \frac{\mu_F^2}{\mu^2}\right) \otimes \mathbf{f}(\mu_F^2) \tag{2.1}$$

where $\mathbf{f}(x, \mu_F^2)$ is a vector of PDFs over flavor space with x the momentum fraction and μ_F^2 the factorization scale. The further scale μ^2 is the one associated to the running of the strong coupling, that can be displaced with respect to μ_F^2, usually keeping a constant ratio of the two scales. For this reason, it is directly involved in the solution of the differential equation (cf. section 2.1.3)[2].

The main ingredients to eq. (2.1) are the Altarelli-Parisi splitting functions $\mathbf{P}\left(a_s(\mu^2), x, \frac{\mu_F^2}{\mu^2}\right)$ Moch, Vermaseren, and Vogt 2004; Vogt, Moch, and Vermaseren 2004, which are matrices over the flavor space, and actually distributions in x (despite the name). Finally, $\otimes$ denotes the multiplicative (or Mellin) convolution.

[2] If $\mu^2 \neq \mu_F^2$ an explicit dependency also appears in the expressions for the splitting functions.

The splitting functions $\mathbf{P}\left(a_s(\mu^2), x, \frac{\mu_F^2}{\mu^2}\right)$ admit a perturbative expansion in the strong coupling $a_s(\mu^2)$:

$$\begin{aligned}\mathbf{P}\left(a_s(\mu^2), x, \frac{\mu_F^2}{\mu^2}\right) = a_s(\mu^2)\mathbf{P}^{(0)}\left(x, \frac{\mu_F^2}{\mu^2}\right) \\ + \left[a_s(\mu^2)\right]^2 \mathbf{P}^{(1)}\left(x, \frac{\mu_F^2}{\mu^2}\right) \\ + \left[a_s(\mu^2)\right]^3 \mathbf{P}^{(2)}\left(x, \frac{\mu_F^2}{\mu^2}\right) \\ + \mathcal{O}\left(\left[a_s(\mu^2)\right]^4\right)\end{aligned} \tag{2.2}$$

which is currently known at NNLO J. Blümlein et al. 2021; Moch, Vermaseren, and Vogt 2004; Vogt, Moch, and Vermaseren 2004 and is under investigation for N^{3}LO Moch, Ruijl, et al. 2022. In a first step, the scale μ, at which the running coupling a_s is evaluated, and the factorization scale μ_F can be assumed to be equal $\mu = \mu_F$. The variation of the ratio μ/μ_F can be considered as an estimate to Missing Higher Order Uncertainties (MHOU) Abdul Khalek et al. 2019c, as it quantifies the dependence of the physical observable value on the unphysical factorization scale (cf. chapter 4).

In order to solve eq. (2.1) a series of steps has to be taken, and we highlight these steps in the following sections.

2.1.1 Mellin space

The presence of the derivative on the left-hand-side and the convolution on the right-hand-side turns eq. (2.1) into a set of coupled integro-differential equations which are non-trivial to solve.

A possible strategy in solving eq. (2.1) is by tackling the problem head-on and iteratively solve the integrals and the derivative by taking small steps: we refer to this as "x-space solution", as the solution uses directly momentum space and this approach is adopted, e.g., by `APFEL` Bertone, Stefano Carrazza, and Rojo 2014, `HOPPET` Salam and Rojo 2009, and `QCDNUM` M. Botje 2011. However, this approach becomes quite cumbersome when dealing with higher-order corrections, as the solutions becomes more and more involved.

We follow a different strategy and apply the Mellin transformation $\mathcal{M}$

$$\tilde{g}(N) = \mathcal{M}[g(x)](N) = \int_0^1 dx\, x^{N-1} g(x) \tag{2.3}$$

where, as well here as in the following, we denote objects in Mellin space by a tilde. This approach is also adopted by `PEGASUS` Vogt 2005 and `FastKernel` Ball,

Del Debbio, Forte, Guffanti, Latorre, Piccione, et al. 2009; Ball, Del Debbio, Forte, Guffanti, Latorre, Rojo, et al. 2010; Del Debbio, Forte, et al. 2007. The numerically challenging step is then shifted to the treatment of the Mellin inverse $\mathcal{M}^{-1}$, as we eventually seek for results in x-space (cf. section 2.1.2).

2.1.2 Interpolation

Mellin space has the theoretical advantage that the analytical solution of the equations becomes simpler, but the practical disadvantage that it requires PDFs in Mellin space. This constraint is in practice a serious limitation since most matrix element generators Buckley et al. 2011 as well as the various generated coefficient function grids (e.g. PINEAPPL S. Carrazza et al. 2020b; Schwan et al. 2022a, `APPLgrid` Carli et al. 2010 and `FastNLO` Daniel Britzger et al. 2012) are not using Mellin space, but rather x-space.

This is bypassed in `PEGASUS` by parametrizing the initial boundary condition with up to six parameters in terms of the Euler beta function. However, this is not sufficiently flexible to accommodate more complex analytic forms, or even parametrizations in form of neural networks.

We are bypassing this limitation by introducing a Lagrange-interpolation Edward 1779; Süli and Mayers 2003 of the PDFs in x-space on arbitrarily user-chosen grids $\mathbb{G}$:

$$f(x) \sim \tilde{f}(x) = \sum_j f(x_j) p_j(x), \quad \text{with } x_j \in \mathbb{G} \tag{2.4}$$

For the usage inside the library we do an analytic Mellin transformation of the polynomials $\tilde{p}_j(N) = \mathcal{M}[p_j(x)](N)$. For the interpolation polynomials p_j we are choosing a subset with $N_{degree} + 1$ points of the interpolation grid $\mathbb{G}$ to avoid Runge's phenomenon Runge 1901; Süli and Mayers 2003 and to avoid large cancellation in the Mellin transform.

2.1.3 Strong coupling

The evolution of the strong coupling $a_s(\mu^2) = \alpha_s(\mu^2)/(4\pi)$ is given by its renormalization group equation (RGE):

$$\beta(a_s) = \mu^2 \frac{da_s(\mu^2)}{d\mu^2} = -\sum_{n=0} \beta_n \left[a_s(\mu^2)\right]^{2+n} \tag{2.5}$$

and is currently known at 5-loop (β_4) accuracy Baikov et al. 2017; K. G. Chetyrkin et al. 2017; Herzog et al. 2017; Luthe, Maier, Marquard, and York Schroder 2017; Luthe, Maier, Marquard, and Schröder 2016.

This is crucial for DGLAP solution, indeed, since the strong coupling a_s is a monotonic function of the scale μ in the perturbative regime, we can actually consider a transformation of eq. (2.1)

$$\frac{d\tilde{\mathbf{f}}}{da_s}(N, a_s) = -\frac{\gamma(N, a_s)}{\beta(a_s)} \tilde{\mathbf{f}}(N, a_s) \tag{2.6}$$

with $\boldsymbol{\gamma} = -\tilde{\mathbf{P}}$ the anomalous dimension and $\beta(a_s)$ the QCD beta function, where the multiplicative convolution is reduced to an ordinary product.

2.1.4 Flavor space

Next, we address the separation in flavor space: formally we can define the flavor space $\mathcal{F}$ as the linear span over all partons (which we consider to be the canonical one):

$$\mathcal{F} = \mathcal{F}_{\mathrm{fl}} = \mathrm{span}\left(g, u, \bar{u}, d, \bar{d}, s, \bar{s}, c, \bar{c}, b, \bar{b}, t, \bar{t}\right) \tag{2.7}$$

The splitting functions $\mathbf{P}$ become block-diagonal in the "Evolution Basis", a suitable decomposition of the flavor space: the singlet sector $\mathbf{P}_S$ remains the only coupled sector over $\{\Sigma, g\}$, while the full valence combination $P_{ns,v}$ decouples completely (i.e. it is only coupling to V), and the non-singlet singlet-like sector $P_{ns,+}$ is diagonal over $\{T_3, T_8, T_{15}, T_{24}, T_{35}\}$, and the non-singlet valence-like sector $P_{ns,-}$ is diagonal over $\{V_3, V_8, V_{15}, V_{24}, V_{35}\}$. The respective distributions are given by their usual definition.

This Evolution Basis is isomorphic to our canonical choice

$$\mathcal{F} \sim \mathcal{F}_{ev} = \mathrm{span}(g, \Sigma, V, T_3, T_8, T_{15}, T_{24}, T_{35}, V_3, V_8, V_{15}, V_{24}, V_{35}) \tag{2.8}$$

but, it is not a normalized basis. When dealing with intrinsic evolution, i.e. the evolution of PDFs below their respective mass scale, the Evolution Basis is not sufficient. In fact, for example, $T_{15} = u^+ + d^+ + s^+ - 3c^+$ below the charm threshold μ_c^2 contains both running and static distributions which need to be further disentangled.

We are thus considering a set of "Intrinsic Evolution Bases" $\mathcal{F}_{iev,n_f}$, where we retain the intrinsic flavor distributions as basis vectors. The basis definition depends on the number of light flavors n_f and, e.g. for $n_f = 4$, we find

$$\mathcal{F} \sim \mathcal{F}_{iev,4} = \mathrm{span}(g, \Sigma_{(4)}, V_{(4)}, T_3, T_8, T_{15}, V_3, V_8, V_{15}, b^+, b^-, t^+, t^-) \tag{2.9}$$

with $\Sigma_{(4)} = \sum_{j=1}^{4} q_j^+$ and $V_{(4)} = \sum_{j=1}^{4} q_j^-$.

2.1.5 Solution Strategies

The formal solution of eq. (2.6) in terms of evolution kernel operators $\tilde{\mathbf{E}}$ is given by

$$\tilde{\mathbf{E}}(a_s \leftarrow a_s^0) = \mathcal{P} \exp\left[-\int_{a_s^0}^{a_s} \frac{\boldsymbol{\gamma}(a_s')}{\beta(a_s')} \mathrm{d}a_s'\right] \tag{2.10}$$

with $\mathcal{P}$ the path-ordering operator. If the anomalous dimension $\boldsymbol{\gamma}$ is diagonal in flavor space, i.e. it is in the non-singlet sector, it is always possible to find an analytical solution to eq. (2.10). In the singlet sector sector, however, this is only true

at LO and to obtain a solution beyond, we need to apply different approximations and solution strategies, on which EKO offers currently eight implementations. For an actual comparison of selected strategies, cf. section 2.2.2.

2.1.6 Matching at Thresholds

EKO can perform calculation in a Fixed Flavor Number Scheme (FFNS) where the number of active or light flavors n_f is constant. This means that both the beta function $\beta^{(n_f)}(a_s)$ and the anomalous dimension $\gamma^{(n_f)}(a_s)$ in eq. (2.6) are constant with respect to n_f. However, this approximation is likely to fail either in the high energy region $\mu_F^2 \to \infty$ for a small number of active flavors, or to fail in the low energy region $\mu_F^2 \to \Lambda^2_{QCD}$ for a large number of active flavors.

This can be overcome by using a Variable Flavor Number Scheme (VFNS) that changes the number of active flavors when the scale μ_F^2 crosses a threshold μ_h^2. This then requires a matching procedure when changing the number of active flavors, and for the PDFs we find

$$\begin{aligned}\tilde{\mathbf{f}}^{(n_f+1)}(\mu^2_{F,1}) = \tilde{\mathbf{E}}^{(n_f+1)}(\mu^2_{F,1} \leftarrow \mu^2_h)\mathbf{R}^{(n_f)}\tilde{\mathbf{A}}^{(n_f)}(\mu^2_h)\tilde{\mathbf{E}}^{(n_f)}(\mu^2_h \leftarrow \mu^2_{F,0}) \\ \times \tilde{\mathbf{f}}^{(n_f)}(\mu^2_{F,0})\end{aligned} \tag{2.11}$$

where the superscript refers to the number of active flavors and we split the matching factor into two parts: the perturbative Operator Matrix Element (OME) $\tilde{\mathbf{A}}^{(n_f)}(\mu_h^2)$, currently implemented at NNLO Buza, Matiounine, Smith, and W. L. van Neerven 1998a, and an algebraic rotation $\mathbf{R}^{(n_f)}$ acting only in the flavor space $\mathcal{F}$.

For backward evolution this matching has to be applied in the reversed order. The inversion of the basis rotation matrices $\mathbf{R}^{(n_f)}$ is simple, whereas this is not true for the OME $\tilde{\mathbf{A}}^{(n_f)}$ especially in case of NNLO or higher order evolution. In EKO we have implemented two different strategies to perform the inverse matching: the first one is a numerical inversion, where the OMEs are inverted exactly in Mellin space, while in the second method, called `expanded`, the matching matrices are inverted through a perturbative expansion in a_s, given by:

$$\begin{aligned}\left(\tilde{\mathbf{A}}^{(n_f)}\right)^{-1}_{exp}(\mu_h^2) = \mathbf{I} - a_s(\mu_h^2)\tilde{\mathbf{A}}^{(n_f),(1)} \\ + a_s^2(\mu_h^2)\left[\tilde{\mathbf{A}}^{(n_f),(2)} - \left(\tilde{\mathbf{A}}^{(n_f),(1)}\right)^2\right] \\ + O(a_s^3)\end{aligned} \tag{2.12}$$

with $\mathbf{I}$ the identity matrix in flavor space.

2.1.7 Running Quark Masses

In the context of PDF evolution, the most used treatment of heavy quarks masses are the pole masses, where the physical values are specified as input

and do not depend on any scale. However for specific applications, such as the determination of MHOU due to heavy quarks contribution inside the proton Ball, Bertone, Bonvini, Stefano Carrazza, et al. 2016, $\overline{\text{MS}}$ masses can also be used. In particular, in S. Alekhin and Moch 2011 it is found that higher order corrections on heavy quark production in DIS are more stable upon scale variation when using the $\overline{\text{MS}}$ scheme. `EKO` allows for this option as it is discussed briefly in the following paragraphs.

Whenever the initial condition for the mass is not given at a scale coinciding with the mass itself (i.e. $\mu_{h,0} \neq m_{h,0}$, being $m_{h,0}$ the given initial condition at the scale $\mu_{h,0}$), `EKO` computes the scale at which the running mass $m_h(\mu_h^2)$ intersects the identity function. Thus for each heavy quark h we solve:

$$m_{\overline{\text{MS}},h}(m_h^2) = m_h \tag{2.13}$$

The value $m_h(m_h)$ is then used as a reference to define the evolution thresholds.

The evolution of the $\overline{\text{MS}}$ mass is given by:

$$m_{\overline{\text{MS}},h}(\mu_h^2) = m_{h,0} \exp\left[-\int_{a_s(\mu_{h,0}^2)}^{a_s(\mu_h^2)} \frac{\gamma_m(a_s')}{\beta(a_s')}\, da_s'\right] \tag{2.14}$$

with $\gamma_m(a_s)$ the QCD anomalous mass dimension available up to N^3LO K. Chetyrkin et al. 2006; Y. Schroder and Steinhauser 2006; Vermaseren, Larin, et al. 1997.

Note that to solve eq. (2.14) $a_s(\mu^2)$ must be evaluated in a FFNS until the threshold scales are known. Thus it is important to start computing the $\overline{\text{MS}}$ masses of the quarks which are closer to the the scale μ_0 at which the initial reference value $a_s(\mu_0^2)$ is given.

Furthermore, to find consistent solutions the boundary condition of the $\overline{\text{MS}}$ masses must satisfy $m_h(\mu_h) \geqslant \mu_h$ for heavy quarks involving a number of active flavors greater than the number of quark flavors $n_{f,0}$ at μ_0, implying that we find the intercept between the RGE and the identity in the forward direction ($m_{\overline{\text{MS}},h} \geqslant \mu_h$). The opposite holds for scales related to fewer active flavors.

2.2 BENCHMARKING AND VALIDATION

Although `EKO` is totally PDF independent, for the sake of plotting we choose NNPDF4.0 Ball et al. 2022a as a default choice for our plots, but for section 2.2.1 where we choose the toy PDF of the Les Houches Benchmarks Dittmar et al. 2005; Giele et al. 2002. We show the gluon distribution $g(x)$ as a representative member of the singlet sector and the valence distribution $V(x)$ as a representative member of the non-singlet sector. Note that PDFs in the same sector have mostly the same behavior, apart from some specific regions (e.g. the T_{15} distribution right after charm matching).

2.2.1 Benchmarks

In this section we present the outcome of the benchmarks between EKO and similar available tools assuming different theoretical settings.

Les Houches Benchmarks

EKO has been compared with the benchmark tables given in Dittmar et al. 2005; Giele et al. 2002. We find a good match except for a list of typos which we list here:

- in table head in Giele et al. 2002 should be $2xL_+ = 2x(\bar{u} + \bar{d})$
- in the head of table 1: the value for α_s in FFNS is wrong (as pointed out and corrected in Dittmar et al. 2005)
- in table 3, part 3 of Giele et al. 2002: $xL_-(x = 10^{-4}, \mu_F^2 = 10^4\,\mathrm{GeV}^2) = 1.0121 \cdot 10^{-4}$ (wrong exponent) and $xL_-(x = 0.1, \mu_F^2 = 10^4\,\mathrm{GeV}^2) = 9.8435 \cdot 10^{-3}$ (wrong exponent)
- in table 15, part 1 of Dittmar et al. 2005: $xd_v(x = 10^{-4}, \mu_F^2 = 10^4\,\mathrm{GeV}^2) = 1.0699 \cdot 10^{-4}$ (wrong exponent) and $xg(x = 10^{-4}, \mu_F^2 = 10^4\,\mathrm{GeV}^2) = 9.9694 \cdot 10^{2}$ (wrong exponent)

Some of these typos have been already reported in Diehl et al. 2022.

In fig. 2.1 we present the results of the VFNS benchmark at NNLO, where a toy PDF is evolved from $\mu_{F,0}^2 = 2\,\mathrm{GeV}^2$ up to $\mu_F^2 = 10^4\,\mathrm{GeV}^2$ with equal values of the factorization and process scales $\mu_F = \mu$. For completeness, we display the singlet $S(x)$ and gluon $g(x)$ distribution (top), the singlet-like $T_{3,8,15,24}(x)$ (middle) and the valence $V(x)$, valence-like $V_3(x)$ (bottom) along with the results from APFEL and PEGASUS. We find an overall agreement at the level of $O(10^{-3})$.

APFEL

APFEL Bertone, Stefano Carrazza, and Rojo 2014 is one of the most extensive tool aimed to PDF evolution and DIS observables calculation. It is provided as a Fortran library, and it has been used by the NNPDF collaboration up to NNPDF4.0 Ball et al. 2022a.

APFEL solves DGLAP numerically in x-space, sampling the evolution kernels on a grid of points up to NNLO in QCD, with QED evolution also available at LO. By construction this method is PDF dependent and the code is automatically interfaced with LHAPDF Buckley et al. 2015. For specific application, the code offers also the possibility to retrieve the evolution operators with a dedicated function.

The program supplies three different solution strategies, with various theory setups, including scale variations and $\overline{\mathrm{MS}}$ masses.

The stability of our benchmark at different perturbative orders is presented in fig. 2.2, using the settings of the Les Houches PDF evolution benchmarks Dittmar

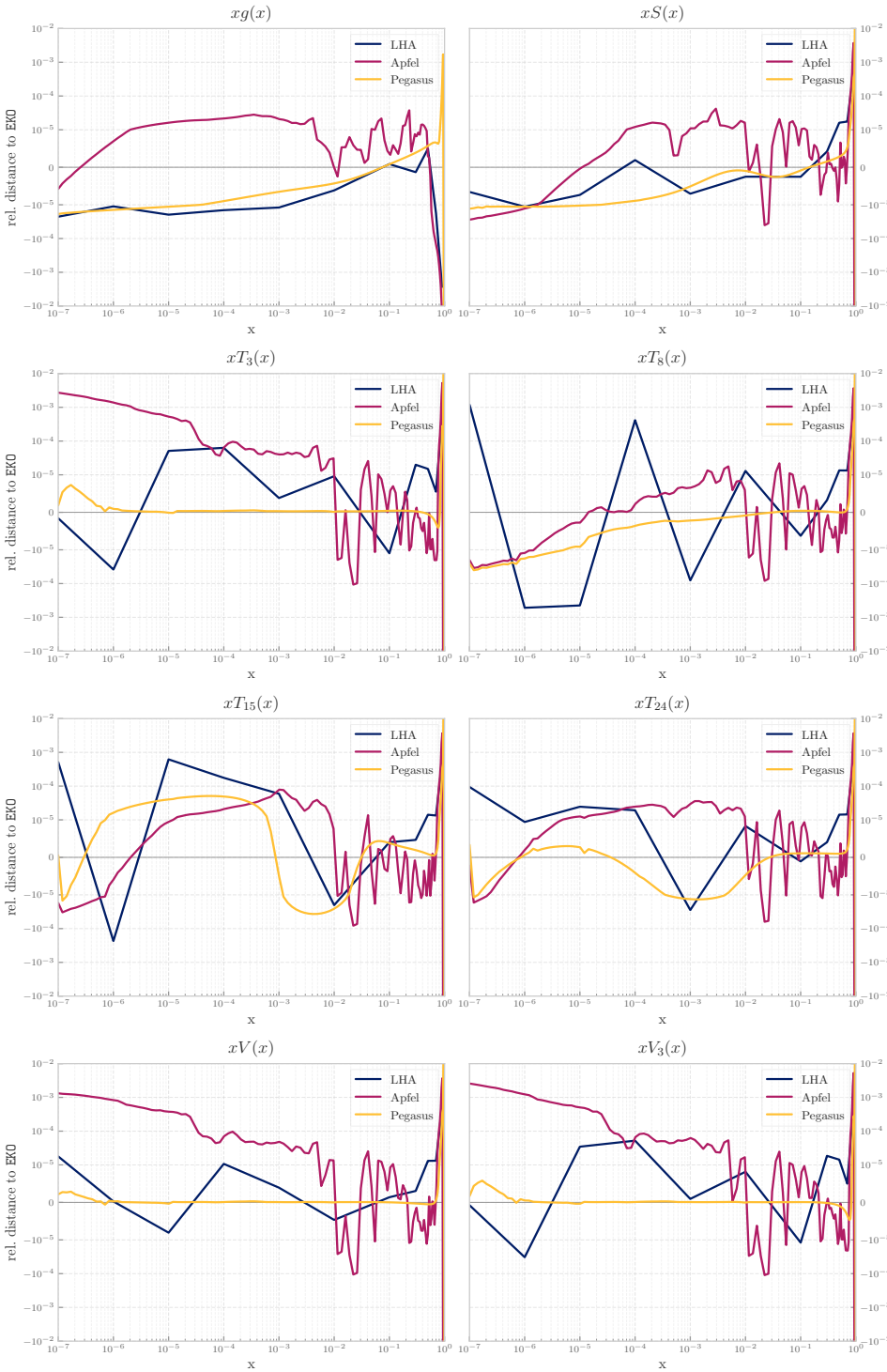

Figure 2.1: Relative differences between the outcome of NNLO QCD evolution as implemented in `EKO` and the corresponding results from Dittmar et al. 2005, `APFEL` Bertone, Stefano Carrazza, and Rojo 2014 and `PEGASUS` Vogt 2005. We adopt the settings of the Les Houches PDF evolution benchmarks Dittmar et al. 2005; Giele et al. 2002.

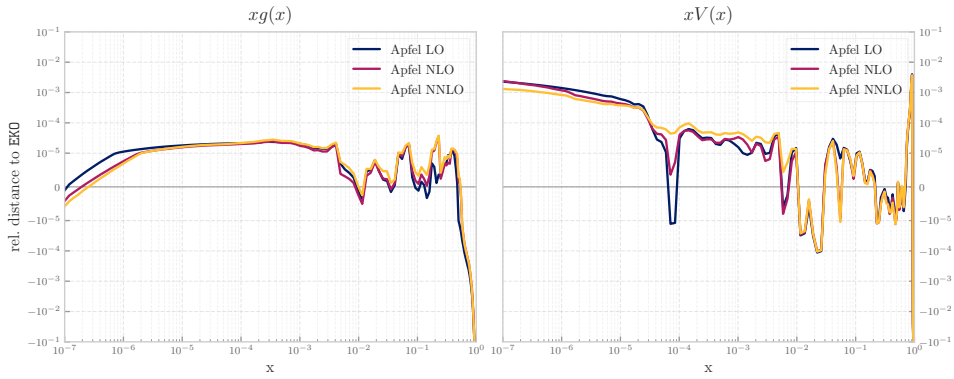

Figure 2.2: Relative differences between the outcome of evolution as implemented in EKO and the corresponding results from APFEL at different perturbative orders. We adopt the same settings of fig. 2.1.

et al. 2005; Giele et al. 2002. The accuracy of our comparison is not affected by the increasing complexity of the calculation.

PEGASUS

PEGASUS Vogt 2005 is a Fortran program aimed for PDF evolution. The program solves DGLAP numerically in N-space up to NNLO. PEGASUS can only deal with PDFs given as a fixed functional form and is not interfaced with LHAPDF.

As shown in fig. 2.1, the agreement of EKO with this tool is better than with APFEL, especially for valence-like quantities, for which an exact solution is possible, where we reach $\mathcal{O}(10^{-6})$ relative accuracy. This is expected and can be traced back to the same DGLAP solution strategy in Mellin space.

Similarly to the APFEL benchmark, we assert that the precision of our benchmark with PEGASUS is not affected by the different QCD perturbative orders, as visible in fig. 2.3. As both, APFEL and PEGASUS, have been benchmarked against HOPPET Salam and Rojo 2009 and QCDNUM M. Botje 2011 we conclude to agree also with these programs.

2.2.2 Solution Strategies

As already mentioned in section 2.1.5, due to the coupled integro-differential structure of eq. (2.1), solving the equations requires in practice some approximations to which we refer as different solution strategies. EKO currently implements 8 different strategies, corresponding to different approximations. Note that they may differ only by the strategy in a specific sector, such as the singlet or non-singlet sector. All the strategies provided agree at fixed order, but differ by higher order terms.

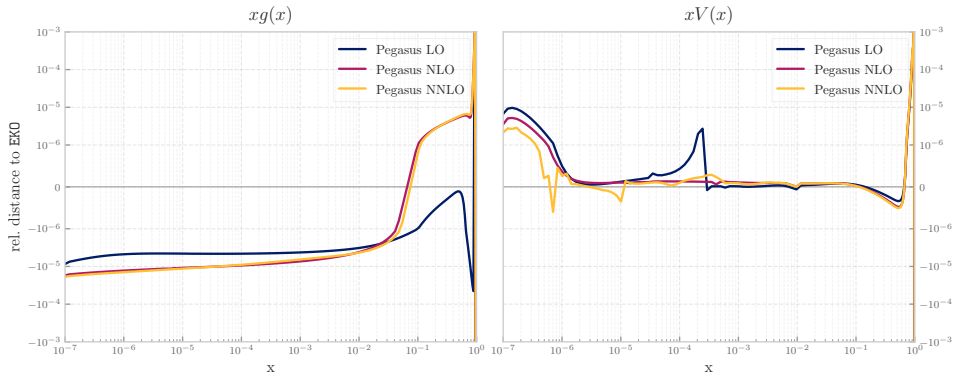

Figure 2.3: Same of fig. 2.2, now comparing to `PEGASUS` Vogt 2005.

In fig. 2.4 we show a comparison of a selected list of solution strategies[3]:

- `iterate-exact`: In the non-singlet sector we take the analytical solution of eq. (2.6) up to the order specified. In the singlet sector we split the evolution path into segments and linearize the exponential in each segment Bonvini 2012. This provides effectively a straight numerical solution of eq. (2.6). In fig. 2.4 we adopt this strategy as a reference.
- `perturbative-exact`: In the non-singlet sector it coincides with `iterate-exact`. In the singlet sector we make an ansatz to determine the solution as a transformation $\mathbf{U}(a_s)$ of the LO solution Vogt 2005. We then iteratively determine the perturbative coefficients of $\mathbf{U}$.
- `iterate-expanded`: In the singlet sector we follow the strategy of `iterate-exact`. In the non-singlet sector we expand eq. (2.6) first to the order specified, before solving the equations.
- `truncated`: In both sectors, singlet and non-singlet, we make an ansatz to determine the solution as a transformation $\mathbf{U}(a_s)$ of the LO solution and then expand the transformation $\mathbf{U}$ up to the order specified. Note that for programs using x-space this strategy is difficult to pursue as the LO solution is kept exact and only the transformation $\mathbf{U}$ is expanded.

The strategies differ most in the small-x region where the PDF evolution is enhanced and the treatment of sub-leading corrections become relevant. This feature is seen most prominently in the singlet sector between `iterate-exact` (the reference strategy) and `truncated`. In the non-singlet sector the distributions also vanish for small-x and so the difference gets artificially enhanced. This is eventually the source of the spread for the valence distribution $V(x)$ making it more sensitive to the initial PDF.

[3] For the full list of available solutions and a detailed descriptions see the online documentation.

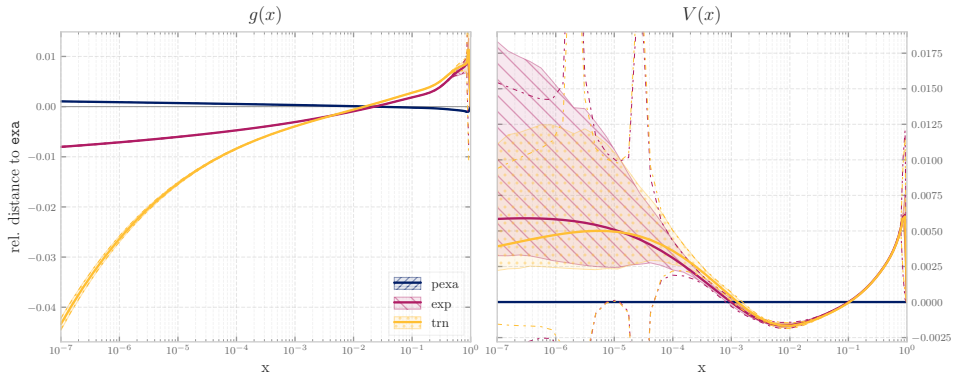

Figure 2.4: Compare selected solutions strategies, with respect to the `iterated-exact` (called `exa` in label) one. In particular: `perturbative-exact` (`pexa`) (matching the reference in the non-singlet sector), `iterated-expanded` (`exp`), and `truncated` (`trn`). The distributions are evolved in $\mu_F^2 = 1.65^2 \to 10^4\ \text{GeV}^2$.

PDF PLOTS The PDF plot shown in fig. 2.4 contains multiple elements, and its layout is in common with figs. 2.5 and 2.7.

All the different entries corresponds to different theory settings, and they are normalized with respect to a reference theory setup (e.g. in fig. 2.4 the `iterative-exact` strategy) and the lines correspond to the relative difference.

Furthermore, an envelope and dashed lines are displayed. To obtain them, the full PDF set is evolved, replica by replica, for each configuration (corresponding to a single evolution operator, that is applied to each replica in turn). Then ratios are taken between corresponding evolved replicas, to highlight the PDF independence of `EKO` rather then any specific set-related features. The upper and lower borders of the envelope correspond respectively to the 0.16 and 0.84 quantiles of the replicas set, while the dashed lines correspond to one standard deviation.

2.2.3 Interpolation

To bridge between the desired x-space input/output and the internal Mellin representation, we do a Lagrange-Interpolation as sketched in section 2.1.2 (and detailed in the online documentation). We recommend a grid of at least 50 points with linear scaling in the large-x region ($x \gtrapprox 0.1$) and with logarithmic scaling in the small-x region and an interpolation of degree four. Also the grids determined by `aMCfast` Bertone, Rikkert Frederix, et al. 2014 perform sufficiently well for specific processes.

For a first qualitative study we show in fig. 2.5 a comparison between an increasing number of interpolation points distributed according to S. Carrazza et al. 2020b, Eq. 2.12. The separate configurations are converging to the solution with the largest number of points. Using 60 interpolation points is almost indistinguishable from using 120 points (the reference configuration in the plot). In the

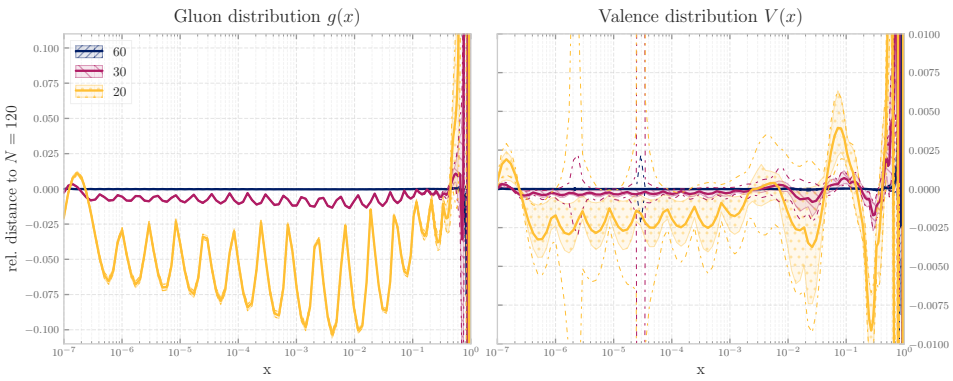

Figure 2.5: Relative differences between the outcome of NNLO QCD evolution as implemented in EKO with 20, 30, and 60 points to 120 interpolation points respectively.

singlet sector (gluon) the convergence is significantly slower due to the more involved solution strategies and, specifically, the oscillating behavior is caused due to these difficulties. The spikes for $x \to 1$ are not relevant since the PDFs are intrinsically small in this region ($\mathbf{f} \to 0$) and thus small numerical differences are enhanced.

Also note that the results of section 2.2.1 (i.e. figs. 2.1, 2.2 and 2.3) confirm that the interpolation error can be kept below the benchmark accuracy.

2.2.4 Matching

We refer to the specific value of the factorization scale at which the number of active flavors is changing from n_f to $n_f + 1$ (or vice-versa) as the threshold μ_h. Although this value usually coincides with the respective quark mass m_h, EKO implements the explicit expressions when the two scales do not match. This variation can be used to estimate MHOU Abdul Khalek et al. 2019c.

In fig. 2.6 we show the strong coupling evolution $a_s(\mu^2)$ around the bottom mass with the bottom threshold μ_b^2 eventually not coinciding with the respective bottom quark mass m_b^2. The dependency on the LO evolution is only due to the change of active flavor in the beta function ($\beta_0 = \beta_0(n_f)$), which can be seen in the ratio plot by the continuous connections of the lines. At NLO evolution the matching condition already becomes discontinuous for $\mu_h^2 \neq m_h^2$, represented in the ratio plot by the offset for the matched evolution. The matching for the NNLO evolution K. Chetyrkin et al. 2006; Y. Schroder and Steinhauser 2006 is intrinsically discontinuous, which is indicated in the ratio plot by the discrete jump at the central scale $\mu^2 = m_b^2$. For $\mu^2 > 2m_b^2$ the evolution is only determined by the reference value $a_s(m_Z^2)$ and the perturbative evolution order. For $\mu^2 < m_b^2/2$ we can observe the perturbative convergence as the relative difference shrinks

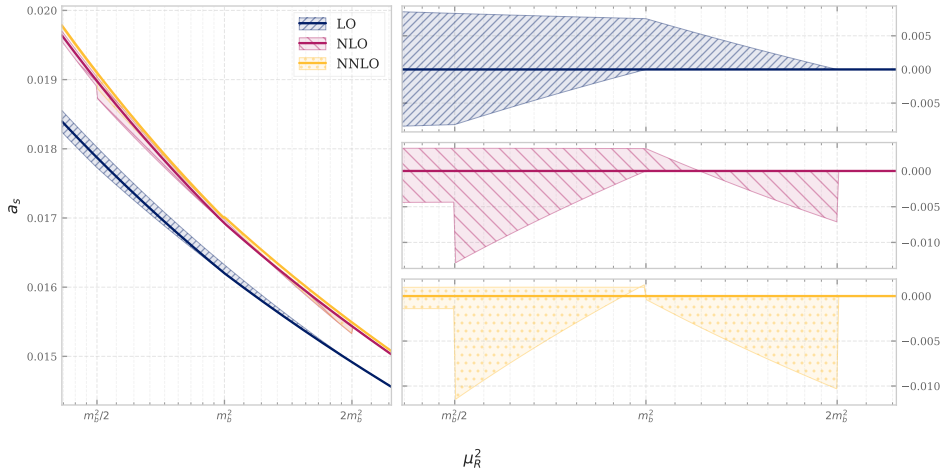

Figure 2.6: Strong coupling evolution $a_s(\mu^2)$ at LO, NLO and NNLO respectively with the bottom matching μ_b^2 at $1/2, 1$, and 2 times the bottom mass m_b^2 indicated by the band. In the left panel we show the absolute value, while on the right we show the ratio towards the central scale choice.

with increasing orders. Since it is converging, the effect of the matching condition should cancel more and more exactly with the difference in running, but the magnitude of both is increasing with the order, since the perturbative expansion of the beta function $\beta(a_s)$ is a same sign series.

In fig. 2.7 we show the relative difference for the PDF evolution with threshold values μ_h^2 that do not coincide with the respective heavy quark masses m_h^2. When matching at a lower scale the difference is significantly more pronounced as the evolution includes a region where the strong coupling varies faster. When dealing with $\mu_h^2 \neq m_h^2$ the PDF matching conditions become discontinuous already at NLO Buza, Matiounine, Smith, and W. L. van Neerven 1998a. These contributions are also available in `APFEL` Bertone, Stefano Carrazza, and Rojo 2014, but not in `PEGASUS` Vogt 2005 and although they are present in the code of `QCDNUM` M. Botje 2011 they can not be accessed there. For the study in Ball, Candido, Cruz-Martinez, et al. 2022 we also implemented the PDF matching at N^3LO Ablinger, Behring, J. Blümlein, De Freitas, Hasselhuhn, et al. 2014; Ablinger, Behring, J. Blümlein, De Freitas, von Manteuffel, and Schneider 2015; Ablinger, Blumlein, et al. 2011; Ablinger, J. Blümlein, De Freitas, Hasselhuhn, von Manteuffel, Round, and Schneider 2014; Ablinger, J. Blümlein, De Freitas, Hasselhuhn, von Manteuffel, Round, Schneider, and Wißbrock 2014; Behring et al. 2014; Bierenbaum et al. 2009a,b; Johannes Blümlein et al. 2017.

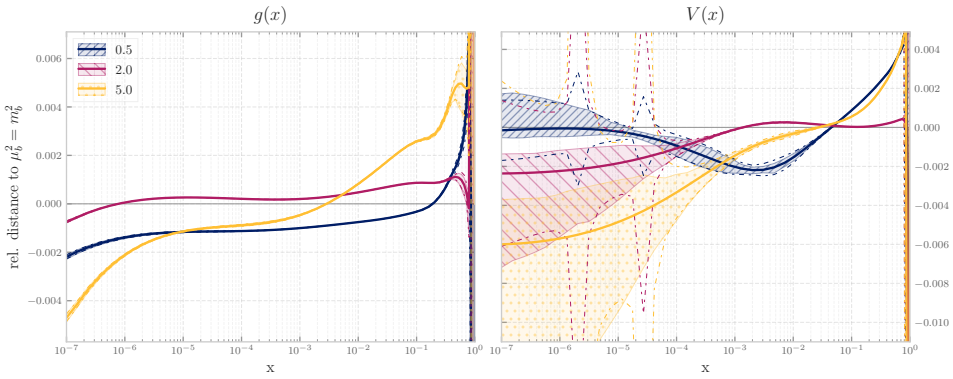

Figure 2.7: Difference of PDF evolution with the bottom matching μ_b^2 at $1/2, 2$, and 5 times the bottom mass m_b^2 relative to $\mu_b^2 = m_b^2$. Note the different scale for the two distributions. All evolved in $\mu_F^2 = 1.65^2 \to 10^4\,\text{GeV}^2$.

2.2.5 Backward

As a consistency check we have performed a closure test verifying that after applying two opposite `EKOs` to a custom initial condition we are able to recover the initial PDF. Specifically, the product of the two kernels is an identity both in flavor and momentum space up to the numerical precision. The results are shown in fig. 2.8 in case of NNLO evolution crossing the bottom threshold scale $\mu_F = m_b$. The differences between the two inversion methods are more visible for singlet-like quantities, because of the non-commutativity of the matching matrix $\tilde{\mathbf{A}}_S^{(n_f)}$.

Special attention must be given to the heavy quark distributions which are always treated as intrinsic, when performing backward evolution. In fact, if the initial PDF (above the mass threshold) contains an intrinsic contribution, this has to be evolved below the threshold otherwise momentum sum rules can be violated. This intrinsic component is then scale independent and fully decoupled from the evolving (light) PDFs. On the other hand, if the initial PDF is purely perturbative, it vanishes naturally below the mass threshold scale after having applied the inverse matching. In this context, `EKO` has been used in a recent study to determine, for the first time, the intrinsic charm content of the proton Ball, Candido, Cruz-Martinez, et al. 2022.

2.2.6 $\overline{\text{MS}}$ masses

In fig. 2.9 we investigate the effect of adopting a running mass scheme onto the respective PDF sets. The left panel shows the $T_{15}(x)$ distribution obtained from the NNPDF4.0 perturbative charm determination Ball et al. 2022a using the pole mass scheme and the $\overline{\text{MS}}$ scheme, respectively. The distributions have been evolved on $\mu_F^2 = 1 \to 10^4\,\text{GeV}^2$. The mass reference values are taken from de

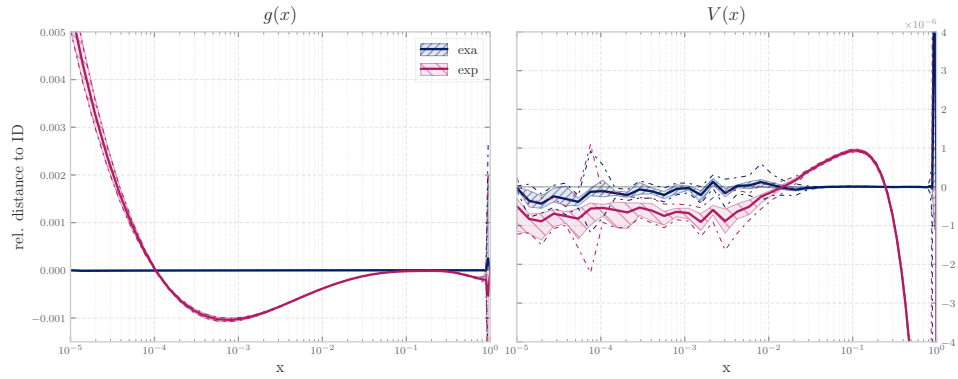

Figure 2.8: Relative distance of the product of two opposite NNLO EKOs and the identity matrix, in case of exact inverse and expanded matching (cf. eq. (2.12)) when crossing the bottom threshold scale $\mu_b^2 = 4.92^2\,\mathrm{GeV}^2$. In particular the lower scale is chosen $\mu_F^2 = 4.90^2\,\mathrm{GeV}^2$, while the upper is equal to $\mu_F^2 = 4.94^2\,\mathrm{GeV}^2$,

Florian et al. 2016b, with the $\overline{\mathrm{MS}}$ boundary condition on the charm mass given as $m_c(\mu_m = 3\,\mathrm{GeV}) = 0.986\,\mathrm{GeV}$, leading to $m_c(m_c) = 1.265\,\mathrm{GeV}$, while the charm pole mass is $m_c^{\mathrm{pole}} \approxeq 1.51\,\mathrm{GeV}$ Ball et al. 2022a. The major differences are visible in the low-x region where the DGLAP evolution is faster and the differences between the charm mass treatment are enhanced: an higher value of the charm mass increases the singlet like distribution $T_{15}(x)$. For the sake of comparison, in the right panel, we plot the relative distance to our result when comparing with the APFEL Bertone, Stefano Carrazza, and Rojo 2014 implementation. As expected the pole mass results are closer due to the same value of the charm mass, while the $\overline{\mathrm{MS}}$ results have a slightly bigger discrepancy which remains in all the x-range around 1‰ accuracy.

In fig. 2.10 we show the evolution of the $\overline{\mathrm{MS}}$ bottom mass $m_b(\mu_m^2)$ using different matching scales μ_b^2 equal to 1/2, 1 and 2 times the mass m_b^2, for each perturbative order (LO, NLO, and NNLO). The curve for $\mu_b^2 = m_b^2$ has been plotted as the central one (bold), while the other two are used as the upper and lower borders of the shaded area (according to their value, point by point). The reference value $m_b(\mu_{b,0}^2)$, has been chosen equal for the three curves, and it has been chosen at $m_b(m_b) = 4.92\,\mathrm{GeV}$. For this reason, above the central matching point $\mu_m^2 \geqslant m_b^2$ two curves coincide ($\mu_b^2 = m_b^2$ and $\mu_b^2 = m_b^2/2$) since they are both running with the same number of flavors ($n_f = 5$) and they have the same border condition. The curve using $\mu_b^2 = 2m_b^2$, however, still runs with a smaller number of flavors ($n_f = 4$) and so does not match the former two. In the lower region $\mu_m^2 < m_b^2$ this is not happening, because even if the number of flavors is now the same, the border condition is specified above matching for $\mu_b^2 = m_b^2$ (in $n_f = 5$). So, starting from m_b^2 and going downward, the central choice $\mu_b^2 = m_b^2$

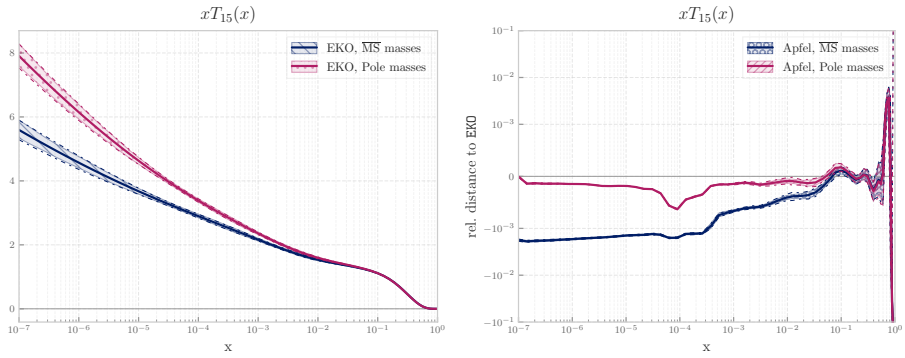

Figure 2.9: (left) The NNPDF4.0 perturbative charm distribution $T_{15}(x)$ Ball et al. 2022a with $\overline{\text{MS}}$ and pole masses NNLO evolution when running on $\mu_F^2 = 1 \to 10^4\,\text{GeV}^2$. (right) Relative difference to `EKO` for the same run with `APFEL` Bertone, Stefano Carrazza, and Rojo 2014.

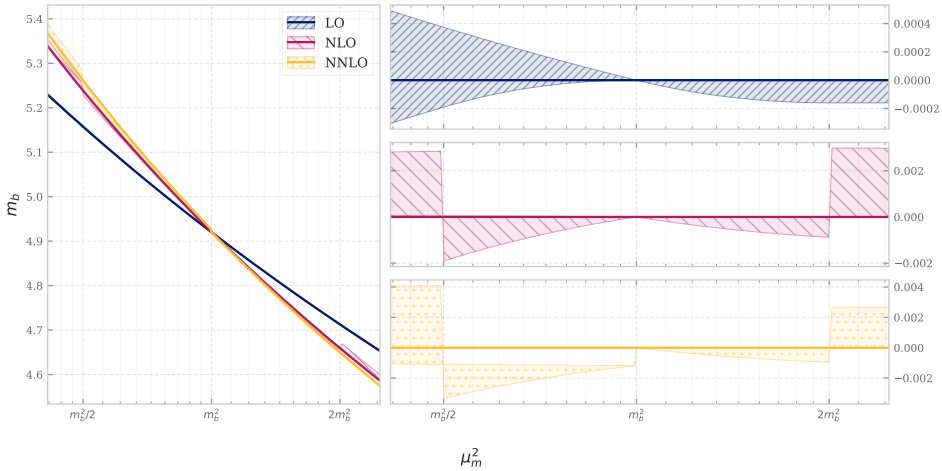

Figure 2.10: Running of the bottom quark mass $m_b(\mu_m^2)$ for different threshold ratios, similar to fig. 2.6. The plot shows how the different choices of matching scales affect the running in the matching region (and slightly beyond) at LO, NLO, and NNLO. The border condition for the running has been chosen at $m_b(m_b) = 4.92\,\text{GeV}$, as it is clear from the plot, since it is the intersection point of all the curves shown.

is matched first and then evolved, while the higher scale choice $\mu_b^2 = 2m_b^2$ immediately runs with four light flavors at m_b^2. Thus the difference consists just in the matching.

2.3 TECHNICAL OVERVIEW

An EKO is effectively a rank 5 tensor $E_{\mu,i\alpha j\beta}$, that evolves a PDF set from one given scale to a user specified list of final scales μ:

$$f_{\mu,i\alpha} = E_{\mu,i\alpha j\beta} f_{j\beta}^{(0)} \tag{2.15}$$

where i and j are indices on flavor, and α and β are indices on the x-grid.

The computation of each rank 4 tensor is almost independent: In a FFNS for each target μ_F^2 an operator $\tilde{\mathbf{E}}(\mu_{F,0}^2 \to \mu_F^2)$ is computed separately. Instead, in a VFNS first a set of operators is computed, to evolve from the initial scale to any matching scales (we call these *threshold operators*). Then, for each target μ_F^2, an operator is computed from the last intermediate matching scale to μ_F^2; finally they are composed together.

2.3.1 Performance Motivations and Operator Specificity

Before diving into the details of EKO performances there is a fundamental point that has to be taken into account: EKO is somehow unique as an evolution program, because its main and only output consists in evolution *operators*.

For this reason, a close comparison on performances with other evolution codes (whose main purpose is the evolution of individual PDFs) would be rather unfair: evolving a single PDF is comparable to the generation of the transformation of a single direction in the PDF space, while the operator acts on the whole function space.

The motivation to primarily look for the operator itself relies on the specific needs of a PDF fit itself. Indeed, a fit requires repeated usage of evolution for the χ^2 evaluation for each fit candidate, and a final evolution step for the generation of the PDF grids to deliver, as those used by LHAPDF Buckley et al. 2015. The first step has been automated long ago, by the generation of the FastKernel tables (formerly done with APFEL evolution, through APFELcomb, inspired to Bertone, Stefano Carrazza, and N. P. Hartland 2017), that store PDF evolution into the grids for predictions, while the second was repeated at any fit, since for each fit is a one-time operation (even though it is actually repeated for the number of Monte Carlo replicas, or Hessian eigenvectors, whose typical sizes are reported in table 2.1).

Actually, both the operations of including evolution in theory grids and PDF grids generation can be further optimized, considering that the evolution only depends on a small number of theory parameters, and so the operator does, such that it can be generated only once, and then used over and over.

PDF set name	members
`CT18NNLO` Hou et al. 2021	59
`MSHT20nnlo_as118` Bailey et al. 2021	65
`NNPDF40_nnlo_as_01180` Ball et al. 2022a	101

Table 2.1: Selected PDF sets with their respective number of members

On top of replicas generation, the search towards an optimal fitting methodology is an iterative process, that involves a large number number of fits. Moreover, whatever program supports the generation of `FastKernel` tables has to create some kind of evolution operator on its own (since the goal of `FastKernel` tables is exactly to be PDF agnostic).

So, the `EKO` work-flow is not a complete novelty, since it was preceded by `APFEL` in-memory operator generation, but it is a further and strong improvement in that direction: being operator-oriented from the beginning, optimizations have been performed for this specific task[4], and maintaining an actual operator format, the operators reuse is possible even across the boundary of `FastKernel` tables generation, and applied with benefit, e.g., for the massive replicas set evolution (consider `NNPDF40_nnlo_as_01180_1000`, that is a single set consisting of 1000 replicas, that can be evolved with a single operator instead of running 1000 times an evolution program, like all the other similar sets), but even repeated fits.

While the benefit is limited for other use cases, any other highly iterative phenomenological study, in which PDF evolution is repeatedly evaluated from different border conditions, would benefit from being backed by `EKO`, since the cost of DGLAP evaluation is paid only once (even though we are conscious that this is mainly beneficial for PDF fitting).

2.3.2 Computation time

As we said above the computation almost happens independently for each target μ_F^2 and the amount of time required scales almost linearly with the number of requested μ_F^2, except for the thresholds operators in VFNS that are computed only once.

We call computing an operator with a fixed number of flavors "evolving in a single patch", since in a VFNS the evolution might span multiple patches. When more than a single patch is involved, operators have to be joined at matching scales μ_h with a non-trivial matching, that has to be computed separately (these are part of the threshold operators).

Typical times required for these calculations in `EKO` are presented in table 2.2. As expected the complexity of the calculation grows with perturbative order, and

[4] E.g. internally integrating the minimal amount of anomalous dimensions required for the operator determination, while still providing a flexible delivery on all the output dimensions (reinterpolating the x dependencies, or rotating into different flavor bases).

so even the time taken increases. At LO no matching conditions are needed, while for NLO and NNLO they are computed once for each matching scale.

	patch	matching
LO	10 s	∅
NLO	45 s	65 s
NNLO	60 s	75 s

Table 2.2: Rough estimates of times taken by EKO, with an average sized x-grid of 50 points and single core.

We consider these time performances satisfactory, even though it is not straightforward to compare EKO with the other evolution codes, as mentioned in section 2.3.1. As an example, NNLO evolution in $\mu_F^2 = 1.65^2\,\text{Gev}^2 \to 100\,\text{GeV}^2$ crossing the bottom matching at $4.92^2\,\text{Gev}^2$ takes $\sim 60\,\text{s} + 135\,\text{s}$ in EKO (135 s for the thresholds operators initialization, 60 s for the last patch). APFEL takes $\sim 25\,\text{s}$ on the same custom interpolation x-grid (APFEL is able to perform significantly better on a pre-defined, built-in grid).

This comparison shows that on the evolution of a single PDF EKO is not really competitive, but the ratio is limited to ~ 7.5. However, we already pointed out that the two programs perform a rather different task: computing a whole operator against a single PDF evolution (on which the benchmarking is done, only because both programs are able to perform this simple task, but it is a worthless task for EKO usage).

The comparison is technically possible, but we do not encourage this kind of benchmarks, because the typical task is actually different, and this motivates the different performances. EKO perform bad in the case of the single task, but with a perfect scaling (negligible work needed for repeated evolution, practically constant), while any other program would perform better for the atomic task, but with a linear scaling in the number of objects to be evolved.

Each program should be selected having in mind the specific usage. EKO is recommended for PDF fitting, and repeated evolution in general.

The time measures in table 2.2 and the rest of this section have been obtained on a regular consumer laptop:

```
OS: Linux 5.13 Ubuntu 21.10 21.10 (Impish Indri)
CPU: (4) x64 Intel(R) Core(TM) i5-6267U CPU @ 2.90GHz
Memory: 7.56 GB
```

No one of them is a careful benchmark, i.e. repeated multiple times, but is mainly meant to show an order of magnitudes comparison.

2.3.3 Memory footprint

Memory usage is dominated by the size of the final object produced, since a much smaller internal representation is used during the computation. The final

object holds information about the rank 5 operator, so it is strictly dependent on the size of the interpolation x-grid and the amount of target μ_F^2 values.

For an average sized x-grid of 50 points, and a single target μ_F^2 the size of the object in memory is of $\sim 7.5\,\text{MB}$, which scales linearly with the amount of μ_F^2 requested. The dependency on the size of the x-grid is roughly quadratic.

2.3.4 Storage

For permanent storage similar considerations applies with respect to the memory object. The main difference is that the object dumped by the EKO functions is always compressed, leading to a size of $\sim 500\,\text{kB}$ for a single μ_F^2, which does not necessarily scales linearly with the amount μ_F^2 since the full rank 5 tensor is compressed all-together (a linear scaling is just the worst case). Similar considerations applies to the dependency on the size of the x-grid.

2.3.5 Possible Improvements

There are a few easy directions to boost the current performances of EKO, leveraging the μ_F^2 splitting.

JOBS To improve the speed of the computation, all the ingredients of the final tensor (patches & matching) can be computed by separate jobs, and dispatched to different processors. They just need to be joined at the very end in a simple linear algebra step.

Notice that the time measures presented in section 2.3.2 are obtained with a fully sequential computation on a single processor, the only one available at present time.

MEMORY Since both the computation and the consumption of an EKO can be done one μ_F^2 at a time, it is possible to store each rank 4 tensor on disk as soon as it is computed, and to load them in memory only while applying them.

Both of these improvements are in the process of being implemented.

2.4 SUMMARY

Most of the work done to develop EKO has been devoted to reproduce known results from other programs (and slightly extending or amending them to have a consistent behavior), in order to have a more flexible framework where to implement new essential features for physics study (more on this in the Outlook at the end of this section). Benchmarks with already existing and widely used codes are shown in section 2.2.1, and demonstrated to be successful. Further, the multiple

options and configurations available are presented in subsequent sections and discussed, all leading to known and understood behaviors.

This does not mean that the current status of EKO does not expose any novelty. table 2.3 summarizes a general comparison on specific features between several evolution programs; we list only tools targeting the same scope of EKO, that is unpolarized PDF fitting. It is exactly for this target (PDF fitting) that EKO is optimized, and among the others three specific features are outstanding: the solution in N-space, the backward VFNS evolution, and the operator-oriented nature.

EKO is the first code to solve DGLAP in Mellin space that has been explicitly designed to be used for PDF fitting, and while this may seem irrelevant, it has been explicitly picked as an improvement for EKO over the similar codes. There are multiple solutions that are only available in x-space by applying numerical approximated procedures, making the exact solution the most reliable one. In N-space this is not required, and the choice of the solution is left completely up to the user, with no numerical deterioration among the alternatives (as it was already for PEGASUS), and thus it can be based on theory considerations. Moreover, the perturbative QCD ingredients used in the evolution (like anomalous dimensions) are often first computed in N-space, making them available for EKO immediately, while a further complex transformation is needed for usage in the other codes.

All the programs listed are able to perform backward evolution in FFNS, that consists in swapping the integral evolution bounds, but the VFNS backward evolution is a unique feature of EKO, which involves the non-trivial inversion of the matching matrix, as outlined in section 2.1.6.

The reason why EKO is an operator-first framework is discussed in detail in section 2.3.1, but essentially it makes EKO particularly suited for our target: repeated evaluation of evolution for PDF fitting. Producing only operators makes EKO less competitive for single one-shot applications, but the optimal scaling with the size of the task (practically constant, since the time consumed is dominated by the operator calculation) makes it extremely good for massive evolution (and already good when evolving $\mathcal{O}(10)$ elements). Two special examples where a massive evolution is required are the post-fit evolution, since it is shared by all the members of the fitted PDF set, and the evolution involved in the comparison of theoretical predictions with experimental data during the fit itself, where evolution is required for each PDF member for each fit step (here the operator is usually embedded in the so-called FK table, discussed in more details in chapter 3).

It should be observed that while the choice of Python as programming language particularly stands out among the other programs (all written in Fortran, either 77 or 95), this is only a benefit from the point of view of project management (being Python much expressive) and third parties contributions (since its syntax is familiar to many). Indeed, we make sure not to experience Python infamous performances, when it comes to the most demanding tasks (like complex kernels evaluation, or Mellin inverse integration) as they either use compiled ex-

Feature	EKO	APFEL	PEGASUS	HOPPET	QCDNUM
input space	x	x	N, x^*	x	x
solution space	N	x	N	x	x
delivery space	x	x	N, x	x	x
delivery	**E**	$\mathbf{f}^a$	$\tilde{\mathbf{f}}$, **f**	$\mathbf{f}^a$	**f**
backward FFNS	✓	✓	✓	✓	✓
backward VFNS	✓			(✓)[b]	
intrinsic evolution	✓				
prog. language	Python	F77	F77	F95	F77
LHAPDF grids	✓	✓			
interpolation grids	✓	✓			

F77 = Fortan 77 F95 = Fortran 95

Table 2.3: Comparison between several evolution programs. The upper part of refers to some physical features: by x we mean the momentum fraction, N the Mellin variables, x^* denotes that PEGASUS is able to deal with x-space input, but only for fixed PDF parametrization (cf. Vogt 2005). **E** and **f** stands for evolution operators and PDFs respectively. The lower part refers to program aspects, such as program language and interface with LHAPDF.

tensions (e.g. those available in scipy Virtanen et al. 2020) or they are compiled Just In Time (JIT), using the numba Lam et al. 2015 package.

While the main purpose of EKO is to evolve PDFs, other QCD ingredients are required to perform the main task, like evolving the strong coupling α_s, running quark masses, or dealing with different flavor bases: they are all provided to the end user.

EKO is an open and living framework, providing all ingredients as a public library, and accepting community contributions, bug reports and feature requests, available through the public EKO repository.

ONGOING DEVELOPMENTS As outlined above EKO implements mostly well-known physics, but we expect a series of upcoming project to build on the provided framework that will extend the current knowledge on PDFs. Several features are already being implemented, and a few of them are already at an advanced stage: the N^3LO solution will be included as soon as it becomes available Moch, Ruijl, et al. 2022, while N^3LO matching conditions and strong coupling are already implemented and used in the recent determination of the intrinsic charm content of the proton Ball, Candido, Cruz-Martinez, et al. 2022.

[a] Both, APFEL and HOPPET, have an interface to access an evolution operator, but no one of the two can be used to store it and reuse it later on (this would require a dedicated interface).

[b] HOPPET is able by default to do backward VFNS, but is not implementing intrinsic matching conditions (i.e. the contributions associated with the presence of an heavy flavor PDF) nor the shifted matching scale.

Another main goal of EKO is to provide a backbone in the determination of MHOU (cf. chapter 4), in the first place by allowing the variation of the various scales used in the determination of evolved PDFs, that can be considered as an approximation to higher orders, implementing the strategies described in Abdul Khalek et al. 2019c. The variation of matching scales involved in VFNS is already implemented and available.

Other planned features, for which development has not yet began, include: polarized evolution J. Blümlein et al. 2022; Vogt, Moch, Rogal, et al. 2008; Vogt, Moch, and Vermaseren 2014, evolution of fragmentation functions Almasy et al. 2012; Mitov et al. 2006; Moch and Vogt 2008, and the QED $\otimes$ QCD evolution of the photon-in-hadron distribution Bertone, Stefano Carrazza, N. P. Hartland, and Rojo 2018; Cridge et al. 2022; Xie et al. 2022, to estimate the impact of electro-weak corrections onto precision predictions.

3 THEORY PINELINE

Performing a PDF fit requires to integrate several elements to be gathered from many different sources: *data* from several experiments, ranging over multiple decades and formats, and competitive *theory predictions*, coming from different providers. Finally, a fitting methodology has to be selected and engineered to implement theory constraints, and to limit not physically motivated bias.

While data are a *static* component in the fit, the theory predictions depend on the candidate PDF, since they are the map that connect the unobserved PDF space, to the observed data space. During the fit, this map will be used a large number of times (at least once for every minimization step), so it is paramount to have an efficient way to evaluate it, otherwise it can become a serious bottleneck.

For this reason, a few interfaces to PDF independent theory predictions have already been implemented D. Britzger et al. 2022; Daniel Britzger et al. 2012; Carli et al. 2010; S. Carrazza et al. 2020a, and they are used in different contexts. They propose different file formats to store the output of a Monte Carlo generator, splitting them by luminosity component, perturbative order, and observables binning. This output can be optimized as an *interpolation grid*, leveraging the fact that the PDF itself is only defined over a finite set of points, and thus including the interpolation basis in the factorized cross-section. Essentially, this recast the partonic cross sections predictions as a *theory array*, for which the Mellin convolution is replaced by a linear algebra contraction over a single or multiple PDF set. This idea can be broadened to apply to any factorized function, describing the structure of an external hadron (both incoming and outgoing).

However, this picture does not exhaust the needs of a PDF fit (or any other hadronic one), because, while the PDF dependence on flavor and x value is folded on the grid, that on factorization scale has to be fixed to the process dependent value. This dependence is not fitted, since it is only determined by perturbative QCD. In order to obtain it, it is required to solve the DGLAP equation with the border condition provided by the fit. But being DGLAP a set of integro-differential equations *linear* in the PDF, this can be converted in the application of a suitable *evolution operator*, solving the same equation, as discussed in section 0.3 and chapter 2. Since the evolution operator can also be computed ahead of time, it is possible to combine the two ingredients (the operator and the grid) in a single fast array interface, that will directly produce the required theory predictions once contracted with the PDF candidate. Thus, the map from

PDF space to data space discussed above, is reduced to a linear algebra product (or more than one, when multiple hadrons are involved). During this operation, there is no loss of generality, since the interpolation basis used for the conversion of the analytic convolutions is already present in many PDF applications due to their non-perturbative nature. Such an interface is called a "Fast Kernel table" (shortened to *FK table*) in the context of the NNPDF collaboration.

To produce the *FK tables* an evolution operator provider is required, and needs to be interfaced with the grids. This was originally done in NNPDF by an internal tool (`FKgenerator`), and then systematized in the `APFELgrid` Bertone, Stefano Carrazza, and N. P. Hartland 2017 package (leveraging the `APFEL` Bertone, Stefano Carrazza, and Rojo 2014 evolution library), later reworked once more taking the name of `APFELcomb` Bertone, Stefano Carrazza, and N. Hartland n.d.

An array interface is extremely useful, since it allows to treat the theory map in the context of many software frameworks, just relying on the data structure of an array. Especially relevant for machine learning software frameworks, but not limited to them, e.g. it allows to create a Bayesian inference-based methodology (cf. chapter 8), without the need of the treatment of further complex functions.

3.1 ARCHITECTURE

As it has been explained in the previous section, a theory map, i.e. an FK table, is made of two main components: a PDF independent interpolation *grid* and an *evolution operator*.

For the second one, we just need a single provider, able to compute the DGLAP solving operator for a variety of theory settings (corresponding to different PDF fits, e.g. NLO and NNLO QCD evolution), able to perform the operator computation as efficient as possible, and to smoothly interface with the grid for convolution. For this reason, the software package `EKO` Candido, Hekhorn, and Magni 2022a has been created, described in details in chapter 2, in order to optimize for this specific task.

`EKO` is very different from `APFEL`, the tool on which the NNPDF framework has until now relied. For instance `APFELgrid` (then `APFELcomb`), the tool which generates `APFEL`-based FK table, introduces an explicit dependency on `APFEL` itself (and thus its internals). `EKO` instead not only exposes a restricted public API (making all the dependent projects decoupled from its very internals), but the dependency is not required at all to consume the `EKO` output, consisting of float arrays stored in a very common `tar` archive, and standard `YAML` metadata. On the other side, the observable grids have to be produced by different generators, in order to cover the full variety of available processes. For this reason, we need an interface to them, with the following targets: standardizing the output and making it reproducible.

The solution we propose is thus based on the concept of interpolation grid, and specifically on PineAPPL as an interface. In particular, PineAPPL exposes APIs to different languages: it is natively written in Rust, but has an API to C/C++,

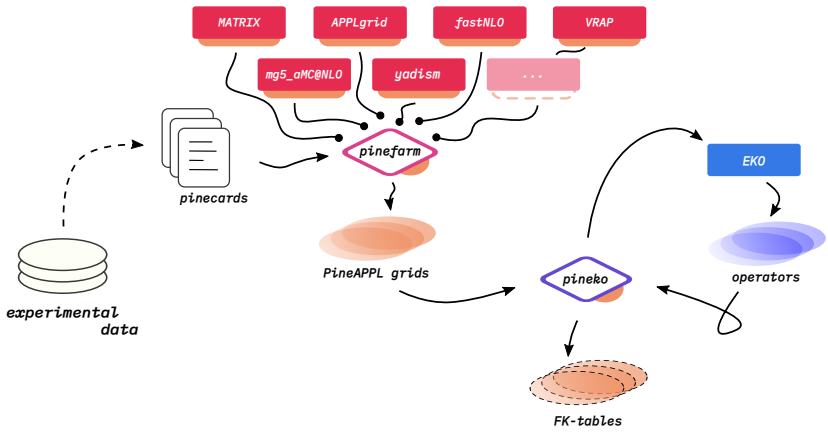

Figure 3.1: Updated version of the flow diagram already appeared in Amoroso et al. 2022, showing the overall pipeline architecture. Arrows in the picture indicate the flow of information (together with the execution order), and the orange insets on other elements indicate an interface to PineAPPL (notice `EKO` not having it). In particular, magenta blocks above `pinefarm` are the providers Anastasiou et al. 2004; Daniel Britzger et al. 2012; Candido, Hekhorn, and Magni 2022b; Carli et al. 2010; R. Frederix et al. 2018; Grazzini et al. 2018.

that can be consumed also by a Fortran application (examples provided for all of them), and a Python API, mostly dedicated to scripting and integration with the rest of the pipeline, but there are providers (essentially `yadism` Candido, Hekhorn, and Magni 2022b, used for DIS at NNLO) already using it to fill grids.

Since different generators require different inputs, we are trying to standardize them into a common format for which other cards can be generated, called *pinecard*. This is still work in progress, nevertheless, it is useful to speak of pinecards, since they are used as inputs for `pinefarm`, that is the unique Python package working as a front-end for the various generators. Essentially, each generator needs dedicated code to run, but this interface has to be written once, and then is part of `pinefarm`, standardizing the input for that generator, and part of the input across all of them (e.g. metadata, like references and observable details, or theory parameters). In fig. 3.1 we summarize our architecture: the generators are directly interfaced with the PineAPPL library, and the output is thus standardized to an interpolation grid (for one or two colliding hadrons), the input instead consists of a *pinecard*.

Once the grid is available, `pineko` (a package dedicated to the final construction of FK tables) can extract the details of the operator needed for the FK table generation from the grid, generate the `EKO` input, and then combine the grid and the operator into the final FK table.

All the components of the pipeline are open source and the code is available in the NNPDF GitHub organization:

- PineAPPL: `https://github.com/NNPDF/pineappl`

- EKO: https://github.com/NNPDF/eko
- pineko: https://github.com/NNPDF/pineko
- pinefarm: https://github.com/NNPDF/runcards (including the relevant pinecards, required for NNPDF fits)

The set of tools does not depend on the NNPDF fitting methodology and can be used in general for any hadronic function fitting[1].

We dubbed this pipeline our *theory **pineline***, that has its own dedicated repository (https://github.com/NNPDF/pinealine), providing a single Python package, acting as a meta-package, in the sense that it does not have any content, but its own version, and it is used as a single front-end for the user of the whole pipeline, reconciling the versions of the individual components. Installation instructions and tutorials are provided on the *pineline* web-page, that is intended to provide a useful introduction and resources for all kind of users:

https://nnpdf.github.io/pineline/

3.2 APPLICATIONS

Components have applications on their own, and part of them have already been used (or are being used) to support other works. Even though here it appears incidental, this is an important design feature: we are building a framework, not just a "pipeline application". The various components should be focused on dedicated tasks and easy to integrate in different architecture (or, more realistically, stand-alone projects), for similar but different goals.

A first example is the study on evidence for an intrinsic charm component in the proton Ball, Candido, Cruz-Martinez, et al. 2022 (cf. chapter 5), based on the NNPDF 4.0 PDF set, latest release of the NNPDF family, and EKO Candido, Hekhorn, and Magni 2022a, the evolution code described in the previous section 3.1. The role of EKO has been to unfold the intrinsic component from the so-called fitted charm, in the 4 flavor number scheme (default scheme at fitting scale for NNPDF), by backward evolving with DGLAP equation in a 3 flavor number scheme PDF set at a lower scale. On top of the required backward evolution, and the proper treatment of intrinsic components, EKO implemented the N^3LO matching conditions between the 4 and 3 flavor schemes, that have been relevant to estimate the perturbative stability of the result obtained.

Another application is the study of the forward backward asymmetry in the Drell–Yan process with a high cut in the invariant mass of the lepton pair Ball, Candido, Forte, et al. 2022 (cf. chapter 6). In particular, the work focuses on the comparison between results obtained with the NNPDF 4.0 PDF set and other contemporary PDF sets from different collaborations. We find that a certain

[1] Generalization of PineAPPL to support fragmentation functions and polarized PDFs is work in progress.

shape in the high cut setting is related to the specific shape of the PDFs in the large-x extrapolation region, and so very sensitive to the possible bias of extending behaviors typical of the central data region. In this context, it has been crucial to have PineAPPL S. Carrazza et al. 2020a Schwan et al. 2022c grids pre-computed to reproduce the results, iterating on the PDF set to investigate different features of the PDF, and trying to trace back the distribution behavior to PDF features.

Finally, a study of the low energy neutrino structure functions is ongoing, where the low Q^2 experimental data is reconciled to the known perturbative calculation at higher energies, based on the PDFs. Here, we use `yadism`, a general inclusive DIS provider interfaced with PineAPPL, to produce perturbative QCD calculation for the structure functions that get matched to experimental data.

4 MISSING HIGHER ORDER UNCERTAINTIES

The first fundamental application of the integrated pineline (cf. chapter 3) will actually the inclusion of Missing Higher Order Uncertainties (MHOU) at NNLO in the `NNPDF4.0` fit.

PDFs are non-perturbative objects, so it may seem counter-intuitive that their accuracy depends on perturbative series truncation. This is a direct consequence of extracting them from high energy collisions data: they are completely determined by physics that happens in the perturbative regime, and the map discussed at the beginning of chapter 3 (the one that connects data to PDFs) is completely determined by pQFT calculations. So, the origin of the perturbative order of PDF sets is exactly determined by the theory predictions used during the extraction: a NNLO set is a PDF set that has been fitted using theory predictions at NNLO. A PDF set directly computed with non-perturbative methods would have no perturbative order associated, even when used in a perturbative calculation [1][2].

The perturbative series enters in the PDF in two different places: the partonic cross section calculations (those encoded in *grids*) and the DGLAP evolution[3]. In principle, these are two different perturbative orders, thus there is not a single truncation, but two of them, and they can happen at two different orders. Still, the two objects are not completely decoupled: DGLAP evolution arise from collinear divergences, subtracted by the chosen factorization scheme. These collinear logarithms appear as well in the partonic cross sections, so it is important to properly

[1] From that point of view would be an *all-order* object, even though it might be subject to other kinds of approximations.

[2] Also consider that DGLAP evolution is perturbative, so, once evolved, it acquires again a dependency on the perturbative truncation.

[3] That technically is used twice: during the fit, to bridge data with the boundary condition candidate, and to evolve the final boundary condition to all scales. But considering the PDF a function of two variables (z and μ_F^2) consistently, the abstract evolution flow used is a single one.

account for them, avoiding double counting. The whole picture of collinear subtractions is deeply connected to treatment of quark masses, better discussed in chapter 1, since a finite value of the mass regulates the collinear divergence on its own. Therefore, the double perturbative order already appears in the partonic cross sections calculations, where the FNS chosen can account for light and heavy quarks at two distinct orders (cf. Forte, Laenen, et al. 2010, in particular the FONLL-B scheme).

4.1 ESTIMATES

The goal of MHOU studies is to give an estimate of the impact of the missing part of the perturbative series, in order to assess the size theory uncertainty propagated on physical observables. There are two categories of possible approaches:

- use all-order information coming from theoretical knowledge of the perturbative series (or properties that applies to the all-order result)
- and extrapolating from the behavior of the known orders.

The *first category* makes use of a similar information of that exploited in *resummation*, with a different goal: resumming the perturbative series produces a new expansion with a better convergence, while in MHOU studies the goal is to estimate the missing part of the initial truncated series. The prominent example of this category is the widespread adoption of *scale variations* as theory uncertainties estimates for perturbative calculations. The physical motivation relies in Callan-Symanzik equations, the same used to obtain DGLAP (cf. section 0.3). These equations encode a property of physical observables: they can not depend on unphysical scales. But this property holds only for the all-order physical observables, and it is spoilt by the perturbative truncation. Therefore, measuring the dependence of the final result on the variation of unphysical scales, it is possible to extract the magnitude of this violation. It is not possible to reconstruct the exact value of an observable from this information alone, since there is no unique solution to the equations, not even conditioning on the known orders, so solving the equations is not a sufficient condition to identify the missing orders and exact observable. Nevertheless, as said before the goal of MHOU investigations is not to upgrade a truncated result to a full one, so capturing the order of magnitude is sufficient. There are cases in which the scale variations approach is known to fail, giving an unreliable estimate also of the order of magnitude. However, most of these cases can be predicted by simple enough properties of the perturbative series. E.g. at low enough orders some partonic channels might not be present yet, like DIS at LO has no gluon channel (and at NLO no quark singlet contribution). There is not a single scale to be varied, but two: the renormalization and the factorization scales, they have been briefly introduced in section 0.3, and they also appeared in chapters 1 and 2. They are linked to the two perturbative truncations described above, so the two of them have to be varied to obtain a

complete estimate. The way the two variations are coordinated is called a scale variations *prescription* and it is illustrated in more details in section 4.4.

The main criticism to *scale variations* is not to capture only a subset of missing terms, that is mostly common to all approaches since the available information is coming from the finite amount of computed terms. Instead, it is the *arbitrariness* connected to the prescriptions themselves. Even for single scale variation there is already a completely free parameter: the amount of the variation. The conventional solution, based on the logarithmic nature of the scale dependence, is to double and halve the value of the scale, usually set to a process scale, to minimize "spurious" contributions (but the chosen scale is also somewhat arbitrary, for the same reason). This does not solve the arbitrariness that remains in the connection between the estimate and the real value, but it gives a way to compare the impact on different calculations, since the two estimates will share the same arbitrary value.

The *second category* began with Cacciari and Houdeau 2011, that formulated a Bayesian model to extrapolate the prediction value for unknown orders, updating it with the known ones. There are two main goals for this kind of approaches: getting rid of the scale variations arbitrariness, and extract more information than a single shift estimate. Indeed, the result of Bayesian inference is always a posterior distribution of some quantities, or something derived from it. The probability density in principle contains more information, and it does not have to rely on some Gaussian assumption (as will be done for example in section 4.2), that while reasonable in most situations, can drastically fail for some edge cases. But in order to infer these quantities, two inputs are required: a prior distribution and a conditional likelihood, and unless some theoretical knowledge about the perturbative series is embedded in their definition, they are arbitrary as well.

In that work in particular, the authors assume that the one and only link between different perturbative orders is a common upper bound, dubbed $\bar{c}$. They assume a certain distribution for the orders, dependent uniquely on this parameter, and motivated only by its simplicity (admittedly forced because of the simplification of the result), and assume a flat non-normalizable prior in $\log(\bar{c})$, and try to estimate the unknown orders passing through the distribution of the $\bar{c}$ parameter. An entire section is spent motivating the various choices, but the guiding principles are just simplicity and technical benefits. So they avoid the choice of quantitative arbitrary parameters, but trading for the choice of qualitatively arbitrary functions. This approach has been further investigated by other authors, proposing empirical Bayes methods to set the optimal scale of the parameter $\bar{c}$ Forte, Isgrò, et al. 2014, considering different models, with more constraining assumptions, or consuming the same information contained in scale variations approach Bonvini 2020. Another option is always to increase the number of parameters Duhr, Huss, et al. 2021, to obtain a different compromise between flexibility and simplicity.

All these approaches are designed to be applied on observables predictions, but they need some further methodological developments to be propagated on the PDF determination. In the following, I will focus on the *theory covariance*

matrix method, introduced by the NNPDF collaboration, but this is not the only option available, since recently other approaches have been proposed, based on probabilistic reweighting and post-fit selection Kassabov et al. 2022 or the introduction of nuisance parameters in the fit McGowan et al. 2022.

4.2 THEORY UNCERTAINTIES IN PDF FITS

All the methods outlined in section 4.1 are suitable to provide an uncertainty estimate on the theoretical prediction of an observable value. With more or less extra assumptions, the exact definition of this uncertainty actually corresponds to *probability distribution* over the value predicted.

During a PDF fit, probability distributions are already involved, but they only appears on the experimental data side. So, the normal workflow consist to compare a probability distribution coming from data with the individual value provided by theoretical predictions, and minimize the χ^2 distance (defined thanks to presence of said data distribution, approximately assumed to be Gaussian). This is already not the whole story, since the NNPDF methodology accounts also for the PDF distribution, that generates a distribution for theory predictions through the theory map of FK tables (cf. chapter 3), i.e. its *push-forward*. But NNPDF, with its MC approach, accounts for this distribution one PDF replica at a time, reducing the problem to minimize the χ^2 distance, with the probability distributions only on experimental side, again.

However, the introduction of theory uncertainties generates once more a distribution for the theoretical prediction values, this time not stemming from the PDF one, but directly from the distribution of the theory map itself[4]. In order to deal with these distributions, another approach is required.

This problem has been faced for the first time by the NNPDF Collaboration in Abdul Khalek et al. 2019b. The strategy adopted is described in section 2 of the reference, and basically consists in assuming a Gaussian distribution for the theory predictions (for a single PDF candidate), finding that this will modify to usual probability distribution for the shifts between theory predictions and experimental data only in the covariance matrix. Essentially, the probability distribution is still a Gaussian, as it is in absence of theory uncertainties, but the full covariance matrix turns out to be simply the sum of the experimental and theoretical covariance matrices.

[4]To get a feeling of what is happening, consider that the theory predictions are the convolution of the PDF f with a theory calculation FK, and taking both of them to depend on some parameters, say respectively ξ, and ζ, so $\sigma(\zeta, \xi,) = \mathrm{FK}(\zeta) \otimes f(\xi,)$, and then the distribution on the observable prediction σ can be generated from both the distribution on ξ,, the PDF distribution, and one on ζ, the theory map one.

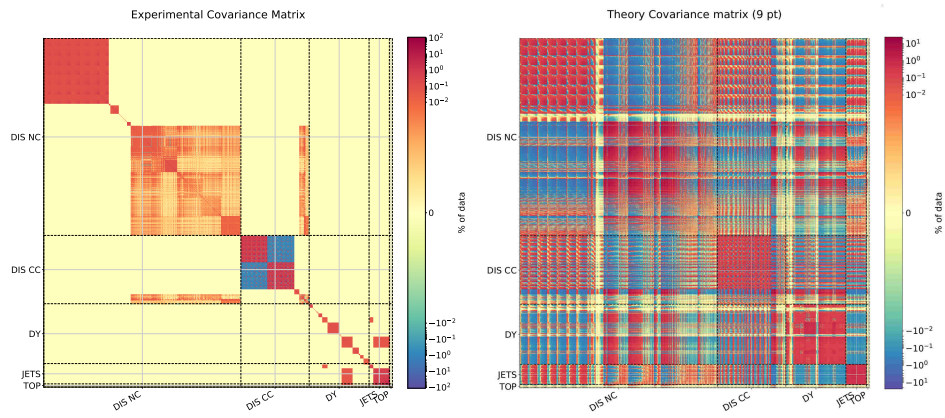

Figure 4.1: Comparison between the experimental covariance matrix and the theoretical one, generated by the 9 point prescriptions, both normalized to central values.

In fig. 4.1 the two matrices are shown for the NNPDF3.1 dataset, using colors to represent the entries as percentages of the central values[5]. The theoretical one is computed using the 9 points prescription, that is explained, together with the other point prescriptions in section 4.4. The sum of the two matrices, i.e. the actual covariance matrix used in the NNPDF3.1th fit, is shown in fig. 4.2.

In Abdul Khalek et al. 2019b the MHOU study stops at NLO, but since NNLO central values are already available (used in the baseline for NNPDF3.1) it is possible to compare the size of the estimated uncertainties (variances of individual observables, i.e. diagonal entries of the combined covariance matrix) to the actual shift between the NLO and NNLO central values for the central value of NNPDF3.1 PDF set. This comparison plot is shown in fig. 4.3, still using the same 9 points prescription adopted for the theory covariance matrices used in previous figures. It is important to note that the NNLO in the PDF is not included in the very same way of NLO, apart for DIS values (still the most part in the NNPDF3.1 dataset). Indeed, in order to produce NNLO values for a generic PDF candidate a NLO interpolation grid (cf. chapter 3) is *upgraded* to an approximate NNLO one by means of K-factors, i.e. under the hypothesis that the only difference between the two orders is only the global normalization, that is obtained by the ratio of predictions computed with an already available PDF set (usually the former release of NNPDF itself, making the process iterative).

The final result for a fit including the combined covariance matrix is shown in fig. 4.4. This set is the NNPDF3.1th release, and it is available on the NNPDF

[5]These plots, and the following ones in this section, are all taken from Abdul Khalek et al. 2019b and used to illustrate the readers the concept discussed. Cf. the reference for further details about the plots themselves.

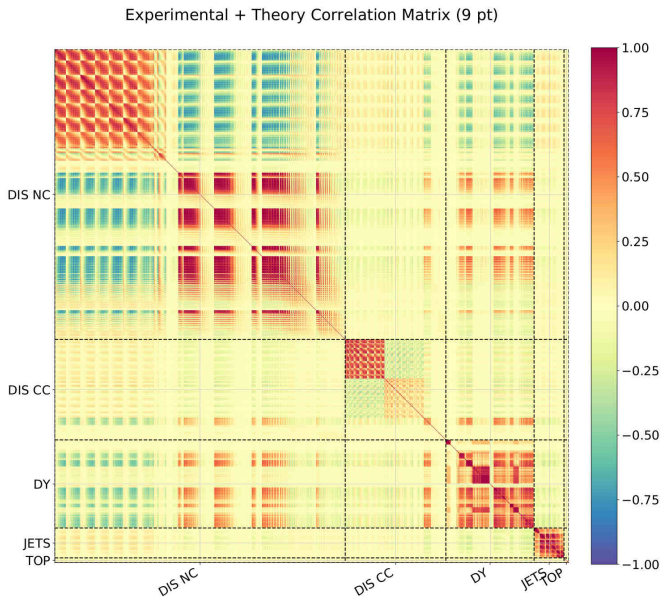

Figure 4.2: Combined covariance matrix (experimental plus theoretical), the actual one used in the `NNPDF3.1th` fit.

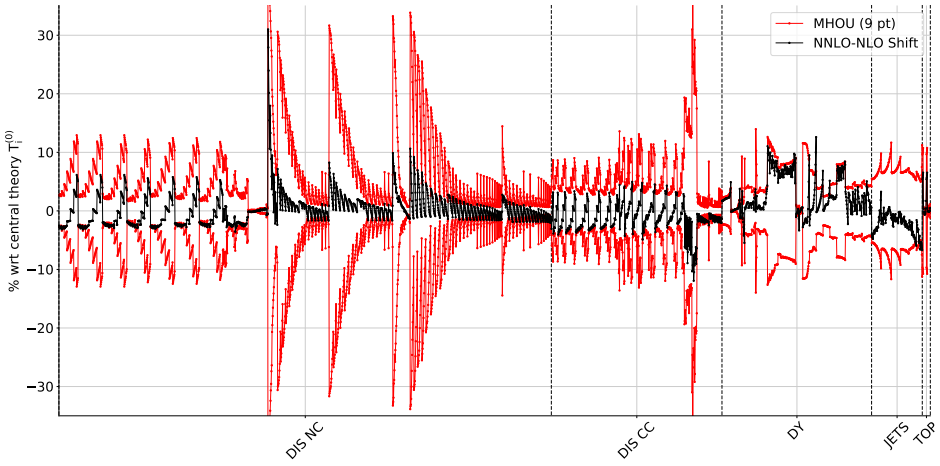

Figure 4.3: The diagonal uncertainties σ_i (red) symmetrized about zero, compared to the shift δ_i for each data-point (black). Values are shown as percentage of the central theory prediction

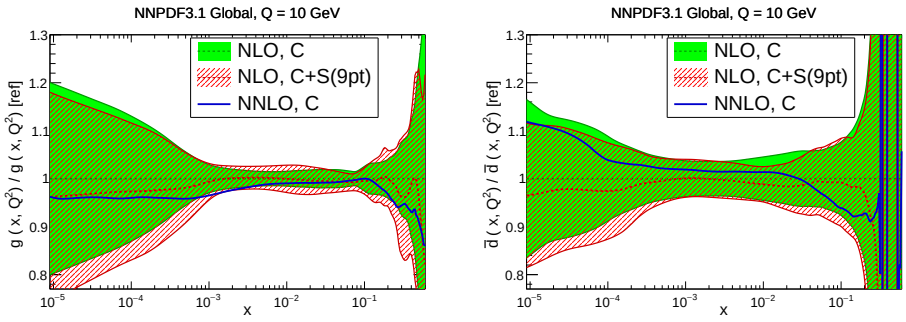

Figure 4.4: `NNPDF3.1th` NLO sets, gluon and anti-down distributions at 10 GeV, the first PDF determination to include MHOU estimates in the fit.

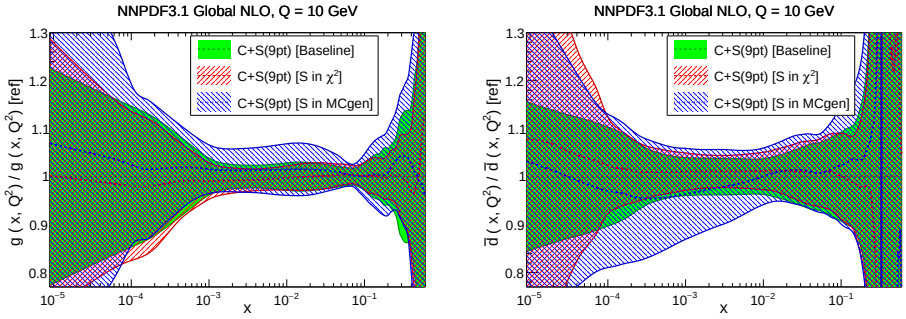

Figure 4.5: Gluon and anti-down distributions comparison, in which it is shown the effect of using the theory covariance matrix in the χ^2 or in the pseudo-data generation only.

website and as an `LHAPDF` website, on the respective website, `https://lhapdf.hepforge.org/pdfsets`.

It is important to remark that the combined covariance matrix has to be used in all the places in which the experimental covariance matrix was used, i.e. all the places in which the distribution enters. Indeed, the effect of the theory distribution, under the assumptions in Abdul Khalek et al. 2019b discussed above, is to modify the covariance matrix in the distribution, and so in all its instances in the PDF fit. In particular, the distribution (and so the covariance matrix) is used in two places:

i. the definition of the χ^2 to be minimized

ii. the generation of pseudo-data, part of the NNPDF methodology (better explained in chapter 8, or in NNPDF literature, such as Ball, Del Debbio, Forte, Guffanti, Latorre, Piccione, et al. 2009)

The effect of using the combined covariance in a single instance, while keeping the experimental one for the other (as in the baseline), is shown in fig. 4.5.

Another appropriate remark is about the inclusion of factorization scale variations. They play with the value of the scale involved in the factorization scheme, that controls which collinear terms factorized in the PDFs. Being the PDFs universal object, by virtue of the factorization theorem, these terms are universal themselves. It is customary, at this point, deciding to include them in partonic cross-sections for each process, or to include them only in the running of DGLAP evolution. Moreover, even when it has been decided to include the variation inside the evolution, it is still possible to expand out the factorization scale dependence, or to resum it as well. These alternatives, together with the choice of refitting or not the border condition for different scales, generates the so-called "schemes" for factorization scale variations, described in section 3.3 of Abdul Khalek et al. 2019b.

The schemes main feature are:

A. including and resumming factorization scale variations in DGLAP evolution, refitting the border condition

B. including them in DGLAP evolution, but expanding them out

C. keeping the variations in the partonic cross-sections, together with renormalization ones

The main disadvantage of scheme A is refitting the border condition, that is a computationally expensive operation (and possibly a delicate one, because of the interplay with fit machinery). But both scheme A and B have the advantage of being on the universal side of factorization, the PDF side, while the other one is process dependent. Indeed, applying scheme C might require to implement scale variations consistently in all the MC generators used to obtain predictions. The default choice of NNPDF has been scheme B, since it has the same advantages of scheme A, concerning universality, but includes expanded scale variations, that should be almost equivalent to scheme C ones, because the expanded corrections are the same, applied on one or the other side of the convolution (PDF or partonic cross-section).

4.3 NEW DEVELOPMENTS

The investigation in Abdul Khalek et al. 2019b stopped at NLO, because of various technical limitations, and the lack of a proper benchmark to assess the reliability of available implementations. An NNLO fit based on the theory covariance matrix formalism is still missing, and this specific target will be achieved with the tools offered by the new theory pineline (cf. chapter 3).

A first update regards the actual implementation of scheme B, introduced in section 4.2. Indeed, scale variations in scheme A-like fashion has been the first and default way they appeared in the evolution programs, starting with programs like `PEGASUS`. In order to apply them, the PDF anomalous dimensions have to be evaluated at a scale shifted from the usual one, and a few extra terms appears.

But, being implemented at the anomalous dimensions level, these contributions are resummed by the solution of the differential equations, and essentially exponentiated (the solution of a linear differential equation is a path ordered exponential of the associated kernel). But this is not the contribution required by scheme B, and violates the scheme C equivalence. To get the expanded result, an extra piece has to be multiplied to the evolved PDF, so after solving the differential equation. For this reason, `EKO` ships both the kind of factorization scale variations (the only scale variations on the evolution side), dubbed *exponentiated* and *expanded*, and the user can choose an actual solution conformal to the scheme B prescriptions. Since `EKO` does not produce evolved PDFs, but evolution operators, the extra piece appearing in the expanded case is also implemented at the anomalous dimensions level, Mellin inverted, and included as an extra factor in the final operator returned as output (i.e. it is not returned individually).

Another delicate point is the role of the strong coupling α_s in the computation of factorization scale variations. Indeed α_s appears in two places: as the parameter of the perturbative series, evaluated at the scale of the process, and in the solution of DGLAP equations. Indeed, the running coupling $\alpha_s(Q^2)$ is a monotonically decreasing function of the scale, and contains the only scale dependence in the anomalous dimensions. Therefore, it is convenient to change variable, and solve DGLAP as function of α_s, instead of the factorization scale. This usage might suggest a relation between the strong coupling and the factorization scale, and then the value used should be affected by the related variations. Instead, this is only used as a monotonic function, so for its mathematical properties and the role in the equation (stemming from anomalous dimensions perturbative expansion), and there is no implication about a physical relation. Essentially, α_s value is only sensitive to renormalization scale variations, that does not affect evolution at all, but it is used there as a function of factorization scale. This usage led to a clash in the options used, and a slightly wrong result when both variations were applied, and it has been fixed in the current pineline.

Moreover, both renormalization scales variations (always coming from the partonic cross-sections) and full scheme C (i.e. including also factorization) can also be accomplished without requiring any information about scale variations from the generators used. The structure of the scale variations contributions consists substantially in scale ratios logarithms multiplying lower order cross-sections, and the coefficients of the perturbative expansion of the beta function of the strong coupling $\beta(\alpha_s)$ (renormalization) or the splitting functions/anomalous dimensions $P(x, \alpha_s)$. If the interpolation grid is stored separately by order, it is then possible to reconstruct the missing dependence on the two scales, as described in section 2.3 of Carli et al. 2010. However, the reconstruction of factorization scale dependence requires convolutions with increasingly complex distributions (with the perturbative order), i.e. more or less the same complexity of a DIS coefficient functions integration, as it is performed by `yadism`. But they are also the universal ingredient. So, we decided for a mixed approach: keep using scheme B there is no need for factorization scale dependence to be stored in or computed from the grids (the original reason to advocate for this scheme, over the C option), but we

can reconstruct the renormalization scale dependence, just relying on a fixed set of numerical coefficients (β function expansion), completely removing the need to obtain scale variations from external providers.

Finally, using both the ingredients provided by the `EKO` and `yadism` libraries, we benchmarked our implementations of scale variations with some analytical results, based on the expressions for some specific DIS contributions, order by order and for specific partonic channels. This already allowed us to resolve the small differences between scheme B and C, and confirm that they are always higher order contributions, even though the difference in the actual values is only negligible in most of the cases, but not all of them, becoming sizeable in specific kinematic corners (from few percent up to $\sim 20\%$). A few channels are still missing at NNLO, and the full benchmark will be presented in a separate publication about the MHOU treatment with the new pineline, possibly the release of the first NNLO NNPDF set accounting for MHOU.

4.4 SCALE VARIATIONS — POINT PRESCRIPTIONS

As introduced in section 4.1, in the *scale variations* approach the scales to be varied are actually two: the *renormalization* and the *factorization* scales.

More precisely, only the kind of scales to be varied are two, but there is one renormalization scale associated to each *process*[6], so the amount of scales is essentially $p + 1$, where p is the amount of processes in the dataset. This is because the factorization scale is the scale associated with the PDF factorization, that, because of universality, it is common to all data. On the other hand, the renormalization scales are associated with the scales arising in renormalization conditions, and this is common for each process, but not unique for all processes.

A consistent way of varying together all these scales is required in a global QCD fit, like those for collinear PDFs, because the amount of processes might quickly scale, and rough choices for prescriptions might result in undesirable features for the theory covariance matrix generated. Also consider that, while dealing with a large number of data points (even for a small amount of processes), covariances become crucial, since they scale as $\mathcal{O}\left(n_{\text{data}}^2\right)$, while the variances are just n_{data}. The specific choices for a coordinated choices of scale values is called a *point prescriptions*, because consist in the selection of a finite set of points (and related weights) in the space of possible values for the unphysical scales, use to estimate the whole observable dependence.

In the following, we present the derivation of suitable point prescriptions, that can be used in the construction of a positive semi-definite theory covariance ma-

[6] Unfortunately, also the process definition is not unambiguous, being essentially a way to gather in groups experimental data point. An indication is given by the theory predictions associated, and the intuitive idea is that each group is identified by a different LO Feynman diagram. Because of this, it might happen that some processes correspond to NLO or further contributions to other simpler processes.

trix. It turns out that two classes of prescriptions are possible, both requiring milder or stronger generalization of equations in Abdul Khalek et al. 2019b. The actual result claimed in the reference can be obtained within the broader generalization, ensuring the positive semi-definiteness of the covariance matrix computed with that class of prescriptions.

There are two conditions that we want to satisfy in constructing the theory covariance matrix, in order to support the interpretation as the covariance matrix of our theory prior distribution.

A. We want the theory covariance to be *generated by some shift vectors* $\Delta_i(\vec{\kappa})$; the vectors should be proportional to the difference of predictions obtained by a theory variation $T_i(\vec{\kappa})$ and the default theory in which $\vec{\kappa} = \vec{\kappa}_0$

$$\Delta_i(\vec{\kappa}) = c_i(\vec{\kappa})\,(T_i(\vec{\kappa}) - T_i(\vec{\kappa}_0)) \tag{4.1}$$

$$S_{ij} = \sum_{\vec{\kappa}\in\mathcal{V}_{ij}} \Delta_i(\vec{\kappa})\Delta_j(\vec{\kappa}) \tag{4.2}$$

B. We want it to be *positive semi-definite*, as required for any covariance matrix

$$v_i S_{ij} v_j > 0 \qquad \forall v \in \mathbb{R}^{n_{\text{data}}} \tag{4.3}$$

4.4.1 Derivation

Once all the elements in eqs. (4.1) and (4.2) are spelled out, we have a clear recipe on how to compute the covariance matrix S_{ij}.

For this reason, we are going to exploit all the properties that are required or desirable (advantageous), in order to limit the available degrees of freedom: anything left, it has to be regarded as being part of the *prescription*.

The current degrees of freedom are:

1. the choice of the $p+1$ dimensional space $\mathcal{V}_{ij}$ of all the accounted variations (p renormalization scales, 1 factorization scale)
2. the choice of normalization coefficients $c_i(\vec{\kappa}) \in \mathbb{R}$ [7]
3. the choice of the default value $\vec{\kappa}_0$

The last element is trivial: it's going to be part of the prescription, but in the following we will always write $\vec{\kappa}_0 = \vec{0}$ for definiteness (it's simple to replace this in the final result with $\vec{\kappa}_0$ in any case).

[7] Not all values of $\mathbb{R}$ make sense, but there is quite a wide range of interesting variations: $\mathbb{N}$ for repeated points, or $\mathbb{Q}^+$ for normalizations (possibly coming from repeated points), or 0 for masking. At this level, we are just not excluding anything that has no special reason to be excluded.

EXTRA SCALES We know that the predictions for each data point only depend on two scales: the common factorization scale, and the related renormalization scale, but not the others. For this reason, it makes no sense to pick the normalization for point i dependent on the other scales, since it would introduce a dependency of the shifts on those scales that was not present in the unnormalized shifts. Thus:

$$c_i(\vec{\kappa}) \equiv c_i(\kappa_F, \kappa_{R,i}) \tag{4.4}$$

PER-PAIR SPACE Next, we claim that the space $\mathcal{V}_{ij}$ can not actually depend on the element ij of the covariance matrix been constructed. Indeed this stems directly for the necessity to prove eq. (4.3) that is done in the following way:

$$\sum_{i,j} v_i S_{ij} v_j = \sum_{i,j} \sum_{\vec{\kappa}\in\mathcal{V}_{ij}} v_i \Delta_i(\vec{\kappa}) \Delta_j(\vec{\kappa}) v_j = \tag{4.5}$$

$$= \sum_{\vec{\kappa}\in\mathcal{V}} \sum_{i,j} v_i \Delta_i(\vec{\kappa}) \Delta_j(\vec{\kappa}) v_j = \tag{4.6}$$

$$= \sum_{\vec{\kappa}\in\mathcal{V}} \left(\sum_i v_i \Delta_i(\vec{\kappa}) \right)^2 > 0 \tag{4.7}$$

If the space $\mathcal{V}$ were actually dependent on ij, it would have not been possible to swap the two sums in the second step.

SPACE CHOICE On the other hand, it is desirable to define the prescription only on the space of relevant scales for the given point ij. This means the factorization scale κ_F and

off-diagonal two renormalization scales $\kappa_{R,i}$ and $\kappa_{R,j}$, or

diagonal even a single one, if the two points are related to the same process, i.e.

$$\kappa_{R,i} = \kappa_{R,j}$$

We would like our expressions not to depend on the number of scales present, and only account for the scale relevant for the pair ij being considered. The easiest choice is to pick the space $\mathcal{V}$ to be fully factorized in the various dimensions of $\vec{\kappa}$. This means that it can be written as

$$\mathcal{V} = \prod_{i=1}^{p+1} v_i, \tag{4.8}$$

with v_i the one-dimensional space representing the variation of the single scale labeled with i.

But this is not the only choice available, it is just the simplest. There is only one more option that guarantees the independence of the projection on the pair ij, i.e. factorize the space for each possible value of κ_F. This option will be explored in section 4.4.3.

In the case of a fully factorized space, the complex choice of the space is reduced on $p+1$ choices for one dimensional spaces. But if there is no reason to distinguish processes at this level, it is reasonable to pick the same space for each renormalization scale.

In practice, the basic one dimensional space will be always the same[8]:

$$\nu = \{-\log(2), 0, \log(2)\} \equiv \{-, 0, +\} \tag{4.9}$$

and the overall space will be just the product:

$$\mathcal{V} = \nu^{p+1} \tag{4.10}$$

NORMALIZATION At this point, all the arbitrariness left for the prescription is encoded in the normalization coefficients. With our simple choice of the space there is no reason to choose complex coefficients, thus we will define the following prescriptions:

$$c_i(\vec{\kappa}) = \begin{cases} 1/\sqrt{N_m} & \kappa \in \mathcal{V}_m^i \\ 0 & \text{else} \end{cases} \tag{4.11}$$

The spaces $\mathcal{V}_m^i$ now defines our point prescription, together with the overall normalization N_m, since the $c_i(\vec{\kappa})$ are acting as *masks* on the points $\vec{\kappa}$ not belonging to the space. For the former we'll choose:

$$\mathcal{V}_m^i = \nu_m^i \times \{-, 0, +\}^{p-1} \tag{4.12}$$

where the two dimensional spaces ν_m^i are always the same space ν_m, but for the scales $(\kappa_F, \kappa_{R,i})$, while the other scales are free to assume any possible value.

For the normalizations instead, there is no strict nor reasonable way to fix it completely, but it is possible to fix the scaling in the case of a space ν_m and ν with an hypothetically large number of point: since we don't want the normalization of the theory covariance matrix to depend on the number of points being in the prescription, we'll choose

$$N_m \propto |\nu_m| \cdot |\nu| = m \cdot 3^{p-1} \tag{4.13}$$

4.4.2 Examples of *full space* prescriptions

Since the presence of many processes have been reconciled at a theoretical (even though abstract) level, here we will focus on fully spelled out examples, in the simplest case of only two data points (1 and 2) belonging to two distinct processes.

[8] The one spelled out is only an option, any other space would work equally well.

Again, the following is in no way a proof, which has been spelled out in details in section 4.4.1, for which considering more than two processes is extremely relevant.

We will show the actual results of the obtained prescriptions for the on-diagonal, S_{11}, and off-diagonal, S_{12} cases.

Notice that, with respect to Abdul Khalek et al. 2019b, here we have not yet introduced the factor s, but it would still be allowed by eq. (4.13). In order to make the comparison with Abdul Khalek et al. 2019b easier, in this section we'll define the actual normalization including this factor, so:

$$N_m = \frac{m \cdot 3^{p-1}}{s_m} \tag{4.14}$$

For convenience, the unnormalized shifts will be called δ, i.e.:

$$\delta_i(\vec{\kappa}) \equiv \Delta_i(\vec{\kappa}) \cdot \sqrt{N_m} \tag{4.15}$$

In general, the expressions for the *diagonal* and *off-diagonal* cases with only two process, $p = 2$, are the following:

diagonal effectively two-dimensional, since both the shifts depend only on two scales

$$S_{11} = \sum_{\vec{\kappa}\in\mathcal{V}} \Delta_1(\vec{\kappa})\Delta_1(\vec{\kappa}) = \tag{4.16}$$

$$= \frac{s_m}{3 \cdot m} \sum_{\vec{\kappa}\in\mathcal{V}^1_m} \delta_1(\vec{\kappa})^2 = \tag{4.17}$$

$$= \frac{s_m}{m} \sum_{(\kappa_F,\kappa_{R,1})\in\mathcal{V}^1_m} \delta_1(\kappa_F, \kappa_{R,1}, 0)^2 \tag{4.18}$$

where in the last step a single value has been chosen for $\kappa_{R,2}$, since δ_1 does not depend on this scale.

off-diagonal effectively three-dimensional, that only for this specific problem coincide with the whole space (for a greater number of processes, would be itself a projection)

$$S_{12} = \sum_{\vec{\kappa}\in\mathcal{V}} \Delta_1(\vec{\kappa})\Delta_2(\vec{\kappa}) = \tag{4.19}$$

$$= \frac{s_m}{3 \cdot m} \sum_{\vec{\kappa}\in\mathcal{V}^1_m\cap\mathcal{V}^2_m} \delta_1(\vec{\kappa})\delta_2(\vec{\kappa}) \tag{4.20}$$

$$= \frac{s_m}{3 \cdot m} \sum_{\vec{\kappa}\in\mathcal{V}^1_m\cap\mathcal{V}^2_m} \delta_{12}(\vec{\kappa}) \tag{4.21}$$

where in the last step we defined $\delta_{12}(\vec{\kappa}) \equiv \delta_1(\vec{\kappa})\delta_2(\vec{\kappa})$.

9 points

The easiest prescription is the *so-called* 9 points prescription, because it corresponds to consider the whole two dimensional space as $\mathcal{V}_9^i$, thus the two elements to be fixed are:

$$\nu_9 = \{-, 0, +\}^2 \tag{4.22}$$

$$N_9 = \frac{8 \cdot 3}{2} = 12 \tag{4.23}$$

with $s_9 = 2$ (naïvely because two scales are involved).

In the following, the expressions for the diagonal and off-diagonal cases are formatted in order to stress the connection with the various pictures in this section. Concerning the *diagonal* expressions they are formatted on three lines, with three terms each, such that each term correspond to one point in the two-dimensional diagram. Since *off-diagonal* would correspond to a three-dimensional picture, this picture is ideally sliced in two-dimensional planes, and each plane is displayed in the equation as a block of terms in square brackets, and slightly indented with respect to previous blocks. In order to preserve the shape, and to stress the effect of zero values in the $c_i(\vec{\kappa})$, missing terms are explicitly marked with zeros.

diagonal for this prescription, we effectively have only 8 shifts, since out of the 9 theory predictions, one shift vanishes, just because it is used as the reference

$$\begin{aligned} S_{11} = \frac{1}{4}\Big[&\delta_1(-,-,0)^2 + \delta_1(-,0,0)^2 + \delta_1(-,+,0)^2 \; + \\ &\delta_1(0,-,0)^2 \; + \qquad 0 \qquad + \delta_1(0,+,0)^2 \; + \\ &\delta_1(+,-,0)^2 + \delta_1(+,0,0)^2 + \delta_1(+,+,0)^2\Big] \end{aligned} \tag{4.24}$$

off-diagonal the two Δ_i combine in three dimensions: each one contains 3 zero elements (relative to the two dimensional central value), but the two are overlapping over the central point $(\kappa_F, \kappa_{R,1}, \kappa_{R,2}) = (0, 0, 0)$, leading to

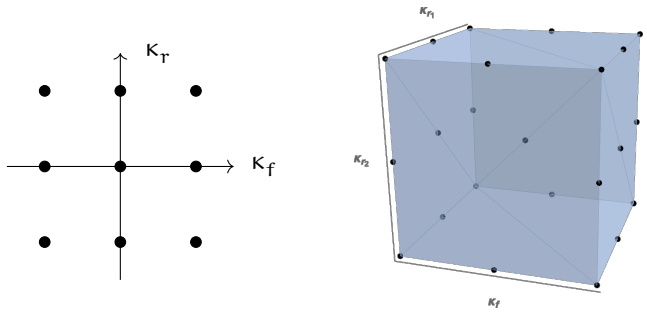

Figure 4.6: Visualization of the 9 points prescription for the diagonal (2 dimensional) and off-diagonal (3 dimensional) elements.

only 5 zero elements out of $3^3 = 27$ total elements, see section 4.4.2; thus the 22 non-vanishing elements are the following:

$$
\begin{aligned}
S_{12} = \frac{1}{12}\Big\{ &\Big[\delta_{12}(-,-,-) + \delta_{12}(-,-,0) + \delta_{12}(-,-,+) \; + \\
&\delta_{12}(-,0,-) \; + \delta_{12}(-,0,0) \; + \delta_{12}(-,0,+) \; + \\
&\delta_{12}(-,+,-) + \delta_{12}(-,+,0) + \delta_{12}(-,+,+)\Big] + \\
&\Big[\delta_{12}(0,-,-) + \quad 0 \quad + \delta_{12}(0,-,+) \; + \\
&\quad 0 \quad + \quad 0 \quad + \quad 0 \quad + \\
&\delta_{12}(0,+,-) + \quad 0 \quad + \delta_{12}(0,+,+)\Big] + \\
&\Big[\delta_{12}(+,-,-) + \delta_{12}(+,-,0) + \delta_{12}(+,-,+) \; + \\
&\delta_{12}(+,0,-) \; + \delta_{12}(+,0,0) \; + \delta_{12}(+,0,+) \; + \\
&\delta_{12}(+,+,-) + \delta_{12}(+,+,0) + \delta_{12}(+,+,+)\Big]\Big\}
\end{aligned}
\tag{4.25}
$$

5 points

Another interesting prescription is the 5 points one, since it is a rather minimal prescription involving both renormalization and factorization scale.

$$\nu_5 = \{(-,0),(0,-),(+,0),(0,+)\} \tag{4.26}$$

$$N_5 = \frac{4 \cdot 3}{2} = 6 \tag{4.27}$$

with $s_5 = 2$ (same reason of eq. (4.22)).

diagonal for this prescription, we effectively have only 4 shifts, since only 5 theory predictions are taken into account[9], and, as for the 9 points, one is used as reference

$$S_{11} = \frac{1}{2}\Big[\quad 0 \quad + \delta_1(-,0,0)^2 + \quad 0 \quad + \delta_1(0,-,0)^2 + \quad 0 \quad + \delta_1(0,+,0)^2 + \quad 0 \quad + \delta_1(+,0,0)^2 + \quad 0 \quad \Big] \tag{4.28}$$

off-diagonal in this case the two two-dimensional normalizations combine into one three-dimensional pattern, where non-zero elements are arranged in the shape of a double square pyramid: only central value is allowed for $\kappa_F \neq 0$, while the four corners are left for $\kappa_F = 0$ (same as the 9 points in this case), see section 4.4.2

$$\begin{aligned} S_{12} = \frac{1}{6}\Big\{\Big[& \quad 0 \quad + \quad 0 \quad + \quad 0 \quad + \\ & \quad 0 \quad + \delta_{12}(-,0,0) + \quad 0 \quad + \\ & \quad 0 \quad + \quad 0 \quad + \quad 0 \quad \Big] + \\ \Big[& \delta_{12}(0,-,-) + \quad 0 \quad + \delta_{12}(0,-,+) + \\ & \quad 0 \quad + \quad 0 \quad + \quad 0 \quad + \\ & \delta_{12}(0,+,-) + \quad 0 \quad + \delta_{12}(0,+,+) \Big] + \\ \Big[& \quad 0 \quad + \quad 0 \quad + \quad 0 \quad + \\ & \quad 0 \quad + \delta_{12}(+,0,0) + \quad 0 \quad + \\ & \quad 0 \quad + \quad 0 \quad + \quad 0 \quad \Big]\Big\} \end{aligned} \tag{4.29}$$

[9] with the shape of a Greek cross, as the + symbol

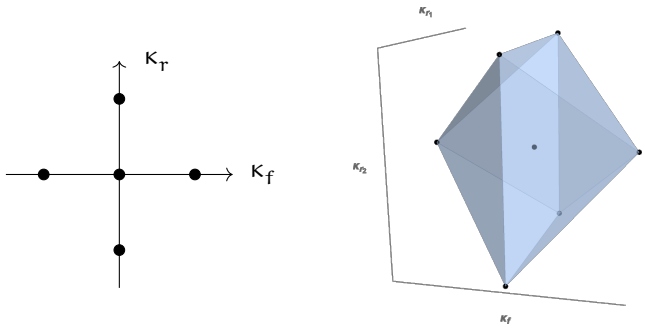

Figure 4.7: Visualization of the 5 points prescription for the diagonal (2 dimensional) and off-diagonal (3 dimensional) elements.

$\bar{5}$ *points*

Just another option with renormalization and factorization scale, with same two dimensional volume, but a different geometry.

$$\nu_5 = \{(-,-),(-,+),(+,-),(+,+)\} \tag{4.30}$$

$$N_5 = \frac{4 \cdot 3}{2} = 6 \tag{4.31}$$

with $s_{\bar{5}} = 2$ (same reason of eq. (4.22)).

diagonal for this prescription, we effectively have only 4 shifts, since only 5 theory predictions are taken into account[10], and, as for the 9 points, one is used as reference

$$\begin{aligned} S_{11} = \frac{1}{2}\Big[&\delta_1(-,-,0)^2 + 0 + \delta_1(-,+,0)^2 + \\ &0 + 0 + 0 + \\ &\delta_1(+,-,0)^2 + 0 + \delta_1(+,+,0)^2\Big] \end{aligned} \tag{4.32}$$

off-diagonal also in this case the two two-dimensional normalizations $c_i(\vec{\kappa})$ have the combined effect of setting to zero a lot of elements in the three

[10] with the shape of St. Andrew's cross, as the $\times$ symbol

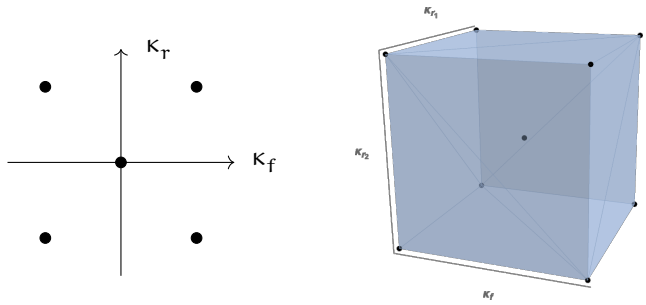

Figure 4.8: Visualization of the $\bar{5}$ points prescription for the diagonal (2 dimensional) and off-diagonal (3 dimensional) elements.

dimensional space, this leaving the shape of an empty cube: the four corners are now left for $\kappa_F \neq 0$, and no point is left for $\kappa_F = 0$

$$
\begin{aligned}
S_{12} = \frac{1}{6}\Big\{ &\Big[\delta_{12}(-,-,-) + \quad 0 \quad + \delta_{12}(-,-,+) \; + \\
&\quad 0 \quad + \quad 0 \quad + \quad 0 \quad + \\
&\delta_{12}(-,+,-) + \quad 0 \quad + \delta_{12}(-,+,+)\Big] + \\
&\Big[\quad 0 \quad + \quad 0 \quad + \quad 0 \quad + \\
&\quad 0 \quad + \quad 0 \quad + \quad 0 \quad + \\
&\quad 0 \quad + \quad 0 \quad + \quad 0 \quad \Big] + \\
&\Big[\delta_{12}(+,-,-) + \quad 0 \quad + \delta_{12}(+,-,+) \; + \\
&\quad 0 \quad + \quad 0 \quad + \quad 0 \quad + \\
&\delta_{12}(+,+,-) + \quad 0 \quad + \delta_{12}(+,+,+)\Big]\Big\}
\end{aligned}
\tag{4.33}
$$

4.4.3 Alternative space: κ_F slices

In section 4.4.1 we made a set choices for the degrees of arbitrariness exposed at the beginning. All of them were yield by a strict requirement (needed to obtain a property, like $S_{ij} \geqslant 0$) or by a reasonable request (e.g. not adding further dependencies with normalizations, which led to eq. (4.4)). Only in one single case we made an assumption based on an unneeded simplicity: the choice of the space as fully factorized.

This choice is sensible for the renormalization scales: why should the space look different seen from the perspective of different processes? Why different processes should be correlated by the space? On the other hand, it is completely arbitrary for the factorization scale. Since factorization scale κ_F is treated separately from renormalization scales $\kappa_{R,i}$, no surprise if even the space symmetry somehow is broken on κ_F[11].

Thus, we can have a different factorized space for each different value of κ_F:

$$\mathcal{V} = \bigsqcup_{\kappa_F \in \nu_F} \mathcal{V}(\kappa_F) \tag{4.34}$$

$$\mathcal{V}(\kappa_F) \equiv \nu(\kappa_F)^p \tag{4.35}$$

where ν_F is the space of possible values of κ_F (usually it will be just ν of eq. (4.9)), and $\nu(\kappa_F)$ is instead the space of renormalization scales related to that single value of the factorization scale.

In this case also the definition of the normalizations $c_i(\vec{\kappa})$ should change with respect to those defined in eq. (4.11) in order to account for this, since the different spaces contain different numbers of points. We decide to normalize the elements such that once the full space is projected over each of the two dimensional spaces $(\kappa_F, \kappa_{R,i})$, the coefficients of the various shifts are equal to one, thus:

$$c_i(\vec{\kappa})^2 \propto \frac{1}{\sum_{\kappa_F'} \nu(\kappa_F')} \frac{|\nu(\kappa_F)|}{|\mathcal{V}(\kappa_F)|} = \frac{1}{m \cdot |\nu(\kappa_F)|^{p-1}} \tag{4.36}$$

since the scales projected are all renormalization scales but a single one, that is the relevant one for the given i, and together with κ_F make the two dimensional space, whose volume is $\sum_{\kappa_F'} \nu(\kappa_F') = m$.

4.4.4 Examples of *sliced space* prescriptions

In this case as well, for better comparison with Abdul Khalek et al. 2019b, we introduce the factor of s in the normalization of eq. (4.36), thus

$$c_i(\vec{\kappa})^2 = \frac{s_m}{m \cdot |\nu(\kappa_F)|^{p-1}} \tag{4.37}$$

Furthermore, same as in section 4.4.2 (on purpose, to stress comparison) we consider the case of only two data points (1 and 2) belonging to two distinct processes. With this limited case it is harder to appreciate the difference in the constructions of section 4.4.1 and section 4.4.3, since it actually lies in the way the different three dimensional shapes for pair of processes are reconciled in the full $p+1$-dimensional space. However, this difference has already been stressed in the abstract construction of the two classes of prescriptions, thus the purpose of

[11] For the $\kappa_{R,i}$, choosing them factorized and uniform as argued, a permutation invariance is present, and makes sense. No reason to extend it to κ_F.

this examples is different: to showcase the different expressions obtained fully explicitly. For this aim the choice of considering just two points is fully satisfactory.

For this second set of examples there is no need to rewrite the full set of terms: they are the exact same of section 4.4.2, the only difference will be in the coefficients, that now might depend on the value of κ_F because of the space structure (and they will always depend on it).

Thus, the expressions for the *diagonal* and *off-diagonal* cases with only two process, $p = 2$, in this second class of prescriptions are the following:

diagonal effectively two-dimensional, since both the shifts depend only on two scales

$$S_{11} = \sum_{\vec{\kappa}\in\mathcal{V}} \Delta_1(\vec{\kappa})\Delta_1(\vec{\kappa}) = \tag{4.38}$$

$$= \sum_{\kappa_F\in\nu_F} \frac{s_m}{|\nu(\kappa_F)|\cdot m} \sum_{\vec{\kappa}_R\in\mathcal{V}(\kappa_F)} \delta_1(\vec{\kappa})^2 = \tag{4.39}$$

$$= \frac{s_m}{m} \sum_{\kappa_F\in\nu_F} \sum_{\kappa_{R,1}\in\nu(\kappa_F)} \delta_1(\kappa_F,\kappa_{R,1},0)^2. \tag{4.40}$$

where in the last step a single value has been chosen for $\kappa_{R,2}$, since δ_1 does not depend on this scale (this trivial sum cancels with the factor of $|\nu(\kappa_F)|$ in the denominator).

Notice that the last sum $\sum_{\kappa_F\in\nu_F}\sum_{\kappa_{R,1}} = \sum_{(\kappa_F,\kappa_{R,1})\in\nu^1_m}$, thus the finally formula for the diagonal case is the same of eq. (4.18). While this is not a proof of the general case, it is simple to show (in essentially the same way of above) that this is the formula obtained for any number of processes p.

off-diagonal effectively three-dimensional, that only for this specific problem coincide with the whole space

$$S_{12} = \sum_{\vec{\kappa}\in\mathcal{V}} \Delta_1(\vec{\kappa})\Delta_2(\vec{\kappa}) = \tag{4.41}$$

$$= \sum_{\kappa_F\in\nu_F} \frac{s_m}{|\nu(\kappa_F)|\cdot m} \sum_{\vec{\kappa}_R\in\mathcal{V}(\kappa_F)} \delta_1(\vec{\kappa})\delta_2(\vec{\kappa}) \tag{4.42}$$

$$= \frac{s_m}{m} \sum_{\kappa_F\in\nu_F} \frac{1}{|\nu(\kappa_F)|} \sum_{\vec{\kappa}_R\in\mathcal{V}(\kappa_F)} \delta_{12}(\kappa_F,\kappa_{R,1},\kappa_{R,2}). \tag{4.43}$$

Since the space of this second class is engineered to give the same terms of the first one (both diagonal and off-diagonal), and the normalizations are chosen such to obtain uniform coefficients for the diagonal case (and then they are the exact same of the first class, as noted above), the only difference will be in the *coefficients of the off-diagonal* case, and they can only depend on the factorization scale κ_F. For this reason, we will not repeat the full construction of the previous

section, but just adopt a concise notation to make the different coefficients explicit in the off-diagonal expressions:

$$S_{12} = \frac{s_m}{m \cdot k_m} \left(c_m(-)\delta_{12}(-, \dots) + c_m(0)\delta_{12}(0, \dots) + c_m(+)\delta_{12}(+, \dots) \right) \tag{4.44}$$

where:

- k_m is the least common multiple of the $|\nu(\kappa_F)|$, in order to leave integer coefficients in the sum
- $c_m(\kappa_F)$ is the leftover the $1/|\nu(\kappa_F)|$ once $1/k_m$ has been factored out
- $\delta_{12}(\kappa_F, \dots)$ is a placeholder for all the terms with that value of κ_F, as they have been spelled out in the corresponding prescription in section 4.4.2

9 points

The specification of this prescription is almost the same of the corresponding one for the first class:

$$\nu_9(-) = \nu_9(+) = \{-, 0, +\} \tag{4.45}$$
$$\nu_9(0) = \{-, +\} \tag{4.46}$$
$$\tag{4.47}$$

Therefore, the resulting *off-diagonal* expression is:

$$S_{12} = \frac{2}{8 \cdot 6} \left(2\, \delta_{12}(+, \dots) + 2\, \delta_{12}(-, \dots) + 3\, \delta_{12}(0, \dots) \right) \tag{4.48}$$
$$= \frac{1}{24} \left(2\, \delta_{12}(+, \dots) + 2\, \delta_{12}(-, \dots) + 3\, \delta_{12}(0, \dots) \right) \tag{4.49}$$

5 points

For this prescription, the difference is a bit more relevant, mainly in terms of the overall factor, since no one of the $\nu_5(\kappa_F)$ spaces has the maximal allowed cardinality, i.e. 3[12]

$$\nu_5(-) = \nu_9(+) = \{0\} \tag{4.50}$$
$$\nu_5(0) = \{-, +\} \tag{4.51}$$
$$\tag{4.52}$$

Therefore, the resulting *off-diagonal* expression is:

$$S_{12} = \frac{2}{4 \cdot 2} \left(2\, \delta_{12}(+, \dots) + 2\, \delta_{12}(-, \dots) + \delta_{12}(0, \dots) \right) \tag{4.53}$$
$$= \frac{1}{4} \left(2\, \delta_{12}(+, \dots) + 2\, \delta_{12}(-, \dots) + \delta_{12}(0, \dots) \right) \tag{4.54}$$

[12] Of course even 3 is completely arbitrary, as explained in eq. (4.9), and the related note, but both classes of prescriptions are perfectly *adaptive* w.r.t. this value, i.e. their definitions work perfectly fine in the general case.

$\bar{5}$ *points*

It is worth to analyze separately also this prescription: the former two are enough to exemplify the regular cases, but this one is slightly *degenerate*. Indeed, one of the spaces is actually empty.

$$\nu_5(-) = \nu_9(+) = \{-, +\} \tag{4.55}$$
$$\nu_5(0) = \{\} \tag{4.56}$$
$$\tag{4.57}$$

We need to generalize a bit the definition given above: k_m is chosen to be the least common multiple of all *non-zero* coefficients. Finally, the *off-diagonal* expression for this prescription is:

$$S_{12} = \frac{2}{4 \cdot 2} \left(\delta_{12}(+, \dots) + \delta_{12}(-, \dots)\right) \tag{4.58}$$
$$= \frac{1}{4} \left(\delta_{12}(+, \dots) + \delta_{12}(-, \dots)\right) \tag{4.59}$$

4.4.5 Summary and final remarks

Two classes of possible prescriptions consistent with the imposed requirements have been identified:

1. full space, with some zero coefficients
2. sliced space, with factorization scale dependent normalizations

No one of the two is strictly allowed by eqs. (4.1) and (4.2) of Abdul Khalek et al. 2019b, since both require the usage of non-trivial normalizations $c_i(\vec{\kappa})$, while the only normalization allowed by eq. (4.2) of the paper is a global one for the whole matrix.

Moreover, eqs. (4.1) and (4.2) of the paper themselves does not coincide with the correct and general eqs. (4.1) and (4.2) in this thesis, because eq. (4.2) in the paper is already defined at the level of a subspace V_m of the large space $\mathcal{V}$, and while this is described in the following, no proof is given of the compatibility of these subspaces as projections of the full one.

Finally, the notation used in Abdul Khalek et al. 2019b is confusing, since eq. (4.2) of the paper gives the impression that the first shift Δ_i and the second shift Δ_j are potentially evaluated on different points, while the point has to be always the same, simply the actual dependence of the shifts is on two different scales. We advocate for a more explicit and transparent syntax, at least while defining the general landscape for prescriptions (while at the individual prescription level a more concise syntax might even be useful, if properly introduced in relation to the general one).

Part II

APPLICATIONS

5 INTRINSIC CHARM

It is unclear whether heavy quarks also exist as a part of the proton wavefunction, which is determined by non-perturbative dynamics and accordingly unknown: so-called intrinsic heavy quarks, S. J. Brodsky, Hoyer, et al. 1980. It has been argued for a long time that the proton could have a sizable intrinsic component of the lightest heavy quark, the charm quark. Innumerable efforts to establish intrinsic charm in the proton (cf. S. J. Brodsky, Kusina, et al. 2015) have remained inconclusive. The study conducted in this work provided evidence for intrinsic charm by exploiting a high-precision determination of the quark-gluon content of the nucleon, Ball et al. 2021b, based on machine learning and a large experimental dataset. We disentangled the intrinsic charm component from charm-anticharm pairs arising from high-energy radiation, Ball, Bertone, Bonvini, Forte, et al. 2016. We established the existence of intrinsic charm at the 3σ level, with a momentum distribution in remarkable agreement with model predictions, S. J. Brodsky, Hoyer, et al. 1980; Hobbs et al. 2014. We also confirmed these findings by comparing to very recent data on Z-boson production with charm jets from the LHCb experiment, Aaij et al. 2021.

5.1 THE INTRINSIC CHARM EVIDENCE

While the successful framework of PDFs has by now been worked through in great detail, several key open questions remain open. One of the most controversial of these concerns the treatment of so-called heavy quarks, i.e. those whose mass is greater than that of the proton ($m_p = 0.94$ GeV). Indeed, virtual quantum effects and energy-mass considerations suggest that the three light quarks

and antiquarks (up, down, and strange) should all be present in the proton wave-function. Their PDFs are therefore surely determined by the low-energy dynamics that controls the nature of the proton as a bound state. However, it is a well-known fact, A. De Roeck and Thorne 2011; Gao et al. 2018; Kovařık et al. 2020; Rojo 2019, that in high enough energy collisions all species of quarks can be excited and hence observed inside the proton, so their PDFs are nonzero. This excitation follows from standard QCD radiation and it can be computed accurately in perturbation theory.

But then the question arises: do heavy quarks also contribute to the proton wave-function? Such a contribution is called "intrinsic", to distinguish it from that computable in perturbation theory, which originates from QCD radiation. Already since the dawn of QCD, it was argued that all kinds of intrinsic heavy quarks must be present in the proton wave-function, Stanley J. Brodsky, J. C. Collins, et al. 1984. In particular, it was suggested, S. J. Brodsky, Hoyer, et al. 1980, that the intrinsic component could be non-negligible for the charm quark, whose mass ($m_c \simeq 1.51$ GeV) is of the same order of magnitude as the mass of the proton.

This question has remained highly controversial, and indeed recent dedicated studies have resulted in disparate claims, from excluding momentum fractions carried by intrinsic charm larger than 0.5% at the 4σ level, Jimenez-Delgado et al. 2015, to allowing up to a 2% charm momentum fraction, Hou et al. 2018. A particularly delicate issue in this context is that of separating the radiative component: finding that the charm PDF is nonzero at a low scale is not sufficient to argue that intrinsic charm has been identified.

Here we present a resolution of this four-decades-long conundrum by providing unambiguous evidence for intrinsic charm in the proton. This is achieved by means of a determination of the charm PDF, Ball et al. 2021b, from the most extensive hard-scattering global dataset analyzed to date, using state-of-the-art perturbative QCD calculations, Heinrich 2021, adapted to accommodate the possibility of massive quarks inside the proton, Ball, Bertone, Bonvini, Forte, et al. 2016; Ball, Bonvini, et al. 2015; Forte, Laenen, et al. 2010, and sophisticated Machine Learning (ML) techniques, Ball et al. 2017b, 2021b; Ball, Bertone, Bonvini, Stefano Carrazza, et al. 2016. This determination is performed at Next-to-Next-to-Leading Order (NNLO) in an expansion in powers of the strong coupling, α_s, which represents the precision frontier for collider phenomenology.

The charm PDF determined in this manner includes a radiative component, and indeed it depends on the resolution scale: it is given in a four-flavor-number scheme (4 FNS), in which up, down, strange and charm quarks are subject to perturbative radiative corrections and mix with each other and the gluon as the resolution is increased. The intrinsic charm component can be disentangled from it as follows. First, we note that in the absence of an intrinsic component, the initial condition for the charm PDF is determined using perturbative matching conditions, J. C. Collins and Tung 1986, computed up to NNLO in Buza, Matiounine, Smith, and W. L. van Neerven 1998b, and recently (partly) extended up to N^3LO, Ablinger, Behring, J. Blümlein, De Freitas, Hasselhuhn, et al. 2014;

Ablinger, Behring, J. Blümlein, De Freitas, von Manteuffel, et al. 2014; Ablinger, Blumlein, et al. 2011; Ablinger, J. Blümlein, De Freitas, Hasselhuhn, von Manteuffel, Round, and Schneider 2014; Ablinger, J. Blümlein, De Freitas, Hasselhuhn, von Manteuffel, Round, Schneider, and Wißbrock 2014; Behring et al. 2014; Bierenbaum et al. 2009a,b; Johannes Blümlein et al. 2017. These matching conditions determine the charm PDF in terms of the PDFs of the three-flavor-number-scheme (3 FNS), in which only the three lightest quark flavors are radiatively corrected. Hence this perturbative charm PDF is entirely determined in terms of the three light quarks and antiquarks and the gluon. However, the 3 FNS charm quark PDF needs not vanish: in fact, if the charm quark PDF in the 4 FNS is freely parametrized and thus determined from the data, Ball, Bertone, Bonvini, Forte, et al. 2016, the matching conditions can be inverted. The 3 FNS charm PDF thus obtained is then by definition the intrinsic charm PDF: indeed, in the absence of intrinsic charm it would vanish, Ball, Bonvini, et al. 2015. Thus unlike the 4 FNS charm PDF, that includes both an intrinsic and a radiative component, the 3 FNS charm PDF is purely intrinsic. In this work we have performed this inversion at NNLO, Buza, Matiounine, Smith, and W. L. van Neerven 1998b, as well as at N^3LO, Ablinger, Behring, J. Blümlein, De Freitas, Hasselhuhn, et al. 2014; Ablinger, Behring, J. Blümlein, De Freitas, von Manteuffel, et al. 2014; Ablinger, Blumlein, et al. 2011; Ablinger, J. Blümlein, De Freitas, Hasselhuhn, von Manteuffel, Round, and Schneider 2014; Ablinger, J. Blümlein, De Freitas, Hasselhuhn, von Manteuffel, Round, Schneider, and Wißbrock 2014; Behring et al. 2014; Bierenbaum et al. 2009a,b; Johannes Blümlein et al. 2017, which as we shall see provides a handle on the perturbative uncertainty of the NNLO result.

Our starting point is the `NNPDF4.0` global analysis, Ball et al. 2021b, which provides a determination of the sum of the charm and anticharm PDFs, namely $c^+(x,Q) \equiv c(x,Q) + \bar{c}(x,Q)$, in the 4 FNS. This can be viewed as a probability density in x, the fraction of the proton momentum carried by charm, in the sense that the integral over all values of $0 \leqslant x \leqslant 1$ of $xc^+(x)$ is equal to the fraction of the proton momentum carried by charm quarks, though note that PDFs are generally not necessarily positive-definite. Our result for the 4 FNS $xc^+(x,Q)$ at the charm mass scale, $Q = m_c$ with $m_c = 1.51$ GeV, is displayed in fig. 5.1 (left). The ensuing intrinsic charm is determined from it by transforming to the 3 FNS using NNLO matching. This result is also shown in fig. 5.1 (left). The bands indicate the 68% Confidence Level (CL) interval associated with the PDF uncertainties (PDFU) in each case. Henceforth, we will refer to the 3 FNS $xc^+(x,Q)$ PDF as the intrinsic charm PDF.

The intrinsic (3 FNS) charm PDF displays a characteristic valence-like structure at large-x peaking at $x \simeq 0.4$. While intrinsic charm is found to be small in absolute terms (it contributes less than 1% to the proton total momentum), it is significantly different from zero. Note that the transformation to the 3 FNS has little effect on the peak region, because there is almost no charm radiatively generated at such large values of x: in fact, a very similar valence-like peak is already found in the 4 FNS calculation.

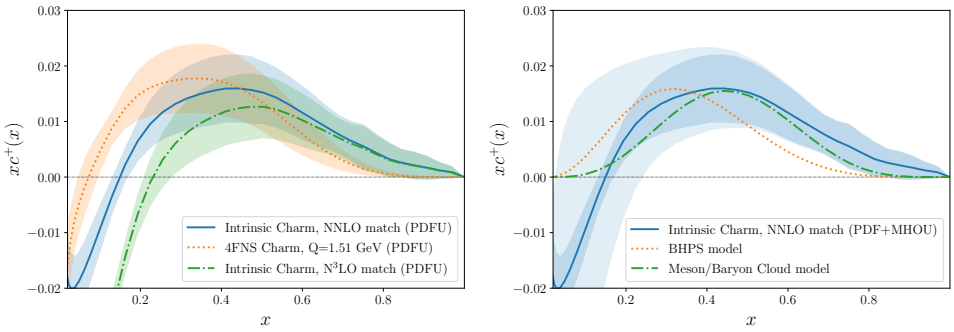

Figure 5.1: The intrinsic charm PDF and comparison with models. Left: the purely intrinsic (3 FNS) result (blue) with PDF uncertainties only, compared to the 4 FNS PDF, that includes both an intrinsic and radiative component, at $Q = m_c = 1.51$ GeV (orange). The purely intrinsic (3 FNS) result obtained using N^{3}LO matching is also shown (green). Right: the purely intrinsic (3 FNS) final result with total uncertainty (PDF + MHOU), with the PDF uncertainty indicated as a dark shaded band; the predictions from the original BHPS model, S. J. Brodsky, Hoyer, et al. 1980, and from the more recent meson/baryon cloud model, Hobbs et al. 2014, are also shown for comparison (dotted and dot-dashed curves respectively).

Because at the charm mass scale the strong coupling α_s is rather large, the perturbative expansion converges slowly. In order to estimate the effect of Missing Higher Order Uncertainties (MHOU), we have also performed the transformation from the 4 FNS NNLO charm PDF determined from the data to the 3 FNS (intrinsic) charm PDF at one order higher, namely at N^{3}LO. The result is also shown fig. 5.1 (left). Reassuringly, the intrinsic valence-like structure is unchanged. On the other hand, it is clear that for $x \lesssim 0.2$ perturbative uncertainties become very large. We can estimate the total uncertainty on our determination of intrinsic charm by adding in quadrature the PDF uncertainty and a MHOU estimated from the shift between the result found using NNLO and N^{3}LO matching.

This procedure leads to our final result for intrinsic charm and its total uncertainty, shown in fig. 5.1 (right). The intrinsic charm PDF is found to be compatible with zero for $x \lesssim 0.2$: the negative trend seen in fig. 5.1 with PDF uncertainties only becomes compatible with zero upon inclusion of theoretical uncertainties. However, at larger x even with theoretical uncertainties the intrinsic charm PDF differs from zero by about 2.5 standard deviations (2.5σ) in the peak region. This result is stable upon variations of dataset, methodology (in particular the PDF parametrization basis) and Standard Model parameters (specifically the charm mass), as demonstrated in sections 5.4 and 5.5.

Our determination of intrinsic charm can be compared to theoretical expectations. Subsequent to the original intrinsic charm model of S. J. Brodsky, Hoyer, et al. 1980 (BHPS model), a variety of other models were proposed, Hobbs et al. 2014; Hoffmann and Moore 1983; Paiva et al. 1998; Pumplin 2006; Steffens et al. 1999, cf. S. J. Brodsky, Kusina, et al. 2015 for a review. Irrespective of their specific

details, most models predict a valence-like structure at large x with a maximum located between $x \simeq 0.2$ and $x \simeq 0.5$, and a vanishing intrinsic component for $x \lesssim 0.1$. In fig. 5.1 (right) we compare our result to the original BHPS model and to the more recent meson/baryon cloud model of Hobbs et al. 2014.

As these models predict only the shape of the intrinsic charm distribution, but not its overall normalization, we have normalized them by requiring that they reproduce the same charm momentum fraction as our determination. We find remarkable agreement between the shape of our determination and the model predictions. In particular, we reproduce the presence and location of the large-x valence-like peak structure (with better agreement, of marginal statistical significance, with the meson/baryon cloud calculation), and the vanishing of intrinsic charm at small-x. The fraction of the proton momentum carried by charm quarks that we obtain from our analysis, used in this comparison to models, is $(0.62 \pm 0.28)\,\%$ including PDF uncertainties only (cf. section 5.6 for details). However, the uncertainty upon inclusion of MHOU greatly increases, and we obtain $(0.62 \pm 0.61)\,\%$, due to the contribution from the small-x region, $x \lesssim 0.2$, where the MHOU is very large, see fig. 5.1 (right). Note that in most previous analyses, Hou et al. 2018 (cf. section 5.7) intrinsic charm models (such as the BHPS model) are fitted to the data, with only the momentum fraction left as a free parameter.

We emphasize that in our analysis the charm PDF is entirely determined by the experimental data included in the PDF determination. The data with the most impact on charm are from recently measured LHC processes, which are both accurate and precise. Since these measurements are made at high scales, the corresponding hard cross-sections can be reliably computed in QCD perturbation theory.

Independent evidence for intrinsic charm is provided by the very recent LHCb measurements of Z-boson production in association with charm-tagged jets in the forward region, Aaij et al. 2021, which were not included in our baseline dataset. This process, and specifically the ratio $\mathcal{R}_j^c$ of charm-tagged jets normalized to flavor-inclusive jets, is directly sensitive to the charm PDF, Boettcher et al. 2016, and with LHCb kinematics also in the kinematic region where the intrinsic component is relevant. Following Aaij et al. 2021; Boettcher et al. 2016, we have evaluated $\mathcal{R}_j^c$ at NLO, Alioli et al. 2010; Sjostrand et al. 2008 (cf. section 5.8 for details), both with our default PDFs that include intrinsic charm, and also with an independent PDF determination in which intrinsic charm is constrained to vanish identically, so charm is determined by perturbative matching (cf. section 5.3).

In fig. 5.2 (top left) we compare the LHCb measurements of $\mathcal{R}_j^c$, provided in three bins of the Z-boson rapidity y_Z, with the theoretical predictions based on both our default PDFs as well as the PDF set in which we impose the vanishing of intrinsic charm. In fig. 5.2 (top right) we also display the correlation coefficient between the charm PDF at $Q = 100$ GeV and the observable $\mathcal{R}_j^c$, demonstrating how this observable is highly correlated to charm in a localized x region that depends on the rapidity bin. It is clear that our prediction is in excellent agreement with the LHCb measurements, while in the highest rapidity bin, which is highly correlated to the charm PDF in the region of the observed valence peak

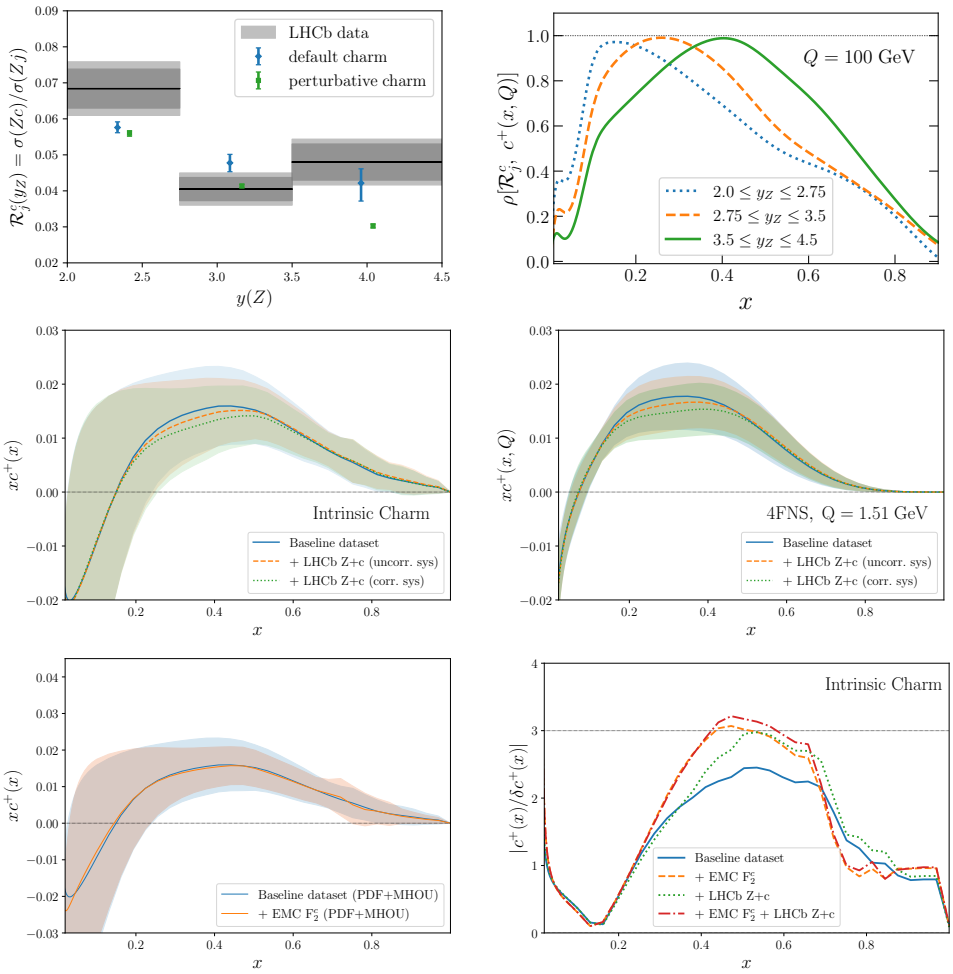

Figure 5.2: Intrinsic charm and Z+charm production at LHCb. Top left: the LHCb measurements of Z boson production in association with charm-tagged jets, $\mathcal{R}_j^c$, at $\sqrt{s} = 13$ TeV, compared with our default prediction which includes an intrinsic charm component, as well as with a variant in which we impose the vanishing of the intrinsic charm component. The thicker (thinner) bands in the LHCb data indicate the statistical (total) uncertainty, while the theory predictions include both PDF and MHOU. Top right: the correlation coefficient between the charm PDF at $Q = 100$ GeV in NNPDF4.0 and the LHCb measurements of $\mathcal{R}_j^c$ for the three y_Z bins. Center: the charm PDF in the 4 FNS (right) and the intrinsic (3 FNS) charm PDF (left) before and after inclusion of the LHCb Z+charm data. Results are shown for both experimental correlation models discussed in the text. Bottom left: the intrinsic charm PDF before and after inclusion of the EMC charm structure function data. Bottom right: the statistical significance of the intrinsic charm PDF in our baseline analysis, compared to the results obtained also including either the LHCb Z+charm (with uncorrelated systematics) or the EMC structure function data, or both.

$x \simeq 0.45$, the prediction obtained by imposing the vanishing of intrinsic charm undershoots the data at the 3σ level. Hence this measurement provides independent direct evidence in support of our result.

We have also determined the impact of these LHCb Z+charm measurements on the charm PDF. Since the experimental covariance matrix is not available, we have considered two limiting scenarios in which the total systematic uncertainty is either completely uncorrelated ($\rho_{sys} = 0$) or fully correlated ($\rho_{sys} = 1$) between rapidity bins. The charm PDF in the 4 FNS before and after inclusion of the LHCb data (with either correlation model), and the intrinsic charm PDF obtained from it, are displayed in fig. 5.2 (center left and right respectively). The bands account for both PDF and MHO uncertainties. The results show full consistency: inclusion of the LHCb $\mathcal{R}_j^c$ data leaves the intrinsic charm PDF unchanged, while moderately reducing the uncertainty on it.

In the past, the main indication for intrinsic charm came from EMC data, Aubert et al. 1983 on deep inelastic scattering with charm in the final state, Harris et al. 1996. These data are relatively imprecise, their accuracy has often been questioned, and they were taken at relatively low scales where radiative corrections are large. For these reasons, we have not included them in our baseline analysis. However, it is interesting to assess the impact of their inclusion. Results are shown in fig. 5.2 (bottom left), where we display the intrinsic charm PDF before and after inclusion of the EMC data. Just like in the case of the LHCb data we find full consistency: unchanged shape and a moderate reduction of uncertainties.

We can summarize our results through their so-called local statistical significance, namely, the size of the intrinsic charm PDF in units of its total uncertainty. This displayed in fig. 5.2 (bottom right) for our default determination of intrinsic charm, as well as after inclusion of either the LHCb Z+charm or the EMC data, or both. We find a local significance for intrinsic charm at the 2.5σ level in the region $0.3 \lesssim x \lesssim 0.6$. This is increased to about 3σ by the inclusion of either the EMC or the LHCb data, and above if they are both included. The similarity of the impact of the EMC and LHCb measurements is especially remarkable in view of the fact that they involve very different physical processes and energies.

5.2 METHODS

The strategy adopted in this work in order to determine the intrinsic charm content of the proton is based on the following observation. The assumption that there is no intrinsic charm amounts to the assumption that all 4 FNS PDFs are determined, J. C. Collins and Tung 1986, using perturbative matching conditions, Buza, Matiounine, Smith, and W. L. van Neerven 1998b, in terms of 3 FNS PDFs that do not include a charm PDF. However, these perturbative matching conditions are actually given by a square matrix that also includes a 3 FNS charm PDF. So the assumption of no intrinsic charm amounts to the assumption that if the 4 FNS PDFs are transformed back to the 3 FNS, the 3 FNS charm PDF is found

to vanish. Hence, intrinsic charm is by definition the deviation from zero of the 3 FNS charm PDF, Ball, Bonvini, et al. 2015. Note that whereas the 3 FNS charm PDF is purely intrinsic, while the 4 FNS charm PDF includes both an intrinsic and a perturbative radiative component, the 4 FNS intrinsic component is not equal to the 3 FNS charm PDF, since matching conditions reshuffle all PDFs among each other.

Intrinsic charm can then be determined through the following two steps, summarized in fig. 5.3. First, all the PDFs, including the charm PDF, are parametrized in the 4 FNS at an input scale Q_0 and evolved using NNLO perturbative QCD to $Q \neq Q_0$. These evolved PDFs can be used to compute physical cross-sections, also at NNLO, which then are compared to a global dataset of experimental measurements. The result of this first step in our procedure is a Monte Carlo (MC) representation of the probability distribution for the 4 FNS PDFs at the input parametrization scale Q_0.

Next, this 4 FNS charm PDF is transformed to the 3 FNS at some scale matching scale Q_c. Note that the choice of both Q_0 and Q_c are immaterial. The former because perturbative evolution is invertible, so results for the PDFs do not depend on the choice of parametrization scale Q_0. The latter because the 3 FNS charm is scale independent, so it does not depend on the value of Q_c. Both statements of course hold up to fixed perturbative accuracy, and are violated by MHO corrections. In practice, we parametrize PDFs at the scale $Q_0 = 1.65$ GeV and perform the inversion at a scale chosen equal to the charm mass $Q_c = m_c = 1.51$ GeV.

The scale-independent 3 FNS charm PDF is then the sought-for intrinsic charm.

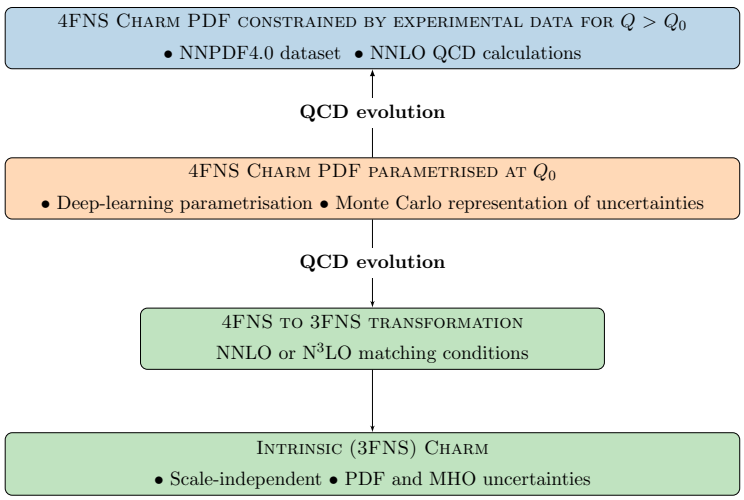

Figure 5.3: The 4 FNS charm PDF is parametrized at Q_0 and evolved to all Q, where it is constrained by the `NNPDF4.0` global dataset. Subsequently, it is transformed to the 3 FNS where (if nonzero) it provides the intrinsic charm component.

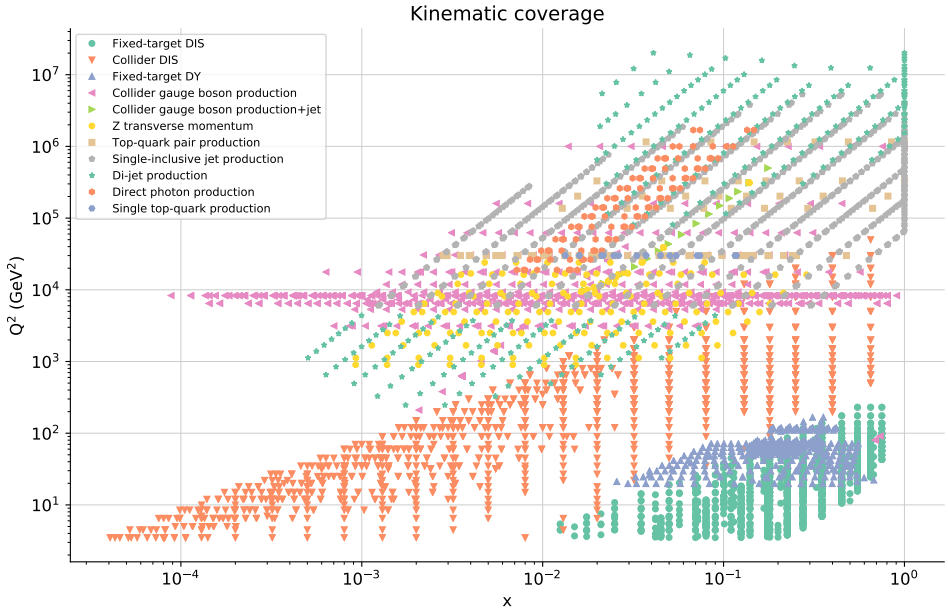

Figure 5.4: The kinematic coverage in the (x, Q) plane covered by the 4618 cross-sections used for the determination of the charm PDF in the present work. These cross-sections have been classified into the main different types of processes entering the global analysis.

GLOBAL QCD ANALYSIS. The 4 FNS charm PDF and its associated uncertainties is determined by means of a global QCD analysis within the NNPDF4.0 framework. All PDFs, including the charm PDF, are parametrized at $Q_0 = 1.65$ GeV in a model-independent manner using a neural network, which is fitted to data using supervised machine learning techniques. The Monte Carlo replica method is deployed to ensure a faithful uncertainty estimate. Specifically, we express the 4 FNS total charm PDF ($c^+ = c + \bar{c}$) in terms of the output neurons associated to the quark singlet Σ and non-singlet T_{15} distributions, cf. Sect. 3.1 of Ball et al. 2021b, as

$$xc^+(x, Q_0; \theta) = \left(x^{\alpha_\Sigma}(1-x)^{\beta_\Sigma}\, \mathrm{NN}_\Sigma(x,\theta) - x^{\alpha_{T_{15}}}(1-x)^{\beta_{T_{15}}}\, \mathrm{NN}_{T_{15}}(x,\theta) \right)/4\,, \tag{5.1}$$

where $\mathrm{NN}_i(x,\theta)$ is the i-th output neuron of a neural network with input x and parameters θ, and (α_i, β_i) are preprocessing exponents. A crucial feature of eq. (5.1) is that no *ad hoc* specific model assumptions are used: the shape and size of $xc^+(x, Q_0)$ are entirely determined from experimental data. Hence, our determination of the 4 FNS fitted charm PDF, and thus of the intrinsic charm, is unbiased.

The neural network parameters θ in eq. (5.1) are determined by fitting an extensive global dataset that consists of 4618 cross-sections from a wide range of different processes, measured over the years in a variety of fixed-target and collider experiments (cf. Ball et al. 2021b for a complete list). Figure 5.4 displays the

kinematic coverage in the (x, Q) plane covered by these cross-sections, where Q is the scale, and x is the parton momentum fraction that correspond to leading-order kinematics. Many of these processes provide direct or indirect sensitivity to the charm content of the proton. Particularly important constraints come from W and Z production from ATLAS, CMS, and LHCb as well as from neutral and charged current Deep Inelastic Scattering (DIS) structure functions from HERA. The 4 FNS PDFs at the input scale Q_0 are related to experimental measurements at $Q \neq Q_0$ by means of NNLO QCD calculations, including the FONLL-C general-mass scheme for DIS, Forte, Laenen, et al. 2010, generalized to allow for fitted charm, Ball, Bertone, Bonvini, Forte, et al. 2016.

We have verified (cf. sections 5.4 and 5.5) that the determination of 4 FNS charm PDF eq. (5.1) and the ensuing 3 FNS intrinsic charm PDF are stable upon variations of methodology (PDF parametrization basis), input dataset, and values of Standard Model parameters (the charm mass). We have also studied the stability of our results upon replacing the current `NNPDF4.0` methodology Ball et al. 2021b with the previous `NNPDF3.1` methodology, Ball et al. 2017a. It turns out that results are perfectly consistent. Indeed, the old methodology leads to somewhat larger uncertainties, corresponding to a moderate reduction of the local statistical significance for intrinsic charm, and to a central value which is within the smaller error band of our current result.

A determination in which the vanishing of intrinsic charm is imposed has also been performed. In this case, the fit quality significantly deteriorates: the values of the χ^2 per data point of 1.162, 1.26, and 1.22 for total, Drell-Yan, and neutral-current DIS data respectively, found when fitting charm, are increased to 1.198, 1.31, 1.28 when the vanishing of intrinsic charm is imposed. The absolute worsening of the total χ^2 when the vanishing of intrinsic charm is imposed is therefore of 166 units, corresponding to a 2σ effects in units of $\sigma_{\chi^2} = \sqrt{2n_{\text{dat}}}$.

CALCULATION OF THE 3 FNS CHARM PDF. The Monte Carlo representation of the probability distribution associated to the 4 FNS charm PDF determined by the global analysis contains an intrinsic component mixed with a perturbatively generated contribution, with the latter becoming larger in the $x \lesssim 0.1$ region as the scale Q is increased. In order to extract the intrinsic component, we transform PDFs to the 3 FNS at the scale $Q_c = m_c = 1.51$ GeV using `EKO`, a novel Python open source PDF evolution framework (cf. chapter 2). In its current implementation, `EKO` performs QCD evolution of PDFs to any scale up to NNLO. For the sake of the current analysis, N^{3}LO matching conditions have also been implemented, by using the results of Ablinger, Behring, J. Blümlein, De Freitas, Hasselhuhn, et al. 2014; Ablinger, Behring, J. Blümlein, De Freitas, von Manteuffel, et al. 2014; Ablinger, Blumlein, et al. 2011; Ablinger, J. Blümlein, De Freitas, Hasselhuhn, von Manteuffel, Round, and Schneider 2014; Ablinger, J. Blümlein, De Freitas, Hasselhuhn, von Manteuffel, Round, Schneider, and Wißbrock 2014; Behring et al. 2014; Bierenbaum et al. 2009a,b; Johannes Blümlein et al. 2017 for $\mathcal{O}(\alpha_s^3)$ operator matrix elements so that the direct and inverse transformations

from the 3 FNS to the 4 FNS can be performed at one order higher. The N^{3}LO contributions to the matching conditions are a subset of the full N^{3}LO terms that would be required to perform a PDF determination to one higher perturbative order, and would also include currently unknown N^{3}LO contributions to QCD evolution. Therefore, our results have NNLO accuracy and we can only use the N^{3}LO contributions to the $\mathcal{O}(\alpha_s^3)$ corrections to the heavy quark matching matching conditions as a way to estimate the the size of the missing higher orders. Indeed, these corrections have a very significant impact on the perturbatively generated component, see section 5.3. They become large for $x \lesssim 0.1$, which coincides with the region dominated by the perturbative component of the charm PDF, and are relatively small for the valence region where intrinsic charm dominates.

Z PRODUCTION IN ASSOCIATION WITH CHARM-TAGGED JETS. The production of Z bosons in association with charm-tagged jets (or alternatively, with identified D mesons) at the LHC is directly sensitive to the charm content of the proton via the dominant $gc \to Zc$ partonic scattering process. Measurements of this process at the forward rapidities covered by the LHCb acceptance provide access to the large-x region where the intrinsic contribution is expected to dominate. This is in contrast with the corresponding measurements from ATLAS and CMS, which only become sensitive to intrinsic charm at rather larger values of p_T^Z than those currently accessible experimentally.

We have obtained theoretical predictions for Z+charm production at LHCb with `NNPDF4.0`, based on NLO QCD calculations using POWHEG-BOX interfaced to PYTHIA8 with the Monash 2013 tune for showering, hadronization, and underlying event. Acceptance requirements and event selection follow the LHCb analysis, where in particular charm jets are defined as those anti-k_T $R = 0.5$ jets containing a reconstructed charmed hadron. The ratio between c-tagged and untagged Z+jet events can then be compared with the LHCb measurements

$$\mathcal{R}_j^c(y_Z) \equiv \frac{N(c \text{ tagged jets}; y_Z)}{N(\text{ jets}; y_Z)} = \frac{\sigma(pp \to Z + \text{ charm jet}; y_Z)}{\sigma(pp \to Z + \text{ jet}; y_Z)}, \tag{5.2}$$

as a function of the Z boson rapidity y_Z (see section 5.8 for details). The more forward the rapidity y_Z, the higher the values of the charm momentum x being probed. Furthermore, we have also included the LHCb measurements in the global PDF determination by means of the Bayesian reweighting (cf. section 5.8).

5.3 THE PERTURBATIVE CHARM PDF

In the absence of intrinsic charm, the charm PDF is fully determined by perturbative matching conditions, i.e. by the matrix $\mathbf{A}^{(n_f)}(Q_c^2)$ in eq. (2.11). We will denote the charm PDF thus obtained "perturbative charm PDF", for short. The PDF uncertainty on the perturbative charm PDF is directly related to that of the

light quarks and especially the gluon, and is typically much smaller than the uncertainty on our default charm PDF, that includes intrinsic charm. Here and in the following we will refer to our final result, as shown in fig. 5.1 (right) as "default". It should be noticed that the matching conditions for charm are nontrivial starting at NNLO: at NLO the perturbative charm PDF vanishes at threshold. Hence, having implemented in EKO also the N^3LO matching conditions, we are able to assess the MHOU of the perturbative charm at the matching scale Q_c, by comparing results obtained at the first two nonvanishing perturbative orders.

As already mentioned, see also fig. 5.2 (top left) in section 5.1, we have constructed a PDF set with perturbative charm, in which the full PDF determination from the global dataset leading to the `NNPDF4.0` PDF set is repeated, but now with the assumption of vanishing intrinsic charm, i.e. with a perturbative charm PDF. This perturbative charm PDF is compared to our default result in fig. 5.5 (left), where the 4 FNS perturbative charm PDF at scale $Q_c = m_c$ obtained using either NNLO or N^3LO under the assumption of no intrinsic charm are shown, together with our result allowing for intrinsic charm. It is clear that while on the one hand, the PDF uncertainty on the perturbative charm PDF is indeed tiny, on the other hand the difference between the result for perturbative charm obtained using NNLO or N^3LO matching is large, and in fact larger at small x than the difference between perturbative charm and our default (intrinsic) result.

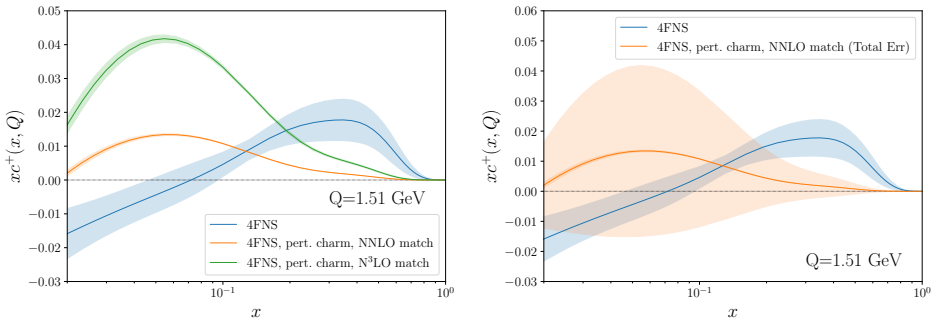

Figure 5.5: Left: the perturbative charm PDF at $Q = 1.51$ GeV obtained from NNLO PDFs using NNLO and N^3LO matching conditions. Right: the NNLO perturbative charm PDF including the MHOU computed as the difference between NNLO and N^3LO matching. In both plots our default (intrinsic) charm PDF is also shown for comparison.

In the same manner as we used the difference between the results obtained from inversion of NNLO and N^3LO matching as an estimate of the MHOU on intrinsic charm, we may use the difference between the 4 FNS perturbative charm obtained from NNLO and N^3LO matching as an estimate of the MHOU on perturbative charm at the scale Q_c. The total uncertainty is found by adding this in quadrature to the PDF uncertainty (which however in practice is negligible).

The result is shown in fig. 5.5 (right). Within this total uncertainty there is now good agreement between our intrinsic charm result and perturbative charm for all $x \lesssim 0.2$. On the other hand, there is a clear deviation for larger x. We may view the difference between the 4 FNS default result and the 4 FNS perturbative charm as the intrinsic component in the 4 FNS, and indeed it is clear from fig. 5.5 that the 4 FNS intrinsic component is sizable and positive at large x. This is of course consistent with our main finding that we only see evidence of intrinsic charm for large $x \gtrsim 0.2$, while for smaller x our result for the charm PDF is compatible with zero, as demonstrated by fig. 5.1 (right) in section 5.1.

5.4 STABILITY OF THE 4 FNS CHARM PDF

The main input to our determination of intrinsic charm is the 4 FNS charm PDF extracted from high-energy data. While this determination comes with an uncertainty estimate, it is important to verify that this adequately reflects the various sources of uncertainty, and that there are no further sources of uncertainty that may be unaccounted for. To this purpose, here we assess the stability of our results first, upon the choice of underlying dataset, next upon changes in methodology, and finally, upon variation of standard model parameters. In each case we verify stability upon the most important possible source of instability: respectively, the use of collider vs. fixed target and deep-inelastic vs. hadronic data (dataset); the choice of parametrization basis (methodology); and the value of the charm quark mass (standard model parameters). As a final consistency check, we compare our result with that which we would have obtained by using the same input dataset, but the previous NNPDF3.1 fitting methodology. Because we are interested in intrinsic charm, in all comparisons we focus on the large-x region in which the intrinsic valence-like peak is found. In this section, the 4 FNS charm PDF is displayed at the scale $Q = 1.65$ GeV so that results for all fit variants, including those with with different m_c values, can be shown at a common scale.

DEPENDENCE ON THE CHOICE OF DATASET. We now study the stability of the 4 FNS charm determination upon variation of the underlying data, which also allows us to identify the datasets or groups of processes that provide the leading constraints on intrinsic charm. To this purpose, we have repeated our PDF determination using a variety of subsets of the global dataset used for our default determination. Results are shown in fig. 5.6, where we compare the result using the baseline dataset to determinations performed by adding to the baseline the EMC charm structure function data (already discussed in the main text); by only including DIS data; by only including collider data (HERA, Tevatron and LHC); and by removing the LHCb W and Z production data.

As already noted in the main text in the case of the 3 FNS result, we find that the extra information provided by the EMC F_2^c data is subdominant in comparison to that from the global dataset. The result is stable and only a moderate

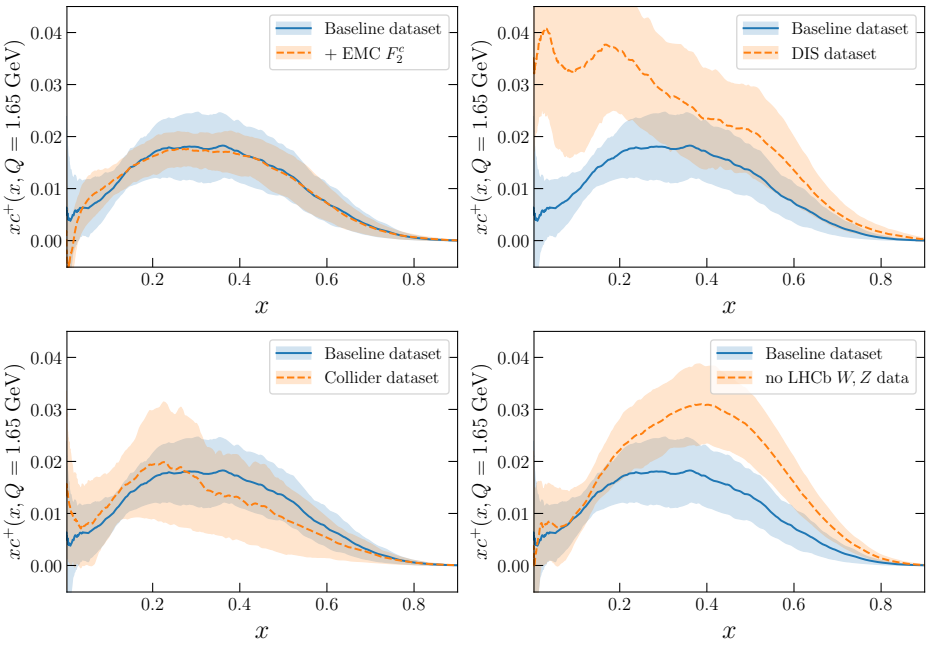

Figure 5.6: The dependence of the 4 FNS charm PDF at $Q = 1.65$ GeV on the input dataset. We compare the baseline result with that obtained by also including EMC F_2^c data (top left), only including DIS data (top right), only including collider data (bottom left) and removing LHCb gauge boson production data (bottom right).

uncertainty reduction at the peak is observed. It is interesting to contrast this with the previous NNPDF study Ball, Bertone, Bonvini, Stefano Carrazza, et al. 2016, in which the global fit provided only very loose constraints on the charm PDF, which was then determined mostly by the EMC data. Indeed, a DIS-only fit (for which most data were already available at the time of the previous determination) determines charm with very large uncertainties. On the other hand, both the central value and uncertainty found in the collider-only fit are quite similar to the baseline result. This shows that the dominant constraint is now coming from collider, and specifically hadron collider data (indeed, as we have seen constraints from DIS data are quite loose). Among these, LHCb data (which are taken at large rapidity and thus impact PDFs at large and small x) are especially important, as demonstrated by the increase in uncertainty when they are removed.

In all these determinations, the charm PDF at $x \simeq 0.4$ remains consistently nonzero and positive, thus emphasizing the stability of our results.

DEPENDENCE ON THE PARAMETRIZATION BASIS. Among the various methodological choices, a possibly critical one is the choice of basis functions. Specifi-

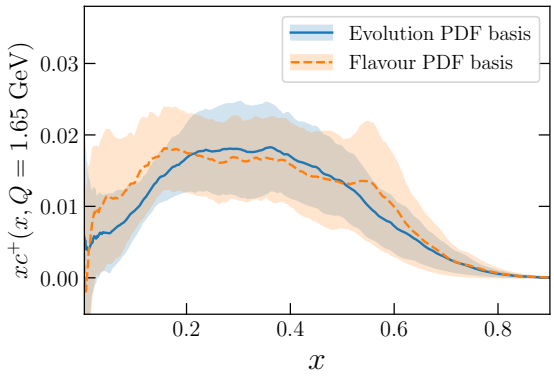

Figure 5.7: The default 4 FNS charm PDF at $Q = 1.65$ GeV compared to a result obtained by parametrizing PDFs in the flavor basis instead of the evolution basis.

cally, in our default analysis, the output of the neural network does not provide the individual quark flavor and antiflavor PDFs, but rather linear combinations corresponding to the so-called evolution basis Ball et al. 2021b. Indeed, our charm PDF is given in eq. (5.1) as the linear combination of the two basis PDFs Σ and T_{15}. One may thus ask whether this choice may influence the final results for individual quark flavors, specifically charm. Given that physical results are basis independent, the outcome of a PDF determination should not depend on the basis choice.

In order to check this, we have repeated the PDF determination, but now using the flavor basis, see Sect. 3.1 of Ball et al. 2021b, in which each of the neural network output neurons now correspond to individual quark flavors, so in particular, instead of eq. (5.1), one has

$$xc^+(x, Q_0; \theta) = (1-x)^{\beta_{c^+}} \, \mathrm{NN}_{c^+}(x, \theta)\,, \tag{5.3}$$

where $\mathrm{NN}_{c^+}(x, \theta)$ indicates the value of the output neuron associated to the charm PDF c^+. The 4 FNS charm PDFs determined using either basis are compared in fig. 5.7 at $Q = 1.65$ GeV. We find excellent consistency, and in particular the valence-like structure at high-x is independent of the choice of parametrization basis.

DEPENDENCE ON THE CHARM MASS. The kinematic threshold for producing charm perturbatively depends on the value of the charm mass. Therefore the perturbative contribution to the 4 FNS charm PDF, and thus the whole charm PDF if one assumes perturbative charm, depends strongly on the value of the charm mass. On the other hand, the intrinsic charm PDF is of nonperturbative origin, so it should be essentially independent of the numerical value of the charm mass that is used in perturbative computations employed in its determination

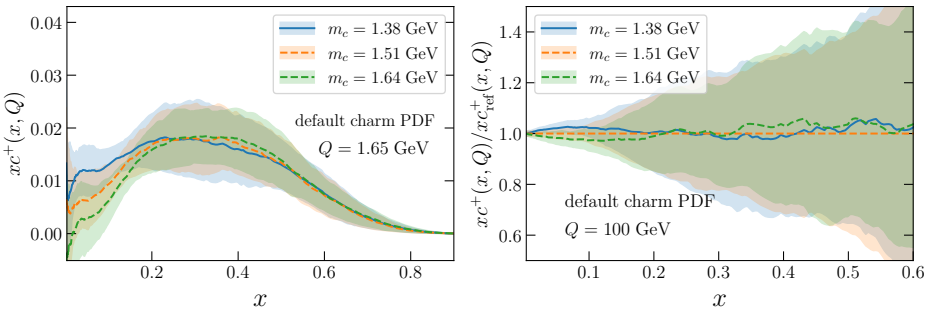

Figure 5.8: The 4 FNS charm PDF determined using three different values of the charm mass. The absolute result (left) is shown at $Q = 1.65$ GeV, while the ratio to the default value $m_c = 1.51$ GeV (right) used elsewhere in this paper is shown at $Q = 100$ GeV.

(though it will of course depend on the true underlying physical value of the charm mass).

In order to study this mass dependence, we have repeated our determination using different values for the charm mass. The definition of the charm mass which is relevant for kinematic thresholds is the pole mass, for which we adopt the value recommended by the Higgs cross-section working group de Florian et al. 2016a based on the study of Bauer et al. 2004, namely $m_c = 1.51 \pm 0.13$ GeV. Results are shown in fig. 5.8, where our default charm PDF determination with $m_c = 1.51$ GeV is repeated with $m_c = 1.38$ GeV and $m_c = 1.64$ GeV. In order to understand these results note that this is the 4 FNS PDF, so it includes both a nonperturbative and a perturbative component. The latter is strongly dependent on the charm mass, but of course the data correspond to the unique true value of the mass and the mass dependence of the perturbative component is present only due to our ignorance of the actual true value. When determining the PDF from the data, as we do, we expect this spurious dependence to be to some extent reabsorbed into the fitted PDF. So we expect results to display a moderate dependence on the charm mass — full independence should hold for the intrinsic (3 FNS) PDF and will be investigated in section 5.5.

In fig. 5.9 the same result is shown, but now for the perturbative charm PDF discussed in section 5.3, so the charm PDF is of purely perturbative origin and fully determined by the strongly mass-dependent matching condition. This dependence is clearly seen in the plots. Furthermore, comparison with fig. 5.8 shows that indeed this spurious dependence is partly reabsorbed in the fit when the charm PDF is determined from the data, so that the residual mass dependence is moderate. In particular, the large-x valence peak, which is dominated by the intrinsic component, is very stable.

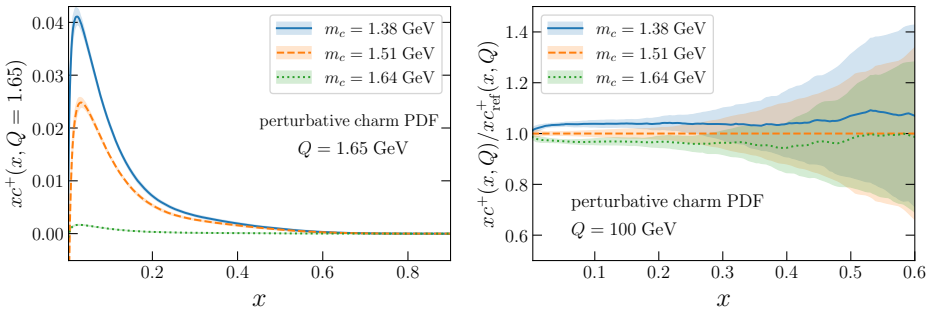

Figure 5.9: The same as fig. 5.8 but now for the perturbative charm PDF.

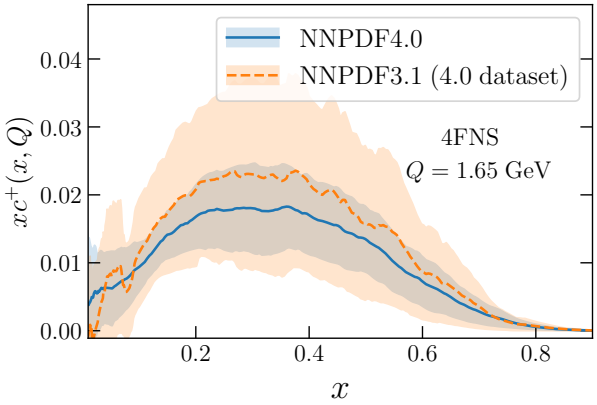

Figure 5.10: Same as fig. 5.7, comparing the baseline determination of the 4 FNS charm PDF, based on `NNPDF4.0`, with that obtained from the same dataset using the `NNPDF3.1` fitting methodology.

COMPARISON WITH `NNPDF3.1`. fig. 5.10 compares the baseline determination of the 4 FNS charm PDF based on `NNPDF4.0` with that obtained from the same input dataset but using instead the `NNPDF3.1` fitting methodology and related settings such those related to positivity and integrability. Results are fully consistent between the two methodologies, with our default determination exhibiting reduced uncertainties due to the various improvements implemented in the `NNPDF4.0` analysis framework.

5.5 STABILITY OF THE 3 FNS CHARM CALCULATION

We now repeat the stability and uncertainty study of the previous section, but for our final result, namely the intrinsic charm PDF. The main difference to be kept in mind is that the uncertainty now also includes the dominant MHOU, due to the matching condition required in order to determine the 3 FNS PDF from the 4 FNS result. In order to get a complete picture, we now add a further set of dataset variations.

DEPENDENCE ON THE INPUT DATASET. fig. 5.11 displays the dataset variations shown in fig. 5.6, now for the intrinsic (3 FNS) charm PDF, but with the total uncertainty now being the sum in quadrature of the PDF and Missing Higher Order Uncertainties, with the latter determined as the difference between results obtained using NNLO and N^3LO matching. Additionally, we also performed a few extra dataset variations: a fit without any W, Z production data from ATLAS and CMS, a fit without jet data, a fit without Z p_T measurements, and a fit without HERA structure function data. Note that the collider-only dataset includes both HERA electron-proton collider data and Tevatron and LHC hadron collider data, but not fixed-target Deep Inelastic Scattering and Drell–Yan production data.

Results are qualitatively very similar to those seen in the 4 FNS, a consequence of the fact that we are focusing on the large-x region where the effect of the matching is moderate, though now the presence of a valence-like peak in all determinations is even clearer, specifically for the DIS-only fit where it was less pronounced in the 4 FNS. Note however that the DIS-only determination exhibits larger uncertainties (up to factor 2) and point-by-point fluctuations, and is dominated by relatively old fixed-target measurements. Comparison of all the dataset variations shows that, in terms of their impact on intrinsic charm, hadron collider data are generally more important that deep-inelastic data, that among the former the LHCb inclusive W, Z data are playing a dominant role, and that jet observables also play a non-negligible role.

It should be stressed that the agreement between results found using DIS data and hadron collider data is highly nontrivial, since in the region relevant for intrinsic charm these determinations are based on disjoint datasets and are affected by very different theoretical and experimental uncertainties: in particular, potential higher-twist effects in the DIS observables are highly suppressed for collider observables. this respect, a DIS-only determination of intrinsic charm is potentially affected by sources of theory uncertainties, such as higher twists, which are not accounted for in global PDF determinations.

We conclude that the characteristic valence-like peak structure at large-x predicted by non-perturbative intrinsic charm models (fig. 5.1 in section 5.1) is always present even under very significant changes of the dataset.

DEPENDENCE ON THE PARAMETRIZATION BASIS. Figure 5.12 displays the comparison between the intrinsic charm PDF determined with the default evolution

basis choice, and the flavor basis. Complete consistency of central values is found, with somewhat larger uncertainties in the case of the flavor basis, due to the more challenging fitting environment in this basis (see the discussion in Ball et al. 2021b).

DEPENDENCE ON THE CHARM MASS VALUE. The study of the charm mass dependence is particularly interesting, because the intrinsic component should be independent of it, hence the residual dependence seen in fig. 5.8 in the 4 FNS, due to the mass dependence of the perturbative component that could not be reabsorbed in the fitting, should no longer be present. Results are shown in fig. 5.13, and it is apparent that indeed the dependence on the charm mass has all but disappeared.

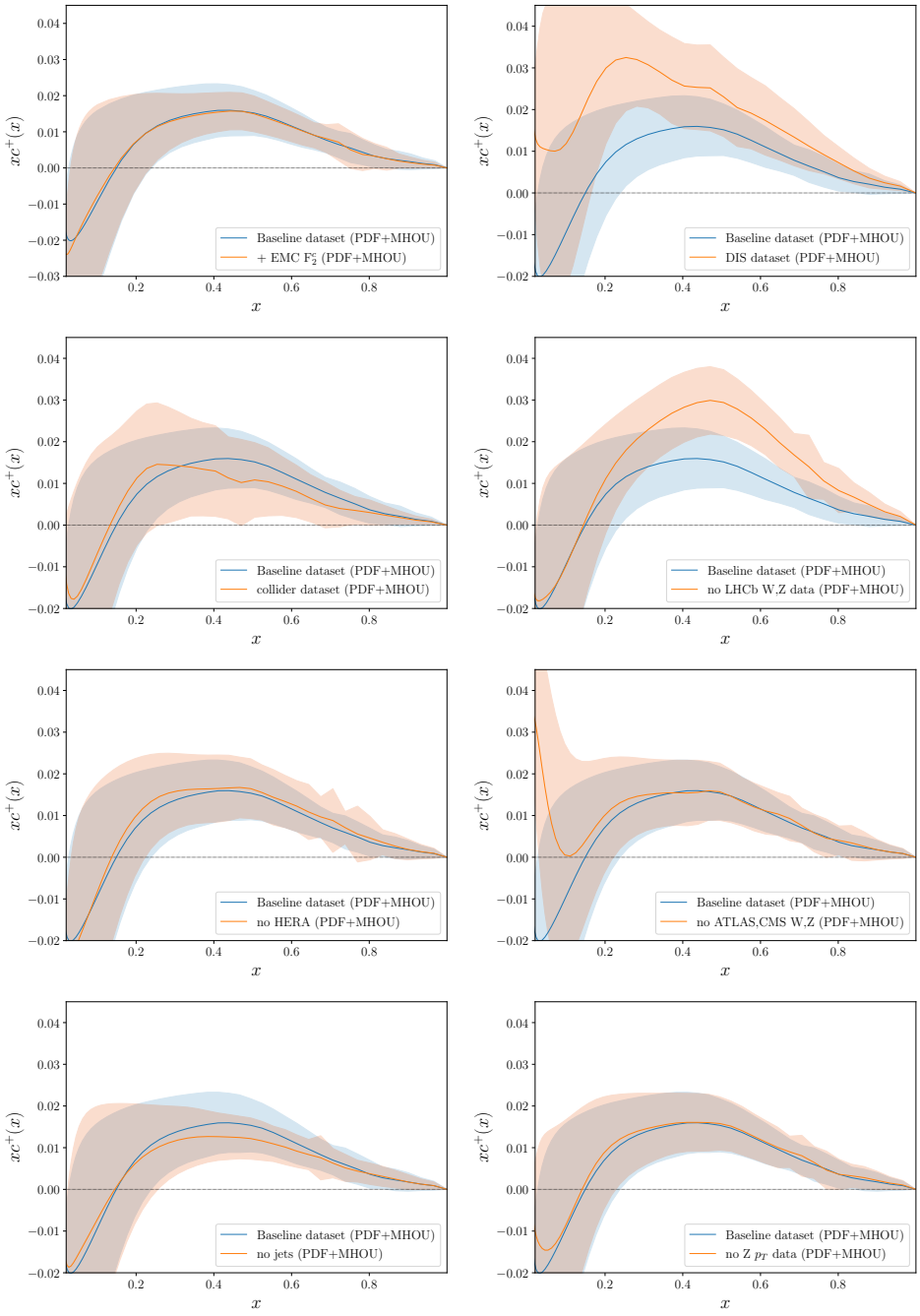

Figure 5.11: Same as fig. 5.6 for the intrinsic charm (3 FNS) PDF (top four plots), now also including four additional dataset variations: no ATLAS and CMS W, Z production data (third row left), no jet data (third row right), no Z p_T measurements (bottom row left), no HERA DIS data (bottom row right). The error band indicates the PDF uncertainties combined in quadrature with the MHOUs.

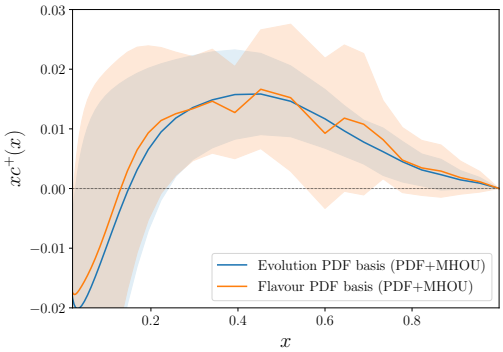

Figure 5.12: Same as fig. 5.7 for the intrinsic (3 FNS) charm.

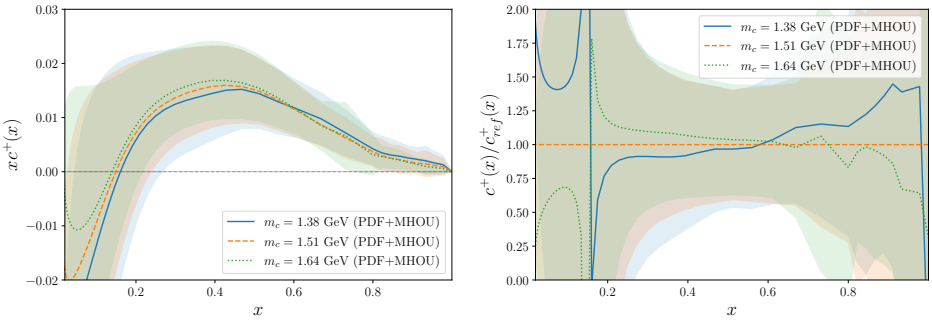

Figure 5.13: Same as fig. 5.8, now for the intrinsic (3 FNS) charm PDF. Note that the intrinsic charm PDF is scale independent.

5.6 THE CHARM MOMENTUM FRACTION

The fraction of the proton momentum carried by charm quarks is given by

$$[c] = \int_0^1 dx\, xc^+(x, Q^2)\,. \tag{5.4}$$

Model predictions, as mentioned, are typically provided up to an overall normalization, which in turn determines the charm momentum fraction. Consequently, the momentum fraction is often cited as a characteristic parameter of intrinsic charm. It should however be borne in mind that, even in the absence of intrinsic charm, this charm momentum fraction is nonzero due to the perturbative contribution.

In table 5.1 we indicate the values of the charm momentum fraction in the 3 FNS for our default charm determination and in the 4 FNS (at $Q = 1.65$ GeV) both for the default result and for perturbative charm PDF (see section 5.3). We provide results for three different values of the charm mass m_c and indicate separately the PDF and the MHO uncertainties. The 3 FNS result is scale-

Scheme	Q	Charm PDF	m_c	$[c]$ (%)
3 FNS	–	default	1.51 GeV	$0.62 \pm 0.28_{\text{pdf}} \pm 0.54_{\text{mhou}}$
3 FNS	–	default	1.38 GeV	$0.47 \pm 0.27_{\text{pdf}} \pm 0.62_{\text{mhou}}$
3 FNS	–	default	1.64 GeV	$0.77 \pm 0.28_{\text{pdf}} \pm 0.48_{\text{mhou}}$
4 FNS	1.65 GeV	default	1.51 GeV	$0.87 \pm 0.23_{\text{pdf}}$
4 FNS	1.65 GeV	default	1.38 GeV	$0.94 \pm 0.22_{\text{pdf}}$
4 FNS	1.65 GeV	default	1.64 GeV	$0.84 \pm 0.24_{\text{pdf}}$
4 FNS	1.65 GeV	perturbative	1.51 GeV	$0.346 \pm 0.005_{\text{pdf}} \pm 0.44_{\text{mhou}}$
4 FNS	1.65 GeV	perturbative	1.38 GeV	$0.536 \pm 0.006_{\text{pdf}} \pm 0.49_{\text{mhou}}$
4 FNS	1.65 GeV	perturbative	1.64 GeV	$0.172 \pm 0.003_{\text{pdf}} \pm 0.41_{\text{mhou}}$

Table 5.1: The charm momentum fraction, eq. (5.4). We show results both in the 3 FNS and the 4 FNS (at $Q = 1.65$ GeV) for our default charm, and also in the 4 FNS for perturbative charm. We provide results for three different values of the charm mass m_c and indicate separately the PDF and the MHO uncertainties.

independent, it corresponds to the momentum fraction carried by intrinsic charm, and it vanishes identically by assumption in the perturbative charm case. The 4 FNS result corresponds to the scale-dependent momentum fraction that combines the intrinsic and perturbative contribution, while of course it contains only the perturbative contribution in the case of perturbative charm. As discussed in section 5.3, the large uncertainty associated to higher order corrections to the matching conditions affects the 3 FNS result (intrinsic charm) in the default case, in which the charm PDF is determined from data in the 4 FNS scheme, while it affects the 4 FNS result for perturbative charm, that is determined assuming the vanishing of the 3 FNS result.

For our default determination, the charm momentum fraction in the 4 FNS at low scale differs from zero at the 3σ level. However, it is not possible to tell whether this is of perturbative or intrinsic origin, because, due to the large MHOU in the matching condition, the intrinsic (3 FNS) charm momentum fraction is compatible with zero. This large uncertainty is entirely due to the small $x \lesssim 0.2$ region, see see fig. 5.1 (right). Accordingly, for perturbative charm the low-scale 4 FNS momentum fraction is compatible with zero. Consistently with the results of section 5.4, the 4 FNS result is essentially independent of the value of the charm mass, but it becomes strongly dependent on it if one assumes perturbative charm.

The 4 FNS charm momentum fraction is plotted as a function of scale in fig. 5.14, both in the default case and for perturbative charm, with the 3 FNS values and the detail of the low-Q 4 FNS results shown in an inset. The dependence on the value of the charm mass is shown in fig. 5.15. The large MHOUs on the 3 FNS result, and on the 4 FNS result in the case of perturbative charm, are apparent. The stability of the default result upon variation of the value of m_c,

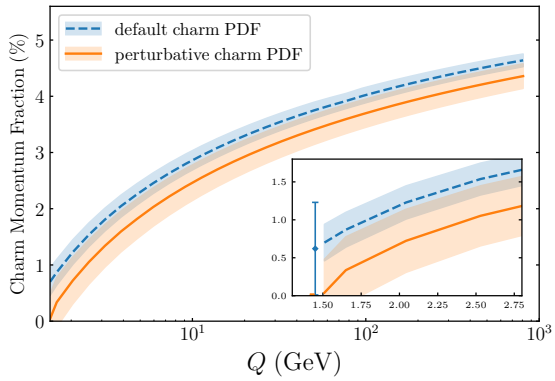

Figure 5.14: The 4 FNS charm momentum fraction in NNPDF4.0 as a function of scale Q, both for the default and perturbative charm cases, for a charm mass value of $m_c = 1.51$ GeV. The inset zooms on the low-Q region and includes the 3 FNS (default) result from table 5.1. Note that the uncertainty includes the MHOU for the 3 FNS default and 4 FNS perturbative charm cases, while it is the PDF uncertainty for the 4 FNS default charm case.

and the strong dependence of the perturbative charm result on m_c, are also clear. Both the large MHOU uncertainty, and the strong dependence on the value of m_c for perturbative charm are seen to persist up to large scales.

It is interesting to understand in detail the impact of the MHOU on the momentum fraction carried by intrinsic charm. To this purpose, we have computed the truncated momentum integral, i.e. eq. (5.4) but only integrated down to some lower integration limit $x_{\rm min}$:

$$[c]_{\rm tr}(x_{\rm min}) \equiv \int_{x_{\rm min}}^{1} dx\, x c^+(x, Q^2)\,. \tag{5.5}$$

Note than in the 3 FNS $xc^+(x)$ does not depend on scale, so this becomes a scale-independent quantity. The result for our default intrinsic charm determination is displayed in fig. 5.16, as a function of of the lower integration limit $x_{\rm min}$. It is clear that for $x_{\rm min} \gtrsim 0.2$ the truncated momentum fraction differs significantly from zero, thereby providing evidence for intrinsic charm with similar statistical significance as the local pull shown in fig. 5.2 bottom left. For $x \lesssim 0.2$ this significance is then washed out by the large MHOUs.

Hence, while the total momentum fraction has been traditionally adopted as a measure of intrinsic charm, our analysis shows that, once MHOUs are accounted for, the information provided by the total momentum fraction is limited, at least with current data and theory.

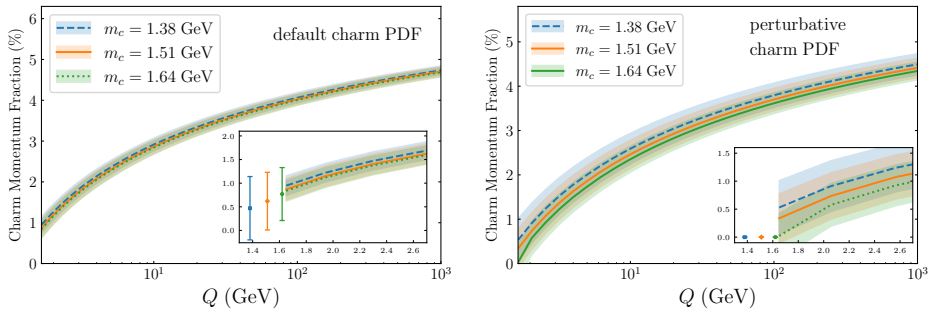

Figure 5.15: Same as fig. 5.14 for different values of the charm mass. Note that the 3 FNS momentum fraction for perturbative charm vanishes identically by assumption.

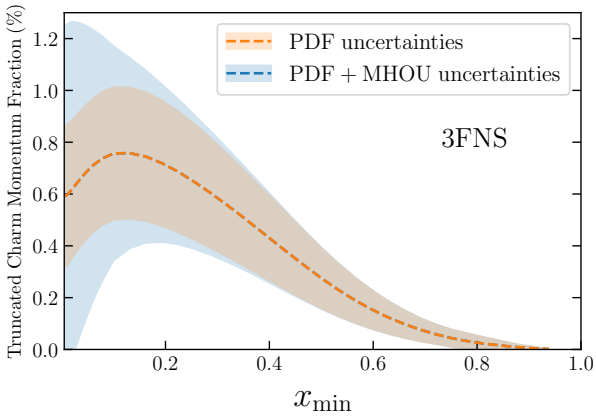

Figure 5.16: The value of the truncated charm momentum integral, eq. (5.5), as a function of the lower integration limit $x_{\min}$ for our baseline determination of the 3 FNS intrinsic charm PDF. We display separately the PDF and the total (PDF + MHOU) uncertainties.

5.7 COMPARISON WITH CT14IC

The possibility of an intrinsic charm component was recently studied in Hou et al. 2018, by modifying the CT14 PDF set, with the initial 4 FNS charm PDF taken equal to the BHPS model S. J. Brodsky, Hoyer, et al. 1980 form with the normalization fitted as a free parameter. A 4 FNS charm PDF with uncertainties at Q = 1.3 GeV was then constructed by taking the BHPS model with best-fit normalization as central value (called the 'BHPS1 model' in Hou et al. 2018); the lower edge of the uncertainty band was taken to coincide with the standard CT14 charm PDF (i.e. the charm PDF determined by perturbative matching from the

3 FNS to the 4 FNS); the upper edge of the uncertainty band was taken as the BHPS model but with normalization fixed to the upper 90% CL limit (called the 'BHPS2 model' in Hou et al. 2018).

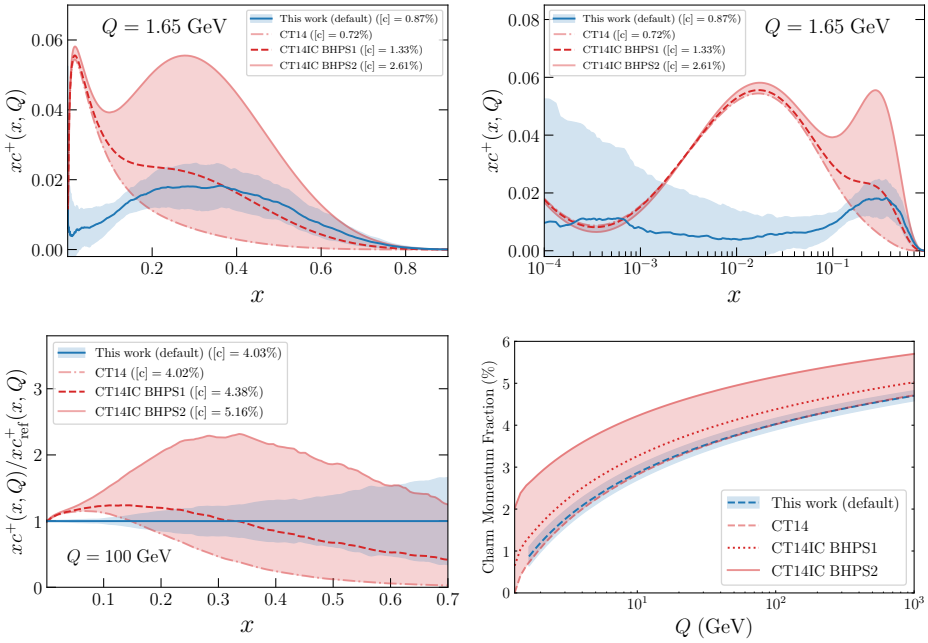

Figure 5.17: The 4 FNS charm PDF from Hou et al. 2018 compared to our result (also in the 4 FNS) at Q = 1.65 GeV on a linear (top left) and logarithmic (top right) scale in x, and at Q = 100 GeV on a linear scale in x and as a ratio to our result (bottom left). The momentum fraction corresponding to either case is also shown as a function of Q (bottom right). Note that for our result the uncertainty band is the 68%CL PDF uncertainty, while for Hou et al. 2018 the central curve (labeled CT14IC BHPS1) corresponds to the BHPS model with best-fit normalization, the lower curve (labeled CT14) corresponds to the default CT14 perturbative charm PDF and the upper curve (labeled CT14IC BHPS2) corresponds to the BHPS model with normalization at the upper 90% CL (see text). The value of the momentum fractions are also provided in each case.

The CT14IC charm PDF is compared to our result in fig. 5.17, at Q = 1.65 GeV and Q = 100 GeV, in the former case on both a logarithmic and linear scale in x and in the latter case on a linear scale only, as a ratio to our default result. Note that the uncertainty band has a different interpretation in the two curves shown: for our result it is the 68% CL PDF uncertainty, while for Hou et al. 2018 it corresponds to the model uncertainty estimated as described above. In fig. 5.17 we also quote the charm momentum fraction in each case, at the corresponding scale Q.

As shown in fig. 5.1 (right), our result for the charm PDF is in good agreement with the BHPS model at large x. Correspondingly, for $x \gtrsim 0.3$ we find reasonably good agreement between our result and the central curve of Hou et al. 2018, which corresponds to a momentum fraction and thus a normalization of the charm PDF not too different from our result (see table 5.1). Both the upper and lower curve from Hou et al. 2018 instead do not agree with our result within uncertainties: indeed the lower edge corresponds to the absence of intrinsic charm (which we exclude) and the upper edge to a momentum fraction which we exclude at more than the 5σ level (see table 5.1).

For intermediate values $3 \cdot 10^{-3} \lesssim x \lesssim 0.3$ our result disagrees with that of Hou et al. 2018, while at very small x all results agree, the intrinsic charm being compatible with zero. The disagreement at intermediate x is mostly due to the fact that in Hou et al. 2018 charm is assumed to take the BHPS form, which vanishes for $x \lesssim 0.1$, in the 4 FNS at the low scale $Q = 1.3$ GeV. Due to perturbative evolution from $Q = 1.3$ GeV to $Q = 1.65$ GeV the charm PDF then develops the large bump that is seen in fig. 5.17, where we instead find that the 4 FNS charm PDF is quite small. This difference persists at large scales as seen in fig. 5.1 (bottom left).

In terms of momentum fractions, shown in fig. 5.1 (bottom right), as already mentioned our result is compatible with the central value of Hou et al. 2018 within uncertainties; and also with the lower edge of Hou et al. 2018 that corresponds to perturbative charm. The upper edge of the prediction from Hou et al. 2018 is instead ruled out at more than 5σ.

5.8 Z+CHARM PRODUCTION IN THE FORWARD REGION

Here we provide full details on our computation of Z+charm production and on the inclusion of the LHCb data for this process in the determination of the charm PDF shown in fig. 5.2.

COMPUTATIONAL SETTINGS. Theoretical predictions for the Z+charm measurements in the forward region by LHCb Aaij et al. 2021 follow the settings described in Boettcher et al. 2016. Z+jet events at NLO QCD theory are generated for $\sqrt{s} = 13$ TeV using the Zj package of the POWHEG-BOX Alioli et al. 2010. The parton-level events produced by POWHEG are then interfaced to PYTHIA8 Sjostrand et al. 2008 with the Monash 2013 tune Skands et al. 2014 for showering, hadronization, and simulation of the underlying event and multiple parton interactions. Long-lived hadrons, including charmed hadrons, are assumed stable and not decayed.

Selection criteria on these particle-level events are imposed to match the LHCb acceptance Aaij et al. 2021. Z bosons are reconstructed in the dimuon final state by requiring $60 \text{ GeV} \leqslant m_{\mu\mu} \leqslant 120 \text{ GeV}$, and only events where these muons satisfy

$p_T^\mu \geqslant 20$ GeV and $2.0 \leqslant \eta_\mu \leqslant 4.5$ are retained. Stable visible hadrons within the LHCb acceptance of $2.0 \leqslant \eta \leqslant 4.5$ are clustered with the anti-k_T algorithm with radius parameter of $R = 0.5$ Cacciari, Salam, et al. 2008. Only events with a hardest jet satisfying 20 GeV $\leqslant p_T^{\rm jet} \leqslant 100$ GeV and $2.2 \leqslant \eta_{\rm jet} \leqslant 4.2$ are retained. Charm jets are defined as jets containing a charmed hadron, specifically jets satisfying $\Delta R(j, c\text{-hadron}) \leqslant 0.5$ for a charmed hadron with $p_T(c\text{-hadron}) \geqslant 5$ GeV. Jets and muons are required to be separated in rapidity and azimuthal angle, so we require $\Delta R(j, \mu) \geqslant 0.5$. The resulting events are then binned in the Z bosom rapidity $y_Z = y_{\mu\mu}$.

The physical observable measured by LHCb is the ratio of the fraction of Z+jet events with and without a charm tag,

$$\mathcal{R}_j^c \equiv \frac{\sigma(pp \to Z + \text{ charm jet})}{\sigma(pp \to Z + \text{ jet})} = \frac{N(c\text{ -tag})}{N(\text{ jets})}\,. \tag{5.6}$$

Here $N(c\text{ -tag})$ and $N(\text{ jets})$ are, respectively, the number of charm-tagged and un-tagged jets, for a Z boson rapidity interval that satisfies the selection and acceptance criteria. The denominator of eq. (5.6) includes all jets, even those containing heavy hadrons. The charm tagging efficiency is already accounted for at the level of the experimental measurement, so it is not required in the theory simulations.

Predictions for eq. (5.6) are produced using our default PDF determination (`NNPDF4.0` NNLO), as well as the corresponding PDF set with perturbative charm (see section 5.3). We have explicitly checked that our results are essentially independent of the value of the charm mass. We have evaluated MHOUs and PDF uncertainties using the output of the POWHEG+PYTHIA8 calculations. We have checked that MHOUs, evaluated with the standard seven-point prescription, essentially cancel in the ratio eq. (5.6). Note that this is not the case for PDF uncertainties, because the dominant partonic subchannels in the numerator and denominator are not the same.

$\chi^2/N_{\rm dat}$	default charm		perturbative charm	
	$\rho_{\rm sys} = 0$	$\rho_{\rm sys} = 1$	$\rho_{\rm sys} = 0$	$\rho_{\rm sys} = 1$
Prior	1.85	3.33	3.54	3.85
Reweighted	1.81	3.14	–	–

Table 5.2: The values of $\chi^2/N_{\rm dat}$ for the LHCb Z+charm data before (prior) and after (reweighted) their inclusion in the PDF fit. Results are given for two experimental correlation models, denoted as $\rho_{\rm sys} = 0$ and $\rho_{\rm sys} = 1$. We also report values before inclusion for the perturbative charm PDFs.

INCLUSION OF THE LHCB DATA. We first compare the quality of the description of the LHCb data before their inclusion. In table 5.2 we show the values of

χ^2/N_{dat} for the LHCb Z+charm data both with default and perturbative charm. Since the experimental covariance matrix is not available for the LHCb data we determine the χ^2 values assuming two limiting scenarios for the correlation of experimental systematic uncertainties. Namely, we either add in quadrature statistical and systematic errors ($\rho_{sys} = 0$), or alternatively we assume that the total systematic uncertainty is fully correlated between y_Z bins ($\rho_{sys} = 1$). Fit quality is always significantly better in our default intrinsic charm scenario than with perturbative charm. As is clear from fig. 5.2 (top left), the somewhat poor fit quality is mostly due to the first rapidity bin, which is essentially uncorrelated to the amount of intrinsic charm (see fig. 5.2, top right).

The LHCb Z+charm data are then included in the PDF determination through Bayesian reweighting Ball et al. 2011; Ball, Bertone, Cerutti, et al. 2012. The χ^2/N_{dat} values obtained using the PDFs found after their inclusion are given in table 5.2. They are computed by combining the PDF and experimental covariance matrix so both sources of uncertainty are included — as mentioned above, MHOUs are negligible. The fit quality is seen to improve only mildly, and the effective number of replicas Ball et al. 2011; Ball, Bertone, Cerutti, et al. 2012 after reweighting is only moderately reduced, from the prior $N_{rep} = 100$ to $N_{eff} = 92$ or $N_{eff} = 84$ in the $\rho_{sys} = 0$ and $\rho_{sys} = 1$ scenarios respectively. This demonstrates that the inclusion of the LHCb Z+charm measurements affects the PDFs only weakly. This agrees with the results shown in fig. 5.2 (center) in section 5.1, where it is seen that the inclusion of the LHCb data has essentially no impact on the shape of the charm PDF, but it moderately reduces its uncertainty in the region of the valence peak.

5.9 PARTON LUMINOSITIES

The impact of intrinsic charm on hadron collider observables can be assessed by studying parton luminosities. Indeed, the cross-section for hadronic processes at leading order is typically proportional to an individual parton luminosity or linear combination of parton luminosities. Comparing parton luminosities determined using our default PDF set to those obtained imposing perturbative charm (see section 5.3) provides a qualitative estimate of the measurable impact of intrinsic charm. Of course this is then modified by higher-order perturbative corrections, which generally depend on more partonic subchannels and thus on more luminosities. In this section we illustrate this by considering the parton luminosities that are relevant for the computation of the Z+charm process in the LHCb kinematics, see section 5.8.

The parton luminosity without any restriction on the rapidity y_X of the final state is

$$\mathcal{L}_{ab}(m_X) = \frac{1}{s}\int_\tau^1 \frac{dx}{x} f_a\left(x, m_X^2\right) f_b\left(\tau/x, m_X^2\right), \qquad \tau = \frac{m_X^2}{s}, \tag{5.7}$$

where a, b label the species of incoming partons, $\sqrt{s}$ is the center-of-mass energy of the hadronic collision, and m_X is the final state invariant mass. For the more realistic situation where the final state rapidity is restricted, $y_{\,\mathrm{min}} \leqslant y_X \leqslant y_{\,\mathrm{max}}$, eq. (5.7) is modified as

$$\mathcal{L}_{ab}(m_X) = \frac{1}{s}\int_\tau^1 \frac{dx}{x} f_a\left(x, m_X^2\right) f_b\left(\tau/x, m_X^2\right) \theta\left(y_X - y_{\,\mathrm{min}}\right) \theta\left(y_{\,\mathrm{max}} - y_X\right) , \tag{5.8}$$

where $y_X = \left(\ln x^2/\tau\right)/2$.

We consider in particular the quark-gluon and the charm-gluon luminosities, defined as

$$\begin{aligned} \mathcal{L}_{qg}(m_X) &\equiv \sum_{i=1}^{n_f} \left(\mathcal{L}_{q_i g}(m_X) + \mathcal{L}_{\bar{q}_i g}(m_X)\right) \\ \mathcal{L}_{cg}(m_X) &\equiv \left(\mathcal{L}_{cg}(m_X) + \mathcal{L}_{\bar{c}g}(m_X)\right) , \end{aligned} \tag{5.9}$$

where n_f is the number of active quark flavors for a given value of $Q = m_X$ with a maximum value of $n_f = 5$. These are the combinations that provide the leading contributions respectively to the numerator ($\mathcal{L}_{cg}$) and the denominator ($\mathcal{L}_{qg}$) of $\mathcal{R}_j^c$ in eq. (5.6).

The luminosities are displayed in fig. 5.18, in the invariant mass region, 40 GeV $\leqslant m_X \leqslant$ 200 GeV which is most relevant for Z+charm production. Results are shown for three different rapidity bins, $-2.5 \leqslant y_X \leqslant 2.5$ (central production in ATLAS and CMS), $2.0 \leqslant y_X \leqslant 2.75$ (forward production, corresponding to the central bin in LHCb), and $3.5 \leqslant y_X \leqslant 4.5$ (highly boosted production, corresponding to the most forward bin in the LHCb selection), as a ratio to our default case.

For central production it is clear that both the quark-gluon and charm-gluon luminosities with our without intrinsic charm are very similar. This means that central Z+charm production in this invariant mass range is insensitive to intrinsic charm. For forward production, corresponding to the central LHCb rapidity bin, $2.0 \leqslant y_X \leqslant 2.75$, in the invariant mass region $m_X \simeq 100$ GeV again there is little difference between results with or without intrinsic charm, but as the invariant mass increases the charm-gluon luminosity with intrinsic charm is significantly enhanced. For very forward production, such as the highest rapidity bin of LHCb, $3.5 \leqslant y_X \leqslant 4.5$, the charm-gluon luminosity at $m_X \simeq 100$ GeV is enhanced by a factor of about 4 in our default result in comparison to the perturbative charm case, corresponding to a $\simeq 3\sigma$ difference in units of the PDF uncertainty, consistently with the behavior observed for the $\mathcal{R}_j^c$ observable in fig. 5.2 (top left) in the most forward rapidity bin. This observation provides a qualitative explanation of the results of section 5.8.

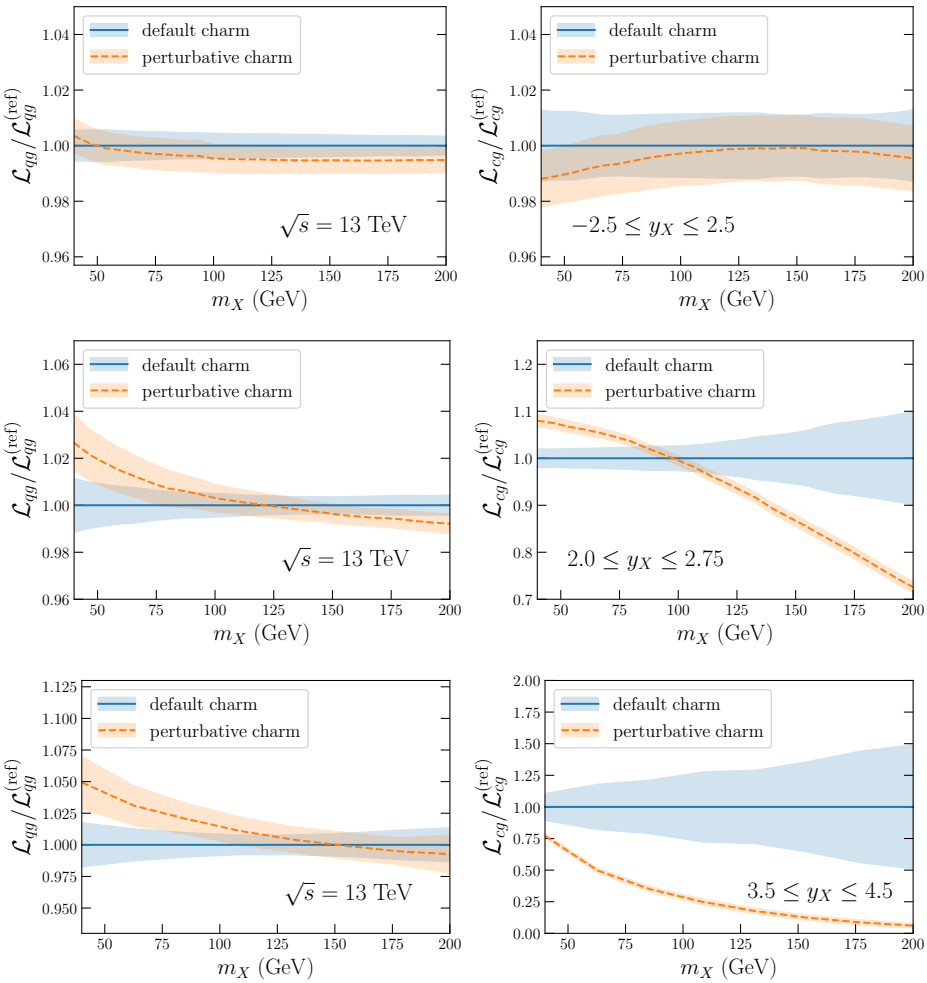

Figure 5.18: The quark-gluon (left) and charm-gluon (right) parton luminosities in the m_X region relevant for Z+charm production and three different rapidity bins (see text). Results are shown both for our default charm PDFs and for the variant with perturbative charm.

5.10 SUMMARY

In this work, long-sought evidence for intrinsic charm quarks in the proton has been presented. These findings close a fundamental open question in the understanding of nucleon structure that has been hotly debated by particle and nuclear physicists for the last 40 years. By carefully disentangling the perturbative component, unambiguous evidence for intrinsic charm is obtained, which turns out to be in qualitative agreement with the expectations from model calculations. The determination of the charm PDF, driven by indirect constraints from the latest high-precision LHC data, is perfectly consistent with direct constraints both from EMC charm production data taken forty years ago, and with very recent Z+charm production data in the forward region from LHCb. Combining all data, local significance for intrinsic charm is found in the large-x region just above the 3σ level. The results motivate further dedicated studies of intrinsic charm through a wide range of nuclear, particle and astro-particle physics experiments, from the high-luminosity LHC Azzi et al. 2019 and the fixed-target programs of LHCb Aaij et al. 2019 and ALICE Dainese et al. 2019, to the Electron-Ion Collider, AFTER Hadjidakis et al. 2021, the Forward Physics Facility Anchordoqui et al. 2021, and neutrino telescopes Halzen and Wille 2016.

Both the `LHAPDF` grids produced in this work and the version of `EKO` with the respective run cards used are made available from `http://nnpdf.mi.infn.it/nnpdf4-0-charm-study/`.

6 FORWARD-BACKWARD ASYMMETRY

An important direction for ongoing and future studies of new physics Beyond the Standard Model (BSM) at the Large Hadron Collider (LHC) is the search for novel heavy resonances. The LHC is uniquely suited to direct searches for these resonances, thanks to its unparalleled center of mass energy, $\sqrt{s} = 13.6\,\text{TeV}$ in the recently started Run III, and the high statistics to be accumulated in the coming years, especially in the high-luminosity (HL) phase. For instance, considering representative benchmark BSM scenarios (cf. Cid Vidal et al. 2019), the HL- LHC is sensitive to searches for sequential Standard Model (SM) W' gauge bosons up to $m_{W'} = 7.8\,\text{TeV}$, E_6 model Z' gauge bosons up to $m_{Z'} = 5.7\,\text{TeV}$, and Kaluza-Klein resonances decaying into a $t\bar{t}$ pair up to $m_{KK} = 6.6\,\text{TeV}$.

The production of such high-mass states proceeds via partonic scattering that involves large values of the momentum fractions x_1 and x_2 of the colliding partons, because the center of mass energy of the partonic collision is $\hat{s} = x_1 x_2 s$. For instance, the on-shell production of a state with invariant mass $m_X = 8\,\text{TeV}$ requires $x_1 x_2 \gtrsim 0.3$, hence for central production at leading order $x_1 = x_2 \approx 0.6$. This is problematic because PDFs are poorly known for $x \gtrsim 0.4$ (cf. Gao et al. 2018; Kovařık et al. 2020), as there is limited data included in current PDF determinations to constrain this kinematic region. Indeed, in the past, claims of possible BSM signals, Abe et al. 1996, were subsequently traced to poor modeling of the PDFs in the large-x region, Lai et al. 1997. The impact of lack of knowledge of the PDFs on BSM searches is thus a delicate issue, Beenakker et al. 2016.

Here we wish to further investigate this by specifically considering Neutral Current (NC) Drell–Yan (DY) dilepton production and associated observables,

frequently used for BSM searches at the LHC. NC Drell–Yan production is one of the cleanest processes in the search for both narrow and broad heavy resonances decaying into dileptons, $pp \to X \to \ell^+\ell^-$, since the two charged leptons can be detected with excellent energy and angular resolution. This also enables the search for smooth, non-resonant distortions with respect to the SM backgrounds, such as those arising in the context of contact interactions or, more generally, induced by Effective Field Theory (EFT) higher-dimensional operators that lead to direct couplings between quarks and leptons, Dawson et al. 2019; J. Ellis et al. 2021; Ethier, Magni, et al. 2021; Greljo et al. 2021. Indeed, both ATLAS and CMS have extensively explored this channel in their BSM search program, Aad et al. 2014, 2019, 2020, 2021; Albert M Sirunyan et al. 2019, 2021. To this purpose, it is mandatory to have a detailed understanding of the dominant SM background, namely dilepton production from quark-antiquark annihilation mediated by a virtual Electroweak (EW) boson, $q\bar{q} \to \gamma^*/Z \to \ell^+\ell^-$, with subleading processes involving the quark-gluon and photon-photon initial states.

Drell–Yan production is one of the SM processes which is known to highest perturbative accuracy: indeed, both N^3LO QCD results, Duhr and Mistlberger 2022, and the full mixed QCD- EW corrections at NNLO, Armadillo et al. 2022; Bonciani, Buccioni, et al. 2020; Bonciani, Buonocore, et al. 2022; Buccioni, Caola, Chawdhry, et al. 2022; Buccioni, Caola, Delto, et al. 2020, have become available recently. Therefore, the main uncertainty on theoretical predictions for this process is mostly due to the PDFs, which, as mentioned, are poorly known at large x. Experimentally, uncertainties are minimized when considering observables in which several systematics cancel in part or entirely. An example relevant for the DY process is the forward-backward asymmetry A_{fb} of the angular distribution of the dilepton pair in the center-of-mass frame of the partonic collision, i.e. the asymmetry in the so-called Collins–Soper angle θ^*, recently measured from the Run II dataset by ATLAS, Aaboud et al. 2017, and CMS, Tumasyan et al. 2022. The sensitivity of this observable to both PDFs and BSM signals has been emphasized recently, Accomando et al. 2019, 2018; Fiaschi et al. 2021, 2022, as well as its relevance to extractions of the weak mixing angle $\sin^2\theta_W$ at the LHC, Albert M. Sirunyan et al. 2018. These studies are mostly restricted to the vicinity of the Z-boson peak, $m_{\ell\bar{\ell}} \sim m_Z$ with $m_{\ell\bar{\ell}}$ being the dilepton mass, though in a recent study by CMS, Tumasyan et al. 2022, the forward-backward asymmetry has been used to obtain a lower mass limit (of 4.4 TeV) on a hypothetical Z' heavy gauge boson.

In this work, we assess to which extent different assumptions on the large-x behavior of PDFs, as well as different estimates of the PDF uncertainty in this region, may affect BSM searches, by specifically studying NC Drell–Yan production, and the forward-backward asymmetry in particular. To this purpose, we explain the dependence of the general qualitative features of the asymmetry on the behavior of PDFs, based on an understanding of the analytic dependence of the asymmetry on the partonic luminosities. We then present detailed computations of the forward-backward asymmetry at the LHC, with realistic experimental cuts, using a variety of PDF sets.

We find that first, the large-x PDF shape and uncertainty can differ considerably between different PDF sets, with NNPDF4.0, Ball et al. 2021b, generally displaying a more flexible shape and a wider uncertainty. And second, that all PDF sets except NNPDF4.0 lead to a qualitative behavior of the asymmetry which in the large-mass multi-TeV region reproduces the shape found around the Z-peak region, even though there is no fundamental reason why this should be the case. We will then trace the observed behavior of the asymmetry to that of the underlying PDFs.

6.1 ANATOMY OF DRELL–YAN PRODUCTION

The aim of this section is to scrutinize the PDF dependence of the NC Drell–Yan differential cross-section and of the associated forward-backward asymmetry by reviewing the LO kinematics, determining LO analytic expressions, and finally comparing these analytical calculations to the results of LO and NLO numerical simulations obtained using MadGraph5_aMC@NLO, Alwall et al. 2014, interfaced to PineAPPL, S. Carrazza et al. 2020a; Schwan et al. 2022b. Specifically, we will relate the behavior of the differential distribution and asymmetry to the relevant parton luminosities.

6.1.1 Drell–Yan kinematics and cross-sections at LO

We consider dilepton production via the exchange of an Electroweak neutral gauge boson Z/γ^* in proton-proton collisions:

$$p(k_1) + p(k_2) \to Z/\gamma^*(q) \to \ell(p_\ell) + \bar{\ell}(p_{\bar{\ell}}) + X. \tag{6.1}$$

The hadronic differential cross-section $d\sigma^{pp\to\ell\bar{\ell}}$ is factorized in terms of PDFs f_i and the partonic cross sections $d\hat{\sigma}_{ij}$ for incoming partons of species i, j as

$$d\sigma^{pp\to\ell\bar{\ell}} = \sum_{ij} \int_0^1 dx_1 dx_2 f_i(x_1, \mu_F^2) f_j(x_2, \mu_F^2) d\hat{\sigma}_{ij}(\hat{k}_1 = x_1 k_1, \hat{k}_2 = x_2 k_2). \tag{6.2}$$

In the sequel we will set the factorization scale μ_F to the invariant mass of the gauge boson, i.e. the dilepton invariant mass, so $\mu_F^2 = m_{\ell\bar{\ell}}^2 = (p_\ell + p_{\bar{\ell}})^2$. The kinematics and Feynman diagram of the LO partonic process in the quark-antiquark channel are shown in fig. 6.1. We do not consider photon-initiated processes, as they do not affect the qualitative features of our discussion.

At LO, the momentum fractions of the two incoming partons are fully fixed by knowledge of the invariant mass and rapidity of the gauge boson, i.e. of the dilepton pair $y_{\ell\bar{\ell}} = (y_\ell + y_{\bar{\ell}})/2$:

$$x_1 = \frac{m_{\ell\bar{\ell}}}{\sqrt{s}} \exp(y_{\ell\bar{\ell}}), \quad x_2 = \frac{m_{\ell\bar{\ell}}}{\sqrt{s}} \exp(-y_{\ell\bar{\ell}}), \tag{6.3}$$

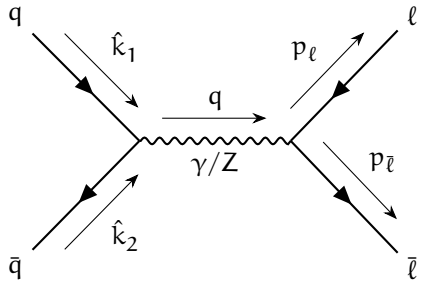

Figure 6.1: Neutral-current Drell–Yan production at LO in the quark-antiquark channel.

where the center of mass energy of the hadronic collision is $s = (k_1 + k_2)^2$ and at LO $m^2_{\ell\bar{\ell}} = \hat{s} = x_1 x_2 s$. The absolute dilepton rapidity thus lies in the range $|y_{\ell\bar{\ell}}| \leqslant \ln(\sqrt{s}/m_{\ell\bar{\ell}})$. Beyond LO there might be extra radiation in the final state, so the LO kinematics provides a lower bound on the momentum fractions of the incoming partons, and all values of the momentum fractions such that $x_{1,2} \geqslant m_{\ell\bar{\ell}}/\sqrt{s}$ are allowed.

It is useful to define the so-called Collins–Soper angle θ^*, J. C. Collins and Soper 1977, which in the hadronic Center of Mass (CoM) frame is defined as

$$\begin{aligned} \cos\theta^* &= \text{sign}(y_{\ell\bar{\ell}})\cos\theta\,, \\ \cos\theta &\equiv \frac{p^+_\ell p^-_{\bar{\ell}} - p^-_\ell p^+_{\bar{\ell}}}{m_{\ell\bar{\ell}}\sqrt{m^2_{\ell\bar{\ell}} + p^2_{T,\ell\bar{\ell}}}}, \quad p^\pm = p^0 \pm p^3. \end{aligned} \tag{6.4}$$

It is easy to show that the Collins–Soper angle θ^* coincides with the scattering angle of the lepton in the partonic CoM frame, $\bar{\theta}$. The latter is defined in terms of the lepton momentum as

$$\cos\bar{\theta} \equiv \frac{p^z_\ell}{m_{\ell\bar{\ell}}}\,, \tag{6.5}$$

where the z axis is along the direction of the incoming quark-antiquark pair. In the partonic CoM frame, of course, $p^z_\ell = -p^z_{\bar{\ell}}$ and $y_{\ell\bar{\ell}} = 0$, so

$$p^\pm_\ell = p^\mp_{\bar{\ell}} = m_{\ell\bar{\ell}}\left(1 \pm \cos\bar{\theta}\right)\,, \tag{6.6}$$

and substituting in eq. (6.4) it immediately follows that, taking the convention $\text{sign}(y_{\ell\bar{\ell}}) = \text{sign}(0) = +1$, $\cos\theta^* = \cos\theta = \cos\bar{\theta}$. The expression of $\cos\theta$ in eq. (6.4) is manifestly invariant upon boosts along the z axis, so the identification of θ with the CoM scattering angle $\bar{\theta}$ remains true in any reference frame.

Note that the definition eq. (6.5) requires a choice for the positive direction of the z axis, which is usually taken along the direction of the incoming fermion (quark). This direction is not experimentally accessible in proton-proton collisions, so the Collins–Soper angle is defined by always taking the positive z axis in the direction of the boosted dilepton pair, i.e., at LO, along the direction of the incoming quark with largest momentum fraction, i.e. by supplementing in

the definition a factor $\mathrm{sign}(y_{\ell\bar{\ell}})$. Hence $\cos\theta^* = \cos\bar{\theta}$ ($\cos\theta^* = -\cos\bar{\theta}$) if the momentum fraction of the incoming quark (antiquark) is the largest.

The hard scattering matrix elements that enter the partonic cross-section in eq. (6.2) are the sum of a pure photon-exchange contribution, a photon-Z interference term, and a pure Z-exchange contribution. Of course, in the region $m_{\ell\bar{\ell}} \gtrsim m_Z$ these contributions are all of the same order. Standard arguments, Peskin and Schroeder 1995, then imply that, because in the Standard Model the photon coupling to leptons is vector while the Z coupling is chiral, the pure photon and pure Z contributions to the cross-section are necessarily even in $\cos\theta^*$ while the interference term is odd.

Specifically, at LO the fully differential hadronic cross-section can be obtained from the well-known result, Peskin and Schroeder 1995, for $e^+e^- \to \mu^+\mu^-$ by replacing the incoming lepton charges with those of the quarks, and accounting for the PDFs, with the result

$$\begin{aligned}\frac{d^3\sigma}{dm_{\ell\bar{\ell}}\, dy_{\ell\bar{\ell}}\, d\cos\theta^*} = \frac{\pi\alpha^2}{3m_{\ell\bar{\ell}}s}\Bigg(&\\ (1+\cos^2(\theta^*))\sum_q S_q\left[f_q(x_1,m^2_{\ell\bar{\ell}})f_{\bar{q}}(x_2,m^2_{\ell\bar{\ell}}) + f_q(x_2,m^2_{\ell\bar{\ell}})f_{\bar{q}}(x_1,m^2_{\ell\bar{\ell}})\right]&\\ +\cos\theta^*\sum_q A_q\,\mathrm{sign}(y_{\ell\bar{\ell}})\left[f_q(x_1,m^2_{\ell\bar{\ell}})f_{\bar{q}}(x_2,m^2_{\ell\bar{\ell}}) - f_q(x_2,m^2_{\ell\bar{\ell}})f_{\bar{q}}(x_1,m^2_{\ell\bar{\ell}})\right]&\\ \Bigg)\end{aligned} \tag{6.7}$$

where α is the QED coupling and the even (symmetric) and odd (antisymmetric) couplings are given by

$$\begin{aligned} S_q &= e_l^2 e_q^2 + P_{\gamma Z}\cdot e_l v_l e_q v_q + P_{ZZ}\cdot(v_l^2+a_l^2)(v_q^2+a_q^2)\\ A_q &= P_{\gamma Z}\cdot 2e_l a_l e_q a_q + P_{ZZ}\cdot 8 v_l a_l v_q a_q\,, \end{aligned} \tag{6.8}$$

in terms of the electric charges e_l, e_q and the vector and axial couplings v_l, v_q and a_l, a_q of the leptons and quarks, and the propagator factors

$$P_{\gamma Z}(m_{\ell\bar{\ell}}) = \frac{2m^2_{\ell\bar{\ell}}(m^2_{\ell\bar{\ell}} - m_Z^2)}{\sin^2(\theta_W)\cos^2(\theta_W)\left[(m^2_{\ell\bar{\ell}}-m_Z^2)^2 + \Gamma_Z^2 m_Z^2\right]} \tag{6.9}$$

$$P_{ZZ}(m_{\ell\bar{\ell}}) = \frac{m^4_{\ell\bar{\ell}}}{\sin^4(\theta_W)\cos^4(\theta_W)\left[(m^2_{\ell\bar{\ell}}-m_Z^2)^2 + \Gamma_Z^2 m_Z^2\right]}, \tag{6.10}$$

with m_Z and Γ_Z respectively the Z mass and width and θ_W the weak mixing angle. In fig. 6.2 we display the symmetric S_q (left) and antisymmetric A_q (right) couplings, eq. (6.8), for up-like and down-like quarks, as a function of the dilepton invariant mass $m_{\ell\bar{\ell}}$. Both couplings are around a factor 2 larger for up-like quarks than for down-like quarks, and become $m_{\ell\bar{\ell}}$-independent for $m_{\ell\bar{\ell}} \gtrsim 1$ TeV, where

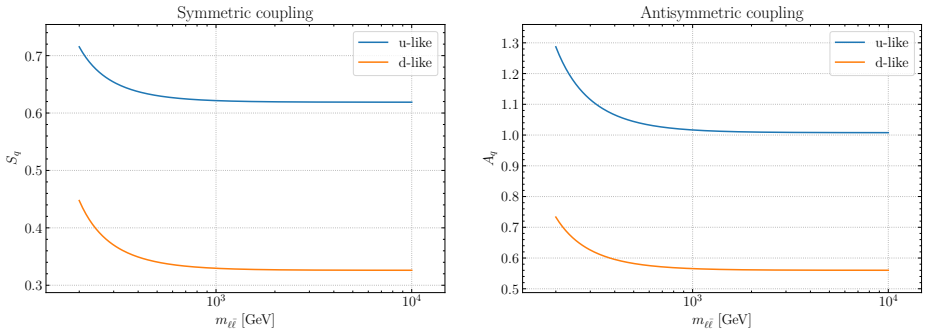

Figure 6.2: The symmetric S_q (left) and antisymmetric A_q (right) couplings, eq. (6.8), for up-like and down-like quarks, as a function of the dilepton invariant mass $m_{\ell\bar{\ell}}$.

they take the asymptotic values $\bar{S}_q$, $\bar{A}_q$ obtained by substituting in eq. (6.8) the large-mass expressions of the propagator factors

$$\bar{P}_{\gamma Z} = \frac{2}{\sin^2(\theta_W)\cos^2(\theta_W)}, \qquad \bar{P}_{ZZ} = \frac{1}{\sin^4(\theta_W)\cos^4(\theta_W)}, \tag{6.11}$$

to which $P_{\gamma Z}$ and P_{ZZ} respectively reduce up to $O(m_Z^2/m_{\ell\bar{\ell}}^2)$ corrections.

The interference term proportional to A_q is odd in the Collins–Soper angle $\cos\theta^*$, leading to a forward-backward scattering asymmetry. In a proton-proton collision the initial state is completely symmetric, so the quark and antiquark contributions to the cross-section eq. (6.7) are necessarily symmetric upon the interchange of the incoming quark and antiquark, with the corresponding momentum fractions fixed at LO by eq. (6.3). However, as mentioned, there is a sign change in the relation between $\cos\theta^*$ and $\cos\theta$ according to whether the incoming parton with largest momentum fraction is a quark or an antiquark, i.e., when interchanging x_1 with x_2 in the argument of the quark and antiquark PDFs, thereby leading to the result of eq. (6.7). This leads to a forward-backward asymmetry whenever the quark and antiquark PDFs have different x dependence.

In order to understand the relation of this forward-backward asymmetry in terms of the behavior of the PDFs, it is convenient to rewrite the PDF combinations that contribute to the differential cross-section eq. (6.7) in terms of symmetric and antisymmetric parton luminosities, defined as

$$\begin{aligned}
\mathcal{L}_{S,q}(m_{\ell\bar{\ell}}, y_{\ell\bar{\ell}}) &\equiv f_q(x_1, m_{\ell\bar{\ell}}^2) f_{\bar{q}}(x_2, m_{\ell\bar{\ell}}^2) + f_q(x_2, m_{\ell\bar{\ell}}^2) f_{\bar{q}}(x_1, m_{\ell\bar{\ell}}^2)\,, \\
\mathcal{L}_{A,q}(m_{\ell\bar{\ell}}, y_{\ell\bar{\ell}}) &\equiv \mathrm{sign}(y_{\ell\bar{\ell}}) \left[f_q(x_1, m_{\ell\bar{\ell}}^2) f_{\bar{q}}(x_2, m_{\ell\bar{\ell}}^2) - f_q(x_2, m_{\ell\bar{\ell}}^2) f_{\bar{q}}(x_1, m_{\ell\bar{\ell}}^2) \right],
\end{aligned} \tag{6.12}$$

where the momentum fractions x_1 and x_2 are given in terms of $m_{\ell\bar{\ell}}$, $y_{\ell\bar{\ell}}$, and $\sqrt{s}$ in eq. (6.3). Note that both parton luminosities are invariant under the inter-

change $x_1 \leftrightarrow x_2$, upon which $y_{\ell\bar{\ell}} \to -y_{\ell\bar{\ell}}$. In terms of these luminosities, the triple differential cross-section eq. (6.7) takes the compact form

$$\frac{d^3\sigma}{dm_{\ell\bar{\ell}}\, dy_{\ell\bar{\ell}}\, d\cos\theta^*} = \frac{\pi\alpha^2}{3m_{\ell\bar{\ell}}s}\left((1+\cos^2(\theta^*))\sum_q S_q \mathcal{L}_{S,q}(m_{\ell\bar{\ell}}, y_{\ell\bar{\ell}}) + \cos\theta^* \sum_q A_q \mathcal{L}_{A,q}(m_{\ell\bar{\ell}}, y_{\ell\bar{\ell}})\right) \tag{6.13}$$

which explicitly displays its symmetry properties upon the transformation $\cos\theta^* \to -\cos\theta^*$, equivalent to a charge conjugation transformation $q \leftrightarrow \bar{q}$ and $\ell \leftrightarrow \bar{\ell}$.

The symmetric and antisymmetric parton luminosities eq. (6.12) can also be expressed in terms of the sum and difference of quark and antiquark PDFs,

$$f_q^{\pm}(x, Q) = f_q(x, Q) \pm f_{\bar{q}}(x, Q)\,, \tag{6.14}$$

where f_q^- is usually called the valence PDF combination, and f_q^+ the total quark PDF . Note that at LO, and more generally in factorization schemes in which PDFs are positive, such as $\overline{\text{MS}}$ (cf. Candido, Forte, et al. 2020), f_q^+ is positive while f_q^- in general is not, and $f_q^+ > |f_q^-|$. We can write the symmetric and antisymmetric parton luminosities in eq. (6.12) as

$$\mathcal{L}_{S,q}(m_{\ell\bar{\ell}}, y_{\ell\bar{\ell}}) = \frac{1}{2}\left(f_q^+(x_1, m_{\ell\bar{\ell}}^2) f_q^+(x_2, m_{\ell\bar{\ell}}^2) - f_q^-(x_2, m_{\ell\bar{\ell}}^2) f_q^-(x_1, m_{\ell\bar{\ell}}^2)\right) \tag{6.15}$$

$$\mathcal{L}_{A,q}(m_{\ell\bar{\ell}}, y_{\ell\bar{\ell}}) = \frac{\text{sign}(y_{\ell\bar{\ell}})}{2}\left(f_q^-(x_1, m_{\ell\bar{\ell}}^2) f_q^+(x_2, m_{\ell\bar{\ell}}^2) - f_q^-(x_2, m_{\ell\bar{\ell}}^2) f_q^+(x_1, m_{\ell\bar{\ell}}^2)\right). \tag{6.16}$$

The symmetric luminosity $\mathcal{L}_{S,q}$ is of course positive, and it is dominated by the $f_q^+(x_1, m_{\ell\bar{\ell}}^2) f_q^+(x_2, m_{\ell\bar{\ell}}^2)$ term, which is always larger than the valence contribution $f_q^-(x_2, m_{\ell\bar{\ell}}^2) f_q^-(x_1, m_{\ell\bar{\ell}}^2)$. The sign of the antisymmetric combination, that in turn drives the sign of the forward-backward asymmetry, is in general not determined uniquely. If x_1 is in the region of the valence peak, and x_2 in the small x region, then $f^-(x_1, m_{\ell\bar{\ell}}^2) \gg f^-(x_2, m_{\ell\bar{\ell}}^2)$, and the antisymmetric luminosity is positive provided only that the valence PDF is positive. As we will discuss in section 6.2, while this is indeed the case in the Z-peak region, it is actually not necessarily the case in the high dilepton mass region relevant for BSM searches.

6.1.2 Single-differential distributions and the forward-backward asymmetry

Starting from the triple differential cross section, eq. (6.13), one can define single differential distributions by integrating the other two kinematic variables over the available phase space. In particular, the single-differential distribution in the Collins–Soper angle θ^* is given by

$$\frac{d\sigma}{d\cos\theta^*} = \int_{m_{\ell\bar{\ell}}^{\min}}^{\sqrt{s}} dm_{\ell\bar{\ell}} \int_{\ln(m_{\ell\bar{\ell}}/\sqrt{s})}^{\ln(\sqrt{s}/m_{\ell\bar{\ell}})} dy_{\ell\bar{\ell}}\, \frac{d^3\sigma}{dm_{\ell\bar{\ell}}\, dy_{\ell\bar{\ell}}\, d\cos\theta^*}\,, \tag{6.17}$$

where $m_{\ell\bar{\ell}}^{\min}$ is a lower kinematic cut in the dilepton invariant mass. Since eq. (6.13) falls off steeply with $m_{\ell\bar{\ell}}$, the region with $m_{\ell\bar{\ell}} \gtrsim m_{\ell\bar{\ell}}^{\min}$ will dominate the integral. Given that the dependence of the fully differential cross-section eq. (6.13) on the Collins–Soper angle factorizes with respect to the PDF dependence, the integration over rapidity and invariant mass does not affect the $\cos\theta^*$ dependence, and the single-differential cross section eq. (6.17) takes the simple form

$$\frac{d\sigma}{d\cos\theta^*} = (1+\cos^2\theta^*)\sum_q g_{S,q} + \cos\theta^* \sum_q g_{A,q}\,, \tag{6.18}$$

where the symmetric and antisymmetric coefficients $g_{S,q}$ and $g_{A,q}$ depend on the quark flavor and on the invariant mass cut $m_{\ell\bar{\ell}}^{\min}$, but not on the Collins–Soper angle itself. The contributions relevant for the forward-backward asymmetry, $g_{A,q}$, are given at LO by

$$g_{A,q} = \frac{\pi\alpha^2}{3s} \int_{m_{\ell\bar{\ell}}^{\min}}^{\sqrt{s}} \frac{dm_{\ell\bar{\ell}}}{m_{\ell\bar{\ell}}} A_q(m_{\ell\bar{\ell}}) \int_{\ln(m_{\ell\bar{\ell}}/\sqrt{s})}^{\ln(\sqrt{s}/m_{\ell\bar{\ell}})} dy_{\ell\bar{\ell}}\, \mathcal{L}_{A,q}(m_{\ell\bar{\ell}}, y_{\ell\bar{\ell}})\,, \tag{6.19}$$

which in the large-$m_{\ell\bar{\ell}}$ region, expressing the longitudinal momentum integration in terms of x_1 (assuming $x_1 \geqslant x_2$), becomes

$$g_{A,q} = \frac{\pi\alpha^2 \bar{A}_q}{3s} \int_{m_{\ell\bar{\ell}}^{\min}}^{\sqrt{s}} \frac{dm_{\ell\bar{\ell}}}{m_{\ell\bar{\ell}}} \int_{m_{\ell\bar{\ell}}/\sqrt{s}}^{1} \frac{dx_1}{x_1}\, \mathcal{L}_{A,q}(m_{\ell\bar{\ell}}, x_1) + \mathcal{O}\left(\frac{m_Z^2}{m_{\ell\bar{\ell}}^2}\right), \tag{6.20}$$

where the $m_{\ell\bar{\ell}}$-independent effective couplings $\bar{A}_q$ are given substituting in eq. (6.8) the expressions for the asymptotic propagator factors eq. (6.11).

Upon integration over the Collins–Soper angle, the antisymmetric contribution vanishes: so for instance the rapidity distribution

$$\frac{d\sigma}{dy_{\ell\bar{\ell}}} = \int_{m_{\ell\bar{\ell}}^{\min}}^{\sqrt{s}} dm_{\ell\bar{\ell}} \int_{-1}^{1} d\cos\theta^* \frac{d^3\sigma}{dm_{\ell\bar{\ell}}\, dy_{\ell\bar{\ell}}\, d\cos\theta^*}\,, \tag{6.21}$$

does not depend on terms proportional to A_q. Hence, for BSM searches in which one is interested in the interference terms, as well as for PDF studies in which one is interested in the valence-sea separation, the forward-backward asymmetry is especially relevant. This observable is defined at the differential level as

$$A_{fb}(\cos\theta^*) \equiv \frac{\frac{d\sigma}{d\cos\theta^*}(\cos\theta^*) - \frac{d\sigma}{d\cos\theta^*}(-\cos\theta^*)}{\frac{d\sigma}{d\cos\theta^*}(\cos\theta^*) + \frac{d\sigma}{d\cos\theta^*}(-\cos\theta^*)}\,, \quad \cos\theta^* > 0\,, \tag{6.22}$$

which in terms of the coefficients introduced in eq. (6.18) is given at LO by

$$A_{fb}(\cos\theta^*) = \frac{\cos\theta^*}{(1+\cos^2(\theta^*))} \frac{\sum_q g_{A,q}}{\sum_{q'} g_{S,q'}}\,, \quad \cos\theta^* > 0\,. \tag{6.23}$$

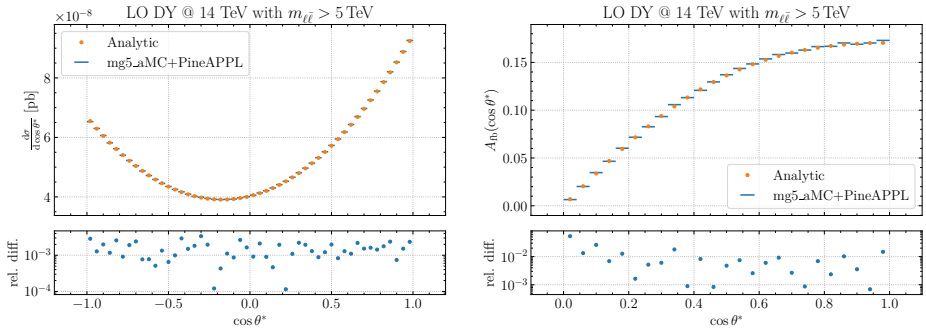

Figure 6.3: The single-inclusive differential distribution in the Collins–Soper angle $\cos\theta^*$, eq. (6.17), and the corresponding forward-backward asymmetry computed at LO, where the analytic calculation eq. (6.22) is compared with the numerical simulation based on `MadGraph5_aMC@NLO` interfaced to PineAPPL. The bottom panels display the relative difference between the analytic and numerical calculations. One of the replicas of the `NNPDF4.0` NNLO PDF set is used as input to the calculation.

This shows that the dependence on $\cos\theta^*$ factorizes and the PDF dependence only appears as an overall normalization factor depending on the ratio of $\sum_q g_{A,q}$ and $\sum_q g_{S,q}$, which in turn depend on the antisymmetric and symmetric partonic luminosities $\mathcal{L}_{A,q}$ and $\mathcal{L}_{S,q}$ respectively. Note that the overall sign of A_{fb} remains in general undetermined.

In order to illustrate concretely these results, in fig. 6.3 we display the single-inclusive differential distribution in $\cos\theta^*$, eq. (6.17), and the corresponding forward-backward asymmetry, eq. (6.22) evaluated at LO for $m_{\ell\bar{\ell}}^{\min} = 5\,\text{TeV}$. The single-differential rapidity distribution eq. (6.21)) is also shown for reference in fig. 6.4. We display both a numerical evaluation based on `MadGraph5_aMC@NLO` interfaced to PineAPPL, as well as analytic results found using the form eq. (6.13) of the triple differential luminosity, with all the values of the parameters entering eqs. (6.8) to (6.10) set to the values used in the `MadGraph5_aMC@NLO` runcard, and performing numerically the integrals in eqs. (6.17) and (6.21). For validation purposes, no kinematic cuts are applied to the rapidities and transverse momenta of final-state leptons. The PDF input is taken to be given, for illustrative purposes, by one of the replicas of the `NNPDF4.0` NNLO set. The relative difference between the analytic and numerical calculation is shown in the bottom panels of fig. 6.3 and demonstrates perfect agreement.

While the discussion so far has been presented at LO, its qualitative features are unaffected by higher-order corrections. To illustrate this, in fig. 6.5 we compare the LO result from fig. 6.3 to the corresponding NLO QCD result. The bottom panels display the NLO K-factor for the $\cos\theta^*$ distribution and the forward-backward asymmetry. Whereas the NLO K-factor in the $\cos\theta^*$ distribution is quite large (around 40%) it exhibits only a mild dependence on the Collins–

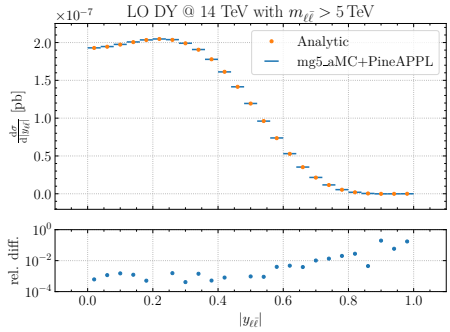

Figure 6.4: Same as fig. 6.3 but now for the absolute dilepton rapidity distribution $|y_{\ell\bar{\ell}}|$

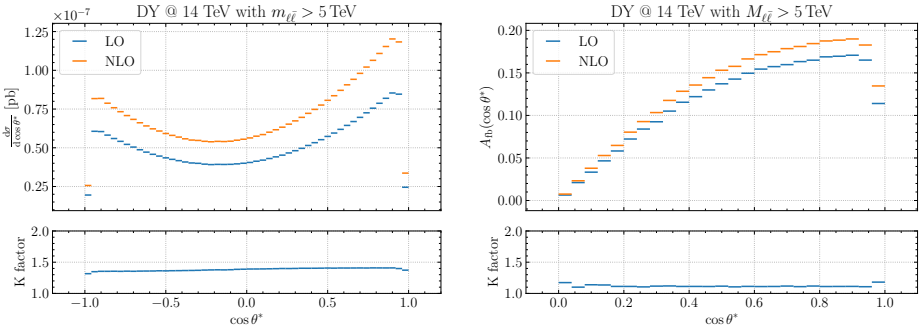

Figure 6.5: Same as fig. 6.3 now comparing the LO result to the NLO QCD result obtained using `MadGraph5_aMC@NLO`. The K-factor is shown in the lower panel.

Soper angle. For A_{fb}, the K-factor is at the 10% level and essentially independent of the value of $\cos\theta^*$.

6.2 THE FORWARD–BACKWARD ASYMMETRY AND THE LARGE-x PDFS

After our general discussion of the Drell–Yan process, we now investigate proton structure at large-x, focusing on its impact on the forward-backward asymmetry $A_{\mathrm{fb}}\left(\cos\theta^*\right)$ at large invariant masses. First, we discuss the dependence of the qualitative features of the asymmetry, and specifically its sign, on the behavior of the underlying PDFs: we illustrate this in a toy model, and compare results to a simple and commonly used approximation. Subsequently, we study the large-x behavior of the PDFs from several recent PDF sets: we compare PDFs, luminosities and the LO asymmetry A_{fb} as a function of the dilepton invariant mass $m_{\ell\bar{\ell}}$.

6.2.1 Qualitative features of A_{fb}

In order to understand the main qualitative features of the $\cos\theta^*$ distribution and of the asymmetry A_{fb} and their dependence on the properties of the underlying PDFs, it is instructive to evaluate predictions based on the same computational setup adopted in section 6.1, namely LO matrix elements without kinematic cuts, using toy PDFs as input. We consider toy quark and antiquark PDF with the form

$$xf_q(x) = A_q x^{-a_q}(1-x)^{b_q}, \quad xf_{\bar{q}}(x) = A_{\bar{q}} x^{-a_{\bar{q}}}(1-x)^{b_{\bar{q}}}, \tag{6.24}$$

where A_q and $A_{\bar{q}}$ are normalization constants, irrelevant for this discussion. For simplicity we neglect the scale dependence of the PDFs. We then compute the single-differential distribution eq. (6.17) and the asymmetry eq. (6.22) with different assumptions on the large x-behavior of these toy PDFs, i.e. different values of the large-x exponents b_q, $b_{\bar{q}}$.

Since the overall normalization does not affect the shape of the distribution, we set $A_q = A_{\bar{q}} = 1$. Furthermore, since we are not interested in the small-x behavior, we set $a_q = a_{\bar{q}} = 1$. Hence, we consider simple scenarios in which

$$xf_q^+(x; b_q, b_{\bar{q}}) = xf_q(x) + xf_{\bar{q}}(x) = x^{-1}\left[(1-x)^{b_q} + (1-x)^{b_{\bar{q}}}\right], \tag{6.25}$$

$$xf_q^-(x; b_q, b_{\bar{q}}) = xf_q(x) - xf_{\bar{q}}(x) = x^{-1}\left[(1-x)^{b_q} - (1-x)^{b_{\bar{q}}}\right], \tag{6.26}$$

with different choices of the parameters b_q and $b_{\bar{q}}$. Specifically, we consider a scenario with $b_q < b_{\bar{q}}$, in particular $(b_q, b_{\bar{q}}) = (3,5)$, which leads to a positive valence combination xf_q^- for all values of x; a scenario with $(b_q, b_{\bar{q}}) = (3,3)$ so xf_q^- vanishes identically; and a third scenario in which the quark PDFs at large-x fall off more rapidly than the antiquarks, $(b_q, b_{\bar{q}}) = (5,3)$, so the valence combination xf_q^- becomes negative.

In fig. 6.6 we display both the $\cos\theta^*$ single-inclusive distribution eq. (6.17) and the asymmetry eq. (6.22). It is apparent that if the antiquark PDFs fall off at large-x faster than the quarks, i.e. when $b_q < b_{\bar{q}}$ the forward-backward asymmetry is positive, while if the converse is true it is negative. Of course if the quark and antiquark PDFs behave in the same way there is no asymmetry. In this simple model, a negative asymmetry corresponds to a negative valence distribution, which conflicts with sum rules and appears to be unphysical. However, the model should be only taken as illustrative of the large-x behavior: it is of course easy to construct PDFs that reproduce this behavior at very large x, while leading to a positive valence PDF as x decreases, consistent with sum rules. One could then argue that Brodsky–Farrar counting rules (cf. Stanley J. Brodsky and Farrar 1973, 1975) imply that $b_q > b_{\bar{q}}$ hence a positive asymmetry is favored. However, counting rules are supposed to only hold asymptotically, so whether they apply in any given region of x is a priori unclear. It is easy to construct generalizations of the model in which the behavior leading to a negative asymmetry is reproduced at large enough x, yet the valence PDFs are positive at smaller x, and the counting rules apply in the strict $x \to 1$ limit.

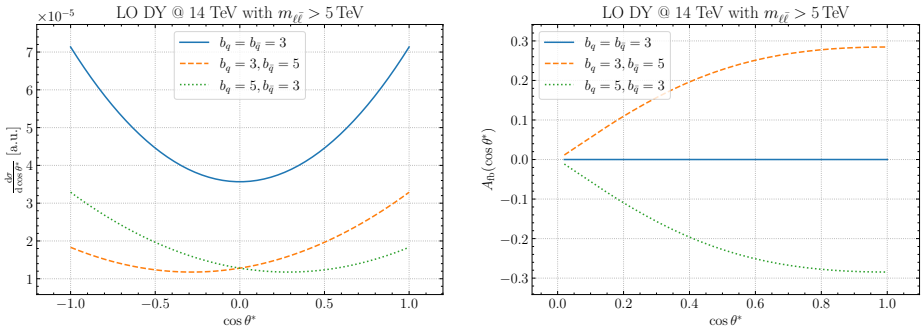

Figure 6.6: The single-inclusive $\cos\theta^*$ distribution eq. (6.17) (left) and the corresponding forward-backward asymmetry (right panel) eq. (6.22) evaluated using the toy PDFs of eq. (6.24). No kinematic cuts are applied except for $m_{\ell\bar{\ell}}^{\min} = 5\,\text{TeV}$.

In fact, whereas in the toy model a negative asymmetry is associated with a negative valence eq. (6.26), the formal condition for a negative asymmetry is (assuming $x_1 > x_2$)

$$\text{sign}\left[\mathcal{L}_{A,q}\right] = \text{sign}\left[\frac{f_q^+(x_2)}{f_q^+(x_1)} - \frac{f_q^-(x_2)}{f_q^-(x_1)}\right] = \text{sign}\left[\frac{f_q(x_2)}{f_q(x_1)} - \frac{f_{\bar{q}}(x_2)}{f_{\bar{q}}(x_1)}\right], \quad x_1 > x_2\,. \tag{6.27}$$

Hence what determines the sign of the antisymmetric luminosity, and thus of the forward-backward asymmetry, is the relative rate of decrease of the quark and antiquark, or valence and total quark PDFs, rather than their sign. Again, it is easy to construct generalizations of the toy model in which the condition eq. (6.27) still holds, yet the valence PDF remains positive.

It is interesting to note that a different conclusion is reached using an approximation to the asymmetry which is quite accurate in the Z peak region. This approximation however turns out to fail at high invariant mass. Indeed, the expression eq. (6.16) of the antisymmetric luminosity in terms of the valence and total PDF combinations f_q^+ and f_q^- PDF combinations suggests an approximation based on the expectation that the valence is dominant at large x and the sea is dominant at small x. Assuming $x_1 > x_2$, one then expects that

$$\mathcal{L}_{A,u}(y_{\ell\bar{\ell}}, m_{\ell\bar{\ell}}) \approx \frac{1}{2} f_u^-(x_1, m_{\ell\bar{\ell}}^2) f_u^+(x_2, m_{\ell\bar{\ell}}^2)\,, \quad x_1 > x_2. \tag{6.28}$$

This is clearly true in the Z-peak region, which motivates the suggestion to use the measurement of A_{fb} as a means to constrain the valence quark combinations, Accomando et al. 2019.

However, while eq. (6.28) provides a satisfactory approximation in the Z-peak region, it fails at larger $m_{\ell\bar{\ell}}$ values. Indeed, for on-shell Z production, with $\sqrt{s} = 14\,\text{TeV}$, for a dilepton rapidity with $y_{\ell\bar{\ell}} \sim 2.5$, the limit of the acceptance region of ATLAS and CMS, the colliding partons have $x_1 = 0.09$ and $x_2 = 6 \times 10^{-4}$. So indeed the contribution in which the valence PDF is evaluated at the smallest x

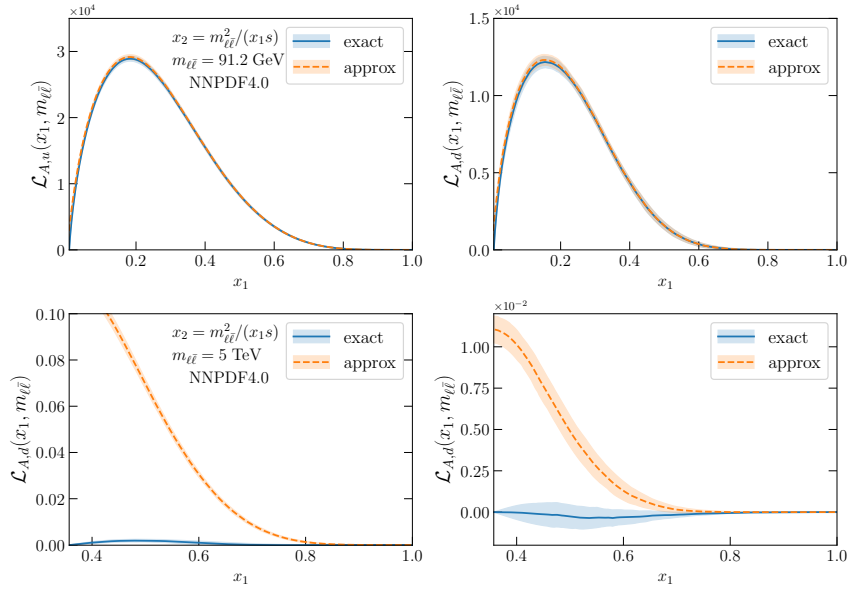

Figure 6.7: The antisymmetric partonic luminosity $\mathcal{L}_{A,q}$, eq. (6.16), for the up and down quarks compared to the approximation eq. (6.28) in the case of NNPDF4.0 at $m_{\ell\bar{\ell}} = m_Z$ (top) and $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ (bottom panels).

value is highly suppressed. But for $m_{\ell\bar{\ell}} = 5\,\text{TeV}$, the smallest value of x_2, attained when $x_1 = 1$, is $x_2 = 0.35$: so both momentum fractions are large and in fact to the right of the valence peak. In such case, there is no obvious hierarchy between the different terms that contribute to to antisymmetric luminosity $\mathcal{L}_{A,q}$.

This is illustrated in fig. 6.7, where we compare the antisymmetric luminosity $\mathcal{L}_{A,q}$ for the up and down quarks to the approximation eq. (6.28), evaluated with NNPDF4.0 NNLO, in the Z-peak region $m_{\ell\bar{\ell}} = m_Z$ and at $m_{\ell\bar{\ell}} = 5\,\text{TeV}$. While indeed for $m_{\ell\bar{\ell}} = m_Z$ eq. (6.28) reproduces the exact luminosity, this is not the case for $m_{\ell\bar{\ell}} \gg m_Z$: both the magnitude and the shape of the luminosity are very different. This qualitative behavior is common to all PDF sets: the approximation fails equally badly regardless of the PDF set.

We conclude that there is no simple relation between the sign of the asymmetry and that of the valence PDF, and that the behavior of the asymmetry must be determined by studying the large-x behavior of the quark and antiquark PDFs.

6.2.2 Parton distributions

We assess now the large-x behavior of the quark and antiquark PDFs in different recent PDF determinations: specifically, we compare ABMP16, CT18, NNPDF4.0, and MSHT20. For completeness, in section 6.4 we also present results obtained with the widely used NNPDF3.1 set, Ball et al. 2017b.

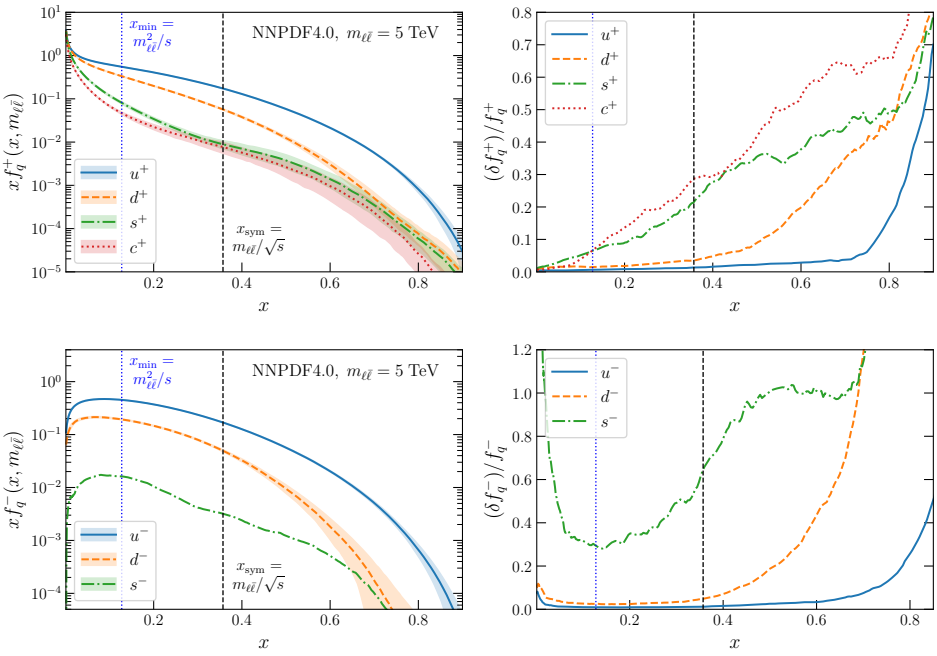

Figure 6.8: Comparison of the xf_q^+ (top) and xf_q^- (bottom) quark PDF combinations for the up, down, strange, and charm quarks, evaluated at $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ for NNPDF4.0 NNLO. The right panels display the relative 68% CL uncertainties. The two vertical lines indicate $x_{\rm min} = m^2_{\ell\bar{\ell}}/s$, the smallest allowed value of x for dilepton DY production for a collider CoM energy $\sqrt{s} = 14\,\text{TeV}$, and the value of x corresponding to a symmetric partonic collision $x_1 = x_2$, namely $x_{\rm sym} = m_{\ell\bar{\ell}}/\sqrt{s}$.

First, we provide a qualitative assessment of the relative size of the PDFs corresponding to individual quark flavors, both for the total and valence PDFs. In fig. 6.8 we compare the total xf_q^+ and valence xf_q^- quark PDF combinations for the up, down, strange, and charm quarks, evaluated at $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ with the NNPDF4.0 NNLO PDF set. The right panels display the corresponding relative 68% CL uncertainties. The leftmost vertical line indicates $x_{\rm min} = m^2_{\ell\bar{\ell}}/s$, the smallest allowed value of x for dilepton DY production with invariant mass $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ for a collider CoM energy $\sqrt{s} = 14\,\text{TeV}$. The rightmost vertical line corresponds to the value of x in a symmetric partonic collision where $x_1 = x_2$, namely $x_{\rm sym} \equiv m_{\ell\bar{\ell}}/\sqrt{s}$.

From fig. 6.8 one can observe that for $x \lesssim 0.3$ there is a clear hierarchy $f_u^+ > f_d^+ > f_s^+ > f_c^+$, while for larger x values the strange and charm PDFs become of comparable magnitude. The up and down quarks, both for xf_q^+ and xf_q^-, are significantly larger than the second-generation quark PDFs until $x \simeq 0.7$, and hence dominate the large-$m_{\ell\bar{\ell}}$ differential distributions in Drell–Yan production. PDF uncertainties grow rapidly with x, reflecting the lack of direct experimental

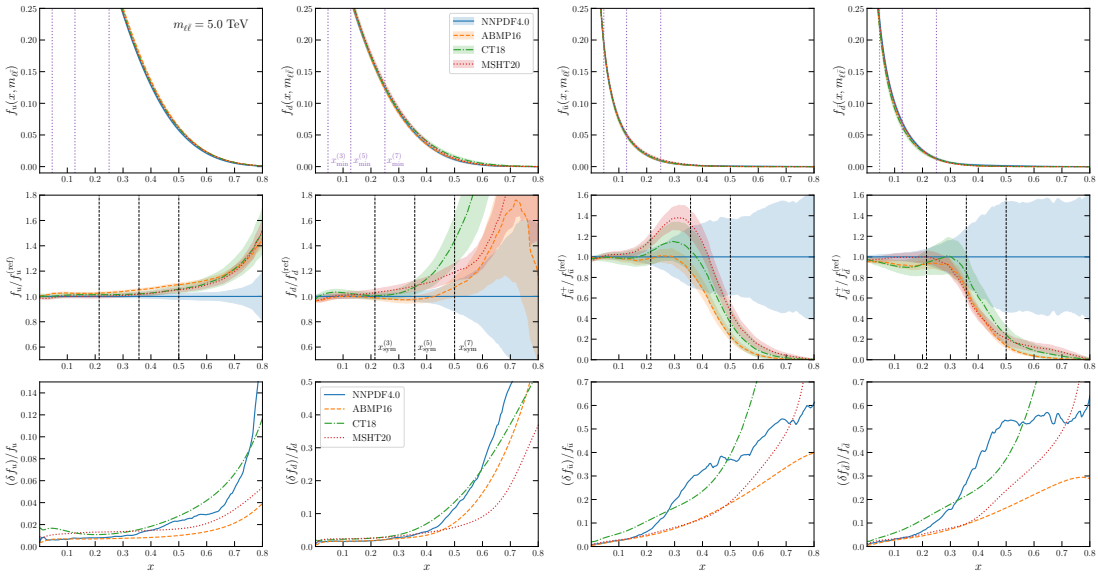

Figure 6.9: The up and down quark and antiquark PDFs evaluated at $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ for NNPDF4.0, CT18, MSHT20, and ABMP16 in the x region relevant for high-mass Drell–Yan production. The upper panels display the absolute PDFs, the middle ones their ratio to the central NNPDF4.0 value, and the bottom panels the relative 68% CL uncertainties. The vertical lines in the top row indicate the values of $x_{\text{min}} = m^2_{\ell\bar{\ell}}/s$ and in the central row those of $x_{\text{sym}} = m_{\ell\bar{\ell}}/\sqrt{s}$ for three different values $m_{\ell\bar{\ell}} = 3, 5, 7\,\text{TeV}$. Note that in the second row the range on the y axis is not the same for quarks and antiquarks, and in the third row also for up and down quarks. Note also that the PDFs, their ratios and their uncertainties are essentially unchanged in the displayed large-x region in the range $1\,\text{TeV} < m_{\ell\bar{\ell}} < 7\,\text{TeV}$.

constraints. The same qualitative behavior of the lighter versus heavier flavor PDFs is observed for other PDF sets. Given the hierarchy $f_u^{\pm}, f_d^{\pm} \gg f_s^{\pm}, f_c^{\pm}$, in the following we will discuss only the behavior of the first-generation quark and antiquark PDFs which are those relevant for the interpretation of neutral-current Drell–Yan production in the kinematic region used for BSM searches.

We next compare the large-x behavior of the four PDF sets ABMP16, CT18, MSHT20, and NNPDF4.0 in fig. 6.9 for $m_{\ell\bar{\ell}} = 5\,\text{TeV}$. We display from top to bottom the absolute PDFs, their ratio to the central NNPDF4.0 value, and their relative 68% CL uncertainties. As in the case of fig. 6.8, we indicate with two vertical lines the values of x_{min} and x_{sym}, both for $m_{\ell\bar{\ell}} = 5\,\text{TeV}$, and for a smaller and a larger value of $m_{\ell\bar{\ell}}$, namely for $m_{\ell\bar{\ell}} = 3\,\text{TeV}$ and $m_{\ell\bar{\ell}} = 7\,\text{TeV}$. For clarity, the values of x_{min} are only shown in the top row of plots, and the values of x_{sym} in the central row. Note that the scale dependence of the PDFs in this range of x and invariant mass is very slight. Indeed, the PDFs shown in fig. 6.8 are

essentially unchanged at $m_{\ell\bar{\ell}} = 3\,\text{TeV}$ or $m_{\ell\bar{\ell}} = 7\,\text{TeV}$; only the corresponding ranges of x_1, x_2 vary significantly.

Good agreement between all PDF sets is found up to around $x \simeq 0.4$. For $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ this corresponds to the value of x_{sym}, i.e. central rapidity. For larger values of $x \gtrsim 0.4$, the up quark PDF xf_u from the NNPDF4.0 set is somewhat suppressed in comparison to the other three sets, which in turn agree among each other. A rather stronger suppression of NNPDF4.0 in comparison to CT18 is observed for the down quark, with MSHT20 and ABMP16 in a somewhat intermediate situation. The opposite behavior is found in the same region $x \gtrsim 0.4$ for antiquark PDFs $xf_{\bar{u}}$ and $xf_{\bar{d}}$: namely, the NNPDF4.0 PDF is significantly larger than that of the other sets. It follows that for a lower invariant mass value $m_{\ell\bar{\ell}} = 3\,\text{TeV}$, all PDF sets are in agreement in the x range in which they are probed, while for a higher value $m_{\ell\bar{\ell}} = 7\,\text{TeV}$ the disagreement between NNPDF4.0 and the other PDF sets is present for most of the $x \geqslant x_{\text{min}}$ range.

It is interesting to observe that in the region with $0.4 \lesssim x \lesssim 0.6$ the PDFs are constrained by some fixed-target DIS structure functions and by forward W and Z production data from LHCb. Hence, at the edge of the data region NNPDF4.0 starts disagreeing with the other global PDF sets considered here, with the disagreement getting more marked as x grows outside the region covered by the data. Qualitatively, NNPDF4.0 is characterized by the fact that the quark PDFs drop faster as a function of x, and the antiquark PDFs drop less fast as x grows towards $x = 1$. As we will show next, this feature will lead to significant differences in the antisymmetric PDF luminosities $\mathcal{L}_{A,q}$ as the value of the dilepton invariant mass $m_{\ell\bar{\ell}}$ is increased.

The relative PDFs uncertainties, shown in the lower panels in fig. 6.9 in all cases grow with x (see also fig. 6.8). The largest PDF uncertainties correspond to either CT18 or NNPDF4.0, depending on the x range and the PDF flavor. Specifically, the NNPDF4.0 uncertainties are largest for f_d in the region $x \gtrsim 0.6$ and for $f_{\bar{u}}$ and $f_{\bar{d}}$ when $0.3 \lesssim x \lesssim 0.5$. The smallest PDF uncertainties are displayed by ABMP16 and MSHT20.

The different behavior of the rate of decrease with x of PDFs in the large x region, specifically comparing NNPDF4.0 to other PDF sets, can be seen most clearly from a comparison off effective asymptotic exponents (cf. Ball, Nocera, et al. 2016)

$$\beta_{a,q}(x,Q) \equiv \frac{\partial \ln |xf_q(x,Q)|}{\partial \ln(1-x)}\,, \tag{6.29}$$

which of course for PDFs of the form of eq. (6.24) just coincide with the exponent b up to $O(1-x)$ corrections. In fig. 6.10 we compare the values of $\beta_{a,q}(x, m_{\ell\bar{\ell}})$ for ABMP16, CT18, MSHT20, and NNPDF4.0 evaluated at $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ for the up and down quark and antiquark PDFs in the x range of fig. 6.8.

It is clear that while all PDF sets have a similar effective asymptotic exponent for $x \lesssim 0.35$, a different behavior of NNPDF4.0 in comparison to other determinations sets in for $x \gtrsim 0.4$. Specifically, for quarks the NNPDF4.0 exponents are always larger, and for antiquarks smaller than those found with other PDF sets. Interestingly, whereas for the up quark the effective exponent $\beta_{a,u}$ is approxi-

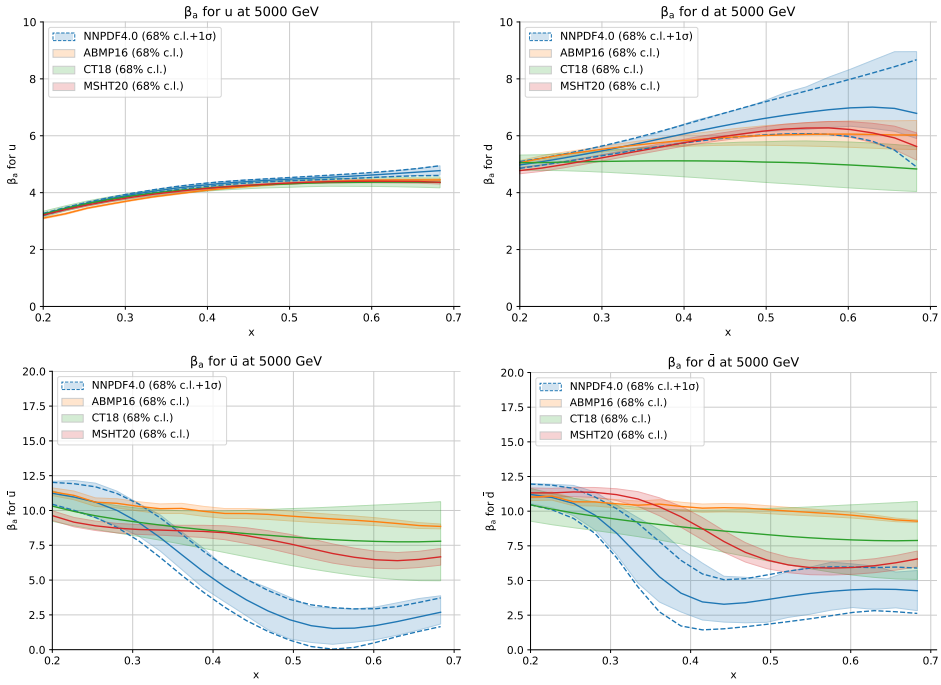

Figure 6.10: The large-x asymptotic exponents $\beta_{a,q}(x, m_{\ell\bar{\ell}})$, defined in eq. (6.29), for ABMP16, CT18, MSHT20, and NNPDF4.0 evaluated at $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ for the up and down quark and antiquark PDFs.

mately constant for all PDF sets when $x \gtrsim 0.4$, with the NNPDF4.0 value being just slightly higher and slowly increasing, for the down quark and all antiquarks this approximately constant behavior is seen for other PDF sets but not for NNPDF4.0. Specifically, for the NNPDF4.0 down quark the exponent slowly but markedly increases for $x \gtrsim 0.3$, together with its uncertainty. In the case of NNPDF4.0 for both antiquarks the exponent rapidly drops in the region $0.3 \lesssim x \lesssim 0.4$. This is consistent with the observation at the PDF level (fig. 6.9) that for NNPDF4.0 at large-x, as compared to the other groups, the up and especially the down quark fall off more rapidly, while the antiquark PDFs drop more slowly. Note in particular that for the down PDF the antiquark effective exponent is significantly smaller than the quark effective exponent for all $x \gtrsim 0.4$.

The fact that a modification in behavior of the effective down quark and especially antiquark PDFs is observed at the edge of the data region for NNPDF4.0, but not for other PDF sets, suggests that this might be related to the fact that NNPDF4.0 generally adopts a more flexible PDF parametrization in comparison to other groups. Also, the uncertainties on the effective exponents $\beta_{a,q}(x, m_{\ell\bar{\ell}})$ tend to be larger for NNPDF4.0 (and also to a lesser extent for CT18) in comparison to those of other groups. Note however that the full PDF uncertainty

contains also a contribution from the overall magnitude, which is not captured by the effective exponents displayed here.

6.2.3 Parton luminosities

We finally turn to the behavior of parton luminosities, with particular regard for the antisymmetric combination which is relevant for the forward-backward asymmetry. As for PDFs, we first assess the qualitative features of the luminosities corresponding to different quark flavors. Specifically, the symmetric $\mathcal{L}_{S,q}$ and antisymmetric $\mathcal{L}_{A,q}$ luminosities eq. (6.12) for individual flavors are displayed in fig. 6.11, evaluated with NNPDF4.0 NNLO for $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ and $\sqrt{s} = 14\,\text{TeV}$. The left panels display the absolute luminosities (in logarithmic and linear scale respectively for the y and x axes) while the right panels show the corresponding PDF uncertainties (relative and absolute for $\mathcal{L}_{S,q}$ and $\mathcal{L}_{A,q}$, respectively). The bottom and top x-axes in each plot show respectively the values of x_1 and x_2 at which the luminosities are being evaluated, within the allowed range $x \geqslant x_{\text{sym}} = m_{\ell\bar{\ell}}/\sqrt{s}$, with the convention $x_1 > x_2$.

The symmetric parton luminosities exhibit of course the same hierarchy between flavors as the corresponding PDF plots of fig. 6.8. The luminosity $\mathcal{L}_{S,q}$ drops rapidly for $x_1 \gtrsim 0.6$. PDF uncertainties depend weakly on x up to $x_1 \gtrsim 0.8$, after which they blow up, and range between $\sim 20\%$ for the up quark luminosity to $\sim 60\%$ for the charm quark one, with down and strange intermediate and of similar magnitude.

As displayed in fig. 6.12, the light quark symmetric luminosities of other global PDF sets are qualitatively similar. We show $\mathcal{L}_{S,u}$, $\mathcal{L}_{S,d}$, and their weighted sum that enters the enters the symmetric coefficient $g_{S,q}$ in eq. (6.18) for the NNPDF4.0, ABMP16, CT18, and MSHT20 at $m_{\ell\bar{\ell}} = 5\,\text{TeV}$. The luminosities are multiplied by the effective charges S_q defined in eq. (6.8), and the bottom panels display the corresponding 68% CL PDF uncertainties. Good agreement between the four sets, with a similar shape of $\mathcal{L}_{S,q}$, is observed. The PDF luminosities for the dominant $\mathcal{L}_{S,u}$ contribution are the largest for NNPDF4.0.

Turning to the antisymmetric PDF luminosities $\mathcal{L}_{A,q}$, we note that, for NNPDF4.0, while the up luminosity is positive, the central value of the down luminosity is negative, though the luminosity is compatible with zero at the one sigma level. Recalling from fig. 6.8 that $x f_{\bar{d}}$ itself is positive for all values of x, this provides an explicit example in which the condition eq. (6.27) is satisfied without the valence combination being negative. We conclude that for NNPDF4.0, the faster drop of the quark distribution and slower drop of the antiquark distribution that was displayed by the effective exponents of fig. 6.10 leads to a negative antisymmetric luminosity, in agreement with eq. (6.27). The absolute PDF uncertainties are of a similar size for $\mathcal{L}_{A,u}$ and $\mathcal{L}_{A,d}$, with a different shape reflecting the underlying central values.

We compare in fig. 6.13 the behavior of the antisymmetric luminosities for all PDF sets for $m_{\ell\bar{\ell}} = 3\,\text{TeV}$ (top) and $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ (bottom). In order to facilitate the understanding of the way the PDF behavior determines that of the asymmetry,

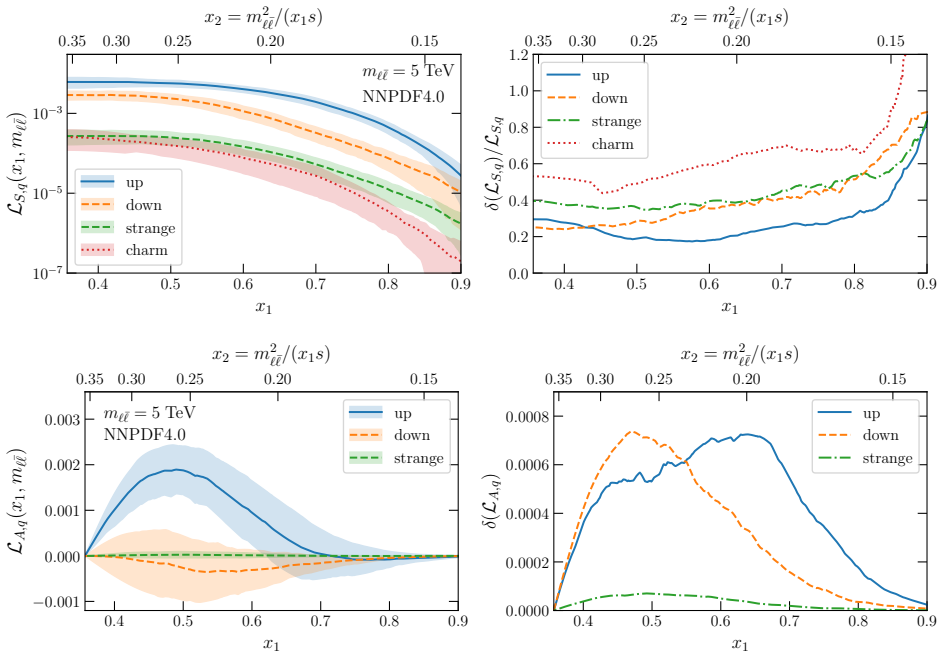

Figure 6.11: The symmetric $\mathcal{L}_{S,q}$ (top) and antisymmetric $\mathcal{L}_{A,q}$ (bottom) parton luminosities (left) and relative uncertainties (right) evaluated with NNPDF4.0 NNLO at $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ and $\sqrt{s} = 14\,\text{TeV}$. The bottom and top x-axes in each plot show respectively the values of x_1 and x_2 at which the luminosities are being evaluated, within the allowed range $x \geqslant x_{\text{sym}} = m_{\ell\bar{\ell}}/\sqrt{s}$, with the convention $x_1 > x_2$.

we show both the contribution of individual flavors and the total contribution to the antisymmetric coefficient $g_{A,q}$ of eq. (6.19). Namely, in fig. 6.13 the luminosities corresponding to individual flavors are multiplied by the corresponding flavor-dependent effective charges A_q defined in eq. (6.8): from left to right we display $\mathcal{L}_{A,u}$, $\mathcal{L}_{A,d}$, and their weighted sum which determines the sign and magnitude of the total forward-backward asymmetry. The corresponding absolute PDF uncertainties for each of the four PDF sets are displayed in fig. 6.14.

fig. 6.13 shows that for ABMP16, CT18, and MSHT20 the antisymmetric parton luminosities depend only mildly on $m_{\ell\bar{\ell}}$, whereas for NNPDF4.0 they exhibit a strong $m_{\ell\bar{\ell}}$ dependence. Indeed, for dilepton invariant masses of $m_{\ell\bar{\ell}} = 3\,\text{TeV}$ there is good agreement between the three groups, but for $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ the NNPDF4.0 up quark luminosity, while preserving a similar valence-like shape, is suppressed by a factor 2 in comparison to other groups, and the down quark luminosity becomes compatible with zero with a negative central value, as already noted. For all PDF sets and $m_{\ell\bar{\ell}}$ values the weighted sum is dominated by the up quark contribution. The strong scale dependence of $\mathcal{L}_{A,q}$ in NNPDF4.0 reflects the underlying PDF behavior seen in fig. 6.9 and highlighted by the effective

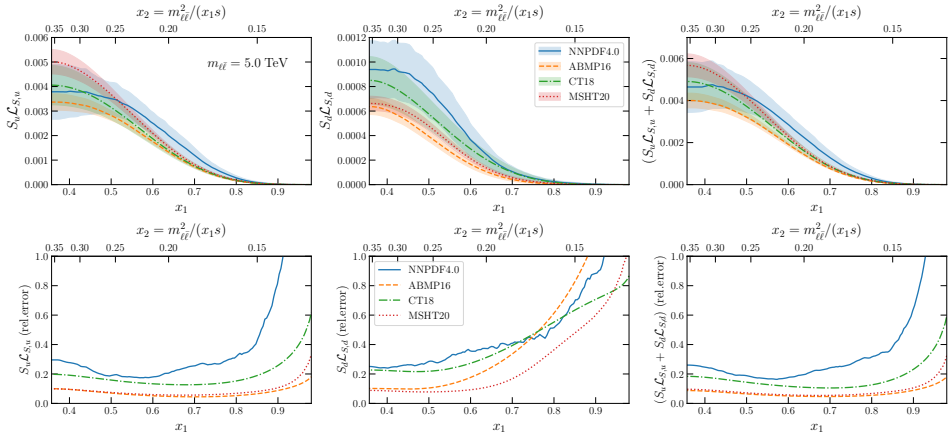

Figure 6.12: The symmetric parton luminosities $\mathcal{L}_{S,q}(x_1, m_{\ell\bar{\ell}})$ for the NNPDF4.0, ABMP16, CT18, and MSHT20 NNLO PDF sets for dilepton invariant masses of $m_{\ell\bar{\ell}} = 5\,\text{TeV}$. The luminosities are multiplied by the effective charges S_q defined in eq. (6.8). From left to right, we display $\mathcal{L}_{S,u}$, $\mathcal{L}_{S,d}$, and their weighted sum that enters the coefficient $g_{S,q}$ in eq. (6.18). The bottom panels display the relative 68% CL PDF uncertainties.

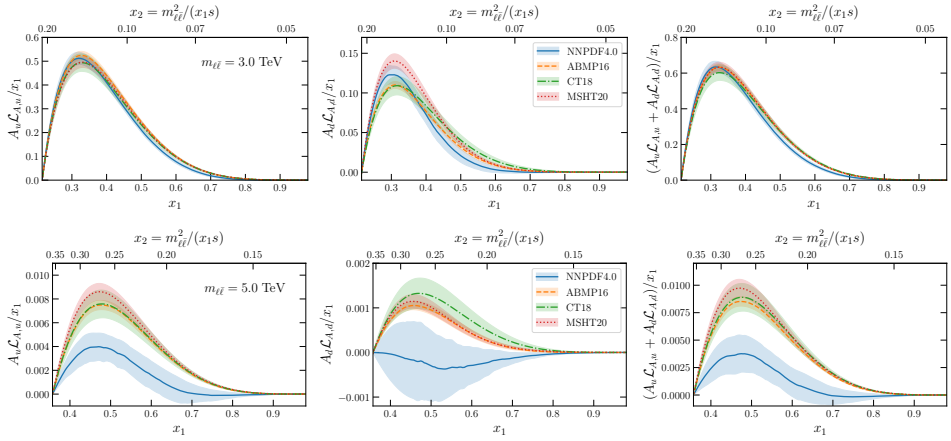

Figure 6.13: The antisymmetric parton luminosities $\mathcal{L}_{A,q}(x_1, m_{\ell\bar{\ell}})$ for the NNPDF4.0, ABMP16, CT18, and MSHT20 NNLO PDF sets for dilepton invariant masses of $m_{\ell\bar{\ell}} = 3\,\text{TeV}$ (top) and $m_{\ell\bar{\ell}} = 5\,\text{TeV}$ (bottom). The luminosities are multiplied by the effective charges A_q defined in eq. (6.8). From left to right, we display $\mathcal{L}_{A,u}$, $\mathcal{L}_{A,d}$, and their weighted sum that enters the coefficient $g_{A,q}$ eq. (6.19).

exponents fig. 6.10. As the scale $m_{\ell\bar{\ell}}$ increases, a range of increasingly large x values is probed, for which, in the case of NNPDF4.0, the quark effective expo-

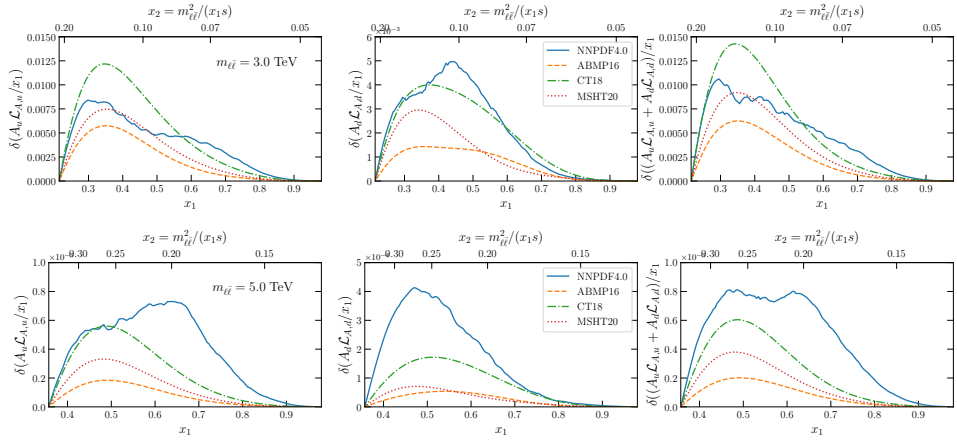

Figure 6.14: Same as fig. 6.13 now for the absolute PDF uncertainties.

nent slowly increases and the antiquark exponent rapidly drops. This leads to a negative asymmetry, following eq. (6.27).

A comparison of the corresponding PDF uncertainties, displayed in fig. 6.14, clearly shows the transition from the data region to the extrapolation region. For $m_{\ell\bar{\ell}} = 3$ TeV the uncertainty $\delta\mathcal{L}_{A,u}$ is generally small for all sets, with CT18 showing a somewhat larger uncertainty for the up quark, and comparable uncertainties for the down quark for all PDF sets. As the scale increases to $m_{\ell\bar{\ell}} = 5$ TeV, where the large-x region is probed, the uncertainty increases, though more markedly for `NNPDF4.0`. For all PDF sets but `NNPDF4.0`, the uncertainty is approximately unchanged when the scale is further increased, while for `NNPDF4.0` it grows markedly.

Finally, in fig. 6.15 we display for all PDF sets the ratio of antisymmetric to symmetric couplings

$$R_{fb} \equiv \frac{\sum_q g_{A,q}}{\sum_{q'} g_{S,q'}}, \tag{6.30}$$

that, according to eq. (6.23), determines at leading order the sign and magnitude of the forward-backward asymmetry distribution $A_{fb}(\cos\theta^*)$. The symmetric and antisymmetric coefficients are obtained by integrating the corresponding symmetric $\mathcal{L}_{S,q}$ and antisymmetric $\mathcal{L}_{A,q}$ partonic luminosities according to eq. (6.19), and the result is shown as a function of the lower integration cut $m^{\min}_{\ell\bar{\ell}}$. In all cases the correlation between PDF uncertainties in the numerator and the denominator are kept into account.

fig. 6.15 shows that, consistently with the behavior of the luminosity of fig. 6.13, for $m^{\min}_{\ell\bar{\ell}} \lesssim 3$ TeV results agree within uncertainties for all PDF sets. The situation is different for higher dilepton invariant masses $m^{\min}_{\ell\bar{\ell}} \gtrsim 3$ TeV: the ratio R_{fb} starts to decrease for `NNPDF4.0`, while it remains approximately constant for the other PDF sets. In particular, for `NNPDF4.0` the coupling ratio vanishes around

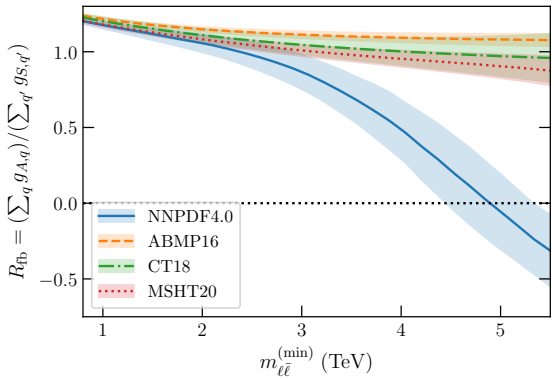

Figure 6.15: The coupling ratio R_{fb}, eq. (6.30), that enters the forward-backward asymmetry $A_{fb}(\cos\theta^*)$ at LO, eq. (6.23), for different PDF sets, as a function of the lower cut in the dilepton invariant mass $m_{\ell\bar{\ell}}^{\rm min}$.

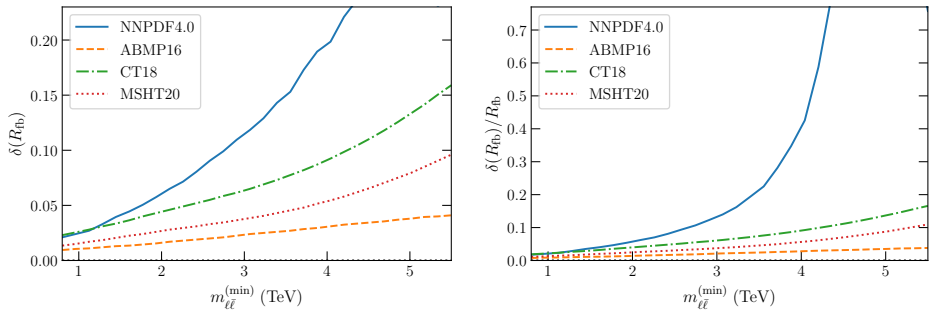

Figure 6.16: The absolute (left) and relative (right panel) uncertainties in the coupling ratio R_{fb} shown in fig. 6.15.

$m_{\ell\bar{\ell}}^{\rm min} \sim 5\,\text{TeV}$, and it becomes negative for yet larger $m_{\ell\bar{\ell}}^{\rm min}$ values. It follows that the forward-backward asymmetry in high-mass Drell–Yan production should decrease and eventually vanish (and possibly even turn negative) in NNPDF4.0 as the $m_{\ell\bar{\ell}}^{\rm min}$ cut is increased, while for CT18, MSHT20, and ABMP16 it should remain positive with a similar magnitude irrespective of the cut $m_{\ell\bar{\ell}}^{\rm min}$ adopted.

fig. 6.16 displays the absolute and relative uncertainties associated to the coupling ratio R_{fb}. We observe that NNPDF4.0 shows the most marked increase of the uncertainties in R_{fb} as $m_{\ell\bar{\ell}}^{\rm min}$ grows. For instance, for $m_{\ell\bar{\ell}}^{\rm min} \gtrsim 4\,\text{TeV}$ the absolute PDF uncertainty in NNPDF4.0 is about twice as large as that found using CT18 four times as large as MSHT20, and about one order of magnitude larger than ABMP16. This trend is magnified for the relative uncertainties due to the decrease in the central value of R_{fb} as $m_{\ell\bar{\ell}}^{\rm min}$ increases.

6.3 THE DRELL-YAN FORWARD-BACKWARD ASYMMETRY AT THE LHC

After the qualitative discussion of the previous sections, here we present results for the $\cos\theta^*$ distributions eq. (6.17) and the forward-backward asymmetry eq. (6.22), with NLO QCD and electroweak corrections included and with realistic selection and acceptance cuts for the LHC at $\sqrt{s} = 14\,\text{TeV}$ and different values of the invariant mass $m_{\ell\bar{\ell}}$ relevant for SM studies and BSM searches.

Computations are performed using `MadGraph5_aMC@NLO`, Alwall et al. 2014, interfaced to PineAPPL, S. Carrazza et al. 2020a; Schwan et al. 2022b, to generate fast interpolation grids. In order to account for realistic detector acceptances, we impose phase-space cuts on the transverse momentum and the pseudo-rapidity of the two leading leptons,

$$p_T^{\ell} > 10\,\text{GeV}\,, \qquad |\eta_\ell| < 2.4\,. \tag{6.31}$$

We then consider various regions of dilepton invariant mass $m_{\ell\bar{\ell}}$: either close to the Z-boson peak ($60\,\text{GeV} < m_{\ell\bar{\ell}} < 120\,\text{GeV}$), relevant for precision SM studies, or the high-mass region relevant for BSM searches, with various choices of a lower mass invariant cutoff ($m_{\ell\bar{\ell}} > 3, 4, 5, 6\,\text{TeV}$). In all cases, in order to facilitate the interpretation of hadron-level results and the connection to the discussion of the PDF features from section 6.2, we also provide results for the two partonic channels that give the largest contribution to the cross-section. As in section 6.2, we compare results obtained using the ABMP16, CT18, MSHT20, and `NNPDF4.0` PDF sets. In all cases, we use the NNLO sets corresponding to the value $\alpha_s(m_Z) = 0.118$ of the strong coupling. Results obtained using the `NNPDF3.1` PDF set are reported in section 6.4.

Before considering the angular distributions, in fig. 6.17 we display the differential distribution in absolute dilepton rapidity $|y_{\ell\bar{\ell}}|$, defined in eq. (6.21), for a dilepton invariant mass of $m_{\ell\bar{\ell}} > 5\,\text{TeV}$. This is the kinematic region relevant for searches of high-mass resonances in the dilepton channel at the LHC, e.g. Aad et al. 2019; Khachatryan et al. 2017. We display the absolute differential distributions with the 68% CL PDF uncertainties (top), the relative PDF uncertainty (center) normalized for each PDF set to the corresponding central prediction, and the pull between the `NNPDF4.0` result, taken as a reference, and other sets (bottom). This pull is defined as

$$\text{Pull}_i = \frac{\sigma^{(0)}_{2,i} - \sigma^{(0)}_{1,i}}{\sqrt{\left(\delta\sigma_{2,i}\right)^2 + \left(\delta\sigma_{1,i}\right)^2}}\,, \qquad i = 1, \ldots, n_{\text{bin}}\,, \tag{6.32}$$

where $\sigma^{(0)}_{1,i}$ and $\sigma^{(0)}_{2,i}$ are the central values of the theory prediction in the i-th bin of the distribution and $\delta\sigma_{1,i}$, $\delta\sigma_{2,i}$ are the corresponding PDF uncertainties. For the central `NNPDF4.0` prediction in the upper panel we also display the contributions from the dominant parton subchannels, namely $u\bar{u} + c\bar{c}$ and $d\bar{d} + s\bar{s} + b\bar{b}$.

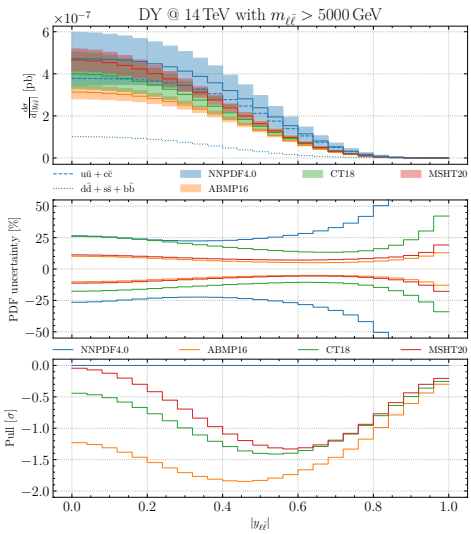

Figure 6.17: The differential distribution in absolute dilepton rapidity $|y_{\ell\bar{\ell}}|$, given in eq. (6.21), for dilepton invariant masses of $m_{\ell\bar{\ell}} > 5\,\text{TeV}$ for neutral current Drell–Yan production at the LHC 14 TeV, obtained using ABMP16, CT18, MSHT20, and NNPDF4.0 NNLO PDFs with $\alpha_s(m_Z) = 0.118$. All uncertainties shown are 68% CL PDF uncertainties, computed at NLO in the QCD and EW couplings with realistic cuts (see text). We show the absolute distributions (top), relative uncertainties (normalized to the central curve of each set, middle) and the pull with respect to the NNPDF4.0 result, eq. (6.32) (bottom). For the central NNPDF4.0 prediction the contributions of the $u\bar{u} + c\bar{c}$ and $d\bar{d} + s\bar{s} + b\bar{b}$ parton subchannels are also shown.

As discussed in section 6.1, the $|y_{\ell\bar{\ell}}|$ distribution depends on the symmetric partonic luminosities $\mathcal{L}_{S,q}$, eq. (6.15), which in turn are driven by the total PDFs xf_q^+. The $|y_{\ell\bar{\ell}}|$ distribution is dominated by the $u\bar{u}$ contribution and its qualitative behaviour is found to be similar for the four PDF sets considered. PDF uncertainties are the largest in NNPDF4.0, ranging between 25% and 50%, and the pull between NNPDF4.0 and CT18 and MSHT20 is at most at the 1.5σ level, and slightly larger with ABMP16. The dependence of the $|y_{\ell\bar{\ell}}|$ distribution on the dilepton mass $m_{\ell\bar{\ell}}$ is moderate, and the same qualitative features are obtained if $m_{\ell\bar{\ell}}$ is lowered down to the Z-peak region, or raised to yet higher values. Hence, for the absolute rapidity distribution there is a reasonable agreement between all PDF sets for all scales considered.

We now turn to the differential distribution in $\cos\theta^*$ and the corresponding forward-backward asymmetry $A_{fb}(\cos\theta^*)$. We first consider the Z-peak region, $60\,\text{GeV} < m_{\ell\bar{\ell}} < 120\,\text{GeV}$, in fig. 6.18. The $\cos\theta^*$ distribution exhibits a small but non-negligible asymmetry, and uncertainties are smallest for NNPDF4.0. The four PDF sets predict a similar behaviour and magnitude of the asymmetry A_{fb}. PDF uncertainties in the asymmetry are comparable for all PDF sets when $\cos\theta^* \approx 0$, and actually largest for NNPDF4.0 when $\cos\theta^* \approx 1$. In all cases the predictions

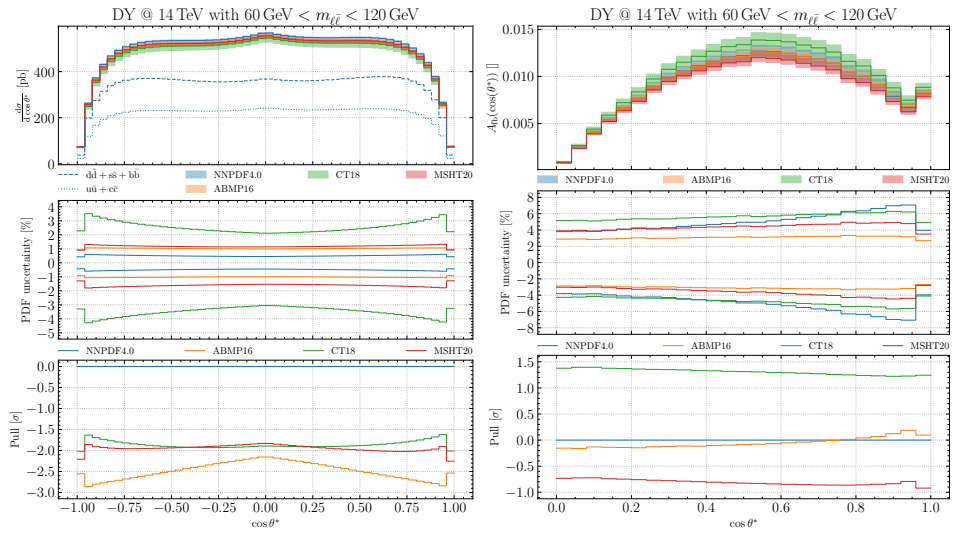

Figure 6.18: Same as fig. 6.17, now for the differential distribution in $\cos\theta^*$ (left) and the corresponding forward-backward asymmetry $A_{fb}(\cos\theta^*)$ (right), in the Z-peak region defined by $60\,\text{GeV} < m_{\ell\bar{\ell}} < 120\,\text{GeV}$.

are compatible within 2σ, with ABMP16 showing larger differences of up to 2.8σ for the $\cos\theta^*$ distribution. Note that the sharp drop-off at the edges $|\cos\theta^*| \approx 1$, appearing in all plots in this section, is a consequence of the phase-space cuts which limit the phase-space volume. Indeed, using LO kinematics

$$|\cos\theta^*| = \tanh\left|\frac{\eta_\ell - \eta_{\bar{\ell}}}{2}\right| = \sqrt{1 - \frac{4(p_T^\ell)^2}{m_{\ell\bar{\ell}}^2}}, \tag{6.33}$$

so $|\cos\theta^*| \approx 1$ requires a lepton pair with either a large rapidity separation, or a very large invariant mass and small transverse momenta.

As expected from the antisymmetric partonic luminosities studied in section 6.2.3, the situation is quite different when considering distributions with a higher dilepton invariant mass range. The angular distribution and forward-backward asymmetry in the high-mass region, for different values of the lower cut in the dilepton invariant mass, namely $m_{\ell\bar{\ell}}^{\text{min}} = 3, 4, 5$ and 6 TeV, are respectively shown in fig. 6.19 and fig. 6.20.

Consistent with the underlying parton luminosities, the $\cos\theta^*$ distribution is dominated by $u\bar{u}$ scattering, while $d\bar{d}$ provides a subdominant contribution. When the lower cut is $m_{\ell\bar{\ell}}^{\text{min}} = 3\,\text{TeV}$ is used, the four PDF sets are in agreement at the 1σ level: they all display a positive forward-backward asymmetry, and exhibit PDF uncertainties ranging between 10% and 15%. As the invariant mass cut is raised, the qualitative behaviour of the angular distribution and asymmetry change substantially for `NNPDF4.0`, while they remain approximately the same for all other PDF sets, consistent with the behaviour of the PDFs and luminosities

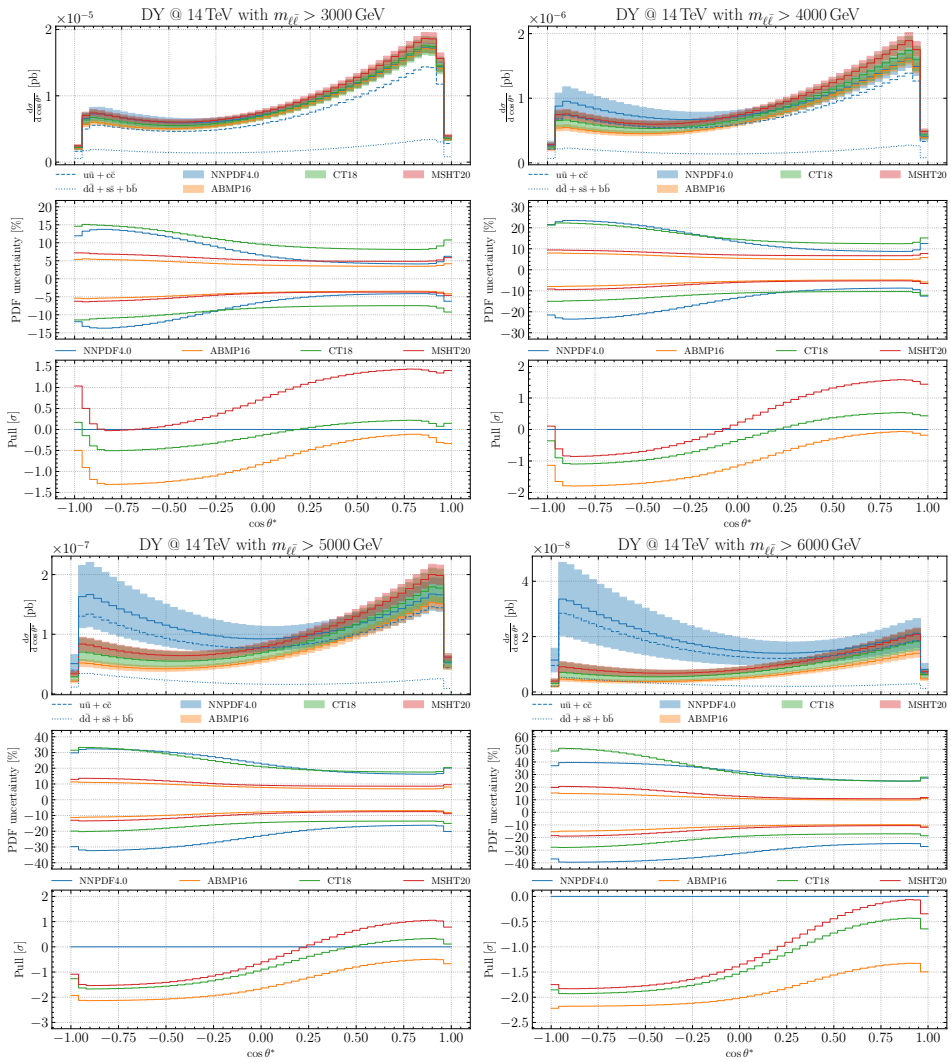

Figure 6.19: Same as fig. 6.18 (left) for different values of the lower cut in the dilepton invariant mass: $m_{\ell\bar{\ell}} \geqslant 3, 4, 5$, and 6 TeV respectively.

discussed in sections 6.2.2 to 6.2.3. Specifically, raising the cut to $m_{\ell\bar{\ell}} \geqslant 4$ TeV, for `NNPDF4.0` the backwards cross-section starts increasing, though the asymmetry remains positive.

For $m_{\ell\bar{\ell}} \geqslant 5$ TeV the central value of the `NNPDF4.0` $\cos\theta^*$ distribution becomes symmetric, though the PDF uncertainty band is rather asymmetric. Also, PDF uncertainties are now the largest for `NNPDF4.0`, reaching up to 30%. Finally, for $m_{\ell\bar{\ell}} \geqslant 6$ TeV the central value of forward-backward asymmetry for `NNPDF4.0` becomes negative, with the PDF uncertainties increasing further so the asymmetry remains compatible with zero at about the 1.1 σ level. For all other PDF sets there

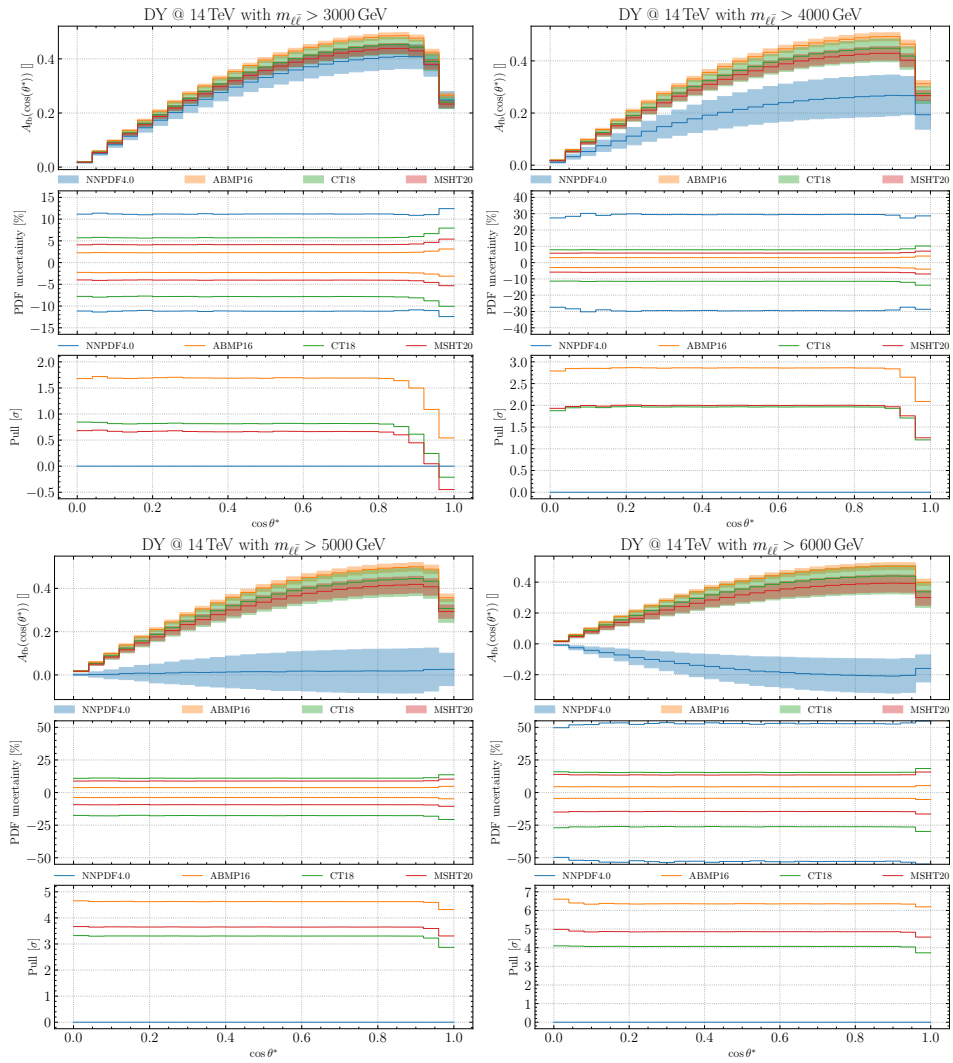

Figure 6.20: Same as fig. 6.18 (right) for different values of the lower cut in the dilepton invariant mass: $m_{\ell\bar{\ell}}^{\min} = 3, 4, 5$, and 6 TeV.

is little change in the shape of the distribution as the dilepton invariant mass cut is increased.

Because of the very large uncertainty on the `NNPDF4.0` result for the $\cos\theta^*$ distribution, even with the highest value of the $m_{\ell\bar{\ell}}^{\min}$ cut, where `NNPDF4.0` finds a symmetric distributions while all other PDF sets find an asymmetry, the pull is always below 2σ. This suggests that the more conservative estimate of `NNPDF4.0` in the extrapolation region might be desirable, and lead to more robust predictions for the forward-backward asymmetry in the high-mass region which is relevant for new physics searches.

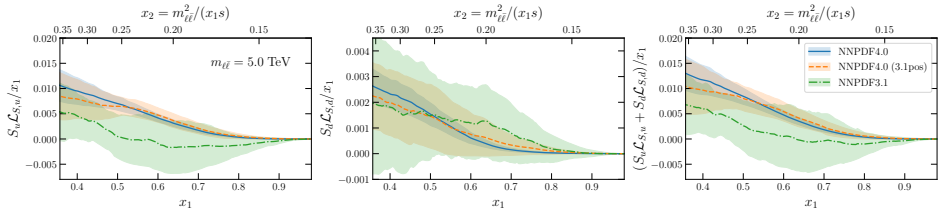

Figure 6.21: Same as fig. 6.12 (upper panels) comparing NNPDF4.0, NNPDF4.0 (3.1pos), and NNPDF3.1.

6.4 A_{FB} IN NNPDF3.1

Finally, we compare partonic luminosities and LHC differential distributions obtained with NNPDF4.0 in section 6.2 and section 6.3 with those based on its predecessor NNPDF3.1, as well as with a variant of NNPDF4.0 where positivity is imposed at the level of observable cross-sections but not at the PDF level, as was the case in NNPDF3.1, which we will denote NNPDF4.0 (3.1pos).

Figure 6.21 compares the symmetric partonic luminosities $\mathcal{L}_{S,q}$ evaluated for $m_{\ell\bar{\ell}} = 5\,\text{TeV}$. The three sets are found to agree within uncertainties, with NNPDF4.0 having the smallest uncertainties. This increase in precision arises only marginally due to the more restrictive positivity constraints imposed, since predictions with the NNPDF4.0 (3.1pos) variant are close to the baseline NNPDF4.0, especially for the $u\bar{u}$ contribution, for both central values and uncertainties. The comparison in fig. 6.21 indicates that phenomenological predictions for high-mass Drell–Yan production based on NNPDF3.1 are expected to be consistent within errors with those of NNPDF4.0 for the contributions symmetric in $\cos\theta^*$, such as the $|y_{\ell\bar{\ell}}|$ distribution.

The antisymmetric luminosities $\mathcal{L}_{A,q}$, relevant for the forward-backward asymmetry, are displayed in fig. 6.22 for $m_{\ell\bar{\ell}} = 3$ and 5 TeV respectively. Their qualitative behavior is similar for all PDF sets, with a marked decrease of PDF uncertainties first from NNPDF3.1 to NNPDF4.0 (3.1pos) then to NNPDF4.0. Specifically, the qualitative $m_{\ell\bar{\ell}}$ dependence of $\mathcal{L}_{A,q}$ remains unchanged. Namely, the positive A_{fb} found for $m_{\ell\bar{\ell}} = 3\,\text{TeV}$ decreases as the dilepton invariant mass is increased. Hence also for the component of the Drell–Yan cross-section which is odd in $\cos\theta^*$ we expect LHC predictions based on NNPDF3.1 to be consistent with those obtained from NNPDF4.0.

These expectations are confirmed by fig. 6.23, which shows the dilepton rapidity $|y_{\ell\bar{\ell}}|$ and the Collins–Soper angle $\cos\theta^*$ distributions for neutral-current DY production at the LHC 14 TeV for dilepton invariant masses of $m_{\ell\bar{\ell}} \geqslant 5\,\text{TeV}$, comparing the baseline NNPDF4.0 predictions with those from NNPDF3.1 and NNPDF4.0 (3.1pos). Indeed, good agreement within the three PDF sets is observed with a significant reduction of PDF uncertainties between NNPDF3.1 and NNPDF4.0, consistent with the behaviour exhibited by the corresponding partonic luminosities.

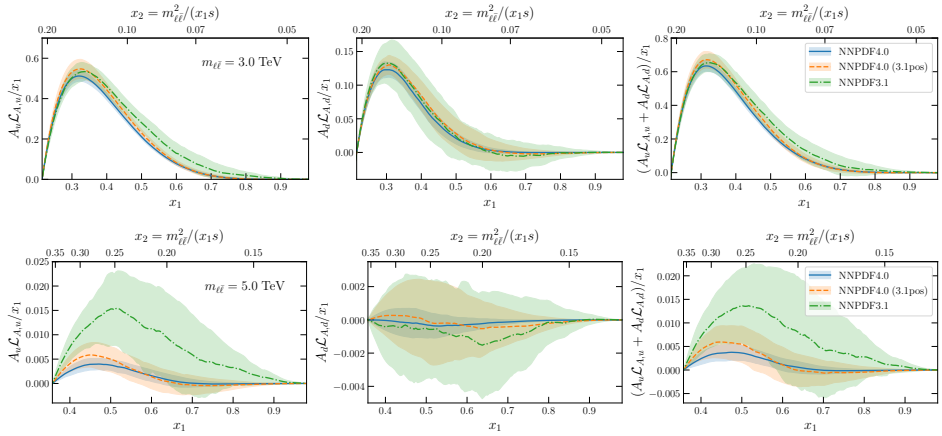

Figure 6.22: Same as fig. 6.13 for the antisymmetric partonic luminosities $\mathcal{L}_{A,q}$, comparing NNPDF4.0, NNPDF4.0 (3.1pos), and NNPDF3.1.

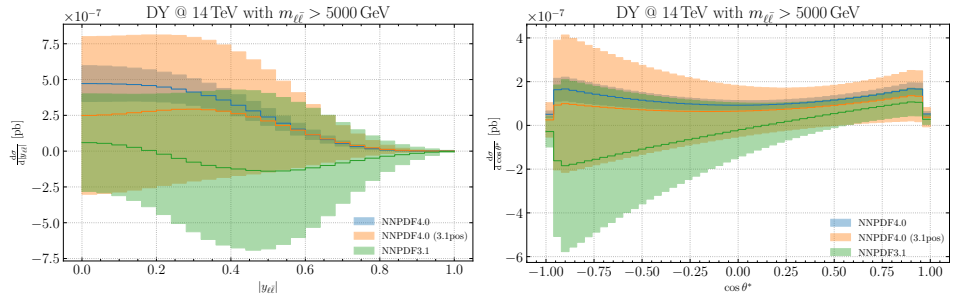

Figure 6.23: Same as figs. 6.17 and 6.19 for the absolute dilepton rapidity $|y_{\ell\bar{\ell}}|$ (left) and the $\cos\theta^*$ (right) distributions for dilepton invariant masses of $m_{\ell\bar{\ell}} \geqslant 5$ TeV comparing NNPDF4.0, NNPDF4.0 (3.1pos), and NNPDF3.1.

6.5 SUMMARY AND OUTLOOK

In this work we have scrutinised the PDF dependence of neutral current Drell–Yan production at large dilepton invariant masses $m_{\ell\bar{\ell}}$, focusing on the behavior of the forward-backward asymmetry A_{fb} in the Collins–Soper angle $\cos\theta^*$, an observable frequently considered in the context of searches for new physics beyond the SM. We have demonstrated that while theoretical predictions for the sign and magnitude of A_{fb} are very similar for all PDF sets in the Z peak region, they depend markedly on the choice of PDF set for large values of $m_{\ell\bar{\ell}}$. We have traced this behavior to that of the PDFs, which agree in the data region, but differ in the large-x region, where PDFs are mostly unconstrained by data.

We have specifically shown that the uncertainty on the asymmetry differs substantially between PDF sets, with NNPDF4.0 displaying a more marked increase as $m_{\ell\bar{\ell}}$ grows, leading to an absolute uncertainty that e.g. for $m^{\text{min}}_{\ell\bar{\ell}} \gtrsim 4$ TeV is

about twice as large as that found using CT18, four times as large as MSHT20, and about one order of magnitude larger than ABMP16. Also, whereas other PDF sets predict a shape of the asymmetry which is unchanged when $m_{\ell\bar{\ell}}$ increases from the Z-peak region to the TeV range, namely a positive asymmetry implying a larger cross-section for $\cos\theta^* \geqslant 0$, NNPDF4.0 finds that as $m_{\ell\bar{\ell}}$ increases, the asymmetry is reduced, and the $\cos\theta^*$ distribution becomes symmetric when $m_{\ell\bar{\ell}}^{\min} \sim 5\,\text{TeV}$.

We have traced this behavior to that of the underlying PDFs in the large-x region, where PDFs are mostly unconstrained by data. Specifically we have seen that in this region NNPDF4.0 has generally wider uncertainties. Also, while for all PDF sets the quark and antiquark distributions vanish as a power of $(1-x)$ as $x \to 1$, for all groups but NNPDF4.0 this power is constant for light quarks to the right of the valence peak, while for NNPDF4.0 it changes as x increases, slowly for up quarks, more rapidly for down quarks and even more rapidly for antiquarks. All this suggests that the different behavior of NNPDF4.0 is due to its more flexible PDF parametrization.

Our general conclusion is that the behavior of the forward-backward asymmetry observed at lower invariant masses is not necessarily reproduced at large masses if flexible enough PDFs are used: the characteristic positive asymmetry observed for low $m_{\ell\bar{\ell}}$ values can be washed out in the high-mass region. Hence, deviations from the traditional expectation of a positive forward-backward asymmetry in high-mass Drell–Yan cannot be taken as an indication of BSM physics, at least based on our current understanding of proton structure in the large-x region.

Turning the argument around, future measurements of the $\cos\theta^*$ distribution and the associated forward-backward asymmetry A_{fb} when included in PDF determinations could help in constraining PDFs at large x. For instance, fig. 6.19 indicates that for $m_{\ell\bar{\ell}}^{\min} = 5\,\text{TeV}$ and $\sqrt{s} = 14\,\text{TeV}$ the asymmetry A_{fb} can be as large as 50% for ABMP16 while it vanishes (within large uncertainties) in the case of NNPDF4.0. By rebinning the $\cos\theta^*$ distribution, for an integrated luminosity of $\mathcal{L} = 6\,\text{ab}^{-1}$, corresponding to the combination at ATLAS and CMS at the end of the HL- LHC data-taking period, $\mathcal{O}(10)$ events are expected in the backward region, with an statistical uncertainty of $\delta_{\text{stat}} \sim 30\%$ which could be sufficient to discriminate between these two limiting scenarios at the 2σ level.

Higher event counts are expected if the $m_{\ell\bar{\ell}}$ cut is loosened, though one is then less sensitive to the large-x region where differences between PDF sets and their uncertainties are maximal. Ultimately, the constraining power of high-mass Drell–Yan in general and of the forward-backward asymmetry in particular can only be addressed by means of a dedicated projections based on binned pseudo-data such as those carried out for the HL- LHC and the Electron Ion Collider in e.g. Abdul Khalek et al. 2018; Khalek et al. 2021. While we leave this exercise for a future study, the investigations presented in this work indicate that A_{fb} at high-invariant masses represents a promising and mostly unexplored channel to pin down large-x light quark and antiquark PDFs at the HL- LHC.

While in this work we have focused on the forward-backward asymmetry in neutral-current Drell–Yan production, similar considerations apply for other processes relevant for BSM searches at high mass at the LHC. Indeed, the HL- LHC will be sensitive to a broad range of hypothetical new massive particles, from resonances in the m_{jj} dijet invariant mass distribution up to 11 TeV, heavy vector triplet resonances decaying into a diboson VV' pair up to 5 TeV, and gluinos with masses up to $m_{\tilde{g}} = 3\,\text{TeV}$ in the minimal supersymmetric standard model (MSSM) with a massless lightest SUSY particle, Cid Vidal et al. 2019. For all these channels, a robust understanding of PDFs and their uncertainties at large x, including the role of methodological and model assumptions, will be necessary to fully exploit the HL- LHC discovery potential for BSM signatures. Conversely, once BSM phenomena have been excluded in some high-energy channel, the corresponding search can be unfolded into a measurement to provide direct constraints on the PDFs in this key large-x region, which in turn will enhance the reach of other searches.

Part III

PROPERTIES AND METHODOLOGY

7 POSITIVITY

As defined in the parton model, PDFs are essentially probability densities, and thus they are positive semi-definite functions over their whole domain. This picture is modified by the factorization scheme, required for NLO or higher order calculations, redefining the PDFs, in such a way that it can assume negative values as well, according to the specific scheme chosen.

While violating the initial intuition associated to the parton model, this is not spoiling any desirable physical property, since PDFs are not observables, but their definition is bound to a factorization scheme (as much as the factorization scale, which is an unphysical scale). Therefore, positivity of physical observables (cross sections and related) is preserved. Nevertheless, might be an interesting question, though academic, to check which schemes yield positive or non-positive PDFs, and whether is possible to tell something about well-known and widespread schemes.

7.1 BACKGROUND AND MOTIVATIONS

The origin of this study has been the observation, brought to NNPDF by external users, that the adoption of `NNPDF3.1` in some searches produced negative results for the central values of physical observables.

The problem was identified in the PDF set having negative values for some flavors, specifically in the large-x region, which is probed by searches, but not covered by data. However, already `NNPDF3.1` imposed the positivity of a set of physical observables, that is a relevant constraint on the PDF shape. But it is not possible to cover, with a finite amount of observable values, the whole spectrum of possible measurements, thus they guarantee more the positivity of standard data-region values (e.g. DY distributions, or DIS structure functions),

but not of any possible BSM search. Moreover, an too strong assumptions in the extrapolation region might also generate some bias towards the SM itself, and the inclusion of constraints on BSM observables would be at the very least arbitrary.

Yet, it is true that in the known cases observables negativity can be traced back to PDFs negativity. Indeed, LO observables are positive by construction (being squared amplitudes), and convoluted with a positive PDF would yield a positive result, though it is not generally true for all orders, since they are affected by factorization subtractions, that can generate negative results also from positive PDFs.

For this reason, we wondered if it were possible to obtain a scheme that would guarantee the positivity of the PDFs, in such a way that two results would be achieved at the same time: LO observables would be positive with these PDFs, and a further theoretical constraint could be imposed on the fit, augmenting the physical information embedded in the fit.

As explained later on, the existence goal is easily obtained, since it is possible to construct "physical" schemes in which the PDF is anchored to an observable, and thus positive by construction. Unfortunately, using PDFs in this schemes would be rather unpractical, since dedicated calculations would be required, adding the details (coefficient functions) of the process linked to the PDFs to any other process as well. So the further requirement we imposed was to obtain a positive subtraction scheme as close as possible from $\overline{\text{MS}}$, such that PDFs could be fitted in that scheme, and tree level observables would become positive anyhow, since they do not require a specific subtraction (thus the difference with respect to $\overline{\text{MS}}$ would be higher order). We started from the assumption that $\overline{\text{MS}}$ itself were not a negative scheme, and we tried to enhance positivity working on the negativity of the coefficient functions in N-space. We found out in the first place that N-space positivity does not coincide with x-space one, and the relation is rather non-trivial. We wanted the PDF values to be positive, not its moments, so we needed to work on n x-space transformation, even though this involved necessarily distributions. Once we had a reasonable candidate, obtained tracing back the structure of $\overline{\text{MS}}$ subtraction in d-dimensions, we started proving its positivity, and studying the relation with the $\overline{\text{MS}}$. There, it strangely appeared from the explicit change of scheme that $\overline{\text{MS}}$ PDFs were more positive than those in the new scheme. Finally, after convincing ourselves of this fact, we turned our argument to prove the positivity of $\overline{\text{MS}}$ scheme itself. This is the most convenient result for the PDF fit, achieving the goal of healing the original *issue* with `NNPDF3.1`, but it is in no way assuring the positivity of resulting observables, that remain beset by subtraction and perturbative truncation.

7.2 POSITIVITY OF PARTONIC CROSS SECTIONS

QCD factorization allows expressing physical cross sections σ as convolutions of partonic cross sections with parton distributions f_i. In the prototypical case

of DIS the cross section is expressed in terms of hadronic structure functions $F(x, Q^2)$, which are then factorized in terms of parton-level structure functions, called coefficient functions C_i:

$$\frac{1}{x}F(x, Q^2) = \sum_i e_i^2 C_i \otimes f_i, \tag{7.1}$$

where the sum runs over all parton species, e_i are quark electric charges, or the sum over all electric charges for the gluon, (for photon-induced DIS, and more in general electroweak charges), $\otimes$ denotes convolution, and we refer to R. K. Ellis et al. 1996 for notations and conventions. The convolution in eq. (7.1) links the three a priori physically distinct scaling variables on which respectively the physical observable F, the partonic cross-section C and the PDF f depend. In the sequel, for clarity, we will denote with x the physically observable variable (Bjorken-x for DIS, or the scaling variable in hadronic collisions), with z the variable on which the coefficient function depends, and with ξ the PDF momentum fraction. Of course, Mellin transformation turns the convolution into an ordinary product and upon transformation all these variables are mapped onto the same N variable.

At LO all factors on the right-hand side of eq. (7.1) are manifestly positive. Indeed, the partonic cross sections (which for DIS at LO are trivial) are defined as the square modulus of amplitudes. The PDFs in turn are defined as operator matrix elements which can be interpreted as probability distributions J. C. Collins and Soper 1982; Curci et al. 1980: for quark PDFs J. C. Collins and Soper 1982

$$\begin{aligned} f_i(\xi) = &\frac{1}{4\pi}\int dy^- e^{-i\xi P^+ y^-} \langle P|\bar{\psi}_i(0, y^-, \vec{0}_T)\gamma^+ \\ &\times \mathcal{P}\exp\left[ig_s\int_0^{y^-} d\bar{y}^- A_a^+(0, \bar{y}^-, \vec{0}_T)\frac{1}{2}\lambda_a\right]\psi_i(0)|P\rangle, \end{aligned} \tag{7.2}$$

where $\mathcal{P}$ denotes path-ordering; P is the four-momentum of the parent hadron in light-cone components and g_s is the strong coupling, with analogous expressions for antiquarks and gluons J. C. Collins and Soper 1982. It can be shown (see e.g. section 6.7 of J. Collins 2013) that the expression eq. (7.2) is a number density, and as such before subtraction of divergences it is positive.

Beyond LO, besides ultraviolet renormalization, both the PDFs and the partonic cross section are beset by collinear singularities which can be factored into the PDF. Before factorization the PDF is a "bare" probability density $f_i^{(0)}$ J. Collins 2013, while after factorization it is a renormalized PDF f_i

$$f_i = \sum_j Z_{ij}^S \otimes f_j^{(0)}. \tag{7.3}$$

In operator language, the factor Z_{ij}^S is a multiplicative renormalization of the operator eq. (7.2), which admits a perturbative expansion

$$Z_{ij}^S(Q^2) = \delta_{ij} + \frac{\alpha_s}{2\pi}\delta_{ij}^S(Q^2) + O(\alpha_s^2), \tag{7.4}$$

where δ_{ij}^S is a counterterm which diverges after regularization is removed, the superscript S denotes the fact that the finite part of the counterterm depends on the choice of a particular subtraction scheme S, and regularization induces a dependence of the counterterm and thus of the renormalization constant on scale.

The counterterm can be determined in a standard way by taking the matrix element of the operator in a state in which the right-hand side of eq. (7.2) is perturbatively computable, such as a free state of a parton i, in which the PDF for finding a parton j is trivially

$$f_j^{i\,(0)}(\xi) = \delta_{ij}\delta(1-\xi), \tag{7.5}$$

imposing a renormalization condition and finally removing the regulator. In practice, this is most easily done J. Collins 2013; Curci et al. 1980 by introducing a probe that couples to the free quark, so for instance computing the structure function eq. (7.1) for deep-inelastic scattering off a free quark. This is the strategy that we will follow in this section, where such a computation will be performed explicitly in a way that fully determines the factorization scheme, both in the $\overline{\text{MS}}$ and in our new positive schemes.

The factorization argument then works as follows. The d-dimensional structure function eq. (7.1) is written as

$$\frac{1}{x}F_i(x,Q^2,\epsilon) = \sum_j e_j^2 C_j \otimes f_j^{i\,(0)} \tag{7.6}$$

$$= \sum_j e_j^2 C_j^S \otimes f_j^{i\,S}; \qquad d = 4-2\epsilon, \tag{7.7}$$

and computed by taking in turn the incoming parton to be each of the parton species, i.e. using eq. (7.5). Of course, the structure function on the l.h.s. then reduces to the unsubtracted, regularized coefficient function, which is essentially the cross-section for scattering off the given incoming free parton. The counterterm is defined by imposing the cancellation of the singularity. Up to NLO, assuming a free incoming parton according to eq. (7.5), substituting in eqs. (7.6) and (7.7)) the perturbative expression eq. (7.4) of the renormalization factor eq. (7.3), and assuming a perturbative expansion of the coefficient functions of the form

$$C_i(z,Q^2) = C_i^{(0)}(z,Q^2) + \frac{\alpha_s}{2\pi}C_i^{(1)}(z,Q^2) + O(\alpha_s^2) \tag{7.8}$$

one gets

$$C_i^S(z,Q^2,\epsilon) = C_i^{(1)}(z,Q^2,\epsilon) - \delta_{qi}^S(z,Q^2,\epsilon)\,, \tag{7.9}$$

where q denotes a quark parton. Note that, up to NLO, imposing finiteness of the DIS structure functions fixes the renormalization in the quark sector because DIS is a probe that only couples to quarks at leading order.

The advantage of determining the counterterms in this way, as opposed to performing a direct computation of the current matrix element eq. (7.2) is that

in operator matrix elements all divergences appear as ultraviolet, while, when computing a structure function for an incoming free parton (or, more generally, a generic partonic cross-section), collinear singularities come from the infrared region of integration over transverse momenta. Hence, one may compute the relevant cross-section using renormalized perturbation theory (i.e., with counterterms already included in the Lagrangian). The only divergences are then of collinear and infrared origin. The regularized partonic cross-section is then finite if the computation is performed with $\epsilon < 0$, and it enjoys the positivity properties of a standard cross-section. This property will be crucial in the argument presented below.

After the subtraction eq. (7.9), the partonic cross-section (coefficient function) is finite in the $\epsilon \to 0$ limit, so one may define the four-dimensional coefficient function as

$$C_i^{(1)\,S}(z) = \lim_{\epsilon \to 0^-} \left(C_i^{(1)}(z, Q^2, \epsilon) - \delta_{qi}^S(z, Q^2, \epsilon) \right), \tag{7.10}$$

where $\epsilon \to 0^-$ denotes the fact that the limit is taken from below, as discussed above. Note that the four-dimensional coefficient function function can depend only on z for dimensional reasons, while the d dimensional one also depends on Q^2 through the combination $\frac{Q^2}{\mu^2}$, where μ^2 is the scale of dimensional regularization. That this subtraction is always possible is the content of factorization theorems J. Collins 2013; Curci et al. 1980. The universal (i.e. process-independent) nature of the collinear singularities ensure that the renormalization conditions on parton distributions, defined as operator matrix elements eq. (7.2) without reference to any specific process, may be determined by the computation of a particular process or set of processes as discussed here.

The finite part of the subtraction is arbitrary and it defines the factorization scheme S. In $\overline{\text{MS}}$ it turns out that in some partonic subchannels the subtracted cross section can be negative: effectively, negative finite parts are factored away from the regularized cross sections, and into the PDFs. These can then also become negative, though whether this happens or not depends on the relative weight of the various subchannels. On the other hand, the residue of the collinear pole is universal – it is given by process-independent splitting functions – and this makes it possible to define its subtraction in a way that preserves positivity of the partonic cross section at the regularized level. If all contributions which are factored away from the partonic cross section and into the PDF remain positive, then the latter also stays positive.

Having explained the general strategy, we now implement it explicitly. We first discuss DIS structure functions. We then turn to double hadronic processes, both quark-induced and gluon induced.

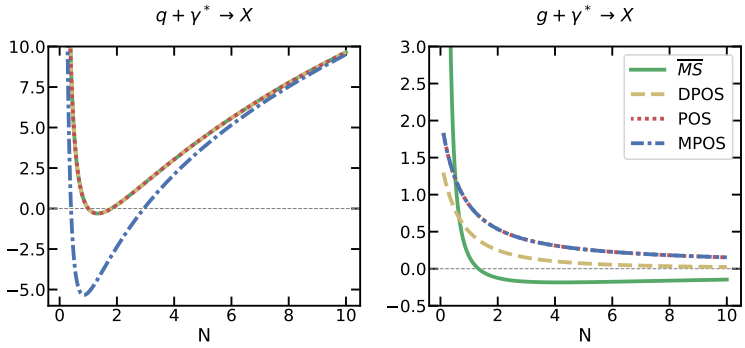

Figure 7.1: Mellin-space NLO contributions to deep-inelastic coefficient functions. The quark (left) and gluon (right) coefficient functions, respectively $C_q^{(1)}$ and $C_g^{(1)}$, eq. (7.11), are shown. The DPOS scheme is defined in eqs. (7.20) and (7.28), the POS scheme is defined in eqs. (7.34) to (7.36), and the MPOS scheme in eqs. (7.81) and (7.82). Results are shown in the $\overline{\text{MS}}$ and DPOS, POS and MPOS schemes. For $C_q^{(1)}$ $\overline{\text{MS}}$, DPOS and POS coincide, and the two curves shown correspond, from top to bottom, to $\overline{\text{MS}}$ and MPOS; for $C_g^{(1)}$ POS and MPOS coincide and the three curves correspond, from bottom to top, to $\overline{\text{MS}}$, DPOS and POS.

7.2.1 Deep-inelastic coefficient functions

At NLO, photon-induced DIS proceeds through the two sub-processes $q + \gamma^* \to X$ and $g + \gamma^* \to X$, in such a way that the contribution of each quark or antiquark flavor to the structure function F_2 can be written as:

$$\frac{1}{x}F_2(x, Q^2) = e_q^2 \left[q + \frac{\alpha_s}{2\pi} \left(C_q^{(1)} \otimes q + C_g^{(1)} \otimes g \right) \right](Q^2), \tag{7.11}$$

where e_q is the electric charge of the quark, on the right-hand side we have omitted the x dependence which arises from the convolution, and the generalization to Z- and W-induced DIS is trivial.

The $\overline{\text{MS}}$ NLO contributions to the coefficient functions C_q and C_g are shown in fig. 7.1 in Mellin space, where the convolution becomes an ordinary product. The Mellin space plot is especially transparent since the x-space cross section is found to high accuracy by computing the inverse Mellin transform in the saddle-point approximation Bonvini et al. 2012: hence, the physical x-space cross section is just the product of the Mellin-space coefficient function and PDF evaluated at the value of N corresponding to the saddle for given kinematics. It is clear from fig. 7.1 that at large N the gluon coefficient function is negative on the real axis: hence, the x-space coefficient function must also be negative because its real moments are negative. This shows that a negative contribution has been factored from the coefficient function into the PDF.

Over-subtraction and the off-diagonal coefficient function

In order to understand what is going on, we look at the dimensionally regularized, unsubtracted gluon coefficient function:

$$C_g^{(1)}(z, Q^2, \epsilon) = \frac{\Gamma(-\epsilon)\left(\frac{\mu_D^2}{\pi\mu^2}\right)^{-\epsilon}\left[8P_{qg}(z) - 16T_R\,\epsilon(3-\epsilon(2-\epsilon))\right]}{16\pi(2-2\epsilon)\Gamma(3-2\epsilon)}\,, \tag{7.12}$$

where

$$\mu_D^2 = \frac{s}{4} = \frac{Q^2(1-z)}{4z}, \tag{7.13}$$

and $s = \frac{Q^2(1-z)}{z}$ is the center-of-mass energy of the $\gamma^* q$ collision. Note that in order to regulate the collinear singularity it is necessary to choose $\epsilon < 0$; it then follows that as ϵ goes to zero from below, $\Gamma(-\epsilon) > 0$ and the unsubtracted coefficient function, eq. (7.12), is positive as it ought to be.

The subtracted $\overline{\text{MS}}$ coefficient function is then given by

$$C_g^{(1)\overline{\text{MS}}}(z) = \lim_{\epsilon\to 0^-}\left[C_g^{(1)}(z, Q^2, \epsilon) - \left(\frac{Q^2}{4\pi\mu^2}\right)^{-\epsilon}\left(-\frac{1}{\epsilon} + \gamma_E\right)P_{qg}(z)\right] \tag{7.14}$$

$$= P_{qg}(z)\left(\ln\left(\frac{1-z}{z}\right) - 4\right) + 3T_R\,, \tag{7.15}$$

where $\epsilon \to 0^-$ denotes the fact that the limit should be taken from below, because the collinear singularity is regulated with $\epsilon < 0$. The P_{qg} splitting function is positive for all z, so for $z > \frac{1}{2}$ the log becomes negative and at large z the coefficient function is negative.

Comparing eqs. (7.12) and (7.14) immediately reveals what happened: the regularized coefficient function contains a term

$$\left(\frac{s/4}{\pi\mu^2}\right)^{-\epsilon} = 1 - \epsilon\ln\left(\frac{Q^2(1-z)/z}{4\pi\mu^2}\right), \tag{7.16}$$

but in the collinear subtraction $\ln\frac{Q^2}{4\pi\mu^2}$ has been subtracted instead. For $z > \frac{1}{2}$, $s < Q^2$ this amounts to over-subtracting, at the larger scale Q^2 instead of the smaller physical scale s. The physical origin of this contribution, and the reason for the mismatch are easy to trace.

Namely, this is the contribution coming from quark emission from the incoming gluon line, and the singularity is due to the collinear singular integration over the transverse momentum of the emitted quark, as revealed by the fact that it is proportional to the corresponding P_{qg} splitting function. The argument of the ensuing collinear log is set by the upper limit of the transverse momentum integration $k_T^{\max}$, which for a $2 \to 2$ process with massless particles in the final state is $k_T^{\max} = \frac{s}{4}$. In $\overline{\text{MS}}$ the collinear subtraction is performed at the scale Q^2, hence

leading to the over-subtraction that we observed, and producing a contribution to the coefficient function which is logarithmically enhanced in the threshold $z \to 1$ limit.

Therefore, this contribution has the same origin as the soft (Sudakov) logarithms which are resummed to all orders when performing threshold resummation S. Catani and Trentadue 1989; Sterman 1987, except that in soft resummation the splitting function is evaluated in the $z \to 1$ limit, and the factor of $\frac{1}{z}$ in the argument of the log is neglected. In fact, threshold resummation can be obtained by identifying (and then renormalization-group improving)

$$|k_T^{\max,\,\mathrm{DIS}}|^2 = \mu_D^2 \tag{7.17}$$

(with μ_D^2 given by eq. (7.13)) as the physical scale in the soft limit Forte and Ridolfi 2003. The over-subtraction is then simply the manifestation of the well-known fact that, in the $\overline{\mathrm{MS}}$ scheme, threshold logs beyond the first are factored in the coefficient function, and not in the PDF Albino and Ball 2001. Indeed, alternative factorization schemes in which these logs are instead included in the PDF have been proposed, in particular the Monte Carlo factorization scheme of Jadach et al. 2016. Note, however, that radiation in off-diagonal parton channels is power-suppressed in the threshold limit, and indeed this contribution is proportional to $\ln(1-z)$, which in Mellin space behaves as $\frac{\ln N}{N}$. This is to be contrasted with the $\left(\frac{\ln(1-z)}{1-z}\right)_+$ behavior, corresponding to $\ln^2 N$, found in diagonal channels, as we shall discuss in section 7.2.1 below. Hence, while it has the same origin, this contribution is not among those included in standard leading-power threshold resummation.

In conclusion, in order to restore positivity it is sufficient to perform the collinear subtraction at the scale $\mu_D^2 = s/4$, eq. (7.17). There is a further subtlety, however. Namely, the factor $2-2\epsilon$ in the denominator of eq. (7.12) is the average over the polarization states of the incoming gluon. Therefore, it should be viewed as an overall prefactor which is common to both the unsubtracted and subtracted coefficient function, and thus must be included in the subtraction term. Because it interferes with a $-\frac{1}{\epsilon}$ pole, not including it, as in $\overline{\mathrm{MS}}$, leads to over-subtraction: the collinear singularity is regulated with $\epsilon < 0$, so $\frac{1}{1-\epsilon} < 1$.

Therefore, we define a modified positivity subtraction as

$$C_g^{(1)\,\mathrm{DPOS}}(z) = \lim_{\epsilon\to 0^-}\left[C_g^{(1)}(z,Q^2,\epsilon) - \frac{1}{1-\epsilon}\left(\frac{\mu_D^2}{\pi\mu^2}\right)^{-\epsilon}\left(-\frac{1}{\epsilon}+\gamma_E\right)P_{qg}(z)\right] \tag{7.18}$$

$$= 3\left[T_R - P_{qg}(z)\right]. \tag{7.19}$$

Note that the normalization of the prefactor is fixed by the requirement of cancellation of the pole. The coefficient function of eq. (7.18) is positive definite, as it is easy to check explicitly. Its Mellin-space form is also shown in fig. 7.1, and it is manifestly positive.

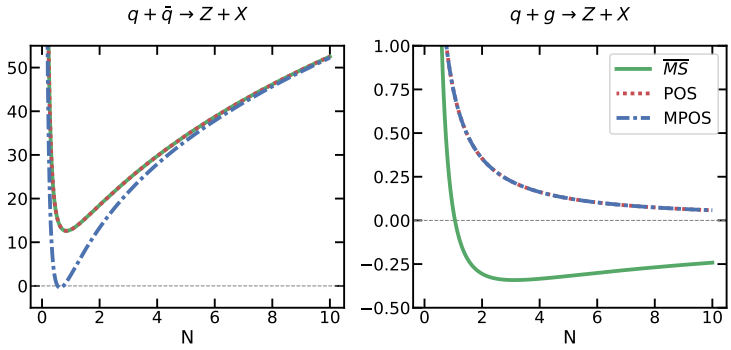

Figure 7.2: Mellin-space NLO contributions to Drell–Yan coefficient functions. The quark (left) and gluon (right) coefficient functions, respectively ${C^q}_q^{(1)}$ and ${C^q}_g^{(1)}$, eq. (7.29), are shown. Results are shown in the $\overline{\text{MS}}$, POS and MPOS schemes. The POS scheme is defined in eqs. (7.34) to (7.36) and eqs. (7.45) to (7.47), and the MPOS scheme in eqs. (7.81) to (7.84)). ${C^q}_q^{(1)}$ $\overline{\text{MS}}$ and POS coincide, and the two curves correspond, from top to bottom, to $\overline{\text{MS}}$ and MPOS; for ${C^q}_g^{(1)}$ POS and MPOS coincide and the two curves correspond, from top to bottom, to $\overline{\text{MS}}$ and POS.

We can rewrite the subtraction which relates the regularized coefficient function, eq. (7.12), to its renormalized counterparts eqs. (7.14) and (7.18) in terms of counterterms according to eq. (7.10), where now $S = \overline{\text{MS}}$, DPOS. We then have

$$C_g^{(1)\,\text{DPOS}}(z) = C_g^{(1)\,\overline{\text{MS}}}(z) - K_{qg}^{\text{DPOS}}(z)\,, \tag{7.20}$$

$$K_{qg}^{\text{DPOS}}(z) = \delta_{qg}^{\overline{\text{MS}}} - \delta_{qg}^{\text{DPOS}} = P_{qg}(z)\left[\ln\left(\frac{1-z}{z}\right) - 1\right]. \tag{7.21}$$

The diagonal coefficient function

We now turn to the diagonal coefficient function: in the $\overline{\text{MS}}$ scheme it is given by

$$C_q^{\overline{\text{MS}}}(z) = \delta(1-z) + \frac{\alpha_s}{2\pi} C_q^{(1)\,\overline{\text{MS}}}(z) \tag{7.22}$$

$$= \delta(1-z)\left(1 + \frac{\alpha_s}{2\pi}\Delta_q^{(1)\,\overline{\text{MS}}}\right) + \frac{\alpha_s}{2\pi}\overline{C}_q^{(1)\,\overline{\text{MS}}}(z)\,, \tag{7.23}$$

where in the last step we have separated off the contribution to $C_q^{(1)\,\overline{\text{MS}}}(z)$ proportional to a Dirac δ (corresponding to a constant in Mellin space) so that

$\overline{C_q}^{(1)\overline{\text{MS}}}(z)$ only contains functions and + distributions. The NLO diagonal coefficient function is given by

$$C_q^{(1)\overline{\text{MS}}}(z) = \lim_{\epsilon\to 0^-}\left[C_q^{(1)}(z,Q^2,\epsilon) - \left(\frac{Q^2}{4\pi\mu^2}\right)^{-\epsilon}\left(-\frac{1}{\epsilon}+\gamma_E\right)P_{qq}(z)\right] \tag{7.24}$$

$$= \lim_{\epsilon\to 0^-}\left[C_q^{(1)}(z,Q^2,\epsilon) - \delta_{qq}^{\overline{\text{MS}}}(z,Q^2,\epsilon)\right] \tag{7.25}$$

$$= C_F\left[\left(p_{qq}(z)\ln\left(\frac{1-z}{z}\right)\right)_+ - \frac{3}{2}\left(\frac{1}{1-z}\right)_+ + 3 + 2z - 4\delta(1-z)\right], \tag{7.26}$$

where $p_{qq}(z)$ is implicitly defined in terms of the quark-quark splitting function $P_{qq}(z)$ as

$$P_{qq}(z) = C_F\left(p_{qq}(z)\right)_+ . \tag{7.27}$$

The Mellin transform of $C_q^{(1)}(z)$ is shown in fig. 7.1. It is clear that the coefficient function is positive for all N: the slightly negative dip of the NLO term in the $N \sim 1$ region is more than compensated by the much larger LO contribution, which in N space is a constant (at $\frac{2\pi}{\alpha_s}$ on the scale of fig. 7.1). As $N \to \infty$, where the NLO contribution diverges (and in principle needs resummation) the growth is actually positive.

A comparison of eq. (7.26) with its off-diagonal counterpart, eq. (7.15), immediately shows what is going on. In this case too, the $\overline{\text{MS}}$ subtraction amounts to an over-subtraction, and indeed the term proportional to $p_{qq}(z)$ in the coefficient function eq. (7.26) has the same origin as the term eq. (7.16), namely, the collinear singularity due to real emission, in this case of a gluon from the incoming quark line. In fact, this is the contribution which is included in standard leading-log threshold resummation. Amusingly, the further (process-dependent) term, proportional to $\left(\frac{1}{1-z}\right)_+$, arises at the next-to-leading log level due to collinear radiation from the outgoing quark line S. Catani and Trentadue 1989, and thus has the same kinematic origin Forte and Ridolfi 2003. One may thus think of generally including these contributions in the PDF by changing the collinear subtraction, as we did above: indeed this is done in the Monte Carlo scheme of Jadach et al. 2016, which aims at including in PDFs all contributions coming from soft radiation.

However, if the goal is ensure positivity, in the diagonal case it is not necessary to modify the $\overline{\text{MS}}$ subtraction prescription. Indeed, in this case over-subtraction actually leads to a more positive coefficient function, due to the fact that the P_{qq} splitting function is negative at large z, where it reduces to a + distribution, i.e., it leads to a negative answer when folded with a positive test function. Of course, this follows from baryon number conservation which requires the vanishing of the first moment of the splitting function. It is in fact easy to check that the $\overline{\text{MS}}$ coefficient function, eq. (7.22), is positive for all $z < 1$. The term proportional to a δ of course has a positive coefficient in the perturbative regime, where it is dominated by the LO term.

We conclude that in order to ensure positivity of the coefficient function it is sufficient to modify the collinear subtraction only in the off-diagonal channel. We therefore set

$$C_q^{(1)\,\mathrm{DPOS}}(z) = C_q^{(1)\,\overline{\mathrm{MS}}}(z)\,. \tag{7.28}$$

eqs. (7.20) and (7.28) thus define the DPOS factorization scheme in the quark channel, in terms of the $\overline{\mathrm{MS}}$ scheme. Note that the considerations underlying the construction of this factorization scheme are based on the structure of the collinear subtraction and the behavior of the splitting functions, and are therefore process-independent.

In order to fully characterize the scheme it is necessary to also consider gluon-induced processes. In Altarelli, Forte, et al. 1998, this was done by considering Higgs production in gluon fusion, with one of the two gluons coming from a proton and the other being taken as a pointlike probe. Equivalently, one might consider Higgs production in photon-gluon fusion. However, the treatment of these processes is essentially the same as that of hadronic processes, to which we thus turn.

7.2.2 Hadronic processes

For hadronic processes[1] the basic factorization formula has the same structure as eq. (7.11), with the structure function replaced by a cross section and the PDF replaced by a parton luminosity $\mathcal{L}_{ij}$: up to NLO

$$\frac{1}{x}\sigma(x,Q^2) = \hat{\sigma}_0\left[\mathcal{L}_{ii} + \frac{\alpha_s}{2\pi}\left(C^i{}_q^{(1)} \otimes \mathcal{L}_{iq} + C^i{}_g^{(1)} \otimes \mathcal{L}_{ig}\right)\right], \tag{7.29}$$

where for simplicity we consider process for which at LO only one partonic channel contributes, so $i = q, g$ labels quark-induced processes (such as Drell–Yan) or gluon-induced processes (such as Higgs production in gluon fusion), $\hat{\sigma}_0$ is the LO partonic cross section and the parton luminosity is

$$\mathcal{L}_{ij} = f_i \otimes f_j\,. \tag{7.30}$$

We first discuss quark-induced processes: their treatment is very close to that of DIS presented in the previous section, so it is sufficient to highlight the differences. We then turn to gluon-induced processes, for which we repeat the analysis of section 7.2.1.

[1] Here *hadronic processes* is used to identify those that in section 0.2 have been called *double hadronic processes*. The more verbose term is used for disambiguation, since also in DIS-like processes an initial hadron is involved, but here it is referred only to those with two hadrons in the initial state.

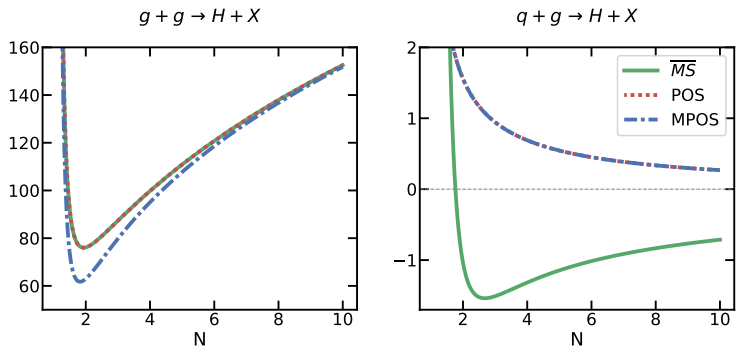

Figure 7.3: Same as fig. 7.2, but now for the Higgs coefficient functions $C^{g}{}_{g}{}^{(1)}$ (left) and $C^{g}{}_{q}{}^{(1)}$ (right).

Quark-induced processes

As a prototype of quark-induced process we consider Drell–Yan production. The NLO coefficient functions (i.e. NLO partonic cross sections normalized to the LO result) are given by

$$C^{q}{}_{q}{}^{(1)\overline{\text{MS}}}(z) = C_F \left[\left(\frac{4\pi^2}{3} - \frac{7}{2} \right) \delta(1-z) + 2 \left(p_{qq}(z) \ln \left(\frac{(1-z)^2}{z} \right) \right)_+ \right]$$
$$= \Delta^{(1)\overline{\text{MS}}}_{qq} \delta(1-z) + 2C_F \left(p_{qq}(z) \ln \left(\frac{(1-z)^2}{z} \right) \right)_+ , \tag{7.31}$$

$$C^{q}{}_{g}{}^{(1)\overline{\text{MS}}}(z) = P_{qg}(z) \left[\ln \left(\frac{(1-z)^2}{z} \right) - 1 \right] + C_F \left[\frac{3}{2} - \frac{3}{2} z^2 + z \right] . \tag{7.32}$$

Comparing the coefficient functions eqs. (7.31) and (7.32) to their DIS counterparts eqs. (7.14) and (7.26) shows that they have the same structure, with a residual logarithmic contribution proportional to the splitting function, due to oversubtraction. The only difference is that the argument of the log is now $\frac{(1-z)^2}{z}$. This is again recognized to be the upper limit of the transverse momentum integral, and to coincide with the argument of the logs whose renormalization-group improvement leads to threshold resummation Forte and Ridolfi 2003: indeed, for a $2 \to 2$ process with a final state particle with mass M^2, and $z = \frac{M^2}{s}$,

$$\mu_h^2 = |k_T^{\text{max, had}}|^2 = \frac{(s-Q^2)^2}{4s} = \frac{Q^2(1-z)^2}{4z} , \tag{7.33}$$

where $Q^2 = M^2$. The coefficient functions, eqs. (7.31) and (7.32), are displayed in fig. 7.2 in Mellin space; their qualitative features are the same as those of the DIS coefficient functions.

Hence, just as in case of DIS, it is possible to define a positive subtraction scheme, which we call POS, and which differs from $\overline{\text{MS}}$ because in the off-diagonal quark-gluon channel the subtraction is performed at the scale μ_h^2, eq. (7.33). Just like for DIS, in the diagonal quark-quark channel there is no need to modify the $\overline{\text{MS}}$ subtraction, which actually makes the coefficient function more positive, so we define a POS factorization of the DY process according to

$$C^{q}{}_{q}{}^{(1)\,\text{POS}}(z) = C^{q}{}_{q}{}^{(1)\,\overline{\text{MS}}}(z)\,, \tag{7.34}$$

$$C^{q}{}_{g}{}^{(1)\,\text{POS}}(z) = C^{q}{}_{g}{}^{(1)\,\overline{\text{MS}}}(z) - K_{qg}^{\text{POS}}(z)\,, \tag{7.35}$$

$$K_{qg}^{\text{POS}}(z) = P_{qg}(z)\left[\ln\left(\frac{(1-z)^2}{z}\right) - 1\right]. \tag{7.36}$$

The quark-gluon coefficient function can be read off eqs. (7.32) and (7.36) and it is easy to check that it is positive definite for all $0 < z < 1$.

Of course, a choice of factorization scheme must be universal. Therefore, it is interesting to check what this choice amounts to if adopted for DIS. Clearly, the hadronic scale eq. (7.33) is always lower than the DIS scale eq. (7.33): $\mu_h^2 < \mu_D^2$. Hence, subtraction in the DPOS scheme amounts to under-subtraction, and if adopted for DIS coefficient function it leads to a DIS coefficient function $C_g^{(1)\,\text{POS}}(z)$ which is actually more positive than that in the DPOS scheme. This is seen in fig. 7.1 (right), where $C_g^{(1)}(z)$ is shown in the $\overline{\text{MS}}$, DPOS and POS schemes.

Gluon-induced processes

In order to fix completely the factorization scheme we turn to gluon-induced hadronic processes. We choose Higgs production in gluon fusion (in the infinite top mass limit) as a prototype, and we repeat the analysis of section 7.2.1, but now for the quark coefficient function $C^{g}{}_{q}{}^{(1)}$. The regularized, unsubtracted expression is (see e.g. Maltoni 2018)

$$C^{g}{}_{q}{}^{(1)}(z, Q^2, \epsilon) = \frac{\Gamma(-\epsilon)\left(\frac{\mu_h^2}{\pi\mu^2}\right)^{-\epsilon}(1-\epsilon)\left[P_{gq}(z) - C_F\frac{(1+z)^2}{2z}\epsilon\right]}{16\pi(2-2\epsilon)\Gamma(3-2\epsilon)}\,, \tag{7.37}$$

where μ_h^2 is given by eq. (7.33), with $Q^2 = M_H^2$, the Higgs square mass. Performing $\overline{\text{MS}}$ subtraction in the usual way we get

$$C^{g}{}_{q}{}^{(1)\,\overline{\text{MS}}}(z) = \lim_{\epsilon\to 0^-}\left[C^{g}{}_{q}{}^{(1)}(z, Q^2, \epsilon) - \left(\frac{Q^2}{4\pi\mu^2}\right)^{-\epsilon}\left(-\frac{1}{\epsilon} + \gamma_E\right)P_{gq}(z)\right] \tag{7.38}$$

$$= P_{gq}(z)\left[\ln\left(\frac{(1-z)^2}{z}\right) - 1\right] + C_F\frac{(1+z)^2}{2z}\,. \tag{7.39}$$

Again, we encounter the same situation that we have seen in the quark channel for DIS, eqs. (7.12) and (7.14): the collinear log has a scale set by the upper limit of the transverse momentum integration, now the hadronic μ_h^2, eq. (7.33), but the $\overline{\text{MS}}$ subtraction is performed at the scale Q^2, which at large z is higher, thus leading to over-subtraction. Indeed, the Mellin-space $\overline{\text{MS}}$ coefficient function ${C^g}_q^{(1)}$, shown in fig. 7.3, is seen to be negative at large N.

As in the quark sector, the problem is fixed by performing the collinear subtraction at the physical scale μ_h^2. Note that also in this case, as for the DIS quark-gluon channel, there is an issue with the sum over gluon polarizations: indeed, because the LO process is in the gluon-gluon channel, even the NLO quark-gluon channel has a gluon in the initial state, leading to a factor $1-\epsilon$ in the denominator of eq. (7.37), which must be accounted for in order to avoid over-subtraction. Hence, we define the POS scheme coefficient function as

$$
{C^g}_q^{(1)\text{POS}}(z) = \lim_{\epsilon\to 0^-}\left[{C^g}_q^{(1)}(z,Q^2,\epsilon) - \frac{1}{1-\epsilon}\left(\frac{\mu_h^2}{\pi\mu^2}\right)^{-\epsilon}\left(-\frac{1}{\epsilon}+\gamma_E\right)P_{qg}(z)\right] \tag{7.40}
$$

$$
= C_F\frac{(1+z)^2}{2z}, \tag{7.41}
$$

with μ_h^2 given by eq. (7.33). The coefficient function is clearly positive. Its Mellin transform is also shown in fig. 7.3.

We finally examine the gluon-gluon NLO coefficient function:

$$
{C^g}_g^{(1)\overline{\text{MS}}}(z) = C_A\left[2\frac{1}{z}\left(zp_{gg}(z)\ln\left(\frac{(1-z)^2}{z}\right)\right)_+ + \left(\frac{473}{36}+\frac{4\pi^2}{3}\right)\delta(1-z) - \frac{11}{3}\frac{(1-z)^3}{z}\right] \tag{7.42}
$$

$$
= \Delta_{gg}^{(1)\overline{\text{MS}}}\delta(1-z) + C_A\left[2\frac{1}{z}\left(zp_{gg}(z)\ln\left(\frac{(1-z)^2}{z}\right)\right)_+ - \frac{11}{3}\frac{(1-z)^3}{z}\right], \tag{7.43}
$$

where, in analogy to eq. (7.27), $p_{gg}(x)$ is implicitly defined by

$$
P_{gg}(z) = C_A\frac{1}{z}\left(zp_{gg}(z)\right)_+ - \frac{n_f}{3}\delta(1-z). \tag{7.44}
$$

As in the diagonal quark channel, the $\overline{\text{MS}}$ subtraction is now multiplied by a splitting function which is negative at large z, for the same physical reason. It therefore leads to a coefficient function which is positive, as seen by inspecting eq. (7.42) and shown in fig. 7.3 (left), so no further scheme change is needed.

Therefore we get

$$C^{g}{}_{g}{}^{(1)\,\mathrm{POS}}(z) = C^{g}{}_{g}{}^{(1)\,\overline{\mathrm{MS}}}(z)\,, \tag{7.45}$$

$$C^{g}{}_{q}{}^{(1)\,\mathrm{POS}}(z) = C^{q}{}_{g}{}^{(1)\,\overline{\mathrm{MS}}}(z) - K^{\mathrm{POS}}_{gq}(z)\,, \tag{7.46}$$

$$K^{\mathrm{POS}}_{gq}(z) = P_{gq}(z)\left[\ln\left(\frac{(1-z)^2}{z}\right) - 1\right]. \tag{7.47}$$

Equations (7.34) to (7.36) and eqs. (7.45) to (7.47) fully define the POS subtraction. We shall see in the next section that they define a positive factorization scheme. Indeed, in the construction presented in this section we have not made use of the detailed from of the partonic cross section, but rather just of the collinear counterterms, expressed in terms of universal splitting functions. Hence, these counterterms, when used in eq. (7.4) define a universal renormalization scheme eq. (7.3) for PDFs, without spoiling PDF universality.

7.3 A POSITIVE FACTORIZATION SCHEME

We will now construct a positive factorization scheme based on the POS subtraction of eqs. (7.34) to (7.36) and eqs. (7.45) to (7.47). We then discuss the scheme transformation from this scheme to the $\overline{\mathrm{MS}}$ scheme and use it to show that PDFs are non-negative in the $\overline{\mathrm{MS}}$ scheme in the perturbative region.

The argument is based on the factorization eqs. (7.6) and (7.7), and, very crudely speaking, amounts to showing that with the POS subtraction, all factors in eq. (7.7) are positive: the left-hand side is positive because it is a physically measurable cross-section, the coefficient C^S function on the right-hand side is positive because the POS subtraction preserves the positivity of the unsubtracted coefficient function C, which is a partonic cross-section, and thus positive before subtraction, but only well-defined in $d > 4$ dimensions.

Taking a Mellin transform of both sides of eqs. (7.6) and (7.7)) all convolutions turn into ordinary products, and it is immediately clear that, because the left-hand side is positive, for the Mellin transformed PDF to be positive it is necessary and sufficient that the coefficient function is positive. However, positivity of the Mellin transform of a function is a necessary condition for its positivity, but not a sufficient one: a negative function may have a positive Mellin transform. The somewhat more complex structure of the discussion below is necessary in order to deal with the necessity of providing an x-space argument.

7.3.1 Positive PDFs

We start by presenting the construction in a simplified setting, namely in the absence of parton mixing. This means that the operators eq. (7.2) whose matrix elements define the PDFs renormalize multiplicatively. This would specifically

correspond to the case of a quark combination that does not mix with the gluon, such as any combination $q^{\rm NS}(Q^2) = q_i(Q^2) - q_j(Q^2)$, where i, j denote generically a quark flavor or antiflavor, with $i \neq j$. We refer to this as a nonsinglet quark combination. We can think of the argument below as applying to such a combination, chosen in such a way that the bare $q^{\rm NS}(Q^2)^{(0)}$, eq. (7.2), is positive – which in general of course will not be true even if q_i and q_j are separately positive. This should be viewed as an academic case – after all, in principle, a positive nonsinglet PDF might not exist – whose purpose is to illustrate the structure of the argument in the absence of parton mixing. We then turn to the realistic case of PDFs that do undergo mixing upon renormalization (which we will refer to as singlet case). The nonsinglet case is simpler, not only because of the absence of mixing, but also because in this case the POS scheme actually coincides with $\overline{\rm MS}$ (i.e., $\overline{\rm MS}$ is already positive).

The nonsinglet case as a toy model

In the nonsinglet case, only the diagonal quark subtraction is relevant: so in the nonsinglet case the DIS structure function eq. (7.11) becomes

$$\frac{1}{x}F_2^{\rm NS}(x,Q^2) = \langle e_i^2\rangle \left[1 + \frac{\alpha_s}{2\pi}C_q^{(1)}\otimes\right] q^{\rm NS}(Q^2)\,, \tag{7.48}$$

where $q^{\rm NS}$ is a difference of two quark or antiquark PDFs, assumed positive and $\langle e_i^2\rangle = \frac{1}{2}\left(e_i^2 + e_j^2\right)$ is the average of their electric charges.

The factorization eqs. (7.6) and (7.7) takes the form

$$\begin{aligned}\frac{1}{x}F_2^{\rm NS}(x,Q^2) &= \langle e_i^2\rangle \lim_{\epsilon\to 0^-}\left[1 + \frac{\alpha_s}{2\pi}C_q^{(1)}(Q^2,\epsilon)\otimes\right]\left[q^{\rm NS}\right]^{(0)} &(7.49)\\ &= \langle e_i^2\rangle \lim_{\epsilon\to 0^-}\left[1 + \frac{\alpha_s}{2\pi}C_q^{(1)\overline{\rm MS}}(Q^2,\epsilon)\otimes\right] \\ &\qquad\qquad \times\left[1 + \frac{\alpha_s}{2\pi}\delta^{\overline{\rm MS}}(Q^2,\epsilon)\otimes\right]\left[q^{\rm NS}\right]^{(0)} &(7.50)\\ &= \langle e_i^2\rangle \left[1 + \frac{\alpha_s}{2\pi}\Delta_q^{(1)\overline{\rm MS}} + \frac{\alpha_s}{2\pi}\bar{C}_q^{(1)\overline{\rm MS}}\otimes\right]\left[q^{\rm NS}\right]^{\overline{\rm MS}}(Q^2)\,, &(7.51)\end{aligned}$$

where $C_q^{(1)\overline{\rm MS}}$, $\overline{C}_q^{(1)\overline{\rm MS}}$, $\Delta_q^{(1)\overline{\rm MS}}$ and $\delta_{qq}^{\overline{\rm MS}}$ have been defined in eqs. (7.23) to (7.25), and the dependence on x on the right-hand side has been omitted because it appears due to the convolution, while the dependence on all other variables has been indicated explicitly.

Now, the discussion of section 7.2.1 shows that, because the bare PDF of eq. (7.2) is a probability density, the three factors which are convoluted in eq. (7.51) are all separately positive when $\epsilon \to 0^-$, i.e. from the negative region, provided only $\mu^2 < \mu_D^2$, with μ_D^2 given by eq. (7.13)[2]. This, as discussed in section 7.2.1

[2]Note that the condition cannot be satisfied in the strict $x \to 1$ limit, but this is as it should be since in the limit the scattering process becomes elastic and it is no longer described by perturbative QCD.

[see in particular eq. (7.26) and fig. 7.1] can be understood as a consequence of the fact that the only region in which the $O(\alpha_s)$ term could overwhelm the LO contribution is the threshold region $z \to 1$, where $\alpha_s \ln(1-z) \sim 1$. However, in this region the $\overline{\text{MS}}$ over-subtraction leads to a coefficient function which is positive because P_{qq} is negative at large z. Consequently, all factors in eq. (7.51) remain positive for all z.

The meaning of the factorization argument eqs. (7.49) and (7.51) can be understood by noting that it is possible to choose a "physical" factorization scheme, Stefano Catani 1996, in which PDFs are identified with physical observables. This means that the coefficient function is set to one to all orders by scheme choice. An example is the "DIS" scheme Diemoz et al. 1988 in which the quark PDF is identified with the DIS structure function, so that eq. (7.48) becomes

$$\frac{1}{x}F_2^{NS}(x,Q^2) = \langle e_i^2\rangle \left[q^{NS}\right]^{DIS}(x,Q^2)\,, \tag{7.52}$$

which holds to all perturbative orders. Comparing this DIS scheme expression of the structure function to the $\overline{\text{MS}}$ expression, eq. (7.11), immediately shows that the quark PDF in the DIS and $\overline{\text{MS}}$ schemes are related by

$$\left[q^{NS}\right]^{DIS}(\xi,Q^2) = \left[1 + \frac{\alpha_s}{2\pi}\Delta_q^{(1)\overline{\text{MS}}} + \frac{\alpha_s}{2\pi}\bar{C}_q^{(1)\overline{\text{MS}}}\otimes\right]\left[q^{NS}\right]^{\overline{\text{MS}}}(Q^2)\,, \tag{7.53}$$

where again we have dropped the ξ, dependence of the convolution on the right-hand side, as in eqs. (7.49) to (7.51).

The $\overline{\text{MS}}$ PDFs can be obtained in terms of the DIS ones by inverting eq. (7.53): perturbative inversion of course gives

$$\left[q^{NS}\right]^{\overline{\text{MS}}}(\xi,Q^2) = \left[1 - \frac{\alpha_s}{2\pi}\Delta_q^{(1)\overline{\text{MS}}} - \frac{\alpha_s}{2\pi}\bar{C}_q^{(1)\overline{\text{MS}}}\otimes\right]\left[q^{NS}\right]^{DIS}(Q^2) + O(\alpha_s^2)\,. \tag{7.54}$$

One may worry that therefore the $\overline{\text{MS}}$ PDFs may turn negative in the large ξ region, where $\alpha_s \ln(1-\xi) \gtrsim 1$ and the last term in square brackets in eq. (7.54), which is negative, may overwhelm the LO contribution term. However, in this region the perturbative inversion is invalid, but it is easy to invert eq. (7.53) exactly in the asymptotic large ξ limit. Letting

$$\begin{aligned}\left[q^{NS}\right]^{DIS}(\xi,Q^2) &= \left[1 + \frac{\alpha_s}{2\pi}\Delta_q^{(1)\overline{\text{MS}}} + \frac{\alpha_s}{2\pi}2C_F\left[\frac{\ln(1-z)}{1-z}\right]_+\otimes\right]\left[q^{NS}\right]^{\overline{\text{MS}}}(Q^2) \\ &\quad + \text{NLL}(1-\xi) \qquad (7.55)\\ &= \left(1 + \frac{\alpha_s}{2\pi}\Delta_q^{(1)\overline{\text{MS}}}\right)\left[1 + c_{LL}\left[\frac{\ln(1-z)}{1-z}\right]_+\otimes\right]\left[q^{NS}\right]^{\overline{\text{MS}}}(Q^2) \\ &\quad + \text{NLL}(1-\xi)\,, \qquad (7.56)\end{aligned}$$

with

$$c_{LL} = \frac{\frac{\alpha_s}{2\pi}2C_F}{1 + \frac{\alpha_s}{2\pi}\Delta_q^{(1)\overline{\text{MS}}}}\,, \tag{7.57}$$

and which holds at the leading $\ln(1-\xi)$ level (LL$(1-\xi)$), inversion can be performed by going to Mellin space and then computing the Mellin inverse term by term in an expansion in powers of α_s. We get

$$\left[q^{NS}\right]^{\overline{MS}}(\xi,Q^2) = \frac{1}{1+\frac{\alpha_s}{2\pi}\Delta_q^{(1)\overline{MS}}}\times$$
$$\left[1-c_{LL}\left[\frac{\ln(1-z)}{\left[1+c_{LL}\ln^2(1-z)/2\right]^2}\frac{1}{1-z}\right]_+\otimes\right]\left[q^{NS}\right]^{DIS}(Q^2)+NLL(1-\xi)\,. \quad (7.58)$$

It is clear that as $\xi\to 1$ the negative LL$(1-\xi)$ contribution actually vanishes.[3]

Now, we observe that $\left[q^{NS}\right]^{DIS}(\xi,Q^2)$ is positive because it is a physical observable. Equation (7.53), which expresses the DIS PDF in terms of the $\overline{MS}$ one, then implies that for $\left[q^{NS}\right]^{\overline{MS}}(\xi,Q^2)$ to be guaranteed to be positive, the $\overline{MS}$ coefficient function must also be positive, otherwise folding a positive $\overline{MS}$ PDF with a negative coefficient function could lead to a negative DIS PDF. So positivity of the $\overline{MS}$ coefficient function is a necessary condition for positivity of the $\overline{MS}$ PDF. However, the inverse of eq. (7.53), expressing the $\overline{MS}$ PDF in terms of the DIS one, implies that the condition is also sufficient, because it gives the $\overline{MS}$ PDF as the convolution of a positive coefficient with a positive PDF. eqs. (7.54) and (7.58) show that the coefficient is indeed positive because in the dangerous $\xi\to 1$ region, where a large negative contribution may arise, inversion can be performed exactly and shown to lead to a positive result. Of course, this argument works for any factorization scheme, and it shows that a necessary and (perturbatively) sufficient condition for the PDFs to be positive is that the coefficient function in that scheme is positive.

The perturbative nature of the argument is worth commenting upon. As discussed at the beginning of this section, the corresponding Mellin space argument is trivial: because in Mellin space the structure function is the product of the PDF times the coefficient function, it follows that positivity of the coefficient function is necessary and sufficient for the positivity of the PDF. However, as already mentioned, Mellin-space positivity is not sufficient for x-space positivity. It is therefore necessary to compute the x-space inverse of the coefficient function, and check that it is still positive.

The inversion is done perturbatively in eq. (7.54), and it leads to a coefficient function which is manifestly positive in most of the z range, except at small and large z, where the coefficient functions blows up, due to high-energy (BFKL) and soft (Sudakov) logs respectively. Consider the large-z case that was discussed above. Upon Mellin transformation, the $z\to 1$ region is mapped onto the $N\to\infty$ region, and specifically, as well known (see e.g. Forte and Ridolfi 2003) powers

[3] A similar argument also applies at small ξ, where the coefficient function also rises, as seen in fig. 7.1. We do not discuss this case in detail since positivity of the $\overline{MS}$ PDF at small ξ is manifest.

of $\ln(1-z)$ are mapped onto powers of $\ln N$. The $\ln N$ logarithmic growth of the coefficient function in this limit is seen in fig. 7.1, where it is apparent that the coefficient function diverges as $N \to \infty$. The N-space inverse of the coefficient function is just its reciprocal, and thus it manifestly vanishes as $N \to \infty$ (while of course remaining positive). One would therefore naively expect that the x-space inverse also vanishes (from the positive side) as $z \to 1$, and this expectation is borne out by the explicit computation presented above in eq. (7.58).[4] Similar arguments apply at higher orders (NNLO and beyond), where the coefficient function grows with a higher order power of $\ln(1-z)$ as $z \to 1$, and at small z, where the coefficient function grows as powers of $\ln \frac{1}{z}$ as $z \to 0$. Hence, either the coefficient function is not logarithmically enhanced, and then the perturbative inverse is manifestly positive, or it is logarithmically enhanced, and then the exact inverse of the enhanced terms can be computed ans also shown to be positive. It is natural to conjecture that an explicit computation of the exact inverse of the full coefficient function would also be positive.

The perturbative assumption is therefore used in two different ways. On the one hand, the NLO correction to the $\overline{\text{MS}}$ coefficient function $\bar{C}_q^{(1)}(z)^{\overline{\text{MS}}}$ is not everywhere positive, as it is apparent from fig. 7.1. However, this is a small correction to the positive coefficient function if $\alpha_s \lesssim 1$, and the overall coefficient function remain positive. This would fail in a region in which α_s blows up. So the full NLO coefficient function remains positive, but only in the perturbative region. On the other hand, the perturbative inversion eq. (7.54) is used to show that positivity of the coefficient function is shared by its inverse, and in regions in which perturbativity would fail it is checked explicitly that this is the case by exact inversion. In this case we conjecture that positivity of the inverse is actually an exact property, even when α_s is arbitrarily large.

The argument based on the physical factorization scheme showing that a positive coefficient function is necessary and (perturbatively) sufficient for a positive PDF is in fact equivalent to the factorization argument eqs. (7.49) to (7.51). Indeed, the operator definition of the quark distribution, eq. (7.2), upon performing a derivative expansion of the Wilson line, leads to the standard expression of its moments in terms of matrix elements of local operators. The interpretation of the bare quark distribution as a probability is then preserved by any physical subtraction scheme such that the matrix elements of Wilson operators are expressed in terms of a measurable quantity. The DIS scheme of eq. (7.48) is of course an example of this scheme. Given the equivalence of the two arguments, one may wonder whether, if at all, perturbativity is used in the argument of eqs. (7.49) to (7.51): specifically, the perturbative inversion of eq. (7.54). The question is an-

[4] Since the Mellin space inverse coefficient function behaves as $[\bar{C}^{(1)}(N)]^{-1} \underset{N\to\infty}{\sim} \frac{1}{\ln^2 N}$ it may appear surprising hat the term in square brackets in eq. (7.58) starts with one. However, it should be born in mind that the Mellin transform of any function which is regular (or indeed integrable) at $x = 1$ vanishes as $\frac{1}{N^k}$, with $k > 0$, hence in Mellin space the suppression of the inverse coefficient function as $N \to \infty$ is a subleading correction to the leading power suppression of $q^{NS}(N)$.

swered in the affirmative: the perturbative inversion is hidden in the step leading from eq. (7.49) to eq. (7.50). Indeed, this step amounts to

$$\left[1+\frac{\alpha_s}{2\pi}C_q^{(1)\overline{\text{MS}}}\otimes\right]^{-1}\left[1+\frac{\alpha_s}{2\pi}C_q^{(1)}(Q^2,\epsilon)\right]=\left[1+\frac{\alpha_s}{2\pi}\delta^{\overline{\text{MS}}}(Q^2,\epsilon)\right]+O(\alpha_s^2)\,, \tag{7.59}$$

i.e. the perturbative inversion of the $\overline{\text{MS}}$ coefficient function. The two arguments are thus seen to coincide. Again, while we only provide a perturbative argument it is natural to conjecture that the argument is in fact exact (i.e. it also holds for large values of α_s).

The POS factorization scheme

Equipped with the results of section 7.3.1 we can turn to the case in which parton mixing is present. This corresponds to the realistic case in which the operators eq. (7.2) mix with the gluon and conversely (at NLO) and with each other at NNLO and beyond. Because at NLO only quark-gluon mixing is present, we refer to this as the singlet case. In order to fully define the factorization scheme at NLO we must thus consider a pair of processes, a quark-induced and a gluon-induced one. The factorization for a pair of hadronic processes can be written as

$$\frac{1}{x}\sigma(x,Q^2)=\hat{\Sigma}_0\otimes\left[1+\frac{\alpha_s}{2\pi}C^{(1)}\otimes\right]f(Q^2)\,. \tag{7.60}$$

In eq. (7.60)

- $\sigma(x,Q^2)$ is a vector of hadronic cross sections

$$\sigma(x,Q^2)=\begin{pmatrix}\sigma^q(x,Q^2)\\ \sigma^g(x,Q^2)\end{pmatrix}, \tag{7.61}$$

  such as the pair of processes of section 7.2.2, namely Drell–Yan and Higgs production in gluon fusion; we are assuming for simplicity and without loss of generality that both are evaluated at the same scale $Q^2=M^2$ (such as when producing an off-shell gauge boson and/or Higgs with the same mass), with a trivial generalization to the case of unequal scales, and the scaling variable is $x=\frac{Q^2}{s}$, with s the hadronic center-of-mass energy;

- $\hat{\Sigma}_0$ is a diagonal matrix of LO partonic cross sections, multiplied by the respective PDFs,

$$\hat{\Sigma}_0(x,Q^2)=\begin{pmatrix}\hat{\sigma}_0^q\,q(x,Q^2) & 0\\ 0 & \sigma_0^g\,g(x,Q^2)\end{pmatrix}, \tag{7.62}$$

  namely the quark and the gluon respectively for Drell–Yan and Higgs;

- $C^{(1)}$ is the two-by-two matrix of NLO coefficient functions $C^{i}{}_{j}{}^{(1)}$ with $i,j=q,g$ defined in eq. (7.29);

- $f(\xi, Q^2)$ is a vector of PDFs that mix upon renormalization:

$$f(\xi, Q^2) = \begin{pmatrix} q(\xi, Q^2) \\ g(\xi, Q^2) \end{pmatrix}. \tag{7.63}$$

Having established a suitable notation, the argument then proceeds in an analogous way as the nonsinglet argument of section 7.3.1, except that now, in order to guarantee positivity of the two-by-two matrix of coefficient functions, we must perform the POS subtraction, which in the diagonal channels (and thus in the nonsinglet case) coincides with $\overline{\text{MS}}$ but in the off-diagonal channel differs from it. Namely, we have

$$\frac{1}{x}\sigma(x, Q^2) = \hat{\Sigma}_0 \otimes \lim_{\epsilon\to 0^-} \left[\mathbb{I} + \frac{\alpha_s}{2\pi} C^{(1)}(Q^2, \epsilon)\otimes\right] f^{(0)} \tag{7.64}$$

$$= \hat{\Sigma}_0 \otimes \lim_{\epsilon\to 0^-} \left[\mathbb{I} + \frac{\alpha_s}{2\pi} C^{(1)\,\text{POS}}(Q^2, \epsilon)\otimes\right] \left[\mathbb{I} + \frac{\alpha_s}{2\pi} \delta^{\text{POS}}(Q^2, \epsilon)\otimes\right] f^{(0)} \tag{7.65}$$

$$= \hat{\Sigma}_0 \otimes \left[\mathbb{I} + \frac{\alpha_s}{2\pi} \Delta^{(1)\,\text{POS}} + \frac{\alpha_s}{2\pi} \overline{C}^{(1)\,\text{POS}}\otimes\right] f^{\text{POS}}(Q^2). \tag{7.66}$$

In eqs. (7.64) to (7.66)

- $\Delta^{(1)\,\text{POS}}$ is the diagonal matrix

$$\Delta^{(1)\,\text{POS}} = \begin{pmatrix} \Delta_{qq}^{(1)\,\overline{\text{MS}}} & 0 \\ 0 & \Delta_{gg}^{(1)\,\overline{\text{MS}}} \end{pmatrix}, \tag{7.67}$$

  with $\Delta_{ii}^{(1)\,\overline{\text{MS}}}$ defined in eq. (7.31) and (7.42) respectively for $i = q$ and $i = g$;

- $\delta^{\text{POS}}(Q^2, \epsilon)$ is a two-by-two matrix of counterterms

$$\delta^{\text{POS}}(z, Q^2, \epsilon) = \left(-\frac{1}{\epsilon} + \gamma_E\right) \begin{pmatrix} \left(\frac{Q^2}{4\pi\mu^2}\right)^{-\epsilon} P_{qq}(z) & \frac{1}{1-\epsilon}\left(\frac{\mu_h^2}{\pi\mu^2}\right)^{-\epsilon} P_{qg}(z) \\ \frac{1}{1-\epsilon}\left(\frac{\mu_h^2}{\pi\mu^2}\right)^{-\epsilon} P_{gq}(z) & \left(\frac{Q^2}{4\pi\mu^2}\right)^{-\epsilon} P_{gg}(z) \end{pmatrix}, \tag{7.68}$$

  with μ_h^2 given by eq. (7.33), so that in the diagonal channels the subtraction is the same as in $\overline{\text{MS}}$, while in the off-diagonal channels it is performed at the physical scale μ_h^2, and also, accounting for the d-dimensional continuation of the average over the polarization of the gluons.

Positivity of the quark and gluon PDF vector $f^{\text{POS}}(Q^2)$, eq. (7.66), now follows from the same argument used to show the positivity of the nonsinglet PDF eq. (7.51). Namely, all factors, which are convoluted in eq. (7.51), are separately

positive when $\epsilon \to 0^-$ and $\mu_h^2 < \mu_D^2$ (with μ_D defined in eq. (7.13)) and in particular, the matrix of POS-scheme coefficient functions is now positive as shown in section 7.2.2.

Also, as in the nonsinglet case, the positivity argument can be formulated in terms of a physical scheme, in which now to all perturbative orders the quark and gluon are defined by

$$\frac{1}{x}\bar{\sigma}(x,Q^2) = f^{\rm PHYS}(x,Q^2)\,, \tag{7.69}$$

where, as in Altarelli, Forte, et al. 1998, the hadronic cross sections $\bar{\sigma}(x,Q^2)$ are computed assuming that one of the two incoming protons is replaced by a beam of antiquarks or a beam of gluons respectively, i.e.

$$\bar{\sigma}(x,Q^2) = \begin{pmatrix} \sigma(x,Q^2)[\bar{q}p \to \gamma^* + X] \\ \sigma(x,Q^2)[gp \to H + X] \end{pmatrix}. \tag{7.70}$$

This hadronic cross section is linear in the PDFs, it coincides with it at LO in any scheme, and, assuming that it coincides with it to all orders, defines the PHYS scheme. Equivalently, one could choose as $\bar{\sigma}$ a DIS structure function in the quark channel, and the cross section for Higgs production in photon-gluon fusion in the gluon channel. The POS and PHYS schemes are then related by

$$f^{\rm PHYS}(x,Q^2) = \left[\mathbb{I} + \frac{\alpha_s}{2\pi}\Delta^{(1)\,\rm POS} + \frac{\alpha_s}{2\pi}\bar{C}^{(1)\,\rm POS}\otimes\right] f^{\rm POS}(Q^2)\,, \tag{7.71}$$

which is perturbatively inverted as

$$f^{\rm POS}(x,Q^2) = \left[\mathbb{I} - \frac{\alpha_s}{2\pi}\Delta^{(1)\,\rm POS} - \frac{\alpha_s}{2\pi}\bar{C}^{(1)\,\rm POS}\otimes\right] f^{\rm PHYS}(Q^2) + O(\alpha_s^2)\,. \tag{7.72}$$

Again, this shows that positivity of the POS-scheme coefficient function is necessary for positivity of the POS-scheme PDFs and sufficient if perturbativity holds. Just like in the case of eq. (7.54), this assumption fails at the endpoints $z \to 0$ and $z \to 1$. However, as well known (cf. R. K. Ellis et al. 1996), and as it is easy to check from the explicit expressions of the matrix elements of $\bar{C}^{(1)}(z)^{\rm POS}$, in both these limits the matrix is diagonal up to power-suppressed corrections. Specifically, in the $z \to 1$ limit the coefficient function matrix is diagonal:

$$\lim_{z\to 1} C^{(1)\,\rm POS}(z,Q^2) = \begin{pmatrix} C^q{}_q^{(1)\,\rm POS} & 0 \\ 0 & C^g{}_g^{(1)\,\rm POS} \end{pmatrix} [1 + O(1-z)]\,. \tag{7.73}$$

Indeed, diagonal coefficient functions grow as $\left(\frac{\ln(1-z)}{(1-z)}\right)_+$ while off-diagonal ones tend to a constant as $z \to 1$. This is clearly seen in the N space plots of figs. 7.2 and 7.3, in which as $N \to \infty$ the diagonal coefficient functions are seen

to grow (as $\ln^2 N$) while the off-diagonal ones vanish (as $\frac{1}{N}$) [5] It follows that at large z the quark and gluon channels decouple, and the perturbativity argument is the same as in the non-singlet case.

Positive PDFs and their scale dependence

In section 7.3.1 we have shown that also in the presence of quark-gluon mixing POS-scheme coefficient functions are positive, and thus in the perturbative regime PDFs are also positive. One can then ask two (closely related) questions. First, at which scale does this conclusion apply, and is it affected by perturbative evolution? And second, which PDF combinations are actually positive? Indeed, as well known, the eigenstates of QCD evolution are the two eigenstates of a mixing matrix between the quark singlet and the gluon, and individual nonsinglet components; any PDF (and thus any observable) can be decomposed into a singlet and nonsinglet component, which evolve independently (see e.g. section 4.3.3 of R. K. Ellis et al. 1996). Of course a difference between two positive quantities is not necessarily positive, so this raises the question of which are actually the positive combinations: the eigenstates of evolution, or individual quark, antiquark and gluons (or indeed something else)?

In order to answer the questions, we start from the observation that the operators whose matrix elements separately define probability densities are the quark operators eq. (7.2), and their antiquark and gluon counterparts. This can be understood physically in a simple way by considering a moment of the PDF: for example, the second moment of the PDF for quark of flavor i is just the matrix element of the energy (Hamiltonian) operator for the corresponding quark, expressed in terms of creation and annihilation operators for the given quark state. Ditto for each antiquark of flavor j, and for the gluon. Hence, at leading order the quantities which are separately positive are individual quark flavors, antiquark flavors, and the gluon.

The argument presented in section 7.3.1 shows that this positivity is preserved for the quark and gluon PDF, which at this order mix to first order in α_s. This argument does not make any assumption about the particular value of Q^2, except that it ought to be in the perturbative region where $\alpha_s(Q^2)$ is small enough. Hence, positivity must necessarily be preserved by QCD evolution.

Actually, that this is the case directly follows from the construction of the positive subtraction scheme. Indeed, QCD evolution of the PDF is a consequence of the Q^2 dependence induced by the factorization into the PDF of scale-dependent collinear logs, i.e., by the scale dependence of the renormalization factor $Z^S_{ij}(Q^2)$ in eqs. (7.3) and (7.4). Indeed, using in these equations the explicit form of the subtraction, as given in eqs. (7.14), (7.24) and (7.38) it follows that upon a change

[5] The same power behavior also holds in the $\overline{\text{MS}}$ scheme, where however the off-diagonal coefficient functions grow as $\ln(1-z)$ as $z \to 1$, corresponding to a $\frac{\ln N}{N}$ behavior of its Mellin transform at large N.

of the scale at which the subtraction is performed, the renormalization factor changes according to

$$Z^S_{ij}(Q'^2) = \left(\delta_{ij} + \frac{\alpha_s(Q'^2)}{2\pi} P_{ij} \ln \frac{Q'^2}{Q^2}\right) \otimes Z_j(Q^2) + O(\alpha_s^2), \tag{7.74}$$

where P_{ij} is the Altarelli-Parisi splitting function. Of course, taken in differential form for infinitesimal scale changes eq. (7.74) is the standard QCD evolution equation.

The POS factorization scheme construction essentially amounts to choosing δ^S_{ij} in eq. (7.4) in such a way that Z^S_{ij} remains positive for all Q^2: in particular, whenever P_{ij} is negative, this will mean that as the scale is increased, the renormalization factor Z^S_{ij} decreases, while (in a positive scheme) remaining positive. Clearly, the condition is more easily satisfied at higher scales because of asymptotic freedom, in agreement with the phenomenological observation Ball et al. 2015; Ball, Del Debbio, Forte, Guffanti, Latorre, Rojo, et al. 2010 that positivity constraints are more restrictive if imposed at low scale and are preserved by evolution.

It is worth noting that a consequence of eq. (7.74) is that, as well known, a scheme change will affect the NLO splitting functions. In particular, in the POS scheme contributions proportional to $\ln \frac{(1-z)^2}{z}$ to the off-diagonal splitting function will now be automatically resummed to all orders when solving the NLO QCD evolution equations. These contributions are actually power-suppressed as $z \to 1$, so this resummation is likely not to have a significant effect: the POS scheme is thus useful as a means to obtain positive PDFs (which is our main goal here), but not necessarily phenomenologically better than the standard $\overline{\text{MS}}$ scheme. On the other hand, in Jadach et al. 2016 a factorization scheme has been advocated, called the Monte Carlo scheme, that is similar in spirit to the POS scheme in the off-diagonal channel, but also modifies the $\overline{\text{MS}}$ subtraction in the diagonal channel by an analogous change of subtraction point. In this Monte Carlo scheme, $\ln(1-z)^2$ contributions in the diagonal channels are also resummed when solving the QCD evolution equation: hence, leading-log threshold (Sudakov) resummation is automatically performed, without having to be added a posteriori. It can be argued that in this Monte Carlo scheme PDFs also resepect positivity Jadach 2020.

7.3.2 Positive schemes vs. $\overline{\text{MS}}$

In the previous section, we have shown that coefficient functions and PDFs in the POS factorization scheme are indeed positive. We would like now to investigate the relation of the POS scheme to other factorization schemes, specifically $\overline{\text{MS}}$, and the related issue of how a positive factorization scheme should be and can be defined.

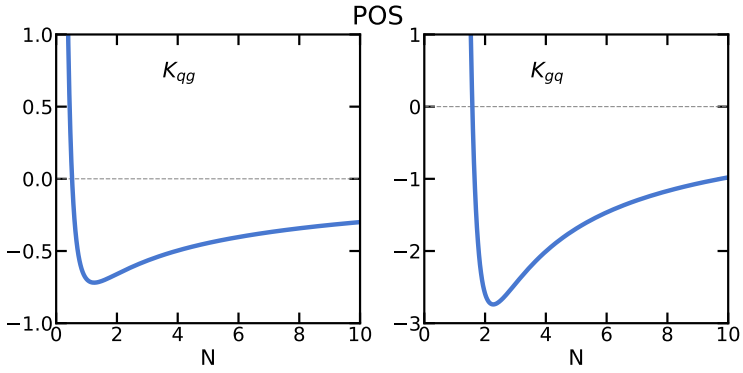

Figure 7.4: The off-diagonal elements of the NLO scheme change matrix K^{POS}, eq. (7.77), in Mellin space.

General positive schemes

The scheme change from POS to $\overline{\text{MS}}$ can be determined using eqs. (7.34) to (7.36) (quark channel) and eqs. (7.45) to (7.47) (gluon channel). We have

$$\left[\mathbb{I}+\frac{\alpha_s}{2\pi}C^{(1)\overline{\text{MS}}}\right] = \left[\mathbb{I}+\frac{\alpha_s}{2\pi}C^{(1)\text{POS}}\right]\otimes\left[\mathbb{I}+\frac{\alpha_s}{2\pi}C^{(1)\text{POS}}\otimes\right]^{-1} \times\left[\mathbb{I}+\frac{\alpha_s}{2\pi}\left(C^{(1)\text{POS}}+K^{\text{POS}}\right)\right] \tag{7.75}$$

$$= \left[\mathbb{I}+\frac{\alpha_s}{2\pi}C^{(1)\text{POS}}\right] \quad \left[\mathbb{I}+\otimes\frac{\alpha_s}{2\pi}K^{\text{POS}}\right], \tag{7.76}$$

where in eq. (7.76) we have written the inverse of the POS scheme coefficient functions in perturbative form according to eq. (7.72). The matrix K^{POS} has the off-diagonal structure

$$K^{\text{POS}} = \left[\ln\left(\frac{(1-z)^2}{z}\right)-1\right]\begin{pmatrix} 0 & P_{qg}(z) \\ P_{gq}(z) & 0\end{pmatrix}. \tag{7.77}$$

The off-diagonal matrix elements of the matrix are displayed in fig. 7.4 in Mellin space. Writing the basic factorization formula eq. (7.66) in the POS and $\overline{\text{MS}}$ schemes, equating the results, and using eq. (7.76) we get

$$f^{\text{POS}}(Q^2) = \left[\mathbb{I}+\frac{\alpha_s}{2\pi}K^{\text{POS}}\otimes\right] f^{\overline{\text{MS}}}(Q^2), \tag{7.78}$$

which gives the scheme change between the $\overline{\text{MS}}$ and POS PDFs.

Inspection of eq. (7.78) immediately shows a possible issue with the POS scheme. Indeed, as well known, momentum conservation implies the pair of relations between the second Mellin moments of splitting functions $\gamma_{qq}(2)+\gamma_{gq}(2)=0$ and

$2n_f\gamma_{qg}(2)+\gamma_{gg}(2)=0$. This relation is verified in the $\overline{\text{MS}}$ scheme: in order for it to remain true in any scheme obtained from $\overline{\text{MS}}$, the scheme change matrix must satisfy

$$K_{qq}+K_{gq}=2n_fK_{qg}+K_{gg}\Big|_{N=2}=0\,, \tag{7.79}$$

where by $K_{ij}\Big|_{N=2}$ we denote the second Mellin moment of the scheme change matrix elements. This relation is not satisfied by the matrix defined in eqs. (7.34) to (7.36) and eqs. (7.45) to (7.47)).

It might therefore be worth considering a variant of the POS scheme, in which momentum conservation is enforced by adding to the diagonal elements of the scheme change matrix a contribution which enforces momentum conservation. This can be done e.g. by adding a soft function, which vanishes both as $z\to 1$ and $z\to 0$. We choose

$$f^{\text{MOM}}(z)=60z^2(1-z)^2\,, \tag{7.80}$$

which has the property that its second Mellin moment equals one: $f^{\text{MOM}}(N=2)=1$. We then define a MPOS scheme as that which is obtained from $\overline{\text{MS}}$ through a scheme change matrix K^{MPOS} whose matrix elements satisfy

$$K^{\text{MPOS}}_{qq}(z)=-f^{\text{MOM}}(z)K^{\text{POS}}_{gq}\Big|_{N=2}\,, \tag{7.81}$$

$$K^{\text{MPOS}}_{qg}(z)=K^{\text{POS}}_{qg}(z)\,, \tag{7.82}$$

$$K^{\text{MPOS}}_{gq}(z)=K^{\text{POS}}_{gq}(z)\,, \tag{7.83}$$

$$K^{\text{MPOS}}_{gg}(z)=-2n_ff^{\text{MOM}}(z)K^{\text{POS}}_{qg}\Big|_{N=2}\,. \tag{7.84}$$

The MPOS scheme then automatically satisfies momentum conservation. Coefficient functions in the MPOS scheme are shown in figs. 7.1 to 7.3. It is clear that coefficient functions, and thus PDFs, remain positive in the MPOS scheme: indeed, the off-diagonal coefficient functions are unchanged, while the diagonal NLO contributions are modified by a small correction which is offset by the large positive LO contribution, and in fact in the hadronic case leaves the NLO correction positive for all z. Hence the MPOS and POS schemes have the same positivity properties. We will thus not discuss the MPOS scheme any further and restrict the discussion for simplicity to the POS scheme.

A further observation is that the POS scheme has been constructed in section 7.2.2 based on the kinematics of hadronic processes, namely by performing the collinear subtraction in off-diagonal channels at the scale μ_h^2, eq. (7.33). As discussed in section 7.2.2, if this scheme is used for the computation of electroproduction processes for which the relevant scale is μ_D^2, eq. (7.17), leads to coefficient functions, and consequently PDFs, that are with stronger reason positive. More in general, the POS scheme has been constructed using universal properties of the collinear emission that only depend on the LO splitting functions and the choice of scale, which is determined by the general kinematics of hadronic processes, but otherwise process-independent. However, the positivity argument presented in this section shows that this choice, whereas theoretically appealing,

is by no means necessary. In fact, any physical scheme choice of the form of eq. (7.69) can be used to construct a positive factorization scheme, by just picking a scheme choice such that the coefficient functions of the processes used to define the PDFs remain positive, and perturbative for all ξ. In any such scheme positivity of the PDFs holds. In fact, the simplest choice would be to pick as a positive factorization scheme the physical scheme itself, in which PDFs are positive by construction, as they are identified with physically observable cross sections.

The $\overline{MS}$ scheme

Having concluded that we can take the POS scheme as representative of a wide class of positive factorization schemes, we now discuss its relation to the $\overline{\text{MS}}$ scheme, and what it tells us about positivity of $\overline{\text{MS}}$ PDFs.

Inverting the scheme change from $\overline{\text{MS}}$ to POS perturbatively (cf. eq. (7.78)) we obtain

$$f^{\overline{\text{MS}}}(Q^2) = \left[\mathbb{I} - \frac{\alpha_s}{2\pi} K^{\text{POS}} \otimes\right] f^{\text{POS}}(Q^2)\,. \tag{7.85}$$

It is then clear that if the POS PDFs are positive, then so are the $\overline{\text{MS}}$ ones, because the matrix K^{POS} vanishes on the diagonal, and it has negative matrix elements off the diagonal, so $-K^{\text{POS}}$ in eq. (7.85) is positive. The perturbative inversion is justified due to the fact that the non-vanishing off-diagonal matrix elements of the K matrix are actually power-suppressed (i.e. next-to-eikonal) in the $z \to 1$ limit.

This can be seen more formally by considering the exact Mellin-space inverse of the scheme change matrix, eq. (7.78):

$$\left[\mathbb{I} + \frac{\alpha_s}{2\pi} K^{\text{POS}}(N)\right]^{-1} = \frac{1}{1 - \left(\frac{\alpha_s}{2\pi}\right)^2 K_{qg}(N) K_{gq}(N)} \left[\mathbb{I} - \frac{\alpha_s}{2\pi} K^{\text{POS}}(N)\right], \tag{7.86}$$

where $K_{ij}^{\text{POS}}(N)$ denote (by slight abuse of notation) the Mellin transforms of the matrix elements K_{ij}^{POS} of the matrix K^{POS}. It is easy to check that the factor $K_{qg}(N)K_{gq}(N)$ is a monotonically decreasing function of N along the real N axis, and in particular it vanishes as $\frac{1}{N^2}$ as $N \to \infty$, hence the prefactor which relates the exact and perturbative inversions, eqs. (7.85) and (7.86), is actually bounded in the region $N \gtrsim 2$ in which the $\overline{\text{MS}}$ coefficient functions, and thus the matrix elements of K, turn negative (cf. figs. 7.2 and 7.3).

We conclude that the light quark and gluon $\overline{\text{MS}}$ PDFs are in fact positive at NLO.

Heavy quarks require a separate discussion, because for heavy quarks $\overline{\text{MS}}$ factorization can be defined in a variety of ways (see e.g. Forte, Laenen, et al. 2010). Specifically, heavy quarks can be treated in a massive scheme, in which collinear singularities associated to them are regulated by their mass, so they decouple from perturbative evolution. In this scheme no collinear subtraction is performed for massive quarks, so their PDF is given by the unsubtracted eq. (7.2) and thus it remains a positive (and scale-independent) probability distribution to all perturbative orders. Note that nothing prevents this heavy quark PDF from having an

"intrinsic" component, of non-perturbative origin: however, in this factorization scheme, the heavy quark PDFs will be scale-independent, and thus positive at all scales.

However, it is also possible to treat the heavy quark in a massless $\overline{\text{MS}}$ scheme, in which the heavy quark is treated like other massless quarks, namely the collinear singularity regulated by its mass is subtracted according to eqs. (7.14) and (7.24), but with μ^2 now replaced by the heavy quark mass. Calculations performed in this scheme, with heavy quark mass effects neglected, are accurate for scales much larger than the quark mass. However, the massless scheme is in principle formally defined for all scales, including at the heavy quark mass. This is sometimes done by using the massless scheme for all flavors, but discontinuously changing the number of flavors at a matching scale chosen equal to (or of order of) the heavy quark mass (zero-mass variable-flavor number scheme, ZM-VFNS Aivazis et al. 1994). Below the matching scale the ZM-VFNS coincides with the massive scheme (with non-evolving heavy quark PDF), and at the matching scale the heavy quark PDF changes discontinuously: the matching condition is the scheme transformation from the massive to the massless $\overline{\text{MS}}$ (computed up to NNLO in Buza, Matiounine, Smith, and W. van Neerven 1998). This scheme transformation accounts for the fact that in the massive scheme the heavy quark decouples from the running, so loop corrections with the massive quark circulating in loops are included in the Wilson coefficient, and not in the operator matrix element, while in the massless scheme they are included in the operator normalization along with all other light quarks, but neglecting the quark mass when computing them.

When $Q^2 \sim m_h^2$ this neglect is not justified, and the corresponding scheme transformation may ruin positivity of the PDF. Specifically, it is often assumed that the massive-scheme PDF vanishes at some scale $Q^2 \sim m_h^2$, and it indeed appears reasonable to expect that the low-scale heavy quark scheme PDF if not vanishing, is rather smaller than light quark PDFs (cf. Ball, Bertone, Bonvini, Stefano Carrazza, et al. 2016; Ball, Bonvini, et al. 2015). However, if one determines the massless-scheme heavy quark PDF by starting with a vanishing massive-scheme PDFs, and using perturbative matching conditions, a negative result can be found – and is indeed found using standard light quark and gluon PDFs Ball et al. 2017b. This is now possible because the massless-scheme heavy quark PDF is not defined by a matrix element of the form of eq. (7.2), but rather, as the transformation of such an operator matrix element to a scheme in which the quark mass is neglected, but in a region in which the quark mass is not negligible. Of course, if $Q^2 \gg m_h^2$ the mass does become negligible, the previous arguments apply, and positivity of the heavy quark PDF is restored. Hence, positivity of the heavy quark PDF in the massless scheme only holds at high enough Q^2 that mass corrections are negligible.

All the discussion so far has been pursued at NLO. However, the main structure of the argument remains true to all perturbative orders. In particular, it is true to all orders that the diagonal splitting functions are negative at large z: in

fact, at large z to all perturbative orders they behave as $\frac{1}{(1-z)_+}$ Albino and Ball 2001. At higher perturbative orders, coefficient functions will contain plus distributions with higher order powers of $\ln(1-z)$, leading to the familiar rise in the partonic cross section which is predicted to all orders by threshold resummation S. Catani and Trentadue 1989; Sterman 1987. Off-diagonal channels, where negative contributions as $z \to 1$ may and indeed are expected to arise, remain power suppressed in this limit. It follows that the off-diagonal structure eq. (7.77) of the matrix relating a positive scheme to $\overline{\text{MS}}$ will hold true to all orders. The positivity argument of section 7.3.2 is a direct consequence of this structure, and it will thus also hold to all orders.

7.4 SUMMARY AND REMARKS

The goal of this work has been the construction of a universal factorization scheme in which PDFs are non-negative. In order to attack the problem, we started from the observation that $\overline{\text{MS}}$ partonic cross sections for typical electro- and hadro-production processes are not positive. This then implies that positivity of the PDFs is not guaranteed, since folding a negative partonic cross section with a positive PDF could lead to a negative physical cross section. We have then traced negative partonic cross sections to the way collinear subtraction is performed in $\overline{\text{MS}}$ and specifically we have shown that it is due to over-subtraction, related to the choice of subtraction scale, and also the treatment of the average over gluon polarizations in d dimensions. This loss of positivity only manifests itself in off-diagonal quark-gluon and gluon-quark channels.

A universal subtraction prescription which preserves positivity of the partonic cross section can then be constructed using hadronic kinematics, and shown to preserve positivity also in electroproduction kinematics. This prescription does not automatically respect momentum conservation, which however can be enforced with a soft modification of the subtraction procedure that does not affect its positivity properties. By performing collinear factorization in the standard approach of J. C. Collins and Soper 1982; Curci et al. 1980 it is then possible to show that positivity of the PDFs, defined as probability distributions, is preserved at all stages, so PDFs remain positive.

In fact, this positivity is a manifestation of the fact that PDFs can always be defined in terms of a physical process: what PDFs do is to allow one to express the perturbative QCD prediction for a process in terms of that for another process. The definition of the PDFs can then be process-independent (as in $\overline{\text{MS}}$) or process-dependent (as in so-called physical schemes Stefano Catani 1996; Diemoz et al. 1988). Its positivity will then be preserved provided only that the renormalization conditions, which fix the value of operator matrix elements that define the PDFs, preserves their interpretation as moments of a probability distribution. Effectively, this corresponds to choosing positive Wilson coefficients.

By considering a scheme in which PDFs are manifestly positive, and the transformation from it to $\overline{\text{MS}}$, we have finally shown that in the $\overline{\text{MS}}$ scheme PDFs

remain positive, despite the fact that off-diagonal partonic cross sections are negative. From a physical point of view, this is a consequence of the fact that the $\overline{\text{MS}}$ subtraction is actually strongly positive in the diagonal channels (where by "strongly" we mean that partonic functions tend to $+\infty$ towards kinematic boundaries). This then overwhelms the negative contribution from off-diagonal channels, while away from kinematic boundaries off-diagonal channels are perturbatively subleading.

Positivity of the PDFs is neither necessary nor sufficient for physical cross sections to be positive, as they ought to: it is not necessary, because it is possible that a negative PDF still leads to a positive hadronic cross section once folded with a suitable coefficient function, and it is not sufficient because in a scheme, such as $\overline{\text{MS}}$, in which some partonic cross sections are negative it could well be that, while the true PDF must necessarily lead to positive measurable cross sections, an incorrectly determined PDF could lead to a negative cross section despite being positive.

In other words, it is not necessarily true that the region in PDF space which is excluded by the requirement of positivity of the PDF is the same as that which is excluded by requiring positivity of the cross sections. However, from the point of view of PDFs determination, knowing that PDFs must be positive in a given factorization scheme does provide a useful constraint, in that it excludes a region which does not have to be explored, though this restriction is not necessarily the most stringent one. It is natural to ask whether the positivity requirement could be more restrictive in some factorization schemes than others, but it is unclear whether and how this question could be answered. The question of optimizing the scheme choice from the point of view of positivity constraints, for the sake of PDFs determination, remains open for future investigation.

8 NEW CANDIDATE METHODOLOGIES

As briefly described in the general introduction (cf. chapter 0) PDF fitting has always been a challenging task, and methodology is decisive in determining the final outcome, but highly non-trivial, driven several arbitrary choices.

Assessing the goodness of methodological choices, and its impact on the final result, is an important and serious topic. To this goal, several discussions among different PDF fitting groups have been devoted (e.g. Albert De Roeck 2009), resulting in publications that assess the status of PDF extraction Sergey Alekhin et al. 2011; Ball et al. 2022b; Michiel Botje et al. 2011; Rojo 2016.

The common ambition would be to pick a methodology that is not *adding information*, such that the resulting PDF is only determined by the data constraints, plus theoretical knowledge (such as sum rules). While this target is certainly desirable, it is not possible to fulfill it completely, because the theory defined object has simply too many degrees of freedom. Indeed, the undetermined PDF is the set of functions introduced in eq. (0.1), and the theoretical knowledge consists in a finite set of linear constraints. Thus, and infinite number of degrees of freedom remains unconstrained.

In order to make inference with a finite amount of data, further assumptions need to be used, to step from the infinite unconstrained directions, that would lead to infinite uncertainty, to a finite space in which optimizing the PDFs distribution for data compatibility with theory predictions. The more traditional approach consists in parametrizing the PDFs with some selected polynomials, fitting some exponents as well, for the behavior about the domain boundaries. This is definitely a sensible choice, and fully compatible with the principle that will be exposed in section 8.3, but with one major drawback: there is no specific reason to prefer a given polynomial basis, so this procedure creates room for some arbitrariness, leading to potential debate about an optimal choice.

In this context, the NNPDF Collaboration proposed an alternative PDF parametrization, based on a Neural Network (NN), that will then be trained with its intrinsic

training algorithm. This led to a series of challenges, that will be described in section 8.1, eventually conducing to release a completely analogue object, but with significant discrepant features, originated by the different methodology.

A number of issues however arose around NNPDF sets, some of them characteristic of the NN determined PDFs, while other shared with other determinations. Some of them will be described in section 8.2.

While attempting to improve the current methodology to address these concerns, the inspection suggested that another paradigm shift could give an easier and more complete answer, keeping the benefit of all or most of the present developments, just minimally (as possible) replacing the fitting *"engine"*. This new proposal is the result of the active discussion on improvements inside the collaboration, but also arising from external proposals, and it will be presented in section 8.3.

8.1 THE NNPDF METHODOLOGY

The methodology adopted by the NNPDF Collaboration has already been described in its publications, Ball, Del Debbio, Forte, Guffanti, Latorre, Piccione, et al. 2009; Forte, Garrido, et al. 2002, and in dedicated reviews, Ethier and Nocera 2020; Forte and Stefano Carrazza 2020. Therefore, it would be redundant to add too many details here, that can be easily found in references, but it is relevant to summarize the main points, for the subsequent discussion.

The primary challenge in switching from polynomial parametrization[1] to NN consists in the uncertainty propagation. Indeed, the first step would be to establish a training algorithm, but, while the fixed parametrizations require an explicit choice, the NN comes with its own efficient training algorithm. It is in no way unique, and there are many options available, but there is no need for a dedicated development. The exact algorithm selection becomes part of the methodology, and it is once more a source of arbitrariness, but there are suitable strategies to educate this choice, and sometimes it is also driven by consistent technical considerations.

Before explaining the uncertainty propagation technique it is worth noticing that the most renowned and usual applications of NNs aim to determine an unknown function, but the details of the learnt function are not particularly relevant, while it is relevant to evaluate it on a specific subset of inputs, settling the *task* to be performed by the trained network. In the case of the PDFs, all values of the functions are relevant, since a large enough set of physical observables is potentially sensible to all its analytical features. Moreover, being able to manually examining the function can help in inferring specific properties of the hadrons. So, the final step of an NN fit is to evaluate the function on a sufficiently complete set of inputs, originating a function set by means of interpolation.

[1] Or parametrization over a fixed basis of functions, more in general.

In this context, fixed parametrization fits are no different, because after determining the parameters in a Hessian fit, they also evaluate the resulting functions. To propagate the uncertainty, an Hessian fit require to determine the main eigenvectors for the minimized quantity, generating a Gaussian distribution that approximate the uncertainty on the fitted parameters in a neighborhood of the best fit.

For an NN fit this is not technically available, because of the large number of parameters involved. And even when it would be possible, it is not recommended to follow the classical approach, since many parameters might be poorly constrained, but with a negligible impact on the value of the PDFs. Therefore, NNPDF proposed a more direct approach: since the distribution in the value of the PDF is derived from the data distribution, it is possible to start from an alternative representation of this distribution. As illustrated in fig. 8.1, the data are then fluctuated, according to their distribution, and many samples of the whole dataset are taken, each one containing one and only one value for each experimental data point (i.e. each sample is a single extraction from the joint distribution of all experimental data). This samples are called *data replicas*. After that, one NN is fitted to each replicate, and the resulting set of trained NNs, called *NN replicas* are gathered, and they are the determined MC representation of the PDFs distribution. In particular, this procedure associates to each point in the PDFs domain a set of values (the NN replicas), that should be interpreted as a sample of the PDFs distribution for that point.

As thoroughly discussed in Del Debbio, Giani, et al. 2022, this boils down to solve the *inverse problem* for the PDFs: the map from PDF space to data space is known, and it consists in the theory prediction. So the fit is essentially inverting this map, constrained with some assumptions. Applying the inverse[2] to each data replica the distribution is propagated from one space to the other.

CLOSURE TESTS However, the training algorithm for the NN is sufficiently sophisticated to introduce a further level of indirectness, that makes harder to understand which features of the input dataset is causing a specific behavior in the output distribution. Because of this partial loss of explainability (part of it is already in the size of the dataset, and the lack of a one-to-one mapping caused by convolutions), the methodology requires to be validated in a more systematic way. NNPDF is doing this employing *closure tests*, explained in details in Del Debbio, Giani, et al. 2022. Basically, a set of fake data is generated by a fake underlying *truth*, chosen to be a reasonable PDF set (usually a set by a different collaboration) applying the same theory predictions used in the fit, and the final result is compared to the starting point.

This exercise is possible at many levels:

[2]The inverse map is not a unique one, since the NN may obtain different minima according to its initialization. If this is done randomly, the distribution of the initialization assigns a certain probability to each possible minimum, creating a probabilistic inverse, that will convolute the data distribution in the result. This becomes clearer in the Bayesian framework in section 8.3, where this further probability can be essentially identified with the prior.

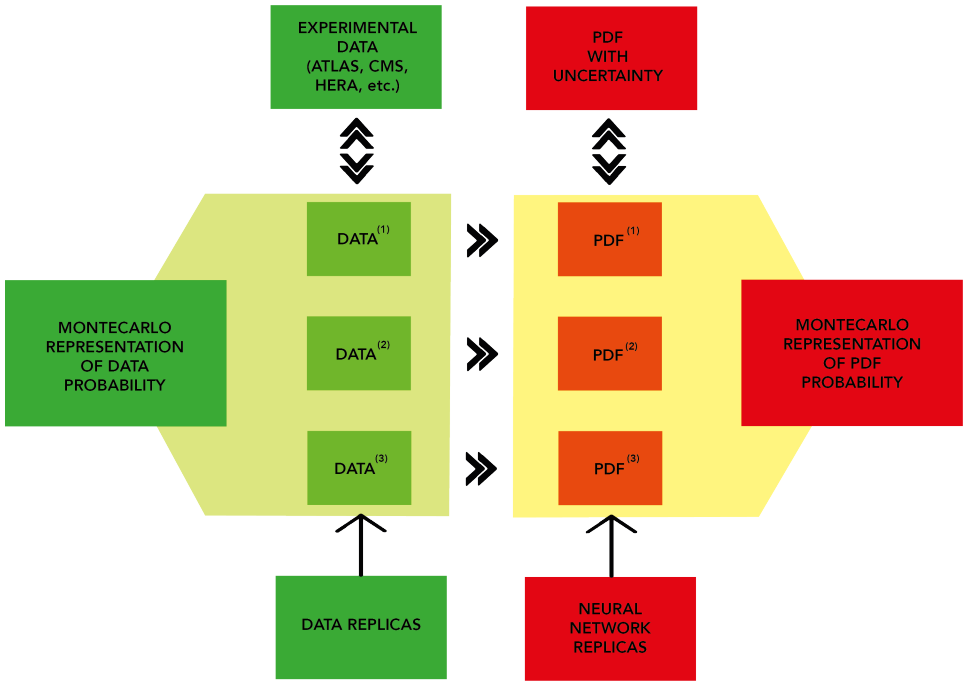

Figure 8.1: NNPDF uncertainty propagation methodology, one of the most important key points of NNPDF methodology. Picture available on the collaboration website http://nnpdf.mi.infn.it/research/general-strategy/.

level 0 using directly the mock truth, or

level 1 fluctuating it, to emulate how the experimental uncertainty is affecting the estimate of central values, or

level 2 fluctuating once more, to obtain the same data replicas used in a regular fit

Closure tests have a great power in assessing the faithfulness of the uncertainties established by a given methodology. It is important to remark that what is tested is not if there would exist an alternative methodology that could led to smaller uncertainties, but only if the chosen one produces an estimate of the uncertainty that is compatible with the expected shift of the central value from the underlying truth. Then, two very different methodologies, one obtaining much bigger errors than the other from the same dataset, can both pass the closure test, if their error is consistent with the actual shift from the truth: big uncertainty associated to a big shift, and conversely.

Nevertheless, also closure tests have some limitations, since they test the methodology with the assumption of perfectly consistent data, and accepting the perturbative theory predictions, truncated at a fixed order, as a prompt for the full theory, generating fake data for that PDF. These two elements introduce an ambiguity in the outcome of the closure tests themselves, that is possible to alleviate studying how it behaves when specific inconsistencies are injected (this is actually a work currently in progress, being performed by NNPDF members).

8.1.1 Generalization

As in other classical machine learning problems, also for PDFs there is a fundamental concern, that goes beyond the loss function minimization: how much does the predictions extend, beyond the region strictly covered by data?

The first instance of this issue is interpolation itself, since an extremely aggressive optimization can generate an incredibly noisy result, matching anyhow all the points in the dataset. This is partially prevented by the structure of the network itself, and its training algorithm: in the case of NNPDF there is no explicit penalty term imposing smoothness in the loss function. The rest is done by the *training-validation* split. As it is customary for many machine learning applications, data are split in the two sets, and only one of the two is passed to the optimizer, while the other is observed during the training, in order to interrupt the process before learning the noise, and prevent *over-fitting*. Despite being very common, there is not a full analytical understanding for this method, or for the NN training more in general, and many counter-intuitive are still being investigated (e.g. double descent, or multiple in general). So, the final and most reliable proof of good interpolation properties comes from the closure tests.

A second instance lies in the extrapolation region. Since this region is not directly controlled by data, or at least very little, it is largely undetermined, but it is important for PDFs uncertainties to be reliable also in this region, usually

heavily affecting BSM searches. Being reliable means to be large enough to cover unobserved features of the PDFs, without completely forgetting the few constraints coming from the sum rules. To check the fidelity of these portions of the final product, a different kind of tests is employed: the dataset is chronologically segmented, and they are incrementally included in the input of the NN, finally comparing the various results. This type of checks have been called *future tests*, and the essential idea consist to check how faithful are the extrapolation uncertainties, since the extrapolation region of a chronologically prior fit will intersect the data region of a later one. In practice, despite NNPDF4.0 being significantly smaller than those in the other PDF set, this is mostly limited to the data region, and the increased flexibility prevents an extreme extrapolation, decoupling the two regions, as much as it is allowed by sum rules. This is shown and exemplified quite well in the Drell–Yan forward-backward asymmetry study, presented in chapter 6.

It is worth to remark once more that there are many possible variations, whose results can span wide space of solutions. In order to avoid arbitrary choices, inspired by subjective opinions or *human experience* (potentially reliable, but certainly hard to quantify and systematically improve), in NNPDF4.0 a large enough space of hyperparameters has been considered, and they have been hyperoptimized (also standard practice, though computationally expensive) based on a grid search. This also contribute to improve the generalization power of the full procedure, since also the methodology parameters have been systematically optimized for this task.

8.1.2 Minor improvements

While it would be desirable that most of the decisions were inspired only by physical or statistical reasons, a common enough pattern is that technical limitations imposed further constraints.

An example of this is the choice of the training algorithm: NNPDF4.0 adopted the Stochastic Gradient Descent (SGD), that incredibly boosted performances with respect to the former Nodal Genetic Algorithm (NGA) used by NNPDF3.1. Both the algorithm are very well-known in literature, with SGD being based on the classical idea of Gradient Descent (GD), applied in optimization tasks since two centuries. Nevertheless, the choice here is mainly driven by its performances, improved by automatic differentiation techniques (implemented in modern frameworks). This allowed further studies, like hyperoptimization cited above, that would have not been possible with the former NGA implementation, only because of performances.

I proposed a couple of minor improvements in this sense, that is worth to briefly report.

NEGATIVE SAMPLING Data replicas generation involves sampling data from the joint distribution of measurements, assuming this to be normally distributed. However, for most of the observables this assumption can only hold approxi-

mately, since they are known to be positive semi-definite. In order to prevent negative values, the current algorithm operates a cut, implementing by redrawing the full sample if a value, supposed to be positive, has actually been drawn negative. Usually this is not causing any issue, since most experimental measurements are rather precise, and thus incompatible with zero. But it is still prone to terrible performances. Indeed, if for any reason there were a fixed rate r for a set of k measurements to be negative, the probability for all of them to be positive would be $(1-r)^k$, and then a number of extractions of order:

$$\frac{1}{(1-r)^k} \tag{8.1}$$

would be required on average to obtain a fully positive sample, so they increase exponentially with the number of points k. If r starts approaching $\mathcal{O}(1/2)$, a huge amount of draws would be required also for a few points. E.g. for $r = 1/2$ and $k = 10$ one thousand extractions would be performed on average, for the full sample.

Following, a list of possible solutions, with related advantages:

DECOUPLE Since the main problem consists the performance, a simple solution is just to redraw only the samples that happened to be negative, in order to prevent the exponential scaling

Slightly more challenging at a technical level, since the code is based on NumPy, and Python iterations should be avoided as much as possible. But it is sufficient to redraw arrays with a length equal to the amount of negative samples, and apply with a mask.

↬ This modifies the distribution to be a truncated Gaussian (obviously rescaled for normalization).

WALL Another option is just to cut all the negative values, and just set them to zero, this limits the procedure to a single draw.

↬ The resulting distribution is the truncated Gaussian, retaining the original normalization, with a discrete point, 0, having a finite weight (the integrated value of the negative tail).

REFLECTION

Keeping the single draw, but avoiding the attribution of finite probabilities to a single point, is possible, e.g. just taking the absolute value of the drawn sample.

↬ In this case the distribution would be the sum of the truncated positive Gaussian, and the mirrored negative tail.

This has been experimented, but it produces a distribution that is significantly different from the original one, leading to unexpected features in the fit.

DISTORTION

To minimally transform the original distribution, it is possible to consider a rescaling that is preserving the right tail of the Gaussian (larger values than average), while smoothly distorting the left one. Shifting the distribution to be centered in 0, $x \to y = x - \bar{x}$, this means the left tail should not exceed a minimal threshold value $\hat{y} = -\bar{x}$. There are many functions that can do this, one possible solution is an Exponential Linear Unit (ELU):

$$\tilde{y} = \begin{cases} y & \text{if } y > 0, \\ \hat{y}\,(\exp(a \cdot y) - 1) & \text{otherwise} \end{cases} \tag{8.2}$$

Finally yielding a value for the transformed x equal to $\tilde{x} = \tilde{y} + \bar{x}$. The value of the parameter a can be set for example to $1/\hat{y}$, such that the derivative is also continuous, arguably the minimal distortion for this functional form. Being the transformation simple and analytical, performances would not be impacted. Other transformations with the same properties would be completely equivalent.

↬ The distribution would be the one described above: a perfect right Gaussian tail, and smoothly warped left one, vanishing for negative values.

NO BOUND Finally, there are applications for which not any of this compromises is required, since it is possible to treat also negative values for the observables. In this case, retaining them has the advantage of keeping a cleaner distribution, avoid more complex procedures, and achieving optimal generation performances the same.

THEORY COVARIANCE MATRIX CONSTRUCTION Another performance issue encountered in the current implementation of NNPDF methodology concerns the construction of the theory covariance matrix. The problem also here consists in extra unneeded iterations. The really superfluous part can just be dropped, while further optimizations are possible if a fully factorized prescription is adopted, in place of the current sliced one (cf. section 4.4). Since this is rather technical, with no impact on anything but performances (as opposed to the former one, also affecting significantly the distribution of a few points), no further details will be provided here, and the interested reader can directly read the documentation of the new proposed implementation:

AleCandido/thcovmat

8.2 NEURAL NETWORKS' PUZZLES

A number of concerns arose during the development of NNPDF, partially from other PDF groups and external parties, the rest from inside the collaboration itself.

A first point is still related to the partial loss of explainability due to the NN, already addressed by closure tests and hyperoptimization. Unfortunately, all the techniques adopted until now always relies on the χ^2 of some subsets of data, that might be a poor quantifier of over-fitting and other possible issues. So, an open question has been posed to the collaboration, whether it is possible to quantify this properties with an alternative metric, that could also be used in the methodology hyperoptimization. A basic idea would be to rely on the arc-length of the PDF replica: if the PDF oscillates more (contains more wiggles) its arc-length is increased, thus minimizing would yield less wiggly PDFs, privileging smoother replicas. While on one side this is already done by the NN itself, on the other it might be worrying to explicitly penalize the arc-length, since the PDF should always be able to reproduce physical oscillations. Some metrics definitions are currently being discussed, based on the statistical properties of replicas ensembles, or refining the idea of detecting wild oscillations, e.g. measuring the local oscillation rate, to detect anomalous regions (this local quantifier has been called *kinetic energy*, for the similarly with the expression of this quantity for a moving particle).

Another direction for improvements lies in the uncertainty propagation itself. At the moment, each replica is determined on its own, irrespectively of the other replicas, and only gathering all of them together at the end will generate the final PDFs distribution. Since the object sought is the distribution, there might be a more direct way to extract it, and once the full distribution (or a part of it) is available during training, more features would be available to the algorithm to construct a better optimization path.

Another subject is the treatment of extrapolation. While NNs are extremely good at interpolating, the extrapolation assumptions remain hidden, and it is more difficult to directly assess the goodness of the extrapolation, and recognize a fit with spurious extrapolation. This is connected to the general lack of analytical insights, but this condition is not intrinsic to the more problem, but rather to the adopted solution. Start thinking this way, it is possible and appropriate to consider if there are more suitable techniques that can make use of the original fully analytical formulation of the problem (some well-behaving functions over a simple real domain), that instead is traditionally lacking in most machine learning applications.

8.3 BAYESIAN PDFs

Starting from the concerns exposed in the previous section, the need for some changes in the methodology became manifest. There is no clear indication that all of the possible issues enumerated might be solved keep using a NN, but especially there is no special need not to consider any other alternative.

One of the points was if there is a way to determine the distribution, rather than the individual elements of the samples, one by one. Following this path, one can start considering having a NN that fits a batch of replicas together, but how

many? All replicas share the same theory, through the FK tables (cf. chapter 3), that already limits the resources required, but fitting the whole set at the same time is prohibitive. Moreover, there was the need of a more direct and insightful approach. Taking mainly into account these two points, an alternative solution start looking promising: why not to directly apply *Bayesian inference* to the PDF posterior determination?

Bayesian inference is sufficiently well-known, and already applied in a series of contexts, including HEP theory (e.g. see Cacciari and Houdeau 2011, as described in chapter 4). The basic essence lies in the so-called Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{8.3}$$

In the context of inference, the Bayes theorem is usually written:

$$P(\theta|\{x\}) = \frac{P(\{x\}|\theta)\, P(\theta)}{\int d\theta\; P(\{x\}|\theta)\, P(\theta)} \tag{8.4}$$

Differences in the notation are minimal, and the integration in the denominator essentially correspond to $P(\{x\})$. But the interpretation is fundamental:

data $\{x\}$ is a set of observations

parameter θ is an unknown parameter, that enters the distribution of the observations, and can thus be inferred from them

prior $P(\theta)$ is the prior knowledge about the parameter θ (also "the *prior*"), before any observation is considered

likelihood $P(\{x\}|\theta)$ is the likelihood of the data observed, given the value of the parameter θ

posterior $P(\theta|\{x\})$ is the probability distribution of the parameter θ, after the observations $\{x\}$ are taken into account

evidence is the normalization of the right-hand side, i.e. the denominator

So, even before the experiment is performed, the observer has a prior knowledge about the parameter, coming from its model or external sources. The effect of the experiment is to update this prior knowledge with the observed results.

There is an extremely important feature of this formula, that applies extremely well to the PDFs extraction task: in order to derive the posterior probability, no inverse map is required: to know the value of the posterior for a given PDF, it is sufficient to be able to evaluate the prior, usually analytic and simple enough to evaluate, the likelihood, that involves the actual connection from PDFs to data space, but only the forward map, and the evidence (usually the complex part, since it involves an integration on a wide space). No minimization is required to determine the full distribution, so no effective inverse map has to be constructed.

It is not the first time that someone attempts to use Bayesian inference for fitting PDFs, Aggarwal et al. 2022; Gbedo and Mangin-Brinet 2017, but there two main novelties that we want to propose:

PARAMETRIZATION

what the other references do is to apply Bayesian inference, in place of an Hessian fit, but retaining a fixed parametrization.

Coming from the experience of NNPDF, it became evident how a sufficiently flexible parametrization is appropriate for the task. But at this point, PDFs are also distributed in a standard format (`LHAPDF`, Buckley et al. 2015), in which a series of points are stored and later interpolated. This approach is successful enough that the distributed grids represent extremely well the fitted object, and essentially loose no feature.

Because of these two considerations, there is an obvious candidate for the parametrization: we want to determine exactly the values of the PDFs that are going to be distributed, i.e. the PDFs evaluated on the points of the chosen grids. Any further value would be lost, and thus provide no information to the user.

DATA SET

all the previous attempts were done at the level of a proof of concept Gbedo and Mangin-Brinet 2017, or to investigate the features of a specific data set Aggarwal et al. 2022

We want to attempt a global fit in which the full data set of NNPDF would be taken into account, to have a meaningful comparison with existing PDF sets.

While it seems too ambitious for a new methodology to aim at the widest data set available immediately (and intermediate steps will be done anyhow), the target is definitely accessible, this because the NNPDF framework provide already all the elements required for a full fit, and only the underlying *engine* has to be switched. In this sense, the road for this study has already been paved by `n3fit` Stefano Carrazza and Cruz-Martinez 2019, that replaced the former NNPDF NN, plugging an isolated inference module in the rest of the framework.

The most expensive part in Gaussian inference is computing the *evidence*: it usually involves one or more integration, and it is possible to perform it analytically only in a handful of cases. Because of this, suitable numerical methods have been developed, in order to make Bayesian inference practically possible for a wide variety of cases. One of the most popular algorithm is Markov Chain Monte Carlo (MCMC), that allows to sample a generic distribution without computing its normalization, since it is based on probability ratios. Many variants of MCMC are known, one of the most popular being the Metropolis–Hastings algorithm Hastings 1970; Metropolis et al. 1953, and more recently Hamiltonian Monte Carlo (HMC) Duane et al. 1987 (also known as *hybrid Monte Carlo*, as it is called in the original paper).

This is what has been used in the past attempts, and what we eventually aim to use as well. However, MCMC are delicate tools, and a series of caveats have

to be considered when analyzing the results, including the role of a first thermalization phase, and the autocorrelations of the chains (new algorithms partially alleviate the main issues, but convergence is still fragile for complex problems). We will then inspect produce some first iterations by means of approximate methods, that are strictly exact for a wide portion of the data set, and are much less computationally demanding, see next section 8.3.1 for some further details.

One clear advantage over fitting replicas is in the algorithm simplicity, and we expect it will eventually drive better performances. Indeed, in NNPDF methodology drawing one more sample requires fitting one NN more, while in MCMC is just an effective step (possibly consisting of multiple actual ones, because of autocorrelations).

Another advantage consists in interpretability, and this happens on many levels. While NN added a level of indirectness, in the case of the Bayesian fit everything is straightforward, since there is no complex architecture or training algorithm involved; the ingredients are the likelihood and PDFs prior, and the composition is simply the normalized product. In order to compare the weight attributed to a PDF candidate, the only operation required is to evaluate the prior on it. This is a criticism that NNPDF had to address in practice, Courtoy et al. 2022: given a set of candidates, constructed with whatever procedure, explain the reason why they are considered unlikely by the chosen methodology, since they are such in the posterior, but still compatible with data and theoretical constraints. What happens is that the NN architecture and initialization is also implementing an *effective* prior, that makes some candidates more unlikely than others, before having seen the observations. But this effective prior is practically impossible to evaluate, so other proxies have to be constructed in order to extract this information from the network. Nothing like that is required if the prior is known analytically, since it can be explicitly motivated and evaluated for comparison.

Moreover, machine learning is best suited to those tasks that is difficult to express analytically, but this is not the case for PDFs: their definitions and theoretical properties are known formulas, so it is possible to study them, and possibly implement, at an analytic level.

One point that is often considered cumbersome in Bayesian inference is the *prior choice*, because it might introduce strong assumptions. As motivated above, assumptions in the case of PDFs are a strong requirement, so explicit *prior* choice is an advantage, to clarify which features in the result are inherited from it, and which are instead produced by data. The former example of direct prior probing to motivate a low weight for apparently reasonable candidates is an instance of this.

There are two classes of assumptions that we want to explicitly include in the prior: theoretical knowledge and smoothness. The first are exact properties of the PDFs predicted by QCD, mainly sum rules dictating the global number of some quark species, or imposing total momentum conservation. The second is a reasonable assumption from an Occam razor-like approach: we want to first resolve fluctuations over large scales, so we deweight high-frequency oscillatory modes. A suitable prior family to encode these assumptions are *Gaussian processes*.

Gaussian processes are an extremely wide and powerful family, also proven to be equivalent to infinitely wide NN. They are characterized by a mean function over their domain $\mathcal{D}$, and a kernel function, associated to the covariance of the process $\mathcal{G}$:

$$\mathrm{E}[\mathcal{G}(x)] = \mu(x) \qquad x \in \mathcal{D} \tag{8.5}$$

$$\mathrm{Cov}[\mathcal{G}(x), \mathcal{G}(y)] = k(x, y) \qquad (x, y) \in \mathcal{D}^2 \tag{8.6}$$

Their defining property is that the marginal distribution over any finite subset of variables is a multi-Gaussian, and consequently the mean and covariance specify the whole distribution. There are many important and useful properties of Gaussian processes, those that are relevant will be described in the related publication, Petrillo 2022.

Just a couple of features that are worth noticing explicitly. Gaussian processes are convenient over continuous domain, since analytic manipulation are possible, and during inference data and constraints can be added on its derivatives as well. Exploiting this, sum rules, that are actually constraints on the primitive, can be imposed analytically in the prior. Another handy property is that, as well as interpolation, extrapolation behavior is determined by the kernel function, and in particular its characteristic extrapolation length. This gives a direct handle to control extrapolation, that will follow quite straightforward from the injected prior knowledge.

Finally, a last part of the methodology applied in NNPDF fits was the mentioned hyper-optimization. In the language of Bayesian determination the hyper-parameters correspond to prior's parameters. This can also be optimized, but a more consistent way of treating them is available: it is possible to obtain a joint posterior distribution $\mathrm{P}(\theta; \alpha|\{x\}))$ over both inferred parameters θ and hyper-parameters α. Once this function is extracted, hyper-optimization would consist in taking the value of α maximizing its marginal posterior. This is called the Maximum A Posteriori (MAP) estimate of α. Though possible, it is not required, and the joint distribution contains more complete information than the output of hyper-optimization itself.

8.3.1 Approximate inference with `lsqfitgp`

Following, the description of the current attempt to achieve approximate Bayesian PDFs is presented. Everything included in this section, and part of the previous one, is actually a summary of the work-in-progress draft, available online Petrillo 2022.

The core of a Gaussian Process regression is rather simple: given some (x_i, f_i) data pairs, and some x'_j points of interest, the posterior distributions for the x'_j set is obtained *conditioning* the joint prior multi-Gaussian on the observed values (basically slicing the distribution in the i dimensions at the f_i values, and normalizing the result). This procedure is exact, and only requires some linear algebra to determine the posterior average and covariance, since the posterior as well will be a Gaussian distribution.

Applying linear transformation changes the Gaussian parameters, but preserves Gaussianity. Thus the fit would be exact also in the case of observations living in a different space, but with a linear map connecting the two spaces. This is exactly the case of DIS data, since a single PDF is involved, and is mapped to data space through convolution, i.e. a linear operation. Sum rules are also linear in the PDFs, and imposed on the primitive. So, a full DIS-only fit, including sum rules, can be performed just:

1. picking a suitable kernel function (defining the prior)
2. conditioning the process on observations
3. evaluating on point of interests[3]

as already explained in Del Debbio, Giani, et al. 2022.

There are two categories of data for which it is not possible to apply the exact algorithm: quadratic and compound data. The first category corresponds to data collected in double hadronic collisions (cf. section 0.2). Since two PDFs are involved, the PDF space and data space are not connected by linear transformations. The same is also true for those data, including DIS ones, that are obtained by applying some operations on the elementary cross-sections, e.g. taking the ratio of two corresponding types of events. A solution that allows for these data is a standard least squares fit, in which the problem is solved iteratively, roughly: linearizing locally, determining the linear solution, and then repeating from there.

The computationally complexity of the fit is dominated by the matrix inversion, required to evaluate the new mean and covariance. Matrix inversion basic complexity is $\mathcal{O}\left(n^3\right)$, for a generic $n \times n$ matrix. Being an extremely widespread task, there are optimized algorithms achieving better performances, but the best improvements are obtained restricting to specific classes of matrices. While this approach might seem quite restrictive, it should be noted that the matrices involved are coming from the evaluation of kernel functions, and they usually involve multiple properties. At the very least they are symmetric and positive semi-definite.

The program implementing the linear and approximate solution is not exclusively developed for this work, but is a generic library to implement least squares fits based on Gaussian processes. Its source is openly available:

`Gattocrucco/lsqfitgp`

and the documentation, user manual, and several examples can be found at:

`https://gattocrucco.github.io/lsqfitgp/docs/`

8.3.2 Status of the project

At the moment of writing, the status of the project is rather inhomogeneous: while the core machinery in `lsqfitgp` is mostly available, and rather advanced,

[3] In practice, this operation is included in the former one.

the general fit is still at the level of a proof of concept, since still being developed on fake data. On the other hand, data and theory predictions are available to be consumed, even though some final improvements are still required. Final evolution and grid generation will be performed in the same way of the NNPDF current fit.

The whole project is public, and available at:

NNPDF/mcpdf

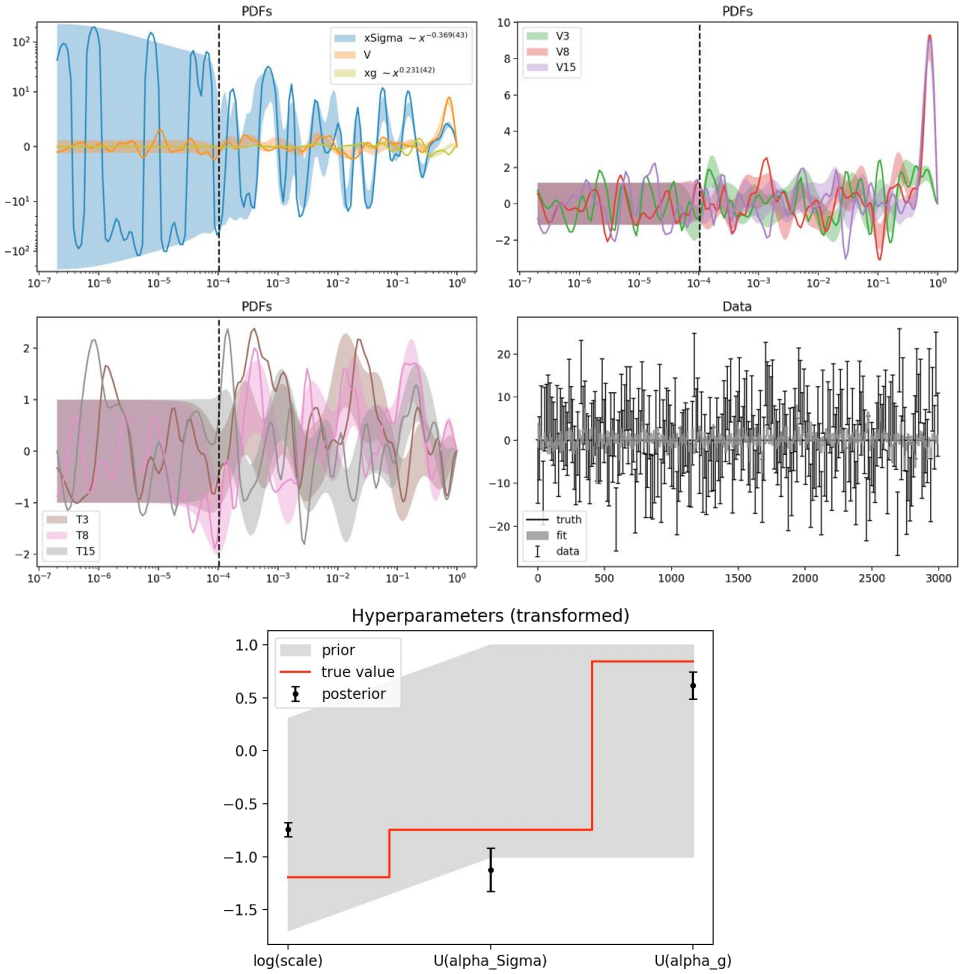

Figure 8.2: Current `lsqfitgp` results on fake data. The lines are the underlying (fake) PDFs used to generate the fake data, while the bands are the $\pm 1\sigma$ intervals of the posterior. The vertical dashed lines mark the start of the x values that are linked to the data. The lower plot shows the inference over the hyperparameters, with the gray band representing the $\pm 1\sigma$ intervals of the prior.

CONCLUSIONS

The material presented cover a range of topics, all related to Parton Distribution Functions, but about different aspects. My PhD itself has been mainly focused on the rework of the theory predictions architecture, in collaboration with other NNPDF members, the number of which increased during the years, together with projects' ambitions: the initial goal of replacing the DIS and DGLAP evolution modules was pursued by me and Felix Hekhorn, but it grew to include a full rework of the theory pipeline, together with other collaborators.

The modular structure of the architecture we created (the backbone of which we inherited from `APFELcomb` and previous NNPDF projects) has been designed to achieve three goals: extensibility, maintainability, and external access. Indeed, while designed and developed by NNPDF members, we hope that the other groups interested in PDF fitting, and related topics, can make use of the tools we developed. The development itself happened completely in public, and still continues this way. Until now, it has not been a big deal: projects were formally public, but not being production-ready little interactions took place in practice. However, we considered important to manifest from the beginning the will to write tools at the disposal of the entire community, accepting feedback, feature requests, and any kind of contributions (within the scope of the projects). This is motivated by the non-negligible expense in the amount of manpower required, and the intention to lower the barrier for further studies, giving free and easy access to common machinery.

We already benefited ourselves from the design choices made: the various projects are the base on which new features are being introduced, as described for the study of Missing Higher Order Uncertainties impact on PDFs in chapter 4, and additional progresses as well, even if they have not been accounted for in this thesis, like the introduction of consistent N^3LO theory predictions and other elements already available in existing software (QED corrections, polarized elements or FF-based predictions, small-x resummed predictions), but all provided by a single framework. The individual modules are already being used in further studies, as it happened for `EKO` in the quest for intrinsic charm in chapter 5, or PineAPPL for Drell–Yan A_{fb} exercise of chapter 6, and also `yadism`, in the determination of low-energy neutrino structure functions (still in preparation, and not described here). As we are doing it, we hope that external users of our modules could the same as well. Also PDF-wise, the new methodology proposal in chapter 8 is an example of how having lowered the barrier through task decoupling allows for more iterations on new ideas, and the consequential innovation.

Any new perturbative order is a completely new challenge, but this translates in more complex calculations, heavier math in the results, and more sophisticated approximations required for them (to keep them fast enough to evaluate), while

the rest of the infrastructure is ready to expand for new contributions. In particular, while N^{3}LO inclusion has already been mentioned as a work-in-progress effort, one of PineAPPL initial motivations has been to account for EW corrections, and this will also be possible with current tools. At the same time, also new data and new processes are becoming available, as well as alternative data sources, like lattice computations, but the advantage of the current approach is that we do not need any major modification of the framework for them, but they will plugged as additional modules.

A specific effort in building the new architecture is being dedicated to reproducibility, not only in making *possible* to reobtain the same result (thus reconstructing and scrutinizing all the choices made), but also making it *simple*, as much as possible, to encourage other people to do it. Running a PDF fit, i.e. a fit accounting for higher orders QCD effects on a global data set, has always been a complex process, but it should become easier over the years, despite more ingredients and further theory being added to the present status, to encourage people to focus on the new developments, instead of the complication of gathering all the pieces from scratch.

It is relevant to note that this last goal is not uniquely pursued by our framework, since it exists at least the remarkable example of `xFitter` https://www.xfitter.org/, with a similar target, but a different perspective, as it privileged an integrated (i.e. monolithic) architecture, distributing theory predictions tools and the fitting machinery altogether. Instead, we want to allow the user (possibly including `xFitter`) to compute its own theory predictions, and use them with their favorite fitting methodology.

All these tools are empowering PDF fitters, in order to let them directing efforts to the final purpose of more accurate and precise determinations, hopefully leading to a fruitful collaboration to reduce the related theoretical uncertainty still very relevant in many hadronic observables.

GLOSSARIES

ACRONYMS

ALICE CMS collaboration (A Large Ion Collider Experiment) 119
API Application Programming Interface 33
ATLAS ATLAS collaboration (A Toroidal LHC Apparatus) 5, 98, 99, 106, 108, 117, 122, 210

BCDMS Bologna-Cern-Dubna-Munich-Saclay 4, 28
BSM Beyond the Standard Model 1, 121, 122, 127, 128, 135, 143, 150, 151, 156, 190

CC Charged Current 13– 15, 17, 22, 28
CHORUS CERN Hybrid Oscillation Research ApparatUS 4, 28
CKM Cabibbo–Kobayashi–Maskawa 22
CL Confidence Level 91, 134
CMS CMS collaboration (Compact Muon Solenoid) 5, 98, 99, 106, 108, 117, 122, 210
CoM Center of Mass 124, 134
CS Collins–Soper 122, 124, 126– 129, 148, 149, 211

DGLAP Dokshitzer–Gribov–Lipatov–Altarelli–Parisi evolution equations vii, 5– 7, 23, 24, 31– 33, 35, 39, 41, 47, 50, 53, 57, 58, 60, 63, 64, 70, 71, 201
DIS Deep Inelastic Scattering vii, 2– 4, 11– 16, 28, 32, 38, 39, 64, 67, 71, 72, 98, 101, 102, 106, 108, 155, 165, 198, 201, 206, 210
DY Drell–Yan iv, 106, 121– 124, 130, 134, 135, 142, 144, 148– 151, 155, 201, 211– 213

EC Electromagnetic Current 15, 16
EFT Effective Field Theory 122
EIC Electron-Ion Collider 119
EKO Evolution Kernel Operators 7, 32
ELU Exponential Linear Unit 192
EM Electromagnetic 12, 17
EMC European Muon Collaboration 94, 95, 101, 102, 119, 209, 210
EW Electroweak 1, 3– 5, 12, 13, 15– 17, 22, 122, 123, 202, 206

FCNC Flavor Changing Neutral Currents 15
FF Fragmentation Functions 2, 201
FFNS Fixed Flavor Number Scheme 22, 24, 25, 27, 37– 39, 49, 53, 54

FNS Flavor Number Scheme iv, 11, 16, 22, 24, 25, 64, 89–92, 94–114, 209–211, 215
FONLL Fixed Order Next-to-Leading Log 17, 23, 25, 28, 98
FPF Forward Physics Facility 119
FTDY Fixed Target Drell–Yan 4

GD Gradient Descent 190
GM-VFNS General Mass Variable Flavor Number Scheme 17, 18, 23, 24, 27, 28
GPD Generalized Parton Distributions 2

HEP High Energy Physics vii, 1, 3, 194
HERA Hadron-Elektron-Ringanlage 4, 28, 98, 101, 106, 108, 210
HL high-luminosity 119, 121
HMC Hamiltonian Monte Carlo 195

LHC Large Hadron Collider iv, 3–6, 93, 99, 101, 106, 119, 121, 122, 143–145, 147, 148, 150, 151, 206, 213
LHCb LHCb collaboration 89, 93–95, 98, 99, 101, 102, 106, 114–117, 119, 136, 209, 210, 216
LO Leading Order iv, 3, 5, 6, 12, 13, 32, 39, 42, 44, 45, 47, 48, 51, 64, 72, 121–131, 142, 145, 156, 206, 207, 211, 213

MAP Maximum A Posteriori 197
MC Monte Carlo 4, 66, 70, 96, 97, 187
MCMC Markov Chain Monte Carlo 195, 196
MHOU Missing Higher Order Uncertainties vii, 17, 28, 34, 38, 44, 55, 63, 64, 67, 69, 72, 92–94, 100, 106, 108, 110–112, 201, 208–211
ML Machine Learning 90

N^{3}LO Next-to-Next-to-Next-to-Leading Order 32, 34, 38, 45, 54, 60, 90–92, 98–100, 106, 201, 202, 209, 210
NC Neutral Current 13–15, 17, 22, 28, 121–123
NGA Nodal Genetic Algorithm 190
NLO Next-to-Leading Order 24, 32, 44, 45, 47, 48, 51, 58, 64, 67, 69, 70, 72, 93, 99, 100, 114, 123, 129, 130, 143, 144, 155, 207, 208, 211, 213
NMC New Muon Collaboration 4
NN Neural Network viii, 185–187, 189, 190, 193, 195–197
NNLO Next-to-Next-to-Leading Order 4, 28, 32, 34, 37, 39–41, 44–48, 51, 58, 63, 67, 70, 72, 90–92, 96, 98–100, 106, 115, 122, 129, 133, 134, 138–140, 143, 144, 206, 207, 210–213
NNPDF NNPDF Collaboration vii–ix, 58, 60, 66, 67, 69, 70, 72, 155, 185–189, 192, 195–197, 199, 201, 214, 242

OME Operator Matrix Element 37

PDF Parton Distribution Functions iv, v, vii, viii, 2–7, 11, 16, 18, 19, 22, 24, 26–28, 31–33, 35–46, 49–51, 53–55, 57, 58, 60, 61, 63, 65–67, 69–72, 89–106, 108–119, 121–123, 125–151, 155, 156, 185–187, 189, 190, 192–199, 201, 202, 206–213, 215, 216
PID Particle IDentifier 5
pQCD perturbative Quantum Chromodynamics 17
pQFT perturbative Quantum Field Theory 2, 63
PS Parton Shower 3

QCD Quantum Chromodynamics 1–4, 6, 13, 32, 33, 36, 38–41, 44, 53–55, 57, 58, 61, 72, 90, 93, 96–99, 114, 122, 129, 130, 196, 202, 206, 207, 211
QED Quantum Electrodynamics 39, 55, 201
QFT Quantum Field Theory 1

RGE Renormalization Group Equations 35, 38
RSL Regular-Singular-Local 19

SGD Stochastic Gradient Descent 190
SLAC Stanford Linear Accelerator Center 4
SM Standard Model 1, 2, 15, 92, 98, 122, 156

TMC Target Mass Corrections 11
TMD Transverse Momentum Distributions 2

VFNS Variable Flavor Number Scheme 37, 39, 49, 50, 53–55

YM Yang–Mills 1

ZM-VFNS Zero-Mass Variable Flavor Number Scheme 23–25, 27

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

Aaboud, M. et al.

2017 "Measurement of the Drell-Yan triple-differential cross section in pp collisions at $\sqrt{s} = 8$ TeV", *JHEP*, 12, p. 059, DOI: 10.1007/JHEP12(2017)059, arXiv: 1710.05167 [hep-ex].

Aad, Georges et al.

2012 "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett. B*, 716, pp. 1-29, DOI: 10.1016/j.physletb.2012.08.020, arXiv: 1207.7214 [hep-ex].

2014 "Search for contact interactions and large extra dimensions in the dilepton channel using proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector", *Eur. Phys. J. C*, 74, 12, p. 3134, DOI: 10.1140/epjc/s10052-014-3134-6, arXiv: 1407.2410 [hep-ex].

2019 "Search for high-mass dilepton resonances using 139 fb^{-1} of pp collision data collected at $\sqrt{s}$ =13 TeV with the ATLAS detector", *Phys. Lett. B*, 796, pp. 68-87, DOI: 10.1016/j.physletb.2019.07.016, arXiv: 1903.06248 [hep-ex].

2020 "Search for new non-resonant phenomena in high-mass dilepton final states with the ATLAS detector", *JHEP*, 11, [Erratum: JHEP 04, 142 (2021)], p. 005, DOI: 10.1007/JHEP11(2020)005, arXiv: 2006.12946 [hep-ex].

2021 "Search for New Phenomena in Final States with Two Leptons and One or No b-Tagged Jets at $\sqrt{s} = 13$ TeV Using the ATLAS Detector", *Phys. Rev. Lett.*, 127, 14, p. 141801, DOI: 10.1103/PhysRevLett.127.141801, arXiv: 2105.13847 [hep-ex].

Aaij, Roel et al.

2019 "First Measurement of Charm Production in its Fixed-Target Configuration at the LHC", *Phys. Rev. Lett.*, 122, 13, p. 132002, DOI: 10.1103/PhysRevLett.122.132002, arXiv: 1810.07907 [hep-ex].

2021 "Study of Z bosons produced in association with charm in the forward region" [Sept. 2021], arXiv: 2109.08084 [hep-ex].

Abdul Khalek, Rabah et al.

2019a "A first determination of parton distributions with theoretical uncertainties", *Eur. Phys. J.*, C, 79:838, DOI: 10.1140/epjc/s10052-019-7364-5, arXiv: 1905.04311 [hep-ph].

Abdul Khalek, Rabah et al.

2019b "Parton Distributions with Theory Uncertainties: General Formalism and First Phenomenological Studies", *Eur. Phys. J. C*, 79, 11, p. 931, DOI: 10.1140/epjc/s10052-019-7401-4, arXiv: 1906.10698 [hep-ph].

2019c "Parton Distributions with Theory Uncertainties: General Formalism and First Phenomenological Studies", *Eur. Phys. J. C*, 79, 11, p. 931, DOI: 10.1140/epjc/s10052-019-7401-4, arXiv: 1906.10698 [hep-ph].

Abdul Khalek, Rabah, Shaun Bailey, Jun Gao, Lucian Harland-Lang, and Juan Rojo

2018 "Towards Ultimate Parton Distributions at the High-Luminosity LHC", *Eur. Phys. J. C*, 78, 11, p. 962, DOI: 10.1140/epjc/s10052-018-6448-y, arXiv: 1810.03639 [hep-ph].

Abe, F. et al.

1996 "Inclusive jet cross section in $\bar{p}p$ collisions at $\sqrt{s} = 1.8$ TeV", *Phys. Rev. Lett.*, 77, pp. 438-443, DOI: 10.1103/PhysRevLett.77.438, arXiv: hep-ex/9601008.

Ablinger, J., A. Behring, J. Blümlein, A. De Freitas, A. Hasselhuhn, A. von Manteuffel, M. Round, C. Schneider, and F. Wißbrock

2014 "The 3-Loop Non-Singlet Heavy Flavor Contributions and Anomalous Dimensions for the Structure Function $F_2(x, Q^2)$ and Transversity", *Nucl. Phys. B*, 886, pp. 733-823, DOI: 10.1016/j.nuclphysb.2014.07.010, arXiv: 1406.4654 [hep-ph].

Ablinger, J., A. Behring, J. Blümlein, A. De Freitas, A. von Manteuffel, et al.

2014 "The 3-Loop Pure Singlet Heavy Flavor Contributions to the Structure Function $F_2(x, Q^2)$ and the Anomalous Dimension", arXiv: 1409.1135 [hep-ph].

Ablinger, J., A. Behring, J. Blümlein, A. De Freitas, A. von Manteuffel, and C. Schneider

2015 "The 3-loop pure singlet heavy flavor contributions to the structure function F2(x,Q2) and the anomalous dimension", *Nuclear Physics B*, 890 [Jan. 2015], pp. 48-151, ISSN: 0550-3213, DOI: 10.1016/j.nuclphysb.2014.10.008.

Ablinger, J., J. Blumlein, S. Klein, C. Schneider, and F. Wissbrock

2011 "The $O(\alpha_s^3)$ Massive Operator Matrix Elements of $O(n_f)$ for the Structure Function $F_2(x, Q^2)$ and Transversity", *Nucl. Phys. B*, 844, pp. 26-54, DOI: 10.1016/j.nuclphysb.2010.10.021, arXiv: 1008.3347 [hep-ph].

Ablinger, J., J. Blümlein, A. De Freitas, A. Hasselhuhn, A. von Manteuffel, M. Round, and C. Schneider

2014 "The $O(\alpha_s^3 T_F^2)$ Contributions to the Gluonic Operator Matrix Element", *Nucl. Phys. B*, 885, pp. 280-317, DOI: 10.1016/j.nuclphysb.2014.05.028, arXiv: 1405.4259 [hep-ph].

Ablinger, J., J. Blümlein, A. De Freitas, A. Hasselhuhn, A. von Manteuffel, M. Round, C. Schneider, and F. Wißbrock

2014 "The transition matrix element $A_{gq}(N)$ of the Variable Flavor Number Scheme at $O(\alpha_s^3)$", *Nuclear Physics B*, 882 [May 2014], pp. 263-288, ISSN: 0550-3213, DOI: 10.1016/j.nuclphysb.2014.02.007.

Accomando, Elena et al.

2019 "PDF Profiling Using the Forward-Backward Asymmetry in Neutral Current Drell-Yan Production", *JHEP*, 10, p. 176, DOI: 10.1007/JHEP10(2019)176, arXiv: 1907.07727 [hep-ph].

Accomando, Elena, Juri Fiaschi, Francesco Hautmann, and Stefano Moretti

2018 "Neutral current forward–backward asymmetry: from θ_W to PDF determinations", *Eur. Phys. J. C*, 78, 8, [Erratum: Eur.Phys.J.C 79, 453 (2019)], p. 663, DOI: 10.1140/epjc/s10052-018-6120-6, arXiv: 1805.09239 [hep-ph].

Aggarwal, Ritu, Michiel Botje, Allen Caldwell, Francesca Capel, and Oliver Schulz

2022 "New constraints on the up-quark valence distribution in the proton" [Sept. 2022], arXiv: 2209.06571 [hep-ph].

Aivazis, M.A.G., John C. Collins, Fredrick I. Olness, and Wu-Ki Tung

1994 "Leptoproduction of heavy quarks. 2. A Unified QCD formulation of charged and neutral current processes from fixed target to collider energies", *Phys. Rev. D*, 50, pp. 3102-3118, DOI: 10.1103/PhysRevD.50.3102, arXiv: hep-ph/9312319.

Albino, Simon and Richard D. Ball

2001 "Soft resummation of quark anomalous dimensions and coefficient functions in MS-bar factorization", *Phys. Lett.*, B513, pp. 93-102, DOI: 10.1016/S0370-2693(01)00742-0, arXiv: hep-ph/0011133 [hep-ph].

Alekhin, S. and S. Moch

2011 "Heavy-quark deep-inelastic scattering with a running mass", *Phys. Lett. B*, 699, pp. 345-353, DOI: 10.1016/j.physletb.2011.04.026, arXiv: 1011.5790 [hep-ph].

Alekhin, Sergey et al.

2011 "The PDF4LHC Working Group Interim Report", arXiv: 1101.0536 [hep-ph].

Alioli, Simone, Paolo Nason, Carlo Oleari, and Emanuele Re

2010 "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX", *JHEP*, 1006, p. 043, DOI: 10.1007/JHEP06(2010)043, arXiv: 1002.2581 [hep-ph].

Almasy, A. A., S. Moch, and A. Vogt

2012 "On the Next-to-Next-to-Leading Order Evolution of Flavour-Singlet Fragmentation Functions", *Nucl. Phys. B*, 854, pp. 133-152, DOI: 10.1016/j.nuclphysb.2011.08.028, arXiv: 1107.2263 [hep-ph].

Altarelli, Guido, Stefano Forte, and Giovanni Ridolfi

1998 "On positivity of parton distributions", *Nucl. Phys.*, B534, pp. 277-296, DOI: 10.1016/S0550-3213(98)00661-0, arXiv: hep-ph/9806345.

Altarelli, Guido and G. Parisi

1977 "Asymptotic Freedom in Parton Language", *Nucl. Phys.*, B126, pp. 298-318, DOI: 10.1016/0550-3213(77)90384-4.

Alwall, J., R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, et al.

2014 "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *JHEP*, 1407, p. 079, DOI: 10.1007/JHEP07(2014)079, arXiv: 1405.0301 [hep-ph].

Amoroso, S. et al.

2022 "Snowmass 2021 whitepaper: Proton structure at the precision frontier" [Mar. 2022], arXiv: 2203.13923 [hep-ph].

Anastasiou, Charalampos, Lance J. Dixon, Kirill Melnikov, and Frank Petriello

2004 "High precision QCD at hadron colliders: Electroweak gauge boson rapidity distributions at NNLO", *Phys. Rev. D*, 69, p. 094008, DOI: 10.1103/PhysRevD.69.094008, arXiv: hep-ph/0312266.

Anchordoqui, Luis A. et al.

2021 "The Forward Physics Facility: Sites, Experiments, and Physics Potential" [Sept. 2021], arXiv: 2109.10905 [hep-ph].

Armadillo, Tommaso, Roberto Bonciani, Simone Devoto, Narayan Rana, and Alessandro Vicini

2022 "Two-loop mixed QCD-EW corrections to neutral current Drell-Yan", *JHEP*, 05, p. 072, DOI: 10.1007/JHEP05(2022)072, arXiv: 2201.01754 [hep-ph].

Aubert, J. J. et al.

1983 "Production of charmed particles in 250-GeV μ^+ - iron interactions", *Nucl. Phys.*, B213, pp. 31-64, DOI: 10.1016/0550-3213(83)90174-8.

Azzi, P. et al.

2019 "Standard Model Physics at the HL-LHC and HE-LHC", arXiv: 1902.04070 [hep-ph].

Baikov, P. A., K. G. Chetyrkin, and J. H. Kühn

2017 "Five-Loop Running of the QCD coupling constant", *Phys. Rev. Lett.*, 118, 8, p. 082002, DOI: 10.1103/PhysRevLett.118.082002, arXiv: 1606.08659 [hep-ph].

Bailey, S., T. Cridge, L. A. Harland-Lang, A. D. Martin, and R. S. Thorne

2021 "Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs", *Eur. Phys. J. C*, 81, 4, p. 341, DOI: 10.1140/epjc/s10052-021-09057-0, arXiv: 2012.04684 [hep-ph].

Ball, Richard D. et al.

2011 "Reweighting NNPDFs: the W lepton asymmetry", *Nucl. Phys.*, B849, pp. 112-143, DOI: 10.1016/j.nuclphysb.2011.03.017, arXiv: 1012.0836 [hep-ph].

2015 "Parton distributions for the LHC Run II", *JHEP*, 04, p. 040, DOI: 10.1007/JHEP04(2015)040, arXiv: 1410.8849 [hep-ph].

2017a "Parton distributions from high-precision collider data", *Eur. Phys. J. C*, 77, 10, p. 663, DOI: 10.1140/epjc/s10052-017-5199-5, arXiv: 1706.00428 [hep-ph].

2017b "Parton distributions from high-precision collider data", *Eur. Phys. J.*, C77, 10, p. 663, DOI: 10.1140/epjc/s10052-017-5199-5, arXiv: 1706.00428 [hep-ph].

2021a "An open-source machine learning framework for global analyses of parton distributions", *Eur. Phys. J. C*, 81, 10, p. 958, DOI: 10.1140/epjc/s10052-021-09747-9, arXiv: 2109.02671 [hep-ph].

2021b "The Path to Proton Structure at One-Percent Accuracy" [Sept. 2021], arXiv: 2109.02653 [hep-ph].

2022a "The path to proton structure at 1% accuracy", *Eur. Phys. J. C*, 82, 5, p. 428, DOI: 10.1140/epjc/s10052-022-10328-7, arXiv: 2109.02653 [hep-ph].

2022b "The PDF4LHC21 combination of global PDF fits for the LHC Run III", *J. Phys. G*, 49, 8, p. 080501, DOI: 10.1088/1361-6471/ac7216, arXiv: 2203.05506 [hep-ph].

Ball, Richard D., Valerio Bertone, Marco Bonvini, Stefano Carrazza, Stefano Forte, Alberto Guffanti, Nathan P. Hartland, Juan Rojo, and Luca Rottoli

2016 "A Determination of the Charm Content of the Proton", *Eur. Phys. J. C*, 76, 11, p. 647, DOI: 10.1140/epjc/s10052-016-4469-y, arXiv: 1605.06515 [hep-ph].

Ball, Richard D., Valerio Bertone, Marco Bonvini, Stefano Forte, Patrick Groth Merrild, Juan Rojo, and Luca Rottoli

2016 "Intrinsic charm in a matched general-mass scheme", *Phys. Lett.*, B754, pp. 49-58, DOI: 10.1016/j.physletb.2015.12.077, arXiv: 1510.00009 [hep-ph].

Ball, Richard D., Valerio Bertone, Francesco Cerutti, Luigi Del Debbio, Stefano Forte, et al.

2012 "Reweighting and Unweighting of Parton Distributions and the LHC W lepton asymmetry data", *Nucl.Phys.*, B855, pp. 608-638, DOI: 10.1016/j.nuclphysb.2011.10.018, arXiv: 1108.1758 [hep-ph].

Ball, Richard D., Marco Bonvini, and Luca Rottoli

2015 "Charm in Deep-Inelastic Scattering", *JHEP*, 11, p. 122, DOI: 10.1007/JHEP11(2015)122, arXiv: 1510.02491 [hep-ph].

Ball, Richard D., Alessandro Candido, Juan Cruz-Martinez, Stefano Forte, Tommaso Giani, Felix Hekhorn, Kirill Kudashkin, Giacomo Magni, and Juan Rojo

2022 "Evidence for intrinsic charm quarks in the proton", *Nature*, 608, 7923, pp. 483-487, DOI: 10.1038/s41586-022-04998-2, arXiv: 2208.08372 [hep-ph].

Ball, Richard D., Alessandro Candido, Stefano Forte, Felix Hekhorn, Emanuele R. Nocera, Juan Rojo, and Christopher Schwan

2022 "Parton Distributions and New Physics Searches: the Drell-Yan Forward-Backward Asymmetry as a Case Study" [Sept. 2022], arXiv: 2209.08115 [hep-ph].

Ball, Richard D., Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Jose I. Latorre, Andrea Piccione, Juan Rojo, and Maria Ubiali

2009 "A Determination of parton distributions with faithful uncertainty estimation", *Nucl. Phys. B*, 809, [Erratum: Nucl.Phys.B 816, 293 (2009)], pp. 1-63, DOI: 10.1016/j.nuclphysb.2008.09.037, arXiv: 0808.1231 [hep-ph].

Ball, Richard D., Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Jose I. Latorre, Juan Rojo, and Maria Ubiali

2010 "A first unbiased global NLO determination of parton distributions and their uncertainties", *Nucl. Phys. B*, 838, pp. 136-206, DOI: 10.1016/j.nuclphysb.2010.05.008, arXiv: 1002.4407 [hep-ph].

Ball, Richard D., Emanuele R. Nocera, and Juan Rojo

2016 "The asymptotic behaviour of parton distributions at small and large x", *Eur. Phys. J.*, C76, 7, p. 383, DOI: 10.1140/epjc/s10052-016-4240-4, arXiv: 1604.00024 [hep-ph].

Barontini, Andrea, Alessandro Candido, Juan Cruz-Martinez, Felix Hekhorn, Christopher Schwan, and Giacomo Magni

2022 "Theory pipeline for PDF fitting", *PoS*, ICHEP2022, p. 784, DOI: 10.22323/1.414.0784, arXiv: 2211.10447 [hep-ph].

Bauer, Christian W., Zoltan Ligeti, Michael Luke, Aneesh V. Manohar, and Michael Trott

2004 "Global analysis of inclusive B decays", *Phys. Rev. D*, 70, p. 094017, DOI: 10.1103/PhysRevD.70.094017, arXiv: hep-ph/0408002.

Beenakker, Wim, Christoph Borschensky, Michael Krämer, Anna Kulesza, Eric Laenen, Simone Marzani, and Juan Rojo

2016 "NLO+NLL squark and gluino production cross-sections with threshold-improved parton distributions", *Eur. Phys. J.*, C76, 2, p. 53, DOI: 10.1140/epjc/s10052-016-3892-4, arXiv: 1510.00375 [hep-ph].

Behring, A., I. Bierenbaum, J. Blümlein, A. De Freitas, S. Klein, and F. Wißbrock

2014 "The logarithmic contributions to the $O(\alpha_s^3)$ asymptotic massive Wilson coefficients and operator matrix elements in deeply inelastic scattering", *Eur. Phys. J. C*, 74, 9, p. 3033, DOI: 10.1140/epjc/s10052-014-3033-x, arXiv: 1403.6356 [hep-ph].

Bellm, Johannes et al.

2016 "Herwig 7.0/Herwig++ 3.0 release note", *Eur. Phys. J. C*, 76, 4, p. 196, DOI: 10.1140/epjc/s10052-016-4018-8, arXiv: 1512.01178 [hep-ph].

Bertone, Valerio, Stefano Carrazza, and Nathan Hartland

n.d. "https://github.com/NNPDF/apfelcomb", Generates FK tables for NNPDF fits.

Bertone, Valerio, Stefano Carrazza, and Nathan P. Hartland

2017 "APFELgrid: a high performance tool for parton density determinations", *Comput. Phys. Commun.*, 212, pp. 205-209, DOI: 10.1016/j.cpc.2016.10.006, arXiv: 1605.02070 [hep-ph].

Bertone, Valerio, Stefano Carrazza, Nathan P. Hartland, and Juan Rojo

2018 "Illuminating the photon content of the proton within a global PDF analysis", *SciPost Phys.*, 5, 1, p. 008, DOI: 10.21468/SciPostPhys.5.1.008, arXiv: 1712.07053 [hep-ph].

Bertone, Valerio, Stefano Carrazza, and Juan Rojo

2014 "APFEL: A PDF Evolution Library with QED corrections", *Comput. Phys. Commun.*, 185, pp. 1647-1668, DOI: 10.1016/j.cpc.2014.03.007, arXiv: 1310.1394 [hep-ph].

Bertone, Valerio, Rikkert Frederix, Stefano Frixione, Juan Rojo, and Mark Sutton

2014 "aMCfast: automation of fast NLO computations for PDF fits", *JHEP*, 08, p. 166, DOI: 10.1007/JHEP08(2014)166, arXiv: 1406.7693 [hep-ph].

Bierenbaum, Isabella, Johannes Blumlein, and Sebastian Klein

2009a "Mellin Moments of the O(alpha**3(s)) Heavy Flavor Contributions to unpolarized Deep-Inelastic Scattering at Q**2 >> m**2 and Anomalous Dimensions", *Nucl. Phys. B*, 820, pp. 417-482, DOI: 10.1016/j.nuclphysb.2009.06.005, arXiv: 0904.3563 [hep-ph].

2009b "The Gluonic Operator Matrix Elements at O(alpha(s)**2) for DIS Heavy Flavor Production", *Phys. Lett. B*, 672, pp. 401-406, DOI: 10.1016/j.physletb.2009.01.057, arXiv: 0901.0669 [hep-ph].

Bierlich, Christian et al.

2022 "A comprehensive guide to the physics and usage of PYTHIA 8.3" [Mar. 2022], arXiv: 2203.11601 [hep-ph].

Bjorken, J. D.

1967 "CURRENT ALGEBRA AT SMALL DISTANCES", *Conf. Proc. C*, 670717, pp. 55-81.

Blümlein, J., P. Marquard, C. Schneider, and K. Schönwald

2021 "The three-loop unpolarized and polarized non-singlet anomalous dimensions from off shell operator matrix elements", *Nucl. Phys. B*, 971, p. 115542, DOI: 10.1016/j.nuclphysb.2021.115542, arXiv: 2107.06267 [hep-ph].

2022 "The three-loop polarized singlet anomalous dimensions from off-shell operator matrix elements", *JHEP*, 01, p. 193, DOI: 10.1007/JHEP01(2022)193, arXiv: 2111.12401 [hep-ph].

Blümlein, Johannes, Jakob Ablinger, Arnd Behring, Abilio De Freitas, Andreas von Manteuffel, Carsten Schneider, and C. Schneider

2017 "Heavy Flavor Wilson Coefficients in Deep-Inelastic Scattering: Recent Results", *PoS*, QCDEV2017, p. 031, DOI: 10.22323/1.308.0031, arXiv: 1711.07957 [hep-ph].

Boettcher, Tom, Philip Ilten, and Mike Williams

2016 "Direct probe of the intrinsic charm content of the proton", *Phys. Rev. D*, 93, 7, p. 074008, DOI: 10.1103/PhysRevD.93.074008, arXiv: 1512.06666 [hep-ph].

Bonciani, Roberto, Federico Buccioni, Narayan Rana, and Alessandro Vicini

2020 "Next-to-Next-to-Leading Order Mixed QCD-Electroweak Corrections to on-Shell Z Production", *Phys. Rev. Lett.*, 125, 23, p. 232004, DOI: 10.1103/PhysRevLett.125.232004, arXiv: 2007.06518 [hep-ph].

Bonciani, Roberto, Luca Buonocore, Massimiliano Grazzini, Stefan Kallweit, Narayan Rana, Francesco Tramontano, and Alessandro Vicini

2022 "Mixed Strong-Electroweak Corrections to the Drell-Yan Process", *Phys. Rev. Lett.*, 128, 1, p. 012002, DOI: 10.1103/PhysRevLett.128.012002, arXiv: 2106.11953 [hep-ph].

Bonvini, Marco

2012 *Resummation of soft and hard gluon radiation in perturbative QCD*, PhD thesis, Genoa U., arXiv: 1212.0480 [hep-ph].

2020 "Probabilistic definition of the perturbative theoretical uncertainty from missing higher orders", *Eur. Phys. J. C*, 80, 10, p. 989, DOI: 10.1140/epjc/s10052-020-08545-z, arXiv: 2006.16293 [hep-ph].

Bonvini, Marco, Stefano Forte, and Giovanni Ridolfi

2012 "The Threshold region for Higgs production in gluon fusion", *Phys. Rev. Lett.*, 109, p. 102002, DOI: 10.1103/PhysRevLett.109.102002, arXiv: 1204.5473 [hep-ph].

Botje, M.

2011 "QCDNUM: Fast QCD Evolution and Convolution", *Comput.Phys.Commun.*, 182, pp. 490-532, DOI: 10.1016/j.cpc.2010.10.020, arXiv: 1005.1481 [hep-ph].

Botje, Michiel et al.

2011 "The PDF4LHC Working Group Interim Recommendations", arXiv: 1101.0538 [hep-ph].

Britzger, D. et al.

2022 "NNLO interpolation grids for jet production at the LHC" [July 2022], arXiv: 2207.13735 [hep-ph].

Britzger, Daniel, Klaus Rabbertz, Fred Stober, and Markus Wobisch

2012 "New features in version 2 of the fastNLO project", in *20th International Workshop on Deep-Inelastic Scattering and Related Subjects*, pp. 217-221, DOI: 10.3204/DESY-PROC-2012-02/165, arXiv: 1208.3641 [hep-ph].

Brodsky, S. J., P. Hoyer, C. Peterson, and N. Sakai

1980 "The Intrinsic Charm of the Proton", *Phys. Lett.*, B93, pp. 451-455, DOI: 10.1016/0370-2693(80)90364-0.

Brodsky, S. J., A. Kusina, F. Lyonnet, I. Schienbein, H. Spiesberger, and R. Vogt

2015 "A review of the intrinsic heavy quark content of the nucleon", *Adv. High Energy Phys.*, 2015, p. 231547, DOI: 10.1155/2015/231547, arXiv: 1504.06287 [hep-ph].

Brodsky, Stanley J., John C. Collins, Stephen D. Ellis, John F. Gunion, and Alfred H. Mueller

1984 "Intrinsic Chevrolets at the SSC", in *1984 DPF Summer Study on the Design and Utilization of the Superconducting Super Collider (SSC) (Snowmass 84)*, p. 227.

Brodsky, Stanley J. and Glennys R. Farrar

1973 "Scaling Laws at Large Transverse Momentum", *Phys. Rev. Lett.*, 31, pp. 1153-1156, DOI: 10.1103/PhysRevLett.31.1153.

Brodsky, Stanley J. and Glennys R. Farrar

1975 "Scaling Laws for Large Momentum Transfer Processes", *Phys. Rev. D*, 11, p. 1309, DOI: 10.1103/PhysRevD.11.1309.

Buccioni, Federico, Fabrizio Caola, Herschel A. Chawdhry, Federica Devoto, Matthias Heller, Andreas von Manteuffel, Kirill Melnikov, Raoul Röntsch, and Chiara Signorile-Signorile

2022 "Mixed QCD-electroweak corrections to dilepton production at the LHC in the high invariant mass region", *JHEP*, 06, p. 022, DOI: 10.1007/JHEP06(2022)022, arXiv: 2203.11237 [hep-ph].

Buccioni, Federico, Fabrizio Caola, Maximilian Delto, Matthieu Jaquier, Kirill Melnikov, and Raoul Röntsch

2020 "Mixed QCD-electroweak corrections to on-shell Z production at the LHC", *Phys. Lett. B*, 811, p. 135969, DOI: 10.1016/j.physletb.2020.135969, arXiv: 2005.10221 [hep-ph].

Buckley, Andy et al.

2011 "General-purpose event generators for LHC physics", *Phys. Rept.*, 504, pp. 145-233, DOI: 10.1016/j.physrep.2011.03.005, arXiv: 1101.2599 [hep-ph].

Buckley, Andy, James Ferrando, Stephen Lloyd, Karl Nordström, Ben Page, Martin Rüfenacht, Marek Schönherr, and Graeme Watt

2015 "LHAPDF6: parton density access in the LHC precision era", *Eur. Phys. J. C*, 75, p. 132, DOI: 10.1140/epjc/s10052-015-3318-8, arXiv: 1412.7420 [hep-ph].

Buza, M., Y. Matiounine, J. Smith, and W. L. van Neerven

1998a "Charm electroproduction viewed in the variable-flavour number scheme versus fixed-order perturbation theory", *The European Physical Journal C*, 1, 1-2 [Mar. 1998], pp. 301-320, ISSN: 1434-6052, DOI: 10.1007/bf01245820.

1998b "Charm electroproduction viewed in the variable-flavour number scheme versus fixed-order perturbation theory", *Eur. Phys. J.*, C1, pp. 301-320, eprint: hep-ph/9612398.

1998 "Charm electroproduction viewed in the variable flavor number scheme versus fixed order perturbation theory", *Eur. Phys. J. C*, 1, pp. 301-320, DOI: 10.1007/BF01245820, arXiv: hep-ph/9612398.

Cacciari, Matteo and Nicolas Houdeau

2011 "Meaningful characterisation of perturbative theoretical uncertainties", *JHEP*, 1109, p. 039, DOI: 10.1007/JHEP09(2011)039, arXiv: 1105.5152 [hep-ph].

Cacciari, Matteo, Gavin P. Salam, and Gregory Soyez

2008 "The Anti-k(t) jet clustering algorithm", *JHEP*, 0804, p. 063, DOI: 10.1088/1126-6708/2008/04/063, arXiv: 0802.1189 [hep-ph].

Callan Curtis G., Jr. and David J. Gross
1969 "High-energy electroproduction and the constitution of the electric current", *Phys. Rev. Lett.*, 22, pp. 156-159, DOI: 10.1103/PhysRevLett.22.156.

Campbell, John, Joey Huston, and Frank Krauss
2017 *The Black Book of Quantum Chromodynamics: A Primer for the LHC Era*, Oxford University Press, ISBN: 978-0-19-965274-7.

Candido, Alessandro, Stefano Forte, and Felix Hekhorn
2020 "Can $\overline{\text{MS}}$ parton distributions be negative?", *JHEP*, 11, p. 129, DOI: 10.1007/JHEP11(2020)129, arXiv: 2006.07377 [hep-ph].

Candido, Alessandro, Felix Hekhorn, et al.
n.d. *yadism: Yet Another DIS module*, in preparation.

Candido, Alessandro, Felix Hekhorn, and Giacomo Magni
2022a "EKO: Evolution Kernel Operators" [Feb. 2022], arXiv: 2202.02338 [hep-ph].
2022b *N3PDF/yadism: FONLL-B*, version v0.11.0, DOI: 10.5281/zenodo.6285149.

Carli, Tancredi, Dan Clements, Amanda Cooper-Sarkar, Claire Gwenlan, Gavin P. Salam, Frank Siegert, Pavel Starovoitov, and Mark Sutton
2010 "A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project", *Eur. Phys. J. C*, 66, pp. 503-524, DOI: 10.1140/epjc/s10052-010-1255-0, arXiv: 0911.2985 [hep-ph].

Carrazza, S., E. R. Nocera, C. Schwan, and M. Zaro
2020a "PineAPPL: combining EW and QCD corrections for fast evaluation of LHC processes", *JHEP*, 12, p. 108, DOI: 10.1007/JHEP12(2020)108, arXiv: 2008.12789 [hep-ph].
2020b "PineAPPL: combining EW and QCD corrections for fast evaluation of LHC processes", *Journal of High Energy Physics*, 2020, 12 [Dec. 2020], ISSN: 1029-8479, DOI: 10.1007/jhep12(2020)108.

Carrazza, Stefano and Juan Cruz-Martinez
2019 "Towards a new generation of parton densities with deep learning models", *Eur. Phys. J. C*, 79, 8, p. 676, DOI: 10.1140/epjc/s10052-019-7197-2, arXiv: 1907.05075 [hep-ph].

Catani, S. and L. Trentadue
1989 "Resummation of the QCD Perturbative Series for Hard Processes", *Nucl. Phys.*, B327, pp. 323-352, DOI: 10.1016/0550-3213(89)90273-3.

Catani, Stefano
1996 "Comment on quarks and gluons at small x and the SDIS factorization scheme", *Z. Phys. C*, 70, pp. 263-272, DOI: 10.1007/s002880050104, arXiv: hep-ph/9506357.

Chatrchyan, Serguei et al.

2012 "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC", *Phys. Lett. B*, 716, pp. 30-61, DOI: 10.1016/j.physletb.2012.08.021, arXiv: 1207.7235 [hep-ex].

Chetyrkin, K. G., G. Falcioni, F. Herzog, and J. A. M. Vermaseren

2017 "Five-loop renormalisation of QCD in covariant gauges", *JHEP*, 10, [Addendum: JHEP 12, 006 (2017)], p. 179, DOI: 10.1007/JHEP10(2017)179, arXiv: 1709.08541 [hep-ph].

Chetyrkin, K.G., Johann H. Kuhn, and Christian Sturm

2006 "QCD decoupling at four loops", *Nucl. Phys. B*, 744, pp. 121-135, DOI: 10.1016/j.nuclphysb.2006.03.020, arXiv: hep-ph/0512060.

Cid Vidal, Xabier et al.

2019 "Report from Working Group 3: Beyond the Standard Model physics at the HL-LHC and HE-LHC", *CERN Yellow Rep. Monogr.*, 7, ed. by Andrea Dainese, Michelangelo Mangano, Andreas B. Meyer, Aleandro Nisati, Gavin Salam, and Mika Anton Vesterinen, pp. 585-865, DOI: 10.23731/CYRM-2019-007.585, arXiv: 1812.07831 [hep-ph].

Collins, John

2013 *Foundations of perturbative QCD*, Cambridge University Press, vol. 32, ISBN: 978-1-107-64525-7, 978-1-107-64525-7, 978-0-521-85533-4, 978-1-139-09782-6.

Collins, John C. and Davison E. Soper

1977 "Angular Distribution of Dileptons in High-Energy Hadron Collisions", *Phys. Rev. D*, 16, p. 2219, DOI: 10.1103/PhysRevD.16.2219.

1982 "Parton Distribution and Decay Functions", *Nucl. Phys.*, B194, pp. 445-492, DOI: 10.1016/0550-3213(82)90021-9.

Collins, John C., Davison E. Soper, and George F. Sterman

1989 "Factorization of Hard Processes in QCD", *Adv. Ser. Direct. High Energy Phys.*, 5, pp. 1-91, DOI: 10.1142/9789814503266_0001, arXiv: hep-ph/0409313.

Collins, John C. and Wu-Ki Tung

1986 "Calculating Heavy Quark Distributions", *Nucl. Phys. B*, 278, ed. by S. C. Loken, p. 934, DOI: 10.1016/0550-3213(86)90425-6.

Courtoy, Aurore, Joey Huston, Pavel Nadolsky, Keping Xie, Mengshi Yan, and C. -P. Yuan

2022 "Parton distributions need representative sampling" [May 2022], arXiv: 2205.10444 [hep-ph].

Cridge, T., L. A. Harland-Lang, A. D. Martin, and R. S. Thorne
2022 "QED parton distribution functions in the MSHT20 fit", *Eur. Phys. J. C*, 82, 1, p. 90, DOI: 10.1140/epjc/s10052-022-10028-2, arXiv: 2111.05357 [hep-ph].

Curci, G., W. Furmanski, and R. Petronzio
1980 "Evolution of Parton Densities Beyond Leading Order: The Nonsinglet Case", *Nucl. Phys.*, B175, pp. 27-92, DOI: 10.1016/0550-3213(80)90003-6.

Dainese, A. et al.
2019 "Physics Beyond Colliders: QCD Working Group Report" [Jan. 2019], arXiv: 1901.04482 [hep-ex].

Dawson, S., P. P. Giardino, and A. Ismail
2019 "Standard model EFT and the Drell-Yan process at high energy", *Phys. Rev. D*, 99, 3, p. 035044, DOI: 10.1103/PhysRevD.99.035044, arXiv: 1811.12260 [hep-ph].

De Florian, D. et al.
2016a "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector", arXiv: 1610.07922 [hep-ph].
2016b "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector", 2/2017 [Oct. 2016], DOI: 10.23731/CYRM-2017-002, arXiv: 1610.07922 [hep-ph].

De Roeck, A. and R. S. Thorne
2011 "Structure Functions", *Prog.Part.Nucl.Phys.*, 66, p. 727, DOI: 10.1016/j.ppnp.2011.06.001, arXiv: 1103.0555 [hep-ph].

De Roeck, Albert
2009 "Outlook: The PDF4LHC initiative", in *HERA and the LHC: 4th Workshop on the Implications of HERA for LHC Physics*, pp. 125-126, DOI: 10.3204/DESY-PROC-2009-02/63.

Del Debbio, Luigi, Stefano Forte, Jose I. Latorre, Andrea Piccione, and Joan Rojo
2007 "Neural network determination of parton distributions: The Nonsinglet case", *JHEP*, 03, p. 039, DOI: 10.1088/1126-6708/2007/03/039, arXiv: hep-ph/0701127.

Del Debbio, Luigi, Tommaso Giani, and Michael Wilson
2022 "Bayesian approach to inverse problems: an application to NNPDF closure testing", *Eur. Phys. J. C*, 82, 4, p. 330, DOI: 10.1140/epjc/s10052-022-10297-x, arXiv: 2111.05787 [hep-ph].

Diehl, Markus, Riccardo Nagar, and Frank J. Tackmann
2022 "ChiliPDF: Chebyshev interpolation for parton distributions", *Eur. Phys. J. C*, 82, 3, p. 257, DOI: 10.1140/epjc/s10052-022-10223-1, arXiv: 2112.09703 [hep-ph].

Diemoz, M., F. Ferroni, E. Longo, and G. Martinelli

1988 "Parton Densities from Deep Inelastic Scattering to Hadronic Processes at Super Collider Energies", *Z. Phys. C*, 39, p. 21, DOI: 10.1007/BF01560387.

Dittmar, M. et al.

2005 "Working Group I: Parton distributions: Summary report for the HERA LHC Workshop Proceedings", *WGI*, arXiv: hep-ph/0511119 [hep-ph].

Dokshitzer, Yuri L.

1977 "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics." *Sov. Phys. JETP*, 46, [Zh. Eksp. Teor. Fiz.73,1216(1977)], pp. 641-653.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth

1987 "Hybrid Monte Carlo", *Phys. Lett. B*, 195, pp. 216-222, DOI: 10.1016/0370-2693(87)91197-X.

Duhr, Claude, Alexander Huss, Aleksas Mazeliauskas, and Robert Szafron

2021 "An analysis of Bayesian estimates for missing higher orders in perturbative calculations", *JHEP*, 09, p. 122, DOI: 10.1007/JHEP09(2021)122, arXiv: 2106.04585 [hep-ph].

Duhr, Claude and Bernhard Mistlberger

2022 "Lepton-pair production at hadron colliders at N^3LO in QCD", *JHEP*, 03, p. 116, DOI: 10.1007/JHEP03(2022)116, arXiv: 2111.10379 [hep-ph].

Edward, Waring

1779 "VII. Problems concerning interpolations", *Phil. Trans. R. Soc.*, 69, pp. 59-67, DOI: http://doi.org/10.1098/rstl.1779.0008.

Ellis, John, Maeve Madigan, Ken Mimasu, Veronica Sanz, and Tevong You

2021 "Top, Higgs, Diboson and Electroweak Fit to the Standard Model Effective Field Theory", *JHEP*, 04, p. 279, DOI: 10.1007/JHEP04(2021)279, arXiv: 2012.02779 [hep-ph].

Ellis, R. Keith, W. James Stirling, and B. R. Webber

1996 *QCD and collider physics*, Cambridge University Press.

2011 *QCD and collider physics*, Cambridge University Press, vol. 8, ISBN: 978-0-511-82328-2, 978-0-521-54589-1, DOI: 10.1017/CBO9780511628788.

Ethier, Jacob J., Giacomo Magni, Fabio Maltoni, Luca Mantani, Emanuele R. Nocera, Juan Rojo, Emma Slade, Eleni Vryonidou, and Cen Zhang

2021 "Combined SMEFT interpretation of Higgs, diboson, and top quark data from the LHC", *JHEP*, 11, p. 089, DOI: 10.1007/JHEP11(2021)089, arXiv: 2105.00006 [hep-ph].

Ethier, Jacob J. and Emanuele R. Nocera

2020 "Parton Distributions in Nucleons and Nuclei", *Ann. Rev. Nucl. Part. Sci.*, 70, pp. 43-76, DOI: 10.1146/annurev-nucl-011720-042725, arXiv: 2001.07722 [hep-ph].

Feynman, R. P.

1969 "The behavior of hadron collisions at extreme energies", *Conf. Proc. C*, 690905, pp. 237-258.

Fiaschi, Juri, Francesco Giuli, Francesco Hautmann, and Stefano Moretti

2021 "Lepton-Charge and Forward-Backward Asymmetries in Drell-Yan Processes for Precision Electroweak Measurements and New Physics Searches", *Nucl. Phys. B*, 968, p. 115444, DOI: 10.1016/j.nuclphysb.2021.115444, arXiv: 2103.10224 [hep-ph].

2022 "Enhancing the Large Hadron Collider sensitivity to charged and neutral broad resonances of new gauge sectors", *JHEP*, 02, p. 179, DOI: 10.1007/JHEP02(2022)179, arXiv: 2111.09698 [hep-ph].

Forte, Stefano and Stefano Carrazza

2020 "Parton distribution functions" [Aug. 2020], arXiv: 2008.12305 [hep-ph].

Forte, Stefano, Lluis Garrido, Jose I. Latorre, and Andrea Piccione

2002 "Neural network parametrization of deep-inelastic structure functions", *JHEP*, 05, p. 062, eprint: hep-ph/0204232.

Forte, Stefano, Andrea Isgrò, and Gherardo Vita

2014 "Do we need N^3LO Parton Distributions?", *Phys.Lett.*, B731, pp. 136-140, DOI: 10.1016/j.physletb.2014.02.027, arXiv: 1312.6688 [hep-ph].

Forte, Stefano, Eric Laenen, Paolo Nason, and Juan Rojo

2010 "Heavy quarks in deep-inelastic scattering", *Nucl. Phys. B*, 834, pp. 116-162, DOI: 10.1016/j.nuclphysb.2010.03.014, arXiv: 1001.2312 [hep-ph].

Forte, Stefano, Davide Napoletano, and Maria Ubiali

2018 "Z boson production in bottom-quark fusion: a study of b-mass effects beyond leading order", *Eur. Phys. J. C*, 78, 11, p. 932, DOI: 10.1140/epjc/s10052-018-6414-8, arXiv: 1803.10248 [hep-ph].

Forte, Stefano and Giovanni Ridolfi

2003 "Renormalization group approach to soft gluon resummation", *Nucl. Phys.*, B650, pp. 229-270, DOI: 10.1016/S0550-3213(02)01034-9, arXiv: hep-ph/0209154 [hep-ph].

Frederix, R., S. Frixione, V. Hirschi, D. Pagani, H. -S. Shao, and M. Zaro

2018 "The automation of next-to-leading order electroweak calculations", *JHEP*, 07, p. 185, DOI: 10.1007/JHEP07(2018)185, arXiv: 1804.10017 [hep-ph].

Gao, Jun, Lucian Harland-Lang, and Juan Rojo
2018 "The Structure of the Proton in the LHC Precision Era", *Phys. Rept.*, 742, pp. 1-121, DOI: 10.1016/j.physrep.2018.03.002, arXiv: 1709.04922 [hep-ph].

Gbedo, Yémalin Gabin and Mariane Mangin-Brinet
2017 "Markov chain Monte Carlo techniques applied to parton distribution functions determination: Proof of concept", *Phys. Rev. D*, 96, 1, p. 014015, DOI: 10.1103/PhysRevD.96.014015, arXiv: 1701.07678 [hep-ph].

Giele, W. et al.
2002 "The QCD / SM working group: Summary report", in *2nd Les Houches Workshop on Physics at TeV Colliders*, pp. 275-426, arXiv: hep-ph/0204316.

Gnendiger, C. et al.
2017 "To d, or not to d: recent developments and comparisons of regularization schemes", *Eur. Phys. J. C*, 77, 7, p. 471, DOI: 10.1140/epjc/s10052-017-5023-2, arXiv: 1705.01827 [hep-ph].

Grazzini, Massimiliano, Stefan Kallweit, and Marius Wiesemann
2018 "Fully differential NNLO computations with MATRIX", *Eur. Phys. J. C*, 78, 7, p. 537, DOI: 10.1140/epjc/s10052-018-5771-7, arXiv: 1711.06631 [hep-ph].

Greljo, Admir, Shayan Iranipour, Zahari Kassabov, Maeve Madigan, James Moore, Juan Rojo, Maria Ubiali, and Cameron Voisey
2021 "Parton distributions in the SMEFT from high-energy Drell-Yan tails", *JHEP*, 07, p. 122, DOI: 10.1007/JHEP07(2021)122, arXiv: 2104.02723 [hep-ph].

Gribov, V. N. and L. N. Lipatov
1972 "Deep inelastic e p scattering in perturbation theory", *Sov. J. Nucl. Phys.*, 15, [Yad. Fiz.15,781(1972)], pp. 438-450.

Hadjidakis, C. et al.
2021 "A fixed-target programme at the LHC: Physics case and projected performances for heavy-ion, hadron, spin and astroparticle studies", *Phys. Rept.*, 911, pp. 1-83, DOI: 10.1016/j.physrep.2021.01.002, arXiv: 1807.00603 [hep-ex].

Halzen, Francis and Logan Wille
2016 "Charm contribution to the atmospheric neutrino flux", *Phys. Rev. D*, 94, 1, p. 014014, DOI: 10.1103/PhysRevD.94.014014, arXiv: 1605.01409 [hep-ph].

Harris, B. W., J. Smith, and R. Vogt
1996 "Reanalysis of the EMC charm production data with extrinsic and intrinsic charm at NLO", *Nucl. Phys. B*, 461, pp. 181-196, DOI: 10.1016/0550-3213(95)00652-4, arXiv: hep-ph/9508403.

Hastings, W. K.

1970 "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika*, 57, 1 [Apr. 1970], pp. 97-109, ISSN: 0006-3444, DOI: 10.1093/biomet/57.1.97, eprint: https://academic.oup.com/biomet/article-pdf/57/1/97/23940249/57-1-97.pdf.

Heinrich, Gudrun

2021 "Collider Physics at the Precision Frontier", *Phys. Rept.*, 922, pp. 1-69, DOI: 10.1016/j.physrep.2021.03.006, arXiv: 2009.00516 [hep-ph].

Hekhorn, Felix

2019 *Next-to-Leading Order QCD Corrections to Heavy-Flavour Production in Neutral Current DIS*, PhD thesis, Tubingen U., Math. Inst., DOI: 10.15496/publikation-34811, arXiv: 1910.01536 [hep-ph].

Herzog, F., B. Ruijl, T. Ueda, J. A. M. Vermaseren, and A. Vogt

2017 "The five-loop beta function of Yang-Mills theory with fermions", *JHEP*, 02, p. 090, DOI: 10.1007/JHEP02(2017)090, arXiv: 1701.01404 [hep-ph].

Hobbs, T. J., J. T. Londergan, and W. Melnitchouk

2014 "Phenomenology of nonperturbative charm in the nucleon", *Phys. Rev. D*, 89, 7, p. 074008, DOI: 10.1103/PhysRevD.89.074008, arXiv: 1311.1578 [hep-ph].

Hoffmann, Erhard and Richard Moore

1983 "Subleading Contributions to the Intrinsic Charm of the Nucleon", *Z. Phys. C*, 20, p. 71, DOI: 10.1007/BF01577720.

Hou, Tie-Jiun et al.

2021 "New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC", *Phys. Rev. D*, 103, 1, p. 014013, DOI: 10.1103/PhysRevD.103.014013, arXiv: 1912.10053 [hep-ph].

Hou, Tie-Jiun, Sayipjamal Dulat, Jun Gao, Marco Guzzi, Joey Huston, Pavel Nadolsky, Carl Schmidt, Jan Winter, Keping Xie, and C. -P. Yuan

2018 "CT14 Intrinsic Charm Parton Distribution Functions from CTEQ-TEA Global Analysis", *JHEP*, 02, p. 059, DOI: 10.1007/JHEP02(2018)059, arXiv: 1707.00657 [hep-ph].

Jadach, S.

2020 *private communications*.

Jadach, S., W. Płaczek, S. Sapeta, A. Siodmok, and M. Skrzypek

2016 "Parton distribution functions in Monte Carlo factorisation scheme", *Eur. Phys. J.*, C76, 12, p. 649, DOI: 10.1140/epjc/s10052-016-4508-8, arXiv: 1606.00355 [hep-ph].

Jimenez-Delgado, P., T.J. Hobbs, J.T. Londergan, and W. Melnitchouk

2015 "New limits on intrinsic charm in the nucleon from global analysis of parton distributions", *Phys.Rev.Lett.*, 114, 8, p. 082002, DOI: 10.1103/PhysRevLett.114.082002, arXiv: 1408.1708 [hep-ph].

Kassabov, Zahari, Maria Ubiali, and Cameron Voisey

2022 "Parton distributions with scale uncertainties: a MonteCarlo sampling approach" [July 2022], arXiv: 2207.07616 [hep-ph].

Khachatryan, Vardan et al.

2017 "Search for narrow resonances in dilepton mass spectra in proton-proton collisions at $\sqrt{s}$ = 13 TeV and combination with 8 TeV data", *Phys. Lett.*, B768, pp. 57-80, DOI: 10.1016/j.physletb.2017.02.010, arXiv: 1609.05391 [hep-ex].

Khalek, Rabah Abdul, Jacob J. Ethier, Emanuele R. Nocera, and Juan Rojo

2021 "Self-consistent determination of proton and nuclear PDFs at the Electron Ion Collider", *Phys. Rev. D*, 103, 9, p. 096005, DOI: 10.1103/PhysRevD.103.096005, arXiv: 2102.00018 [hep-ph].

Kovařık, Karol, Pavel M. Nadolsky, and Davison E. Soper

2020 "Hadronic structure in high-energy collisions", *Rev. Mod. Phys.*, 92, 4, p. 045003, DOI: 10.1103/RevModPhys.92.045003, arXiv: 1905.06957 [hep-ph].

Lai, H. L., J. Huston, S. Kuhlmann, Fredrick I. Olness, Joseph F. Owens, D. E. Soper, W. K. Tung, and H. Weerts

1997 "Improved parton distributions from global analysis of recent deep inelastic scattering and inclusive jet data", *Phys. Rev. D*, 55, pp. 1280-1296, DOI: 10.1103/PhysRevD.55.1280, arXiv: hep-ph/9606399.

Lam, Siu Kwan, Antoine Pitrou, and Stanley Seibert

2015 "Numba: A LLVM-Based Python JIT Compiler", in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, LLVM '15, Association for Computing Machinery, Austin, Texas, ISBN: 9781450340052, DOI: 10.1145/2833157.2833162.

Luthe, Thomas, Andreas Maier, Peter Marquard, and York Schroder

2017 "The five-loop Beta function for a general gauge group and anomalous dimensions beyond Feynman gauge", *JHEP*, 10, p. 166, DOI: 10.1007/JHEP10(2017)166, arXiv: 1709.07718 [hep-ph].

Luthe, Thomas, Andreas Maier, Peter Marquard, and York Schröder

2016 "Towards the five-loop Beta function for a general gauge group", *JHEP*, 07, p. 127, DOI: 10.1007/JHEP07(2016)127, arXiv: 1606.08662 [hep-ph].

Maltoni, F.

2018 "Basics of QCD for the LHC: pp $\to$ H + X as a case study", *CERN Yellow Rep. School Proc.*, 2, pp. 41-67, DOI: 10.23730/CYRSP-2018-002.41.

McGowan, J., T. Cridge, L. A. Harland-Lang, and R. S. Thorne

2022 "Approximate N^{3}LO Parton Distribution Functions with Theoretical Uncertainties: MSHT20aN^{3}LO PDFs" [July 2022], arXiv: 2207.04739 [hep-ph].

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller

1953 "Equation of State Calculations by Fast Computing Machines", *The Journal of Chemical Physics*, 21, 6, pp. 1087-1092, DOI: 10.1063/1.1699114, eprint: https://doi.org/10.1063/1.1699114.

Mitov, A., S. Moch, and A. Vogt

2006 "Next-to-Next-to-Leading Order Evolution of Non-Singlet Fragmentation Functions", *Phys. Lett. B*, 638, pp. 61-67, DOI: 10.1016/j.physletb.2006.05.005, arXiv: hep-ph/0604053.

Moch, S., M. Rogal, and A. Vogt

2008 "Differences between charged-current coefficient functions", *Nucl. Phys. B*, 790, pp. 317-335, DOI: 10.1016/j.nuclphysb.2007.09.022, arXiv: 0708.3731 [hep-ph].

Moch, S., B. Ruijl, T. Ueda, J. A. M. Vermaseren, and A. Vogt

2022 "Low moments of the four-loop splitting functions in QCD", *Phys. Lett. B*, 825, p. 136853, DOI: 10.1016/j.physletb.2021.136853, arXiv: 2111.15561 [hep-ph].

Moch, S. and J. A. M. Vermaseren

2000 "Deep inelastic structure functions at two loops", *Nucl. Phys. B*, 573, pp. 853-907, DOI: 10.1016/S0550-3213(00)00045-6, arXiv: hep-ph/9912355.

Moch, S., J. A. M. Vermaseren, and A. Vogt

2004 "The Three loop splitting functions in QCD: The Nonsinglet case", *Nucl. Phys.*, B688, pp. 101-134, DOI: 10.1016/j.nuclphysb.2004.03.030, arXiv: hep-ph/0403192 [hep-ph].

2005 "The Longitudinal structure function at the third order", *Phys. Lett. B*, 606, pp. 123-129, DOI: 10.1016/j.physletb.2004.11.063, arXiv: hep-ph/0411112.

2009 "Third-order QCD corrections to the charged-current structure function F(3)", *Nucl. Phys. B*, 813, pp. 220-258, DOI: 10.1016/j.nuclphysb.2009.01.001, arXiv: 0812.4168 [hep-ph].

Moch, S. and A. Vogt

2008 "On third-order timelike splitting functions and top-mediated Higgs decay into hadrons", *Phys. Lett. B*, 659, pp. 290-296, DOI: 10.1016/j.physletb.2007.10.069, arXiv: 0709.3899 [hep-ph].

Paiva, S., Marina Nielsen, F. S. Navarra, F. O. Duraes, and L. L. Barz

1998 "Virtual meson cloud of the nucleon and intrinsic strangeness and charm", *Mod. Phys. Lett. A*, 13, pp. 2715-2724, DOI: 10.1142/S0217732398002886, arXiv: hep-ph/9610310.

Peskin, Michael E. and Daniel V. Schroeder

1995 *An Introduction to quantum field theory*, Addison-Wesley, Reading, USA, ISBN: 978-0-201-50397-5.

Petrillo, Giacomo

2022 "Bayesian Parton Distribution Functions fit with Gaussian processes", in, Unpublished report, Third meeting of the Bayesian Analysis Conspiracy Heretics (BACH), eprint: https://www.giacomopetrillo.com/scuola/gppdf.pdf.

Pumplin, Jon

2006 "Light-cone models for intrinsic charm and bottom", *Phys. Rev. D*, 73, p. 114015, DOI: 10.1103/PhysRevD.73.114015, arXiv: hep-ph/0508184.

Rojo, Juan

2016 "PDF4LHC recommendations for Run II", *PoS*, DIS2016, p. 018, DOI: 10.22323/1.265.0018, arXiv: 1606.08243 [hep-ph].

2019 "The Partonic Content of Nucleons and Nuclei" [Oct. 2019], arXiv: 1910.03408 [hep-ph].

Runge, C.

1901 "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten." German, *Schlömilch Z.*, 46, pp. 224-243.

Salam, Gavin P. and Juan Rojo

2009 "A Higher Order Perturbative Parton Evolution Toolkit (HOPPET)", *Comput. Phys. Commun.*, 180, pp. 120-156, DOI: 10.1016/j.cpc.2008.08.010, arXiv: 0804.3755 [hep-ph].

Schroder, Y. and M. Steinhauser

2006 "Four-loop decoupling relations for the strong coupling", *JHEP*, 01, p. 051, DOI: 10.1088/1126-6708/2006/01/051, arXiv: hep-ph/0512058.

Schwan, Christopher, Alessandro Candido, Felix Hekhorn, and Stefano Carrazza

2022a *N3PDF/pineappl: v0.5.0-beta.6*, version v0.5.0-beta.6, DOI: 10.5281/zenodo.5846421.

2022b *N3PDF/pineappl: v0.5.5*, version v0.5.5, Zenodo, DOI: 10.5281/zenodo.7023438.

2022c *NNPDF/pineappl: v0.5.7*, version v0.5.7, DOI: 10.5281/zenodo.7145377.

Sirunyan, Albert M et al.

2019 "Search for contact interactions and large extra dimensions in the dilepton mass spectra from proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP*, 04, p. 114, DOI: 10.1007/JHEP04(2019)114, arXiv: 1812.10443 [hep-ex].

2021 "Search for resonant and nonresonant new phenomena in high-mass dilepton final states at $\sqrt{s} = 13$ TeV", *JHEP*, 07, p. 208, DOI: 10.1007/JHEP07(2021)208, arXiv: 2103.02708 [hep-ex].

2018 "Measurement of the weak mixing angle using the forward-backward asymmetry of Drell-Yan events in pp collisions at 8 TeV", *Eur. Phys. J. C*, 78, 9, p. 701, DOI: 10.1140/epjc/s10052-018-6148-7, arXiv: 1806.00863 [hep-ex].

Sjostrand, Torbjorn, Stephen Mrenna, and Peter Z. Skands

2008 "A Brief Introduction to PYTHIA 8.1", *Comput. Phys. Commun.*, 178, pp. 852-867, DOI: 10.1016/j.cpc.2008.01.036, arXiv: 0710.3820 [hep-ph].

Skands, Peter, Stefano Carrazza, and Juan Rojo

2014 "Tuning PYTHIA 8.1: the Monash 2013 Tune", *European Physical Journal*, 74, p. 3024, DOI: 10.1140/epjc/s10052-014-3024-y, arXiv: 1404.5630 [hep-ph].

Steffens, Fernanda Monti, W. Melnitchouk, and Anthony William Thomas

1999 "Charm in the nucleon", *Eur. Phys. J. C*, 11, pp. 673-683, DOI: 10.1007/s100520050663, arXiv: hep-ph/9903441.

Sterman, George F.

1987 "Summation of Large Corrections to Short Distance Hadronic Cross-Sections", *Nucl. Phys.*, B281, pp. 310-364, DOI: 10.1016/0550-3213(87)90258-6.

Süli, E. and D.F. Mayers

2003 *An Introduction to Numerical Analysis*, An Introduction to Numerical Analysis, Cambridge University Press, ISBN: 9780521007948.

Tumasyan, Armen et al.

2022 "Measurement of the Drell-Yan forward-backward asymmetry at high dilepton masses in proton-proton collisions at $\sqrt{s} = 13$ TeV" [Feb. 2022], arXiv: 2202.12327 [hep-ex].

Van Neerven, W. L. and A. Vogt

2000 "NNLO evolution of deep inelastic structure functions: The Singlet case", *Nucl. Phys.*, B588, pp. 345-373, DOI: 10.1016/S0550-3213(00)00480-6, arXiv: hep-ph/0006154 [hep-ph].

2001 "Nonsinglet structure functions beyond the next-to-next-to-leading order", *Nucl. Phys. B*, 603, pp. 42-68, DOI: 10.1016/S0550-3213(01)00158-4, arXiv: hep-ph/0103123.

Vermaseren, J. A. M., S. A. Larin, and T. van Ritbergen

1997 "The four loop quark mass anomalous dimension and the invariant quark mass", *Phys. Lett. B*, 405, pp. 327-333, DOI: 10.1016/S0370-2693(97)00660-6, arXiv: hep-ph/9703284.

Vermaseren, J. A. M., A. Vogt, and S. Moch

2005 "The Third-order QCD corrections to deep-inelastic scattering by photon exchange", *Nucl. Phys. B*, 724, pp. 3-182, DOI: 10.1016/j.nuclphysb.2005.06.020, arXiv: hep-ph/0504242.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors

2020 "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python", *Nature Methods*, 17, pp. 261-272, DOI: 10.1038/s41592-019-0686-2.

Vogt, A.

2005 "Efficient evolution of unpolarized and polarized parton distributions with QCD-PEGASUS", *Comput. Phys. Commun.*, 170, pp. 65-92, DOI: 10.1016/j.cpc.2005.03.103, arXiv: hep-ph/0408244.

Vogt, A., S. Moch, M. Rogal, and J. A. M. Vermaseren

2008 "Towards the NNLO evolution of polarised parton distributions", *Nucl. Phys. B Proc. Suppl.*, 183, ed. by J. Blumlein, S. Moch, and T. Riemann, pp. 155-161, DOI: 10.1016/j.nuclphysbps.2008.09.097, arXiv: 0807.1238 [hep-ph].

Vogt, A., S. Moch, and J. A. M. Vermaseren

2004 "The Three-loop splitting functions in QCD: The Singlet case", *Nucl. Phys.*, B691, pp. 129-181, DOI: 10.1016/j.nuclphysb.2004.04.024, arXiv: hep-ph/0404111 [hep-ph].

2014 "A calculation of the three-loop helicity-dependent splitting functions in QCD", *PoS*, LL2014, ed. by Martina Mende, p. 040, DOI: 10.22323/1.211.0040, arXiv: 1405.3407 [hep-ph].

Wilson, Kenneth G.

1974 "Confinement of Quarks", *Phys. Rev. D*, 10, ed. by J. C. Taylor, pp. 2445-2459, DOI: 10.1103/PhysRevD.10.2445.

Wobisch, M., D. Britzger, T. Kluge, K. Rabbertz, and F. Stober

2011 "Theory-Data Comparisons for Jet Measurements in Hadron-Induced Processes", arXiv: 1109.1310 [hep-ph].

Xie, Keping, T. J. Hobbs, Tie-Jiun Hou, Carl Schmidt, Mengshi Yan, and C. -P. Yuan

2022 "Photon PDF within the CT18 global analysis", *Phys. Rev. D*, 105, 5, p. 054006, DOI: 10.1103/PhysRevD.105.054006, arXiv: 2106.10299 [hep-ph].

ACKNOWLEDGMENTS

I started my PhD student experience with a suboptimal attitude, since personal and referred experiences made me rather critical about academia as a whole, and I despised some of the dynamics *naturally* associated with the research environment, which many had the *pleasure* to experience.

Nevertheless, I have been extremely lucky: lucky from the very beginning, since even before starting I got some kind advice from members of the institution where I was a student, that have done their best to motivate me and to show some interesting directions worth to investigate. For this, I am grateful to professors Enrico Trincherini and Gigi Rolandi, and my master thesis advisor Massimo D'Elia. It is extremely likely that without them, I would have not even started the PhD adventure.

Yet the beginning of a significant journey is always very far from its end, and I considered many times if just keep going would have been a somewhat optimal choice. There is no such a thing as a straight way, and there has been many non-trivial steps in my own, I got to the end through all the choices done. For this, it has been fundamental not to be alone. I shared this experience with several of my friends from Pisa, that accompanied me for the five years of our Bachelor and Master. Their example and more direct discussions closed the geographical gap between us. Among them, I should thank in particular Marco Costa, since he strongly supported me, in particular at the very beginning. But even more, I should thank the members of our Milan group, in the order I met them: Juan Cruz-Martinez, Christopher Schwan, Tanjona Rabemananjara, and Roy Stegeman, since working with them have been many times the strongest motivation to dedicate my entire self to our common goals. A separate mention has to be dedicated to Felix Hekhorn, since he assisted me while still learning the job almost from zero, and we kept working together every day: literally every day and all day long during the times of the global pandemic. I have never been feeling really isolated, and in practice this is mostly due to him, that lead me and had the patience to teach me everything. Even if they arrived later on, also Giacomo Magni, Andrea Barontini, and Niccolò Laurenti have been optimal companions to share a relevant part of the journey.

It is sometimes a trivial addition to this list the presence of each one's supervisor. I am not going to disobey this consolidated tradition, but I would like to remark how non-trivial this is for me. As already mentioned, even before starting I have been aware of many negative experiences. But this has not been my case, entirely thanks to prof. Stefano Forte. He has been welcoming from the very first day, when I met him in Milan without even knowing I were looking for a PhD position. His door was every day open to my questions, to solve my doubts and

discuss my proposals. He never just dismissed any issue, and carefully listened also to my personal concerns, encouraging me to continue this career.

It has also been incredibly formative to work in the NNPDF collaboration lead by him, and to see various problems arising from working as a group, together with many solutions. It is definitely a complex task to coordinate with each other, but it is the only way to achieve certain results. This entire thesis gather many contributions from NNPDF members, and I am really thankful to all of them. Not only thankful for the work done together, but also for the many lessons I learnt in the process. It took time, but I appreciated different virtues of each one. Diversity is essential for any complex task, and it is completely lost if limiting ourselves to work alone.

Finally, my greatest gratitude goes to my family, in particular my parents, Paolo and Lucia. It might seem like they have done nothing for my PhD, but if I have been able to work on it, it has been thanks to their every day support, both from the practical and emotional point of view. In a very similar way I am grateful to my partner, Viviana. But I owe her something more: she decided to move together with me from Pisa to Milan, to start together a new journey, and to live together every day. She encouraged and supported me so many times and in so many ways. During my education I have been taught to often remind that *"non di solo pane vive l'uomo"*, and definitely research on its own is not enough for me.

Milan, November 2024

Alessandro Candido