Towards a new generation of Parton Distribution Functions: from high-precision collider data to lattice Quantum Chromodynamics

Tommaso Giani

Doctor of Philosophy
The University of Edinburgh
2021

Most of the research presented in this thesis has been done in collaboration with other colleagues, as detailed in the following.

The results presented in chapters 3 and 5 have been produced within the NNPDF collaboration. The original publications are

NNPDF Collaboration, R. Abdul Khalek et al., *A first determination of parton distributions with theoretical uncertainties*, *Eur. Phys. J.* **C** (2019) 79:838, [arXiv:1905.04311]

NNPDF Collaboration, R. Abdul Khalek et al., *Parton Distributions with Theory Uncertainties: General Formalism and First Phenomenological Studies*, *Eur. Phys. J. C* **79** (2019), no. 11 931, [arXiv:1906.10698] ,

in collaboration with Rabah Abdul Khalek, Richard D. Ball, Stefano Carrazza, Stefano Forte, Zahari Kassabov, Emanuele R. Nocera, Rosalyn L. Pearson, Juan Rojo, Luca Rottoli, Maria Ubiali, Cameron Voisey and Michael Wilson.

Results reported in chapter 4 were first presented in

R. Abdul Khalek et al., *Phenomenology of NNLO jet production at the LHC and its impact on parton distributions*, *Eur. Phys. J. C* **80** (2020), no. 8 797, [arXiv:2005.11327] ,

in collaboration with Rabah Abdul Khalek, Stefano Forte, Thomas Gehrmann, Aude Gehrmann-De Ridder, Nigel Glover, Alexander Huss, Emanuele R. Nocera, Joao Pires, Juan Rojo and Giovanni Stagnitto.

Chapter 6 is based on

S. Forte, T. Giani, and D. Napoletano, *Fitting the b-quark PDF as a massive-b scheme: Higgs production in bottom fusion*, *Eur. Phys. J. C* **79** (2019), no. 7 609, [arXiv:1905.02207]

in collaboration with Stefano Forte and Davide Napoletano.

Chapter 7 is based on

L. Del Debbio, T. Giani, and C. J. Monahan, *Notes on lattice observables for parton distributions: nongauge theories*, *JHEP* **09** (2020) 021, [arXiv:2007.02131] ,

in collaboration with Luigi Del Debbio and Christopher J. Monahan.

Chapter 8 is based on

K. Cichy, L. Del Debbio, and T. Giani, *Parton distributions from lattice data: the nonsinglet case*, *JHEP* **10** (2019) 137, [arXiv:1907.06037] ,

in collaboration with Krzysztof Cichy and Luigi Del Debbio.

Finally Chapter 9 is based on

L. Del Debbio, T. Giani, J. Karpie, K. Orginos, A. Radyushkin, and S. Zafeiropoulos, *Neural-network analysis of Parton Distribution Functions from Ioffe-time pseudodistributions*, *JHEP* **02** (2021) 138, [arXiv:2010.03996]

in collaboration with Luigi Del Debbio, Joseph Karpie, Kostas Orginos, Anatoly Radyushkinc and Savvas Zafeiropoulos.

When results are presented I tried to report more details where I believe that I have played a major role, reporting additional material when it is necessary for the general understanding of the discussion.

In order to write the first two chapters, devoted to a general overview of some basic concepts of QCD and factorization, along with the cited references I have also referred to a number of lectures I have attended during the first year of my PhD at the University of Edinburgh. In particular, I want to mention the 2017/2018 QCD course held by Prof. Einan Gardi and the 2017/2018 lattice QCD lectures given by Prof. Antonin Portelli.

Unless specified in the caption, I have produced all the figures reported in this thesis.

Lay Summary

In this thesis we present a number of studies concerning the determination of the Parton Distribution Functions (PDFs) of the proton. PDFs describe the internal structure of the nucleons in terms of their constituent quarks and gluons, and therefore they are important to gain a better understanding of the fundamental structure of the proton. Additionally PDFs are a necessary input to perform a number of relevant computations in high energy-physics phenomenology, which makes their precise knowledge, together with the determination of the corresponding uncertainty, a key ingredient for new physics studies. It is therefore important to produce more and more precise PDFs, researching into new numerical frameworks and physical ideas. In this thesis we present a number of steps in this direction, addressing a number of different topics, which span from the impact of new experimental data from the Large Hadron Collider (LHC) to recent developments from the lattice community.

Abstract

A precise understanding of the proton structure, encoded in Parton Distribution Functions (PDFs), is required to make reliable predictions and analyses at the Large Hadron Collider (LHC), the main source of experimental data probing subnuclear interactions we have today. PDFs have played a central role in the recent discovery of the Higgs boson and, since it is increasingly clear that any effect due to new physics will manifest itself as small deviations from the current theory, a precise determination of PDFs is likely to be a key ingredient for new physics studies. The PDFs are formally defined as matrix elements of renormalized operators in Quantum Chromodynamics (QCD) involving hadronic states. They are inherently non-perturbative quantities, and they are extracted from global QCD analysis over experimental data using the so-called factorization theorems. Producing a new generation of PDFs, satisfying the precision and reliability requirements demanded by the current research, is a challenging task which involves, together with the experimental data input, the development of robust fitting methodologies, along with new physical ideas. In this thesis we present a number of progresses which have been developed in the last few years in the context of PDFs determination, some of which will lead to the next PDFs release by the NNPDF collaboration. In particular we will discuss a new framework for global PDFs determinations, the impact of new experimental data, with particular emphasis on jets data, the inclusion of theory uncertainty in a PDFs fit and the treatment of heavy quarks distributions. We will then discuss a set of recent ideas which would allow to extract PDFs from equal time correlators computable within the framework of lattice QCD, and we will present results regarding data coming from different lattice approaches and collaborations.

A tutti i miei nonni e a Franco Sar

Acknowledgments

First of all, I would like to thank my supervisor, Luigi Del Debbio. I am deeply happy and grateful for the work done together. It has been stimulating, inspiring, exciting and extremely fun, and it has allowed me to grow up quite a lot. I truly consider our long blackboard discussions the most valuable thing of the whole PhD experience, and I am looking forward for more of that in the coming years. I am grateful to Stefano Forte, for the guide and help he has provided me, for the work done together and for always giving a complete answer to any question, usually in less than three minutes. I would like to thank all the members of the particle theory group in Edinburgh, for the great and stimulating environment they have created, in particular Einan Gardi, Roman Zwicky and Richard Ball. I thank all the members of the NNPDF collaboration, with whom I have been working a lot during the last few years. In particular I am deeply in debts with Emanuele R. Nocera, who has helped me a countless amount of times in a number of different tasks and projects, providing me with an example of efficiency, hard work and human kindness which is difficult to match. I am grateful to Juan Rojo, for the support and the opportunity he gave me to work in his group at the end of my PhD. A special thank goes to Rabah, Cameron, Rosalyn, Michael, Zahari, Stefano, Shayan, Emma and Luca. I have learnt a lot from each of them, in terms of physics, coding and teamwork, and it has been a real pleasure to regularly meet them in the different meetings we had during the years. In particular I want to thank Rosalyn and Michael for having shared with me the experience of being PhD students in Edinburgh. It has been great for me to regularly see and work with them and I am grateful for all the trips we had together around the world. I am also very proud of the progresses Michael has done in learning the

basics of the italian language. I want to thank Simone Marzani for the help and support he gave me, Marco Bonvini and Rhorry Gauld, for useful and stimulating discussions about resummation. I am grateful to Nathan P. Hartland for the many questions he answered, to Davide Napoletano for the work done together and for the important help he provided me during the very first days of my PhD. I want to thank Guido Cossu and Ava Khamseh for the work done together during my first year and Anatoly Radyushkin for illuminating discussions, which allowed me to better understand the relation between light-cone and euclidean quantities. Finally, I would like to thank the master students who, when in person teaching was only partially allowed, regularly came to the QFT tutorials, watching me missing minus signs and factors two at the blackboard. Their interest, enthusiasm and questions have been of great support for me, and I have learnt a lot from them.

A number of not-physics friends have played a central role for me during my time in Edinburgh, making me feel home every single day. It wouldn't have been possible for me to do any work without them. A huge thank you to Stefano, for a number of things which wouldn't fit a page. For the pizza, pasta al pesto, creative swearwords and one-arm pull up sessions among other things. But, most of all, for being the best possible flatmate on earth. I am pretty sure that without him I would have already died in some stupid way. I thank Lauren, for having helped me in a number of different ways, for her kindness, her support and for always being there, during both the good and not-so-good times. I thank Roxane for the great time together, for the cycling, for having shown me parts of Scotland I had not seen before and for her help and support in looking for a job. A special thank you to my main British climbing buddies Ed and Patrick with whom I spent countless days and nights pulling on tiny nasty crimps, dishonest slopers and silly pinches, discussing quality climbing videos. And of course huge thanks to the whole Garage Crimpers team, Scott, Colin and Sinclair. I am grateful for the countless sessions on the A2, on the 45.8 in the power garage and of course in the County, which, together with Dumby, has now a very special place among the longlist of my favourite climbing spots. Thanks to Andrea, for having being close during a strange time, when the pandemic started. Finally, I would like to thank Michael, who offered me a room during my very first days in Scotland, without even knowing who I was.

I want to thank a number of friends from Italy, who have constantly supported and helped me in different ways. Huge thanks to Gi, for having shared with me

the first years of this experience and for having always being there for me, despite everything. Thanks to my serious dottorandi friends Fabbri, Tommy, Mario and Anna who in one way or another are always there. I am always looking forward for seeing them every time we are all back to Italy (especially when this happens at Silvi). Thanks to Alice, for the help and support provided during the last months of my PhD. Thanks to Rick, for the chats we have every time we see each other. Huge thanks to Lollo, for being a beast, for all the climbing we are always having together every time I am back to Italy and for the support he is always giving me. Thanks to Ale, for the countless routes tried together and for always taking every single whipper. Thanks to my Passaggio Obbligato friends, in particular to Fede Montagna and Teo Nill, who regularly sent me supportive messages. Finally thanks to Luca, who in his own way has always been there for me.

Grazie ai miei fratelli, ai miei nonni e soprattutto ai miei genitori per l'aiuto e l'affetto incondizionato che ricevo ogni singolo giorno. Infine grazie Sar per essere stato, a tuo modo, di guida e sostegno. Vivi in ogni piccolo traguardo quotidiano.

Contents

Introduction 1

1 Basics of QCD 5

1.1 Lagrangian and its symmetries 6

1.2 The running coupling and asymptotic freedom 8

1.3 Quark masses 10

1.4 Perturbative and non-Perturbative approaches 11

2 Factorization theorems in QCD 15

2.1 The Parton Model 15

2.2 Improved Parton Model and factorization of collinear singularities .. 20

2.2.1 Next-to-Leading-Order QCD corrections 20

2.2.2 Factorization of collinear singularities 22

2.3 Parton Distribution Functions 25

2.3.1 PDFs operator definition 25

2.3.2 DGLAP evolution equations 28

2.4 Heavy quarks 32

3 Global fits of Parton Distribution Functions 36

3.1 NNPDF methodology 37

3.1.1 The Monte Carlo replica method for error propagation 38

3.1.2 Parameterization and Neural Networks 39

3.1.3 Minimization and stopping 42

3.1.4 Fast Kernel tables 43

3.2 Towards NNPDF4.0 44

3.2.1 Architecture and general structure 45

3.2.2 Fit basis 46

3.2.3 Theoretical constraints 47

3.2.4 Results 52

3.2.5 Fit basis independence 53

4 Theory interpretation of jets production at the LHC 58

4.1 Jets data from ATLAS and CMS 59

4.2 Theoretical calculations 61

4.2.1 Scale choice 61

4.2.2 QCD corrections 62

4.2.3 Electroweak corrections 63

4.3 Results 65

4.3.1 Single-inclusive jets 66

4.3.2 Dijets 69

4.3.3 Single-inclusive jets vs. dijets 71

5 Theoretical error in PDFs determination 73

5.1 Theory error as a covariance matrix 74

5.2 MHOU from scale variations 76

5.3 Construction and validation of a theory covariance matrix 80

5.4 NLO PDFs with missing higher order uncertainties 85

5.5 Usage and delivery 90

6 Fitting the b-quark PDF as a massive-b scheme 93

6.1 The FONLL scheme with parametrized heavy quark PDF in hadronic collisions 96

6.1.1 Perturbative ordering 97

6.1.2 Parametrized-b FONLL 98

6.2 Higgs production in b fusion 102

7 PDFs from lattice data: theoretical framework 106

7.1 Light-cone separation 109

7.2 Spatial separation 114

7.3 Factorization theorem 116

7.3.1 Factorization theorem in position space: small-z_3^2 limit 117

7.3.2 Factorization theorem in momentum space: large P_3 limit ... 118

7.4 Smeared distributions ... 121

7.5 Towards PDFs from lattice data ... 125

8 PDFs from quasi-PDFs matrix elements 127

8.1 quasi-PDFs in QCD ... 128

8.2 Nonsinglet distributions from quasi-PDFs Matrix Elements ... 131

8.2.1 Lattice data ... 131

8.2.2 Systematics in matrix elements of quasi-PDFs ... 134

8.2.3 From parton distributions to lattice observables ... 138

8.3 Fit setting and FastKernel implementation ... 140

8.4 Results ... 143

8.4.1 Closure tests ... 144

8.4.2 Fit results ... 146

9 PDFs from pseudo-Ioffe Time Distribution 149

9.1 Lattice data and observables ... 151

9.2 Fits over lattice data: statistical uncertainties only ... 154

9.3 Systematic effects ... 157

9.3.1 Discussion ... 157

9.3.2 Results ... 161

10 Summary 166

A Statistical estimators 168

A.1 PDF distance 168

A.2 ϕ estimator 169

B Impact of the choice of the correlation model 170

C Alternative points prescription for theory error 172

D Massive-b scheme 175

D.1 Matching coefficients 175

D.2 Massive coefficient functions 176

D.2.1 Leading order 176

D.2.2 Next-to-leading order: $b\bar{b}$-channel 177

D.2.3 Next-to-leading order: bg-channel 181

E Lattice observables 183

E.1 Momentum space factorization 183

E.2 equivalence between pseudo- and quasi-PDF approaches 185

E.3 quasi-PDFs and their moments 186

F PDFs from quasi-PDFs matrix elements: matching coefficient and lattice convolution 189

G PDFs from reduced pseudo-ITD data: Pion Mass dependence for 170 ensemble 194

Introduction

The increasing demand for accuracy required nowadays to perform high-energy phenomenology represents one of the main challenges for the particle theory community. The overall precision of theoretical computations has to match the one of the corresponding experimental measurements: more precise experimental data call for more precise computations, together with a better control and understanding of the different sources of errors affecting theoretical predictions.

The computations of high-energy processes involving nucleons are based on factorization, namely on the separation of amplitudes or cross-sections in different contributions, each of which depends on a specific energy scale. While short-distance (or high-energy) contributions can be computed within the framework of perturbation theory, those related to long-distance phenomena and responsible for the internal structure of the nucleons are not directly accessible by means of first principle computations, and are factorized into universal objects of non perturbative origin, denoted as Parton Distribution Functions (PDFs).

PDFs encode our knowledge about the structure of the nucleons in terms of quarks and gluons, and represent an essential ingredient to perform theory computations in collider physics. Using factorization theorems, PDFs can be extracted from global QCD fits over a set of experimental data, and, thanks to their universality, the results can be subsequently used to compute different processes. Unfortunately, they also represent the dominant source of uncertainty in many important computations, including analyses at the Large Hadron Collider (LHC) and other experiments, the determination of standard model parameters, Higgs boson characterisation and searches for New Physics. It is therefore necessary to

push the determination of PDFs to a new level of accuracy, researching for new methodologies and physical ideas to improve such global analyses.

The problem of PDFs determination involves a number of different lines of research which can be investigated using diverse approaches. For phenomenological applications, it is important to develop frameworks which allow to implement and test different numerical and analytical techniques to deal with complex global fits involving a big number of data coming from experimental measurements. Such frameworks need to be flexible enough in order to fit the available data while imposing known physical constraints which have to be satisfied by the final PDFs, and should allow to easily include new data and computations in the analysis. When new experimental data for specific high-energy processes become available, their impact on parton distributions needs to be studied and quantified, to see how the new experimental information changes our knowledge regarding the structure of the nucleons. This involves, on one side, the understanding of the details of the new data (distributions to be included in the analysis, statistical and systematic errors, kinematic cuts to apply in order to use factorization theorems in their domains of validity), on the other the implementation of the most up to date theoretical computation for the process considered. In order to quantify how reliable the theory predictions are, specific studies regarding the theoretical errors and their propagations into PDFs need to be carried out, and a suitable prescription for the combination of experimental and theoretical uncertainties have to be formulated. Alongside this kind of studies, which are usually performed within the high-energy community, other approaches to study parton distributions are possible. Being non-perturbative objects, PDFs are also a natural subject for a lattice QCD investigation. Starting from the formal definition of parton distributions, given in terms of matrix elements between nucleons states of QCD bilocal operators, it is possible to study the detailed structure of the ultraviolet (UV) and infrared (IR) divergences of such objects, and relate them to specific correlators which, in principle, can be obtained through lattice QCD simulations. Such ideas have been studied and developed in recent years, and data from lattice QCD simulations have started appearing. Given the high interest shown by both the lattice and high-energy community, new data for different lattice observables are likely to appear in the coming years. It is therefore important to understand how to extract information about PDFs from them, study their potential for high-energy phenomenology, their interplay with experimental data and understand the different sources of uncertainties affecting the corresponding lattice simulation.

This thesis can be divided into three main parts. The first part, composed by the first two chapters, is devoted to review the basics of QCD and the main concepts underlying factorization theorems. We will introduce parton distributions first from a phenomenological point of view, following the parton model ideas and discussing the general structure of the NLO QCD corrections, and subsequently following a more formal approach, revising the theoretical definition of PDFs in terms of QCD bilocal operators. These chapters are based on a number of standard QCD text books [1–3], quantum field theories lectures and classical references regarding factorization in high-energy processes, such as refs. [4–6], and will be used to set up the main notations adopted in the rest of the thesis. In the second part, made by chapters 3, 4, 5 and 6, we will present a number of phenomenological studies, as detailed in the following. In chapter 3 we will describe the fitting methodology which has been developed within the NNPDF collaboration and its recent implementation within the new `n3fit` framework, with particular emphasis on the numerical techniques adopted to impose physical constraints on PDFs. We will also discuss the so called fit basis, showing how the final results produced within this framework only depend on the experimental and physical inputs, and not on the specific details of our fitting methodology. Such studies will be part of NNPDF4.0, the next public release of the NNPDF collaboration. In chapter 4 we will discuss the impact of jets data in a global PDFs determination. This gives an example of the kind of analyses which need to be done every time an important class of new experimental measurements is available, together with the corresponding theoretical predictions. The results discussed in this chapter have been first presented in ref. [7]. In chapter 5 we will discuss the definition and implementation of a theoretical error accounting for missing higher orders in a PDFs global analysis. The chapter is based on refs. [8, 9] where this study was first presented by the NNPDF collaboration. In chapter 6, based on ref. [10], we discuss an alternative treatment for the bottom PDFs, analysing the specific case of Higgs production in bottom fusion and its potential impact on precise phenomenology. Finally, in the third part of the thesis, made by chapters 7, 8 and 9, we will present a number of studies to understand the relation between PDFs and lattice computable Euclidean correlators. In particular in chapter 7, based on ref. [11], we will revise the main theoretical ideas behind the lattice approach using a nongauge theory as a simple toy example. This chapter is used to introduce the theoretical framework necessary to understand the studies presented in the two following chapters. In chapter 8 we analyse data for quasi-PDFs matrix elements, studying their potential in constraining

PDFs and presenting a detailed analysis of the systematic uncertainties affecting the data and of how these affect the final results. The results reported in this chapter were first presented in ref. [12]. In chapter 9 we present results for PDFs extracted from a different kind of lattice observables, denoted as reduced Ioffe-time pseudodistribution, assessing differences and similarities with the analysis of chapter 8. The chapter is based on ref. [13]. Finally in chapter 10 we briefly summarize the main results and conclusions of this thesis.

The thesis is supplemented with a number of appendices, where some analyses are further expanded and the details and results of the more technical computations are reported for reference.

1 Basics of QCD

In the early '60s it was generally believed that a theory for the strong interaction could not be formulated within the framework of Quantum Field Theory (QFT) [14]. Despite the remarkable success of Quantum Electrodynamics (QED) in describing phenomena such as the anomalous magnetic moment of the electron, the renormalization process was not completely understood yet and renormalizable quantum field theories were still looked at with suspicion.

The experimental observation of Bjorken scaling [15] in Deep Inelastic Scattering (DIS) experiment (SLAC 1960) suggested that the constituents of nucleons may be described as almost-free and point-like objects when observed with high spatial resolution, leading to the formulation of the parton model [16]. Accordingly, the dynamic of partonic systems should have the property of becoming weaker at shorter distances. In 1973 asymptotic freedom of non-Abelian gauge field theories was discovered [17, 18], making possible to embed the partonic model ideas within the framework of a renormalizable QFT. Soon after it was shown how non-Abelian gauge field theories are actually the only ones exhibiting such property in four dimensional space-time [19] and Quantum Chromodynamics (QCD) emerged as a mathematically consistent theory for the strong interaction.

After its formulation, QCD has been successfully used to describe strong interaction processes observed at colliders, and nowadays it represents one of the cornerstone of the Standard Model. In this chapter we present a brief overview of QCD, recalling some basic features of the theory and introducing our notation. For a more detailed treatment of the basics of QCD we refer to standard QFT and QCD textbook, such as refs. [1–3].

1.1 Lagrangian and its symmetries

Quantum Chromodynamics is a non-Abelian gauge theory based on the gauge group $\mathrm{SU}(3)_{\mathrm{color}}$. The classical Lagrangian of QCD, describing the interaction of N_f massive spin-$\frac{1}{2}$ quarks and massless spin-1 gluons, is given by

$$\mathcal{L}_{\mathrm{classical}} = -\frac{1}{4} F^A_{\mu\nu} F^{A\mu\nu} + \sum_{k=1}^{N_f} \overline{\psi}^k_a \left(i\gamma^\mu D_\mu + m_k\right)_{ab} \psi^k_b \,, \tag{1.1}$$

with the field strength tensor and the covariant derivative defined as

$$F^A_{\mu\nu} = \partial_\mu \mathcal{A}^A_\nu - \partial_\nu \mathcal{A}^A_\mu + g\, f^{ABC} \mathcal{A}^B_\mu \mathcal{A}^C_\nu \,, \tag{1.2}$$

$$D_\mu = \partial_\mu - i\, g\, T^A \mathcal{A}^A_\mu \,. \tag{1.3}$$

The summation over k runs over all the quark flavours, with each quark field ψ^k_a belonging to the fundamental representation of the gauge group $\mathrm{SU}(3)_{\mathrm{color}}$ ($a = 1, 2, 3$), while the gauge field $\mathcal{A}^A_\mu$, called gluon, belongs to the adjoint representation ($A = 1, 2, ..., 8$). The quantities g and f^{ABC} are the gauge coupling and the $\mathrm{SU}(3)_{\mathrm{color}}$ structure constants respectively, and T^A are the eight gauge group generators, satisfying

$$\left[T^A, T^B\right] = i f^{ABC} T^C \,, \tag{1.4}$$

$$\mathrm{Tr}\left[T^A T^B\right] = T_R\, \delta^{AB} \,. \tag{1.5}$$

An explicit expression for the generators T^A in the fundamental representation is given by $(T^A)_{ab} = 1/2 \left(\lambda^A\right)_{ab}$ with λ^A representing the eight 3-dimensional Gell-Mann matrices and with the normalization of the generators conventionally chosen to be $T_R = 1/2$. Given the equations above, the colour matrices obey

$$\sum_A \sum_b T^A_{ab} T^A_{bc} = C_F\, \delta_{ac} \,, \tag{1.6}$$

$$\sum_A \sum_C T^A_{BC} T^A_{CD} = \sum_{A,C} f^{ACB} f^{ACD} = C_A\, \delta_{BD} \,, \tag{1.7}$$

where, considering the generic case of $SU(N)$, $C_F = \frac{N^2-1}{2N}$ and $C_A = N$.

The classical Lagrangian of eq. (1.1) does not allow to formulate quantum perturbation theory in a consistent way. The problem cannot be avoided as long as we rely on a gauge invariant Lagrangian, where the gauge field $\mathcal{A}^A_\mu$ has

the freedom to change according to gauge transformations. We can get rid of such freedom by putting constraints on the field $\mathcal{A}^A_\mu$, known as *gauge fixing conditions*, which in general can be expressed as

$$G^\mu \mathcal{A}^A_\mu = B^A \,, \tag{1.8}$$

with G^μ and B^A chosen in some convenient way [1]. Upon functional integration over the arbitrary quantity B^A, such condition is implemented in the theory by adding to the classical Lagrangian the so-called gauge fixing term

$$\mathcal{L}_{gauge-fixing} = -\frac{1}{2\xi}\left(G^\mu \mathcal{A}^A_\mu\right) \,, \tag{1.9}$$

with ξ representing an arbitrary parameter whose specific value defines the gauge. Different choices for the gauge fixing term can be done. Taking $G^\mu = \partial^\mu$ we obtain a class of covariant gauges. In this case the gauge fixing term must be supplemented by an additional term, known as *ghost Lagrangian* [21], describing a complex scalar field η^a (the Faddeev-Popov ghosts) obeying the Grassmann algebra and belonging to the adjoint representation of the gauge group

$$\mathcal{L}_{ghost} = \left(\partial_\alpha \eta^A\right)^* D^\alpha_{AB}\, \eta^B \,. \tag{1.10}$$

Perturbation theory can be formulated starting from the Lagrangian density

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

Another possible gauge fixing term is the one giving the so-called axial gauges, fixed in terms of a chosen vector n such that $G^\mu = n^\mu$. In this case ghost fields decouples and can thus be ignored, but the explicit form of the gluon propagator turns out to be more complicated than the one in the covariant gauges.

The QCD Lagrangian has a number of important symmetries, both exact and approximate, which is worth recalling here. The classical Lagrangian given in eq. (1.1) is invariant under $\mathrm{SU(3)_{color}}$ gauge transformations. After gauge fixing gauge invariance is broken, but the resulting Lagrangian of eq. (1.11) satisfied the BRST symmetry [22, 23], which in turn helps with the renormalizability of the theory. The so-called flavour symmetries are also exact symmetries of

[1] Eq. (1.8) in general admits various solutions, representing different possible gauge configurations known as Gribov copies. Their occurrence is known as the *Gribov copies problem*, or *Gribov ambiguity* [20].

QCD, acting through a global phase transformation of each quark field separately and giving the baryon number conservation. Other symmetries include the discrete global symmetries of parity and time reversal invariance. Finally charge-conjugation is also an exact symmetry of eq. (1.11).

Assuming mass degeneracy for the up and down quarks, the U(1) global symmetry associated with quark number can be extended to a global U(2) = U(1) $\otimes$ SU(2). The new symmetry SU(2) is known as isospin symmetry. We can further enhance the symmetry group to U(1) $\otimes$ SU(3) assuming the strange quark to be also degenerate in mass with the up and the down[2]. In the case of massless quarks, a chiral symmetry $\mathrm{U}(2)_L \otimes \mathrm{U}(2)_R = \mathrm{SU}(2)_V \otimes \mathrm{U}(1)_V \otimes \mathrm{SU}(2)_A \otimes \mathrm{U}(1)_A$ holds, which however is spontaneously broken to $\mathrm{SU}(2)_V \otimes \mathrm{U}(1)_V \otimes \mathrm{U}(1)_A$, with the subscripts V and A denoting the vector and axial combinations. The three pseudo-scalar Goldstone bosons resulting from chiral SU(2) breaking to $\mathrm{SU}(2)_V$ are identified with the three pions π^+, π^- and π^0 in the massless quark limit. While the survived $\mathrm{SU}(2)_V \otimes \mathrm{U}(1)_V$ symmetry is identified with the isospin and baryon number conservation mentioned previously, the remaining $\mathrm{U}(1)_A$, despite not being spontaneously broken, appears to be lost in QCD. The study of what happens to this symmetry is known as the U(1)-problem [25]. The axial symmetry $\mathrm{U}(1)_A$ is actually broken at the quantum level, through the Adler-Bell-Jackiw anomaly, which induces a new term in the QCD Lagrangian proportional to $\epsilon_{\alpha\beta\gamma\delta} \operatorname{Tr} F^{\gamma\delta} F^{\alpha\beta}$. This term would be responsible for a violation of CP in the strong sector and its magnitude is given by the size of the parameter θ, representing an angular variable whose values is fixed to be in the range $\theta < 10^{-9}$ by experimental measures. The unexplained smallness of such parameter is known as the strong CP-problem. Among the proposed solutions, the Peccei-Quinn mechanism was developed [26] which, together with the introduction of additional particles called axions [27, 28], proposes a dynamical explanation for the θ values.

1.2 The running coupling and asymptotic freedom

In analogy with the QED fine structure constant, the QCD coupling constant is defined in terms of the gauge coupling as $\alpha_s = g^2/4\pi$. As a consequence of the renormalization process, the coupling acquires a dependence on the renormalization scale μ, namely the arbitrary scale at which the subtraction of

[2]Such approximate flavour symmetry SU(3) is the basis of the Gell-Mann quark model [24] which was proposed well before the birth of QCD.

the ultraviolet (UV) poles is performed. The resulting renormalization group equations read

$$\mu^2 \frac{\partial \alpha_s}{\partial \mu^2} = \beta\left(\alpha_s\right) . \tag{1.12}$$

The β function can be computed in perturbation theory as a power expansion in α_s. Nowadays results up to five loops have been computed [29]. At next-to-leading order it is given by

$$\beta\left(\alpha_s\right) = -\beta_0 \alpha_s^2 \left(1 + \beta_1 \alpha_s + \mathcal{O}\left(\alpha_s^2\right)\right) , \tag{1.13}$$

with

$$\beta_0 = \frac{33 - 2N_f}{12\pi} , \quad \beta_1 = \frac{153 - 19N_f}{2\pi\left(33 - 2N_f\right)} . \tag{1.14}$$

Using the leading order expression for the β in eq. (1.12) we find the leading-log solution for the running coupling, relating its value at the generic scale Q^2 to the one at a reference scale μ^2

$$\alpha_s\left(Q^2\right) = \frac{\alpha_s\left(\mu^2\right)}{1 + \alpha_s\left(\mu^2\right)\beta_0 \log\frac{Q^2}{\mu^2}} . \tag{1.15}$$

Given the positive sign of β_0, from eq. (1.15) it is evident how, as the scale Q^2 becomes very large, the coupling $\alpha_s\left(Q^2\right)$ decreases to zero. This property, which characterizes non-Abelian gauge theories like QCD, is known as asymptotic freedom and the theory is then said to be asymptotically free. It is customary to introduce a dimensionful parameter directly in the definition of α_s, usually denoted as Λ. It can be defined as

$$\log\frac{\mu^2}{\Lambda^2} = -\int_{\alpha_s(\mu^2)}^{\infty} \frac{dx}{\beta\left(x\right)} , \tag{1.16}$$

with its specific value depending on the choice of the renormalization scheme, on the order of the β function power expansion and on the number of active flavours[3] entering the theory. Any dimensionful quantity in QCD can be expressed in units of Λ. It can be thought as an intrinsic scale of QCD: at scales much larger than Λ the coupling α_s is small and quarks behave as almost free particles.

[3]The notion of active flavours will be discussed in sec. 1.3

1.3 Quark masses

The quark masses represent another parameter of the Lagrangian eq. (1.1). Just like the coupling constant because of the renormalization process they also acquire a dependence on a renormalization scale μ given by

$$\mu^2 \frac{\partial m}{\partial \mu^2} = -\gamma_m \left(\alpha_s\right) m \left(\mu^2\right) \ . \tag{1.17}$$

The quantity γ_m is the mass anomalous dimension which can be computed in perturbation theory as a power expansion in α_s

$$\gamma_m \left(\alpha_s\right) = c\, \alpha_s \left(1 + c' \alpha_s + ...\right) \ , \tag{1.18}$$

with the explicit expression of the coefficients depending on the choice of the renormalization scheme.

In order to study the impact of the quarks mass on a generic physical observable R, consider the situation in which we have only one quark with mass m. Writing $R = R\left(Q^2/\mu^2, \alpha_s\left(\mu^2\right), m\left(\mu^2\right)/Q\right)$ and setting the renormalization scale equal to the physical scale $\mu = Q$, if we assume the first N derivatives of R to be finite in $m = 0$ we can write the expansion

$$\begin{aligned} R\left(1, \alpha_s\left(Q^2\right), m\left(Q^2\right)/Q\right) &\sim R\left(1, \alpha_s\left(Q^2\right), 0\right) \\ &+ \sum_{n=1}^{N} \frac{1}{n!} \left[\frac{m\left(Q^2\right)}{Q^2}\right]^n R^{(n)}\left(1, \alpha_s\left(Q^2\right), 0\right) \ . \end{aligned} \tag{1.19}$$

where $R^{(n)}$ denotes the n-th derivative of R with respect to its third argument. Given the fact that eq. (1.17) leads to a change of the renormalized mass with Q which is at most logarithmic, from eq. (1.19) it is clear how, when considering high energy scales $Q \gg m$, the mass dependence can be dropped, and the quark can be considered massless. On the other hand, when the mass of the quark is much greater than the relevant scale Q it can be shown that the heavy quark mass correction to R are suppressed by inverse powers of m, and therefore they can be ignored when $Q \ll m$. The n_l active flavour introduced in the previous section are the light quarks whose mass is much smaller than the physical scale Q.

Considering the situation in which we have n_l light quarks (i.e. quarks whose mass

is much smaller than Q) and a single heavy quark with mass m, the two values $\alpha_s^{(n_l+1)}$ and $\alpha_s^{(n_l)}$ measured in the two domains $Q \gg m$ and $Q \ll m$ respectively, are usually matched through matching equation of the form

$$\alpha_s^{(n_l+1)}\left(Q^2\right) = \alpha_s^{(n_l)}\left(Q^2\right) + \sum_{k=2}^{\infty} c_k\left(L\right)\left(\alpha_s^{(n_l)}\left(m^2\right)\right)^k , \tag{1.20}$$

where the coefficients $c_k\left(L\right)$ are polynomials in $L = \log Q^2/m^2$ and at the scale $m^2 = Q^2$ the $\mathcal{O}\left(\alpha_s^2\right)$ coefficient vanishes, $c_2\left(0\right) = 0$. Depending on the specific energy scale of interest, one can perform the computation of a generic physical observable considering either n_l or $n_l + 1$ active flavours, each choice being more convenient in a given kinematic region. In sec. 2.4 we will discuss a way in which such computations can be combined in a unique result which is accurate at every energy scale.

1.4 Perturbative and non-Perturbative approaches

The property of asymptotic freedom allows to compute high energy scattering processes as an expansion in the coupling, paving the way to a perturbative treatment of QCD. The Lagrangian given in eq. (1.11) can be written as the sum between the free Lagrangian $\mathcal{L}_0$, describing free fermions and gauge fields, plus an interaction term $\mathcal{L}_I$, containing all the terms proportional to the gauge coupling g: at high energy g becomes small, and all these terms can be treated as perturbative interactions, so that the corresponding contribution in the action can be expanded in a power series of the coupling.

In general, perturbation theory can be seen as a systematic way of approximating the solution of a quantum field theory keeping the error under control. Although its success in describing high energy processes, it is not a full solution of the theory, and there are situations in which it cannot be applied: in the case of QCD, at low energy the coupling becomes large, and a power expansion in α_s is no longer possible. It is therefore important to have a non-perturbative formulation of QCD, based on the classical Lagrangian of eq. (1.1). The framework of lattice QFT represents one of the most studied and developed systematic approaches to study quantum field theories in non-perturbative regimes. In the following we recall the basic ideas underlying the formulation of QCD on an Euclidean lattice, referring to standard textbooks as ref. [30] for a complete discussion.

In lattice QFT the path integral of the theory is defined on a discrete and finite Euclidean space-time, characterized by a finite volume and lattice spacing a, and directly evaluated through Monte-Carlo simulations. The lattice can be defined as a cartesian product

$$\Lambda^4(N) = a\Big(\llbracket 0, N_0 - 1\rrbracket \times \llbracket 0, N_1 - 1\rrbracket \times \llbracket 0, N_2 - 1\rrbracket \times \llbracket 0, N_3 - 1\rrbracket\Big), \tag{1.21}$$

where N is a four vector with integer components and $\llbracket 0, n\rrbracket$ is the set of all the integers j such that $0 \leq j \leq n$. The integer N_μ represents the number of lattice sites in the μ direction, corresponding to a space-time extent equal to aN_μ. From this definition it is clear how for each point x_μ belonging to the lattice there exists a four vector j_μ with integer coordinates such that $x_\mu = aj_\mu$. The zero-component N_0 is usually identified with the temporal extent $T = aN_0$, while the three remaining ones, assumed to be equal, represent the spatial extent in the three spatial directions $L = aN_1 = aN_2 = aN_3$. On such lattice the full Lorentz invariance of the continuum Minkowski space-time is reduced to the hypercubic group, however when considering gauge theories their lattice version is built in such a way to preserve gauge invariance.

Considering for simplicity a theory containing a single scalar field ϕ, taking a generic correlation function in Minkowski space

$$C_n^{(M)}(x_1, ..., x_n) = \frac{1}{Z}\int \mathcal{D}\phi\, \phi(x_1) ... \phi(x_n) \exp\left(iS^{(M)}[\phi]\right), \tag{1.22}$$

one can define the associated Euclidean correlation function by performing a Wick rotation $x_i = (x_i^0, \vec{x}_i) \to \bar{x}_i = (-ix_i^0, \vec{x}_i)$

$$\begin{aligned} C_n^{(E)}(x_1, ..., x_n) &\equiv C_n^{(M)}(\bar{x}_1, ..., \bar{x}_n) = \\ &\frac{1}{Z}\int \mathcal{D}\phi\, \phi(x_1) ... \phi(x_n) \exp\left(-S^{(E)}[\phi]\right), \end{aligned} \tag{1.23}$$

where $S^{(E)}$ is the Euclidean action of the theory. Such Wick rotation effectively transform the Minkowski metric $\eta_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$ into the positive Euclidean metric $\delta_{\mu\nu} = \mathrm{diag}(1, 1, 1, 1)$ [4].

The functional integral of eq. (1.23) can be estimated by averaging the fields

[4] The precise conditions under which the Wick rotation works are stated by the Osterwalder-Schrader axioms [31]. These are a list of properties that correlation functions in Euclidean space have to satisfy to be the analytic continuation of the correlation functions of the original QFT in Minkowski space.

product $\phi(x_1)...\phi(x_n)$ on the probability density $D\phi \exp\left(-S^{(E)}[\phi]\right)$.

In order to write a discrete version of QCD, we need to construct an action for the gauge field and one for the quark fields. In the case of the gauge field, the action can be written in terms of gauge links $U_\mu(x)$, obtaining the so-called Wilson action

$$S_G(U_\mu) = \frac{\beta}{N} \sum_{x \in \Lambda^4} \sum_{\mu > \nu} \mathrm{Re}\left(1 - P_{\mu\nu}(x)\right) , \tag{1.24}$$

$$P_{\mu\nu}(x) = U_\mu(x)\, U_\nu(x + a\hat{\mu})\, U_\mu(x + a\hat{\nu})^\dagger\, U_\nu(x)^\dagger , \tag{1.25}$$

where β is a constant of the theory. It can be shown that in the continuum limit, using $\beta = 2N/g^2$, eq. (1.24) recovers the Euclidean Yang-Mills action.

Considering the fermionic fields, defining the translation operator in the $\hat{\mu}$ direction as $\tau_\mu f(x) = f(x + a\hat{\mu})$, it turns out that a naive discretization of the Dirac action

$$S_{\mathrm{Dirac}}\left[\psi, \bar{\psi}\right] = a^4 \sum_{x \in \Lambda^4} \bar{\psi}(x) \left(\frac{\gamma^\mu}{2a} \left(\tau_\mu - \tau_{-\mu}\right) + m \right) \psi(x) \tag{1.26}$$

would lead to a theory with the wrong continuum limit, describing 16 independent fermion states all having the same energy. This is the so-called *doubling problem*, and it is a consequence of a more general property of fermionic actions known as the Nielsen-Ninomiya no-go theorem [32]. A possible solution to the doubling problem was proposed by Wilson [33], through the addition of another term to the fermionic action to remove the unwanted states

$$S_W\left[\psi, \bar{\psi}\right] = S_{\mathrm{Dirac}}\left[\psi, \bar{\psi}\right] - \frac{a^5}{2} \sum_{x \in \Lambda^4} \bar{\psi}(x)\, \tilde{\delta}^2 \psi(x) , \tag{1.27}$$

where $\tilde{\delta}^2 = a^{-2} \sum_\mu \left(1 - \tau_{-\mu}\right)\left(\tau_\mu - 1\right)$. In the presence of a gauge field, eq. (1.27) becomes

$$S_W\left[\psi, \bar{\psi}, U_\mu\right] = a^4 \sum_{x \in \Lambda^4} \bar{\psi}(x) \left(K[U_\mu] + m\right) \psi(x) , \tag{1.28}$$

where $K[U_\mu]$ is a suitable discretization for the covariant derivative which implements the Wilson prescription. The addition of the Wilson term in the action solves the doubling problem, however it explicitly breaks chiral symmetry, leading to a more complex UV structure of the theory. An example of this appears

when looking at the renormalization of the fermion mass: when a Wilson fermion is coupled to some gauge interaction the mass shift, which would be zero in the continuum and in the case of naive lattice fermions, is not zero anymore. This means that a Wilson fermion with zero bare mass is actually not massless. Its mass is denoted as *critical mass*.

Once the Euclidean action has been defined, the expectation value of a generic observable $O\left[\psi, \bar{\psi}, U_\mu\right]$ can be evaluated through the path integral

$$\langle O \rangle = \frac{1}{Z} \int D\psi \, D\bar{\psi} \, DU_\mu \, O\left[\psi, \bar{\psi}, U_\mu\right] e^{-S[\psi, \bar{\psi}, U_\mu]} \,. \tag{1.29}$$

with

$$S[\psi, \bar{\psi}, U_\mu] = S_G\left(U_\mu\right) + S_W\left[\psi, \bar{\psi}, U_\mu\right] \,. \tag{1.30}$$

Since the fermionic action is quadratic, the functional integrals over ψ and $\bar{\psi}$ can be performed analytically getting

$$\langle O \rangle = \frac{1}{Z} \int DU_\mu \det\left(K + M\right) O_{\text{Wick}}\left[U_\mu\right] e^{-S_G[U_\mu]} \,, \tag{1.31}$$

where O_{Wick} denotes the functional obtained from O performing the Wick contractions of the quark fields. Note that this observable depends on the quark propagator $(K + M)^{-1}$, appearing for every $\psi\bar{\psi}$ contraction. Inverting the term $K + M$ in order to obtain the fermionic propagator represents the elementary computation in lattice QCD simulations. Assuming $\det\left(K + M\right) > 0$ [5], the quantity

$$\frac{1}{Z} \det\left(K + M\right) e^{-S_G[U_\mu]} \,, \tag{1.32}$$

can be interpreted as a probability distribution and an estimation for $\langle O \rangle$ up to a $\mathcal{O}\left(1/\sqrt{N}\right)$ statistical error can be obtained by drawing N samples $U_\mu^{(i)}$ from it and computing

$$\langle O \rangle = \frac{1}{N} \sum_{i=0}^{N} O\left[U_\mu^{(i)}\right] \,. \tag{1.33}$$

[5] In a continuum theory it can be shown that this is true thanks to chiral symmetry, as long as we consider non-zero masses. On the lattice the situation can be more complicated: in the case of Wilson fermions for example chiral symmetry is lost, and the determinant is positive only when the bare masses are bigger than the critical mass.

2

Factorization theorems in QCD

In the previous chapter we have seen how, thanks to asymptotic freedom, the structure of QCD simplifies when considering problems involving short-distance or high-energy scales: the coupling becomes small and the theory can be solved perturbatively. However, cross sections for high energy processes are usually a combination of short- and long-distance effects, and cannot be fully computed within the framework of perturbation theory. Factorization theorems allows to separate (factorize) a cross section or amplitude in separate contributions, each factor containing the dependence of the process on a specific distance scale: while short-distance effects can be computed in perturbation theory, contributions coming from long-distance phenomena have to be extracted from experimental data. In this chapter, starting from the Parton Model ideas we will discuss factorization theorems in QCD, using the case of Deep Inelastic Scattering (DIS) as a basic example. We will then introduce the main subject of the present work, namely Parton Distribution Functions (PDFs).

2.1 The Parton Model

The basic ideas underlying factorization theorems for high energy processes can be described appealing to Feynman's parton model [16, 34, 35]. In this picture fundamental particles called *partons* are bound together to form hadrons. Since the details of the partonic system are unknown, the scattering between a test particle and the hadron as a whole cannot be computed. However we assume that we do know how to describe the scattering with a free parton. For example, let us consider the case of the scattering of a high-energy charged lepton off a

hadron target

$$e^-(k)\, H\,(P)\ \to e^-(k')\, X\,.$$

Such process is known as Deep Inelastic lepton-hadron Scattering (DIS). Looking at this scattering in the centre of mass frame, the hadron will be Lorentz contracted in the direction of the collision and the lifetime of the internal partonic states will be lengthened. As a consequence, the time the electron takes to cross the hadron will be much shorter than the average lifetime of each partonic state. During the time of the interaction the hadron can then be thought as "frozen" in a well defined partonic state, with each parton carrying a definite fraction ξ of the hadron momentum and not interacting with the other ones. If the energy of the collision is high enough, the virtual photon mediating the electron-hadron interaction will interact with a single parton having a given momentum fraction. Likewise, interactions occurring in the final states are assumed to occur on time scales too long to interfere with the hard scattering. Given this picture, it is natural to think about the scattering cross section as classical and incoherent, namely as a sum of probabilities rather than of amplitudes. The parton model ideas can be summarized in the simple formula

$$\begin{aligned}\sigma\left(e^-\,(k)\, H\,(P) \to e^-\,(k') + X\right) = \\ \sum_i \int_0^1 d\xi\, q_{i/H}\,(\xi)\, \hat{\sigma}\left(e^-\,(k)\, q_i\,(\xi P) \to e^-\,(k') + q_f\right)\end{aligned} \tag{2.1}$$

where $\hat{\sigma}$ represents the partonic cross section describing the interaction between a single free parton and the virtual photon, and the set of functions $q_{i/H}\,(\xi)$ are the probability densities of having a parton of kind i inside the hadron H, carrying a fraction $\xi \in (0,1)$ of the total hadron momentum.

In the following we recall in more details how parton model ideas apply to the case of DIS, setting the stage for a more complete discussion of factorization which will be addressed in the next section. DIS experiments are traditionally the main testing ground of perturbative QCD, having been the first processes where pointlike particles were seen inside the hadron, thus motivating the formulation of the parton model. They play a central role in any discussion regarding factorization and provide a simple experimental and theoretical framework to study the strong interaction. As we are going to recall in the following, measurements of DIS structure functions directly probe the structure of hadrons, giving the bulk of experimental measurements at the basis of every phenomenological determination of Parton Distribution Functions.

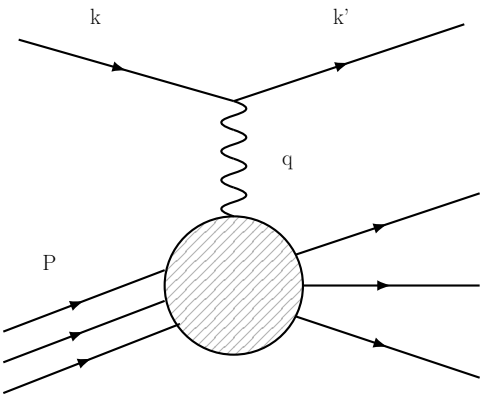

Figure 2.1 *DIS kinematic.*

The kinematic for DIS is reported in fig. 2.1. We specify the discussion to the case of neutral current (NC) unpolarized scattering, considering only the contribution associated to photon exchange (which is the dominant one as long as we consider energy scales below m_Z^2). The space-like momentum of the photon is given by $q = k - k'$, the centre of mass energy square is $s = (P + k)^2$ and we denote the invariant mass square of the final states as $W = (P + q)^2$. It is customary to introduce the kinematic variables

$$Q^2 = -q^2 > 0\,, \quad x = \frac{Q^2}{2P \cdot q}\,, \quad y = \frac{Q^2}{xs}\,. \tag{2.2}$$

In the regime

$$Q^2,\, W^2 \gg m^2_{\text{hadron}} \sim \Lambda^2_{\text{QCD}}\,,$$

leptons and quarks masses can be neglected. It is easy to see that the variable x, known as *Bjorken variable* can take values between 0 and 1, with $x \to 1$ representing the elastic limit in which $W = m^2_{\text{hadron}}$. The Deep Inelastic regime is then defined as $Q^2 \gg \Lambda^2_{\text{QCD}}$ with x fixed and different from 1.

The amplitude corresponding to the diagram in fig. 2.1 reads

$$\mathcal{M} = \frac{e}{q^2} \bar{u}\,(k')\,\gamma^\alpha u\,(k)\,\langle X|j_\alpha\,(0)\,|P\rangle, \tag{2.3}$$

where $|X\rangle$ represents the generic final state with n on-shell particles and j_α is the electromagnetic current through which the photon interacts in the proton. The cross section, which is proportional to the amplitude square, is then found to be

proportional to the product between a leptonic and an hadronic part

$$\frac{d\sigma}{dx\,dQ^2} \propto \int \frac{d^3k'}{2E_{k'}\,(2\pi)^3}\, W^{\mu\nu} L_{\mu\nu}\,. \tag{2.4}$$

The leptonic tensor $L_{\mu\nu}$ can be easily computed within QED, while the hadronic one, containing the information about the interaction between the virtual photon and the hadron, can be parameterized as

$$\begin{aligned} W^{\mu\nu}\,(P,q) = &-\left(g^{\mu\nu} - \frac{q^\mu q^\nu}{q^2}\right) F_1\left(x, Q^2\right) + \\ &\frac{1}{P\cdot q}\left(P^\mu - q^\mu \frac{P\cdot q}{q^2}\right)\left(P^\nu - q^\nu \frac{P\cdot q}{q^2}\right) F_2\left(x, Q^2\right), \end{aligned} \tag{2.5}$$

F_1 and F_2 being scalar functions, called *structure functions*, depending on the invariant quantities of the problem, namely x and Q^2. If, more generally, we allow j_α to be any electroweak current, there will be an additional parity-violating structure function F_3.

Computing explicitly the leptonic tensor and plugging eq. (2.5) in eq. (2.4) we can derive a general expression for the double differential cross section of DIS in the Center of Mass frame

$$\frac{d\sigma}{dx\,dQ^2} = \frac{2\pi\alpha^2}{Q^4}\left[\left[1+(1-y)^2\right] F_T\left(x,Q^2\right) + \frac{2\,(1-y)}{x} F_L\left(x,Q^2\right)\right], \tag{2.6}$$

where $\alpha = e^2/(4\pi)$ is the fine structure constant and the transverse and longitudinal structure functions are defined as

$$F_L = F_2 - 2xF_1\,, \quad F_T = 2F_1\,. \tag{2.7}$$

The partonic cross section $d\hat{\sigma}$ for the scattering of the lepton off a single parton with momentum ξP can be computed through a simple leading order QED computation ($e^- + q \to e^- + q$) getting

$$\frac{d\hat{\sigma}}{dx\,dQ^2} = \frac{4\pi\alpha^2}{Q^4}\left[1+(1-y)^2\right]\frac{1}{2} e^2 \delta\left(x-\xi\right), \tag{2.8}$$

from which we can read the expressions for the partonic structure functions $\hat{F}_1$

and $\hat{F}_2$

$$\hat{F}_2 = xe^2\delta\left(x-\xi\right) = 2x\hat{F}_1\,. \tag{2.9}$$

Finally, using the parton model assumption of eq. (2.1), we can write an explicit expression for the full structure functions

$$F_L\left(x,Q^2\right) = 0\,, \quad F_2\left(x,Q^2\right) = x\sum_a e_a^2\, q_{a/H}\left(x\right)\,. \tag{2.10}$$

Eq. (2.10) makes manifest how DIS experiments probe the structure of the incoming hadron H, giving direct access to the functions $q_{a/H}\left(x\right)$ encoding its internal distribution of quarks and gluons. Also, it shows explicitly how in the parton model the structure functions do not depend on the energy scale. Such property is known as Bjorken scaling, and its experimental observation was taken as a confirmation of the composite nature of hadrons, confirming the basic ideas of the parton model. From the analysis of DIS experimental data within the framework of the parton model the following picture emerged: the proton is composed by two u-valence quarks

$$u_v\left(x\right) = u\left(x\right) - \bar{u}\left(x\right)\,,$$

and one d-valence quark

$$d_v\left(x\right) = d\left(x\right) - \bar{d}\left(x\right)\,,$$

which carry the proton electric charge and baryon number, plus a *sea* of light quarks made of $q\bar{q}$ pairs. Additionally, the following integrals, known as valence sum rules, are satisfied

$$\int_0^1 dx\, u_v\left(x\right) = 2\,, \quad \int_0^1 dx\, d_v\left(x\right) = 1\,. \tag{2.11}$$

Experimentally it was observed that only about the 50% of the proton's momentum is carried by quarks

$$\sum_q \int_0^1 dx\, x\left[q\left(x\right) + \bar{q}\left(x\right)\right] \simeq 0.5\,, \tag{2.12}$$

with the remaining momentum fraction associated to gluons.

2.2 Improved Parton Model and factorization of collinear singularities

In this section, starting from the parton model ideas, we briefly recall how to include Next-to-Leading-Order (NLO) QCD corrections. As we are going to see, when considering QCD corrections two important things happen. First Bjorken scaling is broken, namely the structure functions acquire a non trivial scale dependence. Second, when considering gluons emissions from the initial state particles, infrared collinear singularities arise. The universal factorization of such collinear poles and the subsequent renormalization of parton distribution functions represent the main conceptual point of factorization in high energy processes, and will be discussed in the following.

2.2.1 Next-to-Leading-Order QCD corrections

Considering a generic structure function F, following the parton model ideas we can write it as

$$F\left(x,Q^2\right) = \sum_a \int_x^1 \frac{d\xi}{\xi}\, q_{a/H}^{(0)}\left(\xi\right)\, \hat{F}_a\left(\frac{x}{\xi},Q^2\right)\,, \tag{2.13}$$

with $\hat{F}_a$ representing the partonic level structure function for the scattering of a quark off the virtual photon, and $q_{a/H}^{(0)}$ denoting the parton model PDFs [1]. Such formula is valid up to power corrections, namely up to further terms of non-pertubative origin which are suppressed by powers of Λ_{QCD}^2/Q^2.

From the previous section we know that at Born level $\hat{F}_a$ is proportional to a delta function $e^2\delta\left(1-x\right)$. Considering QCD correction of order α_s, initial and final states emissions of a single gluon, corresponding to Feynman diagrams of fig. (2.2) have to be computed. Accounting also for the corresponding virtual corrections on the quark legs and for the gluon induced vertex correction, the full

[1] Although here we are introducing this formula starting from the ideas of the parton model, eq. (2.13) can be proved in the Bjorken limit, and the bare PDF $q_{a/H}^{(0)}$ can be defined in terms of a bilocal operator matrix element. We will get back to this point in the next sections

result has the following form

$$\hat{F}\left(x,Q^2\right)=e^2\left[\delta\left(1-x\right)+\alpha_s\left(P\left(x\right)\log\frac{Q^2}{Q_0^2}+R\left(x\right)\right)\right], \tag{2.14}$$

with

$$P\left(x\right)=\frac{C_F}{2\pi}\left[\frac{1+x^2}{\left(1-x\right)_+}+\frac{3}{2}\delta\left(1-x\right)\right] \tag{2.15}$$

with the action of the plus distribution over a generic test function $g\left(\xi\right)$ defined as

$$\int_0^1 dx\,\frac{1}{\left(1-x\right)_+}g\left(\xi\right)=\int_0^1 dx\,\frac{1}{\left(1-x\right)}\left[g\left(\xi\right)-g\left(1\right)\right]. \tag{2.16}$$

Two observations can be done. Firstly, as anticipated above, beyond leading order Bjorken scaling is broken by logarithms of Q^2, and the structure function acquires a Q^2 dependence. Secondly, eq. (2.14) contains a logarithmic dependence on an infrared cutoff Q_0^2, pointing out the presence of an infrared divergence.

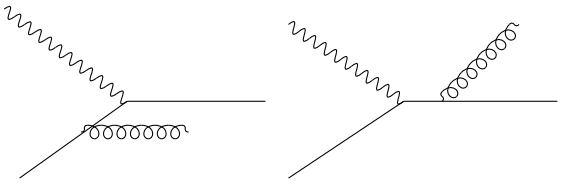

Figure 2.2 *Real NLO QCD corrections.*

Working in a light-cone gauge, such logarithmic divergence can be traced back to the square of the amplitude associated to a gluon emission from the initial state quark. It can be shown that, denoting as $k_\perp$ the longitudinal momentum of the emitted gluon, we end up with a contribution of the form

$$\hat{F}_{q\,\gamma\to q\,g}\left(x,Q^2\right)=\int^{Q^2}\frac{dk_\perp^2}{k_\perp^2}\,\alpha_s\,P\left(x\right)+\ldots, \tag{2.17}$$

where the ellipses stand for finite regular terms. It is clear from eq. (2.17) that such term diverges in the region of small-$k_\perp$. In order to regulate such pole we can introduce the infrared cutoff Q_0^2, getting the logarithmic contribution observed in eq. (2.14). Similarly, when considering multiple gluons emissions from initial state particles, terms of the kind $\left(\alpha_s\left(Q^2\right)\log\frac{Q^2}{Q_0^2}\right)^n$ show up. Since all these terms are of order 1, if we accounted for only some of them we would spoil perturbation theory. In order to get a proper perturbative expansion such terms

have to be resummed at all orders. Such resummation is achieved factorizing the collinear singularities into the parton model PDFs, and solving the resulting renormalization group equations as described in the next sections.

2.2.2 Factorization of collinear singularities

The singularities described in the previous sections arise from the kinematic region where $k_\perp \to 0$, namely when a gluon is emitted parallel to an initial state quark. For this reason they are often called collinear singularities. To understand how to deal with such terms, one needs to realize that the limit of small$-k_\perp$ corresponds to the long-range (low energy) regime of the strong interaction and therefore cannot be treated within perturbation theory. We can then consider the parton distributions introduced through the parton model as bare, unmeasurable quantities, and use them to reabsorb the collinear singularities. In this way, all the dependence on low energy phenomena can be factorized in the parton distribution functions, leaving the hard cross sections free from collinear singularities.

Starting from the logarithmic divergent contribution appearing in eq. (2.14), we can introduce an additional unphysical scale μ_F and write $\log \frac{Q^2}{Q_0^2} = \log \frac{Q^2}{\mu_F^2} + \log \frac{\mu_F^2}{Q_0^2}$. Looking back at eq. (2.13), the infrared divergent partonic structure function can then be written as

$$\hat{F}\left(\xi, Q\right) = \int_\xi^1 \frac{d\eta}{\eta} \Gamma\left(\frac{\xi}{\eta}, \mu_F\right) \hat{F}_{\text{reg}}\left(\eta, \frac{Q}{\mu_F}\right), \tag{2.18}$$

with

$$\Gamma\left(y, \mu_F\right) = \delta\left(1-y\right) + \alpha_s \left[P\left(y\right) \log \frac{\mu_F^2}{Q_0^2} + \Gamma_{finite}\left(y\right)\right], \tag{2.19}$$

$$\hat{F}_{\text{reg}}\left(\eta, \frac{Q}{\mu_F}\right) = \delta\left(1-\eta\right) + \alpha_s \left[P\left(\eta\right) \log \frac{Q^2}{\mu_F^2} + R\left(\eta\right) - \Gamma_{finite}\left(\eta\right)\right]. \tag{2.20}$$

The new scale μ_F introduced above, often referred to as factorization scale, separates long and short distance contributions: everything which is below μ_F is considered to be in a non perturbative regime and it is factorized in the kernel Γ, which therefore contains the infrared poles. The term Γ_{finite} represents finite contributions that can be kept into the subtraction kernel rather than in the hard structure function. Its specific choice is what defines the renormalization scheme.

Substituting eq. (2.18) in eq. (2.13) it is easy to see that we can write

$$F\left(x,Q\right)=\sum_{a}\int_{x}^{1}\frac{d\eta}{\eta}\,q_{a/H}\left(\eta,\mu_{F}\right)\hat{F}_{\mathrm{reg}}\left(\frac{x}{\eta},\frac{Q}{\mu_{F}}\right), \tag{2.21}$$

where the renormalized quark PDFs $q_{a/H}$ is defined as

$$q_{a/H}\left(x,\mu_{F}\right)=\int_{x}^{1}\frac{d\eta}{\eta}\,q_{a/H}^{(0)}\left(\frac{x}{\eta}\right)\Gamma\left(\eta,\mu_{F}\right). \tag{2.22}$$

The collinear poles are therefore factorized from the hard scattering structure function and reabsorbed into the PDFs, following a procedure symilar to the one used for UV renormalization. As a consequence PDFs acquire a non trivial dependence on an unphysical scale μ_F, which will be further described in in the next section.

To sum up, considering higher orders QCD corrections, the DIS structure functions can be written as

$$\begin{aligned}F\left(x,Q^{2}\right)&=\sum_{a}\int_{x}^{1}\frac{d\xi}{\xi}\,C_{a}\left(\frac{x}{\xi},\frac{Q^{2}}{\mu_{F}^{2}},\alpha_{s}\right)q_{a/H}\left(\xi,\mu_{F}^{2}\right)+\mathcal{O}\left(\frac{\Lambda_{QCD}^{2}}{Q^{2}}\right)\\&\equiv\sum_{a}C_{a}\left(\frac{Q^{2}}{\mu_{F}^{2}},\alpha_{s}\right)\otimes q_{a/H}\left(\mu_{F}^{2}\right)+\mathcal{O}\left(\frac{\Lambda_{QCD}^{2}}{Q^{2}}\right),\end{aligned} \tag{2.23}$$

where in the second line we have denoted as $\otimes$ the convolution operation. The coefficients functions C_a appearing in eq. (2.23) correspond to the finite partonic structure functions $\hat{F}_{\mathrm{reg}}$ after renormalization and subtraction of collinear singularities. Their explicit expression will depend on the specific structure function under consideration and on the choice for the renormalization and factorization schemes, defined when removing the UV and collinear singularities respectively. Once properly defined they can be computed order by order in perturbation theory as an expansion in the strong coupling

$$C_{a}\left(x,\alpha_{s}\right)=C_{a}^{(0)}\left(x\right)+\alpha_{s}\,C_{a}^{(1)}\left(x\right)+\alpha_{s}^{2}\,C_{a}^{(2)}\left(x\right)+\ldots, \tag{2.24}$$

where the first contribution (LO) recover the parton model predictions, the second one (NLO) corresponds to the QCD corrections discussed above and the coming ones ($\mathrm{N^{n}NLO}$) will refer to higher order corrections. Differently from the initial formula of eq. (2.13), which was written in analogy to the parton model formulation, the parton distributions have now acquired a scale dependence, which cancel against an analogue dependence in the coefficient functions, leaving

the physical structure function independent from any unphysical scales. Also, even if in our discussion we have only considered the quark channel, the sum over the flavour types a now includes also gluon initiated contributions, which formally start at NNLO.

So far we have discussed factorization for processes with only one hadron in the initial state, but the same ideas and logic apply to inclusive enough high-energy hadron-hadron collisions

$$H_1(p_1) + H_2(p_2) \to W(Q) + X\,,$$

where H_1 and H_2 are the incoming hadrons, having momenta p_1 and p_2, W represents the particle produced in the hard scattering (Higgs or vector bosons, heavy quarks) and X denotes any other particle appearing in the final state. In this case the factorization formula takes the form

$$\sigma(p_1, p_2, Q) = \sum_{a,b} \int_\tau^1 dx_1 dx_2 \\ q_{a/H_1}\left(x_1, \mu_F^2\right) q_{b/H_2}\left(x_2, \mu_F^2\right) \hat{\sigma}_{ab}\left(x_1 p_1, x_2 p_2, \frac{Q^2}{\mu_F^2}, \alpha_s\right) + \mathcal{O}\left(\frac{\Lambda_{QCD}^2}{Q^2}\right), \quad (2.25)$$

where $\tau = \frac{Q^2}{s}$ and $s = (p_1 + p_2)^2$.

The two factorized expressions given in eqs. (2.23), (2.25) allow to connect cross sections for high-energy processes having hadrons in the initial states to hard scattering cross-sections. The former can be measured in collider experiments, while the latter can be computed in perturbation theory. The objects connecting perturbation theory with physical observables are the Parton Distribution Functions. The content of the factorization theorem is that all the dependence on low mass phenomena is entirely contained in the PDFs. Therefore, since they describe the internal structure of a given kind of hadron and have been decoupled from the short-distance details of the specific process we consider, they are nonpertubative and universal objects. This means that the PDFs appearing in the case of DIS must be the same considered for any other high-energy collisions.

2.3 Parton Distribution Functions

In the previous section we have introduced PDFs as some bare quantities, which are then used to reabsorb the infrared collinear poles coming from the fixed order computation of partonic cross sections. Following this approach PDFs are introduced in the discussion through the parton model ideas, and defined as objects containing all the dependence of the physical observables on low energy physics. It is possible to give a rigorous operator definition of parton distributions, which can be applied beyond perturbation theory and makes manifest the universal nature of PDFs. The formalism and notations commonly used to describe PDFs in terms of QCD operators are quite different from those introduced in the previous section, which are usually adopted for phenomenological applications. Since in this work we will present results for which both formalism are required, in this section we briefly revise the formal definition of PDFs, addressing their UV renormalization and renormalization group equations and making contact with the formalism and notations introduced in the previous section.

2.3.1 PDFs operator definition

Working in the Bjorken limit, it can be proved [4, 5] that the bare unpolarized quark PDF appearing in eq. (2.13) is related to the light-cone Fourier transform of a bilocal operator, given by

$$f_{\mathrm{q}}^{(0)}(x) = \int \frac{d\xi^-}{4\pi} e^{-ixP^+\xi^-} \langle \mathrm{P}|\bar{\psi}_{\mathrm{q}}^{(0)}\left(\xi^-\right)\gamma^+ U\left(\xi^-,0\right)\psi_{\mathrm{q}}^{(0)}(0)\,|\mathrm{P}\rangle\,, \tag{2.26}$$

where $|\mathrm{P}\rangle$ denotes a hadronic state with momentum $P^\mu = (P^0, 0, 0, P^z)$, x belongs to the real interval $[-1, 1]$ and $P^\pm = \frac{\left(P^0 \pm P^z\right)}{\sqrt{2}}$ are light-cone coordinates. The index q identifies the parton under investigation. For instance, in a theory where we only consider the four lightest quarks, we have $q = u, d, s, c$. The momentum carried by the parton is $k^\mu = xP^\mu$, $\psi_{\mathrm{q}}^{(0)}$ is the bare quark field operator and the Wilson line U is given by

$$U\left(\xi^-,0\right) = \mathrm{P}\exp\left(-ig\int_0^{\xi^-} d\eta^- A^{(0)+}\left(\eta^-\right)\right)\,. \tag{2.27}$$

An analogous definition can be given for the gluon bare PDFs, denoted as $f_g^{(0)}(x)$. The superscripts (0) in the above expressions identify bare fields: the matrix elements that enter in the definition of $f_{\rm q}^{(0)}$ are ultraviolet divergent, and therefore need to be renormalized. Renormalized parton distributions are usually defined by minimal subtraction, and the relation between the bare and the renormalized quantities is given by

$$f_a^{(0)}(x) = \sum_b \int_x^1 \frac{dy}{y} \, \mathrm{Z}_{ab}\left(\frac{x}{y}, \mu\right) f_b\left(y, \mu^2\right) \,, \tag{2.28}$$

where the indices a and b run over all the parton types (gluon and flavors of quarks) and μ denotes the renormalization scale introduced by the minimal subtraction scheme.

Focusing on the quark PDFs for now, the renormalized distributions introduced above have a compact support given by the interval $[-1, 1]$. To recover the conventions of the previous sections, used for phenomenological applications, it is customary to define the PDFs on the interval $[0, 1]$, and to introduce independent functions for the quarks and the antiquarks, which we have previously denoted as $q(x, \mu^2)$ and $\bar{q}(x, \mu^2)$ respectively. The relation between f_q, q and $\bar{q}$ is

$$f_{\rm q}\left(x, \mu^2\right) = \begin{cases} q(x, \mu^2) \,, & \text{if } x > 0 \,, \\ -\bar{q}(-x, \mu^2) \,, & \text{if } x < 0 \,. \end{cases} \tag{2.29}$$

It is useful to consider the symmetrised and antisymmetrised combinations of $f_{\rm q}$ in the interval $x \in [0, 1]$:

$$f_{\rm q}^{\rm sym}(x, \mu^2) = f_{\rm q}(x, \mu^2) + f_{\rm q}(-x, \mu^2) \,, \tag{2.30}$$

$$f_{\rm q}^{\rm asy}(x, \mu^2) = f_{\rm q}(x, \mu^2) - f_{\rm q}(-x, \mu^2) \,. \tag{2.31}$$

It can be readily shown that

$$f_{\rm q}^{\rm sym}(x, \mu^2) = q(x, \mu^2) - \bar{q}(x, \mu^2) = q^-(x, \mu^2) \,, \tag{2.32}$$

$$f_{\rm q}^{\rm asy}(x, \mu^2) = q(x, \mu^2) + \bar{q}(x, \mu^2) = q^+(x, \mu^2) \,. \tag{2.33}$$

where q^+ and q^- are defined by the equations above. Considering a SU(4) flavour symmetry, we can collect the quark fields in a vector, *e.g.* $\psi = (\psi_u, \psi_d, \psi_s, \psi_c)$,

and define the following nonsinglet bare PDFs:

$$f_A^{(0)}(x) \;\; = \int \tfrac{d\xi^-}{4\pi} e^{-ixP^+\xi^-} \langle \mathrm{P}|\bar{\psi}^{(0)}\left(\xi^-\right) \lambda_A \gamma^+ U\left(\xi^-,0\right) \psi^{(0)}\left(0\right)|\mathrm{P}\rangle \,, \tag{2.34}$$

where $A = 3, 8, 15$, and we have used the 4-dimensional Gell-Mann matrices

$$\lambda_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \lambda_8 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \lambda_{15} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -3 \end{pmatrix} \tag{2.35}$$

In this notation f_3 corresponds to $f^{u-d} = f_u - f_d$, $f_8 = f^{u+d-2s}$, and so on. The symmetrised and antisymmetrised combinations map directly into the so-called *evolution basis* for the PDFs that is normally used in phenomenological studies, see *e.g.* ref. [36] for a detailed definition of the flavor decomposition. More specifically, we have:

$$f_3^{\mathrm{asy}} = u^+ - d^+ = T_3 \,, \tag{2.36}$$

$$f_3^{\mathrm{sym}} = u^- - d^- = V_3 \,, \tag{2.37}$$

$$f_8^{\mathrm{asy}} = u^+ + d^+ - 2s^+ = T_8 \,, \tag{2.38}$$

$$f_8^{\mathrm{sym}} = u^- + d^- - 2s^- = V_8 \,, \tag{2.39}$$

$$f_{15}^{\mathrm{asy}} = u^+ + d^+ + s^+ - 3c^+ = T_{15} \,, \tag{2.40}$$

$$f_{15}^{\mathrm{sym}} = u^- + d^- + s^- - 3c^- = V_{15} \,. \tag{2.41}$$

As mentioned above the bilocal operator products of eq. (2.26) require renormalization. The corresponding renormalization group equations are the Altarelli-Parisi equations for PDFs, or DGLAP, discussed in the next section. Considering on-shell incoming partons, a straightforward 1-loop computation gives

$$f_{a/b}^{(0)}\left(x,\epsilon\right) = \delta_{ab}\,\delta\left(1-x\right) + \alpha_s \left[\frac{1}{\epsilon_{UV}} - \frac{1}{\epsilon_{IR}}\right] P^{(1)}\left(x\right) + \mathcal{O}\left(\alpha_s^2\right) \,, \tag{2.42}$$

where $P^{(1)}\left(x\right)$ represent the collinear splitting function, which will be introduced in sec.2.3.2. To obtain such result, one has to work in dimensional regularization to regularize both the UV and IR divergences. Working in the $\overline{MS}$ scheme, the UV pole $1/\epsilon_{UV}$ is removed through renormalization, and we are left with the

renormalized partonic PDFs

$$f_{a/b}\left(x,\epsilon\right)=\delta_{ab}\,\delta\left(1-x\right)+\alpha_s\left(-\frac{1}{\epsilon}\right)P^{(1)}\left(x\right)+\mathcal{O}\left(\alpha_s^2\right)\,. \tag{2.43}$$

Such object, despite being UV finite, does contain infrared poles.

We can now see how the formal approach followed here recovers the picture given in the previous section. Taking as example the case of DIS structure function, we can apply eq. (2.23) to the case of an incoming parton b[2]

$$\hat{F}_b\left(x,\epsilon\right)=\sum_a\int_x^1\frac{d\xi}{\xi}\,C_a\left(\frac{x}{\xi},\alpha_s\right)f_{a/b}\left(\xi,\epsilon\right)+\mathcal{O}\left(\frac{\Lambda_{QCD}^2}{Q^2}\right)\,. \tag{2.44}$$

Considering the coefficient functions power expansions given in eq. (2.24) and an analogue expansion for $\hat{F}_a$, using the 1-loop expression for the renormalized partonic PDF of eq. (2.43) we get

$$C_b^{(0)}\left(x\right)=\hat{F}_b^{(0)}\left(x\right)\,, \tag{2.45}$$

$$C_b^{(1)}\left(x\right)=\hat{F}_b^{(1)}\left(x,\epsilon\right)+\frac{1}{\epsilon}\sum_a\int_x^1\frac{d\xi}{\xi}P^{(1)}\left(\xi\right)\hat{F}_a^{(0)}\left(\frac{x}{\xi}\right)\,, \tag{2.46}$$

which recovers the prescription introduced in the previous section: in order to compute the hard scattering cross sections, one should calculate the structure function at the parton level and subtract from it certain infrared divergent terms proportional to the splitting kernel and the Born cross section. Such terms, identified as collinear emissions in the previous sections, here are computed in a process independent way starting directly from the formal definition of PDFs. Another way of stating this, is that the infrared subtraction kernel Γ introduced in eq. (2.18) corresponds to the parton level renormalized PDF of eq. (2.43). For a proof of this at every order in perturbation theory see for example [4].

2.3.2 DGLAP evolution equations

As stated in eq. (2.28) the operator defining parton distribution functions requires renormalization. It follows that renormalized PDFs acquire a scale dependence. Including in the discussion also the gluon PDF, the corresponding renormalization

[2]An important property of the hard coefficient functions is that they depend only on the parton type a and not on the specific hadron H, so that they can be computed with the simplest choice of external parton.

group equations read

$$\mu^2 \frac{\partial}{\partial \mu^2} \begin{pmatrix} q_i\left(x,\mu^2\right) \\ g\left(x,\mu^2\right) \end{pmatrix} = \alpha_s \sum_j \int_x^1 \frac{d\xi}{\xi} \begin{pmatrix} P_{q_i q_j}\left(\frac{x}{\xi},\alpha_s\right) & P_{q_i g}\left(\frac{x}{\xi},\alpha_s\right) \\ P_{g q_j}\left(\frac{x}{\xi},\alpha_s\right) & P_{gg}\left(\frac{x}{\xi},\alpha_s\right) \end{pmatrix} \begin{pmatrix} q_j\left(x,\mu^2\right) \\ g\left(x,\mu^2\right) \end{pmatrix} \tag{2.47}$$

with each splitting function P computable in perturbation theory

$$\begin{aligned} P_{q_i q_j}\left(x,\alpha_s\right) &= \delta_{ij} P_{qq}^{(0)}\left(x\right) + \alpha_s P_{q_i q_j}^{(1)}\left(x\right) + ... \\ P_{qg}\left(x,\alpha_s\right) &= P_{qg}^{(0)}\left(x\right) + \alpha_s P_{qg}^{(1)}\left(x\right) + ... \\ P_{gq}\left(x,\alpha_s\right) &= P_{gq}^{(0)}\left(x\right) + \alpha_s P_{gq}^{(1)}\left(x\right) + ... \\ P_{gg}\left(x,\alpha_s\right) &= P_{gg}^{(0)}\left(x\right) + \alpha_s P_{gg}^{(1)}\left(x\right) + ... \end{aligned} \tag{2.48}$$

It is convenient to re-express the DGLAP equations choosing a PDFs basis which maximally diagonalize them. In order to do this the aforementioned evolution basis is particularly useful. Denoting

$$q_i^{\pm} = q_i \pm \bar{q}_i \,, \tag{2.49}$$

and considering the flavours u, d, s, c, b, t the non-singlet sector is defined by the valence distributions

$$V_i = q_i^- \tag{2.50}$$

and by the T_i combinations

$$T_3 = u^+ - d^+ \,, \quad T_8 = u^+ + d^+ - 2s^+ \,, \quad T_{15} = u^+ + d^+ + s^+ - 3c^+ \,, \tag{2.51}$$

$$T_{24} = u^+ + d^+ + s^+ + c^+ - 4b^+ \,, \quad T_{35} = u^+ + d^+ + s^+ + c^+ + b^+ - 5t^+ \,. \tag{2.52}$$

Each non-singlet distribution q^{NS} will then satisfy an independent evolution equation, given by

$$\mu^2 \frac{d}{d\mu^2} q^{NS}\left(x,\mu^2\right) = \int_\xi^1 \frac{d\xi}{\xi} P\left(\xi,\alpha_s\right) q^{NS}\left(\frac{x}{\xi},\mu^2\right) \,. \tag{2.53}$$

The splitting function P for the V_i and T_i distributions is given by P^- and P^+ respectively, which at leading order are

$$P^{-(0)}\left(x\right) = P^{+(0)}\left(x\right) = \frac{C_F}{2\pi} \left(\frac{1+z^2}{1-z}\right)_+ \,. \tag{2.54}$$

Working in the evolution basis, the only distribution which couples to the gluon is the so called singlet combination, defined as

$$\Sigma\left(x,\mu^2\right)=\sum_i\left(q_i\left(x,\mu^2\right)+\bar{q}_i\left(x,\mu^2\right)\right)\,, \tag{2.55}$$

for which we have

$$\mu^2\frac{\partial}{\partial\mu^2}\begin{pmatrix}\Sigma\left(x,\mu^2\right)\\ g\left(x,\mu^2\right)\end{pmatrix}=\alpha_s\int_x^1\frac{d\xi}{\xi}\begin{pmatrix}P_{qq}\left(\frac{x}{\xi},\alpha_s\right) & P_{qg}\left(\frac{x}{\xi},\alpha_s\right)\\ P_{gq}\left(\frac{x}{\xi},\alpha_s\right) & P_{gg}\left(\frac{x}{\xi},\alpha_s\right)\end{pmatrix}\begin{pmatrix}\Sigma\left(x,\mu^2\right)\\ g\left(x,\mu^2\right)\end{pmatrix} \tag{2.56}$$

with the leading order splitting function given by

$$\begin{aligned}
P_{qq}^{(0)}\left(x\right)&=\frac{C_F}{2\pi}\left[\frac{1+x^2}{\left(1-x\right)_+}+\frac{3}{2}\delta\left(1-x\right)\right]\,,\\
P_{gg}^{(0)}\left(x\right)&=\frac{C_A}{\pi}\left[\frac{x}{\left(1-x\right)_+}+\frac{1-x}{x}+x\left(1-x\right)\right]+\delta\left(1-x\right)\frac{\left(11C_A-2n_fT_R\right)}{12\pi}\,,\\
P_{gq}^{(0)}\left(x\right)&=\frac{C_F}{2\pi}\left[\frac{1+\left(1-x\right)^2}{x}\right]\,,\\
P_{qg}^{(0)}\left(x\right)&=\frac{n_f}{2\pi}\left[x^2+\left(1-x\right)^2\right]\,.
\end{aligned} \tag{2.57}$$

Because of charge conjugation invariance and $SU(n_f)$ flavour symmetry, splitting functions are independent on the quark flavour and are the same for quarks and antiquarks. Also, the leading order splitting functions have a nice physical interpretation as probability distributions. Following ref. [37], eq. (2.53) can be written as

$$\begin{aligned}
&q^{NS}\left(x,\mu^2\right)+dq^{NS}\left(x,\mu^2\right)=\\
&\qquad\int_0^1 dy\int_0^1 dz\,\delta\left(zy-x\right)q^{NS}\left(y,\mu^2\right)\left[\delta\left(z-1\right)+\alpha_s\,P\left(z\right)d\log\frac{\mu^2}{\mu_0^2}\right]\,.
\end{aligned} \tag{2.58}$$

The quantity in square brackets can then be interpreted as the probability density of finding a quark inside another quark, with a fraction z of the parent momentum. The quantity $\alpha_s P\left(z\right)$ is then the variation of such probability density per logarithmic unit of the energy. The interpretation of splitting function as probability densities implies that they are positivite for $x<1$ and allows to compute them starting from the QCD vertices for $q\to q\,g$, $g\to q\bar{q}$ and $g\to gg$.

The DGLAP equations can be solved perturbatively by computing an evolution

kernel, which can be subsequently convoluted with PDFs at a given scale Q_0 to evolve them up to the final scale Q. In such evolution kernel the resummation of the large collinear logarithms mentioned at the end of sec. 2.2.1 is achieved by exponentiating them, allowing a consistent definition of PDFs at a generic scale Q. In the following we briefly summarise how the QCD evolution equation can be solved in Mellin space for the nonsinglet sector, yielding such evolution kernel. Denoting the nonsiglet distributions V_i and T_i with $q^{(-)}$ and $q^{(+)}$ respectively, the QCD evolution equations can be written as

$$\mu^2 \frac{\partial}{\partial \mu^2} q^{(\pm)}\left(x, \mu^2\right) = \frac{\alpha_s(\mu^2)}{2\pi} \int_x^1 \frac{d\xi}{\xi} P_{qq}^{(\pm)}\left(\frac{x}{\xi}, \alpha_s(Q^2)\right) q^{(\pm)}\left(\xi, \mu^2\right), \tag{2.59}$$

which in Mellin space becomes

$$\mu^2 \frac{\partial}{\partial \mu^2} q^{(\pm)}\left(N, \mu^2\right) = \gamma^{(\pm)}\left(N, \alpha_s\right) q^{(\pm)}\left(N, \mu^2\right) . \tag{2.60}$$

The distribution at the scale μ^2 is obtained from the distribution at the scale μ_0^2 by introducing the evolution operator Γ

$$q^{(\pm)}\left(N, \mu^2\right) = \Gamma^{(\pm)}\left(N, \alpha_s, \alpha_s^0\right) q^{(\pm)}\left(N, \mu_0^2\right) , \tag{2.61}$$

where $\alpha_s \equiv \alpha_s\left(\mu^2\right)$ and $\alpha_s^0 \equiv \alpha_s\left(\mu_0^2\right)$. Substituting eq. (2.61) in eq. (2.60) and remembering that the dependence of Γ on the scale μ is only through the coupling, we have

$$\beta\left(\alpha_s\right) \frac{\partial}{\partial \alpha_s} \Gamma^{(\pm)}\left(N, \alpha_s, \alpha_s^0\right) = \gamma^{(\pm)}\left(N, \alpha_s\right) \Gamma^{(\pm)}\left(N, \alpha_s, \alpha_s^0\right) . \tag{2.62}$$

Considering for example NLO evolution equations, using eq. (1.13) and the perturbative expansion of the anomalous dimension

$$\gamma^{(\pm)}\left(N, \alpha_s\right) = \frac{\alpha_s}{4\pi} \gamma_0^{(\pm)}(N) + \left(\frac{\alpha_s}{4\pi}\right)^2 \gamma_1^{(\pm)}(N) + \mathcal{O}\left(\alpha_s^3\right) , \tag{2.63}$$

eq. (2.62) can be easily solved getting

$$\Gamma^{(\pm)}\left(N, \alpha_s, \alpha_s^0\right) = 1 + \frac{\alpha_s - \alpha_s^0}{4\pi} \left(\frac{\gamma_1^{(\pm)}(N) - \beta_1 \gamma_0^{(\pm)}(N)}{2\beta_0}\right) . \tag{2.64}$$

The solution in the x-space is obtained by computing the inverse Mellin transform of $\Gamma^{(\pm)}\left(N, \alpha_s, \alpha_s^0\right)$. Having analytically continued the function $\Gamma(N)$ to the complex plane, the inverse Mellin transform is obtained by computing the contour

integral

$$\Gamma^{(\pm)}\left(x,\alpha_s,\alpha_s^0\right)=\int_C\frac{dN}{2\pi i}x^{-N}\,\Gamma^{(\pm)}\left(N,\alpha_s,\alpha_s^0\right)\,. \tag{2.65}$$

2.4 Heavy quarks

Considering a process in perturbative QCD involving heavy quarks[3], the corresponding cross-section can be computed in different renormalization schemes. In a standard minimal subtraction scheme, like $\overline{MS}$, heavy quarks are treated on the same footing as light flavours. In practise this means two things: first, they are endowed with a PDF and second the β function depends on the total number n_f of both light and heavy flavours. Alternatively, in a decoupling scheme heavy quarks are treated as massive particles which fully decouple from QCD evolution equations, so that only the n_l light quarks contribute to the DGLAP and running of α_s, as briefly discussed in sec. 1.3. In the first case, often denoted as massless scheme, collinear logarithms of Q^2/m_h^2 are resummed through DGLAP equations and reabsorbed in the corresponding PDF, but corrections suppressed by powers of m_h^2/Q^2 are neglected. In the second case, denoted as massive scheme, collinear logarithms are only included to fix order, but the full mass dependence is retained. While a minimal subtraction scheme is more precise at high scales $Q^2 \gg m_h^2$, where unresummed collinear logarithms would spoil perturbation theory, a decoupling scheme is more accurate close to the threshold, where mass corrections might be non negligible. Heavy quarks schemes are all based on the idea of combining these two computations, each of which is more accurate in a certain kinematic region, in order to get a single result which is accurate at all scales. Some of the possible options available from the literature are the ACOT [38–40], S-ACOT [41], TR [42] and FONLL schemes. The latter was initially introduce in ref. [43] in the context of hadroproduction of heavy quarks, and subsequently extended to DIS structure functions in refs. [44] and to hadronic processes in ref. [45].

In the following, using the notations of refs. [44, 45] we will briefly recall the main features of the FONLL scheme, which will be used in chapter 6 to construct a new method to deal with initial state heavy quarks in an hadronic process.

[3] considering a process characterized by an hard scale Q^2, we define a quark to be heavy if $m_h^2 \gg Q^2$, with m_h the quark mass. This definition is usually applied to the charm, bottom and top quarks.

Considering an hadronic process involving n_l light quarks q and only one massive quark h of mass m_h, the corresponding cross section in the massless (n_l+1)-flavours scheme is given by

$$\sigma^{(n_l+1)} = \int\int dx_1 dx_2 \sum_{ij=g,q,\bar{q},h,\bar{h}} f_i^{(n_l+1)}\left(x_1,\mu^2\right) f_j^{(n_l+1)}\left(x_2,\mu^2\right) \times \hat{\sigma}_{ij}^{(n_l+1)}\left(x_1,x_2,\alpha_s^{(n_l+1)}\right) . \tag{2.66}$$

The sum in eq. (2.66) runs on both light and heavy flavours, which are all treated in the $\overline{MS}$ scheme. The heavy quark has an associated PDF and contributes to both the DGLAP evolution equations and to the running of α_s, which is therefore denoted as $\alpha_s^{(n_l+1)}$. For simplicity we have omitted the dependence on the factorization and renormalization scales in the hard cross section. The same process can be computed in the massive (n_l)-flavours scheme as

$$\sigma^{(n_l)} = \int\int dx_1 dx_2 \sum_{ij=g,q,\bar{q}} f_i^{(n_l)}\left(x_1,\mu^2\right) f_j^{(n_l)}\left(x_2,\mu^2\right) \times \hat{\sigma}_{ij}^{(n_l)}\left(x_1,x_2,\frac{\mu^2}{m_h^2},\alpha_s^{(n_l)}\right) . \tag{2.67}$$

Unlike the case of the massless computation, here the sum is on light flavours only, there is no PDF corresponding to the heavy quark and the hard cross-section retains the explicit dependence on the heavy quark mass m_h. In order to match the two computations, we express eqs. (2.66), (2.67) in terms of the massless schemes coupling $\alpha_s^{(n_l+1)}$ and light quarks PDFs

$$f_i^{(n_l+1)}, \quad i = g,q,\bar{q}$$

using relations of the form

$$\alpha_s^{(n_l+1)}\left(\mu^2\right) = \alpha_s^{(n_l)}\left(\mu^2\right) + \sum_{k=2}^{\infty} c_k\left(L\right)\left(\alpha_s^{(n_l)}\left(m_h^2\right)\right)^k , \tag{2.68}$$

$$f_i^{(n_l+1)}\left(x,\mu^2\right) = \int_x^1 \frac{dy}{y} \sum_{j=g,q,\bar{q}} K_{ij}\left(\frac{x}{y},L,\alpha_s^{(n_l)}\left(\mu^2\right)\right) f_j^{(n_l)}\left(y,\mu^2\right) , \tag{2.69}$$

where $i = g,q,\bar{q},h,\bar{h}$ and $L = \log\mu^2/m_h^2$. The coefficients $c_k\left(L\right)$ are polynomial in L, and the functions K_{ij} can be expressed as a power expansion in α_s with coefficients that are polynomial in L. The sum over j in eq. (2.69) runs over the n_l light flavours and anti-flavour plus the gluon, therefore the first $2n_l+1$ of these

equations relate the light quarks and gluon PDFs in the two schemes and can be inverted in order to express the massive scheme PDFs in terms of the massless scheme ones. The last two equations allow to express the heavy quark PDF in the massless schemes in terms of the gluon and light flavours PDFs of the massive one. Inverting eqs. (2.68), (2.69) and substituting in eq. (2.67), one can obtain the expression of the massive scheme cross-section in terms of $\alpha_s^{(n_l+1)}$ and $f_i^{(n_l+1)}$, with $i = g, q, \bar{q}$

$$\sigma^{(n_l)} = \int\int dx_1 dx_2 \sum_{ij=g,q,\bar{q}} f_i^{(n_l+1)}\left(x_1,\mu^2\right) f_j^{(n_l+1)}\left(x_2,\mu^2\right) \times B_{ij}\left(x_1,x_2,\frac{\mu^2}{m_h^2},\alpha_s^{(n_l+1)}\right), \tag{2.70}$$

where the coefficient functions B_{ij} can be expressed as a fixed order expansion in $\alpha_s^{(n_l+1)}$

$$B_{ij}\left(x_1,x_2,\frac{\mu^2}{m_h^2},\alpha_s^{(n_l+1)}\left(\mu^2\right)\right) = \sum_{p=0}^{P}\left(\alpha_s^{(n_l+1)}\left(\mu^2\right)\right)^p B_{ij}^p\left(x_1,x_2,\frac{\mu^2}{m_h^2}\right). \tag{2.71}$$

From now on we can use eq. (2.70) to express the massive scheme results, avoiding any further reference to $\alpha_s^{(n_l)}$ and $f_i^{(n_l)}$[4]. Using again eqs. (2.69) to write the massless scheme heavy quarks PDFs in terms of light-quark parton distributions, the massless scheme results eq. (2.66) can be written entirely in terms of light-quark PDFs

$$\sigma^{(n_l+1)} = \int\int dx_1 dx_2 \sum_{ij=g,q,\bar{q}} f_i^{(n_l+1)}\left(x_1,\mu^2\right) f_j^{(n_l+1)}\left(x_2,\mu^2\right) \times A_{ij}\left(x_1,x_2,L,\alpha_s^{(n_l+1)}\right), \tag{2.72}$$

where the coefficients A_{ij} are given by a perturbative expansion of the form

$$A_{ij}\left(x_1,x_2,L,\alpha_s^{(n_l+1)}\left(\mu^2\right)\right) = \sum_{p=0}^{N}\left(\alpha_s^{n_l+1}\left(\mu^2\right)\right)^p \times \sum_{k=0}^{\infty} A_{ij}^{(p),(k)}\left(x_1,x_2\right)\left(\alpha_s^{(n_l+1)}\left(\mu^2\right)L\right)^k, \tag{2.73}$$

where at leading order $N = 0$, at N^{k}LO $N = k$.

[4]Note that eq. (2.70) differs from the original massive scheme expression eq. (2.67) by subleading terms, due to the fact that matching coefficients are known at a given order in perturbation theory.

In order to match the two results given in eqs. (2.72), (2.70) one should notice that the contributions to the massive scheme expression of eq. (2.71) which do not vanish when $\mu^2 \gg m_h^2$, namely all the constant and logarithmic terms, must also be present in the massless scheme computation. The p-th order contribution to the sum of these terms, denoted as $B_{ij}^{(0),p}$, can be implicitly defined as

$$\lim_{m_h \to 0} \left[B_{ij}^{p} \left(x_1, x_2, \frac{\mu^2}{m_h^2} \right) - B_{ij}^{(0),p} \left(x_1, x_2, \frac{\mu^2}{m_h^2} \right) \right] = 0 \,, \tag{2.74}$$

and since it has to be present also in the massless scheme computation it will admit an expansion of the form

$$B_{ij}^{(0),p} \left(x_1, x_2, , \frac{\mu^2}{m_h^2} \right) = \sum_{k=0}^{p} A_{ij}^{(p-k),(k)} \left(x_1, x_2 \right) L^k \,. \tag{2.75}$$

The FONLL method can be expressed as follows: considering the massless scheme coefficients at a given perturbative order p appearing in eq. (2.73) $A_{ij}^{(p),(k)}$, replace all the terms which are also present in the massless limit of the massive scheme $B_{ij}^{(0),p}$ of eq. (2.75) with their fully massive expression B_{ij}^{p} appearing in eq. (2.71). This can be done in a systematic way defining the massless limit of the massive computation as

$$\begin{aligned} \sigma^{(n_l),(0)} = \int \int dx_1 dx_2 \sum_{ij=g,q,\bar{q}} & f_i^{(n_l+1)} \left(x_1, \mu^2 \right) f_j^{(n_l+1)} \left(x_2, \mu^2 \right) \\ & \times B_{ij}^{(0)} \left(x_1, x_2, \frac{\mu^2}{m_h^2}, \alpha_s^{(n_l+1)} \right) \,, \end{aligned} \tag{2.76}$$

with

$$B_{ij}^{(0)} = \sum_{p=0}^{N} \left(\alpha_s^{(n_l+1)} \right)^p B_{ij}^{(0),p} \left(x_1, x_2, \frac{\mu^2}{m_h^2} \right) \,, \tag{2.77}$$

and computing

$$\sigma^{FONLL} = \sigma^{(n_l+1)} + \sigma^{(n_l)} - \sigma^{(n_l),(0)} \,. \tag{2.78}$$

In this way the mass suppressed terms which are not included in a massless computation, but which are known from the massive one, are included in the final results. On the other hand the all order resummation of collinear logarithms L achieved in a massless scheme through DGLAP equations, which would be lost in the massive scheme, is now taken into account.

3

Global fits of Parton Distribution Functions

As seen in the previous chapter, factorization theorems allow for a separation between contributions related to different distance scales: while short distance effects can be obtained through the perturbative computation of partonic matrix elements, long-distance contributions are collected in the universal PDFs $q\left(x, Q^2\right)$. As seen in sec. 2.3.2, knowing the functional form of the PDFs at a given initial scale Q_0^2, their dependence on the generic energy scale Q^2 can be computed by solving the DGLAP evolution equations. However the dependence on x would be computable only solving QCD in a nonperturbative domain. We will come back to this point in chapter 7, when considering lattice QCD observables. In this chapter we will discuss the general approach adopted to extract the x dependence of the PDFs from a discrete set of experimental data, using as basic ingredient the factorization theorems for high-energy processes discussed before.

The general problem is to determine a set of unknown functions given a collection of data instances. Such kind of task has a long history in data science literature, and can be classified as a pattern recognition problem [46]. However the specific problem of PDFs determination has some additional peculiarities that we need to keep in mind. First, PDFs are continuos functions, which makes our problem intrinsically ill-defined: a continuous real function $q\left(x, Q_0^2\right)$ cannot be determined from a discrete set of data, no matter how large such set is. Second, the data are not instances of the functions we are trying to determine, but they are related to them through factorization theorems: each point is determined combining a certain subset of PDFs in a non-linear way, and integrating over a certain range of x, as stated in eq. (2.25). As we will extensively discuss in this and the following chapters, this fact has several practical and conceptual implications.

Another important aspect concerning the problem of PDFs determination is the need for results with well quantified uncertainties and correlations. Universality of PDFs allows to extract them from a set of data for some specific high-energy processes and use the result to make predictions for different observables not included in the analysis. In order for PDFs to be useful as an input to physics predictions one needs to account for the different sources of statistical and systematic uncertainties affecting the experimental data, and propagate them on the resulting PDFs. Ideally one would like to determine a representation of the probability distribution for the unknown PDFs in the whole functional space, so that the full information about uncertainties and correlations are taken into account when making predictions for new observables.

In this chapter we discuss how all these issues have been addressed within the NNPDF collaboration starting from 2002 [47], revising the Monte Carlo replicas generation, the neural networks parameterization and the minimization procedure. Such methodology has been used to produce the last public NNPDF PDFs set NNPDF3.1 [48] . The studies which will be described in chapters 4 and 5, regarding the inclusion of jets data and the formulation of a theoretical error in a global PDFs determination, have been performed within the private `c++` implementation of this framework.

We will then describe how such methodology has been revised and extended within the new `n3fit` code [49], which will be used to produce the next NNPDF release, NNPDF4.0. In particular, we will discuss in detail the implementation of $\overline{MS}$ PDFs positivity, the integrability of the nonsinglet sector and the fit basis independence.

3.1 NNPDF methodology

The NNPDF methodology is based on a Monte Carlo determination of the PDFs error, combined with a neural network parameterization of the unknown x-dependence of the PDFs. So far numerical minimization algorithms have been used, and overfitting has been avoided using a cross-validation technique. In the first part of this section we revise these general ideas, referring to the baseline `c++` code which has been used to produce NNPDF3.1.

3.1.1 The Monte Carlo replica method for error propagation

As mentioned before, in order to make PDFs sets an useful tool for making predictions, a faithful estimation of the errors affecting the analysis is necessary. A possible way to address the problem is to determine the probability distribution of the PDFs set $q(x)$ given a set of data D. Let's denote such probability distribution as $\mathcal{P}(q|D)$. A generic observable $\mathcal{O}$, function of one or more PDFs, will be determined as an expectation value over $\mathcal{P}$. Considering for example the case of a DIS observable, the central value and uncertainty of the prediction will be given by

$$\langle \mathcal{O} \rangle = \int Dq\, \mathcal{O}[q]\, \mathcal{P}(q|D)\ , \qquad \mathrm{Var}[\mathcal{O}] = \int Dq\ \left(\mathcal{O}[q] - \langle \mathcal{O} \rangle\right)^2 \mathcal{P}(q|D)\ . \quad (3.1)$$

The problem is then about how to compute a reliable representation of the probability distribution $\mathcal{P}(q|D)$. In the NNPDF framework a Monte Carlo approach is adopted. In this method an ensemble of N_{rep} artificial data is generated for each experimental point, assuming a multigaussian distribution given by the experimental covariance matrix. More precisely, denoting as $\mathcal{O}_p^{\exp}$ the experimental data for the single point p corresponding to the kinematic variables $\{x_p, Q_p^2\}$, N_{rep} artificial pseudo-data are generated according to [50]

$$\mathcal{O}_p^{(k)} = S_{p,N}^{(k)} \left(\mathcal{O}_p^{\exp} + \sum_l r_{p,l}^{(k)} \sigma_{p,l} + r_p^{(k)} \sigma_{p,s} \right)\ , \quad k = 1, ..., N_{rep} \quad (3.2)$$

where

$$S_{p,N}^{(k)} = \prod_n \left(1 + r_{p,n}^{(k)} \sigma_{p,n}\right)\ , \quad (3.3)$$

takes into account normalization errors. The variable $\sigma_{p,s}$ represents the uncorrelated statistical uncertainty of the datapoint while $\sigma_{p,l}$ and $\sigma_{p,n}$ are the l-th and n-th source of additive and multiplicative systematic uncertainties respectively. The variables $r_p^{(k)}$, $r_{p,l}^{(k)}$ and $r_{p,n}^{(k)}$ are all univariate gaussian random numbers, generating fluctuations of the artificial data around the experimental value. For each replica k, if the l-th additive systematic is correlated between the two experimental points p and p' then $r_{p,l}^{(k)} = r_{p',l}^{(k)}$ with an equivalent condition on $r_{p,n}^{(k)}$ to ensure correlation between multiplicative uncertainties. N_{rep} independent fits are performed, generating a Monte Carlo ensemble of PDFs that faithfully reproduces the statistical features of the original experimental data, providing the

desired representation of the probability density in the space of PDFs. Central values, uncertainties and correlations can then be computed by doing statistics over replicas, so that for example

$$\langle \mathcal{O} \rangle \sim \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} \mathcal{O}^{(k)}, \qquad \text{Var}\left[\mathcal{O}\right] \sim \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} \left(\mathcal{O}^{(k)} - \langle \mathcal{O} \rangle\right)^2 . \tag{3.4}$$

The Monte Carlo method allows to propagate the error from the data to the PDFs set in a natural way, using only the input experimental information and without the need of any further assumption.

3.1.2 Parameterization and Neural Networks

As mentioned above, the determination of a continuos function from a discrete set of data is an intrinsically ill-defined problem, no matter how copious this set is. In order to overcome this issue a particular functional form for the x dependence of the PDFs at a reference scale Q_0 is chosen, given in terms of a set of free parameters. The PDFs at all the other scales Q are determined by solving perturbative evolution equations, and the experimental data are used to determine the optimal values of the free parameters.

The choice of the specific parameterization has been largely debated and investigated, and its final form varies substantially between different fitting groups. The underlying idea is that the PDFs parameterization has to be flexible enough to describe all the data entering the analysis without introducing a bias in the final results. The fact that a too restrictive functional form is likely to introduce a strong bias in the result has been rapidly recognized, and more and more complex models have been adopted in recent PDFs determinations.

The use of neural networks as basic functional form for PDFs was first suggested in 2002 [47], in the context of the determination of the DIS structure function F_2. The idea was further developed in ref. [51] and applied for the first time to quark distributions in ref. [52]. This first suggestion was then developed in the context of global fits of PDFs by the NNPDF collaboration through a series of intermediate steps [50, 53–55], the last of which in 2017, with the release of the PDFs set NNPDF3.1 [48].

Neural networks are a class of non-linear maps between some input $\xi_i^{(1)}$ and some

output $\xi_i^{(L)}$ variables. They are used in several machine learning applications where flexibility and a lack of bias with respect to a conventional fixed parameterization are desired. Like other sets of functions, neural networks, in the limit of an infinite number of parameters, can reproduce any differentiable function. The main advantage of using them is the possibility of dealing with a greater number of parameters than what is usually available using a more standard parameterization. The basic element of a neural net is a *neuron* or *node*, which takes a vector $\vec{x} \in \Re^N$ as input and gives back a scalar output $f(\vec{x}) \in \Re$. A neural network consists of many neurons stacked into layers and can be graphically represented as a direct graph made by input, hidden and output layers. Starting from the input, the output of each layer is used as input for the next one. The specific form of the function f characterizing the nodes is usually given by a linear function composed with a non linear transformation g, called activation function, so that the output of the i-th node of the l-th layer $\xi_i^{(l)}$ is obtained by that of the $(l-1)$-th layer using the relation

$$\xi_i^{(l)} = g\left(\sum_j w_{ij}^{(l)} \xi_j^{(l-1)} + \theta_i^{(l)}\right) \tag{3.5}$$

The *weights* $w_{ij}^{(l)}$ and the *biases* $\theta_i^{(l)}$ are the free parameter of the nets, to be determined during the fit. Different choices can be made for the activation function g. In many cases, including the NNPDF code, it is given by a sigmoid

$$g(x) = \frac{1}{1+e^{-x}} \tag{3.6}$$

for the nodes belonging to the hidden layers, and by a linear function $g(x) = x$ for the output layer.

Starting from the first proof-of-concepts exercises up to the most recent public release, the basic neural network architecture employed in all the NNPDF determination has been the same. The only thing which has been changed is the number of independent parameterized PDFs in a global fit: five in NNPDF1.0 (up and down quarks and antiquarks, gluon), seven in NNPDF1.1 (up, down, strange quarks and antiquarks, gluon) and eight in NNPDF3.1 where the total charm PDF was fitted from data for the first time. Each flavour is independently parameterized using a neural network, with architecture 2-5-3-1, represented in fig. (3.1). The momentum fraction x enters the two input nodes as x and $\log x$, followed by two hidden layers with a sigmoid activation function

and an output layers with a single node, associated with a linear activation function. This architecture is supplemented with a preprocessing polynomial factor $x^{-\alpha}(1-x)^{\beta}$ which controls the PDFs behaviour at large and small x, so that the parameterization for each independent flavour reads

$$x\, q_j\left(x, Q_0^2\right) = x^{1-\alpha_j}(1-x)^{\beta_j}\, \mathrm{NN}_j(x)\ . \tag{3.7}$$

Eq. (3.7) can be supplemented by an additional normalization factor A_j which imposes momentum and valence sum rules. The eight flavours independently parameterized in NNPDF3.1 are

$$q_j = \left[\Sigma, g, V, V_3, V_8, T_3, T_8, c^+\right]\ , \tag{3.8}$$

which in terms of quarks and antiquarks distributions are given by

$$\begin{aligned}
\Sigma &= u + \bar{u} + d + \bar{d} + s + \bar{s} + 2c\,,\\
V &= (u-\bar{u}) + \left(d-\bar{d}\right) + (s-\bar{s})\ ,\\
V_3 &= (u-\bar{u}) - \left(d-\bar{d}\right)\ ,\\
V_8 &= \left(u-\bar{u}+d-\bar{d}\right) - 2\,(s-\bar{s})\ ,\\
T_3 &= (u+\bar{u}) - \left(d+\bar{d}\right)\ ,\\
T_8 &= \left(u+\bar{u}+d+\bar{d}\right) - 2\,(s+\bar{s})\ ,\\
c^+ &= c + \bar{c}\,.
\end{aligned} \tag{3.9}$$

The preprocessing exponents α_i and β_i are randomized, by choosing a different value for each replica within a suitable range. This is determined in a fully self-consistent way: the effective exponents, defined in ref. [55] as

$$\alpha_{eff,i}(x) = \frac{\log q_i(x)}{\log 1/x}\,, \quad \beta_{eff,i}(x) = \frac{\log q_i(x)}{\log(1-x)}\,, \tag{3.10}$$

are computed for each distribution q_j. The 68% confidence level across replicas is determined for each flavour, and the fit is repeated with the exponents randomized in a range taken equal to twice this interval. This procedure is iterated until the range stops changing.

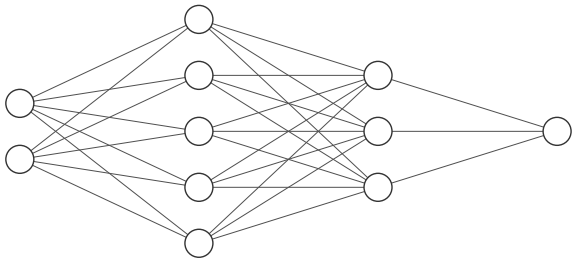

Figure 3.1 *Graphical representation of the neural network used in the NNPDF code. For each PDF of eq.* (3.8) *one independent neural network is implemented.*

3.1.3 Minimization and stopping

The optimal fit is obtained by varying the parameters of $q_j\left(x, Q_0^2\right)$ in such a way that some chosen figure of merit is minimized. Since most of the experimental data are assumed to have a multigaussian distribution, a standard choice for such an object is obtained by taking the standard χ_k^2 for each individual dataset k

$$\chi_k^2 = \sum_{ij}^{N_{dat}} \left(D_{ki} - T_{ki}\right) C_{ij}^{-1} \left(D_{kj} - T_{kj}\right) , \tag{3.11}$$

and then by building the quantity

$$\chi^2 = \sum_k \chi_k^2 \, . \tag{3.12}$$

Here D_{ki} is the i-th experimental datapoint in the k-th dataset; T_{ki} is the corresponding theoretical prediction, computed using a suitable factorization theorem and expressed as a function of the free parameters; C_{ij} is the covariance matrix, which takes into account both statistical and systematic uncertainties, as given by the experimental collaborations. In order to avoid a fitting bias, multiplicative uncertainties required to be handled with a specific method denoted as t_0 prescription, which as been developed in ref. [56] and implemented in all the following NNPDF PDFs determination.

The χ^2 minimization implemented within the NNPDF environment is based on genetic algorithm (GA): after a first random initialization of the neural network parameters, the weights are mutated according to a suitable rule, producing several copies of the original neural net, each one characterized by a different mutation. Mutations with the lowest value of the figure of merit are selected and the procedure is iterated. Different variations of GA have been used in every NNPDF PDFs set, including NNPDF3.1. Another possible option which has been investigated is the CMA algorithm [57], which has been used, for example, in a recent NNPDF determination of fragmentation functions [58]. Nowadays several efficient deterministic methods are available in a number of public libraries. As we are going to discuss in the next section, numerical minimizers are no longer the best possible option and an efficient deterministic minimization is more desirable.

Independently from the specific minimization algorithm implemented, overfitting is avoided employing a cross-validation technique. In this method, the available data are split in two sets. The first, the training set, is used for the minimization of the error function, while the second, the validation set, does not enter the fitting procedure. At each iteration of the minimization algorithm, the error function between the theory predictions from the neural net and the data is computed for both the training and validation set. At an early stage of the training, both these quantities are expected to decrease. However, towards the end of the training, while the error function over the training set will keep decreasing, the same value computed over the validation data will reach a minimal value, and eventually it may even start increasing. This is a signal of overfitting, and the point in parameters space yielding the minimal value of the validation error is the one taken as the fit result.

3.1.4 Fast Kernel tables

In order to get expressions for physical observables, PDFs have to be evolved up to the physical scale Q of the hard processes and finally combined with partonic matrix elements. These steps happen through two separate convolutions, first with the evolution kernel, see sec. 2.3.2, which solves DGLAP evolution equations and second with the hard cross sections. Such convolutions happen by means of FastKernel (FK) tables, introduced and validated in refs. [53, 59] and used in any subsequent NNPDF analysis. Each PDFs is projected on a suitable basis

functions

$$q_i\left(x,Q_0^2\right)=\sum_\alpha q_i\left(x_\alpha,Q_0^2\right)\mathcal{I}^{(\alpha)}\left(x\right) \tag{3.13}$$

so that, considering the case of a DIS observable $\mathcal{O}^{\rm DIS}$, its final value can be expressed as a tensors product of the kind

$$\mathcal{O}^{\rm DIS}=\left[{\rm FK}\right]_{i\alpha}q_{i\alpha}\,. \tag{3.14}$$

with the PDF tensor defined as $q_{i\alpha}=q_i\left(x_\alpha,Q_0^2\right)$. The matrix FK, denoted as FK table, stores the evolution and partonic matrix element and can be precomputed for each process entering the analysis. In the case of hadronic observables $\mathcal{O}^{\rm DY}$ the PDF tensor has to be replaced by a luminosity tensor $\mathcal{L}_{i\alpha j\beta}=q_{i\alpha}q_{j\beta}$ and the FK table becomes a rank-4 tensor,

$$\mathcal{O}^{\rm DY}=\left[{\rm FK}\right]_{i\alpha j\beta}\mathcal{L}_{i\alpha j\beta}\,. \tag{3.15}$$

After computing the value for all the observables entering the fit, the data are split into training and validation sets and the χ^2 function of the training set is minimized.

3.2 Towards NNPDF4.0

The methodology described in the previous section has been completely revised and extended within the new `n3fit` environment, first presented in ref. [49]. This framework will be used to produce NNPDF4.0, the next public PDFs set by the NNPDF collaboration. In this section we describe some of the `n3fit` general features, focusing in particular on the implementation of PDFs positivity and integrability. Additionally, we will present some results regarding the problem of the fit basis independence, addressed here for the first time within this environment.

3.2.1 Architecture and general structure

The `n3fit` code is a python-based framework, written using an object-oriented approach and a number of external libraries. Unlike the previous `c++` code of the NNPDF methodology, which is fully based on an in-house implementation of neural networks and minimization algorithms, in the new framework Keras [60] and Tensorflow [61] have been used to deal with them. This choice greatly simplifies the study of new architectures and techniques recently introduced in the machine learning literature, allowing for a systematic investigation of many of them, and represents an important technological improvement with respect to the previous code.

The two main methodological changes in `n3fit` concern the architecture and the minimization algorithm: rather than using eight independent neural networks, each one giving as output a particular flavour, in the new environment a single net with an eight-dimensional output is used; additionally, gradient descent methods are implemented to replace the genetic algorithm described before. The new architecture, graphically depicted in fig. 3.2, allows to study and take into account cross-correlations between different PDFs, while the new gradient descent minimizers have been proved to produce more stable fits than those obtained using the genetic algorithm.

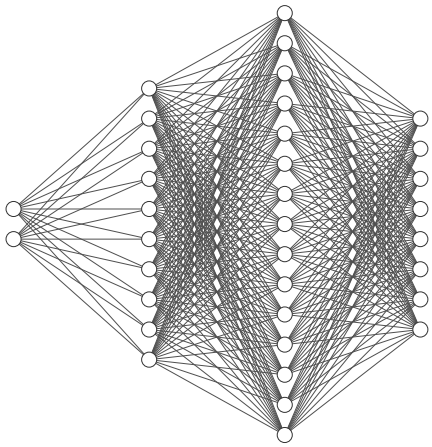

Figure 3.2 *Graphical representation of the neural network used in the* `n3fit` *framework. Each output nodes represents one of the independently parameterized flavours.*

For each dataset entering the fit, a vector of x values is given as input to the neural network and, as in the old methodology, before going through the

intermediate layers it is split into $(x, \log x)$. The eight output nodes of the neural network provide the eight independent PDFs parameterized at the reference scale Q_0. We denote such set of independent parton distributions as *fit basis*. In NNPDF3.1, this is given by eq. (3.8). The new framework allows the user to choose between different options: two standard choices are the so called evolution and flavour basis. While the former represents the equivalent of the one given in eq. (3.8), in the latter each quark, antiquark and the gluon are independently parameterized. The choice of the fit basis should not affect the final results, however different choices might be more convenient from a numerical point of view, and different architecture setups might be required when changing the basis. We will extensively discuss these points in sec. 3.2.2. Depending on our choice for the fit basis, each output can be supplemented with a suitable preprocessing polynomials, and with normalization factors to impose momentum and valence sum rules, as discussed in sec. 3.2.3

3.2.2 Fit basis

The parton distributions used to build the FK tables are the thirteen PDFs

$$q_i = \Sigma, g, V, V_3, V_8, V_{15}, V_{24}, V_{35}, T_3, T_8, T_{15}, T_{24}, T_{35}. \tag{3.16}$$

Eight of these are independently parameterized, while the remaining ones, involving heavy quarks, are determined by perturbative evolution, using the matching relations given in eqs. (2.69). When QED corrections are considered in the analysis, the photon PDF γ is also included and independently parameterized. As mentioned before, the `n3fit` environment allows for different fit basis choices. In other words, the user can choose which distributions should be independently parameterized.

The two most natural choices are the so called *evolution* and *flavour* basis. The former is defined by the eight dimensional subset of (3.16) given by

$$q_k = [g, \Sigma, V, V_3, V_8, T_3, T_8, T_{15}] \ . \tag{3.17}$$

This basis is the equivalent of the one given in eq. (3.8), with the only difference that the distribution c^+ has been replaced by T_{15}, defined as

$$T_{15} = \left(u + \bar{u} + d + \bar{d} + s + \bar{s}\right) - 3\left(c + \bar{c}\right) \ . \tag{3.18}$$

In the case of the flavour basis the neural network outputs up, down, strange and charm quarks and antiquarks distributions (for charm we consider $c = \bar{c}$) plus the gluon

$$\tilde{q}_k = \left[g, u, \bar{u}, d, \bar{d}, s, \bar{s}, c\right] . \tag{3.19}$$

When running a fit, the eight PDFs of the chosen basis are first supplemented with the perturbative heavy quarks distributions and then rotated back into the FK table basis of eq. (3.16).

The final result of a PDFs determination should be independent on the specific basis used in the fit and only determined by experimental data and general physical constraints. The way in which the latter are implemented in the fit can vary depending on the basis, and might define different final methodologies, which in turn could be more or less convenient in terms of numerical performances. It is therefore interesting to study different possible fit basis choices, discussing the differences in performances and methodologies and verifying that, at least in the kinematic regions where experimental data are available, the fit results are independent on the specific PDFs initially parameterized.

3.2.3 Theoretical constraints

As mentioned above, there are several theoretical conditions that can be used to further constrain PDFs. In the following we address each of them, discussing the way in which they are implemented when considering different fit bases.

Sum rules. As seen in sec. 2.1 in the context of the parton model, the valence structure of the proton implies the valence sum rules eq. (2.11). Additionally, energy conservation implies momentum sum rules, given by

$$\int_0^1 dx\, x\left(g\left(x\right) + \Sigma\left(x\right)\right) = 1\,. \tag{3.20}$$

In QCD these sum rules remain valid, and provide additional information to constrain parton distributions. We can rewrite eq. (2.11) in terms of the evolution basis

$$\int_0^1 dx\, V\left(x\right) = \int_0^1 dx\, V_8\left(x\right) = 3\,, \quad \int_0^1 dx\, V_3\left(x\right) = 1\,. \tag{3.21}$$

and implement these conditions in the fit multiplying the distributions V, V_3, V_8 and g by suitable normalization factors A_V, A_{V_3}, A_{V_8} and A_g defined as

$$A_V = A_{V_8} = \frac{3}{\int_0^1 dx\, V(x)}\,,\ A_{V_3} = \frac{1}{\int_0^1 dx\, V_3(x)}\,,\ A_g = \frac{1-\int_0^1 dx\, x\Sigma(x)}{\int_0^1 dx\, xg(x)}\,, \quad (3.22)$$

so that eqs. (3.20), (3.21) are automatically satisfied. Such multiplication happens after rotating the fit basis back into eq. (3.16), so that the procedure remains the same independently on the choice of the fit basis.

Additional information about PDFs can be obtained by the Gottfried sum rule, given by the integral

$$I_G = \int_0^1 \frac{dx}{x}\left[F_2^{lp} - F_2^{ln}\right]\,, \quad (3.23)$$

where F_2^{lp} and F_2^{ln} are the structure functions of lepton-proton and lepton-neutron DIS. Eq. (3.23) can be expressed as

$$\begin{aligned} I_G &= \frac{1}{3}\int_0^1 dx\ \left(u_v(x) - d_v(x)\right) + \frac{2}{3}\int_0^1 dx\ \left(\bar{u}(x) - \bar{d}(x)\right) \\ &= \frac{1}{3} + \frac{2}{3}\int_0^1 dx\ \left(\bar{u}(x) - \bar{d}(x)\right)\,. \end{aligned} \quad (3.24)$$

Experimental data from th NMC collaboration show a deviation of I_G from the nominal value of 1/3 [62], and provide measurements of the quark flavour asymmetry of the sea $\int dx\left(\bar{u} - \bar{d}\right)$. Unlike the case of the valence sum rules, no specific value for the Gottfried sum rule is imposed during the fit. However the value of I_G can be directly related to the integral of distribution T_3

$$I_G = \frac{1}{3}\int_0^1 dx\, T_3(x)\,. \quad (3.25)$$

Eq. (3.25) shows how, in order to have a well defined value for the Gottfried sum rule, the distribution T_3 has to be integrable in the interval $(0, 1)$, providing information for its small-x behaviour.

Large- and small-x behaviour. The large-x behaviour of the PDFs has to be consistent with the elastic limit, namely the condition $q_i(x = 1) = 0$ has to be satisfied by all the distributions of both the flavour and evolution basis. This can be easily implemented in the fit by supplementing the neural network

parameterization of each flavour by the preprocessing factor $(1-x)^{\beta_i}$, just like in the NNPDF methodology described before. The use of an additional polynomial factor $x^{-\alpha_i}$, controlling the small-x behaviour of each flavour, only makes sense when working in the evolution basis. In this case, each flavour is either a singlet or nonsinglet distribution, and as such its small-x behaviour can be classified as integrable (nonsinglet) or not integrable (singlet) in $x = 0$, as discussed in the following paragraphs: a parameterization like the one given in eq. (3.7) can be used, with α randomized in intervals such that $\alpha < 1$ in the former case and $\alpha > 1$ in the latter. On the other hand, when working in the flavour basis each distribution (except the gluon) has both a singlet and a nonsinglet component, which makes it impossible to use a polynomial expression to describe in a consistent way its small-x behaviour. Because of this, when working in the flavour basis only a large-x preprocessing is implemented, so that each distribution of eq. (3.19) is parameterized as

$$x\,\tilde{q}_j\left(x, Q_0^2\right) = (1-x)^{\beta_j}\,\mathrm{NN}_j\left(x\right)\,, \tag{3.26}$$

with $\mathrm{NN}_j\left(x\right)$ denoting the corresponding neural network output, just as in eq. (3.7).

Positivity. As recalled in chapter 2, PDFs are renormalization scheme dependent quantities. Despite at LO they can be interpreted as probabilities distributions, when considering QCD corrections such naive picture doesn't hold any more. This in general prevents PDFs from being positive definite objects. However, regardless of the PDFs sign and shape, cross sections for high-energy processes have to be positive, which means that fit solutions leading to negative cross-sections have to be discarded. This condition might have a non-negligible impact on the PDFs themselves, especially in those kinematic regions where no experimental data are available: functional forms giving negative cross sections cannot be physical solutions of the problem, and therefore should be discarded. In the old `c++` code such requirement is implemented through the use of Lagrange multipliers, penalizing fit solutions for which a set of chosen high-energy cross sections, denoted as positivity observables, result to be negative.

It would be highly beneficial to work in a renormalization scheme in which PDFs are positive definite also beyond LO: this would allow to implement directly the positivity of the distributions, without having to rely on a specific choice of positivity observables. This is indeed the idea of a recent paper [63], where such

positive renormalization scheme for PDFs is built. In the same paper, the Authors work out the relation between such positive renormalization scheme and the standard $\overline{MS}$ scheme, which is the one commonly used in PDFs determinations, and surprisingly they find out that NLO $\overline{MS}$ distributions for quarks, antiquarks and gluon are actually positive definite. This result has been accounted for in `n3fit`, where positivity is imposed at the PDFs level. In the following we describe how this feature is implemented in the code, and we assess the impact of such constraint on the large-x region of the PDFs.

For each distribution $\tilde{q}_k$ which has to be positive, namely for all the PDFs defining the flavour basis of eq. (3.19), we add to the total χ^2 a contribution defined as

$$\chi^2_{k,pos} = \Lambda_k \sum_i \text{Elu}_\alpha \left(-\tilde{q}_k \left(x_i, Q^2\right)\right) , \tag{3.27}$$

with $Q^2 = 5\,\text{GeV}^2$, x_i given by 10 points logarithmically spaced between $5 \cdot 10^{-7}$ and 10^{-1} and 10 points linearly spaced between 0.1 and 0.9. The Elu function is defined as

$$\text{Elu}_\alpha\left(t\right) = \begin{cases} t & \text{if } \; t > 0 \\ \alpha\left(e^t - 1\right) & \text{if } \; t < 0 \end{cases} , \tag{3.28}$$

with $\alpha = 10^{-7}$. When the distribution $\tilde{q}_k\left(x_i, Q^2\right)$ is negative, the total χ^2 will receive a positive contribution (a penalty) proportional to the corresponding Lagrange multipliers Λ_k. Therefore, during the minimization of χ^2_{tot} solutions corresponding to positive distributions q_k will be favoured. Note that such implementation works fine for both the evolution and flavour basis, the only difference being the linear transformation mapping the network outputs to the positive distributions to be used in eq. (3.27).

Integrability. Given the lack of data in the small-x region, when $x < 10^{-4}$ the PDFs are left largely unconstrained. Because of the redundant parameterization employed in the NNPDF methodology, this will result in an artificially big PDFs error in the small-x region. This is an important feature that a reliable PDFs set should have: in those kinematic regions where experimental data are missing, the PDFs error should increase accordingly. There are however some physical considerations that can be made to further constrain PDFs at small-x, which can be used to reduce the huge error band in this kinematic region. As seen in sec. 3.2.3, in order to have well defined momentum and valence sum rules

the distributions V, V_3, V_8, xg, $x\Sigma$ have to be integrable when x decrease to zero. Moreover, in order to have well defined Gottfried sum rules, also the integral of T_3 over the interval $(0, 1)$ has to be finite. The same is true also for the distribution T_8, assuming the same small-x behaviour for all the light sea quarks.

This means that, denoting as $q\left(x, Q_0^2\right)$ a generic integrable PDF at the fitting scale, as x decreases to zero such distribution cannot raise faster than $1/x$. In other words the following limit has to be verified

$$\lim_{x\to 0} xq\left(x, Q_0^2\right) = 0\,. \tag{3.29}$$

In order to satisfy numerically eq. (3.29) we require that, for a given set of points $x_{\text{integ}}^{(i)}$ in the small-x region, the function $xq\left(x\right)$ evaluated in these points is much smaller than its peak value,

$$\sum_i \left|xq\right|_{x=x_{\text{integ}}^{(i)}} \ll xq|_{x=x_{\text{peak}}}\,, \tag{3.30}$$

with x_{peak} denoting the point where the distribution xq reaches its maximum value. Eq. (3.30) can be rewritten introducing a numerical parameter f_q of order 10^{-1}

$$\sum_i \left|xq\right|_{x=x_{\text{integ}}^{(i)}} < f_q * xq|_{x=x_{\text{peak}}}\,. \tag{3.31}$$

Despite eq. (3.31) is not mathematically equivalent to eq. (3.29), the idea is that, if satisfied for small enough x values $x_{\text{integ}}^{(i)}$, the function $xq\left(x\right)$ will decrease to zero as x keeps getting smaller, and the integrals defining the sum rules will be well defined. Also, unlike eq. (3.29), the condition given in eq. (3.31) can be easily verified replica by replica, so that replicas not satisfying it can be discarded.

In a similar way to what done for positivity we can impose integrability. For each distributions q_k which has to be integrable we add to the total χ^2 a new bit defined as

$$\chi^2_{k,integ} = \Lambda_k \sum_i \left[x_i\, q_k\left(x_i, Q^2\right)\right]^2\,, \tag{3.32}$$

In this way, the fit will favour configurations with smaller values of $|x_i\, q_k\left(x_i, Q^2\right)|$, and should therefore produce distributions satisfying eq. (3.31). The points x_i are chosen in different ways depending on the specific distribution and fit basis

we are considering, as detailed in the next section.

3.2.4 Results

In the previous section we have described how theoretical constraints are implemented in the fit. In particular we have seen how, when imposing positivity and integrability, the total experimental χ^2 has to be supplemented by additional contributions proportional to Lagrange multipliers, so that the total χ^2 which is actually minimized during the fit is

$$\chi^2_{tot} = \chi^2_{exp} + \sum_k \chi^2_{k,\text{pos}} + \sum_l \chi^2_{l,\text{integ}} \tag{3.33}$$

with the indices k and l running on the positive and integrable distributions and $\chi^2_{k,\text{pos}}$, $\chi^2_{l,\text{integ}}$ defined in eqs. (3.27), (3.32).

In this section we present results produced with the new `n3fit` methodology, focusing on the effects positivity and integrability constraints have on the final result. Our point here is to show the effects of such theoretical constraints on a given baseline which doesn't include them. We will first present results obtained using the evolution basis, and in the next section we will repeat the exercise in the flavour basis, in order to check explicitly fit basis independence.

The baseline dataset we will consider here is the one of the NNPDF3.1 PDFs set, presented and discussed in ref. [48]. This includes: fixed-target neutral-current (NC) DIS structure function data from NMC [64, 65], SLAC [66] and BCDMS [67]; charged-current (CC) DIS structure function data from CHORUS [68] and NuTeV [69, 70]; HERA data from their combined measurements [71], including charm-production cross sections [72] and b-tagged structure functions [73, 74]; fixed-target Drell-Yan data from E866 [75–77] and E605 [78]; collider Drell-Yan data from CDF [79] and D0 [80–82]; and Drell-Yan, inclusive gauge boson, and top-pair production data from ATLAS [83–90], CMS [91–98] and LHCb [99–102]; ATLAS and CMS single-inclusive jet data [103, 104]. In total this baseline dataset contains $n_{\text{dat}} = 4287$ datapoints, see ref. [48] for more details.

When working in the evolution basis the small-x behaviour is largely controlled by the preprocessing polynomial factor. Eq. (3.31) can then be satisfied mainly by choosing the corresponding preprocessing exponents in an interval which ensures

integrability. Additionally, the Lagrange multiplier term given in eq. (3.32) is added to the total χ^2 using a single small-x point $x = 10^{-9}$. These choices have been proved to produce replicas which mostly satisfy eq. (3.31).

In tab. 3.1 we provide values of $\chi^2/N_{\rm dat}$ for a `n3fit` global fit, reporting results also for each experiment included in the analysis. These are compared to their `n3fit` counterpart produced without positivity and integrability constraints. Inspection of this table shows how the fit quality is basically unaffected by the introduction of the theoretical constraints, with a slightly lower total χ^2 when theory constraints are included. It should be noted that a deterioration of the χ^2 after introducing additional constraints in the fit would be expected, but this is not what observed here, where we observe a slight improvement in the total χ^2. Such mild improvement provides an additional confirmation of the validity of our theoretical constraints: despite these represent additional conditions to be satisfied during the fit, they provide meaningful physical hints, making it easier for the fit to describe the experimental data.

In fig. 3.3 we show the distance between the two fits, quantifying the effect of theoretical constraints at the PDFs level. For the definition of the distance between PDFs sets, see app. A. While in the small-x region the difference between the two fits remains below one-sigma, in the medium- and large-x regions we find differences of up to two-sigmas. The evolution basis flavours V, V_3, V_8 and T_8 are plotted in linear scale in fig. 3.4. Inspection of these plots show an important reduction of the PDFs error for the flavours V, V_3, V_8, and a change in the PDFs shape and central value to satisfy the new positivity constraints, which affects also the distribution T_8.

We conclude that overall after the inclusion of theoretical constraints the fit provides an equivalent, or slightly better, description of the input data. However there are important differences at the level of the PDFs, whose replicas are globally shifted in order to satisfy positivity, resulting in a general decreasing in the PDFs error, and in a non negligible change in their central value.

3.2.5 Fit basis independence

As described in sec. 3.2.2, different choices for the specific PDFs which are independently parameterized are possible. In this section we present results for a fit run in the flavour basis, using the same methodology implemented for the

Experiment	$N_{\rm dat}$	$\chi^2_{\rm evol}$	$\chi^2_{\rm flav}$	Baseline χ^2
NMC	325	1.247	1.258	1.280
SLAC	67	0.733	0.757	0.721
BCDMS	581	1.183	1.130	1.253
CHORUS	832	1.159	1.108	1.122
NTVDMN	76	0.969	1.332	0.961
HERACOMB	1145	1.150	1.125	1.168
HERAF2CHARM	37	1.522	1.572	1.492
F2BOTTOM	29	1.100	1.100	1.112
DYE886	104	1.410	1.534	1.560
DYE605	85	1.143	1.161	1.157
CDF	105	0.969	0.921	0.996
D0	48	1.505	1.362	1.409
ATLAS	360	1.105	1.082	1.126
CMS	408	1.035	1.056	1.052
LHCb	85	1.438	1.320	1.526
Total	4287	1.153	1.136	1.171

Table 3.1 *The values of $\chi^2/N_{\rm dat}$ for each experiment included in the global fit, before and after the inclusion of positivity and integrability constraints. Values are reported for fits in both the evolution and flavour basis. From left to right each column reports: the experiment, the number of datapoints N_{dat}, the value of the total χ^2 in the evolution and flavour basis when positivity and integrability constraints are considered and finally the χ^2 for the baseline fit (evolution basis without positivity and integrability constraints).*

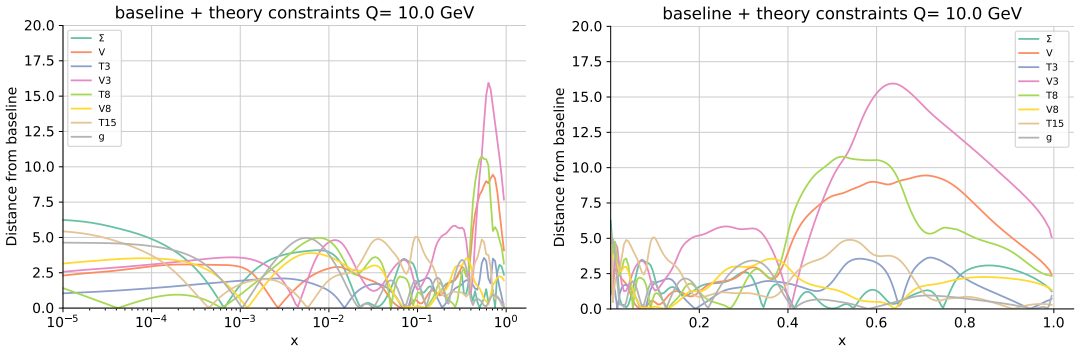

Figure 3.3 *Distance plots between the baseline fit and the one produced using positivity and integrability constraints.*

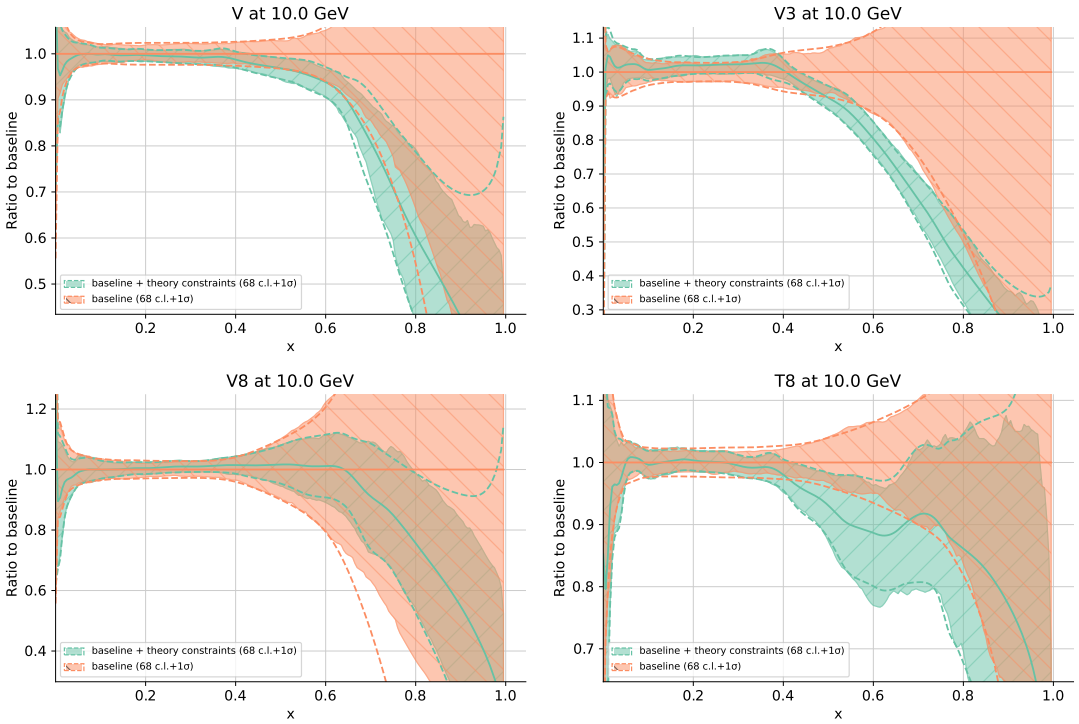

Figure 3.4 *PDFs plots: baseline fit (orange) and fit with positivity and integrability constraints (green). See eq.* (3.9) *for the definition of the distributions plotted here.*

evolution basis (same minimization algorithm and architecture parameters) and imposing the theoretical constraints as described previously.

As mentioned in sec. 3.2.3, when working in the flavour basis no small-x preprocessing is used. This implies that the small-x behaviour of the PDFs, which is unconstrained by experimental data, can only be controlled through Lagrange multipliers. Because of this reason, for integrable distributions more stringent

constraints are used than those implemented for the evolution basis. In particular, in order to get replicas satisfying eq. (3.31), the Lagrange multiplier terms given in eq. (3.32) are built using the three small-x points $x_i = 10^{-5}, 10^{-4}, 10^{-3}$.

The fit quality is again basically unchanged, with a total χ^2 which is slightly better than the corresponding value in the evolution basis: $\chi^2_{\rm flav} = 1.13$ to be compared with the value $\chi^2_{\rm evol} = 1.15$ from tab. 3.1. The resulting PDFs plotted in the evolution basis are reported in fig. 3.5, where they are compared with the evolution basis fit presented in the previous section. The plots are shown in linear scale, in order to highlight the x-region where more experimental data are available. By inspection of fig. 3.5 it is clear that fit basis independence is achieved, showing how, as expected, as long as data are present and physical constraints given by positivity, integrability and sum rules consistently implemented, different choices for the fit basis do not change the final PDFs. It is worth stressing that this result, even if expected, is not trivial. When we change fit basis a number of settings and parameters are changed at the same time. In particular, no small-x preprocessing is used for the flavour basis, which therefore presents a lower number of free parameters. The fact that we still get equivalent results provides a strong validation check of the `n3fit` environment.

In order to determine the best methodology, one should fix the input dataset and the fit basis. Once this is done, a scan of the hyperparameters defining the final methodology has to be performed (neural network architecture, minimization algorithm, positivity and integrability parameters ...), in order to determine the corresponding values which allow the best fit. The methodology used to produce the results presented here has been optimized considering the evolution basis, which will be used in the final NNPDF4.0 release. If results in the flavour basis were to be used to do actual phenomenology, an additional hyperparameters scan should be run, to ensure the best possible performance of the methodology. Here we have shown how, without explicitly doing an additional hyperoptimization, the `n3fit` environment still allows to obtain a good fit using a different basis, providing a proof of concept of basis independence.

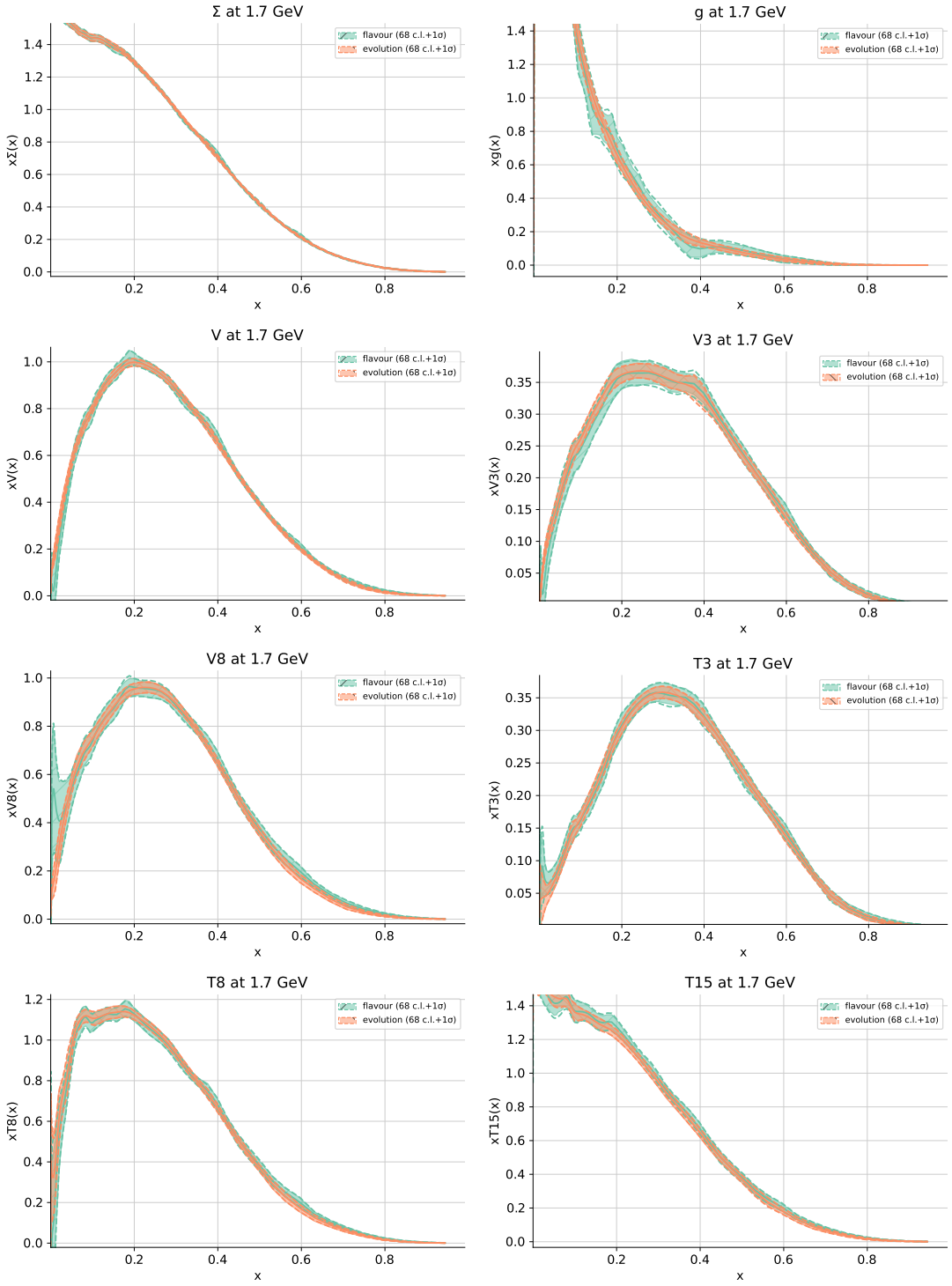

Figure 3.5 *PDFs plots comparing results for fits performed in the evolution (orange) and flavour (green) basis.*

4

Theory interpretation of jets production at the LHC

The inclusion of new experimental data is the central ingredient in any global PDFs determination. New data provide additional constraints on PDFs and allow to get more and more precise results. In this chapter, based on ref. [7], we present a systematic analysis for the inclusion of jets cross-sections in a global parton distributions determination. We will use this specific case to give an example of the general procedure which is usually followed when considering new experimental data, using the recent NNLO QCD computations supplemented by electroweak (EW) corrections to assess the impact of jets and dijets production measurements on the PDFs.

The choice of the most suitable jest observable to be considered in global PDFs determination and, more in general, for precision QCD studies presents a number of open theoretical issues, which makes the inclusion of jest data a particular interesting case. On one hand, the simplest inclusive observable, the single-inclusive jets cross-section [105, 106], turns out to be non-unitary. A possible alternative is offered by the dijets cross-section which however, despite appearing to be unitary and especially well suited for PDFs determination [107], at NLO displays a significant scale dependence. Thanks to the recent NNLO computation for these observables, this last problem has essentially been settled, with the scale dependence of dijets cross-section being under control at NNLO. On the other hand, the single-jet inclusive cross section shows a scale dependence which is not reduced when going to NNLO [108], showing how the perturbative behaviour, the scale dependence [109] and even the definition [110] of this observable are non-trivial.

In this study we address these issues from a phenomenological point of view in the context of PDFs determination. We study the impact of both single-inclusive jets and dijets cross-sections in a global parton distributions fit and we asses which observable leads to better PDFs compatibility with other data, better fit quality, and more stringent constraints on the parton distributions. With this study we aim to provide a guideline for the inclusion of jets observables in a future global fit, like for example NNPDF4.0.

The chapter is structure as follows. In sec. 4.1 we describe the data included in the analysis, together with their kinematic coverage; in sec. 4.2 we briefly discuss the main aspects of the theoretical computation of jets observables, describing the scale choices and the way in which NNLO QCD predictions and EW corrections are implemented in the fit; finally in sec. 4.3 we present our results, consisting in a series of global PDFs fits where different jets observables are included.

4.1 Jets data from ATLAS and CMS

The ATLAS and CMS collaborations have performed a number of measurements of single-inclusive jets and dijets cross sections, with center of mass energies ranging from $\sqrt{s} = 2.76$ to 13 TeV. In this work we will consider data at $\sqrt{s} = 7$ and 8 TeV. Whereas recent global PDFs determinations include some jets data, like for instance NNPDF3.1, which includes ATLAS and CMS single-inclusive data with $\sqrt{s} = 2.76$ and 7 TeV, this is the first time that the full LHC-Run I jet dataset is being considered. In particular dijets data have not been included in any other previous analysis. Also, in the NNPDF3.1 determination, theory predictions for the included jets data were computed by combining NLO coefficient functions with NNLO perturbative evolution. In order to account for the missing NNLO corrections an additional uncertainty was estimated through scale variations and added to the jets data. In this work we will use the full NNLO QCD computation, as detailed in sec. 4.2.

The specific features of the data considered here are summarized in table 4.1: for each dataset we reported the centre of mass energy $\sqrt{s}$, the integrated luminosity $\mathcal{L}$, the jet radius R, the measured differential distribution and the number of datapoints $n_{\rm dat}$. The relevant kinematic variables are defined as follows. For single-inclusive jets we denote as p_T and y the jet transverse momentum and rapidity. For dijets, m_{jj} is the invariant dijet mass, y^* and $|y_{\rm max}|$ are the

absolute rapidity difference and the maximum absolute rapidity of the two leading jets of the event, defined as $y^* = |y_1 - y_2|/2$ and $|y_{\rm max}| = \max(|y_1|, |y_2|)$ respectively. Finally, considering the dijets triple differential distribution, $p_{T,\rm avg} = (p_{T_1} + p_{T_2})/2$ is the average transverse momentum of the two leading jets and $y_b = |y_1 + y_2|/2$ is the boost of the dijets system.

The ATLAS 7 TeV data for single-inclusive jets, given as distributions differential in transverse momentum p_T and rapidity y, cover the kinematic range 100 GeV $\leq p_T \leq$ 1.992 TeV, $0 \leq |y| \leq 3$, while the ATLAS 8 TeV data cover the same range in rapidity and an extended transverse momentum kinematic range 70 GeV $\leq p_T \leq$ 2.5 TeV. In our default fit we include only the central rapidity bin ($y_{jet} \leq 0.5$) of ATLAS 7 TeV, for ease of comparison with the NNPDF3.1 analysis of ref. [48], where the same choice was adopted due to the difficulty in achieving a good description of the complete set of rapidity bins using the default experimental covariance matrix[1]. The CMS 7 TeV data are available for 100 GeV $\leq p_T \leq$ 2.0 TeV, $0 \leq |y| \leq 2.5$, and the CMS 8 TeV cover the extended ranges 74 GeV $\leq p_T \leq$ 2.5 TeV and $0 \leq |y| \leq 3.0$.

Moving to dijets cross-sections, in the case of ATLAS 7 TeV the measurements are double-differential in m_{jj} and y^*, with 260 GeV $\leq m_{jj} \leq$ 4.27 TeV and $0 \leq y^* \leq 3.0$, while for CMS 7 TeV the distributions are differential in m_{jj} and $|y_{\rm max}|$, with 200 GeV $\leq m_{jj} \leq$ 5 TeV and $0 \leq |y_{\rm max}| \leq 2.5$. Finally the CMS 8 TeV data are triple-differential in $p_{T,\rm avg}$, y_b and y^* with ranges 133 GeV $\leq p_{T,\rm avg} \leq$ 1.78 TeV and $0 \leq y_b, y^* \leq 3$. Note that ATLAS dijets measurements are currently available at 7 and 13 TeV but not at 8 TeV.

In addition to the datasets listed in table 4.1 ATLAS and CMS have performed measurements at $\sqrt{s} = 13$ TeV for both single-inclusive jet [112, 113] and dijets [112, 114]. These however have smaller integrated luminosities and for this reason we do not include these datasets in the analysis. Finally several measurements for multijets production are also available, with ATLAS providing differential distributions for three jets cross-sections at 7 TeV [115] and four jets cross-sections at 8 TeV [116] and CMS for three jets at 7 TeV [117]. However theoretical predictions for these observables are currently available only up to NLO, and therefore they will not be considered here.

For all the measurements considered here, the complete set of systematic

[1] In refs. [48, 111] this choice was validated, showing how PDFs determined from each rapidity bin in turn are indistinguishable.

Experiment	Measurement	$\sqrt{s}$ [TeV]	$\mathcal{L}$ [fb^{-1}]	R	Distribution	$n_{\rm dat}$	Reference
ATLAS	Inclusive jets	7	4.5	0.6	$d^2\sigma/dp_T d\lvert y\rvert$	140	[103]
CMS	Inclusive jets	7	4.5	0.7	$d^2\sigma/dp_T d\lvert y\rvert$	133	[118]
ATLAS	Inclusive jets	8	20.2	0.6	$d^2\sigma/dp_T d\lvert y\rvert$	171	[119]
CMS	Inclusive jets	8	19.7	0.7	$d^2\sigma/dp_T d\lvert y\rvert$	185	[120]
ATLAS	Dijets	7	4.5	0.6	$d^2\sigma/dm_{jj} d\lvert y^*\rvert$	90	[121]
CMS	Dijets	7	4.5	0.7	$d^2\sigma/dm_{jj} d\lvert y_{\rm max}\rvert$	54	[118]
CMS	Dijets	8	19.7	0.7	$d^3\sigma/dp_{T,\rm avg} dy_b dy^*$	122	[122]

Table 4.1 *The LHC single-inclusive jet and dijet cross-section data that will be used in this study. For each dataset we indicate the experiment, the measurement, the center of mass energy $\sqrt{s}$, the luminosity $\mathcal{L}$, the jet radius R, the measured distribution, the number of datapoints $n_{\rm dat}$ and the reference.*

uncertainties and correlations available from `HepData` have been used.

4.2 Theoretical calculations

In this section we present the main aspects of the theoretical calculations used to perform our phenomenological study, discussing scale choices and QCD corrections up to NNLO. We also discuss EW corrections and the way in which they are combined with QCD predictions for the purpose of PDFs determination.

4.2.1 Scale choice

As mentioned at the beginning of this chapter, even when considering NNLO predictions the single-inclusive jet cross-sections are in general rather sensible to the choice of central scale. Three possible choices are given by the individual jet transverse momentum p_T, the leading jet transverse momentum $p_{T,1}$ and the scalar sum of the transverse momenta of all the partons in the event

$$\hat{H}_T = \sum_{i \in \text{partons}} p_{T,i} \,. \tag{4.1}$$

Predictions obtained from different scale choices may differ, even at NNLO, by amounts which are comparable to their scale dependence. In ref. [109] the scales $\mu = \hat{H}_T$ and $\mu = 2p_T$ were singled out as optimal ones, according to a number of criteria, such as perturbative convergence and scale uncertainty as error estimate. Here we will consider results for $\mu = \hat{H}_T$.

Turning to dijets observables, also here different scale choices are possible. As mentioned before, at NLO theoretical predictions computed with different choices differ significantly, however the problem is alleviated at NNLO, with $\mu = m_{jj}$ emerging as the preferred choice [123, 124], which therefore will be adopted here.

4.2.2 QCD corrections

Exact NNLO QCD predictions have been computed using `NNLOJET` [125]. As detailed in chapter 3, the partonic matrix elements entering a PDFs fit have to be precomputed in such a way that the numerical convolution with generic input PDFs can be approximated by means of interpolation techniques. To this purpose we use `NLOJET++` [126] interfaced to `FASTNLO` [127]. The computation is performed using the scale choices described above and is validated against the `NNLOJET` computation. This fast interpolation grids are then combined with PDF evolution kernel using `APFELGRID` [59] to obtain FK tables, as described in eq. (3.15). However fast interpolation grids to be used as input for `APFELGRID` are available only at NLO. We therefore implement NNLO QCD corrections by supplementing the NLO grids with QCD K-factors as detailed in the following. We define the NNLO QCD K-factors as

$$K_{\rm NNLO}^{\rm QCD} = \frac{\sum_{ij} \hat{\sigma}_{ij}^{\rm NNLO} \otimes \mathcal{L}_{ij}^{\rm NNLO}}{\sum_{ij} \hat{\sigma}_{ij}^{\rm NLO} \otimes \mathcal{L}_{ij}^{\rm NNLO}}\,, \tag{4.2}$$

where the sum runs over partonic subchannels, $\hat{\sigma}_{ij}$ are partonic matrix elements and $\mathcal{L}$ the corresponding parton luminosity, computed in both the numerator and denominator using NNPDF3.1 NNLO as a fixed input PDF set. NNLO grids for the relevant cross-sections are then obtained from the corresponding NLO grids through the multiplicative prescription

$$\left.\frac{d^2\sigma}{dp_T dy}\right|_{NNLO} = \left.\frac{d^2\sigma}{dp_T dy}\right|_{\rm NLO_{QCD}} \times K_{\rm NNLO}^{\rm QCD}(p_T, y, \sqrt{s})\,. \tag{4.3}$$

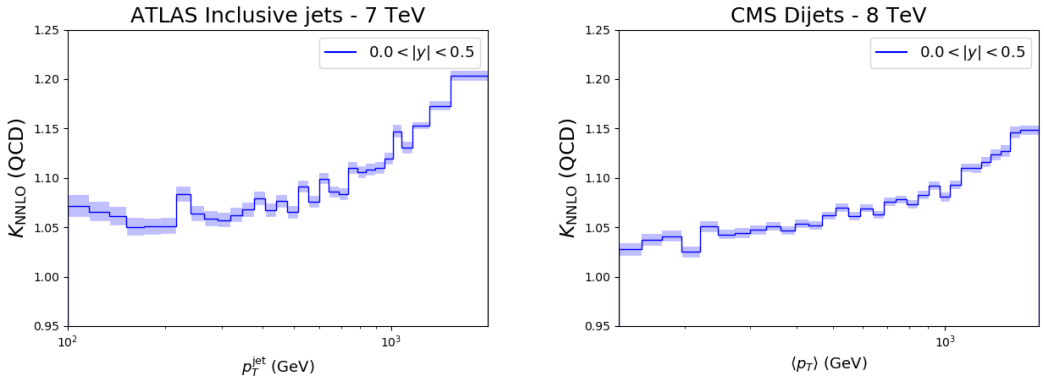

Figure 4.1 *The NNLO QCD K-factors for the central rapidity bins of the ATLAS 7 TeV single-inclusive jets (left) and CMS 8 TeV dijets (right), with the Monte Carlo numerical uncertainties shown as filled bands around the central result. Figure from ref. [7].*

The first term on the right-hand side is the output of the NLO computation, given in terms of fast interpolation grids, while the second term is the bin-by-bin QCD K-factors computing according to eq. (4.2). The result is validated against the full NNLO result from `NNLOJET`. In general QCD K-factors might be affected by point-to-point fluctuations due to underlying numerical uncertainties affecting the NNLO computation, therefore they are provided with a Monte Carlo uncertainty, estimated as described in ref. [128]. This uncertainty is added in quadrature to the experimental one when performing the PDFs fit, fully uncorrelated datapoint by datapoint. For illustration purpose, NNLO QCD K-factors for the central rapidity bins $0 \leq |y| \leq 0.5$ of the ATLAS 7 TeV single-inclusive jet and of the CMS 8 TeV dijet distributions are displayed in fig. 4.1 as functions of p_T, together with the corresponding Monte Carlo uncertainties. In both cases the NNLO QCD K-factors increase monotonically with p_T from about 5% to about 20% for single-inclusive jets, and from about 3% to 15% for dijets.

4.2.3 Electroweak corrections

The EW corrections for all the single-inclusive jet and dijets datasets considered here have been determined using the computation of ref. [129]. These include the $\mathcal{O}(\alpha\alpha_s)$ and $\mathcal{O}(\alpha^2)$ tree level contributions and the $\mathcal{O}(\alpha\alpha_s^2)$ weak radiative corrections, where α and α_s denote the weak and strong coupling respectively. As in the case of NNLO QCD corrections, EW contributions are included by mean

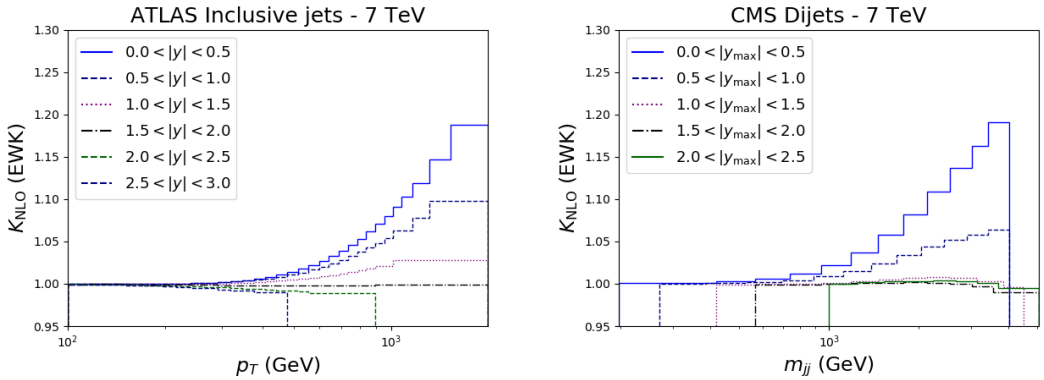

Figure 4.2 *The EW K-factors, eq.* (4.4)*, for the 7 TeV ATLAS and CMS single-inclusive (left) and dijet (right) measurements. For single-inclusive jets the K-factors are shown as a function of jet p_T in six different rapidity bins. For dijets they are shown as a function of the dijet invariant mass m_{jj} for different $y_{\max}$ bins. Figure from ref. [7].*

of K-factor, defined as

$$K^{\text{EW}} = \frac{\sum_{ij} \hat{\sigma}_{ij}^{\text{LO QCD+EW}} \otimes \mathcal{L}_{ij}^{\text{NNLO}}}{\sum_{ij} \hat{\sigma}_{ij}^{\text{LO QCD}} \otimes \mathcal{L}_{ij}^{\text{NNLO}}}\,, \tag{4.4}$$

where the partonic cross sections have been obtained combining the computation of ref. [129] with LO QCD results. Fast interpolation grids accounting for EW corrections can be compute supplementing eq. (4.3) with the EW K-factor of eq. (4.4), getting

$$\left.\frac{d^2\sigma}{dp_T dy}\right|_{NNLO+EW} = \left.\frac{d^2\sigma}{dp_T dy}\right|_{\text{NLO}_{\text{QCD}}} \times K_{\text{NNLO}}^{\text{QCD}}(p_T, y, \sqrt{s}) \times K_{\text{NNLO}}^{\text{EW}}(p_T, y, \sqrt{s})\,. \tag{4.5}$$

Electroweak K-factors have been computed using a proprietary code, with NNPDF3.1 NNLO PDF set as input. In fig. 4.2 representative plots are shown for ATLAS 7 TeV single-inclusive jet and CMS 7 TeV dijest, as functions of p_T (single-inclusive jet) and m_{jj} (dijets), in bins of rapidity y or maximum absolute rapidity $y_{\max}$. In both cases, K-factors are close to one for small values of p_T or m_{jj}, they are mostly flat for large values of the rapidity variable while they grow with p_T or m_{jj} for the central rapidity bin, reaching values as high as 20%. EW K-factors for the other distributions, not displayed here, present similar features.

4.3 Results

In this section we present the main results of our study, consisting in a series of PDF sets in which the global NNPDF3.1 dataset described in sec. 3.2.4 has been supplemented with subsets of the jets and dijets data described in sec. 4.1, after removing all the jets data included in the original NNPDF3.1 PDFs set. We consider a series of different scenarios where we vary the jet observable and the input data. Specifically we have performed fit including either single-inclusive jets or dijets data, in each case considering either the full dataset or the 7 TeV or 8 TeV data only. As described in the previous section we use NNLO QCD theoretical computations supplemented by EW corrections, with the scale choice $\mu = \hat{H}_T$ for single-inclusive jet and $\mu = m_{jj}$ for dijets.

In table 4.2 we report the full list of fits which will be discussed in the following, together with an ID that will be used to identify them. The ID encodes the process used (j for single-inclusive jets and d for dijets); the data used (a for all, 7 or 8 for the 7 TeV or 8 TeV datasets); finally "n" stands for the perturbative accuracy NNLO QCD and "w" reminds that EW corrections are included; Each row of the table corresponds to a different input datasets, with the fit #bn representing the baseline, where no jets data are included.

In all these fits the systematic uncertainties are implemented as fully correlated across bins of different kinematic variables, while statistical uncertainties are correlated only across bins of transverse momentum (for jets) or invariant mass (for dijets). Following the standard NNPDF methodology, multiplicative uncertainties are treated with the t_0-method [56] and all the fits have been iterated once in order to ensure convergence of preprocessing and t_0 method. The fits have been run using the standard NNPDF methodology described in sec. 3.1, employing the `c++` framework used to produce the NNPDF3.1 PDF set. All PDF sets discussed contains $N_{\rm rep} = 100$ Monte Carlo replicas, and the `ReportEngine` software [130] is used to analyze results and compute various fit metrics and statistical estimators.

In table 4.3 we report the χ^2 values for all the fits with single-inclusive jet data listed in table 4.2, while in table 4.4 we report values for dijets fits. In both cases values reported in square brackets are referred to points not included in the corresponding fit. We show the χ^2 values for all the data in the global dataset, grouped by process type (DIS NC, DIS CC, Drell-Yan, Z p_T, top pair) and for

baseline (no jets data)	bn
ATLAS & CMS jets 7-8 TeV	janw
ATLAS & CMS jets 7 TeV	j7nw
ATLAS & CMS jets 8 TeV	j8nw
ATLAS & CMS dijets 7-8 TeV	danw
ATLAS & CMS dijets 7 TeV	d7nw
CMS dijets 8 TeV	d8nw

Table 4.2 *The PDF determinations discussed in this study and their IDs. Each row corresponds to a different choice of input jet dataset, specified in the first column. The ID encodes the process used (j for single-inclusive jets and d for dijets); the data used (a for all, 7 or 8 for the 7 TeV or 8 TeV datasets); the perturbative accuracy (n for QCD NNLO, w for EW corrections). In this and subsequent tables and plots "jets" stands for single-inclusive jets.*

all jets data, both those included in the fit and those which are not.

4.3.1 Single-inclusive jets

We first analyze results concerning the inclusion of single-inclusive jet data only. In fig. 4.3 (left) we show the distances between the central values of the fits #bn and #janw, namely the baseline not including any jets data and the one including all single-inclusive jet cross-sections. From the distances plot it is clear how single-inclusive jet data have an impact only on the gluon distribution, the most affected regions being $x \simeq 0.05$, $0.1 \lesssim x \lesssim 0.2$, and $0.3 \lesssim x \lesssim 0.5$, with the gluon PDF changing by up to almost one sigma. Looking at the PDFs plot in the right panel of fig. 4.3 we notice how at small-x the gluon distribution is suppressed by about 2% and enhanced by about 4% in the large-x region. Looking at the χ^2 values in table 4.3, individual jet datasets show a χ^2 per datapoint of order one, with the only exception for the 8 TeV ATLAS data, for which we get $\chi^2 = 3.22$. It is interesting to note that the inclusion of singe-inclusive jet data leads also to an improvement of the description of dijets data, for which #janw shows better χ^2 than the baseline fit #bn. We observe a mild deterioration of the χ^2 for top pair processes, with a value of 1.25 to be compared to 1.05 of the baseline. A closer investigation shows that this comes from the deterioration in the description of the ATLAS top pair rapidity distributions, whose χ^2 per datapoint increases from 1.22 to 2.01.

Dataset	n_{dat}	bn	janw	j7nw	j8nw
DIS NC	2103	1.17	1.18	1.17	1.18
DIS CC	989	1.10	1.11	1.10	1.11
Drell-Yan	577	1.33	1.30	1.31	1.31
Z p_T	120	1.01	1.02	1.02	1.03
Top pair	24	1.05	1.25	1.02	1.24
Jets (all)	520	[2.60]	1.88	[2.53]	[1.89]
Jets (fitted)		—	1.88	1.12	2.20
ATLAS 7 TeV	31	[1.87]	1.59	1.15	[1.62]
ATLAS 8 TeV	171	[5.01]	3.22	[4.58]	3.25
CMS 7 TeV	133	[1.06]	1.09	1.11	[1.14]
CMS 8 TeV	185	[1.59]	1.25	[1.80]	1.23
Dijets (all)	266	[3.07]	[2.10]	[2.56]	[2.22]
Dijets (fitted)		—	—	—	—
ATLAS 7 TeV	90	[2.47]	[1.95]	[1.97]	[2.01]
CMS 7 TeV	54	[2.40]	[2.08]	[2.12]	[2.15]
CMS 8 TeV	122	[3.81]	[2.21]	[3.20]	[2.39]
Total		1.18	1.28	1.17	1.27

Table 4.3 *The χ^2 per datapoint for all fits of table 4.2 including single-inclusive jet data, with default settings. Results are shown for all datasets, aggregated by process type. For jets data, results are shown both for the sets included in each fit, and also for those not included, enclosed in square brackets. Combined results are also shown for all single-inclusive jet and for all dijet data, both for the full set, and for those included in each fit. The number of datapoints in each dataset is also shown.*

We can asses the impact of different datasets considering fits with 7 TeV or 8 TeV data only, denoted as #j7nw and #j8nw respectively. From table 4.3 we see how the unsatisfactory description of the 8 TeV ATLAS data persists even when no 7 TeV data are included in the analysis, suggesting a possible problem with the dataset itself rather than the presence of internal tension with the 7 TeV measurements. A significant difference between 7 and 8 TeV datasets is that for the former only the central rapidity bin is considered, as mentioned in sec. 4.1, while for the latter all the rapidity bins are included. This suggests that the 8 TeV data may also be affected by similar issues in the treatment of correlations between rapidity bins as those observed in the 7 TeV case in refs. [48]. This problem is addressed in app. B where we will see that this is indeed the case. Looking at the χ^2 for top pair processes, we note how its deterioration with respect to the baseline value comes entirely from the inclusion of 8 TeV data.

Dataset	n_{dat}	bn	danw	d7nw	d8nw
DIS NC	2103	1.17	1.18	1.17	1.18
DIS CC	989	1.10	1.12	1.09	1.12
Drell-Yan	577	1.33	1.29	1.32	1.28
Z p_T	120	1.01	1.07	1.03	1.08
Top pair	24	1.05	1.14	1.04	1.26
Jets (all)	520	[2.60]	[2.06]	[2.70]	[2.14]
Jets (fitted)		—	—	—	—
ATLAS 7 TeV	31	[1.87]	[1.63]	[1.74]	[1.61]
ATLAS 8 TeV	171	[5.01]	[3.36]	[4.65]	[3.55]
CMS 7 TeV	133	[1.06]	[1.06]	[1.14]	[1.07]
CMS 8 TeV	185	[1.59]	[1.64]	[2.17]	[1.68]
Dijets (all)	266	[3.07]	1.65	[2.16]	[1.71]
Dijets (fitted)		—	1.65	1.72	1.68
ATLAS 7 TeV	90	[2.47]	1.76	1.78	[1.78]
CMS 7 TeV	54	[2.40]	1.60	1.63	[1.66]
CMS 8 TeV	122	[3.81]	1.58	[2.67]	1.68
Total		1.18	1.22	1.19	1.20

Table 4.4 *Same as table 4.3, but now for dijets. The baseline is repeated for ease of reference.*

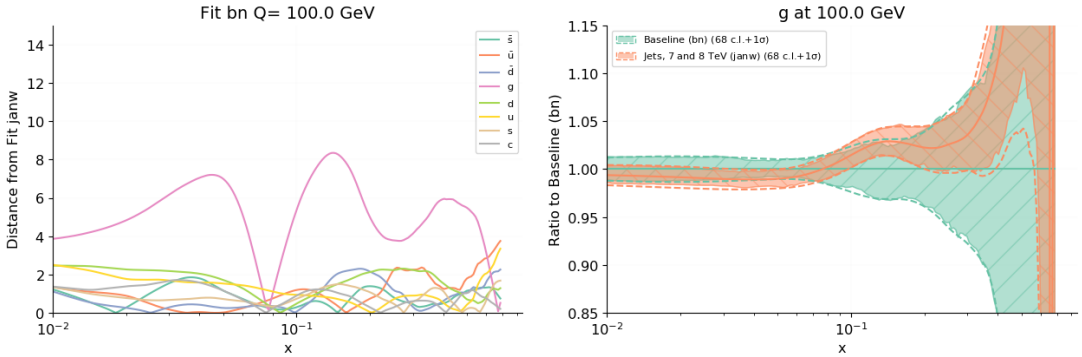

Figure 4.3 *Comparison between the baseline fit with no jet data (#bn) and the fit with all single-inclusive jet data included (#janw). The distance between all PDFs (left) and the ratio of the gluon PDF to the baseline (right) are shown at the scale $Q = 100$ GeV. The shaded band is the 68% confidence interval, while the dashed lines are the edge of one sigma interval.*

In fig. 4.4 we plot the gluon PDF and the corresponding error for the fits with 7 TeV and 8 TeV data only. In both cases the results show an enhancement of the central gluon PDF at large-x and a suppression at small-x, and a general reduction in the PDF uncertainty, which is more marked in the case of the 8

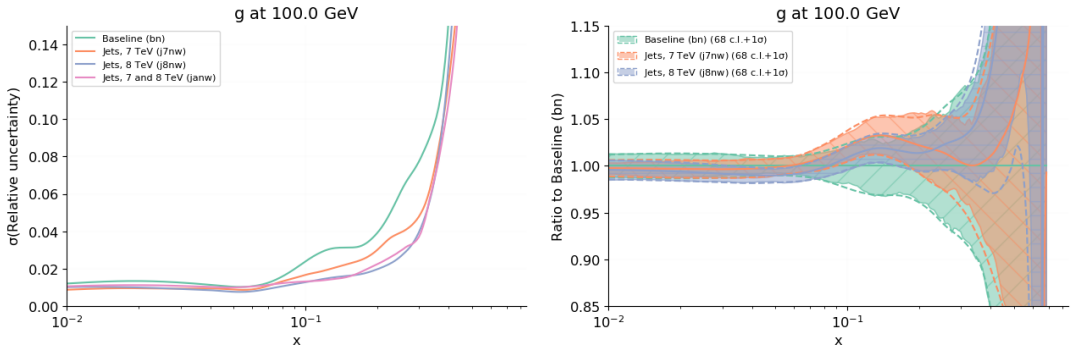

Figure 4.4 *Comparison between the baseline fit with no jet data (#bn), and the fits with only 7 TeV (#j7nw) or only 8 TeV (#j8nw) jet data included. The relative uncertainty on the gluon PDF (left) and the ratio of the gluon PDF to the baseline (right) are shown at $Q = 100$ GeV. All results are shown as ratios to the baseline.*

TeV data. In general, results obtained including 8 TeV data only are very close to those of the fit #janw, showing how 8 TeV datasets provide the dominant contribution driving the impact of single-inclusive jet data on the final PDFs.

4.3.2 Dijets

We now turn to PDFs in which dijets data rather than single-inclusive jet data have been included. As done for the single-inclusive jet data, we start comparing the baseline #bn, where no jets data are considered, to #danw, in which all dijets data are included using NNLO QCD computations with EW corrections. From table 4.4 we see how all the dijets datasets are fairly well described, with χ^2 values for datapoint around 1.6 for each individual dataset. Also, the inclusion of dijets data leads to an improvement in the description of single-inclusive jet data, consistently with what observed in sec. 4.3.1, where we noticed how the inclusion of single-inclusive jet data leads to a better description of dijets data as well. These features suggest that single-inclusive and dijet data have a similar impact on PDFs, and show consistency between data for these two observables. Also, unlike the case of single-inclusive jet data, no tension is observed between dijets data and the baseline dataset, whose χ^2 is left almost unchanged.

In the left panel of fig. 4.5 we show the distances between fits #bn and #danw: again only the gluon PDF is affected by the inclusion of the jets data, with the regions $x \simeq 0.01$ and $0.06 \lesssim x \lesssim 0.4$ being the ones showing the largest effects. Looking at the gluon PDF plot in the right panel of fig. 4.5, we observe a

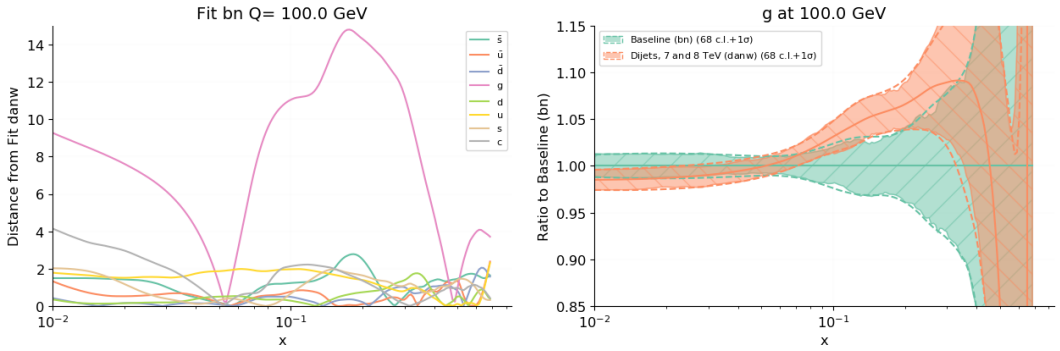

Figure 4.5 *Same as fig. 4.3, but now for dijets.*

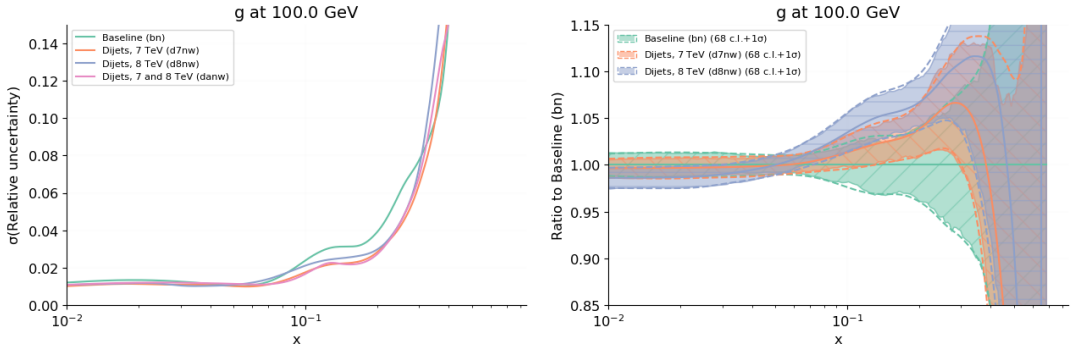

Figure 4.6 *Same as fig. 4.4, but now for dijets.*

suppression in the former region by about 2%, corresponding to a down shift of the central value by about one sigma, and an enhancement by about 10% in the latter, around $x \sim 0.3$, corresponding to an up shift of the central value by more than one sigma. These features are qualitatively similar to those observed in sec. 4.3.1 upon inclusion of single-inclusive jet data, but somewhat more pronounced.

We can study the relative impact of different datasets by studying results for the fits #j7nw and #j8nw, where either the 7 TeV or 8 TeV data only are included. By inspection of fig. 4.6, where we plot the gluon PDF and the corresponding error, we see how the impact of the two datasets on the gluon error and central value is qualitatively the same, and therefore qualitatively equivalent to the one of the full dijets dataset, with the 8 TeV data having a stronger impact. From table 4.4 we observe how the fit quality is equally good for the two fits. However the fit including 8 TeV data leads to a similar description of all the dijets data, including those which are not included in either fits, to the one given by #danw, where all dijets data are included. So once again we conclude that the 8 TeV data provide the dominant contribution.

4.3.3 Single-inclusive jets vs. dijets

Having assessed the impact on PDFs of jets and dijets datasets separately, we now compare results.

The effect on PDFs of the inclusion of jet and dijets in the NNPDF3.1 global dataset is qualitatively the same: they only affect the gluon, by leading to an enhancement of its central value in the region $0.1 \lesssim x \lesssim 0.4$, and to a suppression in the region $0.01 \lesssim x \lesssim 0.1$. The suppression is by about 1%, while the enhancement at the peak, localized at $x \simeq 0.3$, is by about 2.5% for single-inclusive jets, but stronger, by about 7.5% for dijets. These features are clearly visible in fig. 4.7 (right), where the gluon PDF is plotted for the fits #janw (single-inclusive jet only), #danw (dijets only) and the baseline #bn (no jets data).

As for the gluon PDF uncertainty, from the left panel of fig. 4.7 it is clear how the inclusion of either single-inclusive jet or dijets leads to a reduction of the error, with a stronger reduction observed in the case of single-inclusive jets. In this respect it should be observed that ATLAS dijet measurements are not yet available at 8 TeV, while single-inclusive jet measurements are available both from ATLAS and CMS. The constraining power of dijets datasets is therefore more limited.

As for compatibility with the global NNPDF3.1 datasets, the inclusion of jets data does not lead to a deterioration of the description of the rest of the data in comparison to the baseline fit, as we can see looking at the χ^2 values reported in tables 4.3, 4.4. The only exception is the ATLAS top rapidity distribution, which, as mentioned before, seems to be in tension with the 8 TeV single-inclusive jet data.

Concerning the fit quality, the quality of the two fits #janw and #danw to the corresponding jets data is comparable, though slightly better for the latter ($\chi^2 = 1.88$ vs. $\chi^2 = 1.65$). Also, the quality of the fit to dijets when single-inclusive jets are fitted and conversely are almost identical ($\chi^2 = 2.10$ for dijets when fitting single-inclusive jets vs. $\chi^2 = 2.06$ for single-inclusive jets when fitting dijets) and, as observed before, better than what we get for the baseline. This confirms full consistency between the two datasets, with a marginal preference for dijets.

Finally we note how the fit including dijets data is somewhat more internally

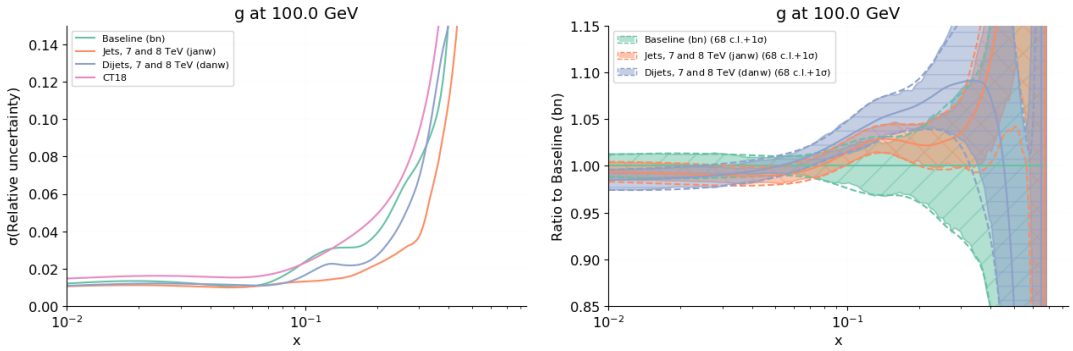

Figure 4.7 *Same as fig. 4.3, but now comparing the fits with all single-inclusive jet data (#janw), and that with all dijet data (#danw). In the gluon comparison (right) results are displayed as a ratio to the baseline with no jet data included (also shown for reference).*

consistent than the one including single-inclusive jets: the χ^2 per datapoint is slightly better (1.22 vs. 1.28) and the χ^2 for individual dataset is generally better, in particular for top production data.

To sum up, in this chapter we have presented a phenomenological investigation of inclusive jets production measurements at LHC in the context of global PDFs determination, exploiting recent NNLO QCD theoretical calculations supplemented by EW corrections, and studying for the first time the impact of the inclusive dijets observables. We have found full consistency between the impact on parton distributions of dijets and single-inclusive jet data, thus establishing the viability of the dijets observable in constraining PDFs. In a comparative assessment of single-inclusive jets vs. dijets we have found how, given the currently available data, the latter has a more marked impact on the central value of the gluon, while the former leads to a more significant reduction of the PDF error. We have also shown evidence of some tension between some single-inclusive jet datasets and the rest of the global dataset, which might be explained by the less stable perturbative behaviour of this observable. Finally we have shown how, both for single-inclusive jets and dijets, the more recent 8 TeV data have a more significant impact than the previous 7 TeV data. We therefore expect that the future availability of more precise measurements from LHC Run-II at 13 TeV data will improve further our knowledge of the gluon PDF.

5 Theoretical error in PDFs determination

In order to get an accurate estimate of the uncertainties affecting the Standard Model predictions, theoretical errors need to be taken into account. For hadron collider processes these are dominated by those due to missing higher order corrections in QCD calculations and to PDFs[1]. It is clear how MHOU will also have an affect on the PDFs themselves, being present in the perturbative predictions of the particular processes used for PDFs determination. Besides from contributing to the overall size of PDFs uncertainty, MHOU might affect the relative weights different points have in a fit: points accurately described by the current perturbative predictions (up to NNLO) should weight more than those poorly described. As discussed in chapter 3, present PDFs uncertainties only account for statistical and systematic errors affecting the experimental data entering the analysis, and typically do not include any source of theory uncertainty. In this chapter, based on refs. [8, 9], we describe how to set up a general formalism for the inclusion of theoretical uncertainties in PDFs determinations, and we specify it to the case of MHOU.

The chapter is structure as follows. In sec. 5.1 we show how a generic source of theory error can be described by means of a covariance matrix; in sec. 5.2 we discuss how, when considering MHO, such covariance matrix can be constructed using scale variations and a suitable prescription, which is validated in sec. 5.3; in sec. 5.4 we present a first NLO PDFs set accounting for MHO, and finally in sec. 5.5 we provide instructions on how to use such result in phenomenological applications.

[1]From now on we will use the acronyms MHO and MHOU to denote missing higher orders and missing higher orders uncertainties respectively.

5.1 Theory error as a covariance matrix

In this section we will show how, by adopting a Bayesian approach and assuming a Gaussian prior probability distribution for the true value of the theory, any missing theoretical uncertainty can be accounted for by adding a contribution to the experimental covariance matrix used in the PDFs fit.

Denoting as D the vector of experimental data entering the analysis and as $\mathcal{T}$ the corresponding vector of "true" unknown values - whose determination is the goal of the experiment - we assume that the experimental results are Gaussianly distributed about this hypothetical true values $\mathcal{T}$

$$P\left(D|\mathcal{T}\right) \propto \exp\left(-\frac{1}{2}\left(D-\mathcal{T}\right)^{T} C^{-1}\left(D-\mathcal{T}\right)\right) . \tag{5.1}$$

The true values $\mathcal{T}$ are unknown, however we can compute the theory predictions T for each experimental data using a theory framework which is generally incomplete, for example because it is based on the fixed-order truncation of a perturbative expansion [2]. Furthermore T depend on PDFs, which are evolved up to the physical scales of the data using again an incomplete theory. The vectors $\mathcal{T}$ and T would coincide if the theory were exact and the PDFs were known with certainty. Writing the difference between the true and the actual value of the theory predictions as

$$\Delta = \mathcal{T} - T\,, \tag{5.2}$$

we can consider this difference as an additional unknown systematic error, accounting for the incomplete theory. If we assume, in the same spirit as when estimating experimental systematic, that the true values $\mathcal{T}$ are Gaussianly distributed about the theory predictions T

$$P\left(\mathcal{T}|T\right) \propto \exp\left(-\frac{1}{2}\left(\mathcal{T}-T\right)^{T} S^{-1}\left(\mathcal{T}-T\right)\right) , \tag{5.3}$$

then the prior probability distribution of Δ will be given by

$$P\left(\Delta\right) \propto \exp\left(-\frac{1}{2}\Delta^{T} S^{-1}\Delta\right) . \tag{5.4}$$

[2] In addition to MHO other effects which could be neglected in the theoretical predictions T are higher twist and nuclear effects.

Eq. (5.1) can be rewritten as

$$P\left(D|\mathcal{T}\right) = P\left(D,\Delta|T\right) \propto \exp\left(-\frac{1}{2}\left(D-T-\Delta\right)^T C\left(D-T-\Delta\right)\right), \qquad (5.5)$$

so that the conditional probability of the data D given the theory predictions T can be obtained using the Bayes theorem and marginalizing over Δ

$$\begin{aligned} P\left(D|T\right) &\propto \int d\Delta\, P\left(\Delta\right) P\left(D,\Delta|T\right) \\ &= \int d\Delta\, \exp\left(-\frac{1}{2}\left(D-T-\Delta\right)^T C^{-1}\left(D-T-\Delta\right) - \frac{1}{2}\Delta^T S^{-1}\Delta\right) \\ &\propto \exp\left(-\frac{1}{2}\left(D-T\right)^T\left(C+S\right)^{-1}\left(D-T\right)\right), \end{aligned} \qquad (5.6)$$

where in the last line we have performed explicitly the Gaussian integral over Δ. Eq. (5.6) defines the likelihood which is usually minimized in a Gaussian fit and shows how theoretical uncertainties can be treated simply as another form of experimental systematic: it is an additional uncertainty to be taken into account when trying to find the truth from the data using a specific theory setting, and it can be accounted for by mean of an additional contribution S to the experimental covariance matrix C. It should be noted that eq. (5.6) has been obtained under the assumption of a Gaussian prior for MHOU, as given in eq. (5.3). In general one could use different models, depending on the nature of the theory error considered, and minimize the resulting likelihood as given in the first line of eq. (5.6). Here we will work using the Gaussian assumption, which will be validated in sec. 5.3.

The problem is then to estimate the theory covariance matrix S. The Gaussian hypothesis eq. (5.3) implies that

$$\int d\mathcal{T}\, P\left(\mathcal{T}|T\right)\left(\mathcal{T}-T\right)_i\left(\mathcal{T}-T\right)_j = \left\langle \Delta_i\, \Delta_j\right\rangle = S_{ij}\,, \qquad (5.7)$$

showing how in general we need to estimate the shifts Δ_i defined in eq. (5.2), in a way that takes into account the theoretical correlations between different points within the same dataset, between different datasets measuring the same physical process and between datasets corresponding to different processes [3].

[3]Unlike experimental correlations, theory correlations will be present even for entirely different processes, through the universal parton distributions, which all share the same theory for DGLAP evolution.

5.2 MHOU from scale variations

The most commonly used method to estimate the theory corrections due to MHO is scale variations. In the following we briefly revise its key ingredients and fix the conventions and terminology used in this work. For simplicity we will discuss the case of electroproduction processes, like DIS, but the same argument can be used to obtain expressions for a generic hadronic process. We refer to ref. [9] for a complete and formal discussion of the topic.

Considering the problem of PDFs determination and remembering the factorized expression for high-energy processes cross-sections, there are two independent source of MHOU: the perturbative expression of the partonic cross-section and the perturbative expression of the anomalous dimensions that determine the evolution of parton distributions. These will be associated with two independent unphysical scales, which here will be denoted as renormalization scale μ_r and factorization scale μ_f. Using RG equations for hard cross-sections and for PDFs it is possible to obtain an estimate of the MHOU by varying independently the two unphysical scales entering the problem.

Considering a generic structure function, denoting as $\overline{F}$ the corresponding scale-dependent theory prediction[4] we have

$$\overline{F}\left(Q^2, \mu_r^2, \mu_f^2\right) = \overline{C}\left(\alpha_s\left(\mu_r^2\right), \frac{\mu_r^2}{Q^2}\right) \otimes q\left(\alpha_s\left(\mu_f^2\right), \frac{\mu_f^2}{Q^2}\right) . \tag{5.8}$$

Following the notations of ref. [9], we introduce the variables $t = \log Q^2/\Lambda^2$, $k_r = \log \mu_r^2/Q^2$ and $k_f = \log \mu_f^2/Q^2$ so that eq. (5.8) can be written as

$$\overline{F}\left(k_r, k_f\right) = \overline{C}\left(\alpha_s\left(t + k_r\right), k_r\right) \otimes \overline{q}\left(\alpha_s\left(t + k_f\right), k_f\right) . \tag{5.9}$$

In the following, we will use the notations $\overline{F}$, $\overline{C}$ and $\overline{q}$ to denote structure functions, Wilson coefficients and PDFs evaluated at the generic scale μ_r and μ_f. When setting such scales equal to the physical one Q, namely when $k_r = k_f = 0$

[4]The structure function $\overline{F}$ depend on μ_r^2 and μ_f^2 in the sense of the RG equation: the dependence on unphysical scales cancels order by order, and the residual dependence can be used to estimate the MHOU.

we define

$$\begin{aligned} F(0,0) &\equiv \overline{F}(0,0)\ , \\ C(t) &\equiv \overline{C}(\alpha_s(t),0)\ , \\ q(t) &\equiv \overline{q}(\alpha_s(t),0)\ . \end{aligned} \tag{5.10}$$

In order to estimate the MHOU due to the truncation of the perturbative expansion of the coefficient function $\overline{C}$ we can fix a specific renormalization scheme and keep $\mu_f = Q$, but varying the renormalization scale μ_r^2 used in the computation of the coefficient function itself. The scale-dependent structure function $\overline{F}$ will then be given by

$$\overline{F}\left(Q^2,\mu_r^2\right) = \overline{C}\left(\alpha_s\left(\mu_r^2\right),\frac{\mu_r^2}{Q^2}\right)\otimes q\left(Q^2\right) = \overline{C}\left(\alpha_s(t+k_r),k_r\right)\otimes q(t)\ . \tag{5.11}$$

Using the RG invariance of the physical cross section

$$\mu_r^2\frac{d}{d\mu_r^2}\overline{F}\left(Q^2,\mu_r^2\right) = \frac{d}{dk_r}\overline{F}(t,k_r) = 0\,, \tag{5.12}$$

it is easy to show that the renormalization scale dependent Wilson coefficients $\overline{C}$ can be written as

$$\overline{C}\left(\alpha_s(t+k_r),k_r\right) = C(t+k_r) - k\frac{d}{dt}C(t+k_r) + \frac{1}{2}k_r^2\frac{d^2}{dt^2}C(t+k_r) + ... \tag{5.13}$$

where according to eq. (5.10) $C(t) = \overline{C}(\alpha_s(t),0)$. In other words, thanks to the RG invariance we can write the renormalization scale dependent Wilson coefficients at a generic scale μ_r in terms of their values at the physical scale $\mu_r = Q$. The log derivatives appearing in eq. (5.13) can be easily evaluated using the perturbative expression of C

$$C(t) = c_0 + \alpha_s(t)\,c_1 + \alpha_s^2(t)\,c_2 + \alpha_s^3(t)\,c_3 + ...\,, \tag{5.14}$$

and of the β function expansion eq. (1.13) getting

$$\begin{aligned} \overline{C}\left(\alpha_s(t+k_r),k_r\right) &= c_0 + \alpha_s(t+k_r)\,c_1 + \alpha_s^2(t+k_r)\left(c_2 + k_r\beta_0 c_1\right) \\ &+ \alpha_s^3(t+k_r)\left(c_3 + k_r\beta_0\left(\beta_1 c_1 + 2c_2 - k_r\beta_0 c_1\right)\right) + ... \end{aligned} \tag{5.15}$$

In the same way, starting again from eq. (5.9), in order to get the scaled varied PDF we can fix $\mu_r = Q$ and vary the scale μ_f at which the PDFs are evaluated.

Setting $\mu_r = Q$ we get

$$\begin{aligned}\overline{F}\left(Q^2, \mu_f^2\right) &= C\left(\alpha_s\left(Q^2\right)\right) \otimes \overline{q}\left(\alpha_s\left(\mu_f^2\right), \frac{\mu_f^2}{Q^2}\right) \\ &= C\left(t\right) \otimes \overline{q}\left(\alpha_s\left(t + k_f\right), k_f\right) \,, \end{aligned} \tag{5.16}$$

and using the RG invariance eq. (5.12) with respect to μ_f we get

$$\overline{q}\left(\alpha_s\left(t + k_f\right), k_f\right) = q\left(t + k_f\right) - k_f \frac{d}{dt} q\left(t + k_f\right) + \frac{1}{2} k_f^2 \frac{d^2}{dt^2} q\left(t + k_f\right) + ... \tag{5.17}$$

where in analogy with what done for the Wilson coefficients we have defined $q\left(t\right) \equiv \overline{q}\left(\alpha_s\left(t\right), 0\right)$. Using the evolution equation[5]

$$\frac{d}{d\mu_f^2} q\left(\mu_f^2\right) = \gamma\left(\alpha_s\left(\mu_f^2\right)\right) q\left(\mu_f^2\right) \,, \tag{5.18}$$

eq. (5.17) can be rewritten as

$$\begin{aligned}\overline{q}\left(\alpha_s\left(t + k_f\right), k_f\right) =& q\left(t + k_f\right) - k\gamma q\left(t + k_f\right) \\ &+ \frac{1}{2} k^2 \left(\gamma^2 + \frac{d}{dt}\gamma\right) q\left(t + k_f\right) + ... \,, \end{aligned} \tag{5.19}$$

which can be further simplified using the perturbative expansion of the anomalous dimension and the expression for the β function[6].

Eqs. (5.15),(5.19) allow to easily perform scale variations for a single process, varying independently the two unphysical scales μ_r and μ_f. In particular, they allow to obtain scaled varied coefficient functions and PDFs in terms of the corresponding values computed at the physical scale Q. Considering a specific process involved in the PDFs fit, one can, for example, perform scale variation in the range $|k_r|, |k_f| < \log 4$, obtaining the scale varied cross section $\overline{F}\left(k_r, k_f\right)$. When considering hadronic processes, the same arguments presented above can be used to obtain a formula similar to the one of eq. (5.15), as shown in ref. [9].

We now consider a situation where we have p different types of processes (like for

[5]for simplicity, in this section all the argument is presented implicitly assuming a Mellin space formalism, so that convolutions are replaced by ordinary products.

[6]In ref. [9] it is explicitly shown that an alternative way of obtaining eq. (5.19) consists in varying the renormalization scale of the anomalous dimension. MHOU due to PDFs evolution can therefore be estimated varying either the PDFs scale or the scale of the anomalous dimension.

example electroproduction processes, hadronic processes, jets ...)

$$\pi_a = \{i_a\}\,, \qquad a = 1, ..., p\,,$$

where i_a labels the datapoints belonging to the a-th process. Each of them is characterized by a factorization scale μ_f (associated to the universal PDFs) and a renormalization scale μ_{r_a} (associated with the hard coefficient functions). Given the i-th point of the a-th process F_{i_a}, we define the corresponding shift Δ_{i_a} as

$$\Delta_{i_a}\left(k_f, k_{r_a}\right) \equiv \overline{F}_{i_a}\left(k_f, k_{r_a}\right) - F_{i_a}\left(0, 0\right)\,, \tag{5.20}$$

where we assume that all scale variations can be performed in the same range $|k_{r_a}|, |k_f| < \log 4$. In practice, for each scale three points can be sampled, corresponding to $k = 0, \pm \log 4$. Note that since the PDFs are universal but the coefficient functions are process dependent, when considering two different processes the scale variations of k_r will be totally independent while those of k_f will be correlated between different processes. In other words, because of PDFs universality the relation between the physical scale of each process (whatever that is) and the factorization scale μ_f is the same for all the processes.

According to eq. (5.7), the theory covariance matrix is then constructed by averaging outer products of the shifts over points in the space of scales

$$S_{i_a j_b} = N \sum_V \Delta_{i_a}\left(k_f, k_{r_a}\right) \Delta_{j_b}\left(k_f, k_{r_b}\right)\,, \tag{5.21}$$

where $i_a \in \pi_a$ and $j_b \in \pi_b$ indicate two data points possibly corresponding to different processes π_a and π_b, V is the set of scale points to be summed over and N is a normalization factor. Note that from this definition it follows immediately that the theory covariance matrix is positive definite: considering a real vector v_i, from eq. (5.21) we have

$$\sum_{ij} v_i S_{ij} v_j = N \sum_V \left(\sum_i v_i \Delta_i \right)^2 \geq 0\,. \tag{5.22}$$

Different prescriptions for the theory covariance matrix definition can be adopted, characterized by a different set of combination of scales which are summed over in eq. (5.21). Here we will discuss results for the so called 9-points prescriptions. In app. C we describe another possible option and we refer to the

original publication ref. [7] for more details about alternative prescriptions. For simplicity, let's first consider the theory covariance matrix entries corresponding to a couple of points belonging to the same process. In this case there are at most two independent scales to be varied, corresponding to the renormalization and factorization scales k_r and k_f. In the 9-points prescription k_r and k_f are varied completely independently, getting the $8 + 1$ points in the scales space reported in fig. 5.1 (left), where the $+1$ refers to the trivial point $k_r = k_f = 0$ for which the shift Δ_i vanishes. The normalization factor appearing in eq. (5.21) is determined by averaging over the number of points associated with the variation of each scale, and adding the contributions from variation of independent scales. So in the case of the 9-points prescriptions we have 8 points and 2 independent scales, giving a normalization factor $N = 1/4$. The corresponding theory covariance matrix entries read

$$\begin{aligned} S_{ij}^{(9\text{pt})} = \frac{1}{4}\{&\Delta_i^{+0}\Delta_j^{+0} + \Delta_i^{-0}\Delta_j^{-0} + \Delta_i^{0+}\Delta_j^{0+} + \Delta_i^{0-}\Delta_j^{0-} \\ &+\Delta_i^{++}\Delta_j^{++} + \Delta_i^{+-}\Delta_j^{+-} + \Delta_i^{-+}\Delta_j^{-+} + \Delta_i^{--}\Delta_j^{--}\}\,. \end{aligned} \tag{5.23}$$

The superscripts $0, \pm$ denote the different variations of k_r and k_f defining the shift, corresponding to $0, \pm\log 4$. Such construction can be generalized to the case of couples of points belonging to two different processes π_1 and π_2. The set V now involves possible variations of three scales k_f, k_{r_1}, k_{r_2}, represented in the right plot of fig. 5.1. Again, varying such scales independently and accounting for the corrects normalization factors, eq. (5.23) can be generalized to the off-diagonal blocks of the theory covariance matrix, giving

$$\begin{aligned} S_{i_1 j_2}^{(9\text{pt})} = \frac{1}{24}\{&2\big(\Delta_{i_1}^{+0} + \Delta_{i_1}^{++} + \Delta_{i_1}^{+-}\big)\big(\Delta_{j_2}^{+0} + \Delta_{j_2}^{++} + \Delta_{j_2}^{+-}\big) \\ &+2\big(\Delta_{i_1}^{-0} + \Delta_{i_1}^{-+} + \Delta_{i_1}^{--}\big)\big(\Delta_{j_2}^{-0} + \Delta_{j_2}^{-+} + \Delta_{j_2}^{--}\big)\} \\ &+3\big(\Delta_{i_1}^{0+} + \Delta_{i_1}^{0-}\big)\big(\Delta_{j_2}^{0+} + \Delta_{j_2}^{0-}\big)\}. \end{aligned} \tag{5.24}$$

5.3 Construction and validation of a theory covariance matrix

In this section we determine the theory covariance matrix at NLO using eqs. (5.23), (5.24) and we validate it against the known NNLO results. As input datasets, we use the same NNPDF3.1 baseline given in tab. 3.1 with

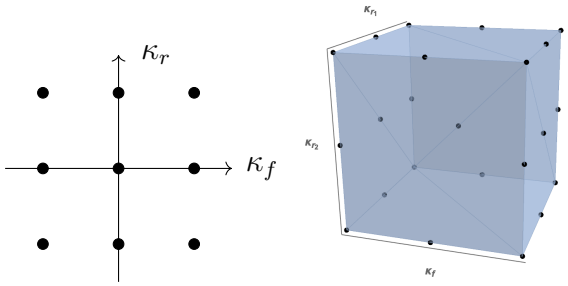

Figure 5.1 *9-points prescription for a single process (left) and for two different processes π_1 and π_2, indicating the sampled values for the factorization scale κ_f and renormalization scale κ_r. Figure from ref. [9].*

two minor differences: the value of the lower kinematic cut has been increased from $Q^2_{min} = 2.69$ GeV^2 to 13.96 GeV^2 in order to ensure the validity of the perturbative QCD expansion when scales are varied downwards, and the HERA F_2^b and fixed-target Drell-Yan cross-sections have been removed, for technical reasons related to difficulties in implementing scale variation. In total we then have $N_{dat} = 2819$ data points. As seen in the previous section, we assume that renormalization scale variation is fully correlated within a given process, but uncorrelated between different processes. Having defined the input experimental data it is then necessary to define what we mean by "process" and divide the input dataset accordingly. Our categorization, summarized in tab. 5.1, involves five distinct processes: charged-current (CC) and neutral-current (NC) deep-inelastic scattering (DIS), Drell–Yan (DY) production of gauge bosons (invariant mass, transverse momentum and rapidity distributions), single-jet inclusive and top pair production cross-sections. Note that such categorization requires and educated guess as to which theory computations share the same higher order corrections, and different choices might be done. We consider the one presented here to be sufficient for a first study. In order to evaluate the theory covariance matrix S_{ij}, it is necessary to be able to evaluate both DIS structure functions and hadronic cross-sections for a range of values of the factorization and renormalization scales, i.e. for $k_f \neq 0$ and $k_r \neq 0$. In this case, the entries of the NLO theory covariance matrix have been constructed by means of the `ReportEngine` software [130] taking the scale-varied NLO theory cross-sections $\overline{F}_i(k_f, k_r)$ as input. These are provided by `APFEL` [131] for the DIS structure functions and by `APFELgrid` [59] combined with `APPLgrid` [132] for the hadronic cross-sections.

In order to get an idea of the structure of the theory-induced correlations, in

Process Type	Datasets
DIS NC	NMC, SLAC, BCDMS, HERA NC
DIS CC	NuTeV, CHORUS, HERA CC
DY	CDF, D0, ATLAS, CMS, LHCb (y, p_T, M_{ll})
JET	ATLAS, CMS inclusive jets
TOP	ATLAS, CMS total+differential cross-sections

Table 5.1 *Classification of datasets into process types.*

fig. 5.2 we compare the experimental correlation matrix, given by

$$\rho_{ij}^{(C)} = \frac{C_{ij}}{\sqrt{C_{ii}}\sqrt{C_{jj}}}\,, \tag{5.25}$$

with the corresponding combined experimental and theoretical correlation matrix

$$\rho_{ij}^{(C+S)} = \frac{(C+S)_{ij}}{\sqrt{(C+S)_{ii}}\sqrt{(C+S)_{jj}}}\,. \tag{5.26}$$

By inspection of fig. 5.2 large positive correlations within individual experiments along the diagonal blocks are apparent, particularly evident for DIS NC and DY data. Within the same process there are large correlations between experiments for the DY, jets and top datapoints and large anticorrelations for the DIS NC points. Correlations and anticorrelations between different processes, despite being present thanks to PDFs universality, are generally weaker.

Next, we wish to construct a validation test for the NLO theory covariance matrix, using the known shift between NNLO and NLO results. In order to do this, we view the set of experimental data as a vector D_i, where $i = 1, ..., N_{dat}$. Such vector lives in a vector space D of dimension N_{dat}, and the theory covariance matrix S_{ij} defines an ellipsoid E belonging to a subspace S of dimension N_{sub} of the full space D. In the context of MHOU we can take the NLO theory predictions evaluated at the central scales $T_i^{NLO}\,(0,0)$ as our best NLO predictions with the ellipsoid E estimating a 68% confidence level for the MHO corrections. We want to check how well the theory covariance matrix S_{ij} predicts both the size and the correlation pattern of the MHO terms. This can be done by testing the extend by which the known shift vector between NNLO and NLO theory predictions $T_i^{NNLO} - T_i^{NLO}$ falls within the ellipsoid E. More in detail, we first normalize the theory covariance matrix S_{ij} to the NLO predictions, so that all its entries are

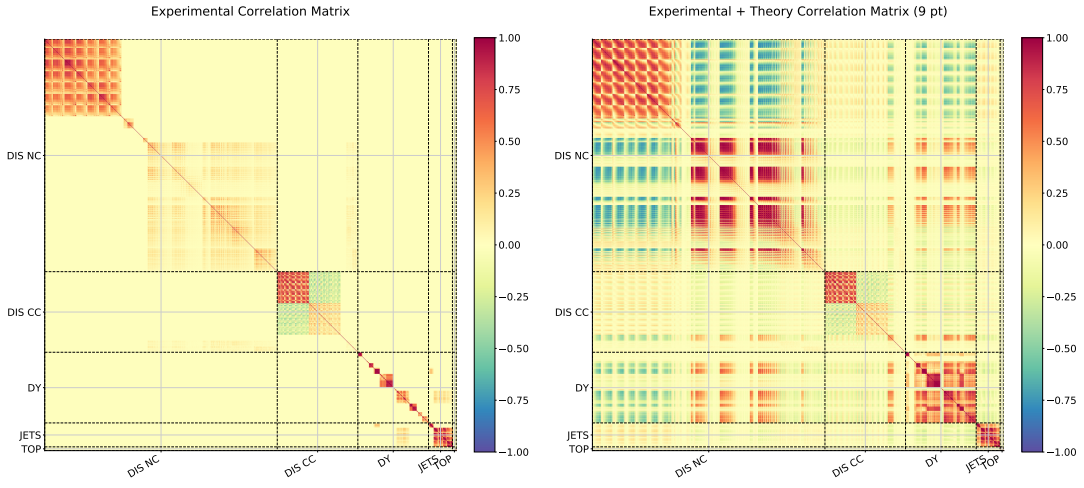

Figure 5.2 *Comparison of the experimental C_{ij} (left) and the combined experimental and theoretical correlation matrices S_{ij}, All entries are normalized to the central experimental value. The data are grouped by process and, within a process, by experiment. Figure from ref. [9].*

dimensionless allowing for a meaningful comparison

$$\hat{S}_{ij} = S_{ij} / \left(T_i^{NLO} T_j^{NLO} \right) , \tag{5.27}$$

and likewise we define a normalized shift vector as

$$\delta_i = \left(T_i^{NNLO} - T_i^{NLO} \right) / T_i^{NLO} . \tag{5.28}$$

The NNLO predictions used to define the shift δ_i are computed using the NNLO matrix elements and anomalous dimensions but the same NLO PDF set used to compute the NLO theory predictions. In this way the shift δ_i only accounts for the perturbative effects due to NNLO corrections, without including additional effects due to refitting. A first test to check whether the overall size of the scale variation is of the right order of magnitude consists into comparing the diagonal entries $\hat{S}_{ii} = \sigma_i^2$ to the normalized shift δ_i. This check is performed in fig. 5.3: in all cases δ_i turns out to be smaller or comparable to σ_i, showing how the overall size of the estimated uncertainties, obtained by varying the renormalization and factorization scales by a factor two in either directions, gives a qualitative reliable (if somewhat conservative) estimate of the true MHOU.

The validation of the full covariance matrix $\hat{S}_{ij}$ requires some more work. In order to identify the subspace S we diagonalize the matrix $\hat{S}_{ij}$, getting a set of

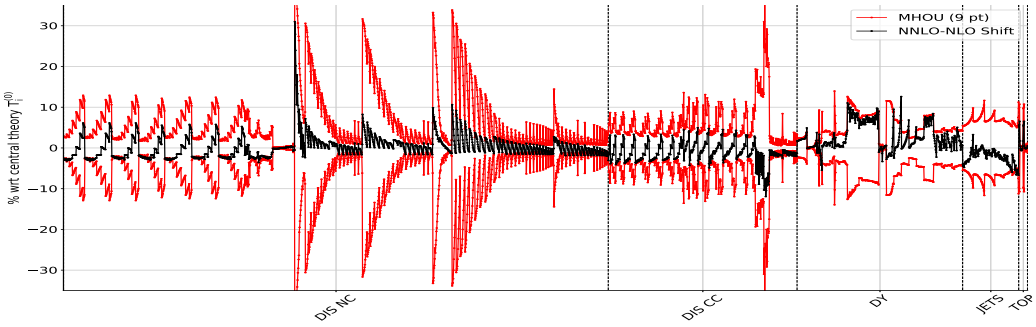

Figure 5.3 *The diagonal uncertainties σ_i (red) symmetrized about zero, compared to the shift δ_i (black) for each datapoint. Figure from ref. [9].*

orthonormal eigenvectors e_i^α and the corresponding non-zero eigenvalues λ_α with $\alpha = 1, ..., N_{sub}$. There is also a set of $N_{dat} - N_{sub}$ zero eigenvalues, corresponding to eigenvectors spanning the space D/S. In general the shift vector δ will live in the space D. For a successful test we expect most of δ to lie within S. In other words, denoting as δ^s the projection of the shift over the subspace S

$$\delta_i^s = \sum_{\alpha=1,...,N_{sub}} \delta^\alpha e_i^\alpha \,, \tag{5.29}$$

we expect the angle θ between δ and δ^s

$$\theta = \arccos\left(\frac{|\delta^s|}{|\delta|}\right) \tag{5.30}$$

to be reasonably small. This geometric relation is represented graphically in fig. 5.4, where the space D is drawn as a three dimensional space and the subspace S as a two dimensional space. For individual processes we find

$$\theta = 3^o, 14^o, 22^o, 32^o, 16^o$$

for top, jets, DY, NC and CC DIS respectively, while for the complete dataset we find $\theta = 26^o$. It is clear from these numbers how processes with larger numbers of data points, having a wider kinematic range and thus more structure to predict, are much harder to describe than those with only few data, which translates into bigger values of θ for bigger datasets. However in general the θ values we get for each specific process and for the global dataset result reasonably small, validating our definition and construction of a NLO theory covariance matrix.

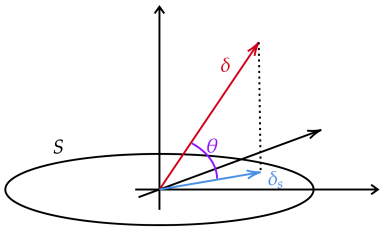

Figure 5.4 *Schematic representation of the geometric relation between the shift vector $\delta \in D$ (here drawn as a three dimensional space), and the component δ^S of the shift vector which lies in the subspace S (here drawn as a two dimensional space, containing the ellipse E defined by the theory covariance matrix). The angle θ between δ and δ^S is also shown. Figure from ref. [9].*

5.4 NLO PDFs with missing higher order uncertainties

In this section we present the first determination of the proton PDFs which systematically accounts for MHOU, using the theory covariance matrix formalism described in the previous sections. We will present only a NLO fit, leaving a full NNLO analysis for a future work. Note that a NLO PDFs fit offers a nontrivial validation of our methodology, by comparing the results with and without MHOU to a standard NNLO PDF set (obtained starting from the same input datasets).

As discussed in sec. 5.1, the theory uncertainties are included by replacing the experimental covariance matrix C_{ij} with the sum $C_{ij} + S_{ij}$. The NNPDF methodology described in chapter 3 is otherwise unchanged. It is then clear how the inclusion of a theory-induced contribution in the covariance matrix affects only two steps of the fit: the pseudodata generation and the definition of the χ^2 to be minimized. In particular, denoting as $D_i^{(k)}$ the k-th replicas for the i-th datapoint entering the analysis, we will now have

$$\lim_{N_{rep}\to\infty} \frac{1}{N_{rep}-1} \sum_{k=1}^{N_{rep}} \left(D_i^{(k)} - \langle D_i \rangle\right)\left(D_j^{(k)} - \langle D_j \rangle\right) = C_{ij} + S_{ij}\,, \tag{5.31}$$

with $\langle D_i \rangle$ denoting the average over the N_{rep} Monte Carlo pseudodata. Each

PDF replica is then fitted by minimizing

$$\chi^2 = \frac{1}{N_{dat}} \sum_{i,j=1}^{N_{dat}} (D_i - T_i) (C + S)^{-1}_{ij} (D_j - T_j) \ , \tag{5.32}$$

where the theory predictions T_i are computed using the central scales choice. In the following, in order to asses the fit quality and to study the impact of MHOUs on the final PDF uncertainties, we will provide values for the total and partial χ^2 and for the estimator ϕ, defined in ref. [55] as

$$\phi = \sqrt{\langle \chi^2_{\rm exp} \left[T^{(k)} \right] \rangle - \chi^2_{\rm exp} \left[\langle T^{(k)} \rangle \right]} \, , \tag{5.33}$$

where by $\chi^2_{\rm exp} \left[T^{(k)} \right]$ we denote the value of the χ^2 computed using the k-th PDF replica and only including the experimental covariance matrix. The average χ^2 values entering eq. (5.33) are the χ^2 averaged over the replicas and the χ^2 computed using the central PDF, which is obtained as an average of all replicas. As shown in app. A.2, eq. (5.33) can be written as

$$\phi = \left(\frac{1}{N_{dat}} \sum_{i,j=1}^{N_{dat}} (C)^{-1}_{ij} \, T_{ij} \right)^{1/2} \ , \tag{5.34}$$

where $T_{ij} = \langle T^{(k)}_i T^{(k)}_j \rangle - \langle T^{(k)}_i \rangle \langle T^{(k)}_j \rangle$. In words, ϕ gives the average over all the datapoints of the ratio of the uncertainties of the predictions to the uncertainties of the original experimental data, taking correlations into account. For a purely diagonal covariance matrix, this would be the ratio of the uncertainty of the predictions using the output PDFs to that of the original data. Note that ϕ is defined in such a way that the uncertainty in the prediction is always normalized to the experimental uncertainty, rather than to the combined experimental and theoretical uncertainty. By comparing ϕ values for fits with and without MHOU we can then get a quantitative idea of the effect of theory uncertainty on the final PDF error.

In order to asses the effect of MHOU, in addition to fits with the theory covariance matrix, two baseline NLO and NNLO fits based on the experimental covariance matrix C only have been produced, using the same input datasets described in sec. 5.3. As mentioned previously, including a new contribution to the covariance matrix of the fit will affect both the PDFs central value and uncertainty. In order to disentangle these two different effects, we also study PDFs determined by only partially including the theory covariance matrix S in the analysis, either only in

Label	Order	Cov. Mat.	Comments
NNPDF31_nlo_as_0118_kF_1_kR_1	NLO	C	baseline NLO
NNPDF31_nlo_as_0118_scalecov_9pt	NLO	$C+S$	
NNPDF31_nlo_as_0118_scalecov_9pt_fit	NLO	$C+S$	S only in χ^2 definition
NNPDF31_nlo_as_0118_scalecov_9pt_sampl	NLO	$C+S$	S only in sampling
NNPDF31_nnlo_as_0118_kF_1_kR_1	NNLO	C	baseline NNLO

Table 5.2 *Summary of the PDF sets discussed in this section. The perturbative order and nature of the treatment of uncertainties for each fit are indicated.*

the data generation or in the fitting. The PDF sets which will be discussed in the following are reported in table 5.2. For each fit we indicate its label, the perturbative order and the covariance matrix used.

The χ^2 and ϕ values are shown in tables 5.3 and 5.4 respectively, for both the total dataset and the individual processes of table 5.1. Considering the fit where the theory covariance matrix is included in both the χ^2 definition and in the Monte Carlo replicas generation, for all the processes the χ^2 decreases, improving by about 3% when considering the total dataset. Additionally the total χ^2 almost coincides with the NNLO χ^2, suggesting that indeed the theory uncertainty is correctly accounting for the missing NNLO corrections. Looking at the value of ϕ, we notice how, interestingly, this only increases by around 30%, much less than what one might expect looking at the relative size of the NLO MHOU and experimental uncertainties. These numbers suggest that the main effect of the inclusion of the theory covariance matrix is that, in the data region, tensions which are otherwise present in the global dataset due to the MHO are partially resolved, leading to a better fit quality without any major effect on the final PDFs error. Looking at the fits where the theory error is included in the χ^2 but not in the replicas generation, it is clear how the inclusion of the theory covariance matrix in the χ^2 definition only leads to a final χ^2 value very close to that of the fit where the MHOU are fully included. This means that, as we would expect, the MHOU affect mostly the central value of the fit, since the relative weight carried by each point is altered during the fit according to their relative size of their MHOU. On the other hand, considering the case where the theory covariance matrix is included only in the replica generation, the χ^2 goes up and ϕ increases dramatically, pointing out a much more prominent effect in PDFs

Process	$n_{\rm dat}$	$\chi^2/n_{\rm dat}$ in the NNPDF3.1 global fits				
		NLO				NNLO
		C	$C+S^{(9\rm pt)}$	$C+S^{(9\rm pt)}_{\rm fit}$	$C+S^{(9\rm pt)}_{\rm sampl}$	C
DIS NC	1593	1.088	1.079	1.081	1.227	1.084
DIS CC	552	1.012	0.928	0.929	1.036	1.079
DY	484	1.486	1.447	1.461	1.434	1.231
JETS	164	0.907	0.839	0.848	0.911	0.950
TOP	26	1.260	1.012	1.001	1.264	1.068
Total	2819	1.139	1.109	1.113	1.217	1.105

Table 5.3 *The values of the $\chi^2/N_{\rm dat}$ in the NNPDF3.1 global fits with the theory covariance matrix S, compared to the results based on including only the experimental covariance matrix C. We also show results obtained including the theory covariance matrix only in the χ^2 definition but not in the data generation and conversely. Values corresponding to the NNLO fit with experimental covariance matrix C only are also shown.*

uncertainty. This behaviour is expected: due to the inclusions of MHOU in the pseudodata generation, the replica fluctuations are wider, leading to an increase in the PDFs error. Since the theory error is not included in the χ^2, such increase in the PDFs error is now uncompensated by a rebalancing of the datasets in the fits.

In fig. 5.5 we show plots for the NLO PDFs of the gluon, the total quark singlet, the anti-down quark and the strange, comparing results for fits based on C and $C+S$. All the PDFs are plotted at $Q = 10$ GeV and normalized to the fit results without MHOU. The central value of the NNLO fit based on the experimental covariance matrix only is shown as well. We find that in the data region the increase in PDF uncertainties is very moderate, while the central values can be shifted by up to one sigma. On the other hand, in the regions where the PDFs are loosely constrained by the experimental data, the PDF uncertainties increases significantly. These features are in agreement with the observation that the estimator ϕ, whose values are reported in table 5.4, increases only by a moderate amount when including the theory error, and provide further evidence that in the data region the inclusion of the theory covariance matrix has resolved

Process	ϕ in the NNPDF3.1 global fits				
	NLO				NNLO
	C	$C + S^{(9\text{pt})}$	$C + S^{(9\text{pt})}_{\text{fit}}$	$C + S^{(9\text{pt})}_{\text{sampl}}$	C
DIS NC	0.266	0.412	0.414	1.137	0.305
DIS CC	0.389	0.408	0.388	0.502	0.471
DY	0.361	0.377	0.378	0.603	0.380
JETS	0.295	0.359	0.336	0.461	0.392
TOP	0.375	0.443	0.382	0.612	0.363
Total	0.314	0.405	0.400	0.932	0.362

Table 5.4 *Same as Table 5.3, but for the values of the ϕ estimator.*

tensions due to MHO. This point can be further supported by observing the improvement in the agreement between the best-NLO fit and the NNLO PDFs (the latter with experimental covariance matrix only): looking at the central value of the NNLO fit, first it is fully compatible with the uncertainty band of the NLO fit, second it is evident how upon inclusion of the NLO MHOU the central best fit moves towards the correct NNLO result. This is particularly evident in the case of the gluon and of the strangeness, where inclusion of MHOU leads to a suppression at large x of the first and to an enhancement in the whole x region of the second.

Finally in fig. 5.6 we look at PDFs where the theory covariance matrix has been included in the χ^2 definition but not in the Monte Carlo replicas generation and conversely. It is clear from the plots how, when S is included in the data replica generation only, uncertainties increased significantly. This is in agreement with the numbers observed in tables 5.3 and 5.4: the wider fluctuations in the data generation are not matched by the χ^2 definition, resulting in an overall bigger error and a worse fit quality. On the other hand, when S is included only in the χ^2 definition, the effect of theory error on the central value of the fit is singled out: the central value of the fit is very close to that obtained when including the MHOU in both data generation and fit, and, consistently with table 5.4, the change of the PDFs error in the data region is very small. This confirms our previous statement according to which, while the addition of a theory covariance matrix in replicas generation increases the fluctuations of the data replicas, this

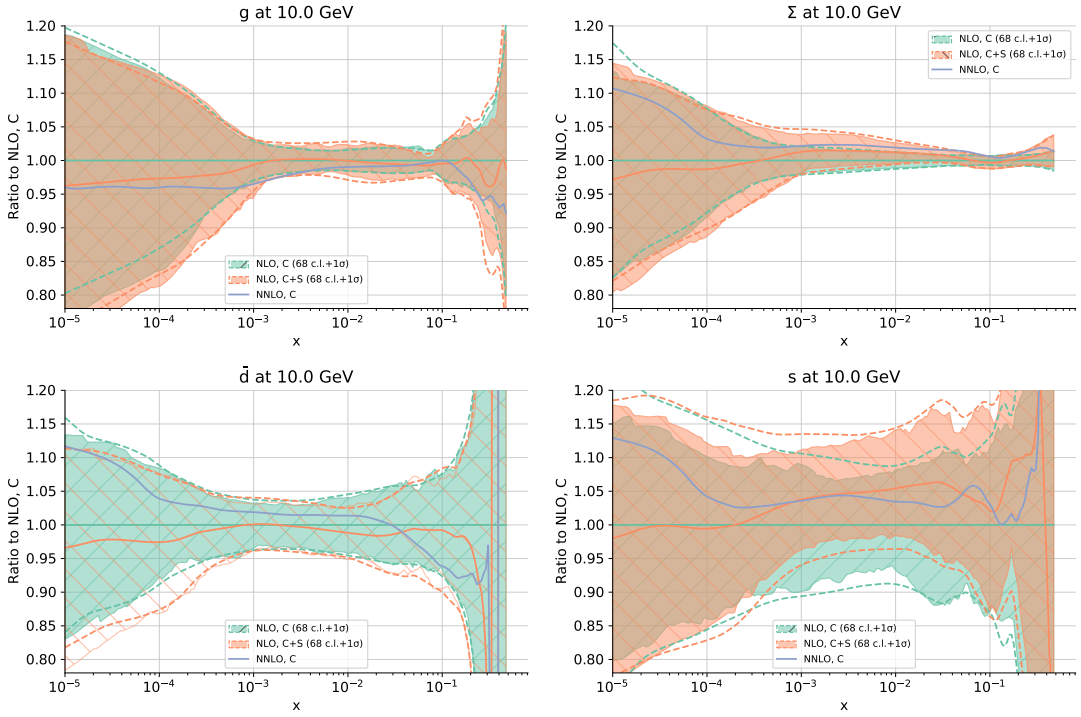

Figure 5.5 *Results of the NLO fits based on C and $C + S$ normalized to the former, as well as the central value of the NNLO fit based on C for the gluon, the total quark singlet, the anti-down quark, and the strange PDFs, all at $Q = 10$ GeV.*

is compensated by the inclusion of MHOU in the fit, which releases tensions between dataset, with the net results that, while the central values are shifted, the uncertainties in the data region do not increase much.

5.5 Usage and delivery

The PDFs with theory error can be used in the same way as a standard Monte Carlo PDF sets. In this section we briefly discuss how to combine the PDF theory error with that in the hard matrix elements, providing detailed instructions to use the results presented in sec. 5.4.

The MHOU discussed here arise from the fact that the PDFs are determined using finite order perturbative computations: parton distributions obtained by using different perturbative orders in the computations for the input processes will be different (so that for example NLO PDFs differ from NNLO PDFs), and the formalism developed here provides a way to take this into account when working

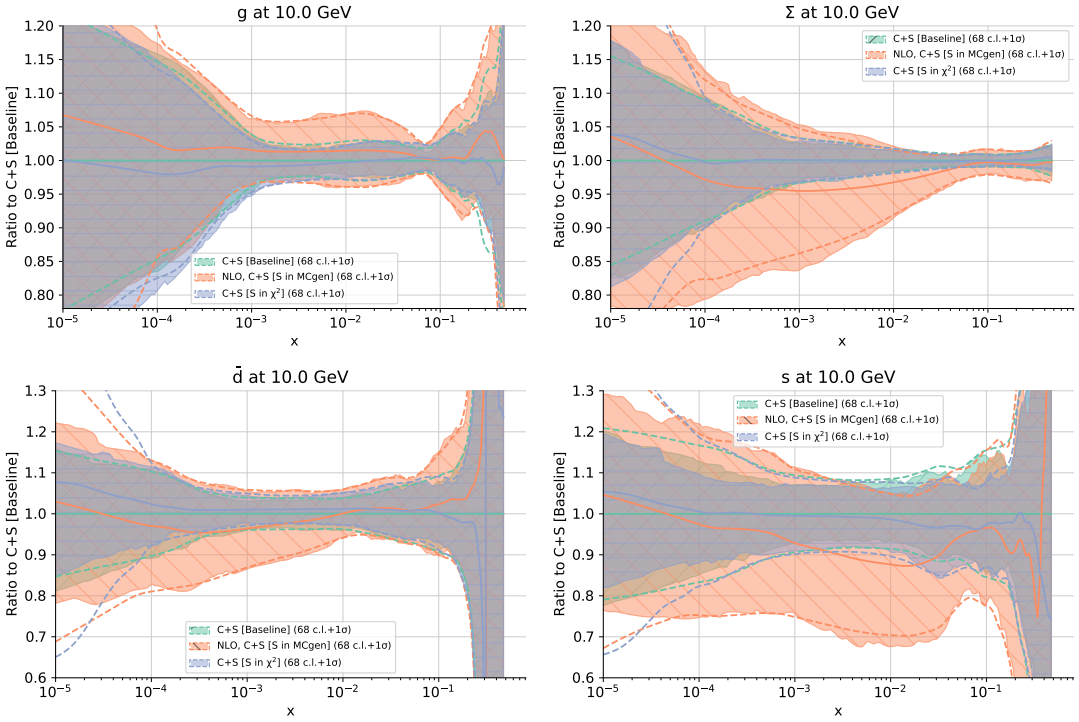

Figure 5.6 *Same as fig. 5.5, now comparing the results of the baseline $C+S$ fit with those in which the theory covariance matrix S is included either in the χ^2 definition or in the generation of Monte Carlo replicas, but not in both*

at some finite order in perturbation theory. We have further seen how there exist two distinct sources of MHOU in PDFs, the first due to the computation of the hard cross-sections for partonic processes, the second due to the computation of the anomalous dimension. Considering a factorized prediction for some other process not used during the PDFs determination, an additional source of MHOU is the computation of the hard process itself, which in turn will carry MHOU related to the computation of the hard cross-section and MHOU related to the evolution of the PDF from the initial scale (at which the PDFs are delivered) to the scale of the process. In summary, each theory prediction for a factorized cross section carry two different MHOU, a PDFs uncertainty, discussed in this work, and an uncertainty arising in the calculation of the prediction itself. We will ignore any possible correlation between these two different source of MHOU, considering the two theory error as completely independent.

In the following we summarize our procedure to compute the total uncertainty for a given factorized cross-section $\mathcal{F}$. The PDF uncertainty $\sigma_{\mathcal{F}}^{\text{PDF}}$ can be evaluated in the same way as usually done for a standard Monte Carlo PDFs set, as the

standard deviation over the replicas set

$$\sigma_{\mathcal{F}}^{\mathrm{PDF}} = \left(\frac{1}{N_{\mathrm{rep}} - 1} \sum_{k=1}^{N_{\mathrm{rep}}} \left(\mathcal{F}\left[\{q^{(k)}\}\right] - \langle \mathcal{F}\left[\{q\}\right] \rangle \right)^2 \right)^{\frac{1}{2}} . \tag{5.35}$$

Considering the PDFs presented in this work, $\sigma_{\mathcal{F}}^{\mathrm{PDF}}$ will account for uncertainties related to both experimental data and MHO in the PDFs determination. On the other hand, when eq. (5.35) is used with a standard PDFs set (like NNPDF3.1), the resulting uncertainty only includes statistical and systematic errors from the data. Turning to MHOU on the hard matrix element $\sigma_{\mathcal{F}}^{\mathrm{th}}$, this can be estimated using any prescription preferred by the user. A commonly used procedure is given by the 7-points scale variation presented in ref. [133]. Alternatively, one can use the theory covariance matrix used for the PDFs determination: the uncertainty on the cross-section $\mathcal{F}$ will be the corresponding diagonal entry of the covariance matrix

$$\sigma_{\mathcal{F}}^{\mathrm{th}} = \left[S_{\mathcal{F}\mathcal{F}} \right]^{\frac{1}{2}} , \tag{5.36}$$

with the shift Δ_{ij} computed for $i = j = \mathcal{F}$. The PDF uncertainty eq. (5.35) and the uncertainty on the hard matrix element eq. (5.36) can now be combined in quadrature, giving the total uncertainty $\sigma_{\mathcal{F}}^{\mathrm{tot}}$ for the cross section $\mathcal{F}$

$$\sigma_{\mathcal{F}}^{\mathrm{tot}} = \left(\left(\sigma_{\mathcal{F}}^{\mathrm{PDF}} \right)^2 + \left(\sigma_{\mathcal{F}}^{\mathrm{th}} \right)^2 \right)^{\frac{1}{2}} . \tag{5.37}$$

6

Fitting the b-quark PDF as a massive-b scheme

It has been recently shown [48] that for accurate phenomenology at the LHC it is advantageous to treat the charm parton distribution (PDF) on the same footing as light-quark PDFs, namely, to parametrize it and extract it from data, rather than to take it as radiatively generated from the gluon using perturbative matching conditions. This is likely to be due to the fact that the matching conditions eq. (2.69) are only known to the lowest nontrivial order, which may well be subject to large higher order corrections, as revealed by the strong dependence of results on the choice of matching scale. On top of this, of course, the starting low-scale heavy quark PDFs could in principle also have a non-perturbative "intrinsic" component [134, 135]. It is important to note that whether or not the heavy quark PDF has a nonperturbative component, and whether it is advantageous to parametrize the heavy quark PDF are separate issues: in fact in ref. [48] it was shown that the main phenomenological advantage in parametrizing and fitting the charm PDF comes from a region in which any nonperturbative contribution to charm is likely to be extremely small.

The case of the bottom quark PDF is, in this respect, particularly interesting. On the one hand, one may think that the problem of large higher order corrections to the matching conditions is alleviated in this case by the larger value of the mass. However, on the other hand, there is a more subtle consideration. Namely, there are b-initiated hadron collider processes — some of which are especially relevant for new physics searches — such as Higgs production in bottom fusion, for which b quark mass effects might be non-negligible [136–138]. This suggests the use of a scheme in which the b quark is treated as a massive final-state parton — hence not endowed with a PDF. In such a scheme the b-induced process necessarily starts at a higher perturbative order than in a scheme in which there exists a b

PDF, because the b production process is included in the hard matrix element. As a consequence, the computation of the b-induced process itself is more difficult and it can typically only be performed with a lower perturbative accuracy than in a scheme in which the b quark is a massless parton.

The problem is somewhat alleviated if the massive-scheme and massless-scheme computations are combined, with the b-PDF in the massless scheme assumed to be produced by perturbative matching conditions. We henceforth refer to such a computational framework as "matched-b". However, in a matched-b framework the massive computation is still beset by the need to start at high perturbative order. As a possible way out, the use of a "massive five flavor scheme" has been suggested recently [139, 140], in which there is a b PDF (hence five flavors), yet b quark mass effects are included (possibly, at least in part, also in parton showering). The use of an independently parametrized b quark PDF within a framework in which massive and massless computations are combined offers a simpler way of dealing with the same problem. We refer to this as a "parametrized b" computational framework. Such an approach has been developed for electroproduction in refs. [141, 142], and it has been used in order to produce PDF sets with parametrized charm [48, 143], including the recent NNPDF3.1 set. Because the only data currently used for PDF determination in which heavy quark mass effects have a significant impact are deep-inelastic scattering data close to the charm production threshold, in these references only electroproduction was considered and only the parametrization of the charm was studied.

In these previous studies, an independently parametrized heavy quark PDF is included in the FONLL method [43], which, as discussed in sec. 2.4, allows for the matching of a scheme in which the heavy quark mass is included but the heavy quark decouples from QCD evolution equations, and a massless scheme in which the heavy quark mass is neglected, but the heavy quark PDF couples to perturbative evolution. In this parametrized heavy quark version of the FONLL scheme, the heavy quark PDF is present both in the massive and massless scheme, though decoupled from evolution in the massive scheme; unlike in conventional matched heavy quark computations in which the number of PDFs is different, with one more PDF in the massless scheme. The rationale for FONLL with a parametrized heavy quark is to simultaneously include heavy quark mass effects at lower scales and the resummation of collinear mass logarithms in the heavy quark PDFs at higher scales. This has the important byproduct that one ends

up with a computational framework in which there are heavy quarks in the initial state even in the scheme in which mass effects are retained.

Therefore, in a parametrized-b FONLL framework, problems related to the fact that the relevant processes in a massive scheme start at higher order is thus completely evaded, since the heavy quark PDF is always present. Mass effects are then included to finite perturbative order, along with the resummation of mass logarithms, though (unlike in some "massive five-flavor scheme") mass corrections to resummed perturbative evolution are not included. On the other hand, any possible nonperturbative corrections to the b PDF, including, say, the effective value of the b mass at which the matching should happen, are then included in the PDF itself and thus extracted from the data.

In this chapter, following ref. [10], we explicitly construct the parametrized-b FONLL method, by generalizing to hadronic processes the construction of refs. [141, 142] of FONLL with parametrized heavy quark PDF. We specifically consider the application to Higgs production in bottom fusion. This process has been computed at the matched level both using the FONLL method [45, 144] and EFT-based methods [145, 146], with the respective results benchmarked in ref. [133] and found to be in good agreement with each other. All these computations were performed in a matched-b approach, in which the b PDF is absent in the massive (four-flavor) scheme, and determined by matching condition in the massless (five-flavor) scheme. Here we take this process as a prototype for the use of a parametrized-b scheme for hadron-collider processes.

First, we discuss how the counting of perturbative orders changes in the presence of a parametrized-b PDF, and redefine suitable matched schemes based on this new counting. We then work out the generalization to hadronic processes of FONLL with parametrized heavy quark PDF of refs. [141, 142], we discuss in which sense it effectively provides an alternative to the massive five-flavor scheme, and then we work out explicit expressions for Higgs production in bottom fusion to the matched next-to-leading order - next-to leading log (NLO-NLL) level and NLO-NNLL level. We finally compare predictions obtained within this approach with some plausible choices of the b-quark PDF to those obtained in the approach of refs. [45, 144], and argue that results with similar or better phenomenological accuracy can be obtained in a much simpler way within this new approach.

6.1 The FONLL scheme with parametrized heavy quark PDF in hadronic collisions

Even though we have the general goal of constructing a parametrized-b FONLL scheme for hadronic processes, we always specifically refer to Higgs production in gluon fusion, in order to have a concrete reference case, and test scenario. We recall that, as discussed in sec. 2.4, the FONLL method matches two calculations of the same process performed in two different renormalization schemes: a massive scheme in which the heavy quark mass is retained, but the heavy quark decouples from the running of α_s and from QCD evolution equations, and a massless scheme in which the heavy quark contributes to the running of α_s and QCD evolution equations, but the heavy quark mass is neglected. In the computation of a hard process at scale Q^2, in the former scheme mass effects $O\left(\frac{m_b^2}{Q^2}\right)$ are retained, but mass logarithms $\ln\frac{Q^2}{m_b^2}$ are only kept to finite order in α_s (where m_b denotes generically the mass of the heavy quark). In the latter scheme, mass effects are neglected but mass logarithms are resummed to all orders in α_s. Hence by matching the two calculations one retains accuracy both at low scales where quark mass effects are important, and at high scales where mass logarithms are large.

The general idea of the FONLL method is to realize that these are just two different renormalization schemes: the massive scheme is a decoupling scheme, and the massless scheme is a minimal subtraction scheme. So the two calculations can be simply matched by re-expressing both in the same renormalization scheme using eqs. (2.68), (2.69) and then subtracting common contributions. In practice, this is done by expressing the massive scheme computation in terms of the PDFs and α_s of the massless scheme, and then adding to it the difference $\sigma^{\rm d}$ between the massless calculation and the massless limit of the massive one. Schematically

$$\sigma^{\rm FONLL} = \sigma^{\rm massive} + \sigma^{\rm d} \tag{6.1}$$

$$\sigma^{\rm d} = \sigma^{\rm massless} - \sigma^{\rm massive,\,0}. \tag{6.2}$$

This corresponds to replacing all the terms in the massless computation which are included to finite order in α_s in the massive computation with their massive counterpart.

In the simplest (original) version of FONLL, as discussed in ref. [43] for b production in hadronic collisions, and in ref. [44] for deep-inelastic scattering,

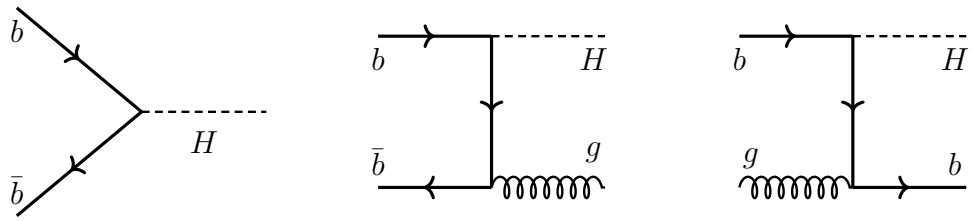

Figure 6.1 *Feynman diagrams for the leading (left) and next-to-leading order real emission contributions to Higgs production in bottom fusion which are present in the massive scheme when the b quark PDF is independently parametrized, but absent otherwise.*

in the massive scheme there is no heavy quark PDF, and the heavy quark can only appear as a final-state particle. In the massless scheme the heavy quark PDF is determined by matching conditions in terms of the light quarks and the gluon. These conditions match the massless scheme at a scale μ such that the heavy quark PDF only appears for scales above μ. Specifically, at order $O(\alpha_s)$, the heavy quark PDF just vanishes at the scale $\mu = m_b$ and it is generated by perturbative evolution at higher scales, while at $O(\alpha_s^2)$ it has a nontrivial gluon-induced matching condition at all scales.

When introducing a parametrized PDF both the massive and massless scheme computations change. The massless scheme changes, somewhat trivially, in that the heavy quark PDF, at the matching scale, instead of being given by a matching condition, is freely parametrized. The massive scheme changes nontrivially in that there is now a heavy quark PDF also in this scheme, only it does not evolve with the scale. The consequences of this were worked out in refs. [141, 142] in the case of electroproduction, and we study them here for hadroproduction for the first time.

6.1.1 Perturbative ordering

Because there is now a b PDF also in the massive scheme, the counting of perturbative orders in this scheme changes substantially. Specifically, for Higgs production in bottom fusion the diagrams of fig. 6.1 are present only when the b PDF is independently parametrized. This means that while in the massive scheme the process in the matched-b approach of refs. [45, 144] starts at $O(\alpha_s^2)$, in a parametrized-b approach it starts at $O(\alpha_s^0)$. As discussed in detail in refs. [44, 45, 144], the FONLL method allows the consistent combination of computations performed at different perturbative orders either in

the massive or massless scheme: various combinations were defined and discussed in refs. [45, 144] for Higgs production in bottom fusion.

With the new counting of perturbative orders which is relevant for a parametrized-b framework it is convenient to define some new combinations. We consider in particular the combination of the massive scheme $O(\alpha_s)$ computation with the standard five-flavor scheme next-to-leading log (NLL) and next-to-next-to-leading log computations. We call these combinations respectively FONLL-AP (hence corresponding to NLO-NLL) and FONLL-BP (corresponding to NLO-NNLL).

6.1.2 Parametrized-b FONLL

The construction of the parametrized-b FONLL for hadronic processes closely follows the corresponding construction for electroproduction, presented in refs. [141, 142], to which we refer for more details. In comparison to the matched-b FONLL of refs. [45, 144] the massive scheme contribution to eq. (6.1) includes an extra contribution:

$$\begin{aligned} \sigma_{\text{FONLL}_P} &= \sigma_{\text{FONLL}_M} + \delta\sigma_P \\ \delta\sigma_P &= \sigma_P^{\text{massive}} - \sigma_P^{\text{massive},0}, \end{aligned} \tag{6.3}$$

where $\sigma_P^{\text{massive}}$ is the massive-scheme contribution to the given process with initial-state heavy quarks and $\sigma_P^{\text{massive},0}$ its massless limit (which subtract its double counting with the massless-scheme contribution). This massive scheme contribution has to be re-expressed in terms of massless-scheme PDFs, as explained in detail in refs. [43–45, 141, 142, 144].

For Higgs production in bottom fusion, up to NLO, this extra contribution is given by the real diagrams of fig. 6.1, supplemented by the corresponding virtual correction and thus it has the form

$$\begin{aligned} \delta\sigma_P^{\text{massive}}\left(\frac{m_H^2}{m_b^2}\right) &= 2\int_{\tau_0}^1 \frac{\mathrm{d}x}{x} \int_{\frac{\tau_0}{x}}^1 \frac{\mathrm{d}y}{y^2} \\ & f_b^{(4)}(x)\, f_{\bar{b}}^{(4)}\left(\frac{\tau_0}{xy}\right) \left[\sigma_{b\bar{b}}^{(4),(0)}\left(y, \frac{m_H^2}{m_b^2}\right) + \alpha_s\, \sigma_{b\bar{b}}^{(4),(1)}\left(y, \frac{m_H^2}{m_b^2}\right)\right] \\ & + 4\,\alpha_s\, f_b^{(4)}(x)\, f_g^{(4)}\left(\frac{\tau_0}{xy}\right)\, \sigma_{bg}^{(4),(1)}\left(y, \frac{m_H^2}{m_b^2}\right), \end{aligned} \tag{6.4}$$

where the subscript P denotes the fact that this contribution is only present

when the b PDF is independently parametrized, and the superscript (4) is used to denote the massive factorization scheme, as in refs. [45, 144]. Note that even though, with a parametrized b there are five flavors also in the massive scheme, only the four lightest ones contribute to the running of α_s and perturbative evolution. The massive cross-sections $\sigma_{ij}^{(4),(k)}$ were computed e.g. in ref. [139] based on corresponding QED results from ref. [147] and are collected in appendix D.2 after scheme change as we discuss below.

Note that in the matched-b computation of ref. [45, 144] this process in the massive scheme starts at $O(\alpha_s^2)$, hence up to NLO (with our new counting) the contribution given in eq. (6.4) is the only one to $\sigma^{\rm massive}$ eq. (6.1): so in actual fact in this case

$$\sigma^{\rm massive,\,NLO} = \sigma_P^{\rm massive,\,NLO}. \tag{6.5}$$

The expression of $\sigma^{\rm massive,\,NLO}$ suitable for use in the FONLL formula eq. (6.1) is obtained, as mentioned, by re-expressing the massive scheme PDFs and α_s in terms of massless-scheme ones. For simplicity we assume that this is done at a matching scale $\mu_b = m_b$. The matching condition for α_s is, as well known,

$$\alpha_s^{(4)}(Q^2) = \alpha_s^{(5)}(Q^2)\left[1 - \alpha_s \frac{T_R}{2\,\pi}\,\log\frac{Q^2}{m_b^2} + \mathcal{O}(\alpha_s^2)\right] \tag{6.6}$$

while to $O(\alpha_s)$ the only nontrivial matching condition is that for the b PDF:

$$\begin{aligned} f_b^{(4)}(x) = f_b^{(5)}(x,Q^2) - \alpha_s \int_x^1 \frac{\mathrm{d}z}{z}\Bigg[& K_{bb}^{(1)}\left(z,Q^2\right) f_b^{(5)}\left(\frac{x}{z},Q^2\right) \\ & + K_{bg}^{(1)}(z,Q^2)\, f_g^{(5)}\left(\frac{x}{z},Q^2\right)\Bigg] + \mathcal{O}(\alpha_s^2), \end{aligned} \tag{6.7}$$

where again the superscripts (4) and (5) denote respectively the massive- and massless-scheme expressions, and K_{ij} are the matching coefficients

$$f_i^{(5)}(Q^2) = \sum_j K_{ij}(Q^2) \otimes f_j^{(4)}(Q^2), \tag{6.8}$$

where the sum runs over all partons (including the heavy quark), and

$$K_{ij}(Q^2) = \delta_{ij}\delta(1-z) + \sum_{n=1} \alpha_s^n(Q^2) K_{ij}^{(n)}(Q^2). \tag{6.9}$$

Note that, of course, since there is a heavy quark PDF also in the massive scheme, K_{ij} is a square matrix, so that, to $O(\alpha_s)$, $K_{ij}^{-1}(Q^2) = \delta_{ij} - \alpha_s(Q^2) K_{ij}^{(1)}(Q^2)$. The

matching function $K_{bb}^{(1)}$ was calculated in ref. [148]. Its explicit expression is given for ease of reference in appendix D.1 together with that of the splitting functions P_{ij}.

Substituting eqs. (6.6-6.7) in eq. (6.4) we get the desired expression for the massive-scheme cross section:

$$\sigma_P^{\text{massive}}\left(\frac{m_H^2}{m_b^2}\right) = \int_{\tau_H}^1 \frac{dx}{x}\int_{\frac{\tau_H}{x}}^1 \frac{dy}{y^2} \sum_{ij=b,g} f_i^{(5)}(x,Q^2) f_j^{(5)}\left(\frac{\tau_H}{xy},Q^2\right) B_{ij}\left(y,\alpha_s^{(5)}(Q^2),\frac{Q^2}{m_b^2}\right), \tag{6.10}$$

where to $O(\alpha_s)$ the non-vanishing coefficients are

$$B_{b\bar{b}}^{(0)}\left(y,\frac{m_H^2}{m_b^2}\right) = \sigma_{b\bar{b}}^{(4),(0)}\left(y,\frac{m_H^2}{m_b^2}\right) \tag{6.11}$$

$$B_{b\bar{b}}^{(1)}\left(y,\frac{m_H^2}{m_b^2}\right) = \sigma_{b\bar{b}}^{(4),(1)}\left(y,\frac{m_H^2}{m_b^2}\right) - 2\,\sigma_0 \int_y^1 \mathrm{d}z\, z\, \delta(z-y)\, K_{bb}^{(1)}\left(z,\ln\frac{m_H^2}{m_b^2}\right) \tag{6.12}$$

$$B_{bg}^{(1)}\left(y,\frac{m_H^2}{m_b^2}\right) = \sigma_{bg}^{(4),(1)}\left(y,\frac{m_H^2}{m_b^2}\right) - \sigma_0 \int_y^1 \mathrm{d}z\, z\, \delta(z-y)\, K_{bg}^{(1)}\left(z,\ln\frac{m_H^2}{m_b^2}\right), \tag{6.13}$$

whose explicit expressions are collected, as mentioned, in appendix D.2.

In order to construct the FONLL expression eq. (6.1), the massive scheme contribution must be combined with the difference term σ^{d} eq. (6.2). However, it is easy to check that, just like in the case of electroproduction[141, 142], this term, which is subleading when using matched b, vanishes identically with parametrized b. This is due to the fact that the massless limit of the massive-scheme calculation only differs from the massless-scheme calculation because of the resummation of mass logarithms $\ln\frac{Q^2}{m_b^2}$ beyond the accuracy of the massive-scheme result (so at $O(\alpha_s^2)$ and beyond, in our case). However, when re-expressing the massive-scheme calculation in terms of massless-scheme PDFs the evolution of the b-PDF is only removed up to the same accuracy as that of the massive scheme calculation. This is seen explicitly in eq. (6.7), in which mass logarithms $\ln\frac{Q^2}{m_b^2}$ are only removed up to $O(\alpha_s)$. Therefore, the higher order logarithms remain unsubtracted in the expression of $f_b^{(5)}(x,Q^2)$ and thus cancel exactly between σ^{massless} and $\sigma^{\text{massive},0}$.

The FONLL result thus reduces to the expression eq. (6.10):

$$\sigma^{\text{FONLL-AP}} = \sigma_P^{\text{massive}} \left(\frac{m_H^2}{m_b^2} \right). \tag{6.14}$$

We can thus write the FONLL result in the form

$$\sigma^{\text{FONLL-AP}} = \sum_{i,j} \sum_{l,m} \sigma_{ij}^{\text{massive}} \left(\frac{m_h^2}{m_b^2} \right) \otimes K_{il}^{-1} \otimes f_l^{(5)} \left(Q^2 \right) K_{jm}^{-1} \otimes f_m^{(5)} \left(Q^2 \right), \tag{6.15}$$

where K_{il}^{-1} is the inverse of the matching matrix defined in eq. (6.6), perturbatively defined order by order according to eq. (6.9). This is of course well defined with a parametrized b because K_{ij} is a square matrix. As discussed in detail in refs. [141, 142] the effect of the inverse matching matrices in eq. (6.15) is to remove collinear logarithms related to the evolution of the b PDF from the massless scheme PDFs $f_i^{(5)}$, since these are already included in the massive-scheme matrix cross-section $\sigma_{ij}^{\text{massive}}$, where they would appear as mass logarithms $\ln \frac{Q^2}{m_b^2}$ in the large Q^2 limit (in actual facts here $Q^2 = m_H^2$). As a consequence, the result eq. (6.15) is completely independent of the matching scale m_b (i.e. the scale at which the b PDF is parametrized), as it must be, given that once the b PDF is parametrized there is no matching scale left. We will check this cancellation explicitly (see fig. 6.2 below).

Eq. (6.15) shows that FONLL effectively acts as a massive five-flavor scheme, in which standard five-flavor PDFs are combined with the massive-scheme cross-section, with massive quarks included in the initial state: it is in fact akin to five-flavor scheme of ref. [139], though in this reference mass effects are also included in parton showering, which we do not consider here. In FONLL corrections are consistently included to the order at which the massive-scheme cross-section is computed, with collinear and mass logarithms resummed to the logarithmic order to which PDFs are used. The structure of the result eq. (6.15) is universal, and so are the PDFs which appear in it. Therefore, to the extent that the PDF is correctly fitted, mass corrections (i.e. all terms suppressed by powers of m_b/Q) are then fully included up to the order of the massive-scheme calculation: $O(\alpha_s)$ for FONLL-AP and FONLL-BP. Of course these latter corrections are not universal and will have to be computed separately for each process.

As mentioned, the FONLL framework allows for the combination of massive- and massless-scheme computations performed at arbitrary, independent accuracy. We discuss specifically the two cases defined in sect. 6.1.1, FONLL-AP and

FONLL-BP. In FONLL-AP, the massive-scheme partonic cross sections $\sigma_{ij}^{\text{massive}}$ are computed up to NLO (i.e. $O(\alpha_s)$), while the PDFs are evolved using NLO (more properly, NLL) evolution equations. Hence, in this case eq. (6.15), with $\sigma_{ij}^{\text{massive}}$ computed up to $O(\alpha_s)$ (i.e. including the diagrams of fig. 6.1), and NLO PDFs.

In FONLL-BP, the massive-scheme computation is still performed up to NLO, but now the massless-scheme computation is performed up to NNLO. This has two consequences. The first is that NNLO PDFs are now used. The other is that hard cross-sections are now computed up to NNLO i.e. up to $O(\alpha_s^2)$. Because massive terms are included only up to $O(\alpha_s)$, eq. (6.15) must now be supplemented by a purely massless $O(\alpha_s^2)$ contribution:

$$\sigma^{\text{FONLL-BP}} = \sigma^{\text{FONLL-AP}} + \sum_{l,m} \sigma_{lm}^{(5),(2)} \otimes f_l^{(5)}\left(Q^2\right) f_m^{(5)}\left(Q^2\right), \tag{6.16}$$

where $\sigma^{\text{FONLL-AP}}$ is given by eq. (6.15). Note that because the matching functions K_{ij}^{-1} are used to re-express the massive-flavor scheme cross-section in the massless scheme, they are accordingly computed to the same accuracy as the massive-scheme partonic cross-section itself: so here to $O(\alpha_s)$. The difference term σ^{d} eq. (6.2) always vanishes identically. It is clear that the computation is considerably streamlined in comparison to the standard FONLL framework of refs. [45, 144].

6.2 Higgs production in b fusion

We now present explicit results for Higgs production in b-quark fusion within the FONLL-AP and FONLL-BP scheme, and compare them to previous results of refs. [45, 144]. Results presented in this section are obtained using the following set-up. For the calculation of the 5F scheme coefficient functions, we use the interface to the `bbh@nnlo` code [149] as implemented in the public `bbhfonll` code [150]. The subtraction terms needed for the FONLL-B calculation of refs. [45, 144] is obtained using `bbhfonll`. The standard contributions to the 4F scheme are computed using the MG5_aMC@NLO package [151, 152], while we have implemented the new terms $\delta\sigma_P^{\text{massive}}$ eq. (6.4) and their massless limit in a new version of `bbhfonll`, following the expressions reported in appendix D.2. Both codes use the LHAPDF [153] package.

We use the NNPDF3.1 NNLO set of parton distributions with $\alpha_s(M_z) = 0.118$ [48]. For a first default comparison we just use the default NNPDF3.1 set, including the b PDF. From the point of view of a computational framework in which the b PDF is fitted, this can be thought of as the b PDF that one would get if initial PDFs were parametrized at $Q_0 = m_b$, and the fitted b PDF were to turn out to be exactly equal to that given by the matching condition at this scale. Furthermore, in order to get a feeling for effects related to the size of the b-PDF we then consider, for the sake of argument, a b PDF equal to that which would be obtained by using the matching condition at $\mu_b = 2/3m_b$ or $\mu_b = 1/2m_b$, and then evolving up to $Q = m_b$ where the initial PDF is given.

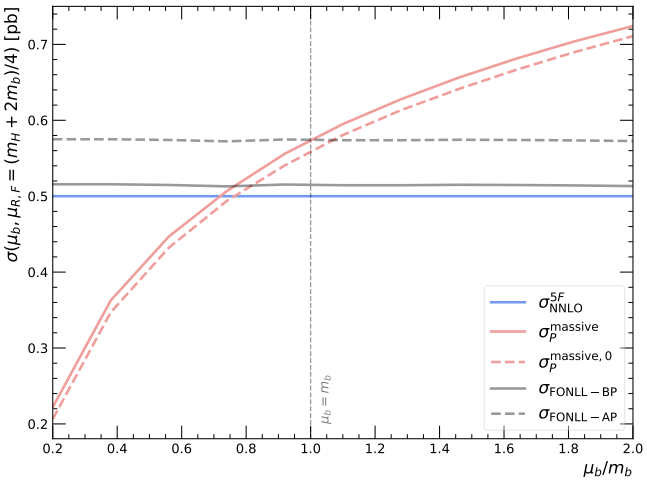

Figure 6.2 *Cancellation of the dependence on the matching scale in the FONLL-AP and FONLL-BP schemes.*

First, as a consistency check, in fig. 6.2 we verify that indeed the dependence on μ_b cancels when constructing the FONLL result with parametrized b according to eq. (6.3). In this figure the massive-scheme result has been constructed using a fixed b PDF (that which corresponds to the standard matching condition at $\mu_b = m_b$) and then re-expressing results in terms of the massive scheme PDFs and α_s in terms of massless-scheme ones. This is done using eq. (6.8), which contains the matching coefficients K_{ij} which depend on the matching scale μ_b, and thus the massive-scheme result becomes μ_b-dependent. However, this dependence cancels exactly in the final FONLL result.

In fig. 6.3 we show the factorization and renormalization scale dependence of the cross-section computed in various schemes, with the other scale kept fixed at the preferred [45, 144] value $\mu = \frac{m_H+2m_b}{4}$. Specifically, we compare results obtained using the FONLL-AP and FONLL-BP schemes discussed above, the pure five-

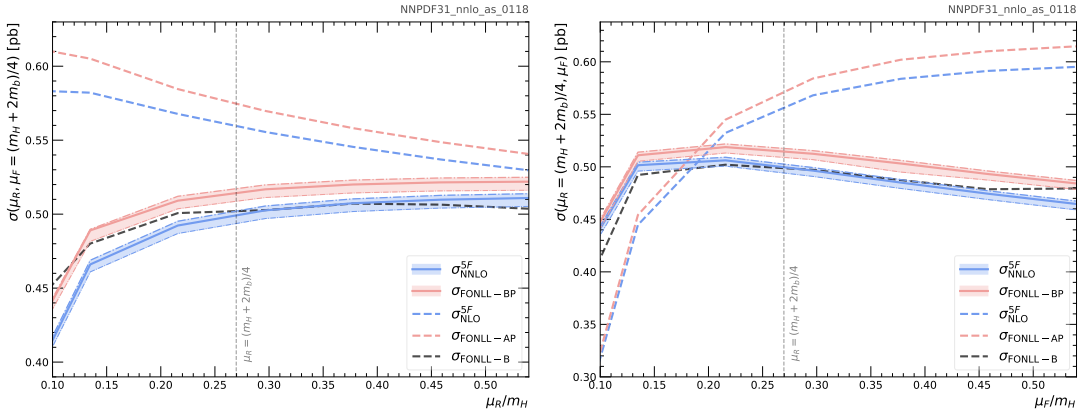

Figure 6.3 *Renormalization (left) and factorization (right) scale variation of the cross-section for Higgs production in bottom fusion. The pure five-flavor scheme computation is compared to the FONLL-AP and FONLL-BP results presented here and to the FONLL-B result of ref. [45]. For the pure five flavor NNLO and the FONLL-BP three curves are shown, corresponding to three choices of initial b PDF (see text).*

flavor scheme , and the FONLL-B result of ref. [144], all using the same PDFs (including the b PDF) as discussed above. For the pure five-flavor scheme and for FONLL-BP we also show the three curves corresponding to the three different choices for the b PDF discussed above, with a corresponding band: the central, thick, solid line represents the default $\mu_b = m_b$ choice, while the edges of the band are drawn with dot-dash curves with decreasing thickness, with the thicker of the two corresponding to $\mu_b = 2\frac{m_b}{3}$, and the other two $\mu_b = \frac{m_b}{2}$.

Note that the FONLL-BP computation eq. (6.16) and the FONLL-B result [144] are directly comparable: indeed, they both include the five-flavor scheme computation up to NNLO, and combine it with the first two orders of the massive-scheme computation. The difference is that in FONLL-B in the massive-scheme computation refers to the process $gg \to b\bar{b}H$, while in FONLL-BP it refers to $b\bar{b} \to H$. If the b PDF is the same as given by perturbative matching, the difference is then only that, in the latter case, only mass effects related to the $b\bar{b}$ which fuses into the Higgs are included, while in the former, also those related to the further unobserved $b\bar{b}$ pair are present. In a realistic situation, in which FONLL-BP is used while parametrizing and fitting the b PDF these mass effects should be reabsorbed in the fitted b PDF. In our comparisons, they appear as a certain enhancement of FONLL-BP in comparison to FONLL-B due to the opening of phase space.

Otherwise, the qualitative features of the comparison between FONLL and the pure five-flavor scheme remain essentially the same as discussed in ref. [144]: FONLL is quite close to the five-flavor scheme, with mass effects a non-negligible, but small, positive correction. Indeed, the difference between FONLL-AP and FONLL-BP, i.e., the impact of NNLO corrections in the five-flavor scheme, is much more significant than that of mass corrections. The impact of varying the b PDF by an amount which is comparable to a reasonable variation of the matching scale is clearly comparable to that of the mass corrections. This provides evidence for the fact that fitting the b PDF is likely to have a significant impact on precision phenomenology.

Note that results for the FONLL-B scheme differ at the percent level from those of ref. [144] because there a different PDF set and m_b value were used, for the sake of benchmarking with ref. [145, 146]. This further highlights the fact that the size of effects due to the b PDF is comparable to that of mass corrections.

In summary, we have shown how the FONLL matching of massive- and massless-scheme treatment of computations involving heavy quarks can be generalized to the case in which the heavy quark PDF is freely parametrized for hadronic processes. We have show that this effectively provides us with a massive heavy quark scheme, in which the heavy quark is endowed with a standard PDF satisfying QCD evolution equations, yet it is treated as massive in hard matrix elements. A first application to Higgs production in bottom fusion shows that effects related to the b PDF (whose size is estimated by the pink band of fig. 6.3) are quite likely to be comparable to mass corrections (whose size is given by the difference between the FONLL-B and 5FS results plotted in fig. 6.3): both are small, but non-negligible corrections to a purely massless NNLO calculation in which the b PDF is obtained from perturbative matching conditions. Determining the b PDF from data is thus likely to be necessary for a description of b-induced hadron collider processes at percent or sub-percent accuracy.

As a direction for further study, it should be noticed that extending our results to NNLO — thereby allowing the construction of a FONLL-CP result, in the terminology of sec. 6.1.1 (NNLO+NNLL) — is beyond current knowledge. Indeed, starting at NNLO the cancellation between real and virtual corrections is no longer trivial, and is spoiled by massive quarks in the initial state [154, 155]. This problem has been recently revised in ref. [156]. Hence, such an extension to NNLO would require conceptual advances in the understanding of QCD factorization in the presence of massive quarks, which are left for future studies.

7

PDFs from lattice data: theoretical framework

Given the central role parton distributions have in the analysis of experimental data at hadronic colliders, it would be highly beneficial to be able to use lattice QCD to determine these crucial ingredients in our current understanding of nucleon structure. The non-perturbative nature of PDFs makes them a natural candidate for a lattice investigation, however it has been known for a long time that it is not possible to obtain them directly from first principle computations, due to the Euclidean metric of the lattice, which does not allow to access the light-cone matrix element of eq. (2.26). To overcome this issue, several methods have been recently formulated, and in recent years, there has been a significant effort within the lattice community to study and compute specific Euclidean quantities that, in turn, can be related to PDFs through a factorization theorem. Two examples of these are quasi-PDFs[1] and pseudo-PDFs. They were introduced in refs. [159, 160], and since then numerous publications have appeared, addressing the main theoretical issues for these approaches: the definition and renormalization of the non-local operators involved in the lattice simulation [161–171], the proof of the factorization theorem between PDFs and Euclidean matrix elements [157, 172–176], the computation of the matching coefficients relating lattice-computable quantities to PDFs in different renormalization schemes [172, 173, 175–184]. For recent reviews, we refer the reader to refs. [158, 185–190].

This program has often been referred to as the "first principles computation of PDFs", generating different reactions among the lattice and high-energy physics communities: on the one hand it has been welcomed with enthusiasm, triggering

[1]Quasi-PDFs are one example of the more general LaMET formalism [157, 158], but here we focus on the collinear x-dependent distributions.

several dedicated studies; on the other hand it has been criticized in refs. [191, 192] on the basis that equal-time correlators do not give access to the full non-perturbative PDF. Both reactions are healthy and show the importance of the original proposal in [159]. This criticism mentioned above has, in turn, been addressed in refs. [169, 193]. Given the increasing number of lattice calculations, there is a need to revise and clarify the main conceptual questions: that is, how do we extract information on PDFs from quasi- and pseudo-PDFs, and what is the interplay between quasi- and pseudo-PDFs with experimental data?

In this chapter, following ref. [11], we introduce and study these topics in the context of a renormalizable scalar theory, setting the stage for the phenomenological studies presented in the next two chapters. Scalar field theory is a valuable model for understanding the essential theoretical issues in a simple framework, as shown in the pioneering study of PDFs by Collins in ref. [4]. We follow the ideas presented there, which we extend to account for quasi- and pseudo-PDFs. Our aim is to investigate, clarify and highlight some subtle points using scalar field theory as a simple playground, and to assess how the lattice QCD results that are currently available can be used to extract PDFs.

We will consider a massive scalar field theory, with a ϕ^3 interaction term in $d = 6$ dimensions, in order to have a marginal interaction. The bare Lagrangian $\mathcal{L}$ is given by

$$\mathcal{L} = \frac{1}{2}\left(\partial\phi\right)^2 - \frac{m^2}{2}\phi^2 - \frac{g}{3!}\phi^3. \tag{7.1}$$

Working within this model allows us to analyze the conceptual framework for quasi- and pseudo-PDFs in a clean and straightforward way, avoiding complications associated with QCD that are unnecessary for understanding the basics of these approaches. We focus in particular on the matrix element of a field bilinear between "nucleon" states:

$$\mathcal{M} = \langle P|\phi\left(z\right)\phi\left(0\right)|P\rangle\,, \tag{7.2}$$

when the separation z between the fields is either light-cone like, $z^2 = 0$, or purely spatial, $z^2 = -z_3^2$. In the first case, we obtain the matrix element that underlies the formal definition of collinear PDFs [4, 5], which are obtained as the Fourier

transform along a light-cone direction of the matrix element in eq.(7.2) [2]:

$$f(x) = xP^+ \int \frac{dz^-}{2\pi}\, e^{-ixP^+z^-} \,\langle P|\phi\,(z)\,\phi\,(0)\,|P\rangle\,, \tag{7.3}$$

where P^+ and z^- are the usual light-cone coordinates of the four-vectors P and z respectively. Eq. (7.3) is the scalar analogous of eq. (2.26), giving the PDF definition in QCD. In the second case we obtain an equal-time correlator that can be computed on the lattice. We address the problem of the renormalization of these quantities and study the relation between them at one loop in perturbation theory, both in position and momentum space. As we shall see, the main features of the computation are the same as in QCD. This allows us to understand easily the main concepts, relations and limitations of the quasi- and pseudo-PDF approaches. With a clear picture of the theoretical background and of what is currently available in the literature, we then propose a general framework to extract collinear PDFs from the available lattice data, based on the optimization of a parametric form of the PDFs within the `NNPDF` framework. Such program is carried on in chapters 8, 9.

We address, in turn, a number of questions that have been raised in the context of QCD, and analyze the lessons that we can draw from the scalar model. First we discuss issues that are related to the analysis of ultraviolet (UV) divergences of the bilinear operator and their subtraction through the renormalization process. In particular in sec. 7.1 we perform the computation of $\mathcal{M}$ in the case of a light-cone separation, recovering the results of ref. [4] through a position space calculation. In sec. 7.2 we perform the same exercise outside the light-cone, choosing a purely spatial separation between fields, and we discuss the main differences with respect to the light-cone case. In both cases, we define quantities that are free of divergences when the regulator is removed, and then focus on the relation between light-cone and equal-time correlators. In sec. 7.3 we work out this relation explicitly at one loop in perturbation theory, and analyze the limits leading to a factorization theorem, in both position and momentum space, and in sec. 7.4 we extend the discussion to include smeared equal-time correlators. In sec. 7.5 we summarize, discuss how these ideas can be used in a fitting framework to extract PDFs, and draw our conclusions. In app. E we report the technical details of the computations and we address the objections raised in refs. [191, 192].

[2] The field bilinear needs to undergo a proper renormalization, which we explore in detail in this Chapter.

7.1 Light-cone separation

As stressed in ref. [160], the matrix element defined in eq. (7.2) is a function of the Lorentz invariants z^2 and $\nu = P \cdot z$, the "Ioffe time", so that we can write $\mathcal{M} = \mathcal{M}(\nu, z^2)$. In this section we focus on the perturbative renormalization of $\mathcal{M}(\nu, z^2)$ at the one-loop level, in the light-cone separation case, $z^2 = 0$. We work in perturbation theory, denoting the bare field of our theory as ϕ, and we consider partonic matrix elements

$$\widehat{\mathcal{M}}\left(\nu, z^2\right) = \langle p|\phi(z)\,\phi(0)\,|p\rangle \tag{7.4}$$

between on-shell quark states with four-momentum p, with $p^2 = m^2_{\text{pole}}$. Throughout this calculation, we denote partonic quantities with a "hat", while the lower-case p refers to the momentum of the parton. In what follows the Lorentz invariant ν is defined as $\nu = p \cdot z$. Restricting ourselves to the case for which $z_0 \geqslant 0$, we have

$$\begin{aligned}\widehat{\mathcal{M}}\left(\nu, z^2\right) &= \langle p|T\left[\phi(z)\,\phi(0)\right]|p\rangle \\ &= \lim_{p^2 \to m^2_{\text{pole}}} \left(p^2 - m^2_{\text{pole}} + i\epsilon\right)^2 \int dz_1\, dz_2\, e^{-ip\cdot z_1} e^{ip\cdot z_2} \langle 0|T\left[\phi(z)\,\phi(0)\,\phi(z_1)\,\phi(z_2)\right]|0\rangle\,,\end{aligned} \tag{7.5}$$

where m^2_{pole} is defined by the location of the pole in the scalar propagator, and can be computed at each order in perturbation theory. At tree level we have $m^2_{\text{pole}} = m^2$, while in general $m^2_{\text{pole}} - m^2 = \mathcal{O}(g^2)$.

When computing the 4-point function entering eq. (7.5), we will not consider diagrams like those in fig. 7.1. Following ref. [4], we are only interested in the contribution proportional to $\exp(-ip \cdot z)$, and therefore discard topologies like the one in diagram (a). Diagram (b) is removed by considering the connected contribution only.

Therefore the only Feynman diagrams contributing to eq. (7.5) up to one-loop order are those shown in fig. 7.2. Denoting the propagator in position space as

$$\langle 0|T\left[\phi(x)\,\phi(y)\right]|0\rangle = \overbrace{\phi_x \phi_y}\,, \tag{7.6}$$

the Wick contraction that contributes to the tree level diagram (a) of fig. 7.2 is

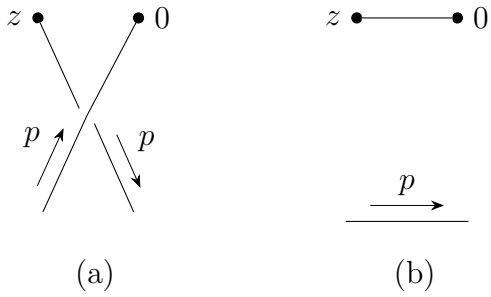

Figure 7.1 *Contractions that are not considered in the present discussion. Diagram (a) is excluded when considering contributions proportional to* $\exp(-ip\cdot z)$*, while diagram (b) cancels when looking at the connected correlator.*

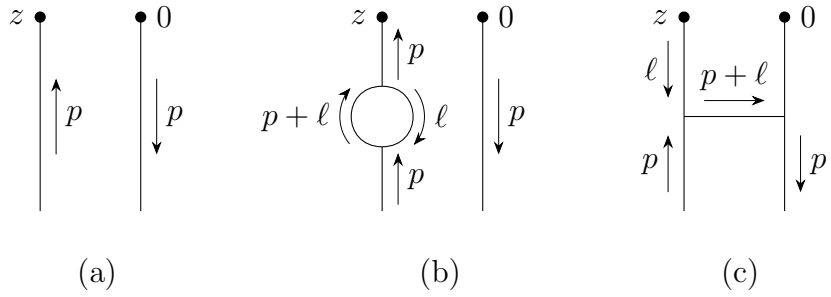

Figure 7.2 *Feynman diagrams up to one loop for the matrix element* $\langle 0|T\left[\phi\left(z\right)\phi\left(0\right)\phi\left(z_1\right)\phi\left(z_2\right)\right]|0\rangle$*.*

given by

$$\overbrace{\phi_z\phi_{z_1}}\overbrace{\phi_{z_2}\phi_0} = \int_{l_1}\frac{i\,e^{-il_1\cdot(z-z_1)}}{l_1^2-m^2+i\epsilon}\int_{l_2}\frac{i\,e^{-il_2\cdot z_2}}{l_2^2-m^2+i\epsilon}\,, \tag{7.7}$$

where we use the notation

$$\int_k = \int\frac{\mathrm{d}^dk}{(2\pi)^d}\,. \tag{7.8}$$

Plugging eq. (7.7) in eq. (7.5) we obtain the tree level expression for $\widehat{\mathcal{M}}\left(\nu,z^2\right)$

$$\widehat{\mathcal{M}}^{(0)}\left(\nu,z^2\right) = -e^{-i\nu} \equiv \widehat{\mathcal{M}}^{(0)}\left(\nu,0\right). \tag{7.9}$$

Note that the tree level result does not depend on the invariant separation z^2 and therefore we can set $z^2=0$ in the second equality above.

At one-loop order the self-energy diagram (b) yields the mass and wave function

renormalization. Its contribution to eq. (7.5) is

$$\widehat{\mathcal{M}}_{\text{self}}\left(\nu, z^2\right) = R\,\widehat{\mathcal{M}}^{(0)}\left(\nu, 0\right) , \tag{7.10}$$

where R is the $\mathcal{O}\left(g^2\right)$ contribution to the residue of the propagator at the pole mass. In $d = 6 - 2\epsilon$ dimensions, we have

$$R = \frac{d\Pi\left(l^2\right)}{dl^2}_{l^2=p^2_{\text{pole}}} = \alpha\left[\frac{1}{12}\log\frac{m^2}{\mu^2} + \frac{1}{12}\frac{1}{\epsilon} + \frac{b}{2}\right], \tag{7.11}$$

where $b/2$ is a finite contribution and $\alpha = g^2/(64\pi^3)$. The same $\mathcal{O}\left(\alpha\right)$ contribution is obtained from the diagram with the self energy corrections on the second leg, so that the total contribution coming from the tree level plus self-energy corrections is

$$\widehat{\mathcal{M}}_{\text{self}}\left(\nu, z^2\right) = \left[1 + \alpha\left(\frac{1}{6}\log\frac{m^2}{\mu^2} + \frac{1}{6}\frac{1}{\epsilon} + b\right)\right]\widehat{\mathcal{M}}^{(0)}\left(\nu, 0\right) + \mathcal{O}\left(\alpha^2\right) . \tag{7.12}$$

Note the absence of any z^2 dependence: as far as the first two diagrams of fig. 7.2 are concerned, there are no differences between the light-cone and the pure spatial case. This is to be expected, since the one-loop diagrams (b) simply implement the mass and wave function renormalization.

We can now move to the computation of the remaining $\mathcal{O}\left(\alpha\right)$ term, *i.e.* diagram (c). This contraction is given by

$$\begin{aligned}
&\int dw_1\, dw_2\, \overbrace{\phi_z\phi_{w_1}}\overbrace{\phi_{w_1}\phi_{z_1}}\overbrace{\phi_{w_2}\phi_{w_1}}\overbrace{\phi_{w_2}\phi_0}\overbrace{\phi_{z_2}\phi_{w_2}} = \\
&\quad = (-ig)^2\int dw_1\, dw_2 \int_{l_1}\frac{ie^{-il_1\cdot(z-w_1)}}{l_1^2 - m^2 + i\epsilon}\int_{l_2}\frac{ie^{-il_2\cdot(w_1-z_1)}}{l_2^2 - m^2 + i\epsilon}\int_{l_3}\frac{ie^{-il_3\cdot(w_2-w_1)}}{l_3^2 - m^2 + i\epsilon}\times \\
&\qquad\qquad \times\int_{l_4}\frac{ie^{-il_4\cdot w_2}}{l_4^2 - m^2 + i\epsilon}\int_{l_5}\frac{ie^{-il_5\cdot(z_2-w_2)}}{l_5^2 - m^2 + i\epsilon}.
\end{aligned} \tag{7.13}$$

Plugging this into eq. (7.5), we have

$$\begin{aligned}
\widehat{\mathcal{M}}^{(1)}\left(\nu, z^2\right) &= -i\,g^2\int_k \frac{e^{-ik\cdot z}}{\left(k^2 - m^2 + i\epsilon\right)^2}\frac{1}{\left(p-k\right)^2 - m^2 + i\epsilon} \\
&= g^2\int_0^1 d\xi\ \left(1-\xi\right)K\left(z^2, M^2\right)\ \widehat{\mathcal{M}}^{(0)}\left(\xi\nu, 0\right) ,
\end{aligned} \tag{7.14}$$

where we have introduced a Feynman parameter ξ and defined

$$K\left(z^2, M^2\right) = 2i \int_q \frac{e^{-iq\cdot z}}{\left(q^2 - M^2 + i\epsilon\right)^3}, \tag{7.15}$$

with

$$q = k - \xi p\,, \tag{7.16}$$

$$M^2 = m^2\left(1 - \xi + \xi^2\right)\,. \tag{7.17}$$

The integral $K\left(z^2, M^2\right)$ can be computed by performing a Wick rotation $z_E^\mu = (iz^0, \vec{z})$ and using

$$\frac{1}{\left(q_E^2 + m^2\right)^\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^\infty dT\, T^{\alpha-1} e^{-T\left(q_E^2 + m^2\right)}\,. \tag{7.18}$$

We obtain

$$K\left(z^2, M^2\right) = 2\int \frac{d^d q_E}{(2\pi)^d} \frac{e^{iq_E z_E}}{\left(q_E^2 + M^2\right)^3} = \int_0^\infty dT\, T^2 e^{-TM^2} \int \frac{d^d q_E}{(2\pi)^d}\, e^{iq_E z_E - Tq_E^2}$$

$$= \frac{1}{(4\pi)^{\frac{d}{2}}} \int_0^\infty \frac{dT}{T} T^{3-\frac{d}{2}}\, e^{-TM^2} e^{-\frac{z_E^2}{4T}}\,, \tag{7.19}$$

where in the last line we have performed the Gaussian integral over $d^d q_E$.

Since we are considering the case of a light-cone separation $z_E^2 = -z^2 = 0$, $K\left(0, M^2\right)$ in $d = 6$ dimensions is logarithmically divergent. The divergence arises from the lower end of the integral over T, as the exponential suppression factor in the integrand vanishes on the light-cone. We apply dimensional regularization, taking $d = 6 - 2\epsilon$ and introducing the $\overline{\text{MS}}$ scale μ through the rescaling of the coupling $g^2 \to g^2 e^{\gamma_E}\mu^2/(4\pi)$. We find

$$K\left(0, M^2; \mu^2\right) = \int_0^\infty \frac{dT}{T}\, \left(T\mu^2 e^{\gamma_E}\right)^\epsilon e^{-TM^2} = \Gamma(\epsilon)\left(\frac{\mu^2 e^{\gamma_E}}{M^2}\right)^\epsilon = \frac{1}{\epsilon} + \log\frac{\mu^2}{M^2}\,, \tag{7.20}$$

where the pole in $1/\epsilon$ reflects the original logarithmic divergence in dimensional regularization. Putting everything together, we obtain the full one-loop expres-

sion of the bare position space matrix element in dimensional regularization

$$\widehat{\mathcal{M}}(\nu,0) = \left[1+\alpha\left(\frac{1}{6}\log\frac{m^2}{\mu^2}+\frac{1}{6}\frac{1}{\epsilon}+b\right)\right]\widehat{\mathcal{M}}^{(0)}(\nu,0) + \alpha\int_0^1 d\xi\,(1-\xi)\left(\frac{1}{\epsilon}+\log\frac{\mu^2}{m^2(1-\xi+\xi^2)}\right)\widehat{\mathcal{M}}^{(0)}(\xi\nu,0)\ . \quad (7.21)$$

The structure of the divergences in eq. (7.21) shows that this quantity can be renormalized by convolution with a renormalization kernel $\mathcal{K}$. Denoting the renormalized matrix element as $\widehat{\mathcal{M}}_R(\nu,0,\mu^2)$, we have

$$\widehat{\mathcal{M}}_R\left(\nu,0,\mu^2\right) = \int_0^1 dy\,\mathcal{K}(y)\,\widehat{\mathcal{M}}(y\nu,0)\ . \quad (7.22)$$

The specific choice of the finite terms that appear in the kernel $\mathcal{K}(y)$, together with subtraction of the $1/\epsilon$ poles, defines the renormalization scheme. For example, in the $\overline{\text{MS}}$ scheme, the renormalization kernel is

$$\mathcal{K}(y) = \delta(1-y) - \alpha\left[\frac{1}{6\,\epsilon}\delta(1-y)+\frac{1}{\epsilon}(1-y)\right]\ , \quad (7.23)$$

and the corresponding renormalized quantity is

$$\widehat{\mathcal{M}}_R\left(\nu,0,\mu^2\right) = \left[1+\alpha\left(\frac{1}{6}\log\frac{m^2}{\mu^2}+b\right)\right]\widehat{\mathcal{M}}^{(0)}(\nu,0) + \alpha\int_0^1 d\xi\,(1-\xi)\log\frac{\mu^2}{m^2(1-\xi+\xi^2)}\,\widehat{\mathcal{M}}^{(0)}(\xi\nu,0)\ . \quad (7.24)$$

We conclude this derivation with a comment on the form of the renormalization kernel $\mathcal{K}$ given in eq. (7.23): the contribution proportional to a delta function is a multiplicative renormalization term, implementing the subtraction of the singularities generated by diagram (b) of fig. 7.2, which is basically the wave function renormalization. The second contribution, $-\frac{\alpha}{\epsilon}(1-y)$, implements the renormalization of the one-loop diagram (c) of fig. 7.2, and because this contribution is not proportional to a delta function, the renormalization of this term is not multiplicative, but requires a convolution.

Taking the log derivative of eq. (7.24) we obtain

$$\mu^2\frac{d}{d\mu^2}\widehat{\mathcal{M}}_R\left(\nu,0,\mu^2\right) = \alpha\int_0^1 d\xi\,P(\xi)\,\widehat{\mathcal{M}}_R\left(\xi\nu,0,\mu^2\right)+\mathcal{O}\left(\alpha^2\right)\ , \quad (7.25)$$

where the $\mathcal{O}(\alpha)$ Altarelli-Parisi splitting kernel is given by

$$P(\xi) = (1-\xi) - \frac{1}{6}\delta(1-\xi) = (1-\xi)_+ + \frac{1}{3}\delta(1-\xi) \ . \tag{7.26}$$

The renormalized collinear PDF is defined from the renormalized matrix element,

$$\widehat{\mathcal{M}}_R\left(\nu, 0, \mu^2\right) = \int_0^1 dx\, e^{ix\nu} \widehat{f}\left(x, \mu^2\right) \ , \tag{7.27}$$

and therefore, from eq. (7.25),

$$\mu^2 \frac{d}{d\mu^2} \widehat{f}\left(x, \mu^2\right) = \alpha \int_x^1 \frac{d\xi}{\xi} P(\xi)\ \widehat{f}\left(\frac{x}{\xi}, \mu^2\right) \ , \tag{7.28}$$

which yields the standard DGLAP evolution equations, which were already obtained in ref. [4] for the scalar theory. As detailed in sec. 2.3.2, the solution of eq. (7.28) in perturbation theory is given by an evolution kernel $\Gamma(x, \mu, \mu_0, \alpha)$, which resums the large collinear logarithms and allows the PDF at a generic scale μ to be computed in terms of the PDF at the scale μ_0 as

$$\widehat{f}\left(x, \mu^2; \theta\right) = \int_x^1 \frac{d\xi}{\xi} \Gamma\left(\frac{x}{\xi}, \mu, \mu_0, \alpha_s\right) \widehat{f}\left(\xi, \mu_0^2; \theta\right) \ . \tag{7.29}$$

7.2 Spatial separation

We now consider the case in which the separation between the fields is purely spatial $z_E^2 = z_3^2$. As seen in the previous section, the z^2 dependence enters only through diagram (c) of fig. 7.2. Considering this contribution, the kernel $K(z^2, M^2)$ defined in eq. (7.19) is no longer divergent for $z_3 \neq 0$, as the term $\exp\left[-z_E^2/(4T)\right]$ regulates the small-T behaviour. The integral can evaluated directly in $d = 6$ dimensions, yielding

$$K\left(-z_3^2, M^2\right) = \frac{1}{64\pi^3} \int_0^\infty \frac{dT}{T} e^{-T} e^{-\frac{(Mz_3)^2}{4T}} = \frac{1}{64\pi^3}\, 2K_0\left(Mz_3\right), \tag{7.30}$$

where K_0 is the modified Bessel function. Plugging eq. (7.30) into eq. (7.14) we obtain the contribution from diagram (c) in the case of purely spatial separation:

$$\widehat{\mathcal{M}}^{(1)}\left(\nu, -z_3^2\right) = \alpha \int_0^1 d\xi\ (1-\xi)\, 2K_0\left(Mz_3\right) \widehat{\mathcal{M}}^{(0)}\left(\xi\nu, 0\right). \tag{7.31}$$

Note that, as long as $z_3 \neq 0$, this contribution does not contain any UV divergences. For $Mz_3 \to 0$ the Bessel function diverges logarithmically, and we recover the UV divergence of the light-cone case.

The full one-loop bare matrix element is then given by

$$\widehat{\mathcal{M}}\left(\nu, -z_3^2\right) = \left[1 + \alpha\left(\frac{1}{6}\log\frac{m^2}{\mu^2} + \frac{1}{6}\frac{1}{\epsilon} + b\right)\right]\widehat{\mathcal{M}}^{(0)}\left(\nu, 0\right) + \alpha\int_0^1 d\xi\ (1-\xi)\, 2K_0\left(Mz_3\right)\widehat{\mathcal{M}}^{(0)}\left(\xi\nu, 0\right). \tag{7.32}$$

As before, this quantity can be renormalized by convolution,

$$\widehat{\mathcal{M}}_R\left(\nu, -z_3^2; \mu^2\right) = \int_0^1 dy\, \tilde{\mathcal{K}}\left(y\right)\widehat{\mathcal{M}}\left(y\nu, -z_3^2\right). \tag{7.33}$$

However, since the only UV pole comes from the self-energy contributions, the kernel $\tilde{\mathcal{K}}\left(y\right)$ is proportional to a delta function. For example, in the $\overline{\text{MS}}$ scheme we can take

$$\tilde{\mathcal{K}}\left(y\right) = \delta\left(1-y\right)\left[1 - \alpha\frac{1}{6\,\epsilon}\right]. \tag{7.34}$$

In other words, in the case of purely spatial separation the renormalization of the matrix element is purely multiplicative [167]. The additional UV divergence we had to remove in the light-cone case is substituted here by a finite contribution $K_0\left(Mz_3\right)$. The corresponding renormalized quantity is

$$\widehat{\mathcal{M}}_R\left(\nu, -z_3^2; \mu^2\right) = \left[1 + \alpha\left(\frac{1}{6}\log\frac{m^2}{\mu^2} + b\right)\right]\widehat{\mathcal{M}}^{(0)}\left(\nu, 0\right) + \alpha\int_0^1 d\xi\ (1-\xi)\, 2K_0\left(Mz_3\right)\widehat{\mathcal{M}}^{(0)}\left(\xi\nu, 0\right). \tag{7.35}$$

Note also that both eqs. (7.24) and (7.35) contain an infrared (IR) divergence regularized by the mass m: in the former the mass is manifest in the log, while in the latter the mass appears in the Bessel function, which diverges logarithmically for $m \to 0$.

7.3 Factorization theorem

Having defined the renormalized correlators in the previous sections, let us investigate the one-loop relation between the light-cone and the equal-time correlators. Combining eqs. (7.24) and (7.35) we write

$$\begin{aligned}\widehat{\mathcal{M}}_R\left(\nu,-z_3^2;\,\mu^2\right) =&\widehat{\mathcal{M}}_R\left(\nu,0,\mu^2\right)\\ &+\alpha\int_0^1 d\xi\;(1-\xi)\left(2K_0\left(Mz_3\right)-\log\frac{\mu^2}{M^2}\right)\widehat{\mathcal{M}}_R\left(\xi\nu,0,\mu^2\right)\,,\end{aligned} \tag{7.36}$$

and using eq. (7.27) we find

$$\widehat{\mathcal{M}}_R\left(\nu,-z_3^2;\mu^2\right)=\int_0^1 dx\,\tilde{C}\left(x\nu,mz_3,\frac{\mu^2}{m^2}\right)\widehat{f}\left(x,\mu^2\right)\,, \tag{7.37}$$

with

$$\tilde{C}\left(x\nu,mz_3,\frac{\mu^2}{m^2}\right)=e^{ix\nu}-\alpha\int_0^1 d\xi\;(1-\xi)\left(2K_0\left(Mz_3\right)-\log\frac{\mu^2}{M^2}\right)e^{i\xi x\nu}\,. \tag{7.38}$$

This expression shows the connection between the collinear PDFs and an equal-time correlator, through a convolution with a perturbative kernel. In general, the latter contains a logarithmic dependence on m^2, namely the kernel contains IR singularities. However, as we will see, these singularities cancel exactly when taking a specific limit, leaving an expression free from IR poles, which therefore has the form of a proper factorization theorem. Before discussing this in detail, we recall that, although eq. (7.37) has been worked out in perturbation theory, considering matrix elements between on-shell quark states, the renormalization of the bilocal operators discussed so far does not depend on our choice of specific external states. It follows that eq. (7.37) holds also for external proton states. From now on we will refer to full proton matrix elements rather than partonic ones, removing the symbol '$\widehat{\ }$' used so far to denote partonic quantities.

7.3.1 Factorization theorem in position space: small-z_3^2 limit

The behavior of the coefficient $\tilde{C}$ in the small-z_3^2 limit is obtained by expanding the Bessel function as

$$2K_0\left(Mz_3\right) = -\log\left(M^2 z_3^2\right) + 2\log\left(2e^{-\gamma_E}\right) + \mathcal{O}\left(M^2 z_3^2\right) , \tag{7.39}$$

so that eq. (7.37) becomes

$$\mathcal{M}_R\left(\nu, -z_3^2; \mu^2\right) = \int_0^1 dx\, \tilde{C}\left(x\nu, \mu^2 z_3^2\right) f\left(x, \mu^2\right) , \tag{7.40}$$

with

$$\tilde{C}\left(x\nu, \mu^2 z_3^2\right) = e^{ix\nu} - \alpha \int_0^1 d\xi\ (1-\xi) \log\left(\mu^2 z_3^2 \frac{e^{2\gamma_E}}{4}\right) e^{i\xi x\nu} + \mathcal{O}\left(m^2 z_3^2\right) . \tag{7.41}$$

We note that in this limit the logarithmic behaviour of the Bessel function matches that of the light-cone quantity, so that the two matrix elements display the same IR behaviour: as a result the coefficient $\tilde{C}$ is IR safe, and eq. (7.40) represents a proper factorization theorem connecting a lattice computable quantity on the left hand side with a collinear PDF on the right hand side.

We note that this factorization also applies to the so-called reduced distributions [194, 195], the quantities usually determined in lattice calculations in the pseudo-PDF approach, first introduced in ref. [160]. They were originally defined as

$$\mathfrak{M}\left(\nu, -z_3^2\right) = \frac{\mathcal{M}_R\left(\nu, -z_3^2; \mu^2\right)}{\mathcal{M}_R\left(0, -z_3^2; \mu^2\right)} , \tag{7.42}$$

although a double ratio was proposed in [196]. Here we restrict our attention to the ratio defined in eq. (7.42). In the context of our model, using the small-z_3^2 limit of eq. (7.36) we have

$$\begin{aligned}
\mathfrak{M}\left(\nu, -z_3^2\right) &= \mathcal{M}_R\left(\nu, 0, \mu^2\right) \\
&\quad - \alpha \log\left(\mu^2 z_3^2 \frac{e^{2\gamma_E}}{4}\right) \int_0^1 d\xi\ (1-\xi) \left[\mathcal{M}_R\left(\xi\nu, 0, \mu^2\right) - \mathcal{M}_R\left(\nu, 0, \mu^2\right)\right] \\
&= \mathcal{M}_R\left(\nu, 0, \mu^2\right) - \alpha \log\left(\mu^2 z_3^2 \frac{e^{2\gamma_E}}{4}\right) \int_0^1 d\xi\ (1-\xi)_+ \mathcal{M}_R\left(\xi\nu, 0, \mu^2\right) ,
\end{aligned} \tag{7.43}$$

and therefore

$$\mathfrak{M}\left(\nu,-z_3^2\right)=\int_0^1 dx\,\tilde{C}_+\left(x\nu,\mu^2 z_3^2\right)f\left(x,\mu^2\right)\,, \tag{7.44}$$

with

$$\tilde{C}_+\left(x\nu,\mu^2 z_3^2\right)=e^{ix\nu}-\alpha\,\log\left(\mu^2 z_3^2\frac{e^{2\gamma_E}}{4}\right)\int_0^1 d\xi\,\left(1-\xi\right)_+ e^{i\xi x\nu}+\mathcal{O}\left(m^2 z_3^2\right)\,. \tag{7.45}$$

Note the absence of any μ^2 dependence on the left hand side of eqs. (7.42) and (7.44): the perturbative dependence on the renormalization scale μ^2 cancels exactly in the ratio, leaving a quantity that depends only on the scale z_3^2. More precisely, eqs. (7.44), (7.45) show how, in the small-z_3^2 limit, the renormalization scale dependence of $\mathcal{M}_R\left(\nu,0,\mu^2\right)$ generated by diagram (c) is replaced by an equal z_3^2 dependence that can be obtained from the former through the substitution

$$\mu^2\to\frac{4e^{-2\gamma_E}}{z_3^2}\,.$$

In other words, the factorization formula worked out in this section predicts a logarithmic dependence on z_3^2 for the equal-time correlator, which replaces the analogous logarithmic behaviour of the PDFs on the renormalization scale μ^2, predicted by the one-loop DGLAP. Such dependence on z_3^2 should be visible in real lattice QCD data when working in the factorization regime, and indeed it was observed in refs. [195, 196].

7.3.2 Factorization theorem in momentum space: large P_3 limit

A factorization theorem can also be established working in momentum rather than in position space. Taking the Fourier transform of eq. (7.40) with respect to z_3 and defining

$$q\left(y,\mu^2,P_3^2\right)=\frac{P_3}{2\pi}\int_{-\infty}^{\infty}dz_3\,e^{-iyP_3z_3}\,\widehat{\mathcal{M}}\left(P_3z_3,-z_3^2\right)\,, \tag{7.46}$$

$$C\left(\eta,\frac{m^2}{x^2P_3^2},\frac{\mu^2}{m^2}\right)=\int_{-\infty}^{\infty}\frac{d\theta}{2\pi}\,e^{-i\theta\eta}\,\tilde{C}\left(\theta,\frac{m\theta}{xP_3},\frac{\mu^2}{m^2}\right)\,, \tag{7.47}$$

we obtain

$$q\left(y,\mu^2,P_3^2\right)=\int_0^1\frac{dx}{x}\,f\left(x,\mu^2\right)C\left(\frac{y}{x},\frac{m^2}{x^2P_3^2},\frac{\mu^2}{m^2}\right)\,, \tag{7.48}$$

with

$$C\left(\eta,\frac{m^2}{x^2P_3^2},\frac{\mu^2}{m^2}\right)=\int_0^1 d\xi\;(1-\xi)\left[\frac{1}{\sqrt{(\eta-\xi)^2+\frac{M^2}{x^2P_3^2}}}-\delta\left(\xi-\eta\right)\log\frac{\mu^2}{M^2}\right]\,. \tag{7.49}$$

Note that taking the Fourier transform, as in eq. (7.46), involves an integration of the Bessel function $K_0\left(z_3M\right)$ through its singularity at $z_3=0$, which is discussed in detail in app. E.1. Looking at eq. (7.49), we note that, again, the coefficient C contains explicit logarithms of the mass, rendering it infrared divergent. However, these divergences cancel when considering the large P_3 regime, by expanding the Fourier transform of the Bessel function in the limit $\frac{M^2}{\xi^2P_3^2}\to 0$. If $\eta>1$ or $\eta<0$, then looking at eq. (7.49) we have

$$\lim_{P_3\to\infty}C\left(\eta,\frac{m^2}{x^2P_3^2},\frac{\mu^2}{m^2}\right)=C\left(\eta\right)=\pm\int_0^1 d\xi\,\frac{1-\xi}{\eta-\xi}=\pm\left[(1-\eta)\log\frac{\eta}{\eta-1}+1\right]\,, \tag{7.50}$$

where the solution with the plus refers to $\eta>1$, and the one with the minus to $\eta<0$. On the other hand, if $\eta\in(0,1)$, the factor $1/|\eta-x|$ generated in this limit produces a non-integrable singularity at $\eta=x$ [197]. To overcome this issue, as detailed in app. E.1, we can write

$$\frac{1}{\sqrt{(\eta-\xi)^2+\frac{M^2}{x^2P_3^2}}}=\log 4\eta\,(1-\eta)\,\frac{x^2P_3^2}{M^2}\delta\left(\eta-\xi\right)+\frac{1}{|\eta-\xi|_+}+\mathcal{O}\left(\frac{M^2}{P_3^2}\right)\,, \tag{7.51}$$

so that in the region $\eta\in(0,1)$ we find

$$\begin{aligned}
&C\left(\eta,\frac{M^2}{x^2P_3^2},\frac{\mu^2}{M^2}\right)\overset{P_3\to\infty}{\sim}C\left(\eta,\frac{\mu^2}{x^2P_3^2}\right)\\
&\quad=\int_0^1 d\xi\;(1-\xi)\left[\frac{1}{|\eta-\xi|_+}+\delta\left(\eta-\xi\right)\log 4\eta\,(1-\eta)\,\frac{x^2P_3^2}{\mu^2}\right]+\mathcal{O}\left(\frac{m^2}{P_3^2}\right)\\
&\quad=2\eta-1+(1-\eta)\log 4\eta\,(1-\eta)\,\frac{x^2P_3^2}{\mu^2}+\mathcal{O}\left(\frac{m^2}{P_3^2}\right)\,.
\end{aligned} \tag{7.52}$$

Note the cancellation of the logarithmic dependence on the mass, which leads again to a proper factorization formula, this time in momentum space. We conclude that, in momentum space, the factorization theorem is realized in the limit $P_3 \to \infty$ and in our model this factorization theorem takes the form

$$q\left(y,\mu^2,P_3^2\right) = \int_0^1 \frac{dx}{x}\, f\left(x,\mu^2\right) C\left(\frac{y}{x},\frac{\mu^2}{x^2P_3^2}\right) + \mathcal{O}\left(\frac{m^2}{P_3^2}\right), \tag{7.53}$$

with

$$C\left(\eta,\frac{\mu^2}{x^2P_3^2}\right) = \delta\left(1-\eta\right) + \alpha \begin{cases} \left(1-\eta\right)\log\frac{\eta}{\eta-1} + 1 & \eta > 1 \\ \left(1-\eta\right)\log 4\eta\left(1-\eta\right)\frac{x^2P_3^2}{\mu^2} + 2\eta - 1 & 0 < \eta < 1 \\ -\left(1-\eta\right)\log\frac{\eta}{\eta-1} - 1 & \eta < 0 \end{cases} . \tag{7.54}$$

Factorization in position space, given in eqs. (7.40) and (7.41), is equivalent to factorization in momentum space, given in eqs. (7.53) and (7.54). In other words, taking the small-z_3^2 limit in position space is entirely equivalent to taking the large-P_3 limit in momentum space. This can be easily verified by computing the Fourier transfom of the small-z_3^2 coefficient $\tilde{C}$ of eq. (7.41), and checking that it is equal to the high-P_3 coefficient C of eq. (7.54)

$$\begin{aligned} \frac{1}{x}\, C\left(\eta,\frac{\mu^2}{x^2P_3^2}\right) &= \frac{P_3}{2\pi}\int_{-\infty}^{\infty} dz_3\, e^{-iyP_3z_3}\, \tilde{C}\left(x\nu,\mu^2z_3^2\right) \\ &= \frac{1}{x}\int_{-\infty}^{\infty}\frac{d\theta}{2\pi}\, e^{-i\theta\eta}\, \tilde{C}\left(\theta,\frac{\mu^2\theta^2}{x^2P_3^2}\right) \quad \text{with} \quad \eta = \frac{y}{x}. \end{aligned} \tag{7.55}$$

This check, despite being conceptually straightforward, does require some care [176]. We provide the details of the computation in app. E.1. The implementation of the factorization theorem in position space, together with the definition of reduced distributions, are the typical approach followed in nonperturbative calculations of pseudo-PDFs [160, 196, 198–201], while the realization of the factorization in momentum space characterizes the quasi-PDF approach [159, 181, 202–204]. The latter will be addressed from a phenomenological point of view in chapter 8, the former in chapter 9.

As we have shown in this section in the simplified context of our model, these two approaches are conceptually equivalent, and related by a Fourier transform: in one case the lattice calculation needs to provide the correlators for small values of

z_3, while in the other large values of P_3 are required. In both scenarios, however, the object that is actually computed is the matrix element of spatially-separated fields. This is the only quantity of interest, without the need to define either pseudo- or quasi-PDFs.

7.4 Smeared distributions

In ref. [205, 206], the gradient flow was proposed as an approach to control the power divergence associated with the Wilson-line operator that defines the Ioffe time distribution in QCD. The gradient flow [207–209] is a classical, gauge-invariant, one-parameter mapping of the theory that exponentially damps the UV fluctuations. This corresponds to smearing in real space, with a smearing scale that is parametrised by the flow time. In the limit of small flow time, the matrix elements of smeared fields can be related to those at vanishing flow time by a short flow-time expansion [210].

In Yang-Mills theories, gauge invariance ensures that no new divergences are introduced at finite flow time. Thus, provided the boundary theory is properly renormalized, the matrix elements of composite operators composed of fields at finite flow time are guaranteed to be finite. In the absence of gauge symmetries, the simplest method for maintaining this property is to exclude interactions from the flow time evolution of the fields, in which case this evolution corresponds to simple Gaussian smearing [211–213].

The flow time can be viewed as a non-perturbative regulator that does not affect the infrared properties of correlation functions. The smeared Ioffe-time matrix elements, constructed from fields at finite flow time, therefore satisfy the same factorization theorems as the original Ioffe-time matrix elements [205]. In the scalar case, the boundary fields $\phi(x)$ in eq. (7.4) are replaced by fields at finite flow time $\rho(t;x)$, so that the partonic matrix element becomes

$$\widehat{\mathcal{M}}_t\left(\nu,\overline{z}^2\right) = \langle p|\rho\left(t;z\right)\rho\left(t;0\right)|p\rangle\,. \tag{7.56}$$

Here the subscript indicates that the fields are evaluated at flow time t, and $\overline{z}^2 = z^2/t$.

The gradient flow is only well-defined in Euclidean space, but for $z^2 < 0$, the matrix elements are signature independent [174]. The tree-level and one-loop

diagrams that contribute to this matrix element are exactly those given in fig. 7.2, with $\phi(x)$ replaced by $\rho(t;x)$. Working in the small flow-time regime, where contributions of $\mathcal{O}(t)$ can be neglected, the only diagram that must be calculated is diagram (c) of fig. 7.2. Therefore, we can deduce the factorization properties of the smeared matrix element directly from the analogue of eq. (7.14) at nonzero flow time

$$\begin{aligned}\widehat{\mathcal{M}}_t^{(1)}\left(\nu,-\overline{z}_3^2\right) &= g^2\int_{k_E} e^{-2k_E^2 t}\,\frac{e^{-ik_{\mathrm{E}}z_3}}{\left(k_{\mathrm{E}}^2+m^2\right)^2}\frac{1}{\left(p_{\mathrm{E}}-k_{\mathrm{E}}\right)^2+m^2}\\ &= g^2\int_0^1 d\xi\;(1-\xi)\,K_t\left(-\overline{z}_3^2,\overline{M}^2\right)\,\widehat{\mathcal{M}}^{(0)}\left(\xi\nu,0\right)\,, \end{aligned} \tag{7.57}$$

where the exponential damping is the result of the smearing of the fields and we have introduced $\overline{M}^2 = M^2 t$. Here the kernel $K_t\left(-\overline{z}_3^2,\overline{M}^2\right)$ is given by

$$K_t\left(-\overline{z}_3^2,\overline{M}^2\right) = \frac{\mu^{6-d}}{(4\pi)^{d/2}}e^{-2m^2t\xi}\int_0^\infty \mathrm{d}T\,\frac{T^2}{(T+2t)^{d/2}}\,e^{-TM^2}e^{(4\xi t p_{\mathrm{E}}-iz_{\mathrm{E}})^2/(4(T+2t))}\,, \tag{7.58}$$

which reduces to the kernel in eq. (7.19) when $t=0$.

By introducing the further dimensionless variables $\overline{\mu}^2=\mu^2 t$, $\overline{m}^2=m^2t$, and

$$\beta^2 = -\frac{1}{t}\left(\xi t p_{\mathrm{E}}^\mu - \frac{iz_{\mathrm{E}}^\mu}{2}\right)^2 = \xi^2\overline{m}^2 + i\xi\nu + \frac{\overline{z}_3^2}{4}\,, \tag{7.59}$$

and changing variables to $u = T/t+2$, the integral becomes

$$K_t\left(-\overline{z}_3^2,\overline{M}^2\right) = \frac{\overline{\mu}^{6-d}}{(4\pi)^{d/2}}e^{-2(\xi-1)^2\overline{m}^2}\int_2^\infty \mathrm{d}u\,\frac{(u-2)^2}{u^{d/2}}e^{-u\overline{M}^2-\beta^2/u}\,. \tag{7.60}$$

This integral can be solved in terms of *incomplete Bessel functions* [214–216], which can be studied in various asymptotic regimes. In particular,

$$\begin{aligned}K_t\left(-\overline{z}_3^2,\overline{M}^2\right) &= \frac{2\,\overline{\mu}^{6-d}}{(4\pi)^{d/2}}e^{-2\frac{(\xi-1)^2}{1-\xi+\xi^2}\overline{M}^2}\\ &\quad\times\left[K_0(2\,|\overline{M}\beta|,2) - 4\frac{\overline{M}}{|\beta|}K_1(2\,|\overline{M}\beta|,2) + 4\frac{\overline{M}^2}{\beta^2}K_2(2\,|\overline{M}\beta|,2)\right]\,,\end{aligned} \tag{7.61}$$

where

$$K_n(y,a) = K_n(y) - J(y,n,a)\,, \tag{7.62}$$

with $J(y,n,a)$ the finite integral

$$J(y,n,a) = \int_0^a \mathrm{d}v\, e^{-y\cosh(v)} \cosh(nv)\,. \tag{7.63}$$

This result is finite in six dimensions, because the incomplete Bessel functions are finite at finite flow time and quark mass. Indeed, one can evaluate these integrals numerically by imposing a cutoff. For sufficiently large cutoff, the results are independent of the cutoff value. Using eq. (7.62), eq. (7.61) can be written as

$$\begin{aligned} K_t\left(-\overline{z}_3^2, \overline{M}^2\right) = \frac{2}{(4\pi)^3} e^{-2\frac{(\xi-1)^2}{1-\xi+\xi^2}\overline{M}^2} \Big\{ & K_0(2\,|\overline{M}\beta|) - 4\frac{\overline{M}}{|\beta|} K_1(2\,|\overline{M}\beta|) + 4\frac{\overline{M}^2}{\beta^2} K_2(2\,|\overline{M}\beta|) \\ & - J(2\,|\overline{M}\beta|, 0, 2) + 4\frac{\overline{M}}{|\beta|} J(2\,|\overline{M}\beta|, 1, 2) - 4\frac{\overline{M}^2}{\beta^2} J(2\,|\overline{M}\beta|, 2, 2) \Big\}\,. \end{aligned} \tag{7.64}$$

In the limit where

$$\frac{t^2 m^2}{z_E^2} \ll 1\,, \tag{7.65}$$

the argument of the Bessel functions, $|\overline{M}\beta|$, can be expressed as

$$2|\overline{M}\beta| = M|z_E| + \mathcal{O}\left(\frac{t^2m^2}{z_E^2}\right) = Mz_3 + \mathcal{O}\left(\frac{t^2m^2}{z_E^2}\right)\,, \tag{7.66}$$

so that, in the limit of small z_3 we can expand them as

$$2K_0(2\,|\overline{M}\beta|) = -\log\left(M^2 z_3^2\right) + 2\log\left(2e^{-\gamma_E}\right) + \mathcal{O}\left(m^2 z_3^2, \frac{t^2m^2}{z_E^2}\right)\,, \tag{7.67}$$

$$2\frac{\overline{M}}{|\beta|} K_1(2\,|\overline{M}\beta|) = 0 + \mathcal{O}\left(m^2 z_3^2, \frac{t^2m^2}{z_E^2}, 1/\overline{z}^2\right)\,, \tag{7.68}$$

$$2\frac{\overline{M}^2}{\beta^2} K_2(2\,|\overline{M}\beta|) = 0 + \mathcal{O}\left(m^2 z_3^2, \frac{t^2m^2}{z_E^2}, 1/\overline{z}^2\right)\,. \tag{7.69}$$

Care must be taken when matching these expressions to the light-cone case. The limits need to be taken in the right order so that the quantity $\frac{t^2m^2}{z_E^2}$ remains small in the process. One must first consider the small flow time regime at fixed z_3, in which $\overline{z} \gg 1$, and then consider the limit in which $m^2 z_3^2$ goes to zero. Taking the limit of small $m^2 z_3^2$ at fixed t would violate the condition above and invalidate the factorization theorem, *viz.* data for values of t and z_3 that correspond to large

values of t^2m^2/z_E^2 are not described by the factorization theorems discussed here. With this in mind, the only logarithmic infrared divergence occurs in the first Bessel function, which has been expanded using eq. (7.39). Thus, in the small flow-time regime eq. (7.64) becomes

$$K_t\left(-\overline{z}_3^2, \overline{M}^2\right) = \frac{1}{(4\pi)^3}\left[-\log\left(M^2 z_3^2\right) + 2\log\left(2e^{-\gamma_E}\right) + \mathcal{R}(Mz_3)\right] + \mathcal{O}\left(m^2 z_3^2, \frac{t^2m^2}{z_E^2}, 1/\overline{z}^2\right), \tag{7.70}$$

where the rational function $\mathcal{R}(Mz_3)$ contains the IR finite contributions generated by the J functions of eq. (7.63). The logarithmic IR divergence in eq. (7.70), regularized by the mass m, matches those in eqs. (7.24) and (7.35).

In the short flow-time regime, the one-loop contributions to eq. (7.56) from diagrams (a) and (b) are just those given in eq. (7.12). The corresponding renormalized quantity at one loop is therefore

$$\widehat{\mathcal{M}}_t\left(\nu, -z_3^2; \mu^2\right) = \left[1 + \alpha\left(\frac{1}{6}\log\frac{m^2}{\mu^2} + b\right)\right]\widehat{\mathcal{M}}^{(0)}\left(\nu, 0\right) + \alpha\int_0^1 d\xi\ (1-\xi)\left[-\log\left(M^2 z_3^2\right) + 2\log\left(2e^{-\gamma_E}\right) + \mathcal{R}(Mz_3)\right]\widehat{\mathcal{M}}^{(0)}\left(x\nu, 0\right). \tag{7.71}$$

We can now directly relate this quantity, via a factorization relation, to the light-cone quantity $f(x, \mu^2)$ using eq. (7.27). We obtain

$$\widehat{\mathcal{M}}_t\left(\nu, -z_3^2; \mu^2\right) = \int_0^1 dx\, \overline{C}\left(x\nu, \mu^2 z_3^2\right)\widehat{f}\left(x, \mu^2\right), \tag{7.72}$$

with

$$\overline{C}\left(x\nu, \mu^2 z_3^2\right) = e^{ix\nu} - \alpha\int_0^1 d\xi\ (1-\xi)\left[\log\left(\mu^2 z_3^2 \frac{e^{2\gamma_E}}{4}\right) - \mathcal{R}(Mz_3)\right]e^{i\xi x\nu} + \mathcal{O}\left(m^2 z_3^2, \frac{t^2m^2}{z_E^2}, 1/\overline{z}^2\right). \tag{7.73}$$

This factorization relation provides the explicit connection between the collinear PDFs and an equal-time correlator at nonzero flow time, through a convolution with a perturbative kernel. Equal-time correlators at nonzero flow time therefore represent an additional class of lattice observables related to PDFs through a factorization theorem.

7.5 Towards PDFs from lattice data

We have addressed the definition and renormalization of equal-time correlators whose computation can be performed on the lattice, studying their relation with the corresponding light-cone matrix elements underlying the definition of collinear PDFs via factorization theorems. To highlight and clarify the most important aspects of the factorization theorems, we have studied them in the context of a nongauge theory. This allows us to avoid the formal complications that arise in QCD, which can obscure the key concepts. We derive the relation between the light-cone and Euclidean matrix elements at the one-loop level, and then study the limits that lead to well-defined factorization theorems. These relations express suitable correlators that are evaluated by Monte Carlo calculations in terms of a convolution between a collinear PDF and an infrared safe coefficient function, which can be evaluated in perturbation theory. We obtain factorization theorems in both position and momentum space, by considering the regimes of small-z_3^2 and large-P_3 respectively, and show that these limits are equivalent at one loop, which highlights the formal equivalence of the pseudo- and quasi-PDFs approach. In addition, we demonstrate that the gradient flow can be used to define a new class of lattice observables that satisfy factorization.

These ideas naturally suggest that the lattice data should be used in a fitting framework to extract PDFs, in the same way experimental data are usually included in global QCD analyses: the unknown x-dependence of the PDF at a specific fitting scale is parametrized by introducing a suitable functional form. The PDF at a generic scale can be computed in terms of its parametric form at the fitting scale, which then leads to a theoretical prediction for the lattice observable when working in either the small-z_3^2 or large-P_3 limit. Assuming that we have a set of lattice results for the real and imaginary part of the Ioffe-time matrix elements, a standard minimum-χ^2 fit yields the values of the free parameters that best describe such data. As in any other PDF determination, we highlight the importance of having a robust estimate of the full covariance matrix that enters the χ^2 definition, and this should be provided by the lattice group performing the calculation. It is important to stress that this procedure is exactly the one described in chapter 3, with the lattice matrix elements playing the same role as the cross-sections for high-energy processes. Given a discrete set of points for quantities that are connected to collinear PDFs through a factorization theorem, we can use them to perform a fit, thereby obtaining an estimate of the PDFs and

their corresponding error.

We conclude this chapter with an important remark. We have discussed the conceptual equivalence of the pseudo and quasi distribution methods, however we emphasize that conceptual equivalence may not translate to equivalence in practice. On the one hand, the LaMET approach relies on large hadronic momenta to suppress higher twist contamination. On the other hand, the pseudo distribution approach uses small spatial separations to suppress higher twist effects, but requires large momenta to cover a range of Ioffe times. In both cases, large values of the hadron momentum can lead to significant signal-to-noise challenges and discretization effects of the form $(aP)^n$. The interplay of higher twist contamination and discretization effects is nontrivial and will depend both on the details of the distribution itself and on the specific choice of discretization. These effects must be studied systematically, across a wide range of observables, to pin down systematic uncertainties and strengthen the role that lattice QCD can play in the determination of hadron structure.

In the next chapters we will address these issues from a phenomenological point of view, considering lattice QCD data for quasi- and pseudo-PDFs matrix elements and performing a first study of them within the `NNPDF` framework.

8

PDFs from quasi-PDFs matrix elements

Data for equal-time correlators coming from first lattice QCD simulations have started appearing and have gotten into already a relatively advanced stage over the last few years [164, 181–184, 202, 217–228]. Results from these studies give an idea of what PDFs from the lattice might look like, not only for nonsinglet quark PDFs of the nucleon, but also for the pion PDF and distribution amplitude, as well as for the gluon PDF of the nucleon. Given the general interest shown by the community, a quick improvement in the technologies involved in such lattice simulations is to be expected in the next few years. A great quantity of increasingly precise lattice data is then likely to be available in the near future, requiring detailed studies about the possible impact they might have on the overall precision of PDFs determination.

Despite the increasing number of numerical results becoming available, an optimal strategy for reconstructing the PDFs from these data has not been entirely addressed yet. The approach which has been initially used within the lattice community (and that is still employed in many analyses) consists in approximating the quasi-PDFs by mean of a discrete Fourier transform, starting from the limited number of points for the corresponding equal-time correlator available from numerical simulations. The continuous function resulting from this procedure is subsequently convoluted with the perturbative matching coefficients relating euclidean and light-cone distributions, (see eq. (7.54)) in order to obtain the final PDF. The numerical error introduced by this procedure is rather large and difficult to control, so that it generally provides unstable and inaccurate results. This problem was first addressed within the lattice community in ref. [229] where a series of possible approaches to tackle the problem of incomplete and discretized Fourier transform has been presented.

In this chapter, based on ref. [12], following the ideas of sec. 7.5 and the formalism described in sec. 7.3.2, we exploit the momentum space factorization of quasi-PDFs in PDFs and perturbatively computable coefficients to extract nonsinglet distributions from the data of refs. [181, 202], using the `NNPDF` framework described in chapter 3 and treating the lattice data on the same footing as experimental ones.

8.1 quasi-PDFs in QCD

In this section we extend the formalism introduced in chapter 7 to the full QCD case, giving formulas and expressions which have to be used when working with data coming from lattice QCD simulations, and referring to the original publications where they were first presented to further details. Denoting by Γ a generic Dirac structure and by the suffix A the specific nonsinglet distribution we want to consider, we consider the matrix element between nucleon states with momentum P given by

$$M^{(0)}_{\Gamma,A}(z,P) = \langle P|\mathcal{M}^{(0)}_{\Gamma,A}(z)|P\rangle\,, \tag{8.1}$$

with

$$\mathcal{M}^{(0)}_{\Gamma,A}(z) = \bar{\psi}^{(0)}(z)\,\lambda_A\Gamma\,U(z,0)\,\psi^{(0)}(0)\,. \tag{8.2}$$

with λ_A denoting the flavour structure and the gauge-link $U(z,0)$ given by eq. (2.27). eqs. (8.1), (8.2) represent the QCD generalization of the scalar quantity defined in eq. (7.2).

The vector bilocal operator obtained for $\Gamma = \gamma^\mu$ can be decomposed in terms of two form factors which only depend on the Lorentz invariants z^2 and $\nu \equiv -z\cdot P$ as

$$M^{(0)}_{\gamma^\mu,A}(z,P) \equiv M^{(0)}_{\mu,A}(z,P) = 2P_\mu\,\mathcal{M}^{(0)}_A\left(\nu,z^2\right) + z_\mu\,\mathcal{N}^{(0)}_A\left(\nu,z^2\right)\,. \tag{8.3}$$

By choosing a light-cone separation $z = (0,z^-,0_\perp)$ together with $\gamma^\mu = \gamma^+$ and $P = (P^+,0,0_\perp)$ we get

$$M^{(0)}_{+,A}(z,P) = 2P_+\,\mathcal{M}^{(0)}_A(\nu,0) = 2P_+\int_{-1}^{1} dx\,e^{ix\nu}f^{(0)}_A(x) \tag{8.4}$$

with $f_A^{(0)}(x)$ being the bare collinear nonsinglet parton distribution given in eq. (2.34). Because of the light-cone separation z involved in its definition, $M_+^{(0)}$ is not directly computable on a Euclidean lattice. We can define a different quantity that is amenable to lattice simulations by choosing a purely spatial separation, $z = (0, 0, 0, z_3)$, together with $\gamma^\mu = \gamma^0$ and $P = (E, 0, 0, P_3)$. Then taking the time component of eq. (8.3) we get

$$M_{0,A}^{(0)}(z, P) = 2E\, \mathcal{M}_A^{(0)}\left(\nu, -z_3^2\right) . \tag{8.5}$$

The correlators defined in eqs. (8.4) and (8.5) are known in the literature as (bare) *Ioffe-time* distribution (ITD) and pseudodistribution (pseudo-ITD) respectively [160, 230].

Taking the Fourier transform with respect to z, we obtain the definition of the quasi-PDF

$$\tilde{f}_A^{(0)}(x, P_z) = 2E \int_{-\infty}^{\infty} \frac{dz}{4\pi}\, e^{ix\mathrm{P}_z z}\, \mathcal{M}_A^{(0)}\left(zP_z, -z_3^2\right) . \tag{8.6}$$

As in the case of standard PDFs, the matrix elements defining the quasi-PDFs contain UV divergences, and need to be renormalized. The main features of the perturbative renormalization of a bilocal operator, as the one appearing in eq. (8.6), have been described in chapter 7 in the context of the scalar theory. When considering the QCD case, for $z_3^2 \neq 0$ in addition to usual ultraviolet (UV) divergences described in the scalar QFT case, specific link-related UV divergences arise, which are regularized by a finite lattice spacing a. Thus, $\mathcal{M}_A^{(0)}(\nu, -z_3^2)$ is in fact $\mathcal{M}_A^{(0)}(\nu, -z_3^2; a^2)$. As we found out looking at the scalar QFT, the position space operator appearing in eq. (8.6) can be multiplicatively renormalized [167], according to

$$\mathcal{M}_A\left(zP_z, -z_3^2, \mu\right) = Z_A(z_3^2)\, e^{\delta m|z|/a} \mathcal{M}_A^{(0)}\left(zP_z, -z_3^2; a^2\right) . \tag{8.7}$$

The only difference with respect to the scalar model is given by the exponential factor $e^{\delta m|z|/a}$, which reabsorbs the power divergence from the Wilson line which appears in the QCD case. The position-dependent factor $Z_A(z_3^2)$ takes care of the remaining UV logarithmic divergences. Importantly, we recall that the quasi-PDFs retain a dynamical dependence on the hadron momentum P, unlike PDFs, which are defined to be invariant under Lorentz boosts. Also, their support is defined to be the full real axis.

The interest in quasi-PDFs comes from the potential to relate them to light-cone PDFs in the limit of high values of P_z, as detailed in sec. 7.3.2; factorization allows us to rewrite the quasi-PDFs as a convolution of the light-cone PDFs with a coefficient function that can be computed in perturbation theory, up to corrections that are suppressed by inverse powers of P_z. It follows that they can be written as

$$\tilde{f}_A\left(x,\mu^2\right) = \int_{-1}^{1} \frac{dy}{|y|} C_A\left(\frac{x}{y}, \frac{\mu}{|y|P_z}, \frac{\mu}{\mu'}\right) f_A\left(y,\mu'^2\right) + \mathcal{O}\left(\frac{M^2}{P_z^2}, \frac{\Lambda^2_{\mathrm{QCD}}}{P_z^2}\right) , \quad (8.8)$$

where the terms $\mathcal{O}\left(\frac{M^2}{P_z^2}, \frac{\Lambda^2_{\mathrm{QCD}}}{P_z^2}\right)$ include the power corrections suppressed by the hadron momentum. The functions C_A, usually called matching coefficients, depend on the choice of the renormalization scheme, and on the kind of quasi-PDF under consideration. The first matching expressions, for all Dirac structures, were derived in ref. [172], using a simple transverse momentum cutoff scheme. In later works, matching coefficients were derived that relate the quasi-PDFs in different renormalization schemes to light-cone PDFs in the $\overline{\mathrm{MS}}$ scheme. The matching from $\overline{\mathrm{MS}}$ quasi-PDFs was first considered in ref. [179], both for non-singlet and singlet quark PDFs, as well as for gluons. Even though one can choose operators for the latter that do not mix with singlet quark quasi-PDFs under renormalization [170, 171], mixing under matching is inevitable. No mixing of the flavour nonsinglet sector with flavour singlet or gluon sectors occurs, as stated in eq. (8.8). Ref. [179] did not, however, address the known issue of self-energy corrections, exhibiting a logarithmic UV divergence. This was resolved in ref. [176] by adding terms outside of the plus prescription in the matching coefficient. As noticed in ref. [181], such prescription for renormalizing this divergence violates vector current conservation, *i.e.* the integral of the matched PDF is different from the integral of the input quasi-PDF, and not necessarily equal to 1 over the whole integration range. As a remedy, a modified matching expression, which is given explicitly in eq. (F.0.1) of app. F, was proposed in ref. [181]. It consists in resorting to pure plus functions when subtracting the logarithmic divergence in self-energy corrections. However, this is an additional subtraction with respect to the minimal subtraction of the $\overline{\mathrm{MS}}$ scheme and thus, defines a modified $\overline{\mathrm{MS}}$ scheme, the so-called M$\overline{\mathrm{MS}}$ scheme. As such, it requires the quasi-PDF to be expressed in this modified scheme. The expression for the conversion of $\overline{\mathrm{MS}}$-renormalized matrix elements to the M$\overline{\mathrm{MS}}$ scheme was worked out in ref. [202] and we refer to it for the details of the procedure. Nevertheless, this modification is numerically very small, as also shown in

ref. [202]. An alternative modification of the $\overline{\text{MS}}$ scheme that guarantees vector current conservation was derived in an updated version of ref. [176]. This defines the so-called "ratio" scheme. In this scheme, only pure plus functions are used, like in the M$\overline{\text{MS}}$ scheme, but the modification is done also for the "physical" region of $0 < z < 1$ (in the notation of eq. (F.0.1)). Thus, the expected numerical effect of this modification is larger, as shown explicitly in ref. [202]. For this reason, we choose to use the M$\overline{\text{MS}}$ procedure, with details of the lattice computation of the bare matrix elements and the renormalization in the M$\overline{\text{MS}}$ scheme outlined in the next section. Yet another possibility of performing the matching consists in directly relating the quasi-PDFs in the intermediate RI-type scheme to $\overline{\text{MS}}$ light-cone PDFs. This was proposed in ref. [180] for the unpolarized case. Obviously, such procedure is equivalent to the one adopted here, with possibly different systematic effects. All of the discussed papers considered the matching to only first order in perturbation theory, but NNLO results for the matching coefficients have recently become available in both position and momentum space [231–235].

8.2 Nonsinglet distributions from quasi-PDFs Matrix Elements

In this section, we describe the lattice data we will use in this chapter, presenting briefly the quasi-PDFs matrix elements (MEs) computed in refs. [181, 202]. Using the results recalled in the previous sections, we show that we can factorize such matrix elements into two nonsinglet distributions and a perturbatively computable coefficient, just as if they were experimental data for high-energy cross sections.

8.2.1 Lattice data

The field of nucleon isovector $(u - d)$ quasi-PDFs has matured in recent years. Exploratory studies for all types of collinear PDFs – unpolarized, helicity and transversity – were performed in 2014-2016 [217–220]. They used lattice ensembles with non-physical pion masses and the results had unsubtracted divergences, due to the lack of a well-defined renormalization procedure. The latter was proposed and applied for the first time in refs. [163, 164], utilizing a variant of the regularization-independent momentum subtraction scheme (RI'-

MOM) [236]. Moreover, another major progress for unpolarized PDFs was the identification of a lattice-induced mixing between the bilinear operator used in the first exploratory studies, which was defined using γ_z to determine the Dirac structure, and the scalar bilinear operator (in spin space) [163]. Even though in principle it is possible to compute the matrix elements of the latter and a mixing renormalization matrix to properly subtract the divergences [164], this is bound to lead to much deteriorated precision, due to the rather poor signal for the scalar operator. Instead, it is preferable to define the quark bilinear using the γ_0 Dirac matrix, since this procedure does not give rise to mixing. Moreover the quasi-PDF computed with it converges faster in powers of $1/P_z^2$ to the light-cone PDF, as argued in ref. [237]. Summarising in just one sentence, we could say that the major progresses with respect to the early works for unpolarized quasi-PDFs came from: (1) change of the Dirac structure in order to avoid the mixing, (2) non-perturbative renormalization procedure, (3) simulations at the physical pion mass. Matrix elements corresponding to such setup were computed in refs. [181, 202, 224] and they are briefly described below. For a recent review of other available results for quasi-PDF matrix elements, see *e.g.* ref. [188].

The data used in this chapter were computed by the Extended Twisted Mass Collaboration (ETMC)[1]. They used one ensemble of gauge field configurations with two degenerate light quarks [238] with masses chosen to reproduce the physical value of the pion mass ($m_\pi \approx 130$ MeV, *i.e.* slightly below the actual physical value). The lattice spacing is $a = 0.0938(3)(2)$ fm [239] and the lattice has $48^3 \times 96$ sites, corresponding to the spatial extent L of around 4.5 fm and $m_\pi L = 2.98$. ETMC calculated bare quasi-PDF matrix elements for the unpolarized, helicity and transversity cases, but we concentrate only on the unpolarized one. The lattice data are available for three nucleon boosts, $P_z = 6\pi/L$, $8\pi/L$ and $10\pi/L$ (0.83 GeV, 1.11 GeV and 1.38 GeV in physical units) and for four values of the temporal separation between the nucleon creation and annihilation operators, t_s/a=8, 9, 10, 12 (0.75, 0.84, 0.94, 1.13 fm). As shown in refs. [181, 202], there are signs of convergence in the nucleon momentum (the largest two momenta give compatible results), indicating that the boost is already enough to suppress higher-twist effects below statistical precision. Moreover, as pointed out in ref. [202], excited-states contamination at the largest source-sink separation is small, *i.e.* the single-state fits at this t_s are compatible with two-state fits including all four values of t_s. Hence, for the purpose of this study, we consider only the data at the largest nucleon boost and at the largest source-sink

[1]Until 2018 known as the European Twisted Mass Collaboration.

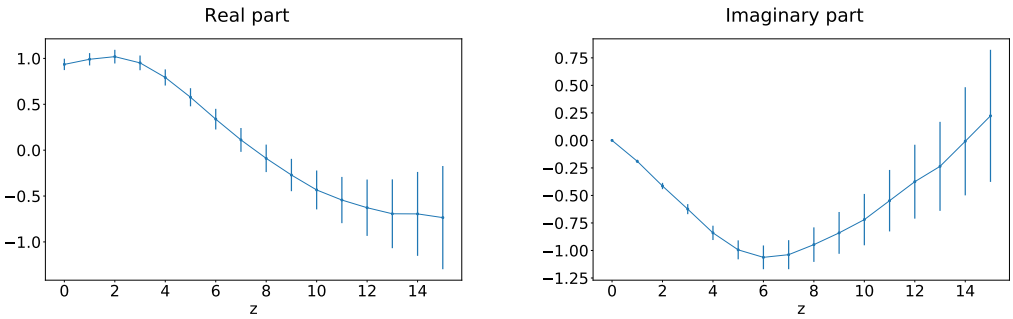

Figure 8.1 *Real (left) and imaginary (right) part of the quasi-PDF ME for the data used in this Chapter, computed in refs. [181, 202]. The error band displayed accounts only for statistical uncertainty. The separation z is expressed in units of the lattice spacing a.*

separation. They are shown in fig. 8.1.

As seen previously, the bare lattice data contain two types of divergences. First of all, there are standard logarithmic divergences with respect to the regulator, *i.e.* terms that behave like $\log(a\mu)$. Additionally, for non-zero Wilson line lengths, further power-like divergences appear. They resum into a multiplicative exponential factor, $e^{\delta m|z|/a}$, where δm is operator-independent. The desired renormalization scheme for the final results is the $\overline{\text{MS}}$ scheme of dimensional regularization. However, obviously, the latter is impossible on a lattice, restricted to integer dimensions. Thus, the usage of an intermediate lattice renormalization scheme is required. In ref. [164], it was proposed to use an RI'-type prescription. The renormalization conditions are enforced on the amputated vertex functions of operators with different Wilson line lengths z, setting them to their tree-level values. A similar renormalization condition is applied for the quark propagator. This results in a set of matrix elements renormalized in the RI' scheme. Thus, a perturbative conversion from the RI' to the $\overline{\text{MS}}$ scheme is needed. Such a conversion was derived in ref. [163] to one-loop order and was applied to the RI'-renormalized matrix elements. As we discussed in the previous section, to guarantee vector current conservation, we use a modified $\overline{\text{MS}}$ scheme, termed the M$\overline{\text{MS}}$ scheme. Thus, another perturbative conversion of the $\overline{\text{MS}}$-renormalized matrix elements is required, according to the formula given in ref. [202]. After this conversion, renormalized matrix elements in the M$\overline{\text{MS}}$ scheme are the starting point of the current analysis.

It is important to emphasize that despite having numerical evidence for the smallness of the effects of the nucleon momentum and of excited states, matrix

elements from lattice studies come with a variety of other systematic effects. We discuss them in the next subsection. For more details about the lattice computation of the matrix elements, we refer the reader to ref. [202].

8.2.2 Systematics in matrix elements of quasi-PDFs

A proper investigation of systematic effects in matrix elements evaluated in lattice QCD simulations is a difficult task, necessitating dedicated efforts. Such efforts consist in simulating with varied parameter values, such as the lattice spacing, the lattice volume, or the temporal separation between the source and the sink in nucleon three-point functions. Moreover, unrelated to the lattice regularization, there are theoretical uncertainties intrinsic to the quasi-distribution approach whose impact should also be assessed [2]. For an extensive review of these different uncertainties, we refer to refs. [186, 188], while a discussion of the systematic effects in the ETMC quasi-PDFs computation can be found in ref. [202]. The latter contains the analysis of the effects investigated so far and a discussion of directions that need to be pursued to fully quantify all the relevant systematics.

Here, we briefly summarize the conclusions reached up to the present stage. It is important to emphasize that while the impact of some systematics is already known to a reasonable degree, reliable estimates of certain types of effects are still largely unknown. Nevertheless, rough assessments can be made even in the case of missing lattice data, by looking at the behaviour of related observables such as the average quark momentum fraction or nucleon charges that have a long history of evaluations on the lattice [240–244]. This allows us to build scenarios describing the potential impact of the systematics on the matrix elements of quasi-PDFs. We consider three scenarios where the systematic effect is a given percentage of the central value of the matrix element and three further ones where it is a given additive shift. We always exclude from the analysis the imaginary part of the matrix element at $z = 0$, equal to 0 by antisymmetry with respect to the sign change of z.

Cut-off effects. One of the most obvious systematic effects in lattice computations comes from the finite value of the lattice spacing, a, *i.e.* the ultraviolet cut-off imposed for the regularization of the theory. While a proper investigation of this uncertainty requires explicit simulations at a few values of the lattice

[2]Note that theoretical uncertainties can be included in global fits of PDFs as detailed in chapter 5

spacing, which are still missing for quasi-PDFs, we may assume that discretization effects are not excessive. This expectation is based on two indirect, but related premises. First, one of the manifestations of large cut-off effects is the violation of the continuum relativistic dispersion relation, which is, however, not observed in the lattice data in ref. [202]. Second, the first moment of the unpolarized $u-d$ PDF gives the quark momentum fraction $\langle x \rangle_{u-d}$. This quantity was intensively investigated on the lattice and we may take the typical size of discretization effects found in such studies. Looking at a summary plot including data from different lattice groups, such as fig. 12 from ref. [241], we see that cut-off effects at lattice spacings comparable to the one of the present work are typically at the 5-15% level in a fixed lattice setup (same discretization, pion mass, volume etc.). Thus, we investigate 6 plausible choices for the magnitude of cut-off effects: 10%, 20%, 30% of the matrix element and additive effects of 0.1, 0.2 and 0.3.

Finite volume effects (FVE). Another natural source of uncertainty in all lattice simulations is the finite size of the box, L, which acts as an infrared regularization. Similarly to discretization effects, a robust investigation of these effects necessitates running the computations for a few values of the lattice size. However, the difference with respect to the lattice spacing effects, typically linear in a or a^2 in the asymptotic scaling regime, is that FVE are usually suppressed as $\exp(-m_\pi L)$, where m_π is the pion mass. This leads to typically $\mathcal{O}(1-5\%)$ effects in hadron structure observables if $m_\pi L \geq 3$. For the matrix elements used in this work, $m_\pi L \approx 3$ – thus, the reasonable assumption about the size of FVE is approx. 5%. In addition to these "standard" FVE of lattice computations, it has been recently pointed out that the usage of a spatially extended operator, including a Wilson line, may lead to additional FVE [245]. The analysis of ref. [245] pertains to a toy scalar theory and predicts a FVE of the form $\exp(-M(L-z))$ (possibly with a polynomial amplifying prefactor), with M being the analogue of the mass of the investigated hadron in the quasi-PDF approach. Given that the nucleon mass is at the physical point around 7 times larger than the pion mass, that would lead to totally irrelevant effects, since the maximum considered z is more than 3 times smaller than L. However, it can not be excluded that in QCD, the form of this FVE can be more severe, e.g. $\exp(-m_\pi(L-z))$. With the physical m_π and $z_{\max} \approx L/3$, this could lead to the amplification of FVE from $\mathcal{O}(5\%)$ to even above 10% at large z. We remark that ETMC has investigated FVE in the renormalization functions for the matrix elements and found no sign of excessive FVE coming at large z (total FVE not larger than around 3%) [202]. We investigate 3 scenarios for fixed percentage effects: constant FVE of 2.5% and

5%, as well as z-dependent ones of the form $\exp(-3 + 0.062z/a)\%$, where 0.062 is the pion mass value for the present ensemble, expressed in lattice units. In addition, we consider 3 shifts: 0.025, 0.05 and $\exp(-3 + 0.062z/a)$.

Excited states contamination. One of the key uncertainties in nucleon structure calculations is whether the ground state hadron state is isolated. If the temporal separation between the interpolating operators creating the nucleon and annihilating it is too small, uncontrolled excited states contamination may appear, leading to a bias in the results. In the context of quasi-PDFs, an important aspect is that this contamination strongly depends on the boost, causing a delicate interplay between the need of large momentum, required for robust matching to light-cone distributions, and excited states contamination, larger for high boost. Thus, a careful analysis is needed to ensure ground state dominance. Such an analysis was performed for the matrix elements used in this work [202]. The conclusion that we use for the present case is that these matrix elements are safe against excited states effects at the level of their statistical precision. In this way, we choose three values of uncertainty from excited states: 5%, 10% and 15%. When the renormalized matrix elements are close to zero, the relative uncertainty can be larger and thus, we consider also three additive scenarios with magnitude 0.05, 0.1 and 0.15.

Truncation effects. The perturbative ingredients of the quasi-PDF approach are of two kinds. One of them is related to the fact that the lattice approach works in integer dimensions and thus, dimensional regularization of the $\overline{\text{MS}}$ scheme is impossible. Instead, as discussed above, a non-perturbative renormalization programme has been proposed by ETMC [164], utilizing a variant of the regularization-independent momentum subtraction scheme (RI'-MOM). The renormalization correlators obtianed in this way can then be translated perturbatively to the $\overline{\text{MS}}$ scheme and finally to the M$\overline{\text{MS}}$ scheme, using formulae derived in refs. [163, 202]. These formulae are currently available to the one-loop level and thus subject to a truncation effect from unknown higher orders. A manifestation of this effect is the fact that the Z-factors have a non-vanishing imaginary part even after conversion to $\overline{\text{MS}}$, where they should be purely real. To evaluate the impact of this uncertainty, we compare the renormalized matrix elements with the ones obtained from applying only the real parts of the Z-factors. We find that the matrix elements obtained by this alternative procedure are compatible with the actual ones within statistical uncertainties, with relatively larger effects observed for small z/a in the imaginary part (up to $\mathcal{O}(5\%)$) and

intermediate z/a in the real part (the real part is small there – thus, the observed absolute effects of around 0.2 can be a large percentage of the value). Apart from the scheme conversion truncation effects, the necessary perturbative ingredient of the approach is the matching between quasi-PDFs and light-cone distributions, also known to one loop [172, 176, 202]. We observe that comparing the quasi-distribution and the resulting light-cone PDF, the numerical magnitude of the matching factor can be significant and thus, the higher order effects may be sizable. The "natural" size of such truncation effects is of $\mathcal{O}(\alpha_s^2)$, which amounts to around 10% at the renormalization scale we consider. However, they are rather uncertainties of the procedure, so they can not be translated to uncertainties of the matrix elements. These uncertainties are the analogue of the theoretical uncertainties that come from a truncated perturbative expansion in the description of observables in phenomenological fits of the PDFs, which have been described in chapter 5. Finally, we consider 6 scenarios for truncation effects pertinent to matrix elements (*i.e.* originating from the perturbative uncertainty in Z-factors): 10%, 20%, 30% of the central value of the matrix element, as well as shifts of 0.1, 0.2 and 0.3.

Higher twist effects. For the current analysis we decide to ignore the effect of higher twists, *i.e.* the presence of power-like corrections to the factorization formula. At this preliminary stage, we are not concerned by their effects, but a more precise phenomenological analysis should definitely take those into account. In particular, it should be kept in mind that, as argued in refs. [176], power corrections can be enhanced for both $x \to 0$ and $x \to 1$, limiting the quasi-PDFs approach in the small-x and large-x regime. We will come back to this point in the conclusions.

Other effects. Apart from the systematics mentioned above, there are some other effects that potentially affect the results. One of them is the usage of a setup including two degenerate light quarks. However, this effect, working in the isospin limit instead of taking into account the different masses and electric charges of the light quarks, is expected to be much below the level of the current precision – of the order of the proton-neutron mass splitting, *i.e.* at the per mille level. A similar magnitude can be expected for the contribution of the neglected sea quark loops from heavier quarks. Such effects can at present be safely ignored and will become important only when aiming at an $\mathcal{O}(1\%)$ precision or better.

Final scenarios. In the end, we define 6 scenarios of possible impact of systematic effects, summarized in tab. 8.1. Scenarios S1-S3 include uncertainties

Scenario	Cut-off	FVE	Excited states	Truncation
S1	10%	2.5%	5%	10%
S2	20%	5%	10%	20%
S3	30%	$e^{-3+0.062z/a}\%$	15%	30%
S4	0.1	0.025	0.05	0.1
S5	0.2	0.05	0.1	0.2
S6	0.3	$e^{-3+0.062z/a}$	0.15	0.3

Table 8.1 *Scenarios of the impact of different systematic effects in the renormalized matrix elements of quasi-PDFs. Percentage values for scenarios S1-S3 should be understood as a given fraction of the central value of the matrix element, while absolute values for S4-S6 are shifts independent from the matrix element.*

that are a fixed percentage of the central value of the matrix element, while for S4-S6, the uncertainties are additive shifts independent from the value of the matrix element. Scenarios S1, S4 can be considered as the most "optimistic" ones. More realistic estimates of uncertainties are included in S2 and S5. Finally, S3 and S6 are "pessimistic", *i.e.* assume largest plausible estimates of the various systematic effects.

8.2.3 From parton distributions to lattice observables

In this chapter, we aim at fitting the data presented in the previous subsection; further studies with different data and treatments of systematic errors are presented in the next chapter. Hence, we specialize our discussion to the case of the unpolarized isovector parton distribution. Following the notations of sec. 2.3.1, the parton distribution f_3 is defined as

$$f_3\left(x,\mu^2\right) = \begin{cases} u\left(x,\mu^2\right) - d\left(x,\mu^2\right) , & \text{if} \quad x > 0 \\ -\bar{u}\left(-x,\mu^2\right) + \bar{d}\left(-x,\mu^2\right) , & \text{if} \quad x < 0 \end{cases} \tag{8.9}$$

where the support is given by $x \in [-1,1]$. The factorization theorem in eq. (8.8) becomes

$$\tilde{f}_3\left(x,\mu,P_z\right) = \int_{-1}^{+1} \frac{dy}{|y|} C_3\left(\frac{x}{y}, \frac{\mu}{|y|P_z}\right) f_3\left(y,\mu^2\right) , \tag{8.10}$$

where the quasi-PDF is the one given by $\Gamma = \gamma^0$ and the explicit expression of the matching coefficients is given in app. F. Starting from the definition of quasi-PDFs given in eq. (8.6), it is clear that the lattice ME is given by the inverse Fourier transform of eq. (8.10), which yields an equation relating the light-cone PDFs on the right hand side to the lattice observable:

$$\mathcal{M}_3\left(zP_z, z^2, \mu\right) = \int_{-\infty}^{\infty} dx\, e^{-i(xP_z)z} \int_{-1}^{+1} \frac{dy}{|y|} C_3\left(\frac{x}{y}, \frac{\mu}{|y|P_z}\right) f_3\left(y, \mu^2\right). \tag{8.11}$$

Since C_3 is purely real, we can split the above complex identity into two real equations, relating the real and imaginary part of the ME $\mathcal{M}_3(z)$ to the light-cone distribution f_3. For the purpose of this work, we introduce two lattice observables, denoted by $\mathcal{O}^{\mathrm{Re}}_{\gamma^0}(z,\mu)$ and $\mathcal{O}^{\mathrm{Im}}_{\gamma^0}(z,\mu)$, defined as

$$\begin{aligned}
\mathcal{O}^{\mathrm{Re}}_{\gamma^0}(z,\mu) &\equiv \mathrm{Re}\left[\mathcal{M}_3\left(zP_z, z^2, \mu^2\right)\right] \\
&= \int_{-\infty}^{\infty} dx\, \cos(xP_z z) \int_{-1}^{+1} \frac{dy}{|y|} C_3\left(\frac{x}{y}, \frac{\mu}{|y|P_z}\right) f_3\left(y, \mu^2\right), \qquad (8.12) \\
\mathcal{O}^{\mathrm{Im}}_{\gamma^0}(z,\mu) &\equiv \mathrm{Im}\left[\mathcal{M}_3\left(zP_z, z^2, \mu^2\right)\right] \\
&= -\int_{-\infty}^{\infty} dx\, \sin(xP_z z) \int_{-1}^{+1} \frac{dy}{|y|} C_3\left(\frac{x}{y}, \frac{\mu}{|y|P_z}\right) f_3\left(y, \mu^2\right), \qquad (8.13)
\end{aligned}$$

where we have only included z and μ in the arguments of $\mathcal{O}^{\mathrm{Re}}_{\gamma^0}$ and $\mathcal{O}^{\mathrm{Im}}_{\gamma^0}$ in order to simplify the notation – since we are working here with only one value of P_z there is little advantage in keeping all the arguments. The explicit expression of C_3 contains plus distributions. Making them explicit we can write the equations above as

$$\mathcal{O}^{\mathrm{Re}}_{\gamma^0}(z,\mu) = \int_0^1 dx\, \mathcal{C}^{\mathrm{Re}}_3\left(x, z, \frac{\mu}{P_z}\right) V_3(x,\mu) = \mathcal{C}^{\mathrm{Re}}_3\left(z, \frac{\mu}{P_z}\right) \circledast V_3\left(\mu^2\right), \tag{8.14}$$

$$\mathcal{O}^{\mathrm{Im}}_{\gamma^0}(z,\mu) = \int_0^1 dx\, \mathcal{C}^{\mathrm{Im}}_3\left(x, z, \frac{\mu}{P_z}\right) T_3(x,\mu) = \mathcal{C}^{\mathrm{Im}}_3\left(z, \frac{\mu}{P_z}\right) \circledast T_3\left(\mu^2\right) \tag{8.15}$$

where V_3 and T_3 are the nonsinglet distributions defined by

$$V_3(x) = u(x) - \bar{u}(x) - \left[d(x) - \bar{d}(x)\right], \tag{8.16}$$

$$T_3(x) = u(x) + \bar{u}(x) - \left[d(x) + \bar{d}(x)\right], \tag{8.17}$$

where, for simplicity, the μ dependence has been omitted. The equations above relate the position space matrix elements computable on the lattice with the collinear PDFs. Similar expressions were worked out in ref. [246] in the context of

the pion distribution amplitude. The proof of eqs. (8.14), (8.15) does require some care, and it is fully worked out in app. F. The coefficients $\mathcal{C}_3^{\mathrm{Re,Im}}$ are related to the real and imaginary part of the Fourier transform of the matching coefficient C_3 appearing in eq. (8.10). Since the latter is defined in terms of plus distributions, the computation is quite involved, and the explicit expression of $\mathcal{C}_3^{\mathrm{Re,Im}}$ is obtained in app. F, making the action of the plus distributions explicit before taking the Fourier transform. A discussion about the convergence of the integrals involved is also reported there. The above results show how fixed z matrix elements defining the quasi-PDF in position space give direct access to two independent nonsinglet distributions, through the integration of the parton distribution over its full support with a perturbatively computable coefficient. We will denote this operation as $\circledast$.

It is useful at this point to recall the form of the QCD factorization formula for the DIS nonsinglet structure function, given in eq. (2.23). Comparing eqs. (8.14), (8.15) with eq. (2.23), we see that the lattice observables introduced above can be treated on the same footing as experimental data for DIS structure functions, as they are related to the nonsinglet distributions through a convolution with a coefficient that can be computed in perturbation theory. However, the form of such convolution, denoted by $\circledast$, is quite different from the one appearing in the DIS case, denoted by $\otimes$: the former involves a DIS-like convolution first, to go from the PDFs to quasi-PDFs, followed by an integration over the full x-range to go to position space. This suggests that this kind of convolution, if implemented in a PDFs fit, may constrain the output much more than what the standard DIS convolution can do.

8.3 Fit setting and FastKernel implementation

The main point of the discussion in sec. 7.5 is that the lattice equal-time correlators are just another possible observable connected to PDFs through some kind of factorization theorem. From a practical point of view, this means that we can treat the lattice data on exactly the same footing as the experimental ones, allowing a smooth and natural way to introduce them in a parton distributions fit. The results presented in this and in the following chapter have therefore been produced using the `c++` fitting framework of the `NNPDF` collaboration described in chapter 3: lattice data and the corresponding systematics are implemented in the code, just as they were data for DIS structure functions, and the fitting

code can be run using the standard methodology, based on neural network parameterization, Monte Carlo replicas generation, numerical minimization of the χ^2 and cross validation. The χ^2 minimized during the fit is given by eq (3.11), where the explicit form of the covariance matrix is

$$C_{ij} = \sigma_{i,s}^2 \, \delta_{ij} + \sum_k \sigma_{i,k} \sigma_{j,k} \tag{8.18}$$

where i and j run over the lattice points and $\sigma_{i,s}$, $\sigma_{i,k}$ are respectively the total statistical and the set of systematical uncertainties of the i-th lattice point, described in sec. 8.2.2. The covariance matrix enters both the definition of the χ^2 and the generation of Monte Carlo replicas, being therefore important for both the central value of the fit and the final PDFs error. A solid knowledge of the covariance matrix is therefore an essential ingredient to get reliable results. The `NNPDF` methodology has been used to produce PDF sets for many years now, and provides a flexible environment within which it has been possible to fit more than 4000 experimental points, coming from a variety of different high energy processes in different kinematic ranges. Therefore it represents a reliable framework which can be used to study and analyze the available lattice data, to assess how well these are able to constrain the PDFs and to compare lattice results with those coming from standard PDF sets.

In order to get theoretical predictions for the data entering the fit, the parton distributions have to be evolved from the fitting scale up to the observable scale, and then they have to be convoluted with the correct coefficient function. As discussed in sec. 3.1.4, these two steps are performed by mean of the FastKernel tables. We show an example of this also in chapter 4, when describing the implementation of jets data in a global PDFs determination. The same procedure has to be implemented for lattice data as well. As seen in sec. 8.2.3, in this case the integration of the parton distributions over their full support is needed. This makes the form of the convolution $\circledast$ more complicated than the one we usually have for high-energy observables, which makes the general implementation of the FastKernel tables slightly different from the standard case. This has been achieved using a proprietary code and in the following we summarize the main steps followed in the implementation. It is important to emphasise once again that in this analysis, once the FastKernel tables have been generated, the lattice data are treated exactly on the same footing as any other data, viz. the exact same methodology and code are used for fitting experimental and lattice data.

The lattice observables $\mathcal{O}^{\mathrm{Re,Im}}_{\gamma^0}(z,\mu^2)$ are determined at a given renormalization scale μ^2. They can be written in terms of the nonsinglet distributions at a given reference scale μ_0^2, by first evolving the parton distribution up to the scale μ^2, and then convoluting it with the coefficents $\mathcal{C}_3^{\mathrm{Re,Im}}$ defined in eqs. (8.14), (8.15) and worked out in app. F. For the nonsinglet distributions considered in this work, the evolution is given by

$$T_3\left(x,\mu^2\right) = \int_x^1 \frac{dy}{y}\,\mathrm{K}_3^{(+)}\left(\frac{x}{y},\alpha_s,\alpha_s^0\right) T_3\left(y,\mu_0^2\right)\,, \tag{8.19}$$

$$V_3\left(x,\mu^2\right) = \int_x^1 \frac{dy}{y}\,\mathrm{K}_3^{(-)}\left(\frac{x}{y},\alpha_s,\alpha_s^0\right) V_3\left(y,\mu_0^2\right)\,. \tag{8.20}$$

where the kernels $\mathrm{K}^{(\pm)}$ are obtained by solving the DGLAP evolution equations in the nonsinglet sector, as described in sec. 2.3.2. For V_3 and T_3 we have two different nonsinglet evolution kernels, denoted by $\mathrm{K}_3^{(-)}$ and $\mathrm{K}_3^{(+)}$ respectively. Eqs. (8.19) and (8.20) can be rewritten expressing the parton distribution in terms of an interpolation basis [53], for instance for the case of T_3

$$T_3\left(x,\mu_0^2\right) = \sum_\beta T_3\left(x_\beta,\mu_0^2\right)\mathcal{I}^{(\beta)}(x) + \mathcal{O}\left[(x_{\beta+1}-x_\beta)^p\right], \tag{8.21}$$

where p is the lowest order neglected in the interpolation. In other words, the interpolating functions act by picking up the value of the PDF at some point x_β of a predefined x-grid. Substituting in the evolution equation eq. (8.19) we get

$$T_3\left(x_\alpha,\mu^2\right) = \sum_\beta \mathcal{K}^{(+)}_{\alpha\beta}\, T_3\left(x_\beta,\mu_0^2\right)\,. \tag{8.22}$$

with

$$\mathcal{K}^{(+)}_{\alpha\beta} = \int_{x_\alpha}^1 \frac{dy}{y}\,\mathrm{K}^{(+)}\left(\frac{x_\alpha}{y},\alpha_s,\alpha_s^0\right)\mathcal{I}^{(\beta)}(y)\,. \tag{8.23}$$

The interpolation basis used at the initial scale can also be used to interpolate the parton distributions at the scale μ^2 appearing in eqs. (8.14), (8.15). For the imaginary part of the lattice observable we get

$$\mathcal{O}^{\mathrm{Im}}_{\gamma^0}(z,\mu) = \sum_\alpha C^{\mathrm{Im}}_{z\alpha}\, T_3\left(x_\alpha,\mu^2\right)\,, \tag{8.24}$$

with

$$C^{\mathrm{Im}}_{z\alpha} = \int_0^1 dx\, \mathcal{C}^{\mathrm{Im}}_3 \left(x, z, \frac{\mu}{P_z} \right) \mathcal{I}^{(\alpha)}(x) \ . \tag{8.25}$$

Putting together eqs. (8.22) and (8.24) we get

$$\mathcal{O}^{\mathrm{Im}}_{\gamma^0}(z,\mu) = \sum_\beta \mathcal{H}^{\mathrm{Im}}_{z\beta}\, T_3\left(x_\beta, \mu_0^2\right) \ , \tag{8.26}$$

where

$$\mathcal{H}^{\mathrm{Im}}_{z\beta} = \sum_\alpha C^{\mathrm{Im}}_{z\alpha}\, \mathcal{K}^{(+)}_{\alpha\beta} \ . \tag{8.27}$$

Eq. (8.27) defines the FastKernel table which enters the computation of the χ^2 during the fit. It connects the parton distribution at the fitting scale to the lattice observable, taking into account the QCD evolution, the matching and the Fourier transform, expressing them through a single matrix vector multiplication. Clearly a similar set of equations defines a FastKernel table that yields the real part of the lattice observable, $\mathcal{O}^{\mathrm{Re}}_{\gamma^0}$, as a function of the valence parton distribution V_3.

8.4 Results

Let us now proceed to presenting and discussing our numerical results. First, we study the way the available lattice data might constrain the parton distributions in a fit, by mean of closure tests: fake data for the real and imaginary part of the ME are generated according to eqs. (8.14), (8.15) using as input a chosen PDFs set. The fitting code is then run over these pseudo-data, using exactly the same setting used in a common fit. By comparing the output of such fits with the known input PDFs sets, we can assess the accuracy we may expect to get from the current knowledge of the lattice data and their systematics.

Then we present results for fits run over the data presented in sec. 8.2.1, studying the 6 different scenarios for the treatment of the systematic errors described in sec. 8.2.2 and summarized in tab. 8.1. The results presented here have been produced using the `NNPDF` fitting code [48] and the `ReportEngine` software [130].

8.4.1 Closure tests

As shown in sec. 8.2.3, we can relate PDFs to lattice observables through the matching convolution of eq. (8.10) followed by a Fourier transform. As already pointed out at the end of sec. 8.2.3, the resulting convolution $\circledast$ is quite different from the one entering standard QCD fits. In this section, we assess how much this operation together with the available lattice data from refs. [181, 202] are able to constrain parton distributions in a fit, running some preliminary closure tests. For a detailed description of the closure test procedure, we refer to sec. 4 of ref. [55]. We generate pseudo-data corresponding to the data of ref. [181] using `NNPDF31_nlo_as_0118` as our input PDFs set, and we run the fitting code over them. The outcome of the closure test fit is then used to assess how well the input PDFs can be reconstructed starting from the 16 position space ME points and their uncertainties.

In order to get an idea of the impact on the fit of the statistical and systematic ME errors, we consider three different scenarios: first we generate fake data assuming no systematic uncertainties and a small uncorrelated statistical uncertainty for each point, constant for all of them and of the order of the smallest real one. From the results of this closure test we can estimate the real constraining power of the convolution $\circledast$, assuming an ideal scenario where all the systematics are under control and the statistical error is kept small. Second, we repeat the exercise but using the real statistical uncertainties, to assess how much the real statistics of the current simulations affect the conclusions of the previous case. Finally we look at the effect of the systematics, considering as a specific example the scenario S2 of table 8.1. The three cases are summarized in table 8.2 and the results are shown in figs. 8.2, 8.3, 8.4.

Closure test	Statistics	Systematics
`CT1`	fake	-
`CT2`	real	-
`CT3`	real	S2

Table 8.2 *Closure tests with different choices of the statistical and systematic error. The results for each option above is summarised in the plots below.*

Looking at the results for CT1, fig. 8.2, it is worth stressing that the lattice data entering the fit are just 16 for the real part and 15 for the imaginary part of the

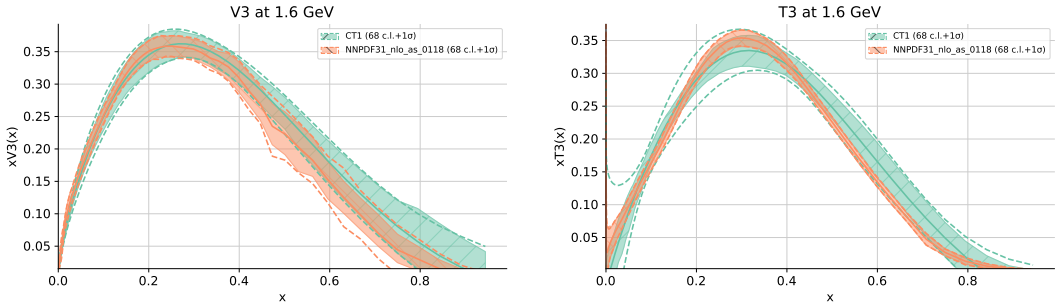

Figure 8.2 *Closure test fit with fixed small statistical error and no systematics (CT1) compared to the input PDFs set. V_3 (top line) and T_3 (lower line) combinations in linear and logarithmic scale are shown. The input PDFs set is fully reconstructed within 1-sigma level, getting PDFs with an error band comparable to the input one.*

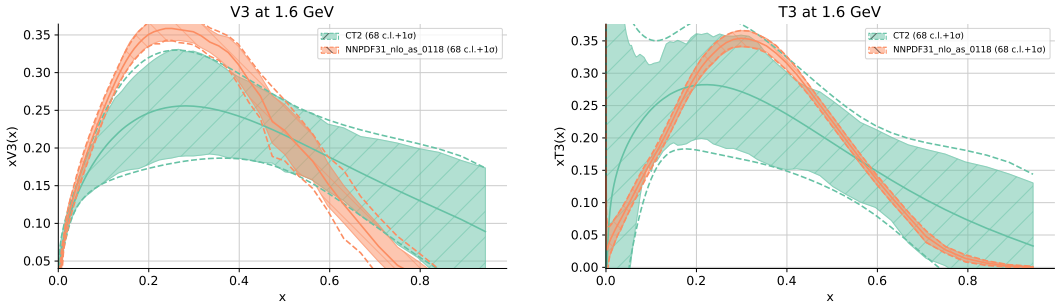

Figure 8.3 *Closure test fit with real statistical error and no systematics (CT2) compared to the input PDFs set. Top line: V_3 combination in linear and logarithmic scale. V_3 (top line) and T_3 (lower line) combinations in linear and logarithmic scale are shown. The error band of the reconstructed set is way bigger than the one of the input PDFs, showing a non negligible impact of the current statistics over the final PDFs error.*

matrix element. Just half of them are actually used in the training procedure, while the other ones are used to build the validation set. In a standard NLO QCD global fit, like the one used as input PDF here, the number of points entering the analysis is $\mathcal{O}(4000)$. fig. 8.2 shows how good the convolution $\circledast$ is in constraining the PDFs, assuming an ideal scenario where all the systematics are under control, and the statistics are kept small. Looking at the results for CT2 and CT3 in figs. 8.3 and 8.4, it is clear how big the impact of the statistical and systematic uncertainties of the ME is on the PDFs error: in both cases the input PDFs set is reconstructed within 1-sigma level, with some tension for V_3 at medium x in the first case. The PDFs error is increasingly big, becoming huge

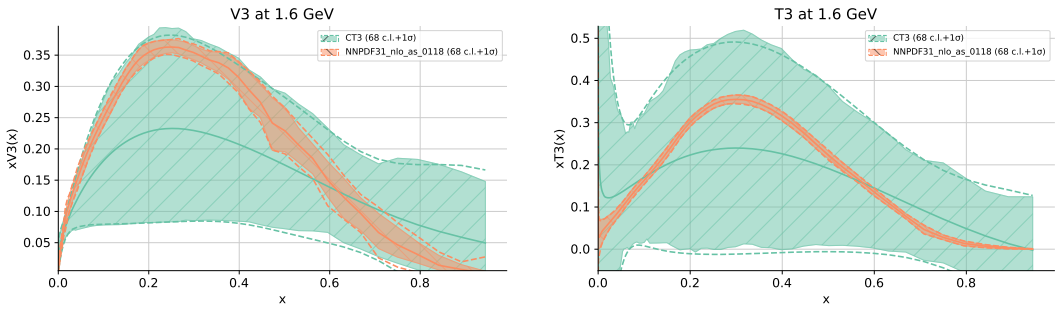

Figure 8.4 *Closure test fit with real statistical and systematic error (CT3) compared to the input PDFs set. V_3 (top line) and T_3 (lower line) combinations in linear and logarithmic scale are shown. The errors of the reconstructed PDFs are huge.*

when the full systematics are considered. Fig. 8.4 shows what we may expect in a real life scenario.

To sum up, the results from CT1 show how promising this kind of lattice data might be in constraining PDFs. On the other hand, the results from CT2 and CT3 highlight the importance of having a good control over both the statistical and systematical uncertainties in the lattice simulations of the ME. It is worth noticing, however, that the overall error band of the reconstructed PDFs, even in presence of the full systematic errors, would surely be reduced when new data are available.

8.4.2 Fit results

In this section, we present our results for fits ran over the data from refs. [181, 202], described in sec. 8.2.1. As mentioned before, we consider 6 different scenarios for the treatment of the systematic errors, summarized in table 8.1. We show results for "optimistic" (S1,S4), "realistic" (S2,S5) and "pessimistic" (S3,S6) scenarios, the difference between the elements of each couple being the nature of the systematic errors: an additive shift given by a percentage of the ME for the first, a constant shift for all the ME points for the second one.

The results of the fit for the two optimistic scenarios are shown in fig. 8.5. S1 is sligthly more conservative than S4, but overall there is not much difference between them. The situation changes for the more realistic scenarios (fig. 8.6), where S2 is much more conservative than S5. In the former case the tension with

NNPDF31_nlo_0118 is smaller than what we observe in the previous scenarios, due to the increase in the error band and to a slight shift of the central replica of the fit. Similar comments can be made for the most pessimistic scenarios, shown in fig. 8.7, having S3 with a huge error band and a more remarkable shift of the central replica towards the one of NNPDF31. Overall, we notice how, when the systematics are given by a percentage of the ME, we get qualitatively different results moving from one scenario to the other, while in the case we consider constant shifts there is no much difference between different cases.

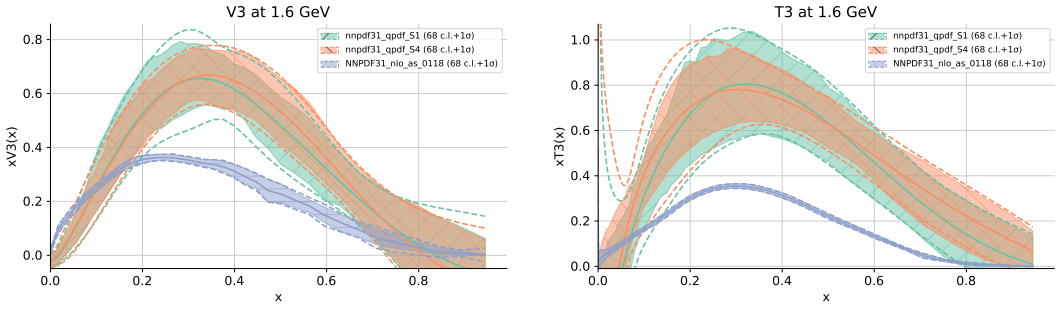

Figure 8.5 *S1 vs. S4: S1 results are sligthly more conservative than the S4 ones, but overall there is no significant difference between the two optimistic scenarios.*

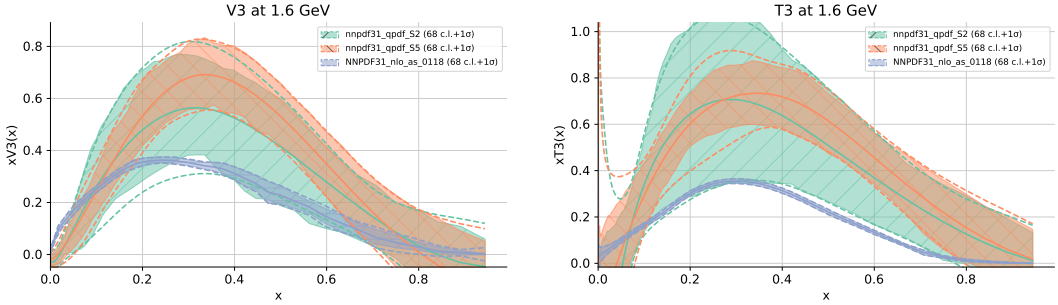

Figure 8.6 *S2 vs. S5: S2 results are more conservative than the S5 ones, showing also a small shift of the replica 0 towards the light-cone PDFs. Overall, S2 results are comptible with* NNPDF31_nlo_0118 *within 1-sigma level.*

To sum up, in this chapter, we have used the momentum space factorization of quasi-PDFs in order to relate the unpolarized isovector parton distribution to well-defined matrix elements computable on the lattice. Using some of the currently available lattice data, we have used such result to extract the nonsinglet distributions V_3 and T_3 within the NNPDF framework, studying also different

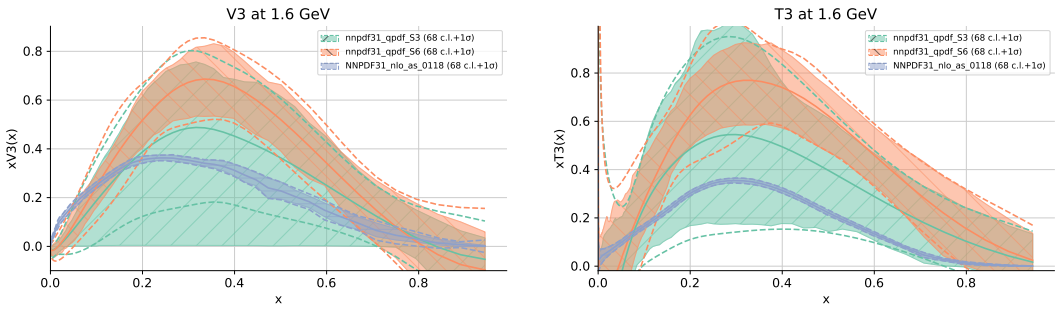

Figure 8.7 *S3 vs. S6: S3 results are extremely conservative, while those for S6 do not show a qualitative difference with respect to S4 and S5.*

possible scenarios for the treatment of the systematic uncertainties from lattice QCD simulations.

Our first results from closure tests show how effective these lattice data might be in constraining PDFs, allowing a consistent determination of the target distribution starting from $\mathcal{O}(15)$ ME points. On the other hand, we show that a consistent treatment of the lattice systematics is extremely important, and how the final result of the fit strongly depends on the specific systematics scenario we consider. Considering the most realistic ones, agreement with the phenomenological PDFs is observed within 1 sigma level, for both the nonsiglet distributions considered here. The error bands are, however, very large with respect to the corresponding phenomenological PDFs, showing again how important the control over the lattice systematics is.

Despite having focused on the quasi-PDFs case, the framework we implemented is general enough to allow for the treatment of different lattice data. In the next chapter we will discuss a similar analysis addressing this time the pseudo-PDFs approach. In this case more data are available and, as we are going to see, position space factorization formulas allows for a number of additional advantages.

9

PDFs from pseudo-Ioffe Time Distribution

In the previous chapter, starting from the momentum space factorization formula connecting quasi-PDFs to collinear PDFs, upon numerical implementation of the Fourier transform we obtained an expression relating parton distributions directly to position space quasi-PDFs matrix elements. Such factorized expression has been subsequently used in the NNPDF framework in order to extract two nonsinglet distributions from data for quasi-PDFs matrix elements produced in refs. [181, 202]. A similar analysis was recently performed by the JAM collaboration in ref. [247] for the spin-averaged and spin-dependent PDFs employing quasi-PDF lattice data. In this chapter, base on ref. [13] we extend such analysis to the case of the pseudo-PDFs approach, which relies on the position space formulation of the same factorization theorem. In particular, we will consider data for reduced pseudo-ITD, introduced in sec. 7.3.1 in the context of the scalar field theory and that we now revise in the context of QCD.

As we saw in sec. 8.1 the UV divergences of the Ioffe-time pseudodistribution (pseudo-ITD) are multiplicatively renormalizable [166, 167]. The relevant renormalization factor $Z(z_3^2)\, e^{\delta m|z|/a}$ does not depend on ν and, for small z_3^2, is known at one loop. Its explicit form is inessential if one introduces the so-called reduced Ioffe-time pseudo-distributions first defined in ref. [160] as

$$\mathfrak{M}\left(\nu, z_3^2\right) = \frac{\mathcal{M}\left(\nu, -z_3^2, \mu^2\right)}{\mathcal{M}\left(0, -z_3^2, \mu^2\right)}. \tag{9.1}$$

The Z-factors of the numerator and denominator are the same and cancel in the ratio leaving the reduced distribution on the left-hand side without any residual dependence on unphysical scales.

Working in the small-z_3^2 limit, the pseudo-ITD can be matched at one-loop level to the corresponding ITD through a finite perturbative kernel, expressing the pseudo-ITD in terms of the collinear PDFs through a factorization formula based on the operator product expansion (OPE). The computation of the relevant QCD diagrams has been performed in a number of independent papers. The original QCD computation is reported, for example, in refs. [165, 176, 197, 237]. The result reads

$$\mathfrak{M}\left(\nu, z_3^2\right) = \int_{-1}^{1} dx\, C\left(x\nu, \mu^2 z_3^2\right) f\left(x, \mu^2\right) + \mathcal{O}\left(z_3^2 \Lambda^2\right) , \tag{9.2}$$

with

$$C\left(\xi, \mu^2 z_3^2\right) = e^{i\xi} - \frac{\alpha_s}{2\pi} C_F \int_0^1 dw \left[\frac{1+w^2}{1-w} \log\left(z_3^2 \mu^2 \frac{e^{2\gamma_E+1}}{4}\right) + 4\frac{\log(1-w)}{1-w} - 2(1-w)\right]_+ e^{i\xi w} + \mathcal{O}\left(\alpha_s^2\right). \tag{9.3}$$

Eqs. (9.2), (9.3) allow to relate collinear PDFs to quantities which are computable in lattice QCD simulations, through a factorized expression similar to those relating collinear PDFs to physical cross sections. Just as described in the previous chapter, this formula can be used in a fitting framework to extract PDFs from lattice data.

This, besides being a complementary exercise to the one presented in chapter 8, has also some practical advantages. First, when working in the pseudo-ITD approach, the factorization is realized in the limit of small-z^2. Unlike in the quasi-PDFs approach, where the factorization is realized for high values of P, here we are allowed to keep in the analysis data coming from a wide range of momentum values, without having to remove those with lower P. This advantage is particularly important, because in lattice QCD, the low momentum data are significantly more precise for a fixed computational cost. Second, we can directly use the position space factorization formula of eq. (9.2), relying on the analytical expression for the perturbative coefficient of eq. (9.3) and without having to perform the numerical Fourier transform described in app. F.

The structure of the chapter is as follows. In sec. 9.1 we define the lattice observable considered in the fit, describe the corresponding data and briefly recall the main features of the NLO terms entering the factorization formulas. In sec. 9.2 we present the first set of results: we consider the fits where only the

statistical uncertainties of the lattice data are taken into account. Analyzing data from different lattice ensembles we show that, in general, without accounting for systematic effects it is not possible to obtain a good fit. In sec. 9.3 we discuss and quantify some of the systematic uncertainties affecting the reduced pseudo-ITD data. We include such systeatics in the analysis and we study their impact on the final PDFs and on the fit quality.

9.1 Lattice data and observables

In this section we describe the lattice observables we will consider in the following, together with the corresponding data. As in the case of the previous chapter we will consider two different observables corresponding to the real and imaginary part of the reduced pseudo-ITD defined in eq. (9.1).

Considering the case of the unpolarized isovector parton distribution, taking the real and complex parts of eq. (9.2) and using eq. (9.3), we can define the two lattice observables

$$\mathrm{Re}\left[\mathfrak{M}\right]\left(\nu,-z_3^2\right)=\int_0^1 dx\, C^{\mathrm{Re}}\left(x\nu,\mu^2 z_3^2\right) V_3\left(x,\mu^2\right)\,, \tag{9.4}$$

$$\mathrm{Im}\left[\mathfrak{M}\right]\left(\nu,-z_3^2\right)=\int_0^1 dx\, C^{\mathrm{Im}}\left(x\nu,\mu^2 z_3^2\right) T_3\left(x,\mu^2\right)\,, \tag{9.5}$$

with

$$\begin{aligned} C^{\mathrm{Re}}\left(\xi,\mu^2 z_3^2\right) &= \cos\left(\xi\right) \\ &\quad -\frac{\alpha_s}{2\pi}C_F\int_0^1 dw\left[B\left(w\right)\log\left(z_3^2\mu^2\frac{e^{2\gamma_E+1}}{4}\right)+L\left(w\right)\right]\cos\left(\xi w\right), \end{aligned} \tag{9.6}$$

$$\begin{aligned} C^{\mathrm{Im}}\left(\xi,\mu^2 z_3^2\right) &= \sin\left(\xi\right) \\ &\quad -\frac{\alpha_s}{2\pi}C_F\int_0^1 dw\left[B\left(w\right)\log\left(z_3^2\mu^2\frac{e^{2\gamma_E+1}}{4}\right)+L\left(w\right)\right]\sin\left(\xi w\right)\,, \end{aligned} \tag{9.7}$$

where the kernels $B\left(w\right)$ and $L\left(w\right)$, according to eq. (9.3), are given by

$$B\left(w\right)=\left[\frac{1+w^2}{1-w}\right]_+\,, \tag{9.8}$$

$$L\left(w\right)=\left[4\frac{\log\left(1-w\right)}{1-w}-2\left(1-w\right)\right]_+\,. \tag{9.9}$$

It is worth recalling some important features of the NLO coefficients given in eqs. (9.6), (9.7). The contributions proportional to the two kernels $B(w)$ and $L(w)$ of eqs. (9.8), (9.9) can be seen as an evolution and a scheme change term respectively [195, 198]: while the former is responsible for the evolution from the PDF scale $\hat{z}^{-2} = \mu^2 \frac{e^{2\gamma_E+1}}{4}$ to the pseudo-ITD scale z^2, the latter takes into account the finite terms characterizing the specific choice of the renormalization scheme. They are plotted in fig. 9.1 for both the real and imaginary part, using the PDFs set `NNPDF31_nlo_as_0118` as input. The evolution term $B(w)$ also connects

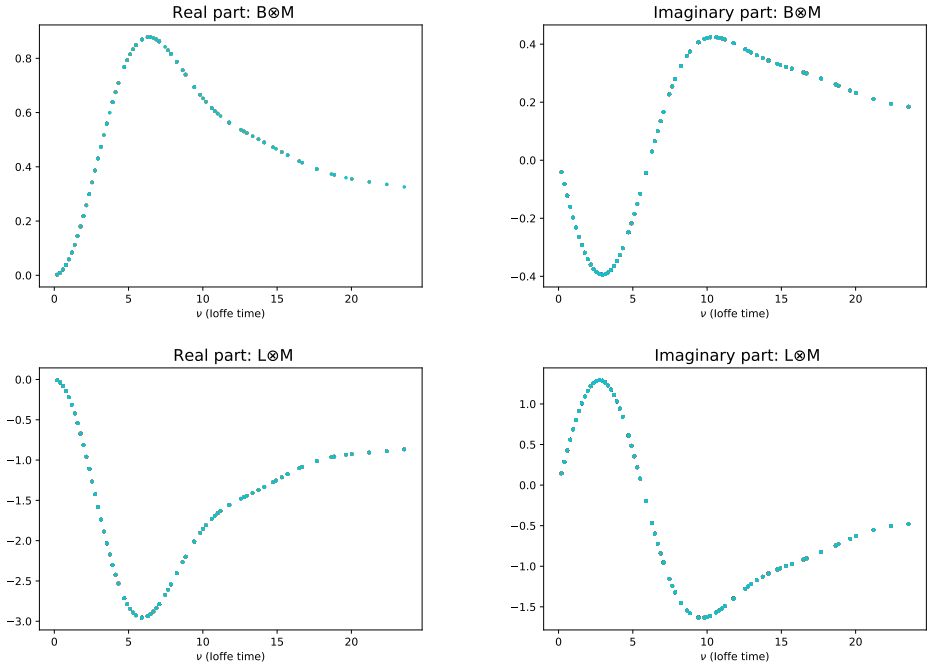

Figure 9.1 *Upper plot: The NLO evolution term for the real (left) and imaginary part (right). Lower plot: The NLO scheme change term for the real (left) and imaginary part (right).*

pseudo-ITD points having different values of z^2: considering for example the real part, from eqs. (9.4), (9.6) it follows

$$\begin{aligned}\mathrm{Re}\left[\mathcal{M}\right]\left(\nu, z_0^2\right) = \mathrm{Re}\left[\mathcal{M}\right]\left(\nu, z^2\right) \\ - C_F \frac{\alpha_s}{2\pi} \log \frac{z_0^2}{z^2} \int_0^1 dx \left[\int_0^1 dw\, B(w) \cos\left(x\nu w\right)\right] V_3\left(x, \mu^2\right) ,\end{aligned} \tag{9.10}$$

which relates the real part of the pseudo-ITD point at the scale z^2 with the one having the same Ioffe time at the scale z_0^2 [196, 248].

We will consider the data for reduced pseudo-ITD from refs. [198, 200]: the

Lattice ensemble	a(fm)	M_π (MeV)	$L^3 \times T$	$n_{\rm dat}$	Reference
fine	0.094(1)	358(3)	$32^3 \times 64$	48	
big	0.127(2)	415(23)	$32^3 \times 96$	48	[198, 200]
coarse	0.127(2)	415(23)	$24^3 \times 64$	36	
280	0.094(1)	278(3)	$32^3 \times 64$	64	[200]
170	0.091(1)	172(6)	$64^3 \times 128$	80	

Table 9.1 *Lattice data details*

datasets presented in ref. [198] have been produced starting from three different lattice ensembles, denoted as *fine*, *big* and *coarse* and which differ for the volume and lattice spacing used in the simulations. They have been produced using values of the pion mass ranging from 358 MeV (*fine*) to 415 MeV (*coarse* and *big*). In the present work we will focus on the datapoints produced from the *fine* ensemble, while those from the *coarse* and *big* ones will be used to estimate systematic effects due to continuum limit and finite lattice volume. We will also consider pseudo-ITD points presented in ref. [200], produced using pion mass equal to 172 MeV. Following the original convention of ref. [200] we will denote the corresponding lattice ensemble as *170*. Points from the ensemble *280*, presented in the same paper and produced using similar lattice spacing and pion mass 278 MeV, will be used to estimate the pion mass effects in the analyses for the ensembles fine and 170. These five ensembles of $2+1$ flavor lattice QCD were generated by the JLab/W&M collaboration using clover Wilson fermions and a tree level tadpole-improved Symanzik gauge action. One iteration of stout smearing with the weight $\rho = 0.125$ for the staples is used in the fermion action. A direct consequence of the stout smearing is that the value of the tadpole corrected tree-level clover coefficient $c_{\rm SW}$ used is very close to the non-perturbative value determined, a posteriori, using the Schrödinger functional method. The detailed features of these ensembles are reported in tab. 9.1, together with the number of reduced pseudo-ITD datapoints $n_{\rm dat}$ computed from each of them.

Given a set of lattice data for the real and imaginary part of the reduced pseudo-ITD, the distributions T_3 and V_3 can again be extracted from them through a standard minimum-χ^2 fit. Here we will implement them in the `NNPDF` framework following the same approach as the one described in chapter 8.

9.2 Fits over lattice data: statistical uncertainties only

In this section, we will present results for fits performed over the lattice data computed from the ensembles *fine* and *170*, denoted as *fine-stat* and *170-stat* respectively. Such fits have been produced considering statistical uncertainties only. We will show how, in general, without having the complete information regarding the lattice systematic uncertainties it is not always possible to obtain a good fit. In the next section, taking as example the case of the *fine* ensemble, we will discuss and estimate some of the possible systematic effects, studying their impact on the fit quality and on the resulting PDFs.

Parton distributions resulting from fits *fine-stat* and *170-stat*, together with the corresponding error bands, are plotted in the upper and lower plots of fig. 9.2, and the χ^2 values are reported in tab. 9.2: despite the PDFs extracted from the two datasets are compatible within one σ, the error band of the fit *fine-stat* appears to be slightly smaller than the other, with an average χ^2 value per datapoint equal to 8.36, pointing out a possible underestimation of the error and a bad fit quality. This could be caused by inconsistencies between different datapoints, due to unknown systematic uncertainties affecting them. On the other hand, the fit *170-stat* shows better χ^2 values, with an average value per datapoint equal to 1.38.

Focusing on the more problematic case of the fine ensemble results, in order to assess which points are more likely to be affected by large systematic errors, we will study the contribution to the χ^2 coming from each datapoint

$$\delta_i = \frac{(D_i - T_i)^2}{\sigma_i^2}\,, \tag{9.11}$$

D_i and T_i being the i-th lattice point and the corresponding prediction from the fit respectively, and find out which points D_i are more than 4σ (or 3σ) off from the fitted distribution T_i. These are the points that, most likely, do not belong to the fitted distribution and which therefore might be affected by larger systematic effects.

The contributions $\sqrt{\delta_i}$ are plotted in the upper plot of fig. 9.3 as a function of the Ioffe-time ν, with the red and yellow lines highlighting the 4σ and 3σ cut

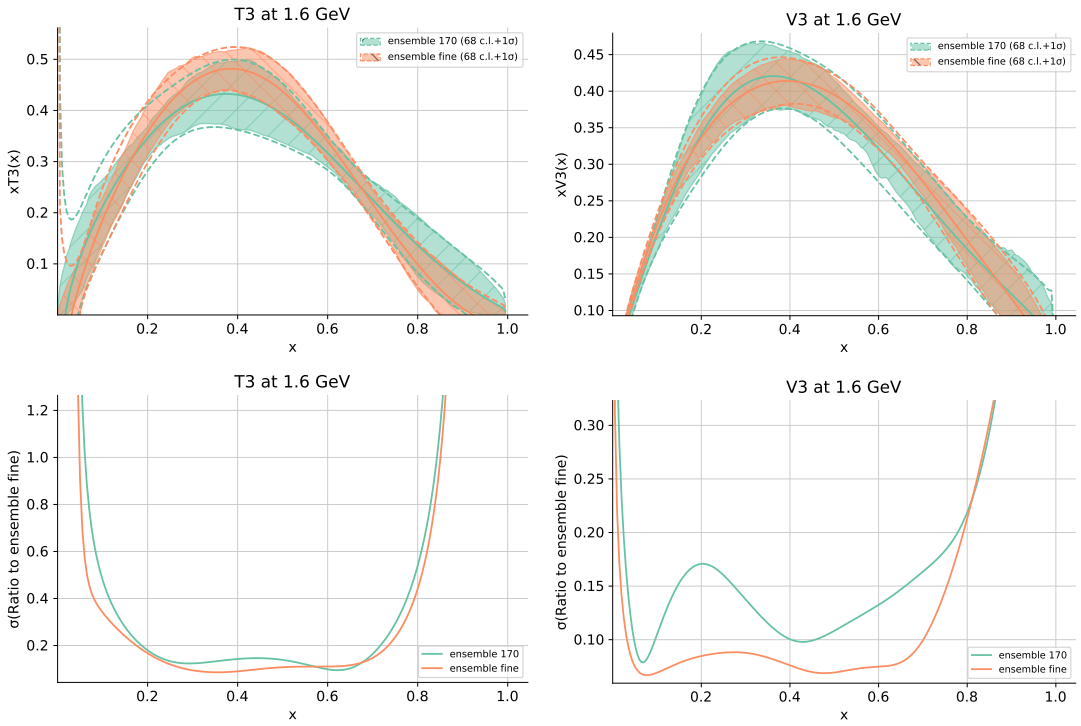

Figure 9.2 *Upper plots: PDFs from datapoints computed from the ensembles fine and 170. The shaded bands represent the PDFs error computed as the 68 c.l. of the fit replicas, while the dashed line is obtained by computing the standard deviation point by point in x. Lower plots: corresponding PDFs errors, computed as standard deviation over fit replicas and displayed in function of x.*

respectively: it is clear that a bunch of points having small Ioffe-time values are those giving the highest contribution to the total χ^2, being more than 3σ or 4σ off. We can implement 4σ and 3σ cuts, removing the problematic points from the dataset and producing new fits, denoted as *fine-stat-*3σ and *fine-stat-*4σ: the new fits show more reasonable χ^2 values, reported in tab. 9.2, showing how, upon removing the outliers, the remaining points, coming from a wide range of momentum p and Euclidean separation z_3, are fitted reasonably well. The PDFs resulting from the 3σ cut are plotted in the lower plot of fig. 9.3, normalized to the fit without any cuts: it is clear how, despite spoiling the total χ^2, the problematic points do not seem to have a big impact on the final PDFs.

We conclude that, depending on the specific lattice ensemble we consider, quite a high number of small Ioffe-time points do not belong to the fitted distribution. In order to get reasonable χ^2 values, such points have to be removed from the fit. This highlights possible tensions between datapoints and may point out the presence of systematic effects. In order to avoid any underestimation of the PDFs

Ensemble	fit	Obs	n_{dat}	χ^2	χ^2_{tot}
fine	fine-stat	Re $[\mathfrak{M}]$	48	7.94	8.36
		Im $[\mathfrak{M}]$	48	8.77	
	fine-stat-4σ	Re $[\mathfrak{M}]$	39	2.68	3.28
		Im $[\mathfrak{M}]$	39	3.89	
	fine-stat-3σ	Re $[\mathfrak{M}]$	34	1.45	1.86
		Im $[\mathfrak{M}]$	32	2.27	
170	170-stat	Re $[\mathfrak{M}]$	80	0.68	1.38
		Im $[\mathfrak{M}]$	80	2.07	

Table 9.2 *Details of fits with statistical uncertainties only. From left to right we report the lattice ensemble, the fit name, the observables included in the analysis, the number of datapoints and finally the partial and total χ^2.*

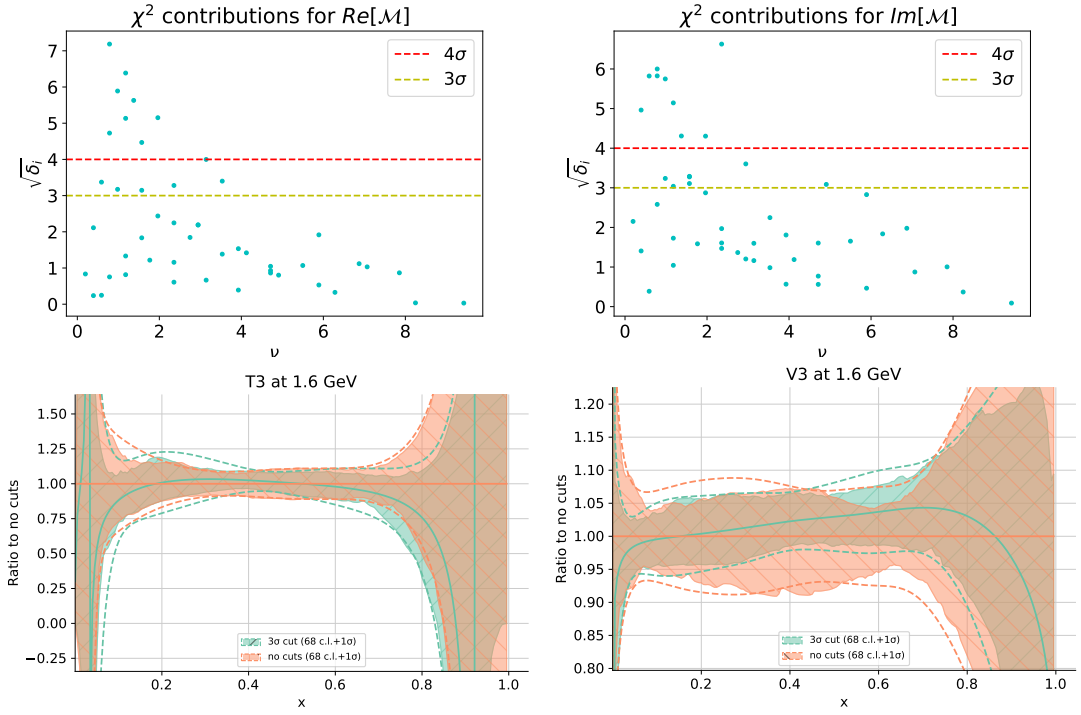

Figure 9.3 *Upper plots: $\sqrt{\delta_i}$ contributions for each datapoint of the fine ensemble. The red and yellow lines highlight the 4σ and 3σ cut respectively. Lower plots: PDFs from fits fine-stat (orange) and fine-stat-3σ (green), normalized to the former.*

error and to introduce back in the analysis all the available points, systematic uncertainties need to be quantified and implemented in the fit.

9.3 Systematic effects

9.3.1 Discussion

The high χ^2 values of the fits presented in the previous section might point out the presence of some tensions between datapoints. In the following, focusing on the case of the fine ensemble results, we will show that this is indeed the case, and we will investigate possible sources of systematic uncertainties and their numerical values.

The matrix element defining the pseudo-ITD is a function of the Ioffe-time ν and of the scale z^2. Points having the same Ioffe-time but different Euclidean separation can be related through eq. (9.10), which can be used to evolve each pseudo-ITD point up to a chosen reference scale $z_0^2 = (0.7\,a)^2$. Looking at fig. 9.1 it is clear that, given this choice for z_0, the sign of the NLO correction of eq. (9.10) will be positive for every datapoint, so that the evolution increases the real part of the pseudo-ITD. Considering the imaginary part, the sign of the NLO evolution term is initially negative, and it turns positive at bigger values of ν. Such effects can be seen in fig. 9.4, where the pseudo-ITD points computed from the fine ensemble are plotted before (blue) and after evolution (red). After evolution,

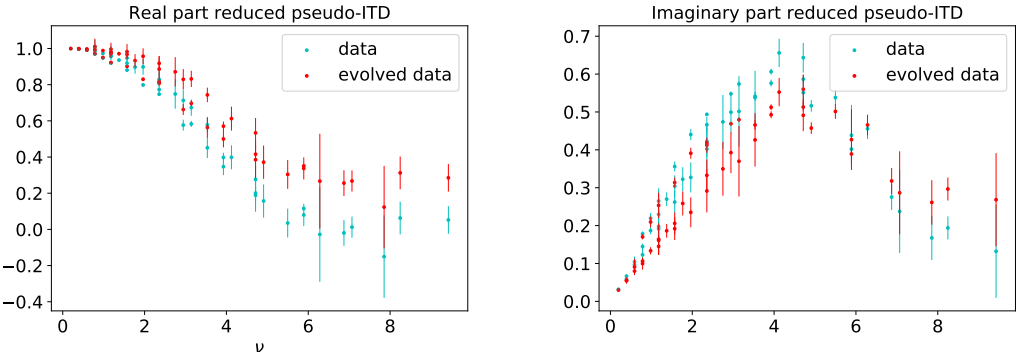

Figure 9.4 *Data for the real part of the pseudo-ITD at their original scale z^2 and evolved at the common scale z_0^2.*

points having the same Ioffe time should have the same value. In practice, they should be compatible within errors. Looking at the red points of fig. 9.4, where each point is plotted with the corresponding statistical uncertainty, it is clear how, expecially in the small Ioffe time region, this is not always the case: after evolution, some points having the same Ioffe time are not compatible between each other. Such discrepancies might be explained by the presence of systematic

effects we are not accounting for.

As already mentioned in the previous chapter, a proper investigation of the systematic effects affecting the computation of the equal time correlators underlying the definition of pseudo-PDFs is a difficult and expensive task which would require to run different lattice simulations varying a set of parameters, like for example the lattice spacing, the lattice volume, the pion mass. Alongside systematic effects due to the lattice simulation, other sources of errors are those connected to the theoretical framework of the pseudo-PDFs approach, like the presence of higher twist effects and perturbative matching truncation effects, as detailed in sec. 8.2.2.

As mentioned in sec. 9.1, in ref. [198] additional pseudo-ITD points were computed starting from other two lattice ensembles, with pion mass similar to that of the fine one, but having different volume and lattice spacing, denoted as *big* and *coarse*, whose features are reported in tab. 9.1. Systematic uncertainties due to the continuum limit (CL) and finite volume (FV) can be directly estimated using these additional results as detailed in ref. [198]: the real and imaginary components of the pseudo-ITD are fitted to a polynomial as a function of the Ioffe-time ν; the difference between coarse and fine ensemble results is taken as an estimate for lattice spacing effects as a function of ν, while the analogous difference considering the coarse and big ensembles gives an estimate for uncertainties due to finite lattice volume. Systematic effects due to the pion mass (PM) can be estimated in a similar way: as mentioned in sec. 9.1, in ref. [200] the data of the ensembles fine and 170 have been supplemented with additional pseudo-ITD results produced from a third ensemble having pion mass equal to 278 MeV, denoted as ensemble *280*. The difference between polynomial fits for the ensembles fine and 280 is taken as an estimate for pion mass effects. These differences will be considered as three independent sources of correlated systematic, affecting each datapoint entering the analysis. They are shown in the upper plots of fig. 9.5 as functions of the Ioffe-time, denoted as FV (finite volume), CL (continuum limit) and PM (pion mass).

It is important to understand whether or not these systematic uncertainties are enough to account for the discrepancies described at the beginning of the section. In the lower plots of fig. 9.5 FV, CL and PM systematic effects are plotted for the relevant Ioffe-time values, together with the aforementioned discrepancies. Consistently with what observed previously, the latter seem to affect mostly low Ioffe-time points, which are also those for which the estimated systematics reach

their minimum values. Therefore from fig. 9.5 it follows that FV, CL and PM systematics cannot be considered responsible for the big contributions to the χ^2 noted in the fits of the previous section. In other words, they are likely not enough to account for the observed discrepancies affecting low Ioffe-time points. It should be noted that a study of more than 2 ensembles for each systematic error may be necessary for a more definitive conclusion.

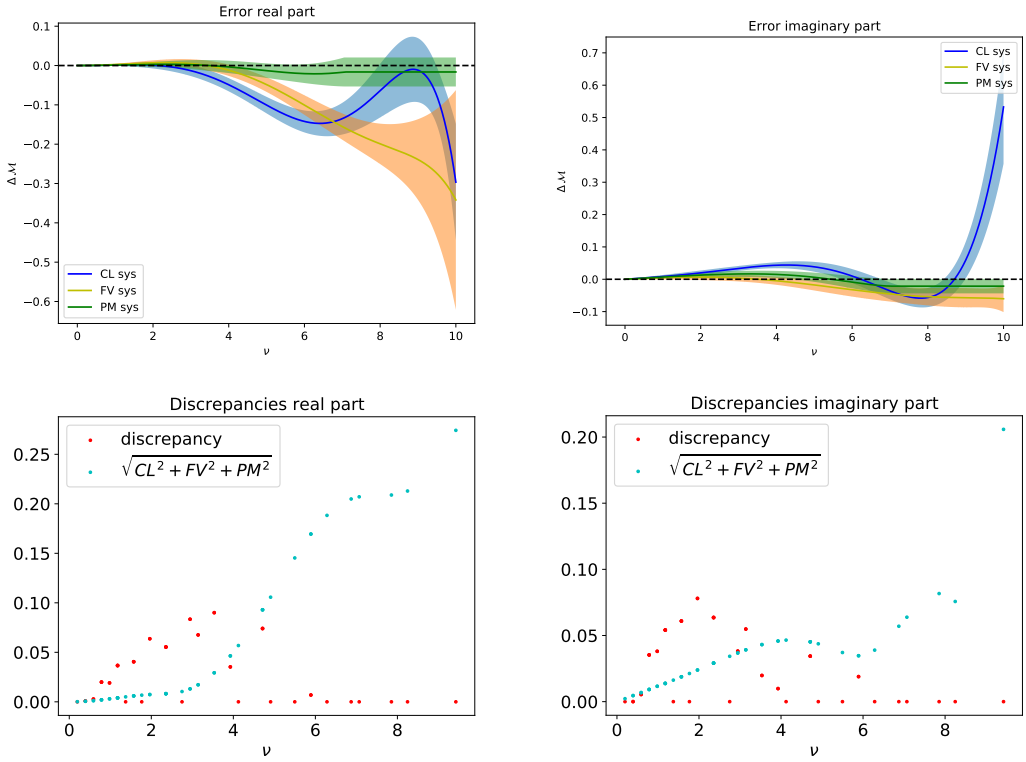

Figure 9.5 *Upper plots: finite volume (FV), continuum limit (CL) and pion mass (PM) systematics provided as functions of the ioffe-time ν for the real (left) and imaginary (right) part of the matrix element. Lower plots: discrepancies between data having the same ioffe-time (red) together with the total FV, CL and PM systematic effects (blue).*

Excited states contaminations might represent another possible source of systematic effects. Also missing higher orders in perturbation theory and higher twist effects could in principle be treated as additional systematic uncertainties. Unlike the case of the FV, CL and PM systematic uncertainties discussed above, we cannot estimate the size of such effects using the current lattice results. One could then follow the approach adopted in the previous chapter, where different scenarios for the size of such systematics have been considered, and try to draw conclusions about their impact on the PDFs and on the fit quality. Here we will follow a different approach, trying to quantify an additional uncertainty

which accounts for the unknown missing systematic effects, following a Bayesian approach. The gaussian Bayesian approach that we are going to use here is the same described in chapter 5 in the case of global QCD analysis in order to take into account the theoretical error due to missing higher orders[1].

As discussed in sec. 5.1, the figure of merit which is minimized during a Gaussian fit is defined as the probability of the data D given the model parameters θ, namely the likelihood

$$\mathcal{P}\left(D|\theta\right) = e^{-\frac{1}{2}(D-T(\theta))^T\,\Sigma^{-1}\,(D-T(\theta))}. \tag{9.12}$$

where Σ is the covariance matrix of the data D, accounting for the known statistical and systematic uncertainties, and $T\left(\theta\right)$ is the theoretical prediction, function of the model parameters. If we assume the presence of unknown systematic effect Δ affecting the datapoints D, eq. (9.12) can be modified as

$$\mathcal{P}\left(D,\Delta|\theta\right) = e^{-\frac{1}{2}(D+\Delta-T(\theta))^T\,\Sigma^{-1}\,(D+\Delta-T(\theta))}. \tag{9.13}$$

Assuming a Gaussian prior distribution $\mathcal{P}\left(\Delta\right) = \exp\left[-\frac{1}{2}\Delta^T\,\hat{\Sigma}^{-1}\,\Delta\right]$ we can marginalize over Δ getting

$$\int d\Delta\,\mathcal{P}\left(\Delta\right)\mathcal{P}\left(D,\Delta|\theta\right) \propto e^{-\frac{1}{2}(D-T(\theta))\left(\Sigma+\hat{\Sigma}\right)^{-1}(D-T(\theta))}, \tag{9.14}$$

which defines the relevant likelihood to be minimized. Eq. (9.14) shows how the presence of unknown systematic effects can be accounted for by introducing in the likelihood an additional contribution to the covariance matrix, denoted by $\hat{\Sigma}$, which defines the prior probability distribution of these systematics. Its specific definition is of course arbitrary, and depends on the knowledge of the missing uncertainties we have. This Bayesian approach, despite not providing a general method to estimate the missing systematics, allows to include in the analysis the partial information we may have about them.

In the present case, we only know the discrepancies observed at the beginning of this section, not described by continuum limit, finite volume and pion mass effects. We can look at such discrepancies as an indication of the minimal size of the systematic effects affecting the data and use them to construct a suitable $\hat{\Sigma}$: for each couple of points having a given Ioffe-time value, we will define the two corresponding diagonal components of $\hat{\Sigma}$ as half of the distance between

[1] In ref. [249] a similar approach was applied to cosmological data.

evolved points, setting the off diagonal elements to zero. Each point sharing the same Ioffe time value with at least another one will therefore be affected by an additional, uncorrelated systematic such that, after evolution, datapoints having the same Ioffe-time will be compatible between each other. Clearly, this global, uncorrelated systematic will be the dominant one for small Ioffe-time points, where most of the problematic points are, while for higher value of ν lattice spacing, finite volume and mass effects will dominate.

9.3.2 Results

To sum up, in sec. 9.3.1, we have discussed and estimated four different source of systematics: the first three, accounting for finite volume, lattice spacing and pion mass effects, can be computed directly from the available lattice results as a function of the Ioffe-time ν, and will be implemented in the fit as three independent sources of correlated systematics; the fourth one has been estimated using the size of the discrepancies observed between points having the same Ioffe-time, and will be considered as an additional uncorrelated uncertainty, in order to take into account the minimal size of all the remaining systematic effects we have not directly computed. As mentioned in sec. 9.1, such systematics enter the definition of the covariance matrix responsible for both replicas generation and the χ^2 definition, and therefore it has a central role in both the determination of the fit central value and its error band. The new fit is denoted as *fine-sys* and the resulting PDFs are plotted in fig. 9.6, together with the results from the fit *fine-stat* presented sec. 9.2: the distribution T_3 is only marginally affected by the introduction of the systematic errors, showing a mild down shift of its central value in the medium and large x regions; on the other hand both the central value and the error band of V_3 change, with an overall down shift of the former and a sizable increase of the latter. The χ^2 values are reported in tab. 9.3: the average value per datapoint is now 1.15, showing a good fit quality. It should be noted that after the inclusion of systematic uncertainties in the analysis, the effect on the final result could be different depending on the specific situation we are considering. In other words, it is not always the case that the inclusion of new systematic effects leads to an increase of the final PDFs error, as we also noticed in chapter 5. This can be seen for example in the case of the distribution T_3 plotted in fig. 9.6, from which it is clear how the error of the fit *fine-sys* has not increased with respect to the one of *fine-stat*. The reason for this can be traced back to the fact that the covariance matrix defined in eq. (8.18) enters

Ensemble	fit	Obs	n_{dat}	χ^2	χ^2_{tot}
fine	no cuts	Re [$\mathfrak{M}$]	48	1.00	1.15
		Im [$\mathfrak{M}$]	48	1.30	

Table 9.3 *Details of the fit with systematic uncertainties. From left to right we report the lattice ensemble, the fit name, the observables included in the analysis, the number of datapoints and finally the partial and total χ^2.*

both the Monte Carlo replicas generation and the χ^2 definition: while the former mostly controls the final PDFs error, the latter is responsible for the relative weights different points have in the analysis. Points affected by bigger errors will give smaller contributions to the χ^2 and therefore will count less in the fit. Each replica will be shifted by a certain amount, which takes into account both the new replicas distribution and the different weights of the data entering the χ^2, so that the net effect on the final PDFs is non-trivial, and might consist in a global shift of the central value of the replicas distribution rather than in an increase of its error band.

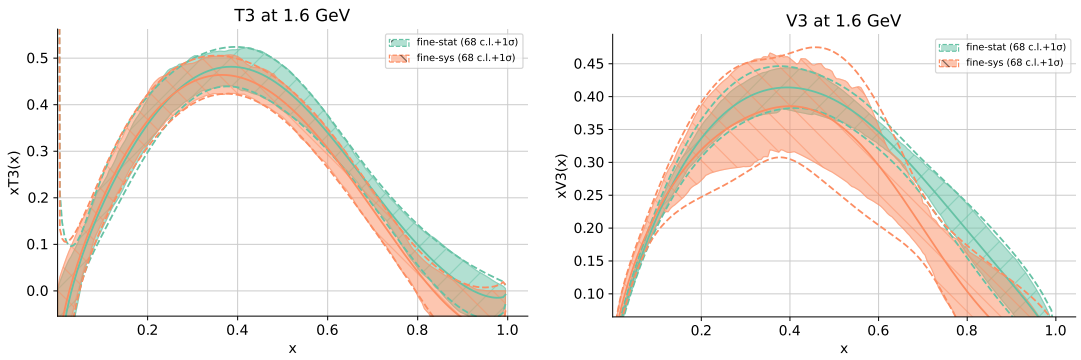

Figure 9.6 *PDFs from the fits fine-stat and fine-sys.*

Despite it is probably too early to draw comparisons between our results and phenomenological distributions, it is interesting to see how they look when plotted together, as we did in the previous chapter for quasi-PDFs results: given the fact that nowadays V_3 and T_3 are very well constrained by experimental data, the discrepancies we observe between lattice and phenomenological results might be a good indication of the size of the systematic we are still missing, highlighting specific x-region where the lattice PDFs error might have been underestimated. In fig. 9.7, our result *fine-sys* and the corresponding distributions from the NLO PDF sets NNPDF31 [48] are plotted together (orange and green curves respectively), both as absolute values (upper plots) and normalized to NNPDF31 (lower plots).

Looking at results from *fine-sys*, in the case of both V_3 and T_3 the two distributions are compatible up to medium (~ 0.25 and ~ 0.45) and for large values of x (> 0.8), showing a probable underestimation of the PDFs error for the intermediate x ranges.

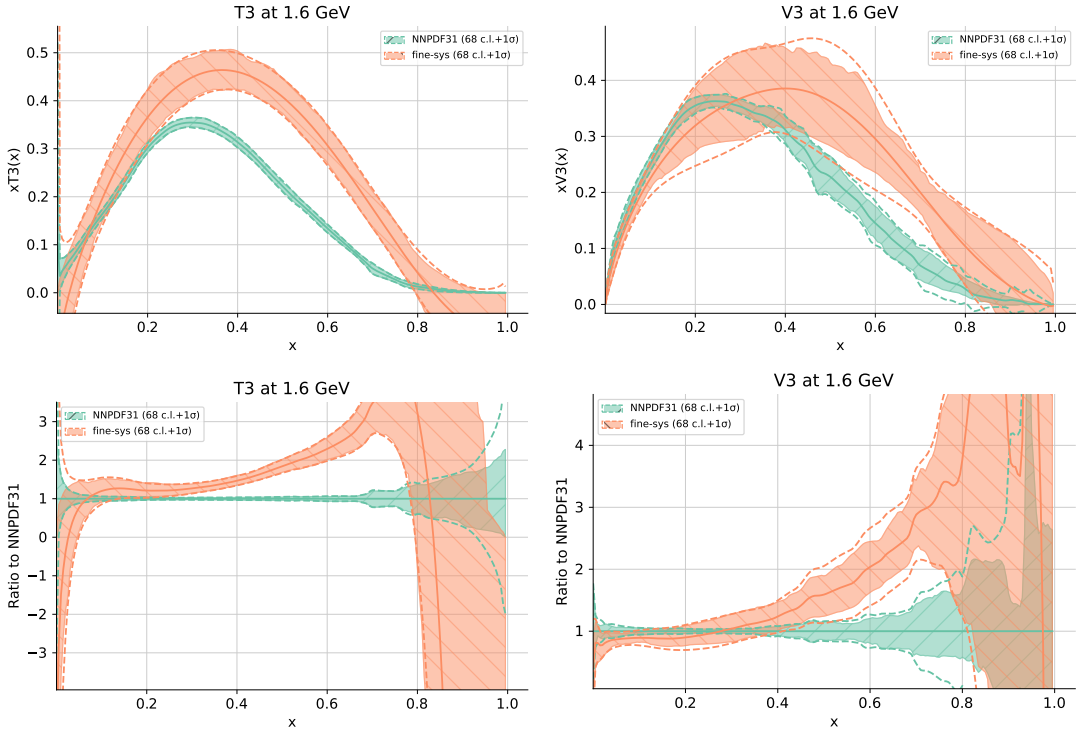

Figure 9.7 *PDFs from the fits fine-sys compared with the corresponding distributions from NNPDF31. In the lower plots results are normalized to NNPDF31 PDFs.*

To sum up, in this chapter we have considered the pseudo-ITD data produced in refs. [198, 200]. Using the position space factorization theorem relating such data to collinear PDFs, we have extracted two nonsinglet distributions within the `NNPDF` framework.

After extracting PDFs from different data sets and considering statistical uncertainties only, we have shown that in one of the cases considered, the fit quality appears to be really poor, pointing out the need for a detailed knowledge of the systematic effects. Using the results of ref. [198, 200] we have directly estimated those connected to finite volume, lattice spacing and pion mass effects. As for systematic uncertainties which cannot be directly computed from lattice results (like for example truncation effects and higher twist corrections), starting from the observed discrepancies between low Ioffe-time points we have used a Bayesian approach to introduce an additional systematic which allows us to

mitigate the tensions between the problematic datapoints, using the partial pieces of information which are available to us.

The Bayesian approach however is not completely satisfying, since it relies on a partial knowledge of the missing uncertainties and requires to make a number of assumptions about them. More work has to be done to achieve a detailed knowledge of the systematic uncertainties in lattice simulations: without a stringent control over them, it is not possible to draw reliable conclusions and to make comparisons with phenomenological distributions.

Finally, we stress once more that the analysis performed in this paper is complementary to that presented in the previous chapter, where quasi-PDFs matrix elements where considered instead, starting from the momentum space version of the factorization theorem. In both cases, results have been produced within the `NNPDF` environment, running the same machinery used for global QCD analysis over experimental data. It is therefore interesting to compare our best result of this chapter *fine-sys* with one of the more realistic cases of the previous one, like for example those produced using the systematic scenario S2. Both PDFs sets have been obtained using the same NNPDF methodology, the only difference being the input data (pesudo-ITD and quasi-PDFs data) and the corresponding errors, discussed in details in sec. 9.3 and 8.2.2 respectively. Quasi-PDFs and

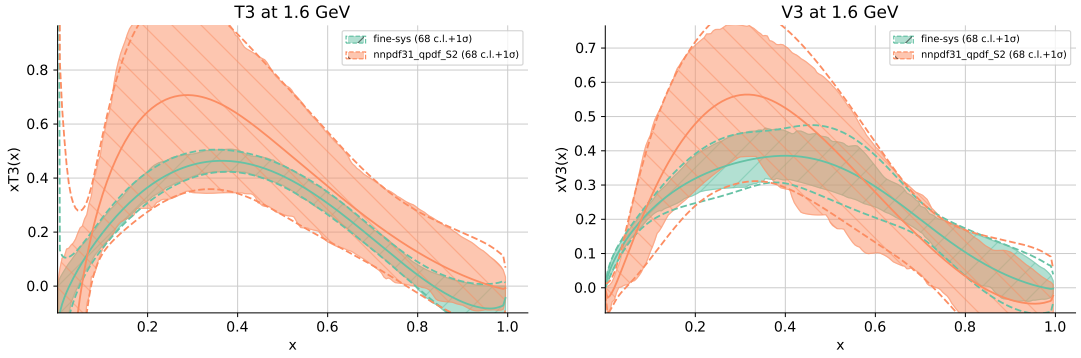

Figure 9.8 *PDFs from the fits fine-sys compared with the corresponding distributions from the fit presented in Chapter 8 produced with the systematics scenario S2.*

pseudo-ITD results are plotted together in fig. 9.8: both T_3 and V_3 distributions appear to be in good agreement, the main difference being a huge decrease in the PDFs error when considering results presented in this work. This difference can be partially traced back to the number of points included in the analysis: while in the analysis of chapter 8 16 points for quasi-PDFs matrix element where included, in this chapter data corresponding to all momentum values are

considered, for a total of 48 pseudo-ITD points. Clearly, having more points in the analysis allows to better constraint the fit results, giving final PDFs with smaller error. The size of the statistical and systematic uncertainties affecting the points entering the two analyses is of course another reason for different PDFs error. In general it is expected that, given equivalent computational cost, the low momenta matrix elements, which are used in the pseudo-PDF approach, are exponentially more precise than the large momenta matrix elements, to which the quasi-PDF approach are restricted. However a detailed study of such differences is left for a future study.

The next logical step might be a global lattice QCD fit within this same framework, where data for multiple lattice observables coming from different simulations are simultaneously included in the analysis. Given the results presented here it is clear that, when combining the currently available data for quasi-PDFs and reduced pseudo-ITD, the fit will be driven by the latter, and the results won't be much different from those of the fit *fine-sys*. However the situation might change when some level of maturity in terms of precision and systematic effects is achieved. One could then combine data from all pertinent lattice formalisms including results from the so-called "Good Lattice Cross-Sections" (LCS) approach, which is described in ref. [175] and represents a general framework, where one computes matrix elements that can be factorized into PDFs at short distances [246, 250–253], on the same lines of what described in chapter 7.

Clearly a global analysis only makes sense after having scrutinised each set of data individually, and having understood the systematics that affect them. Chapters 8, 9 represent a first step in this direction.

10 Summary

In this thesis we have presented a number of studies connected to the general topic of the proton Parton Distribution Functions. PDFs represent an essential input to perform computations of processes involving hadrons in collider physics, and nowadays they are responsible for the dominant source of theoretical uncertainties in many important analyses. In order to achieve better accuracy in theoretical predictions entering phenomenological studies, a better control of PDFs uncertainties is therefore necessary. This can be achieved through both improving the numerical frameworks and techniques commonly used to extract PDFs from experimental data, and exploring new approaches and physical ideas.

In the first part of the thesis we have revised the general methodology adopted by the NNPDF collaboration in global PDFs determination, describing its implementation within the `n3fit` environment. In particular, we have described the implementation of positivity and integrability constraints, studying their impact in a global fit. Additionally, we have presented studies regarding fit basis independence, showing how the final results are driven by the experimental input and do not depend on the methodological details, such as the choice of the distributions that are independently parameterized.

We have then presented a systematic study of the inclusion of inclusive jet production measurements at LHC in the context of global PDFs determination. Single-inclusive jets were considered and, for the first time in a PDFs fit, dijets data. In order to perform this study we have used recent NNLO QCD computations supplemented by EW corrections, both implemented in the fit by mean of K-factors.

Using a Bayesian approach, we have set up a general formalism to include different

sources of theoretical error in a global PDFs determination by means of a covariance matrix, and we have considered the case of missing higher order corrections in the QCD calculations entering the global analysis. After constructing and validating the corresponding covariance matrix, we have presented a first NLO global PDFs fit including missing higher order uncertainties, studying the impact of the theory error on the PDFs central value and uncertainty.

We have generalized the FONLL matching of massive and massless computations for hadronic processes involving heavy quarks to the case in which the heavy quark PDF is freely parameterized. As a first application we have studied the case of Higgs production in bottom fusion, showing how b PDF effect are likely to be comparable to mass effect. The determination of the bottom PDF from experimental data is therefore likely to be necessary for precision studies of b-induced hadron collider processes.

In the final chapters of the thesis we have addressed the problem of PDFs determination from a different point of view, exploring a number of recent ideas which have been developed within the lattice community. We have first revised and clarified the main conceptual points of such ideas in the context of a simple nongauge theory, proposing a general strategy to extract PDFs from lattice simulation in a systematic way. We have then considered some of the data which are currently available from lattice QCD simulations, and we have presented a first study of PDFs determination from lattice QCD observables adopting the NNPDF methodology. Our results show how promising lattice data might be in constraining PDFs, but also highlight the necessity for a deeper understanding and control of the different systematic errors entering lattice simulations.

A

Statistical estimators

A.1 PDF distance

Considering a set of $N_{\rm rep}$ replicas q_i of a given parton distribution q, the estimator for the expected true value of q is given by

$$\langle q \rangle = \frac{1}{N_{\rm rep}} \sum_{i=1}^{N_{\rm rep}} q_i \,. \tag{A.1.1}$$

The square distance between two estimates for the expected true value obtained from two different fits is given by [53]

$$d^2\left(\langle q^{(1)} \rangle, \langle q^{(2)} \rangle\right) = \frac{\left(\langle q^{(1)} \rangle - \langle q^{(2)} \rangle\right)^2}{\sigma^2\left[\langle q^{(1)} \rangle\right] + \sigma^2\left[\langle q^{(2)} \rangle\right]} \,, \tag{A.1.2}$$

with the variance of the mean given by

$$\sigma^2\left[\langle q^{(k)} \rangle\right] = \frac{1}{N_{\rm rep}} \sigma^2\left[q^{(k)}\right] \,, \tag{A.1.3}$$

with $\sigma^2\left[q^{(k)}\right]$ the variance of the variable $q_i^{(k)}$

$$\sigma^2\left[q^{(k)}\right] = \frac{1}{N_{\rm rep} - 1} \sum_{i=1}^{N_{\rm rep}} \left(q_i^{(k)} - \langle q^{(k)} \rangle\right)^2 \,. \tag{A.1.4}$$

According to this definitions, $d \simeq 1$ corresponds to statistically equivalent fits, while, considering a fit with $N_{\rm rep} = 100$ replicas , $d \simeq 10$ corresponds to a difference of one-sigma in units of the corresponding variance.

A.2 ϕ estimator

In chapter 5 we introduced the estimator ϕ, defined as

$$\phi = \sqrt{\langle \chi^2_{\rm exp}\left[T^{(k)}\right]\rangle - \chi^2_{\rm exp}\left[\langle T^{(k)}\rangle\right]}\,. \tag{A.2.1}$$

Here, following ref. [55], we show how such quantity measures the standard deviation over the replica sample in units of the data uncertainty. Using the χ^2 definition the first term of eq. A.2.1 can be written as

$$\begin{aligned} N_D\langle \chi^2_{\rm exp}\left[T^{(k)}\right]\rangle = &\sum_{IJ}\langle T_I^{(k)} C^{-1}_{IJ} T_J^{(k)}\rangle - \sum_{IJ} D_I C^{-1}_{IJ}\langle T_J^{(k)}\rangle \\ &- \sum_{IJ}\langle T_I^{(k)}\rangle C^{-1}_{IJ} D_J + \sum_{IJ} D_I C^{-1}_{IJ} D_J\,, \end{aligned} \tag{A.2.2}$$

so that subtracting $\chi^2_{\rm exp}\left[\langle T^{(k)}\rangle\right]$ we get

$$N_D\left(\langle \chi^2_{\rm exp}\left[T^{(k)}\right]\rangle - \chi^2_{\rm exp}\left[\langle T^{(k)}\rangle\right]\right) = \sum_{IJ}\langle T_I^{(k)} C^{-1}_{IJ} T_J^{(k)}\rangle - \sum_{IJ}\ \langle T_I^{(k)}\rangle C^{-1}_{IJ}\langle T_J^{(k)}\rangle \tag{A.2.3}$$

from which

$$\phi^2 = \frac{1}{N_D}\sum_{IJ} C^{-1}_{IJ} T_{JI}\,, \tag{A.2.4}$$

i.e. the average over all the data points of the uncertainties and correlations of the theoretical predictions, T_{IJ}, normalized according to the corresponding uncertainties and correlations of the data as expressed through the covariance matrix C_I.

B

Impact of the choice of the correlation model

In this appendix we discuss the impact of different decorrelation models for the 8 TeV single-inclusive jet data from ATLAS. As discussed in sec. 4.3.1, these data appear to be fully consistent with the corresponding 7 TeV ones, yet they show a poor χ^2 when included in the global fit. The problem might be due to some issues in the covariance matrix of these data, in a similar way to what discussed in ref. [254] for 7 TeV dataset.

To check whether or not this is indeed the case, starting from the default fit (#janw) we produce two variants by modifying the treatment of three (out of 659) correlated systematic uncertainties related to the jet energy scale, as suggested in ref. [119]. Specifically, in the first variation, denoted as #janw-8dec, these three uncertainties are completely decorrelated; in the second variation, denoted as #janw-8pcor, they are partially decorrelated, splitting each uncertainty into three components and decorrelating one of them. From Table. B.1 we see how, upon decorrelation the χ^2 for ATLAS improves considerably, leaving the values for the other datasets almost unaffected. Very similar results are obtained when fully or partially decorrelating the relevant sources of systematics, thus validating the prescription of ref. [119]. At the PDFs level the results are very stable, as we see by inspection of fig. , where the gluon PDF for #janw and #janw-8dec are compared.

We conclude that the decorrelation models suggested in ref. [119] solve the issue observed in sec. 4.3, leading to a good fit quality of the ATLAS single-inclusive jet data at 8 TeV without significant change in the PDFs.

Dataset	n_{dat}	janw	janw-8dec	janw-8pcor
ATLAS 7 TeV	31	1.59	1.59	1.61
ATLAS 8 TeV	171	3.22	0.83	0.98
CMS 7 TeV	133	1.09	1.12	1.12
CMS 8 TeV	185	1.25	1.42	1.42
ATLAS 7 TeV	90	[1.95]	[1.98]	[1.98]
CMS 7 TeV	54	[2.08]	[2.19]	[2.17]
CMS 8 TeV	122	[2.21]	[2.96]	[3.04]

Table B.1 *Same as Table 4.3 for fits performed with alternative choices of decorrelation models. Now only χ^2 values for jet data are shown. Results for the fits with default settings #janw already shown in Table 4.3 are included for ease of reference.*

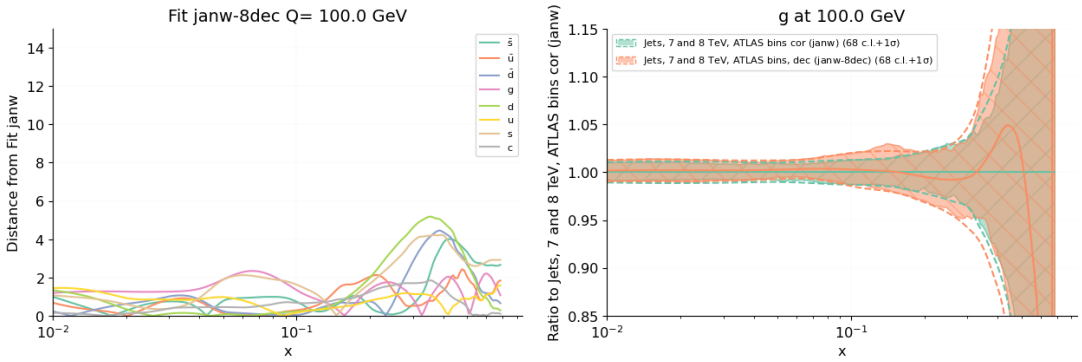

Figure B.1 *Same as fig. 4.3, but now comparing the default fit to single-inclusive jet data (fit #janw), to a fit in which selected systematic uncertainties are decorrelated in the ATLAS 8 TeV data (fit #janw-8dec). The gluon is shown as ratio to the former fit.*

C

Alternative points prescription for theory error

In sec. 5.2 we have discussed how to construct the theory covariance matrix using the so-called 9-points prescription, where the renormalization and factorization scales of a specific process are varied completely independently. In ref. [9] a series of alternative options are investigated, and the 9-points prescription is selected as the most suitable one according to the validation procedure described in sec. 5.3. In this appendix we present an alternative simpler choice, denoted as 3-points prescription, studying how the fit and validation results change, and showing how the 9-points prescription is indeed a better option.

In the 3-points prescription both the renormalization and factorization scales are varied coherently by a fixed amount about the central value. In other words, considering a single process we set $k_f = k_r$ and we only vary the single resulting scale, obtaining the 2+1 points in the scales space shown in fig. C.1. The resulting entries for the theory covariance matrix are

$$S_{ij}^{(\mathrm{3pt})} = \frac{1}{2}\left\{\Delta_i^{++}\Delta_j^{++} + \Delta_i^{--}\Delta_j^{--}\right\}, \tag{C.0.1}$$

where the indices i, j label point belonging to the same process π. Considering two different processes π_1 and π_2 we set $k_f = k_{r_1}$ for π_1 and $k_f = k_{r_1}$ for π_2 and then vary k_{r_1} and k_{r_2} independently. Note that in this way the correlation in the variation of k_f between π_1 and π_2 is necessarily ignored so that the variations for each process are entirely uncorrelated. eq. (C.0.1) is generalized to the off-diagonal entries of the theory covariance matrix getting

$$S_{i_1 j_2}^{(\mathrm{3pt})} = \frac{1}{4}\left\{\left(\Delta_{i_1}^{++} + \Delta_{i_1}^{--}\right)\left(\Delta_{j_2}^{++} + \Delta_{j_2}^{--}\right)\right\}, \tag{C.0.2}$$

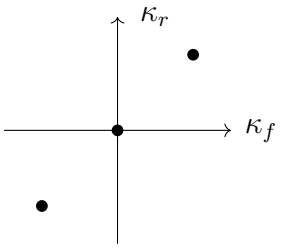

Figure C.1 *3-points prescription for a single process.*

where the indices i_1, j_2 now label point belonging to the processes π_1 and π_2 respectively. The validation of such covariance matrix proceeds like described in sec. 5.3, through the computation of the angle θ between the shift δ and its component on the subspace S, defined in eq. (5.30). In Table C.1 we compare the values of such angle for each process in the case of the 3- and 9-points prescriptions. We note how the former performs rather poorly in comparison to the 9-points one, suggesting that the lack of correlation in the factorization scale between processes in this prescription implies that much of the correlation pattern in the MHOU due to universal PDF evolution has been missed.

Presc.	DIS NC	DIS CC	DY	JET	TOP
9-pt	32°	16°	22°	14°	3°
3-pt	54°	36°	39°	24°	12°

Table C.1 *Comparison between the θ values for the 3- and 9-points prescriptions.*

In Table. C.2 we compare the values for the total χ^2 and ϕ estimator for fits performed with the 9-points, already shown in sec. 5.4, and the 3-points prescription, together with values for the baseline NLO fit produced without any theory error. Unlike the 9-points prescription case, for which as observed in sec. 5.4 the χ^2 improves with respect to the baseline, in the case of the 3-points prescription the fit quality remains unchanged after the inclusion of the MHOU. Both prescriptions show an increase in the value of ϕ, which is bigger in the case of the 9-points one.

Finally in fig. C.2 we study the dependence of the fit results on the choice of the prescription for the theory covariance matrix, taking as example the gluon distribution. In the same plot we report also the central value of the NNLO fit with experimental uncertainties only and all the distributions are normalized to the 3-points prescription results. In general the two results are consistent,

	C	$C + S^{(3\text{pt})}$	$C + S^{(9\text{pt})}$
χ^2	1.139	1.139	1.109
ϕ	0.314	0.394	0.405

Table C.2 *Comparison between χ^2 and ϕ total values of 3- and 9-points prescriptions*

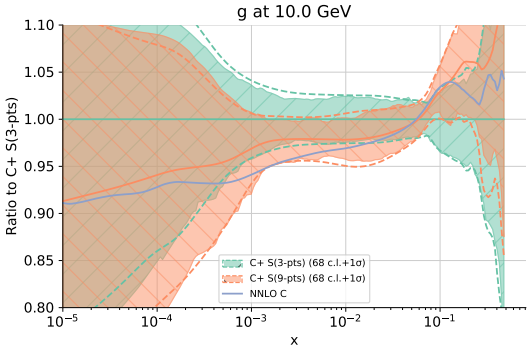

Figure C.2 *Comparison between the gluon PDF produced using the 3-points and 9-points prescription for the theory covariance matrix. The central value of the NNLO fit without any theory error is also shown. All the distributions are normalized to the 3-points prescription results.*

but the central value of the 9-points prescription result is much closer to the NNLO central value, providing further confirmation for preferring the 9-point prescription.

D

Massive-b scheme

D.1 Matching coefficients

We collect for ease of reference the well-known matching coefficients which relate the four and five scheme PDFs. Up to $O(\alpha_s)$

$$K_{ij}(z,Q^2) = \delta_{ij}\delta(1-z) + \alpha_s(Q^2)\, K_{ij}^{(1)}(z,Q^2) + \mathcal{O}(\alpha_s^2)\,. \tag{D.1.1}$$

so that

$$K_{ij}^{-1}(z,Q^2) = \delta_{ij}\delta(1-z) - \alpha_s(Q^2)\, K_{ij}^{(1)}(z,Q^2) + \mathcal{O}(\alpha_s^2)\,. \tag{D.1.2}$$

The only non-zero contributions at order $\mathcal{O}(\alpha_s)$ are the heavy quark-heavy quark and the heavy quark-gluon matching functions, which are respectively given by

$$\begin{aligned} K_{bb}^{(1)}\left(x,\frac{Q^2}{\mu_b^2}\right) &= \frac{C_F}{2\,\pi}\left\{P_{qq}(x)\left[\ln\frac{Q^2}{\mu_b^2} - 2\ln(1-x) - 1\right]\right\}_+ \\ K_{bg}^{(1)}\left(x,\frac{Q^2}{\mu_b^2}\right) &= \frac{T_R}{2\,\pi}P_{qg}(x)\,\ln\frac{Q^2}{\mu_b^2} \end{aligned} \tag{D.1.3}$$

where

$$P_{qg}(x) = \left(1\,-\,2\,x\,+\,2\,x^2\right) \quad \text{and} \quad P_{qq}(x) = \frac{2}{1-x} - (1+x)\,. \tag{D.1.4}$$

D.2 Massive coefficient functions

In this appendix we summarize the computation of the coefficient functions in the massive scheme and of their massless limit up to $O(\alpha_s)$. The NLO corrections are computed using the extension of Catani-Seymour subtraction for massive initial states developed in ref. [147] and extended to QCD in ref. [139]. This way of preforming the computation has the main advantage of following closely that of the five-flavor massive scheme, so that a direct comparison is much easier to at the analytic level. Indeed, strictly speaking because of eq. (6.15) the massless limit is not needed. However, we have computed it explicitly in order to check that it matches the massless-scheme result (thereby verifying eq. (6.14) explicitly), and also in order to produce fig. 6.2, which provides a further consistency check. Another advantage of this way of performing the computation (though we do not use it here) is that it allows for the computation of the fully differential cross section in this scheme.

D.2.1 Leading order

The leading order partonic cross section for the production of a Higgs boson, accounting for the mass of the initial state b and $\bar{b}$, is given by

$$\hat{\sigma}_0(xs) = \left(\frac{g^2_{b\bar{b}H}\,\beta_0\,\pi}{6}\right)\delta(xs - m_H^2) = \sigma_0\, x\,\delta\left(x - \frac{m_H^2}{s}\right) \tag{D.2.1}$$

where

$$\sigma_0 = \frac{g^2_{b\bar{b}H}\,\beta_0\,\pi}{6\,m_H^2}\,, \quad \text{and} \quad \beta_0 = \sqrt{1 - \frac{4\,m_b^2}{m_H^2}}\,. \tag{D.2.2}$$

where $g_{b\bar{b}H}$ is the coupling of the b quark to the Higgs boson, obtained as the mass of the quark divided by vacuum expectation value of the Higgs sector:

$$g_{b\bar{b}H} = \frac{m_b}{v}\,. \tag{D.2.3}$$

In the following we will also use the notation

$$\mathcal{B}(x) \equiv \hat{\sigma}_0(xs)\,, \quad \text{and} \quad \mathcal{B} \equiv \hat{\sigma}_0(s)\,. \tag{D.2.4}$$

D.2.2 Next-to-leading order: $b\bar{b}$-channel

The next to leading order corrections to the Higgs production in bottom quark fusion consist in virtual corrections ($\mathcal{V}$) to the left diagram of fig. 6.1, as well as of real emission corrections ($\mathcal{R}$) , represented by the central diagram of fig. 6.1. Both this contributions are separately divergent when the additional gluon, real or virtual, becomes soft, though the final result remains finite. In order to handle these soft divergences we employ the subtraction scheme defined in [139]. This implies that we need two more ingredients: a subtraction term, $\mathcal{S}$, and its integral over the gluon phase space, $\mathcal{I} = \int \mathrm{d}\Phi_g \mathcal{S}$. Our final result is then given by:

$$\hat{\sigma}_{\mathrm{NLO}} = \int \mathrm{d}\Phi_1 \mathcal{B} + \mathcal{V} + \mathcal{I} + \int \mathrm{d}\Phi_2 \mathcal{R} - \mathcal{S}\,. \tag{D.2.5}$$

Real corrections, and subtraction term

The real emission partonic differential cross section, is given by

$$\int \mathrm{d}\Phi_2 \mathcal{R} = \int \mathrm{d}\Phi_2 \left|\overline{\mathcal{M}}_{b\bar{b}Hg}\right| (s,t,u)\,, \tag{D.2.6}$$

where

$$\mathrm{d}\Phi_2 = \frac{1}{32\,\pi\,\beta\,s}\mathrm{d}\cos\theta\,\Theta(1+\cos\theta)\,\Theta(1-\cos\theta)\,, \quad \beta = \sqrt{1-\frac{4\,m_b^2}{s}}\,, \tag{D.2.7}$$

and

$$\begin{aligned}\left|\overline{\mathcal{M}}_{b\bar{b}Hg}\right|(s,t,u) = \frac{4}{3}\pi g_{b\bar{b}H}^2 C_F \alpha_s \Bigg\{ \left(s-m_H^2\right)\left[\frac{1}{m_b^2-t}+\frac{1}{m_b^2-u}\right] \\ +(m_H^2-4\,m_b^2)\left[\frac{2\,(s-2\,m_b^2)}{(m_b^2-t)(m_b^2-u)} - \frac{2\,m_b^2}{(m_b^2-t)^2} - \frac{2\,m_b^2}{(m_b^2-u)^2}\right]\Bigg\}\,.\end{aligned} \tag{D.2.8}$$

The Mandelstam variables in terms of scalar products and $\cos\theta$ are given by

$$\begin{cases} t = m_b^2 - \dfrac{s-m_H^2}{2}\left(1-\beta\,\cos\theta\right) \\ u = m_b^2 - \dfrac{s-m_H^2}{2}\left(1+\beta\,\cos\theta\right) \end{cases}. \tag{D.2.9}$$

In order to remove the soft divergence which appears in the $s \to m_H^2$ limit we

need to construct a suitable subtraction term. Using the relevant equations in ref. [139] we find

$$\mathcal{S} = \frac{2}{3}\,\pi\,\alpha_s\,C_F\,g_{b\bar{b}H}^2\,\beta_0^2\,m_H^2\,\frac{1}{\tilde{x}}\left[\frac{2}{m_b^2-t}\left(P_{qq}(\tilde{x}) - \frac{2\,\tilde{x}\,m_b^2}{m_b^2-t}\right) + \frac{2}{m_b^2-u}\left(P_{qq}(\tilde{x}) - \frac{2\,\tilde{x}\,m_b^2}{m_b^2-u}\right)\right], \tag{D.2.10}$$

where

$$\tilde{x} = \frac{m_H^2 - 2\,m_b^2}{s - 2\,m_b^2}\,. \tag{D.2.11}$$

Combining eqs. (D.2.8) and (D.2.10) and factoring the trivial $\frac{\alpha_s\,C_F\,\sigma_0}{\pi}$ dependence we get

$$\begin{aligned}\frac{\alpha_s\,C_F\,\sigma_0}{\pi}\int \mathrm{d}\Phi_2\,[\mathcal{R}-\mathcal{S}] &= \frac{\alpha_s\,C_F\,\sigma_0}{\pi}\frac{m_b^2}{2}\int_{-1}^{1}\mathrm{d}\cos\theta\left[\frac{s\,(s-m_H^2)^2}{(m_H^2-2\,m_b^2)(m_b^2-t)(m_b^2-u)}\right] \\ &= -\frac{\alpha_s\,C_F\,\sigma_0}{\pi}\frac{1}{\beta_0}\left(\frac{1-\beta^2}{\beta^2}\right)\frac{x}{(1-2\,x-\beta^2)}\ln d\,,\end{aligned} \tag{D.2.12}$$

where we defined

$$d \equiv \frac{1+\beta}{1-\beta}\,, \quad \text{and} \quad x \equiv \frac{m_H^2}{s}\,. \tag{D.2.13}$$

Virtual corrections, and integrated subtraction term

QCD virtual corrections to the Born process in this simple case completely factorize in a vertex form factor:

$$\mathcal{V} = \frac{\alpha_s\,C_F}{\pi}\mathcal{B}\,\delta_g\,, \tag{D.2.14}$$

with

$$\begin{aligned}\delta_g = &-1 - L_\lambda + \frac{(1-\beta_0^2)}{\beta_0}\ln d_0 \\ &- \frac{1+\beta_0^2}{2\,\beta_0}\left[-\ln d_0\,L_\lambda + \ln^2 d_0 + \mathrm{Li}_2\left(1-\frac{1}{d_0}\right) - \frac{\pi^2}{2}\right],\end{aligned} \tag{D.2.15}$$

where

$$L_\lambda \equiv \frac{1}{\epsilon} + \ln\frac{4\,\pi\,\mu_R^2}{m_b^2} + \mathcal{O}(\epsilon^2)\,. \tag{D.2.16}$$

The integrated subtraction term $\mathcal{I}$ is obtained by integrating $\mathcal{S}$, eq. (D.2.10), over the phase space of the emitted gluon. This term can be separated into two pieces: a term proportional to $\delta(1-x)$, which contains the singularity, and a plus distribution:

$$\mathcal{I} = \delta(1-x)\, I + \{\mathcal{G}(x)\}_+ \,, \tag{D.2.17}$$

where

$$\begin{aligned} I =& 2 + L_\lambda - \ln \frac{(1+\beta_0^2)^2}{1-\beta_0^2} + \frac{1-3\,\beta_0^2}{4\beta_0} \ln d_0 \\ &+ \frac{1+\beta_0^2}{2\,\beta_0} \left[\frac{1}{2} \ln^2 d_0 - \ln d_0 \, \ln \frac{4\,\beta_0^2}{(1+\beta)^2} - L_\lambda \, \ln d_0 - 1 + 2\,\mathrm{Li}_2\left(\frac{1}{d_0}\right) - \frac{\pi^2}{3} \right] , \end{aligned} \tag{D.2.18}$$

and

$$\{\mathcal{G}(x)\}_+ = \left\{ P_{qq}(x) \left[\frac{1+\beta^2}{2\beta} \ln d - 1 \right] + (1-x) \right\}_+ . \tag{D.2.19}$$

Final formulae, mass and PDF renormalization

We now combine the various partial results obtained in the previous subsections into the full expression for the $b\bar{b}$-channel coefficient functions. First, however, we need to adjust b-quark mass and the PDFs. Renormalization of the b mass leads to the replacement

$$g^2_{hb\bar{b}} = g^2_{hb\bar{b}}(\mu_R^2) \left(1 - \frac{\alpha_s\, C_F}{\pi} \left(\frac{3}{2} \ln \frac{m_b^2}{\mu_R^2} - 2 \right) \right) . \tag{D.2.20}$$

in σ_0, eq. (D.2.2).

The massive b PDF is free of collinear singularities and thus it does not have to undergo subtraction: indeed it is scale independent. However, we must perform the change of renormalization scheme eq. (6.11) which relates the massive and massless schemes. Up to $\mathcal{O}(\alpha_s)$ we get

$$\begin{aligned} B_{b\bar{b}}\left(x, \mu_R^2, \mu_F^2, \mu_b^2\right) = \Big[& \sigma_0(\mu_R^2)\delta(1-x) + \\ & \alpha_s(\mu_R^2) B^{(1)}_{b\bar{b}}\left(x, \mu_R^2, \mu_F^2, \mu_b^2\right) \Big] + \mathcal{O}(\alpha_s^2) \end{aligned} \tag{D.2.21}$$

where

$$B^{(1)}_{b\bar{b}}\left(x,\mu_R^2,\mu_F^2,\mu_b^2\right) = \frac{\sigma_0(\mu_R^2)\,C_F}{\pi}\left\{\left[\frac{3}{2}\ln\frac{\mu_R^2}{\mu_b^2}+2+I+\delta_g\right]\delta(1-x) + \int_0^1 \mathrm{d}z\{\mathcal{G}(z)-2\,K^{(1)}_{b\bar{b}}(z)\}_+ z\,\delta(z-x)+\int \mathrm{d}\Phi_2\,[\mathcal{R}-\mathcal{S}]\right\}. \tag{D.2.22}$$

Performing the z integration gives the final result

$$\begin{aligned} &B^{(1)}_{b\bar{b}}\left(x,\mu_R^2,\mu_F^2,\mu_b^2\right) = \\ &\quad\frac{\sigma_0(\mu_R^2)\,C_F}{\pi}\Bigg\{\delta(1-x)\left[\xi-2+\frac{3}{2}\left(\gamma_0\ln\frac{(1+\beta)^2}{4}-\gamma_0\ln\frac{m_H^2}{m_b^2}+\ln\frac{\mu_R^2}{\mu_F^2}\right)\right] \\ &+4\,\mathcal{D}_1(1-x)+2\left[\gamma\ln\frac{(1+\beta)^2}{4}+\gamma\ln\frac{m_H^2}{m_b^2}+\ln\frac{\mu_b^2}{\mu_F^2}\right]\mathcal{D}_0(1-x) \\ &-(2+x+x^2)\left[\gamma\ln\frac{(1+\beta)^2}{4}+\gamma\ln\frac{m_H^2}{m_b^2}-\gamma\ln x+\ln\frac{\mu_b^2}{\mu_F^2}+2\ln(1-x)\right]+x\,(1-x) \\ &-\frac{2\,\gamma\,\ln x}{1-x}-\frac{1}{\beta_0}\left(\frac{1-\beta^2}{\beta^2}\right)\frac{x}{(1-2\,x-\beta^2)}\ln d\Bigg\}, \end{aligned} \tag{D.2.23}$$

where

$$\begin{aligned} \xi &= 1+\ln\left(\frac{1-\beta_0^2}{(1+\beta_0)^2}\right)+\frac{(5-7\beta_0^2)}{4\,\beta_0}\ln d_0 \\ &\quad+\frac{(\beta_0^2+1)}{\beta_0}\left(2\,\mathrm{Li}_2\left(\frac{1}{d_0}\right)+\frac{\pi^2}{6}-\ln d_0\ln\frac{4\beta_0^2}{(1+\beta_0^2)(1+\beta_0)}\right), \end{aligned} \tag{D.2.24}$$

and

$$\gamma=\frac{1+\beta^2}{2\,\beta}, \quad \gamma_0=\frac{1+\beta_0^2}{2\beta_0} \quad \text{and} \quad \mathcal{D}_n(x)=\left(\frac{\ln^n(1-x)}{1-x}\right)_+. \tag{D.2.25}$$

Massless limit

The massless limit of the $b\bar{b}$-channel can be computed directly from eq. (D.2.23), by setting $\beta = 1$ everywhere except in the logarithms, where one can use the simple expansion

$$\beta \sim 1 - \frac{2\,x\,m_b^2}{m_H^2}+\mathcal{O}\left(\frac{m_b^4}{m_H^4}\right). \tag{D.2.26}$$

We get

$$B_{b\bar{b}}^{(1),(0)}\left(x,\mu_R^2,\mu_F^2,\mu_b^2\right)=\frac{\alpha_s\,C_F\,\sigma_0(\mu_R^2)}{\pi}\Bigg\{\delta(1-x)\left[-1+\frac{\pi^2}{3}+\frac{3}{2}\ln\frac{\mu_R^2}{\mu_F^2}\right]$$
$$+4\,\mathcal{D}_1(1-x)+2\left(\ln\frac{m_H^2}{\mu_F^2}+\ln\frac{\mu_b^2}{m_b^2}\right)\mathcal{D}_0(1-x)-\frac{2\ln x}{1-x} \tag{D.2.27}$$
$$-\left(2+x+x^2\right)\left[\ln\frac{m_H^2}{\mu_F^2}+\ln\frac{\mu_b^2}{m_b^2}+\ln\frac{(1-x)^2}{x}\right]+x\,(1-x)\Bigg\}.$$

As it can be easily verified, this exactly corresponds to its massless scheme equivalent, which can be found in eq. (A6) of ref. [255].

D.2.3 Next-to-leading order: bg-channel

In the presence of initial-state massive quarks, the cross-section for the bg-channel is free of soft or collinear divergences, and no subtraction is accordingly necessary. Also in this case, however, we must perform the scheme change eq. (6.11). We get

$$\begin{aligned}B_{bg}^{(1)}(x,\mu_R^2,\mu_F^2,\mu_b^2)&=\hat{\sigma}_{bg}(x,\mu_R^2)-\alpha_s\int_0^1\mathrm{d}z\,K_{bg}^{(1)}(z,\mu_F^2)\sigma(zs)\\&=\hat{\sigma}_{bg}(x,\mu_R^2)-\frac{\alpha_s\,T_R\,\sigma_0}{\pi}\left[\frac{x}{2}\,P_{qg}(x)\ln\frac{\mu_F^2}{\mu_b^2}\right]\Bigg|_{x=\frac{m_H^2}{s}},\end{aligned} \tag{D.2.28}$$

where

$$\hat{\sigma}_{bg}(x,\mu_R^2)=\int\mathrm{d}\Phi_2^{(b)}\left|\overline{\mathcal{M}}_{bgHb}\right|^2(s,t,u)\,, \tag{D.2.29}$$

and the subscript (b) in $\Phi_2^{(b)}$ denotes the fact that now the phase-space has a massive b instead of a massless gluon, in the final state. The color- and helicity-averaged square matrix element, can be obtained from eq. (D.2.8) using crossing symmetry. In addition, we have to take into account that the gluon can have 8 possible colors (as opposed to 3 for a quark),

$$\left|\overline{\mathcal{M}}_{bgHb}\right|^2(s,t,u)=-\frac{3}{8}\left|\overline{\mathcal{M}}_{b\bar{b}Hg}\right|^2(t,s,u)\,, \tag{D.2.30}$$

where now the Mandelstam invariants are given by

$$\begin{cases} t = 2\,m_b^2 + \dfrac{s}{32}\left((5-\beta^2)(\beta^2+4\,x-5) - (3+\beta^2)\,\Lambda\,\cos\theta\right) \\ u = m_b^2 + \dfrac{s}{32}\left((5-\beta^2)(\beta^2+4\,x-5) + (3+\beta^2)\,\Lambda\,\cos\theta\right), \end{cases} \tag{D.2.31}$$

where

$$\Lambda = \sqrt{(3+\beta^2)^2 + 16x^2 - 8\,x\,(5-\beta^2)}\,, \tag{D.2.32}$$

while the phase-space $\mathrm{d}\Phi_2^{(b)}$ is given by

$$\mathrm{d}\Phi_2^{(b)} = \frac{\Lambda\,x}{32\,\pi\,(3+\beta^2)\,m_H^2}\,\mathrm{d}\cos\theta\,\Theta(1+\cos\theta)\,\Theta(1-\cos\theta)\,. \tag{D.2.33}$$

Performing the $\cos\theta$ integration gives

$$\begin{aligned} \hat{\sigma}_{bg}(x,\mu_R^2) =& \frac{\alpha_s\,T_R\,\sigma_0(\mu_R^2)}{\pi}\frac{x}{16\beta_0\,(\beta^2+3)^3} \\ &\times\Bigg\{-64\left(9\beta^4+(40x-42)\beta^2+8x(4x-9)+49\right)\operatorname{arctanh}\left(\frac{\Lambda}{\beta^2+4x-5}\right) \\ &\frac{4096\,\Lambda\,(1-\beta^2)\,(\beta^2+x-1)}{(-\Lambda+\beta^2+4x-5)\,(\Lambda+\beta^2+4x-5)} \\ &+\Lambda\,\left(5-\beta^2\right)\left(\beta^4+(4x+22)\beta^2+44x-71\right)\Bigg\}. \end{aligned} \tag{D.2.34}$$

Massless limit

As in the case of the $b\bar{b}$ channel, taking the massless limit requires setting $\beta = 1$ everywhere except in the logarithms where one can use eq. (D.2.26), which gives

$$\begin{aligned} B_{bg}^{(1),(0)}(x,\mu_R^2,\mu_F^2,\mu_b^2) = \frac{T_R}{\pi}\Bigg\{&\frac{x}{2}\,P_{qg}(x)\left[\ln\left(\frac{(1-x)^2}{x}\right)+\ln\frac{m_H^2}{\mu_F^2}+\ln\frac{\mu_b^2}{m_b^2}\right] \\ &-\frac{x}{4}(1-x)(3-7\,x)\Bigg\}. \end{aligned} \tag{D.2.35}$$

Once again, one can explicitly check that this exactly corresponds to it massless limit counterpart, which can be found in eq. (A9) of ref. [255].

E

Lattice observables

E.1 Momentum space factorization

In this appendix we report in detail some of the computations performed in sec. 7.3.2, to obtain the coefficient C of eq. (7.49) and its high momentum limit of eq. (7.54). In order to compute the Fourier transform of the coefficient $\tilde{C}$ entering eq. (7.37), we perform a change variable, $\theta = \xi P_3 z_3$, and define $\eta = \frac{y}{\xi}$, so that

$$\begin{aligned}\frac{P_3}{2\pi}\int_{-\infty}^{\infty} dz_3\, e^{-iyP_3z_3}\,\tilde{C}\left(xP_3z_3, mz_3, \frac{\mu^2}{m^2}\right) &= \frac{1}{x}\int_{-\infty}^{\infty}\frac{d\theta}{2\pi}\, e^{-i\eta\theta}\,\tilde{C}\left(\theta, \frac{m\theta}{xP_3}, \frac{\mu^2}{m^2}\right) = \\ &= \frac{1}{x}\left[\delta\left(\eta-1\right) - \alpha\int_0^1 d\xi\,\left(1-\xi\right)\int_{-\infty}^{\infty}\frac{d\theta}{2\pi}\, e^{-i(\eta-\xi)\theta}\left(2K_0\left(\frac{M\theta}{xP_3}\right) - \log\frac{\mu^2}{M^2}\right)\right].\end{aligned} \tag{E.1.1}$$

The Fourier transform of the Bessel function, obtained also in ref. [237], can be computed using the integral representation in eq. (7.30), computing the gaussian integral over θ first:

$$\int_{-\infty}^{\infty}\frac{d\theta}{2\pi}\, e^{-i(\eta-\xi)\theta}\int_0^{\infty}\frac{dT}{T}\, e^{-T} e^{-\left(\frac{M\theta}{xP_3}\right)^2\frac{1}{4T}} = \frac{1}{\sqrt{\left(\eta-\xi\right)^2 + \frac{M^2}{x^2P_3^2}}}\,, \tag{E.1.2}$$

so that the $\mathcal{O}(\alpha)$ contribution to (E.1.1) can be written as

$$\int_0^1 d\xi\,\left(1-\xi\right)\left[\frac{1}{\sqrt{\left(\eta-\xi\right)^2 + \frac{M^2}{x^2P_3^2}}} - \delta\left(\xi-\eta\right)\log\frac{\mu^2}{M^2}\right]. \tag{E.1.3}$$

As mentioned in sec. 7.3.2, the computation of the large-P_3 limit when $\eta \in (0,1)$ requires additional care, since the integrand develops a non-integrable divergence for $\xi = \eta$ when $M^2/(x^2 P_3^2) \to 0$. This issue was first addressed and solved in ref. [197] in the context of QCD. Since in the scalar theory the same kind of Bessel function appears, its Fourier transform leads to an analogous singularity. In order to elucidate this problem, given a generic test function $\phi(\xi)$, we consider the integral

$$\int_0^1 d\xi \, \frac{\phi(\xi)}{\sqrt{(\eta-\xi)^2+\kappa^2}} \tag{E.1.4}$$

in the limit where $\kappa \to 0$. Defining

$$G\left(\eta,\kappa^2\right) = \int_0^1 \frac{d\xi}{\sqrt{(\xi-\eta)^2+\kappa^2}} \tag{E.1.5}$$

allows us to rewrite eq. (E.1.4) above as

$$\int_0^1 d\xi \, \frac{\phi(\xi)}{\sqrt{(\eta-\xi)^2+\kappa^2}} = \phi(\eta) G\left(\eta,\kappa^2\right) + \int_0^1 d\xi \, \frac{1}{\sqrt{(\eta-\xi)^2+\kappa^2}} \left(\phi(\xi)-\phi(\eta)\right) . \tag{E.1.6}$$

The divergence of the original integral is encoded in the function $G(\eta,\kappa^2)$, which can be readily evaluated:

$$G\left(\eta,\kappa^2\right) = \log\left(4\eta(1-\eta)\frac{1}{\kappa^2}\right) + \mathcal{O}\left(\kappa^2\right) . \tag{E.1.7}$$

The integral on the RHS of eq. (E.1.6) is convergent for $\kappa \to 0$, and we have

$$\begin{aligned}
&\int_0^1 d\xi \, \frac{1}{\sqrt{(\eta-\xi)^2+\kappa^2}} \left(\phi(\xi)-\phi(\eta)\right) = \\
&\quad = \int_0^1 d\xi \, \frac{1}{|\xi-\eta|} \left(\phi(\xi)-\phi(\eta)\right) + \mathcal{O}(\kappa^2) \\
&\quad = \int_0^1 d\xi \, \frac{1}{|\xi-\eta|_+} \phi(\xi) + \mathcal{O}(\kappa^2) .
\end{aligned} \tag{E.1.8}$$

Therefore, collecting both contributions,

$$\frac{1}{\sqrt{(\eta-\xi)^2+\kappa^2}} = \delta(\eta-\xi)\log\left(4\eta(1-\eta)\frac{1}{\kappa^2}\right) + \frac{1}{|\eta-\xi|_+} + \mathcal{O}\left(\kappa^2\right) . \quad \text{(E.1.9)}$$

E.2 equivalence between pseudo- and quasi-PDF approaches

As discussed at the end of sec. 7.3, taking the small-z_3^2 limit in position space is equivalent to taking the large-P_3 limit in momentum space. This can be verified at 1-loop by showing that the coefficent functions of eqs. (7.41) and (7.54) are related through a Fourier transform, as stated in eq. (7.55). Here we report the details of the computation. Taking the Fourier transform of the small-z_3^2 coefficient of eq.(7.41) and defining $\eta = y/x$ we have

$$\begin{aligned}\frac{P_3}{2\pi}\int_{-\infty}^{\infty} dz_3\, e^{-iyP_3z_3}\,\tilde{C}\left(x\nu,\mu^2z_3^2\right) &= \frac{1}{x}\int_{-\infty}^{\infty}\frac{d\theta}{2\pi}\,e^{-i\theta\eta}\,\tilde{C}\left(\theta,\frac{\mu^2\theta^2}{x^2P_3^2}\right)\\ &= \frac{1}{x}\Bigg[\delta(\eta-1)+\alpha\log\frac{4\left(xP_3\right)^2}{\mu^2e^{2\gamma_E}}\int_0^1 d\xi\,\delta(\xi-\eta)(1-\xi)\\ &\qquad -\alpha\int_0^1 d\xi\,(1-\xi)\int_{-\infty}^{\infty}\frac{d\theta}{2\pi}e^{-i(\eta-\xi)\theta}\,\log\theta^2\Bigg] . \end{aligned} \quad \text{(E.2.1)}$$

Following ref. [176], the Fourier transform of $\log\theta^2$ can be defined as

$$\begin{aligned}\int\frac{d\theta}{2\pi}e^{-it\theta}\log\theta^2 &= \left[\frac{d}{d\tau}\int\frac{d\theta}{2\pi}e^{-it\theta}\left(\theta^2\right)^\tau\right]_{\tau=0}\\ &= -2\gamma_E\,\delta(t) - \frac{\theta\left(1-|t|\right)}{|t|_{(+0)}} - \frac{\theta\left(|t|-1\right)}{|t|_{(+\infty)}} + \frac{1}{(t)^2}\,\delta\left(\frac{1}{|t|}\right) , \end{aligned} \quad \text{(E.2.2)}$$

with

$$\frac{1}{|t|}_{(+0)} = \lim_{a\to0}\left[\frac{\theta\left(|t|-a\right)}{|t|}+\delta\left(|t|-a\right)\log a\right] , \quad \text{(E.2.3)}$$

$$\frac{1}{|t|}_{(+\infty)} = \frac{1}{(t)^2}\lim_{a\to0}\left[\theta\left(\frac{1}{|t|}-a\right)|t|+\delta\left(\frac{1}{|t|}-a\right)\log a\right] , \quad \text{(E.2.4)}$$

$$\delta\left(\frac{1}{|t|}\right) = \lim_{a\to0}\delta\left(\frac{1}{|t|}-a\right) . \quad \text{(E.2.5)}$$

The proof of eq. (E.2.2) can be found, for example, in the appendix A and C of ref. [176], to which we refer for more details. Setting $t = \eta - \xi$ and plugging

everything in eq. (E.2.1), remembering that $\xi \in [0,1]$, we get different answers depending on the value of η. For $\eta \in [0,1]$, just the first two terms in eq. (E.2.2) contribute, giving

$$\int_0^1 d\xi \left[2\gamma_E \, \delta(\eta - \xi) - \lim_{a \to 0} \left(\frac{\theta(|\eta - \xi| - \beta)}{|\eta - \xi|} + \delta(|\eta - \xi| - a) \log a \right) \right] (1 - \xi)$$
$$= \log e^{2\gamma_E} (1 - \eta) + (1 - \eta) \log \eta (1 - \eta) + 2\eta - 1 \,, \tag{E.2.6}$$

while for $\eta > 1$ or $\eta < 0$ the third contribution in eq. (E.2.2) gives simply

$$- \int_0^1 d\xi \; (1 - \xi) \, \frac{|\eta - \xi|}{(\eta - \xi)^2} \,. \tag{E.2.7}$$

Looking at the last term in eq. (E.2.2), considering its contribution to the convolution integral with the PDF and doing the integral over x first we find

$$\lim_{a \to 0} \int_0^1 \frac{dx}{x} \int_0^1 d\xi \; (1 - \xi) \, \delta \left(\frac{1}{|\frac{y}{x} - \xi|} - a \right) f(x) \propto \lim_{a \to 0} a^2 f(a) = 0. \tag{E.2.8}$$

Using eqs. (E.2.6), (E.2.7), (E.2.8) in eq. (E.2.1) we find back the expression for $C\left(\eta, \frac{\mu^2}{x^2 P_3^2}\right)$ as in eq. (F.0.1), which completes our check.

E.3 quasi-PDFs and their moments

As mentioned at the beginning of chapter 7, the works where the concept of quasi-PDF was first introduced have been criticized in refs. [191, 192], where it was argued that such approach does not give access to the full nonperturbative PDF. In support of their argument, the Authors have shown that moments of quasi-PDFs are divergent: since the moments of parton distributions should reproduce the (finite) matrix elements of the renormalized local DIS operator, they conclude that the quasi-PDF cannot be considered as an euclidean generalization of the light-cone PDF. The problem has been addressed in several independent papers, see e.g. refs. [165, 169, 193]. In this appendix we revise these criticisms in the framework of the scalar model: first we show how the points raised in ref. [191, 192] can be easily seen and understood within the toy model presented in chapter 7, showing explicitly how all the moments of quasi-PDFs are indeed divergent; second we discuss how such feature does not invalidate the programme presented in sec. 7.5, based on the determination of a parametric form of the

light-cone PDF based on a discrete set of data for the euclidean matrix element.

We start this section by computing the moments of the quasi-PDF. From eq. (7.31), using the integral representation of the Bessel function, the $\mathcal{O}(\alpha)$ contribution to the euclidean matrix element reads

$$\hat{\mathcal{M}}^{(1)}\left(\nu, -z_3^2\right) = \alpha \int_0^1 d\xi\ (1-\xi) \int_0^\infty \frac{dT}{T} e^{-T} e^{-\frac{z_3^2 M^2}{4T}} e^{-i\xi P_3 z_3}\,. \tag{E.3.1}$$

The corresponding contribution to the quasi-PDF is found by taking the Fourier transform of the expression above:

$$\hat{q}^{(1)}(y) = \frac{P_3}{2\pi} \int_{-\infty}^{\infty} dz_3\, e^{-iyP_3 z_3} \hat{\mathcal{M}}^{(1)}\left(\nu, -z_3^2\right) \tag{E.3.2}$$

$$= \alpha \frac{P_3}{\sqrt{\pi}} \int_0^1 d\xi\ (1-\xi) \frac{1}{M} \int_0^\infty \frac{dT}{\sqrt{T}} e^{-T} e^{-T(y+\xi)^2 \frac{P_3^2}{M^2}}\,, \tag{E.3.3}$$

where in the last line we have computed the gaussian integral over z_3. Taking the n-th moment of $\hat{q}^{(1)}(y)$ yields

$$\int_{-\infty}^{\infty} dy\, y^n \hat{q}^{(1)}(y) = \alpha \frac{P_3}{\sqrt{\pi}} \int_0^1 d\xi\ (1-\xi) \frac{1}{M} \int_0^\infty \frac{dT}{\sqrt{T}} e^{-T} \int_{-\infty}^{\infty} dy\ (y-\xi)^n\, e^{-Ty^2 \frac{P_3^2}{M^2}}\,. \tag{E.3.4}$$

We can expand the polynomial term as

$$(y-\xi)^n = \sum_{k=0}^{n} \binom{k}{n} y^{n-k} \xi^k \tag{E.3.5}$$

and evaluate each contribution in turn. The term with $k = n$, performing the integral over y first, yields

$$\alpha \frac{P_3}{\sqrt{\pi}} \int_0^1 d\xi\ (1-\xi)\, \xi^n \frac{1}{M} \int_0^\infty \frac{dT}{\sqrt{T}}\, e^{-T} \int_{-\infty}^{\infty} dy\, e^{-Ty^2 \frac{P_3^2}{M^2}} \tag{E.3.6}$$

$$= \alpha \int_0^1 d\xi\ (1-\xi)\, \xi^n \int_0^\infty \frac{dT}{T} e^{-T}\,. \tag{E.3.7}$$

The integral over T is divergent, with the divergence originating from the lower end of the integration region, i.e. when $T \to 0$. Introducing a cutoff a^2 for small values of T [1] and considering the limit $a^2 \to 0$, we get the logarithmic divergent

[1] The cutoff a has dimensions of length and can be thought of as a lattice spacing if the theory were regulated on a lattice.

contribution

$$\alpha \int_0^1 d\xi\ (1-\xi)\,\xi^n \int_{a^2}^{\infty} \frac{dT}{T} e^{-T} \ \overset{a^2\to 0}{\sim}\ -\alpha \int_0^1 d\xi\ (1-\xi)\,\xi^n \log a^2\,. \tag{E.3.8}$$

Similarly we can consider contributions coming from even values of $n-k$. Using

$$\begin{aligned}\int_{-\infty}^{\infty} dy\, y^{2m} e^{-Ty^2\frac{P_3^2}{M^2}} &= \frac{M}{P_3}\left(-\frac{M^2}{P_3^2}\frac{d}{dT}\right)^m \int_{-\infty}^{\infty} dy\, e^{-Ty^2} \\ &= \frac{M\sqrt{\pi}}{P_3}\left(-\frac{M^2}{P_3^2}\frac{d}{dT}\right)^m \frac{1}{\sqrt{T}} \propto \frac{M\sqrt{\pi}}{P_3}\frac{1}{T^{m+\frac{1}{2}}}\,,\end{aligned} \tag{E.3.9}$$

and considering $n-k=2m$, we get

$$\begin{aligned}&\alpha\,\frac{P_3}{\sqrt{\pi}} \int_0^1 d\xi\ (1-\xi)\,\xi^{n-2m}\frac{1}{M}\int_0^{\infty}\frac{dT}{\sqrt{T}}\,e^{-T}\int_{-\infty}^{\infty} dy\, y^{2m} e^{-Ty^2\frac{P_3^2}{M^2}} \\ &\qquad \propto \alpha \int_0^1 d\xi\ (1-\xi)\,\xi^{n-2m}\int_{a^2}^{\infty}\frac{dT}{T^{m+1}}\,e^{-T} \\ &\qquad \overset{a^2\to 0}{\sim}\ \alpha \int_0^1 d\xi\ (1-\xi)\,\xi^{n-2m}\frac{1}{m}\left(\frac{1}{a^2}\right)^m\,,\end{aligned} \tag{E.3.10}$$

where again we have introduced a cutoff a^2 for small values of T and considered the limit $a^2 \to 0$. Contributions from odd values of $n-k$ vanish. Looking at eqs. (E.3.8), (E.3.10) it is then clear that all the moments of the quasi-PDFs will be at least logarithmically divergent with the cutoff a^2, with higher moments affected by higher power divergences.

This relatively simple calculation shows that we obtain divergent contributions for the moments of the quasi-PDF and therefore quasi-PDFs cannot be considered as the proper euclidean generalization of the light-cone parton distribution. This, however, does not invalidate the approach described in sec. 7.5: as mentioned, what really matters is the existence of a factorization theorem connecting the collinear PDF with a renormalizable quantity that can be computed on the lattice, which in our case will be the euclidean matrix element of eq. (7.35), computed for fixed values of P_3 and z_3. As long as z_3 is kept small and different from 0, the factorization formula (7.40) holds, and can be used to fit the light-cone PDF using the available lattice data. How well such data can constrain the PDF is something which should be investigated, just as in the same way the constraints from new experimental measurements are usually analyzed.

F

PDFs from quasi-PDFs matrix elements: matching coefficient and lattice convolution

As detailed at the end of sec. 8.2.1, the matching coefficients to be used to relate the data of refs. [181, 202] to the light-cone PDFs are those expressed in the M$\overline{\text{MS}}$ scheme. Their explicit expression is given by [181, 202]

$$C_3\left(\xi,\eta\left(\xi\right)\right)=\delta\left(1-\xi\right)+C_3^{\text{NLO}}\left(\xi,\eta\left(\xi\right)\right)_+ ,$$

$$C_3^{\text{NLO}}\left(\xi,\eta\left(\xi\right)\right)_+=\frac{\alpha_s}{2\pi}C_F\begin{cases}\left[\frac{1+\xi^2}{1-\xi}\log\frac{\xi}{\xi-1}+1+\frac{3}{2\xi}\right]_{+(1)}^{[1,+\infty]} & \xi>1\\ \left[\frac{1+\xi^2}{1-\xi}\log\left[\frac{1}{\eta^2(\xi)}\left(4\xi\left(1-\xi\right)\right)\right]-\frac{\xi(1+\xi)}{1-\xi}\right]_{+(1)}^{[0,1]} & 0<\xi<1\\ \left[-\frac{1+\xi^2}{1-\xi}\log\frac{\xi}{\xi-1}-1+\frac{3}{2(1-\xi)}\right]_{+(1)}^{[-\infty,0]} & \xi<0\end{cases} . \tag{F.0.1}$$

where the superscripts indicate the domain over which the plus prescription acts. The matching coefficients relate the light-cone PDF to the quasi-PDF up to power suppressed terms according to

$$\tilde{f}_3\left(xP_z,\mu^2\right)=\int_{-1}^{1}\frac{dy}{|y|}C_3\left(\frac{x}{y},\frac{\mu}{yP_z}\right)f_3\left(y,\mu^2\right). \tag{F.0.2}$$

In the following, we work out the full expression of the coefficients appearing in eqs. (8.14), (8.15). Starting from eq. (F.0.2) we have

$$\tilde{f}_3\left(x,\mu^2,P_z\right)=\int_{-1}^{1}\frac{dy}{|y|}\delta\left(1-\frac{x}{y}\right)f_3\left(y,\mu^2\right)+\int_{-1}^{1}\frac{dy}{|y|}C_3^{\text{NLO}}\left(\frac{x}{y},\frac{\mu}{yP_z}\right)_+ f_3\left(y,\mu^2\right) . \tag{F.0.3}$$

Let us focus on the next-to-leading order term, making the plus distribution explicit. In order to do so, we find it useful to split the integral in the two contributions for $y < 0$ and $y > 0$. A change of variables, $\frac{x}{y} = \xi$, yields

$$\begin{aligned}\tilde{f}_3^{\mathrm{NLO}}\left(x,\mu^2,P_z\right) &\equiv \int_{-1}^{1}\frac{dy}{|y|}\,C_3^{\mathrm{NLO}}\left(\frac{x}{y},\frac{\mu}{yP_z}\right)_+ f_3\left(y,\mu^2\right)\\ &= \int_{|x|}^{\infty} d\xi\, C_3^{\mathrm{NLO}}\left(\xi,\frac{\mu\xi}{xP_z}\right)_+ \frac{1}{|\xi|} f_3\left(\frac{x}{\xi},\mu^2\right) +\\ &\quad + \int_{-\infty}^{-|x|} d\xi\, C_3^{\mathrm{NLO}}\left(\xi,\frac{\mu\xi}{xP_z}\right)_+ \frac{1}{|\xi|} f_3\left(\frac{x}{\xi},\mu\right). \end{aligned} \tag{F.0.4}$$

The plus distribution appearing in the matching coefficients is defined in ref. [176] and implemented as follows

$$\int_D d\xi\, C\left(\xi, g\left(\xi\right)\right)_+ f\left(\xi\right) = \int_D d\xi\, \left[C\left(\xi, g\left(\xi\right)\right) f\left(\xi\right) - C\left(\xi, g\left(1\right)\right) f\left(1\right)\right], \tag{F.0.5}$$

with $g\left(\xi\right) = \frac{\mu\xi}{xP_z}$ and D representing a generic integration domain, which in our case will be, according to eq. (F.0.4), either $(-\infty, -|x|)$ or $(|x|, +\infty)$. It follows

$$\begin{aligned}\tilde{f}_3^{\mathrm{NLO}}\left(x,\mu^2,P_z\right) &= \int_{|x|}^{\infty} d\xi\, \left[C_3^{\mathrm{NLO}}\left(\xi,\frac{\mu\xi}{xP_z}\right)\frac{f_3\left(\frac{x}{\xi},\mu^2\right)}{|\xi|} - C_3^{\mathrm{NLO}}\left(\xi,\frac{\mu}{xP_z}\right) f_3\left(x,\mu^2\right)\right]\\ &\quad + \int_{-\infty}^{-|x|} d\xi\, \left[C_3^{\mathrm{NLO}}\left(\xi,\frac{\mu\xi}{xP_z}\right)\frac{f_3\left(\frac{x}{\xi},\mu^2\right)}{|\xi|} - C_3^{\mathrm{NLO}}\left(\xi,\frac{\mu}{xP_z}\right) f_3\left(x,\mu^2\right)\right]. \end{aligned} \tag{F.0.6}$$

It can be easily verified that the two contributions appearing in the above equation are indeed well defined for every fixed x: the singularity in $\xi = +1$ is cured by the plus prescription, while for $\xi \to \pm\infty$ the matching coefficient behaves like $C\left(\xi\right) \sim \frac{1}{\xi^2}$, which is enough to guarantee the convergence of all the integrals above. For numerical stability we find it useful to avoid the singularity in $\xi = +1$ introducing a suitable small parameter $\delta \sim 10^{-6}$, and rewriting the above

equation as

$$\tilde{f}_3^{\rm NLO}\left(x,\mu^2,P_z\right) = \int_{|x|}^{1-\delta} d\xi\, C_3^{\rm NLO}\left(\xi,\frac{\mu\xi}{xP_z}\right)\frac{f_3\left(\frac{x}{\xi},\mu\right)}{\xi} - f_3\left(x,\mu\right)\int_{|x|}^{1-\delta} d\xi\, C_3^{\rm NLO}\left(\xi,\frac{\mu}{xP_z}\right)$$
$$+ \int_{1+\delta}^{\infty} d\xi\, C_3^{\rm NLO}\left(\xi,\frac{\mu\xi}{xP_z}\right)\frac{f_3\left(\frac{x}{\xi},\mu\right)}{\xi} - f_3\left(x,\mu\right)\int_{1+\delta}^{\infty} d\xi\, C_3^{\rm NLO}\left(\xi,\frac{\mu}{xP_z}\right)$$
$$- \int_{-\infty}^{-|x|} d\xi\, C_3^{\rm NLO}\left(\xi,\frac{\mu\xi}{xP_z}\right)\frac{f_3\left(\frac{x}{\xi},\mu\right)}{\xi} - f_3\left(x,\mu\right)\int_{-\infty}^{-|x|} d\xi\, C_3^{\rm NLO}\left(\xi,\frac{\mu}{xP_z}\right). \tag{F.0.7}$$

In order to obtain the lattice ME, we need to compute the real and imaginary part of the Fourier transform of eq. (F.0.3) as shown in eq. (8.11). Starting from the leading-order contribution, we get

$$\int_{-\infty}^{\infty} dx\, \cos\left(xP_z z\right)\int_{-1}^{1}\frac{dy}{|y|}\,\delta\left(1-\frac{x}{y}\right) f_3\left(y,\mu^2\right)$$
$$= \int_0^1 dy \cos\left(yP_z z\right)\left(f_3\left(y,\mu^2\right)+f_3\left(-y,\mu^2\right)\right) = \int_0^1 dx \cos\left(xP_z z\right) V_3\left(x,\mu^2\right)$$
$$= \int_0^1 dx\, \mathrm{A}^{\rm Re,\,LO}\left(xP_z z\right) V_3\left(x,\mu^2\right) \tag{F.0.8}$$

where we have integrated in x first, re-expressed the integral $\int_{-1}^{1} dy$ as $\int_0^1 dy$, used

$$f_3\left(x\right)+f_3\left(-x\right) = f_3^{\rm sym}\left(x\right) = u^-\left(x\right) - d^-\left(x\right) = V_3\left(x\right) \tag{F.0.9}$$

and finally changed variables back to x. Moving now to the next-to-leading order part, we analyse each of the six contributions listed in eq. (F.0.7), defining for each lattice observable six integrals to be computed, denoted as $\mathrm{I}_i^{\rm Re}$, $\mathrm{I}_i^{\rm Im}$ $i = 1,..,6$. Starting from the first contribution to the real part we get

$$\mathrm{I}_1^{\rm Re} = \int_{-\infty}^{\infty} dx\, \cos\left(xP_z z\right)\int_{|x|}^{1-\delta} d\xi\, C_3^{\rm NLO}\left(\xi,\frac{\mu\,\xi}{xP_z}\right)\frac{f_3\left(\frac{x}{\xi},\mu\right)}{\xi}$$
$$= \int_0^{\infty} dx\, \cos\left(xP_z z\right)\int_x^{1-\delta}\frac{d\xi}{\xi}\, C_3^{\rm NLO}\left(\xi,\frac{\mu\,\xi}{xP_z}\right)\left(f_3\left(\frac{x}{\xi},\mu\right)+f_3\left(-\frac{x}{\xi},\mu\right)\right)$$
$$= \int_0^1 dx\, \cos\left(xP_z z\right)\int_{x/(1-\delta)}^{1}\frac{dy}{y}\, C_3^{\rm NLO}\left(\frac{x}{y},\frac{\mu}{yP_z}\right) V_3\left(y,\mu\right), \tag{F.0.10}$$

where in the last line we have changed variables back to $\frac{x}{\xi} = y$. Also, the integration range for x becomes $(0,1)$, since $x < y < 1$. Renaming variables, we

have

$$\begin{aligned}\mathrm{I}_1^{\mathrm{Re}} &= \int_0^1 dx \left[\frac{1}{x}\int_0^1 dy\,\Theta\left(x-\frac{y}{1-\delta}\right)\cos\left(yP_z z\right) C_3^{\mathrm{NLO}}\left(\frac{y}{x},\frac{\mu}{xP_z}\right)\right] V_3\left(x,\mu^2\right) \\ &= \int_0^1 dx\,\mathrm{A}_1^{\mathrm{Re,\,NLO}}\left(x,z,\frac{\mu}{P_z}\right) V_3\left(x,\mu^2\right)\,. \end{aligned} \tag{F.0.11}$$

Analogously, we find out that the other five contributions can be written as

$$\mathrm{I}_i^{\mathrm{Re}} = \int_0^1 dx\,\mathrm{A}_i^{\mathrm{Re,\,NLO}}\left(x,z,\frac{\mu}{P_z}\right) V_3\left(x,\mu^2\right). \tag{F.0.12}$$

with

$$\mathrm{A}_2^{\mathrm{Re,\,NLO}} = \cos(xP_z z)\int_x^{1-\delta} d\xi\, C_3^{\mathrm{NLO}}\left(\xi,\frac{\mu}{xP_z}\right), \tag{F.0.13}$$

$$\mathrm{A}_3^{\mathrm{Re,\,NLO}} = \frac{1}{x}\int_0^{\infty} dy\,\Theta\left(\frac{y}{1+\delta}-x\right)\cos\left(yP_z z\right) C_3^{\mathrm{NLO}}\left(\frac{y}{x},\frac{\mu}{xP_z}\right), \tag{F.0.14}$$

$$\mathrm{A}_4^{\mathrm{Re,\,NLO}} = \cos(xP_z z)\int_{1+\delta}^{\infty} d\xi\, C_3^{\mathrm{NLO}}\left(\xi,\frac{\mu}{xP_z}\right), \tag{F.0.15}$$

$$\mathrm{A}_5^{\mathrm{Re,\,NLO}} = -\frac{1}{x}\int_0^{\infty} dy\,\cos\left(yP_z z\right) C_3^{\mathrm{NLO}}\left(-\frac{y}{x},\frac{\mu}{xP_z}\right), \tag{F.0.16}$$

$$\mathrm{A}_6^{\mathrm{Re,\,NLO}} = \cos(xP_z z)\int_{-\infty}^{-x} d\xi\, C_3^{\mathrm{NLO}}\left(\xi,\frac{\mu}{xP_z}\right) \tag{F.0.17}$$

Collecting all the terms yields eq. (8.14)

$$\mathcal{O}_{\gamma^0}^{\mathrm{Re}}\left(z,\mu\right) = \int_0^1 dx\,\mathcal{C}_3^{\mathrm{Re}}\left(x,z,\frac{\mu}{P_z}\right)\mathrm{V}_3\left(x,\mu^2\right), \tag{F.0.18}$$

where

$$\mathcal{C}_3^{\mathrm{Re}}\left(x,z,\frac{\mu}{P_z}\right) = \mathrm{A}^{\mathrm{Re,\,LO}} + \mathrm{A}^{\mathrm{Re,\,NLO}} \tag{F.0.19}$$

with

$$\mathrm{A}^{\mathrm{Re,\,NLO}} = \mathrm{A}_1^{\mathrm{Re,\,NLO}} - \mathrm{A}_2^{\mathrm{Re,\,NLO}} + \mathrm{A}_3^{\mathrm{Re,\,NLO}} - \mathrm{A}_4^{\mathrm{Re,\,NLO}} - \mathrm{A}_5^{\mathrm{Re,\,NLO}} - \mathrm{A}_6^{\mathrm{Re,\,NLO}}. \tag{F.0.20}$$

We now turn to the imaginary part of the Fourier transform. The computation is exactly the same as in the previous case, with the only difference that now we have a sin instead of a cos. Because of this, when re-expressing the integral

$\int_{-\infty}^{\infty} dx$ as $\int_{0}^{\infty} dx$, we get an additional minus sign, which gives the combination

$$f(x) - f(-x) = f_3^{\text{asy}}(x) = u^+(x) - d^+(x) = T_3(x). \tag{F.0.21}$$

Therefore, the results for the imaginary part can be obtained from those for the real part simply by replacing cos with sin and V_3 with T_3.

G

PDFs from reduced pseudo-ITD data: Pion Mass dependence for 170 ensemble

Similarly to what done for the fine ensemble in sec. 9.3.1, the data for the ensemble 280 presented in ref. [200] can also be used to estimate pion mass effects for results concerning the ensemble 170. The corresponding polynomial curves are plotted in fig. G.1 as functions of the Ioffe-time.

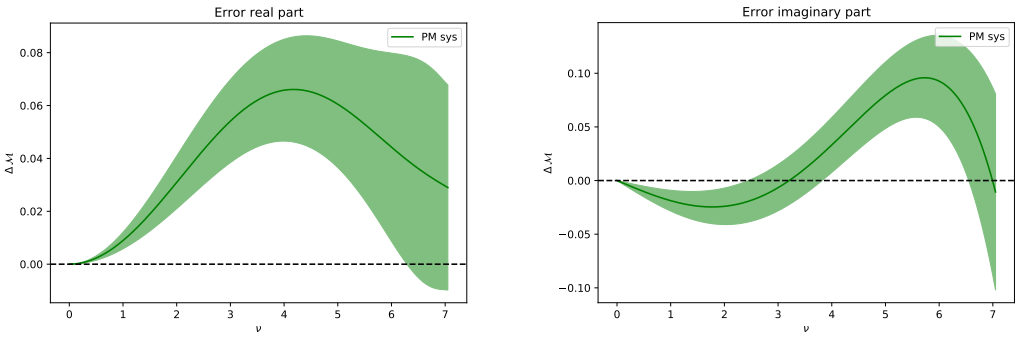

Figure G.1 *Pion mass (PM) systematic provided as functions of the ioffe-time ν for the real (left) and imaginary (right) part of the matrix element.*

As in the case of the analysis for the fine ensemble, the curves in fig. G.1 are used to define a source of correlated systematic. The resulting PDFs, denoted as *170-sys*, are plotted in fig. G.2 together with the results for the ensemble 170 presented in sec. 9.2, where only statistical uncertainties have been considered. From fig. G.2 it is clear how introducing pion mass systematic effects in the analysis has very little impact on the distributions, the major effect being a mild down shift of the central value of V_3 in the medium x region. We conclude that the mild pion mass dependence observed in pseudo-ITD data of ref. [200] has no

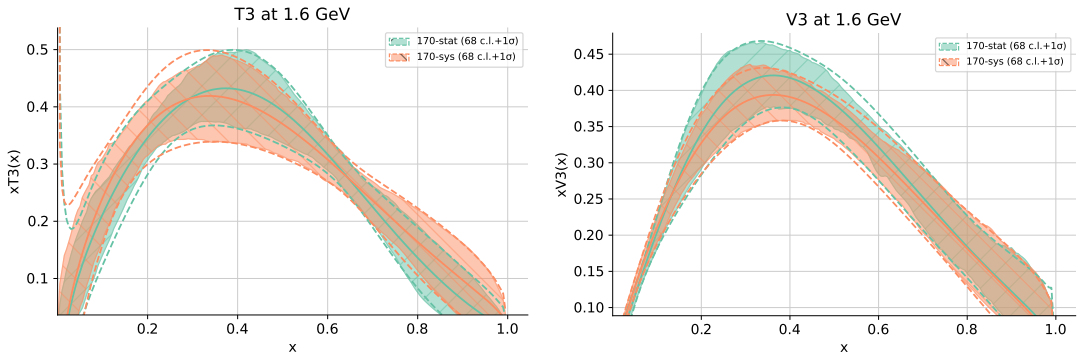

Figure G.2 *PDFs from the fits 170-stat and 170-sys.*

sizable impact on the final PDFs.

Bibliography

[1] R. Ellis, W. Stirling, and B. Webber, *QCD and collider physics*, vol. 8. Cambridge University Press, 2, 2011.

[2] T. Muta, *Foundations of Quantum Chromodynamics: An Introduction to Perturbative Methods in Gauge Theories, (3rd ed.)*, vol. 78 of *World scientific Lecture Notes in Physics*. World Scientific, Hackensack, N.J., 3rd ed., 2010.

[3] J. C. Collins, *Renormalization: An Introduction to Renormalization, The Renormalization Group, and the Operator Product Expansion*, vol. 26 of *Cambridge Monographs on Mathematical Physics*. Cambridge University Press, Cambridge, 1986.

[4] J. C. Collins, *Intrinsic transverse momentum: Nongauge theories*, *Phys. Rev.* **D21** (1980) 2962.

[5] J. C. Collins and D. E. Soper, *Parton Distribution and Decay Functions*, *Nucl. Phys.* **B194** (1982) 445–492.

[6] J. C. Collins, D. E. Soper, and G. F. Sterman, *Factorization of Hard Processes in QCD*, *Adv. Ser. Direct. High Energy Phys.* **5** (1989) 1–91, [hep-ph/0409313].

[7] R. Abdul Khalek et al., *Phenomenology of NNLO jet production at the LHC and its impact on parton distributions*, *Eur. Phys. J. C* **80** (2020), no. 8 797, [arXiv:2005.11327].

[8] **NNPDF** Collaboration, R. Abdul Khalek et al., *A first determination of parton distributions with theoretical uncertainties*, *Eur. Phys. J.* **C** (2019) 79:838, [arXiv:1905.04311].

[9] **NNPDF** Collaboration, R. Abdul Khalek et al., *Parton Distributions with Theory Uncertainties: General Formalism and First Phenomenological Studies*, *Eur. Phys. J. C* **79** (2019), no. 11 931, [arXiv:1906.10698].

[10] S. Forte, T. Giani, and D. Napoletano, *Fitting the b-quark PDF as a massive-b scheme: Higgs production in bottom fusion*, *Eur. Phys. J. C* **79** (2019), no. 7 609, [arXiv:1905.02207].

[11] L. Del Debbio, T. Giani, and C. J. Monahan, *Notes on lattice observables for parton distributions: nongauge theories*, *JHEP* **09** (2020) 021, [arXiv:2007.02131].

[12] K. Cichy, L. Del Debbio, and T. Giani, *Parton distributions from lattice data: the nonsinglet case*, *JHEP* **10** (2019) 137, [arXiv:1907.06037].

[13] L. Del Debbio, T. Giani, J. Karpie, K. Orginos, A. Radyushkin, and S. Zafeiropoulos, *Neural-network analysis of Parton Distribution Functions from Ioffe-time pseudodistributions*, *JHEP* **02** (2021) 138, [arXiv:2010.03996].

[14] G. 't Hooft, *When was asymptotic freedom discovered? or the rehabilitation of quantum field theory*, *Nucl. Phys. B Proc. Suppl.* **74** (1999) 413–425, [hep-th/9808154].

[15] J. D. Bjorken, *Asymptotic sum rules at infinite momentum*, *Phys. Rev.* **179** (Mar, 1969) 1547–1553.

[16] R. P. Feynman, *Very high-energy collisions of hadrons*, *Phys. Rev. Lett.* **23** (Dec, 1969) 1415–1417.

[17] H. D. Politzer, *Reliable perturbative results for strong interactions?*, *Phys. Rev. Lett.* **30** (Jun, 1973) 1346–1349.

[18] D. J. Gross and F. Wilczek, *Ultraviolet behavior of non-abelian gauge theories*, *Phys. Rev. Lett.* **30** (Jun, 1973) 1343–1346.

[19] S. Coleman and D. J. Gross, *Price of asymptotic freedom*, *Phys. Rev. Lett.* **31** (Sep, 1973) 851–854.

[20] V. N. Gribov, *Quantization of Nonabelian Gauge Theories*, *Nucl. Phys. B* **139** (1978) 1.

[21] L. Faddeev and V. Popov, *Feynman Diagrams for the Yang-Mills Field*, *Phys. Lett. B* **25** (1967) 29–30.

[22] C. Becchi, A. Rouet, and R. Stora, *Renormalization of Gauge Theories*, *Annals Phys.* **98** (1976) 287–321.

[23] I. Tyutin, *Gauge Invariance in Field Theory and Statistical Physics in Operator Formalism*, arXiv:0812.0580.

[24] M. Gell-Mann, *A Schematic Model of Baryons and Mesons*, *Phys. Lett.* **8** (1964) 214–215.

[25] S. Weinberg, *The U(1) Problem*, *Phys. Rev. D* **11** (1975) 3583–3593.

[26] R. Peccei and H. R. Quinn, *Constraints Imposed by CP Conservation in the Presence of Instantons*, *Phys. Rev. D* **16** (1977) 1791–1797.

[27] S. Weinberg, *A new light boson?*, *Phys. Rev. Lett.* **40** (Jan, 1978) 223–226.

[28] F. Wilczek, *Problem of strong p and t invariance in the presence of instantons*, *Phys. Rev. Lett.* **40** (Jan, 1978) 279–282.

[29] F. Herzog, B. Ruijl, T. Ueda, J. Vermaseren, and A. Vogt, *The five-loop beta function of Yang-Mills theory with fermions*, *JHEP* **02** (2017) 090, [arXiv:1701.01404].

[30] J. Smit, *Introduction to quantum fields on a lattice*, .

[31] K. Osterwalder and R. Schrader, *Axioms for Euclidean Green's functions*, *Communications in Mathematical Physics* **31** (1973), no. 2 83 – 112.

[32] H. B. Nielsen and M. Ninomiya, *No Go Theorem for Regularizing Chiral Fermions*, *Phys. Lett. B* **105** (1981) 219–223.

[33] K. G. Wilson, *Confinement of quarks*, *Phys. Rev. D* **10** (Oct, 1974) 2445–2459.

[34] R. Feynman, *Photon-hadron interactions*, .

[35] J. D. Bjorken and E. A. Paschos, *Inelastic Electron Proton and gamma Proton Scattering, and the Structure of the Nucleon*, *Phys. Rev.* **185** (1969) 1975–1982.

[36] A. Vogt, *Efficient evolution of unpolarized and polarized parton distributions with QCD-PEGASUS*, *Comput. Phys. Commun.* **170** (2005) 65–92, [hep-ph/0408244].

[37] G. Altarelli and G. Parisi, *Asymptotic freedom in parton language*, *Nuclear Physics B* **126** (1977), no. 2 298 – 318.

[38] M. Aivazis, J. C. Collins, F. I. Olness, and W.-K. Tung, *Leptoproduction of heavy quarks. 2. A Unified QCD formulation of charged and neutral current processes from fixed target to collider energies*, *Phys. Rev. D* **50** (1994) 3102–3118, [hep-ph/9312319].

[39] M. Aivazis, F. I. Olness, and W.-K. Tung, *Leptoproduction of heavy quarks. 1. General formalism and kinematics of charged current and neutral current production processes*, *Phys. Rev. D* **50** (1994) 3085–3101, [hep-ph/9312318].

[40] W.-K. Tung, S. Kretzer, and C. Schmidt, *Open heavy flavor production in QCD: Conceptual framework and implementation issues*, *J. Phys. G* **28** (2002) 983–996, [hep-ph/0110247].

[41] M. Krämer, F. I. Olness, and D. E. Soper, *Treatment of heavy quarks in deeply inelastic scattering*, *Phys. Rev. D* **62** (2000) 096007, [hep-ph/0003035].

[42] R. Thorne and R. Roberts, *An Ordered analysis of heavy flavor production in deep inelastic scattering*, *Phys. Rev. D* **57** (1998) 6871–6898, [hep-ph/9709442].

[43] M. Cacciari, M. Greco, and P. Nason, *The P(T) spectrum in heavy flavor hadroproduction*, *JHEP* **9805** (1998) 7, [hep-ph/9803400].

[44] S. Forte, E. Laenen, P. Nason, and J. Rojo, *Heavy quarks in deep-inelastic scattering*, *Nucl.Phys.* **B834** (2010) 116–162, [arXiv:1001.2312].

[45] S. Forte, D. Napoletano, and M. Ubiali, *Higgs production in bottom-quark fusion in a matched scheme*, *Phys. Lett.* **B751** (2015) 331–337, [arXiv:1508.01529].

[46] S. Forte and S. Carrazza, *Parton distribution functions*, arXiv:2008.12305.

[47] S. Forte, L. Garrido, J. I. Latorre, and A. Piccione, *Neural network parametrization of deep inelastic structure functions*, *JHEP* **05** (2002) 062, [hep-ph/0204232].

[48] **NNPDF** Collaboration, R. D. Ball et al., *Parton distributions from high-precision collider data*, *Eur. Phys. J.* **C77** (2017), no. 10 663, [arXiv:1706.00428].

[49] S. Carrazza and J. Cruz-Martinez, *Towards a new generation of parton densities with deep learning models*, *Eur. Phys. J. C* **79** (2019), no. 8 676, [arXiv:1907.05075].

[50] **NNPDF** Collaboration, R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, A. Piccione, J. Rojo, and M. Ubiali, *A Determination of parton distributions with faithful uncertainty estimation*, *Nucl. Phys.* **B809** (2009) 1–63, [arXiv:0808.1231]. [Erratum: Nucl. Phys.B816,293(2009)].

[51] **NNPDF** Collaboration, L. Del Debbio, S. Forte, J. I. Latorre, A. Piccione, and J. Rojo, *Unbiased determination of the proton structure function F(2)**p with faithful uncertainty estimation*, *JHEP* **03** (2005) 080, [hep-ph/0501067].

[52] **NNPDF** Collaboration, L. Del Debbio, S. Forte, J. I. Latorre, A. Piccione, and J. Rojo, *Neural network determination of parton distributions: The Nonsinglet case*, *JHEP* **03** (2007) 039, [hep-ph/0701127].

[53] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali, *A first unbiased global NLO determination of parton distributions and their uncertainties*, *Nucl. Phys.* **B838** (2010) 136–206, [arXiv:1002.4407].

[54] R. D. Ball et al., *Parton distributions with LHC data*, *Nucl. Phys. B* **867** (2013) 244–289, [arXiv:1207.1303].

[55] **NNPDF** Collaboration, R. D. Ball et al., *Parton distributions for the LHC Run II*, *JHEP* **04** (2015) 040, [arXiv:1410.8849].

[56] **NNPDF** Collaboration, R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali, *Fitting Parton Distribution Data*

with Multiplicative Normalization Uncertainties, *JHEP* **05** (2010) 075, [arXiv:0912.2276].

[57] N. Hansen, *The CMA evolution strategy: A tutorial*, *CoRR* **abs/1604.00772** (2016) [arXiv:1604.00772].

[58] **NNPDF** Collaboration, V. Bertone, S. Carrazza, N. P. Hartland, E. R. Nocera, and J. Rojo, *A determination of the fragmentation functions of pions, kaons, and protons with faithful uncertainties*, *Eur. Phys. J.* **C77** (2017), no. 8 516, [arXiv:1706.07049].

[59] V. Bertone, S. Carrazza, and N. P. Hartland, *APFELgrid: a high performance tool for parton density determinations*, *Comput. Phys. Commun.* **212** (2017) 205–209, [arXiv:1605.02070].

[60] F. Chollet et al., *Keras*, 2015.

[61] M. Abadi et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015. Software available from tensorflow.org.

[62] **New Muon** Collaboration, P. Amaudruz et al., *The Gottfried sum from the ratio F2(n) / F2(p)*, *Phys. Rev. Lett.* **66** (1991) 2712–2715.

[63] A. Candido, S. Forte, and F. Hekhorn, *Can* $\overline{\text{MS}}$ *parton distributions be negative?*, *JHEP* **11** (2020) 129, [arXiv:2006.07377].

[64] **New Muon** Collaboration, M. Arneodo et al., *Accurate measurement of F2(d) / F2(p) and R**d - R**p*, *Nucl. Phys. B* **487** (1997) 3–26, [hep-ex/9611022].

[65] **New Muon** Collaboration, M. Arneodo et al., *Measurement of the proton and deuteron structure functions, F2(p) and F2(d), and of the ratio sigma-L / sigma-T*, *Nucl. Phys. B* **483** (1997) 3–43, [hep-ph/9610231].

[66] L. W. Whitlow, E. M. Riordan, S. Dasu, S. Rock, and A. Bodek, *Precise measurements of the proton and deuteron structure functions from a global analysis of the SLAC deep inelastic electron scattering cross-sections*, *Phys. Lett.* **B282** (1992) 475–482.

[67] **BCDMS** Collaboration, A. C. Benvenuti et al., *A High Statistics Measurement of the Proton Structure Functions* $F_2(x, Q^2)$ *and R from Deep Inelastic Muon Scattering at High* Q^2, *Phys. Lett.* **B223** (1989) 485.

[68] **CHORUS** Collaboration, G. Onengut et al., *Measurement of nucleon structure functions in neutrino scattering*, *Phys. Lett.* **B632** (2006) 65–75.

[69] **NuTeV** Collaboration, M. Goncharov et al., *Precise measurement of dimuon production cross-sections in ν_μFe and $\bar{\nu}_\mu$Fe deep inelastic scattering at the Tevatron*, *Phys. Rev.* **D64** (2001) 112006, [hep-ex/0102049].

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

[71] **ZEUS, H1** Collaboration, H. Abramowicz et al., *Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data*, *Eur. Phys. J.* **C75** (2015), no. 12 580, [arXiv:1506.06042].

[72] **H1 , ZEUS** Collaboration, H. Abramowicz et al., *Combination and QCD Analysis of Charm Production Cross Section Measurements in Deep-Inelastic ep Scattering at HERA*, *Eur.Phys.J.* **C73** (2013) 2311, [arXiv:1211.1182].

[73] **H1** Collaboration, F. D. Aaron et al., *Measurement of the Charm and Beauty Structure Functions using the H1 Vertex Detector at HERA*, *Eur. Phys. J.* **C65** (2010) 89–109, [arXiv:0907.2643].

[74] **ZEUS** Collaboration, H. Abramowicz et al., *Measurement of beauty and charm production in deep inelastic scattering at HERA and measurement of the beauty-quark mass*, *JHEP* **09** (2014) 127, [arXiv:1405.6915].

[75] **NuSea** Collaboration, J. C. Webb et al., *Absolute Drell-Yan dimuon cross sections in 800-GeV/c p p and p d collisions*, hep-ex/0302019.

[76] J. C. Webb, *Measurement of continuum dimuon production in 800-GeV/c proton nucleon collisions*, hep-ex/0301031.

[77] **FNAL E866/NuSea** Collaboration, R. S. Towell et al., *Improved measurement of the anti-d/anti-u asymmetry in the nucleon sea*, *Phys. Rev.* **D64** (2001) 052002, [hep-ex/0103030].

[78] G. Moreno et al., *Dimuon production in proton - copper collisions at $\sqrt{s}$ = 38.8-GeV*, *Phys. Rev.* **D43** (1991) 2815–2836.

[79] **CDF** Collaboration, T. A. Aaltonen et al., *Measurement of $d\sigma/dy$ of Drell-Yan e^+e^- pairs in the Z Mass Region from $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV*, *Phys. Lett.* **B692** (2010) 232–239, [arXiv:0908.3914].

[80] **D0** Collaboration, V. M. Abazov et al., *Measurement of the shape of the boson rapidity distribution for $p\bar{p} \to Z/\gamma^* \to e^+e^- + X$ events produced at $\sqrt{s}$=1.96-TeV*, *Phys. Rev.* **D76** (2007) 012003, [hep-ex/0702025].

[81] **D0** Collaboration, V. M. Abazov et al., *Measurement of the muon charge asymmetry in $p\bar{p} \to W+X \to \mu\nu + X$ events at $\sqrt{s}$=1.96 TeV*, *Phys.Rev.* **D88** (2013) 091102, [arXiv:1309.2591].

[82] **D0** Collaboration, V. M. Abazov et al., *Measurement of the electron charge asymmetry in $\boldsymbol{p\bar{p} \to W + X \to e\nu + X}$ decays in $\boldsymbol{p\bar{p}}$ collisions at $\boldsymbol{\sqrt{s} = 1.96}$ TeV*, *Phys. Rev.* **D91** (2015), no. 3 032007, [arXiv:1412.2862]. [Erratum: Phys. Rev.D91,no.7,079901(2015)].

[83] **ATLAS** Collaboration, G. Aad et al., *Measurement of the high-mass Drell–Yan differential cross-section in pp collisions at $\sqrt{s}$=7 TeV with the ATLAS detector*, *Phys.Lett.* **B725** (2013) 223, [arXiv:1305.4192].

[84] **ATLAS** Collaboration, G. Aad et al., *Measurement of the low-mass Drell-Yan differential cross section at $\sqrt{s}$ = 7 TeV using the ATLAS detector*, *JHEP* **06** (2014) 112, [arXiv:1404.1212].

[85] **ATLAS** Collaboration, G. Aad et al., *Measurement of the inclusive $W^\pm$ and Z/γ^* cross sections in the electron and muon decay channels in pp collisions at $\sqrt{s}$= 7 TeV with the ATLAS detector*, *Phys.Rev.* **D85** (2012) 072004, [arXiv:1109.5141].

[86] **ATLAS** Collaboration, M. Aaboud et al., *Precision measurement and interpretation of inclusive W^+, W^- and Z/γ^* production cross sections with the ATLAS detector*, arXiv:1612.03016.

[87] **ATLAS** Collaboration, G. Aad et al., *Measurement of the transverse momentum and ϕ^*_η distributions of Drell–Yan lepton pairs in proton–proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *Eur. Phys. J.* **C76** (2016), no. 5 291, [arXiv:1512.02192].

[88] **ATLAS** Collaboration, G. Aad et al., *Measurement of the $t\bar{t}$ production cross-section using $e\mu$ events with b-tagged jets in pp collisions at $\sqrt{s}$ = 7*

and 8 TeV *with the ATLAS detector*, *Eur. Phys. J.* **C74** (2014), no. 10 3109, [arXiv:1406.5375]. [Addendum: Eur. Phys. J.C76,no.11,642(2016)].

[89] **ATLAS** Collaboration, M. Aaboud et al., *Measurement of the $t\bar{t}$ production cross-section using $e\mu$ events with b-tagged jets in pp collisions at $\sqrt{s}$=13 TeV with the ATLAS detector*, *Phys. Lett.* **B761** (2016) 136–157, [arXiv:1606.02699].

[90] **ATLAS** Collaboration, G. Aad et al., *Measurements of top-quark pair differential cross-sections in the lepton+jets channel in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*, *Eur. Phys. J.* **C76** (2016), no. 10 538, [arXiv:1511.04716].

[91] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the electron charge asymmetry in inclusive W production in pp collisions at $\sqrt{s} = 7$ TeV*, *Phys.Rev.Lett.* **109** (2012) 111806, [arXiv:1206.2598].

[92] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the muon charge asymmetry in inclusive pp to WX production at $\sqrt{s} = 7$ TeV and an improved determination of light parton distribution functions*, *Phys.Rev.* **D90** (2014) 032004, [arXiv:1312.6283].

[93] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the differential and double-differential Drell-Yan cross sections in proton-proton collisions at $\sqrt{s} = 7$ TeV*, *JHEP* **1312** (2013) 030, [arXiv:1310.7291].

[94] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the differential cross section and charge asymmetry for inclusive $\mathrm{pp} \to \mathrm{W}^{\pm} + X$ production at $\sqrt{s} = 8$ TeV*, *Eur. Phys. J.* **C76** (2016), no. 8 469, [arXiv:1603.01803].

[95] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the Z boson differential cross section in transverse momentum and rapidity in proton–proton collisions at 8 TeV*, *Phys. Lett.* **B749** (2015) 187–209, [arXiv:1504.03511].

[96] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the t-tbar production cross section in the e-mu channel in proton-proton collisions at sqrt(s) = 7 and 8 TeV*, *JHEP* **08** (2016) 029, [arXiv:1603.02303].

[97] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the top quark pair production cross section in proton-proton collisions at* $\sqrt{(s)} = 13$ *TeV*, *Phys. Rev. Lett.* **116** (2016), no. 5 052002, [arXiv:1510.05302].

[98] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the differential cross section for top quark pair production in pp collisions at* $\sqrt{s} = 8$ *TeV*, *Eur. Phys. J.* **C75** (2015), no. 11 542, [arXiv:1505.04480].

[99] **LHCb** Collaboration, R. Aaij et al., *Inclusive W and Z production in the forward region at* $\sqrt{s} = 7$ *TeV*, *JHEP* **1206** (2012) 058, [arXiv:1204.1620].

[100] **LHCb** Collaboration, R. Aaij et al., *Measurement of the cross-section for* $Z \to e^+e^-$ *production in pp collisions at* $\sqrt{s} = 7$ *TeV*, *JHEP* **1302** (2013) 106, [arXiv:1212.4620].

[101] **LHCb** Collaboration, R. Aaij et al., *Measurement of the forward Z boson production cross-section in pp collisions at* $\sqrt{s} = 7$ *TeV*, *JHEP* **08** (2015) 039, [arXiv:1505.07024].

[102] **LHCb** Collaboration, R. Aaij et al., *Measurement of forward W and Z boson production in pp collisions at* $\sqrt{s} = 8$ *TeV*, *JHEP* **01** (2016) 155, [arXiv:1511.08039].

[103] **ATLAS** Collaboration, G. Aad et al., *Measurement of the inclusive jet cross-section in proton-proton collisions at* $\sqrt{s} = 7$ *TeV using 4.5* fb^{-1} *of data with the ATLAS detector*, *JHEP* **02** (2015) 153, [arXiv:1410.8857].

[104] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the inclusive jet cross section in pp collisions at* $\sqrt{s} = 2.76$ *TeV*, *Eur. Phys. J. C* **76** (2016), no. 5 265, [arXiv:1512.06212].

[105] S. D. Ellis, Z. Kunszt, and D. E. Soper, *The One Jet Inclusive Cross-section at Order $\alpha^{-}s \hat{} \{3\}$ Quarks and Gluons*, *Phys. Rev. Lett.* **64** (1990) 2121.

[106] F. Aversa, P. Chiappetta, M. Greco, and J. Guillet, *Higher Order Corrections to QCD Jets*, *Phys. Lett. B* **210** (1988) 225.

[107] W. Giele, E. Glover, and D. A. Kosower, *The inclusive two jet triply differential cross-section*, *Phys. Rev. D* **52** (1995) 1486–1499, [hep-ph/9412338].

[108] J. Currie, E. Glover, T. Gehrmann, A. Gehrmann-De Ridder, A. Huss, and J. Pires, *Single Jet Inclusive Production for the Individual Jet p_{T} Scale Choice at the LHC*, *Acta Phys. Polon. B* **48** (2017) 955–967, [arXiv:1704.00923].

[109] J. Currie, A. Gehrmann-De Ridder, T. Gehrmann, E. N. Glover, A. Huss, and J. a. Pires, *Infrared sensitivity of single jet inclusive production at hadron colliders*, *JHEP* **10** (2018) 155, [arXiv:1807.03692].

[110] M. Cacciari, S. Forte, D. Napoletano, G. Soyez, and G. Stagnitto, *Single-jet inclusive cross section and its definition*, *Phys. Rev. D* **100** (2019), no. 11 114015, [arXiv:1906.11850].

[111] E. R. Nocera and M. Ubiali, *Constraining the gluon PDF at large x with LHC data*, *PoS* **DIS2017** (2018) 008, [arXiv:1709.09690].

[112] **ATLAS** Collaboration, M. Aaboud et al., *Measurement of inclusive jet and dijet cross-sections in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *JHEP* **05** (2018) 195, [arXiv:1711.02692].

[113] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the double-differential inclusive jet cross section in proton–proton collisions at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **76** (2016), no. 8 451, [arXiv:1605.04436].

[114] **CMS** Collaboration, A. M. Sirunyan et al., *Dependence of inclusive jet production on the anti-k_{T} distance parameter in pp collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **12** (2020) 082, [arXiv:2005.05159].

[115] **ATLAS** Collaboration, G. Aad et al., *Measurement of three-jet production cross-sections in pp collisions at 7 TeV centre-of-mass energy using the ATLAS detector*, *Eur. Phys. J. C* **75** (2015), no. 5 228, [arXiv:1411.1855].

[116] **ATLAS** Collaboration, G. Aad et al., *Measurement of four-jet differential cross sections in $\sqrt{s} = 8$ TeV proton-proton collisions using the ATLAS detector*, *JHEP* **12** (2015) 105, [arXiv:1509.07335].

[117] **CMS** Collaboration, V. Khachatryan et al., *Measurement of the inclusive 3-jet production differential cross section in proton–proton collisions at 7 TeV and determination of the strong coupling constant in the TeV range*, *Eur. Phys. J. C* **75** (2015), no. 5 186, [arXiv:1412.1633].

[118] **CMS** Collaboration, S. Chatrchyan et al., *Measurements of differential jet cross sections in proton-proton collisions at $\sqrt{s} = 7$ TeV with the CMS detector*, *Phys.Rev.* **D87** (2013) 112002, [arXiv:1212.6660].

[119] **ATLAS** Collaboration, M. Aaboud et al., *Measurement of the inclusive jet cross-sections in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *JHEP* **09** (2017) 020, [arXiv:1706.03192].

[120] **CMS** Collaboration, V. Khachatryan et al., *Measurement and QCD analysis of double-differential inclusive jet cross sections in pp collisions at $\sqrt{s} = 8$ TeV and cross section ratios to 2.76 and 7 TeV*, *JHEP* **03** (2017) 156, [arXiv:1609.05331].

[121] **ATLAS Collaboration** Collaboration, G. Aad et al., *Measurement of dijet cross sections in pp collisions at 7 TeV centre-of-mass energy using the ATLAS detector*, *JHEP* **1405** (2014) 059, [arXiv:1312.3524].

[122] **CMS** Collaboration, A. M. Sirunyan et al., *Measurement of the triple-differential dijet cross section in proton-proton collisions at $\sqrt{s} = 8$ TeV and constraints on parton distribution functions*, *Eur. Phys. J.* **C77** (2017), no. 11 746, [arXiv:1705.02628].

[123] J. Currie, A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, A. Huss, and J. Pires, *Precise predictions for dijet production at the LHC*, *Phys. Rev. Lett.* **119** (2017), no. 15 152001, [arXiv:1705.10271].

[124] J. Currie, A. Gehrmann-De Ridder, T. Gehrmann, N. Glover, A. Huss, and J. Pires, *Jet cross sections at the LHC with NNLOJET*, *PoS* **LL2018** (2018) 001, [arXiv:1807.06057].

[125] A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, A. Huss, and J. Pires, *Triple Differential Dijet Cross Section at the LHC*, *Phys. Rev. Lett.* **123** (2019), no. 10 102001, [arXiv:1905.09047].

[126] Z. Nagy, *Three jet cross-sections in hadron hadron collisions at next-to-leading order*, *Phys. Rev. Lett.* **88** (2002) 122003, [hep-ph/0110315].

[127] **fastNLO** Collaboration, M. Wobisch, D. Britzger, T. Kluge, K. Rabbertz, and F. Stober, *Theory-Data Comparisons for Jet Measurements in Hadron-Induced Processes*, arXiv:1109.1310.

[128] A. Gehrmann-De Ridder, T. Gehrmann, N. Glover, A. Huss, and T. A. Morgan, *NNLO QCD corrections for Z boson plus jet production*, *PoS* **RADCOR2015** (2016) 075, [arXiv:1601.04569].

[129] S. Dittmaier, A. Huss, and C. Speckner, *Weak radiative corrections to dijet production at hadron colliders*, *JHEP* **11** (2012) 095, [arXiv:1210.0438].

[130] Z. Kassabov, "Reportengine: A framework for declarative data analysis." https://doi.org/10.5281/zenodo.2571601, Feb., 2019.

[131] V. Bertone, S. Carrazza, and J. Rojo, *APFEL: A PDF Evolution Library with QED corrections*, *Comput. Phys. Commun.* **185** (2014) 1647–1668, [arXiv:1310.1394].

[132] T. Carli, D. Clements, A. Cooper-Sarkar, C. Gwenlan, G. P. Salam, F. Siegert, P. Starovoitov, and M. Sutton, *A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project*, *Eur. Phys. J. C* **66** (2010) 503–524, [arXiv:0911.2985].

[133] **LHC Higgs Cross Section Working Group** Collaboration, D. de Florian et al., *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, arXiv:1610.07922.

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

[135] S. J. Brodsky, C. Peterson, and N. Sakai, *Intrinsic heavy-quark states*, *Phys. Rev. D* **23** (Jun, 1981) 2745–2757.

[136] F. Maltoni, G. Ridolfi, and M. Ubiali, *b-initiated processes at the LHC: a reappraisal*, *JHEP* **07** (2012) 022, [arXiv:1203.6393]. [Erratum: JHEP 04, 095 (2013)].

[137] M. Lim, F. Maltoni, G. Ridolfi, and M. Ubiali, *Anatomy of double heavy-quark initiated processes*, *JHEP* **09** (2016) 132, [arXiv:1605.09411].

[138] E. Bagnaschi, F. Maltoni, A. Vicini, and M. Zaro, *Lepton-pair production in association with a $b\bar{b}$ pair and the determination of the W boson mass*, *JHEP* **07** (2018) 101, [arXiv:1803.04336].

[139] F. Krauss and D. Napoletano, *Towards a fully massive five-flavor scheme*, *Phys. Rev.* **D98** (2018), no. 9 096002, [arXiv:1712.06832].

[140] D. Figueroa, S. Honeywell, S. Quackenbush, L. Reina, C. Reuschle, and D. Wackeroth, *Electroweak and QCD corrections to Z-boson production with one b jet in a massive five-flavor scheme*, *Phys. Rev.* **D98** (2018), no. 9 093002, [arXiv:1805.01353].

[141] R. D. Ball, V. Bertone, M. Bonvini, S. Forte, P. Groth Merrild, J. Rojo, and L. Rottoli, *Intrinsic charm in a matched general-mass scheme*, *Phys. Lett.* **B754** (2016) 49–58, [arXiv:1510.00009].

[142] R. D. Ball, M. Bonvini, and L. Rottoli, *Charm in Deep-Inelastic Scattering*, *JHEP* **11** (2015) 122, [arXiv:1510.02491].

[143] R. D. Ball, V. Bertone, M. Bonvini, S. Carrazza, S. Forte, A. Guffanti, N. P. Hartland, J. Rojo, and L. Rottoli, *A Determination of the Charm Content of the Proton*, *Eur. Phys. J.* **C76** (2016), no. 11 647, [arXiv:1605.06515].

[144] S. Forte, D. Napoletano, and M. Ubiali, *Higgs production in bottom-quark fusion: matching beyond leading order*, *Phys. Lett.* **B763** (2016) 190–196, [arXiv:1607.00389].

[145] M. Bonvini, A. S. Papanastasiou, and F. J. Tackmann, *Resummation and matching of b-quark mass effects in $b\bar{b}H$ production*, *JHEP* **11** (2015) 196, [arXiv:1508.03288].

[146] M. Bonvini, A. S. Papanastasiou, and F. J. Tackmann, *Matched predictions for the $b\bar{b}H$ cross section at the 13 TeV LHC*, *JHEP* **10** (2016) 53, [arXiv:1605.01733].

[147] S. Dittmaier, *A general approach to photon radiation off fermions*, *Nucl. Phys.* **B565** (2000) 69–122, [9904440].

[148] M. Buza, Y. Matiounine, J. Smith, and W. L. van Neerven, *Charm electroproduction viewed in the variable flavor number scheme versus fixed order perturbation theory*, *Eur. Phys. J.* **C1** (1998) 301–320, [hep-ph/9612398].

[149] R. V. Harlander and W. B. Kilgore, *Next–to-next–to–leading order Higgs production at hadron colliders*, *Phys.Rev.Lett.* **88** (2002) 201801, [hep-ph/0201206].

[150] S. Forte, D. Napoletano, and M. Ubiali, "bbhfonll." http://bbhfonll.hepforge.org/, 2017.

[151] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, *JHEP* **1407** (2014) 79, [arXiv:1405.0301].

[152] M. Wiesemann, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, and P. Torrielli, *Higgs production in association with bottom quarks*, *JHEP* **02** (2015) 132, [arXiv:1409.5301].

[153] A. Buckley, J. Ferrando, S. Lloyd, K. Nordström, B. Page, M. Rüfenacht, M. Schönherr, and G. Watt, *LHAPDF6: parton density access in the LHC precision era*, *Eur. Phys. J.* **C75** (2015) 132, [arXiv:1412.7420].

[154] R. Doria, J. Frenkel, and J. C. Taylor, *Counter Example to Nonabelian Bloch-Nordsieck Theorem*, *Nucl. Phys.* **B168** (1980) 93–110.

[155] S. Catani, M. Ciafaloni, and G. Marchesini, *Non-cancelling infrared divergences in QCD coherent state*, *Nucl. Phys.* **B264** (1986) 588–620.

[156] F. Caola, K. Melnikov, D. Napoletano, and L. Tancredi, *Noncancellation of infrared singularities in collisions of massive quarks*, *Phys. Rev. D* **103** (2021), no. 5 054013, [arXiv:2011.04701].

[157] X. Ji, *Parton Physics from Large-Momentum Effective Field Theory*, *Sci. China Phys. Mech. Astron.* **57** (2014) 1407–1412, [arXiv:1404.6680].

[158] X. Ji, Y.-S. Liu, Y. Liu, J.-H. Zhang, and Y. Zhao, *Large-Momentum Effective Theory*, arXiv:2004.03543.

[159] X. Ji, *Parton physics on a euclidean lattice*, *Phys. Rev. Lett.* **110** (Jun, 2013) 262002.

[160] A. V. Radyushkin, *Quasi-parton distribution functions, momentum distributions, and pseudo-parton distribution functions*, *Phys. Rev.* **D96** (2017), no. 3 034025, [arXiv:1705.01488].

[161] X. Ji and J.-H. Zhang, *Renormalization of quasiparton distribution*, *Phys. Rev.* **D92** (2015) 034006, [arXiv:1505.07699].

[162] T. Ishikawa, Y.-Q. Ma, J.-W. Qiu, and S. Yoshida, *Practical quasi parton distribution functions*, arXiv:1609.02018.

[163] M. Constantinou and H. Panagopoulos, *Perturbative renormalization of quasi-parton distribution functions*, *Phys. Rev.* **D96** (2017), no. 5 054506, [arXiv:1705.11193].

[164] C. Alexandrou, K. Cichy, M. Constantinou, K. Hadjiyiannakou, K. Jansen, H. Panagopoulos, and F. Steffens, *A complete non-perturbative renormalization prescription for quasi-PDFs*, *Nucl. Phys.* **B923** (2017) 394–415, [arXiv:1706.00265].

[165] X. Ji, J.-H. Zhang, and Y. Zhao, *More On Large-Momentum Effective Theory Approach to Parton Physics*, *Nucl. Phys.* **B924** (2017) 366–376, [arXiv:1706.07416].

[166] X. Ji, J.-H. Zhang, and Y. Zhao, *Renormalization in Large Momentum Effective Theory of Parton Physics*, *Phys. Rev. Lett.* **120** (2018), no. 11 112001, [arXiv:1706.08962].

[167] T. Ishikawa, Y.-Q. Ma, J.-W. Qiu, and S. Yoshida, *Renormalizability of quasiparton distribution functions*, *Phys. Rev.* **D96** (2017), no. 9 094019, [arXiv:1707.03107].

[168] J. Green, K. Jansen, and F. Steffens, *Nonperturbative Renormalization of Nonlocal Quark Bilinears for Parton Quasidistribution Functions on the Lattice Using an Auxiliary Field*, *Phys. Rev. Lett.* **121** (2018), no. 2 022004, [arXiv:1707.07152].

[169] A. V. Radyushkin, *Structure of parton quasi-distributions and their moments*, *Phys. Lett.* **B788** (2019) 380–387, [arXiv:1807.07509].

[170] J.-H. Zhang, X. Ji, A. Schäfer, W. Wang, and S. Zhao, *Accessing Gluon Parton Distributions in Large Momentum Effective Theory*, *Phys. Rev. Lett.* **122** (2019), no. 14 142001, [arXiv:1808.10824].

[171] Z.-Y. Li, Y.-Q. Ma, and J.-W. Qiu, *Multiplicative Renormalizability of Operators defining Quasiparton Distributions*, *Phys. Rev. Lett.* **122** (2019), no. 6 062002, [arXiv:1809.01836].

[172] X. Xiong, X. Ji, J.-H. Zhang, and Y. Zhao, *One-loop matching for parton distributions: Nonsinglet case*, *Phys. Rev.* **D90** (2014), no. 1 014051, [arXiv:1310.7471].

[173] Y.-Q. Ma and J.-W. Qiu, *Extracting Parton Distribution Functions from Lattice QCD Calculations*, *Phys. Rev.* **D98** (2018), no. 7 074021, [arXiv:1404.6860].

[174] R. A. Briceño, M. T. Hansen, and C. J. Monahan, *Role of the Euclidean signature in lattice calculations of quasidistributions and other nonlocal matrix elements*, *Phys. Rev.* **D96** (2017), no. 1 014502, [arXiv:1703.06072].

[175] Y.-Q. Ma and J.-W. Qiu, *Exploring Partonic Structure of Hadrons Using ab initio Lattice QCD Calculations*, *Phys. Rev. Lett.* **120** (2018), no. 2 022003, [arXiv:1709.03018].

[176] T. Izubuchi, X. Ji, L. Jin, I. W. Stewart, and Y. Zhao, *Factorization Theorem Relating Euclidean and Light-Cone Parton Distributions*, *Phys. Rev.* **D98** (2018), no. 5 056004, [arXiv:1801.03917].

[177] X. Ji, A. Schäfer, X. Xiong, and J.-H. Zhang, *One-Loop Matching for Generalized Parton Distributions*, *Phys. Rev.* **D92** (2015) 014039, [arXiv:1506.00248].

[178] X. Xiong and J.-H. Zhang, *One-loop matching for transversity generalized parton distribution*, *Phys. Rev.* **D92** (2015), no. 5 054037, [arXiv:1509.08016].

[179] W. Wang, S. Zhao, and R. Zhu, *Gluon quasidistribution function at one loop*, *Eur. Phys. J.* **C78** (2018), no. 2 147, [arXiv:1708.02458].

[180] I. W. Stewart and Y. Zhao, *Matching the quasiparton distribution in a momentum subtraction scheme*, *Phys. Rev.* **D97** (2018), no. 5 054512, [arXiv:1709.04933].

[181] C. Alexandrou, K. Cichy, M. Constantinou, K. Jansen, A. Scapellato, and F. Steffens, *Light-Cone Parton Distribution Functions from Lattice QCD*, *Phys. Rev. Lett.* **121** (2018), no. 11 112001, [arXiv:1803.02685].

[182] C. Alexandrou, K. Cichy, M. Constantinou, K. Jansen, A. Scapellato, and F. Steffens, *Transversity parton distribution functions from lattice QCD*, *Phys. Rev. D. (Rapid Communication), in production,* (2018) [arXiv:1807.00232].

[183] Y.-S. Liu, J.-W. Chen, L. Jin, H.-W. Lin, Y.-B. Yang, J.-H. Zhang, and Y. Zhao, *Unpolarized quark distribution from lattice QCD: A systematic analysis of renormalization and matching*, arXiv:1807.06566.

[184] Y.-S. Liu, J.-W. Chen, L. Jin, R. Li, H.-W. Lin, Y.-B. Yang, J.-H. Zhang, and Y. Zhao, *Nucleon Transversity Distribution at the Physical Pion Mass from Lattice QCD*, arXiv:1810.05043.

[185] L. Del Debbio, *Parton distributions in the LHC era*, *EPJ Web Conf.* **175** (2018) 01006.

[186] C. Monahan, *Recent Developments in x-dependent Structure Calculations*, *PoS* **LATTICE2018** (2018) 018, [arXiv:1811.00678].

[187] Y. Zhao, *Unraveling high-energy hadron structures with lattice QCD*, *Int. J. Mod. Phys. A* **33** (2019), no. 36 1830033, [arXiv:1812.07192].

[188] K. Cichy and M. Constantinou, *A guide to light-cone PDFs from Lattice QCD: an overview of approaches, techniques and results*, *Adv. High Energy Phys.* **2019** (2019) 3036904, [arXiv:1811.07248].

[189] A. V. Radyushkin, *Theory and applications of parton pseudodistributions*, arXiv:1912.04244.

[190] H.-W. Lin et al., *Parton distributions and lattice QCD calculations: toward 3D structure*, arXiv:2006.08636.

[191] G. C. Rossi and M. Testa, *Note on lattice regularization and equal-time correlators for parton distribution functions*, *Phys. Rev.* **D96** (2017), no. 1 014507, [arXiv:1706.04428].

[192] G. Rossi and M. Testa, *Euclidean versus Minkowski short distance*, *Phys. Rev.* **D98** (2018), no. 5 054028, [arXiv:1806.00808].

[193] J. Karpie, K. Orginos, and S. Zafeiropoulos, *Moments of Ioffe time parton distribution functions from non-local matrix elements*, *JHEP* **11** (2018) 178, [arXiv:1807.10933].

[194] J.-H. Zhang, J.-W. Chen, and C. Monahan, *Parton distribution functions from reduced Ioffe-time distributions*, *Phys. Rev. D* **97** (2018), no. 7 074508, [arXiv:1801.03023].

[195] A. Radyushkin, *One-loop evolution of parton pseudo-distribution functions on the lattice*, *Phys. Rev.* **D98** (2018), no. 1 014019, [arXiv:1801.02427].

[196] K. Orginos, A. Radyushkin, J. Karpie, and S. Zafeiropoulos, *Lattice QCD exploration of parton pseudo-distribution functions*, *Phys. Rev.* **D96** (2017), no. 9 094503, [arXiv:1706.05373].

[197] A. V. Radyushkin, *Quark pseudodistributions at short distances*, *Phys. Lett.* **B781** (2018) 433–442, [arXiv:1710.08813].

[198] B. Joó, J. Karpie, K. Orginos, A. Radyushkin, D. Richards, and S. Zafeiropoulos, *Parton Distribution Functions from Ioffe time pseudo-distributions*, arXiv:1908.09771.

[199] B. Joó, J. Karpie, K. Orginos, A. V. Radyushkin, D. G. Richards, R. S. Sufian, and S. Zafeiropoulos, *Pion valence structure from Ioffe-time parton pseudodistribution functions*, *Phys. Rev. D* **100** (2019), no. 11 114512, [arXiv:1909.08517].

[200] B. Joó, J. Karpie, K. Orginos, A. V. Radyushkin, D. G. Richards, and S. Zafeiropoulos, *Parton Distribution Functions from Ioffe time pseudo-distributions from lattice caclulations; approaching the physical point*, arXiv:2004.01687.

[201] A. V. Radyushkin, *Generalized parton distributions and pseudodistributions*, *Phys. Rev. D* **100** (2019), no. 11 116011, [arXiv:1909.08474].

[202] C. Alexandrou, K. Cichy, M. Constantinou, K. Hadjiyiannakou, K. Jansen, A. Scapellato, and F. Steffens, *Systematic uncertainties in parton distribution functions from lattice QCD simulations at the physical point*, *Phys. Rev.* **D99** (2019), no. 11 114504, [arXiv:1902.00587].

[203] Y. Chai et al., *Parton distribution functions of Δ^+ on the lattice*, arXiv:2002.12044.

[204] M. Bhat, K. Cichy, M. Constantinou, and A. Scapellato, *Parton distribution functions from lattice QCD at physical quark masses via the pseudo-distribution approach*, arXiv:2005.02102.

[205] C. Monahan and K. Orginos, *Quasi parton distributions and the gradient flow*, *JHEP* **03** (2017) 116, [arXiv:1612.01584].

[206] C. Monahan, *Smeared quasidistributions in perturbation theory*, *Phys. Rev. D* **97** (2018), no. 5 054507, [arXiv:1710.04607].

[207] R. Narayanan and H. Neuberger, *Infinite N phase transitions in continuum Wilson loop operators*, *JHEP* **03** (2006) 064, [hep-th/0601210].

[208] M. Luscher and P. Weisz, *Perturbative analysis of the gradient flow in non-abelian gauge theories*, *JHEP* **02** (2011) 051, [arXiv:1101.0963].

[209] M. Luscher, *Chiral symmetry and the Yang–Mills gradient flow*, *JHEP* **04** (2013) 123, [arXiv:1302.5246].

[210] M. Lüscher, *Future applications of the Yang-Mills gradient flow in lattice QCD*, *PoS* **LATTICE2013** (2014) 016, [arXiv:1308.5598].

[211] C. Monahan and K. Orginos, *Locally smeared operator product expansions in scalar field theory*, *Phys. Rev. D* **91** (2015), no. 7 074513, [arXiv:1501.05348].

[212] C. Monahan, *The gradient flow in simple field theories*, *PoS* **LATTICE2015** (2016) 052, [arXiv:1512.00294].

[213] K. Fujikawa, *The gradient flow in $\lambda\phi^4$ theory*, *JHEP* **03** (2016) 021, [arXiv:1601.01578].

[214] R. Cichetti and A. Faraone, *Incomplete Hankel and Modified Bessel Functions: A Class of Special Functions for Electromagnetics*, *IEEE Transactions on Antennas and Propagation* **52** (2004) 3373.

[215] D. Jones, *Incomplete Bessel Functions. I*, *Proc. of the Edinburgh Math. Soc.* **50** (2007) 173.

[216] F. E. Harris, *Incomplete Bessel, generalized incomplete gamma, or leaky aquifer functions*, *J. Comp. App. Math.* **215** (2008) 260.

[217] H.-W. Lin, J.-W. Chen, S. D. Cohen, and X. Ji, *Flavor Structure of the Nucleon Sea from Lattice QCD*, *Phys. Rev.* **D91** (2015) 054510, [arXiv:1402.1462].

[218] C. Alexandrou, K. Cichy, V. Drach, E. Garcia-Ramos, K. Hadjiyiannakou, K. Jansen, F. Steffens, and C. Wiese, *Lattice calculation of parton distributions*, *Phys. Rev.* **D92** (2015) 014502, [arXiv:1504.07455].

[219] J.-W. Chen, S. D. Cohen, X. Ji, H.-W. Lin, and J.-H. Zhang, *Nucleon Helicity and Transversity Parton Distributions from Lattice QCD*, *Nucl. Phys.* **B911** (2016) 246–273, [arXiv:1603.06664].

[220] C. Alexandrou, K. Cichy, M. Constantinou, K. Hadjiyiannakou, K. Jansen, F. Steffens, and C. Wiese, *Updated Lattice Results for Parton Distributions*, *Phys. Rev.* **D96** (2017), no. 1 014513, [arXiv:1610.03689].

[221] J.-H. Zhang, J.-W. Chen, X. Ji, L. Jin, and H.-W. Lin, *Pion Distribution Amplitude from Lattice QCD*, *Phys. Rev.* **D95** (2017), no. 9 094514, [arXiv:1702.00008].

[222] **LP3** Collaboration, H.-W. Lin, J.-W. Chen, T. Ishikawa, and J.-H. Zhang, *Improved parton distribution functions at the physical pion mass*, *Phys. Rev.* **D98** (2018), no. 5 054504, [arXiv:1708.05301].

[223] **LP3** Collaboration, J.-H. Zhang, L. Jin, H.-W. Lin, A. Schäfer, P. Sun, Y.-B. Yang, R. Zhang, Y. Zhao, and J.-W. Chen, *Kaon Distribution Amplitude from Lattice QCD and the Flavor SU(3) Symmetry*, *Nucl. Phys.* **B939** (2019) 429–446, [arXiv:1712.10025].

[224] J.-W. Chen, L. Jin, H.-W. Lin, Y.-S. Liu, Y.-B. Yang, J.-H. Zhang, and Y. Zhao, *Lattice Calculation of Parton Distribution Function from LaMET at Physical Pion Mass with Large Nucleon Momentum*, arXiv:1803.04393.

[225] J.-W. Chen, L. Jin, H.-W. Lin, Y.-S. Liu, A. Schäfer, Y.-B. Yang, J.-H. Zhang, and Y. Zhao, *First direct lattice-QCD calculation of the x-dependence of the pion parton distribution function*, arXiv:1804.01483.

[226] H.-W. Lin, J.-W. Chen, X. Ji, L. Jin, R. Li, Y.-S. Liu, Y.-B. Yang, J.-H. Zhang, and Y. Zhao, *Proton Isovector Helicity Distribution on the Lattice at Physical Pion Mass*, *Phys. Rev. Lett.* **121** (2018), no. 24 242003, [arXiv:1807.07431].

[227] Z.-Y. Fan, Y.-B. Yang, A. Anthony, H.-W. Lin, and K.-F. Liu, *Gluon Quasi-Parton-Distribution Functions from Lattice QCD*, *Phys. Rev. Lett.* **121** (2018), no. 24 242001, [arXiv:1808.02077].

[228] T. Izubuchi, L. Jin, C. Kallidonis, N. Karthik, S. Mukherjee, P. Petreczky, C. Shugert, and S. Syritsyn, *Valence parton distribution function of pion from fine lattice*, arXiv:1905.06349.

[229] J. Karpie, K. Orginos, A. Rothkopf, and S. Zafeiropoulos, *Reconstructing parton distribution functions from ioffe time data: from bayesian methods to neural networks*, *Journal of High Energy Physics* **2019** (Apr, 2019) 57.

[230] V. Braun, P. Gornicki, and L. Mankiewicz, *Ioffe - time distributions instead of parton momentum distributions in description of deep inelastic scattering*, *Phys. Rev.* **D51** (1995) 6036–6051, [hep-ph/9410318].

[231] Z.-Y. Li, Y.-Q. Ma, and J.-W. Qiu, *Extraction of Next-to-Next-to-Leading-Order PDFs from Lattice QCD Calculations*, arXiv:2006.12370.

[232] L.-B. Chen, W. Wang, and R. Zhu, *Next-to-next-to-leading order corrections to quark Quasi parton distribution functions*, arXiv:2006.14825.

[233] V. Braun, K. Chetyrkin, and B. Kniehl, *Renormalization of parton quasi-distributions beyond the leading order: spacelike vs. timelike*, arXiv:2004.01043.

[234] L.-B. Chen, W. Wang, and R. Zhu, *Quasi parton distribution functions at NNLO: flavor non-diagonal quark contributions*, arXiv:2005.13757.

[235] L.-B. Chen, W. Wang, and R. Zhu, *Master Integrals for two-loop QCD corrections to Quasi PDFs*, arXiv:2006.10917.

[236] G. Martinelli, C. Pittori, C. T. Sachrajda, M. Testa, and A. Vladikas, *A General method for nonperturbative renormalization of lattice operators*, *Nucl. Phys.* **B445** (1995) 81–108, [hep-lat/9411010].

[237] A. Radyushkin, *Nonperturbative Evolution of Parton Quasi-Distributions*, *Phys. Lett.* **B767** (2017) 314–320, [arXiv:1612.05170].

[238] **ETM** Collaboration, A. Abdel-Rehim et al., *First physics results at the physical pion mass from $N_f = 2$ Wilson twisted mass fermions at maximal twist*, *Phys. Rev.* **D95** (2017), no. 9 094515, [arXiv:1507.05068].

[239] C. Alexandrou and C. Kallidonis, *Low-lying baryon masses using $N_f = 2$ twisted mass clover-improved fermions directly at the physical pion mass*, *Phys. Rev.* **D96** (2017), no. 3 034511, [arXiv:1704.02647].

[240] S. Syritsyn, *Review of Hadron Structure Calculations on a Lattice*, *PoS* **LATTICE2013** (2014) 009, [arXiv:1403.4686].

[241] M. Constantinou, *Hadron Structure*, *PoS* **LATTICE2014** (2015) 001, [arXiv:1411.0078].

[242] M. Constantinou, *Recent progress in hadron structure from Lattice QCD*, *PoS* **CD15** (2015) 009, [arXiv:1511.00214].

[243] C. Alexandrou, *Selected results on hadron structure using state-of-the-art lattice QCD simulations*, in *Proceedings, 45th International Symposium on Multiparticle Dynamics (ISMD 2015): Kreuth, Germany, October 4-9, 2015*, 2015. arXiv:1512.03924.

[244] J. Green, *Systematics in nucleon matrix element calculations*, in *36th International Symposium on Lattice Field Theory (Lattice 2018) East Lansing, MI, United States, July 22-28, 2018*, 2018. arXiv:1812.10574.

[245] R. A. Briceño, J. V. Guerrero, M. T. Hansen, and C. J. Monahan, *Finite-volume effects due to spatially nonlocal operators*, *Phys. Rev.* **D98** (2018), no. 1 014511, [arXiv:1805.01034].

[246] G. S. Bali et al., *Pion distribution amplitude from Euclidean correlation functions*, *Eur. Phys. J.* **C78** (2018), no. 3 217, [arXiv:1709.04325].

[247] J. Bringewatt, N. Sato, W. Melnitchouk, J.-W. Qiu, F. Steffens, and M. Constantinou, *Confronting lattice parton distributions with global QCD analysis*, arXiv:2010.00548.

[248] J. Karpie, K. Orginos, A. Radyushkin, and S. Zafeiropoulos, *Parton distribution functions on the lattice and in the continuum*, *EPJ Web Conf.* **175** (2018) 06032, [arXiv:1710.08288].

[249] J. Luis Bernal and J. A. Peacock, *Conservative cosmology: combining data with allowance for unknown systematics*, *JCAP* **07** (2018) 002, [arXiv:1803.04470].

[250] G. S. Bali, V. M. Braun, B. Gläßle, M. Göckeler, M. Gruber, F. Hutzler, P. Korcyl, A. Schäfer, P. Wein, and J.-H. Zhang, *Pion distribution amplitude from Euclidean correlation functions: Exploring universality and higher-twist effects*, *Phys. Rev. D* **98** (2018), no. 9 094507, [arXiv:1807.06671].

[251] R. S. Sufian, J. Karpie, C. Egerer, K. Orginos, J.-W. Qiu, and D. G. Richards, *Pion Valence Quark Distribution from Matrix Element*

Calculated in Lattice QCD, *Phys. Rev.* **D99** (2019), no. 7 074507, [arXiv:1901.03921].

[252] **RQCD** Collaboration, G. S. Bali et al., *Light-cone distribution amplitudes of octet baryons from lattice QCD*, *Eur. Phys. J. A* **55** (2019), no. 7 116, [arXiv:1903.12590].

[253] R. S. Sufian, C. Egerer, J. Karpie, R. G. Edwards, B. Joó, Y.-Q. Ma, K. Orginos, J.-W. Qiu, and D. G. Richards, *Pion Valence Quark Distribution from Current-Current Correlation in Lattice QCD*, arXiv:2001.04960.

[254] L. A. Harland-Lang, A. D. Martin, and R. S. Thorne, *The Impact of LHC Jet Data on the MMHT PDF Fit at NNLO*, *Eur. Phys. J. C* **78** (2018), no. 3 248, [arXiv:1711.05757].

[255] R. V. Harlander and W. B. Kilgore, *Higgs boson production in bottom quark fusion at next-to-next-to leading order*, *Phys. Rev.* **D68** (2003) 13001, [hep-ph/0304035].