UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

Corso di Laurea Magistrale in Fisica (LM-17)

Validation criteria in the determination of parton distributions.

Relatore:
Chiar.mo Prof. Stefano FORTE
Correlatore:
Dott. Andrea BARONTINI

Autore:
Dott. Samuele VOLTAN
Matricola: 983976

Anno Accademico 2021/2022

Dedicato ai miei genitori.

ABSTRACT

We address the problem of the validation of fitting methodologies that determine Parton Distribution Functions (PDFs) from experimental data.

Knowledge of PDFs is essential to perform measurements of high energy physics processes at hadron colliders. According to the standard theory of strong interactions – Quantum Chromodynamics (QCD) – cross sections are produced from the combination of PDFs with quantities that can be computed using perturbation theory in QCD. Since PDFs themselves cannot be measured directly, their determination is an inverse problem in the sense that it consists in reconstructing their functional form from experimental data.

In the Bayesian approach, PDF determination corresponds to understanding posterior probability distributions in PDF space through the analysis of data space probabilities. Within this framework, tests can be performed on a fitting methodology by comparing its output to a known underlying true set of PDFs, which can be guessed from previous determinations. This is done by fitting artificial data, rather than experimental, generated from chosen underlying PDFs with realistic uncertainties. The procedure takes the name of closure test and it has been systematically used by the NNPDF collaboration since 2012.

In this thesis, we exploit closure tests to determine the impact of inconsistent data on the NNPDF4.0 methodology. By definition, inconsistent data are such that their real uncertainty is larger than their nominal, which is determined through the composition of statistical and systematic errors given by experimental collaborations. It has been suggested that the presence of inconsistent data can impact the measure of a PDF fit quality, the χ^2, and therefore explain the large χ^2 values obtained by several collaborations in the latest determinations. This follows from the reasonable assumption that, when presented with inconsistent datasets, a methodology would either follow the trend of consistent data, thereby increasing the inconsistent dataset's χ^2, or behave in the opposite way and therefore increase the χ^2 of consistent datapoints.

We perform a direct measure of the impact of inconsistent data exploiting the closure test setup. Being trained on artificially generated data, a closure test fit is in principle free from inconsistencies. Inconsistent data can therefore be introduced manually in the fitting framework by manipulations of the experimental covariance matrix, which accounts for both systematic and statistical uncertainties. This is done through the rescaling of a number of eigenvalues in the systematic uncertainty matrices, which correspond to the measurement of specific observables. In particular, we study four situations corresponding to measurements of neutral current DIS, single inclusive jet production and electroweak Drell-Yan (DY) boson production.

Contrary to a standard PDF fit, the output of an inconsistent closure test fit can be compared to the underlying law selected to generate artificial training data. This eases the introduction of a family of statistical estimators that are used to determine whether the results of the fit are comparable with what expected. In particular, we investigate the closure test performance both for the PDF central val-

ues and uncertainties delivered by the closure test, and for the expected deviations from the hypotesis made on the prior probability distributions.

ACKNOWLEDGEMENTS

First and foremost, I wish to thank my supervisor Prof. Stefano Forte for the opportunity to work alongside him and for being a source of inspiration. I am extremely grateful for the constant support and precious suggestions He provided me with during this year. I consider myself lucky to have met a man of great culture that is profoundly dedicated to forming the minds of younger generations.

I would also like to thank the members of the NNPDF collaboration with who I enjoyed sharing my journey, starting from my co-supervisor Dott. Andrea Barontini. Not only he is a brilliant scientist. Above all, he has the rare ability to be able to listen and help by seeing possibilities where others see problems.

A special thank goes to Dott. Roy Stegeman and Dott. Juan Cruz Martinez for helping me out during the early stages of my work. Without their patient advice, it would have been much difficult for me to embrace both technical and conceptual knowledge that made me accomplish this project.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

QED	Quantum Electrodynamics
QFT	Quantum Field Theory
QCD	Quantum Chromodynamics
DIS	Deep Inelastic Scattering
RGE	Renormalization Group Equation
PDF	Parton Distribution Function
GA	Genetic Algorithm
UV	Ultraviolet
IR	Infrared
LO	Leading Order
NLO	Next-to-Leading Order
NNLO	Next-to-Next-to-Leading Order
SM	Standard Model
MAP	Maximum A Posteriori
MHOU	Missing Higher Order Uncertainty
CLT	Central Limit Theorem
GD	Gradient Descent
SGD	Stochastic Gradient Descent
FFNN	Feed Forward Neural Network
VFNS	Variable Flavour Number Scheme
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
DY	Drell-Yan
KL	Kullback-Leibler

Part I

PRELUDE

This part is devoted to the systematic review of known theoretical aspects of the subjects that concern the determination of parton distribution functions through deep learning techniques.

1 STRONG INTERACTIONS

This chapter is devoted to the review of theoretical aspects of the theory of strong interactions, that is, Quantum Chromodynamics (QCD). For what concerns the topics that will be discussed in the rest of the thesis, we shall focus our attention on two advanced aspects of the theory, both usually referred to as factorization. We will extensively mention and exploit several notions of Quantum Field Theory (QFT) – such as couplings and renormalizablilty – that are assumed to be consolidated by the reader. If this is not the case, we refer to the book by Peskin and Schroeder [1] for an introduction of QFT.

By factorization one means the property of QCD for which an observable can be written as the composition of two separate families of contributions, respectively describing the perturbative and non-perturbative regimes of QCD. From this point of view, factorization can be thought as an application of the Wilson operator expansion, as we will discuss in Section 1.2.

Another feature of QCD that is commonly referred to as factorization is the subtraction of collinear singularities. It follows from the fact that singularities associated with collinear emission of real partons in strong processes cannot be renormalized or cancelled systematically. The universality of such singularities allows for their subtraction into the non-perturbative contributions of QCD cross sections, where they are treated as fictitious consequences of the application of perturbation theory at low energies. Their discussion is carried out from Section 1.2.2

As a consequence of the factorization of collinear singularities, Wilson coefficients acquire a dependence on a energy scale, called factorization scale. The evolution of physical quantities with such scale is described by a set of differential equations, called the Altarelli-Parisi or DGLAP equations, that will be presented in Section 1.3

The final part, Section 1.4, is reserved to a discussion on the masses of quarks, which are the fundamental fields of QCD with the notation that will be introduced in Section 1.1. The latter section features a brief introduction to the QCD Lagrangian and the running of its coupling constant in the context of the renormalization of the theory.

1.1 QUANTUM CHROMODYNAMICS

As many fundamental theories, strong interactions are described within the framework of a QFT. This paradigm takes his roots in many independent discoveries made in the 1960s and early 70s in two separate fields: Deep Inelastic Scattering (DIS) experiments and the quark model.

The latter aimed at explaining the discovery of a huge number of strongly interacting particles over a short period of time with the assumption that these particles were nothing more than resonances of more fundamental constituents, the quarks [2, 3]. The classification of such particles was based on the irreducible representations of the flavour symmetry group $SU(3)$. In this picture, quarks are the elements of the fundamental representation and the observed resonances – among which are baryons and mesons – can be described as bound states of quarks.

Few years later, evidence of the presence of pointlike constituents inside hadrons was found in the first DIS experiments. Such particles, called partons, were understood to behave as free in the high energy limit and form hadronic matter as their bound states.

The success of both descriptions of strongly interacting processes suggested the identification of partons with quarks. It was not until a systematic treatment of renormalization, leading to the discovery of asymptotic freedom [4, 5], that the two ideas could be incorporated into a QFT. This framework provides the correct interpretation of quarks as fundamental fields.

1.1.1 *Lagrangian formulation*

QCD is the theory that describes strong interactions. Its fundamental fields are spin-1/2 fermions of fractional electric charge – either $+2/3$ or $-1/3$ – coming in three families, each one containing two of them and their anti-particles, for a total of six flavours. The need for six flavours of quarks comes from the fact that the theory must take into account all the known hadrons as their bound states. Quarks are described by Dirac spinor fields ψ_A in the fundamental representation of $SU(3)$, that is, $A = 1, 2, 3$. Note that, contrary to the flavour symmetry introduced by the quark model, the $SU(3)$ gauge group of QCD is identified with a new charge, called color charge. We say that the color $SU(3)$ symmetry is exact, in the sense that it is the fundamental symmetry of QCD.

The force carriers, called gluons, are spin-1 bosons described by the gauge fields A^a_μ in the adjoint representation of $SU(3)$, thus $a = 1, \dots, 8$. Interaction between quarks and gluons is given by the covariant derivative

$$D_\mu = \partial_\mu + igT^a A^a_\mu, \tag{1}$$

where T^a are the eight generators of $SU(3)$ and g is the QCD coupling constant. The following commutation relations hold:

$$[T^a, T^b] = i f^{abc} T^c, \tag{2}$$

f^{abc} being the $SU(3)$ structure constants. The generators live in the fundamental representation and thus, dressed with all its indices, the covariant derivative reads $(D_\mu)_{AB}$. It is quite common in literature to fix the normalization of the generators with the identification $T^a = \lambda^a/2$, where λ^a are the Gell-Mann matrices. This choice implies that $\mathrm{tr}\left(T^a T^b\right) = \delta^{ab}/2$.

Given the interaction of Equation 1, the QCD Lagrangian is

$$\mathcal{L} = \sum_{\text{flavours}} \overline{\psi}_A (i\not{D} - m_f)_{AB} \psi_B - \frac{1}{4} F^a_{\mu\nu} F^{\mu\nu a}. \tag{3}$$

The gauge fields are hidden inside the covariant derivative by means of Equation 1, and into the field strenght tensor

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g f^{abc} A^b_\mu A^c_\nu. \tag{4}$$

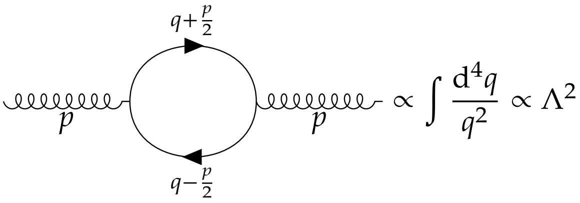


Figure 1: Gluon self energy (one-loop radiative correction of the gluon propagator). The superficial divergence resulting from the integration over loop momentum q is of order Λ^2.

We see in Equation 3 what we meant by saying that the color symmetry is exact, while the flavour symmetry is not. Indeed, while the Lagrangian can be factorized as a sum of flavours, interaction is described by the color indices and, consequently, different flavours of quark do not interact with each other. The reason for this choice – which is the reason for the introduction of color – is that the presence of flavour mixing terms would prevent the QCD Lagrangian from being renormalizable.

In addition, since QCD's gauge group is non-abelian, the classical Lagrangian of Equation 3 does not account for a gluon mass term. As it happens for the Standard Model (SM) of electroweak interactions, any experimental mass correction must be realized upon possible spontaneous symmetry breaking mechanism: however, gluons are believed to have zero mass and that is not necessary at all in QCD. Another feature of non-abelianity is the presence of gauge fields product in Equation 4, which gives rise to trilinear and quadrilinear interaction between gluons. This means that, contrary to what happens for photons in Quantum Electrodynamics (QED), gluons carry color charge. This difference arises at a classical level, i. e. in the Lagrangian itself, and its consequences can be seen in the quantum theory after renormalization through opposite runnings of the QED and QCD couplings.

1.1.2 *Consequences of renormalization*

A brief discussion concerning the renormalization of field theories is necessary when studying theoretical aspects of particle physics such as parton distributions. Renormalization handles what seem to be conceptual obstacles by turning them into fundamental and predictive features of the theory. Since the methods introduced by renormalization have a common ground with factorization in QCD, this section can ease its introduction while summarizing key concepts of QCD such as asymptotic freedom and the properties of its vacuum.

QFTs provide computations of scattering amplitudes through perturbative expansion of the operators involved in the processes. Assuming that perturbation theory holds, i. e. growing powers of the coupling constants give smaller corrections, this is usually done by writing Feynman diagrams. Higher order computations involve diagrams with increasingly complicated topologies that give rise to divergent integrals. For instance, consider QCD's gluon self energy: Figure 1 shows that integration over the loop momentum gives a quadratic superficial divergence in the region of high momenta, called Ultraviolet (UV) region. As a consequence, every observable quantity computed with this diagram loses its physical sense, meaning that it cannot be interpreted as a probability. Altough sometimes divergent contributions cancel each other out, there remains an infinite number of diagrams yielding such unphysical results.

The problem is cured by assuming that the Lagrangian of Equation 3 is built upon bare quantities that are related to physical ones through scale dependent relations, so that the poles of the physical Green functions, i. e. expectation values of the field operators, are systematically subtracted by such dependence. This introduces a dependency of physical observables on a energy scale, called renormalization scale λ_R, that is described by a differential equation, called Renormalization Group Equation (RGE).

It can be shown that Green functions that depend on the coupling constant $\alpha_S = g^2/4\pi$ solve the RGE if the coupling behaves according to the following differential equation, called the running coupling equation:

$$\begin{cases} \dfrac{d\alpha_S}{d\log Q^2} = \beta(\alpha_S) \\ \alpha_S(\lambda_R^2) = \alpha_S. \end{cases} \tag{5}$$

This equation states that there exists a universal function β, related to the shifts in the coupling constant, that compensates for the shifts in the renormalization scale. In other words, it describes the rate of the renormalization group flow of the coupling constant.

The β function is related to the derivative of the renormalized coupling with respect to the scale, which depends only on counterterms adopted during renormalization and is therefore a property of the theory. Moreover, since renormalizable theories feature dimensionless coupings, β can be written as a perturbative series with numerical coefficients:

$$\beta(\alpha_S) = \alpha_S^2 \left(-\beta_0 + \sum_{k=0}^{+\infty} \beta_k \alpha_S^k \right). \tag{6}$$

For QCD, results up to five loops [6] have been computed: at Leading Order (LO), we find

$$\beta_0 = \frac{1}{12\pi}(33 - 2N_{\text{flav}}), \tag{7}$$

where N_{flav} stands for the number of flavours that are considered at the scale Q.

Solutions of the RGE equation are found integrating Equation 5 over α_S:

$$\log \frac{Q^2}{\lambda_R^2} = \int_{\alpha_S(Q^2)}^{\alpha_S} \frac{d\alpha}{\beta(\alpha)}. \tag{8}$$

We can adopt the perturbative expansion given by Equation 6 to find

$$\begin{aligned} \log \frac{Q^2}{\lambda_R^2} &= -\int_{\alpha_S(Q^2)}^{\alpha_S} \frac{d\alpha}{\beta_0 \alpha^2} \left(1 + \sum_{k=0}^{+\infty} \frac{\beta_k}{\beta_0} \alpha^k \right) = \\ &= \frac{1}{\beta_0} \left(\frac{1}{\alpha_S(Q^2)} - \frac{1}{\alpha_S} \right) - \frac{\beta_1}{\beta_0^2} \log \frac{\alpha_S(Q^2)}{\alpha_S} + \mathcal{O}(\alpha_S). \end{aligned} \tag{9}$$

At LO in the coupling constant, the running coupling is therefore

$$\alpha_S(Q^2) = \frac{\alpha_S}{1 + \beta_0 \alpha_S \log(Q^2/\lambda_R^2)} (1 + O(\log \alpha_S)). \tag{10}$$

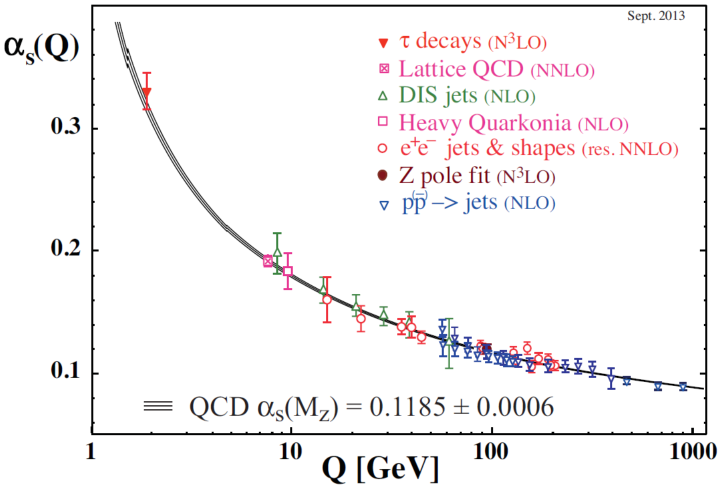


Figure 2: Summary of measurements of α_S as a function of the energy scale Q/Λ_{QCD}. The respective degree of QCD perturbation theory used in the extraction of α_S is indicated in brackets. Taken from [7].

The expression shows that, depending on the sign of the β_0 coefficient, the running coupling increases (or decreases) at logarithmic rate with Q. This means that a positive value for β_0 implies an effective coupling that becomes stronger at large momenta and weaker at small ones, while a negative β_0 determines the opposite behavior. Since β_0 is a universal parameter, such distinction allows us to separate QFTs into two families.

For instance, QED belongs to the group of field theories for which the coupling becomes smaller at small momenta since its β_0 is positive. This can be understood as a dieletric property of QED's vacuum: at large distances – i. e. at small energies – the primary electric charged is masked by the infinite particle-antiparticle pairs created by the photon self-energy contributions. On the other hand, non-abelian gauge theories such as QCD always account for a trilinear contributions to their gauge boson's self-energy. This behavior of the vacuum is paramagnetic and belongs to the family of field theories whose coupling strenght decreases at small distances. This effect, shown in Figure 2, is called asymptotic freedom.

The fact that the coupling grows at small momenta means that there exists a energy scale at which α_S is of order one and perturbation theory ceases to hold. This scale is denoted by Λ_{QCD} and clearly depends on the choice of the renormalization scheme adopted, the order of the β-series and the number of flavors considered when solving Equation 5. By all means we can say that Λ_{QCD} is the characterstic scale of QCD and any dimensionful quantity can be expressed in units of it. It is sometimes stated that Λ_{QCD} is the characteristic scale of confinement. This is reasonably true in the sense that confinement is a property of the low momentum sector of QCD for which two color-charged particles cannot be observed as isolated. However, color confinement is a property of the inter-quark potential rather than of the coupling of the theory and its criteria are typical of non perturbative QCD.

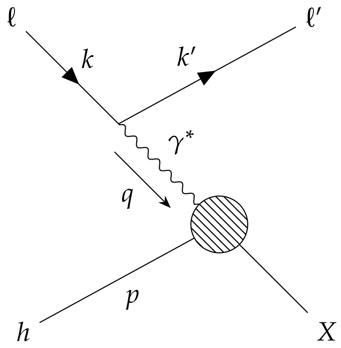


Figure 3: A pictorial diagram of a DIS process. The incoming lepton ℓ scatters off a hadronic target h with the exchange of a virtual photon γ with momentum q.

1.2 FACTORIZATION

As we have seen in the previous section, computation of scattering amplitudes for non-abelian gauge theories are hindered by the presence of a kinematic region that cannot be treated within perturbation theory. With the exception of inclusive cross sections without hadrons in the initial state, such as $\sigma(e^+ + e^- \to \text{hadrons})$ [8], we can state that a general strong amplitude is a combination of short and long-distance behaviors. Its computation can therefore be carried out through factorization of the non perturbative effects from the perturbative high energy contributions, as indicated in the introduction.

The section is devoted to the presentation of factorization through the discussion of the LO and Next-to-Leading Order (NLO) treatment of DIS in perturbative QCD, respectively in Section 1.2.1 and Section 1.2.2. Since real emission diagrams only arise ad NLO, the latter will deliver the introduction the factorization of collinear singularities and the factorization scale that have been mentioned at the beginning of this chapter. The final part, Section 1.2.3, will instead focus on extending factorization to hadronic processes, i. e. strong processes with hadrons both in the initial and final states.

1.2.1 *Leading order DIS*

As anticipated, perturbative QCD describes strong processes in terms of Parton Distribution Functions (PDFs), assuming that quarks and gluons – the partons – are the fundamental components of hadrons.

A process that involves scattering of bound states of such particles is therefore described in the hypothesis that partons interact according to QCD, carrying a fraction $x < 1$ of the total momentum of the hadron. The probability for the i-th parton to enter the interaction is proportional to x and there exists a probability density function $f_i(x)$, the PDF, such that $P_i \propto f_i(x)\,dx$. In this light, a general hadronic contribution to an observable σ can always be computed within perturbative QCD by combinations of such probability densities with the parton level amplitudes.

In order to better understand how factorization works, we shall now derive it for the LO DIS cross sections. DIS is a leptonic strong process consisting in the inelastic scattering of a lepton with a hadronic target, such as a proton, as pictured in Figure 3. The final state is composed by the scattered lepton and a complicated state of products of the disintegrated hadron. We shall restrict ourselves to the case of a photon induced unpolarized scattering, therefore discarding the contribution of the Z channel in the assumption that the energy scale of the process is significantly lower than m_Z^2.

The form of the total cross section features a leptonic contribution $L_{\mu\nu}$, from the upper half of the diagram, and a hadronic part $H_{\mu\nu}$ expressed in terms of the matrix elements

$$\begin{aligned} L_{\mu\nu} &\propto \sum_{\sigma,\sigma'} \langle \ell(k,\sigma)| \, j_\mu^{\text{QED}} \, |\ell'(k',\sigma')\rangle \, \langle \ell'(k',\sigma')| \, j_\nu^{\text{QED}} \, |\ell(k,\sigma)\rangle \\ H_{\mu\nu} &\propto \sum_\sigma \sum_X \langle h(p,\sigma)| \, j_\mu^{\text{had}} \, |X\rangle \, \langle X| \, j_\nu^{\text{had}} \, |h(p,\sigma)\rangle \, , \end{aligned} \tag{11}$$

such that $d\sigma \propto L_{\mu\nu} H^{\mu\nu}$. The hadronic contribution has to be extracted from comparison between experimental data and a suitable parameterization in terms of structure functions $F_a(x, Q^2)$. The number of such functions is determined by the polarization states of the virtual boson exchanged in the process: within the approximations made before and discarding parity violation contributions, $H_{\mu\nu}$ is parametrized by two functions, F_1 and F_2.

At high energy, the double differential cross section for the considered DIS process is

$$\frac{d^2\sigma}{dx\, dQ^2} = \frac{4\pi\alpha^2}{xQ^4}\left[y^2 x F_1(x,Q^2) + (1-y)F_2(x,Q^2)\right], \tag{12}$$

with $Q^2 = -q^2$ and $y = Q^2/xs$, where $\sqrt{s}$ is the centre-of-mass energy. As we anticipated, structure functions can be computed upon combinations of their parton level counterparts with PDFs. The combination is carried out with a mathematical operation called convolution, defined as

$$f(x) \otimes g(x) = \int_x^1 \frac{d\xi}{\xi} \, f\left(\frac{x}{\xi}\right) g(\xi). \tag{13}$$

The operation is clearly symmetric as one can check with the substitution $\xi' = x/\xi$.

Factorization is the property by which the two structure functions F_1 and F_2 can be written as follows:

$$F_1(x,Q^2) = \sum_i \widehat{F}_1^i(x,Q^2) \otimes f_i(x) \tag{14}$$

and

$$F_2(x,Q^2) = \sum_i \frac{x}{\xi} \widehat{F}_2^i(x,Q^2) \otimes f_i(x). \tag{15}$$

where $\widehat{F}_1$ and $\widehat{F}_2$ are parton-level structure functions that can be computed within perturbative QCD through the study of the subprocesses displayed in Figure 4.

At LO, the process is a QED vertex and therefore the interaction between the photon and the quark is purely electromagnetic. Note that, for this reason, contributions of gluon-initiated processes are not present at LO and will eventually start at NLO associated with real emissions of quarks. The partonic cross section $\widehat{\sigma}(\ell + q \to \ell + q)$ follows from trivial QED calculations and reads

$$\frac{d\widehat{\sigma}}{dx\, dQ^2} = \frac{4\pi\alpha^2}{2Q^4} e^2 \left[1 + (1-y)^2\right] \delta(x - \xi). \tag{16}$$

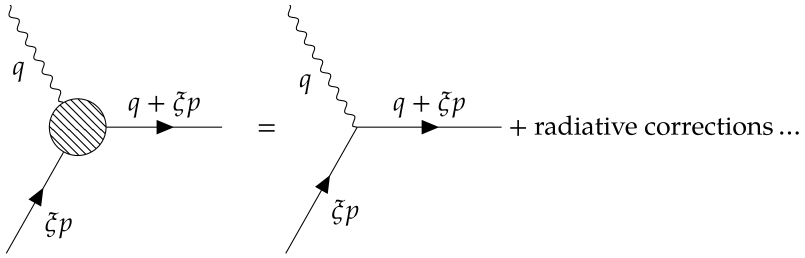


Figure 4: The LO partonic scattering of the virtual photon for a DIS subprocess with fermionic parton in the initial states.

Therefore, we can make the following identifications for the partonic structure functions:

$$\widehat{F}_2^i = e_i^2 \delta(\xi - x) \tag{17}$$

and

$$\widehat{F}_1^i = \frac{\xi^2}{2x} \widehat{F}_2^i. \tag{18}$$

Finally, we can recover from Equation 17 and Equation 18 the values of the total structure functions F_1 and F_2 which parametrized the hadronic tensor through a direct integration of Equation 14 and Equation 15: we find

$$F_1(x, Q^2) = \sum_i \int_x^1 \frac{d\xi}{\xi} \frac{\xi^2 e_i^2 \delta(\xi - x)}{2x} f_i(\xi) = \frac{1}{2} \sum_i e_i^2 f_i(x) \tag{19}$$

and, similarly,

$$F_2(x, Q^2) = x \sum_i e_i^2 f_i(x). \tag{20}$$

The structure functions are multiplicatively dependent on the electric charge of the quark, which follows from the assumption that the LO interaction between the virtual photon and quarks is purely electromagnetic.

In conclusion, we found that the LO DIS cross section can be computed from the combination of a leptonic tensor $L_{\mu\nu}$, which is easily found through trivial QCD calculations, and a hadronic tensor $H^{\mu\nu}$ parametrized with structure functions. Factorization consists in determining the values of these structure functions through the convolution of PDFs with partonic cross sections that can be computed with the Feynman rules deriving from Equation 3. The problem of describing the hadronic contribution is therefore solved by the assumption that it is entirely determined by more elementary processes combined together.

1.2.2 *Factorization of collinear singularities*

We now review the NLO treatment of DIS by computing the radiative corrections of Figure 4 within perturbative QCD. Corrections to the LO approximation can be split into two families: real emissions, displayed in Figure 5, and virtual loop diagrams, as in Figure 6.

Aside from UV poles, that can be cured with renormalization at all orders in the coupling constant, QCD's virtual contributions suffer both UV and Infrared (IR) singularities due to the former family. Indeed, the Lorentz invariant phase space

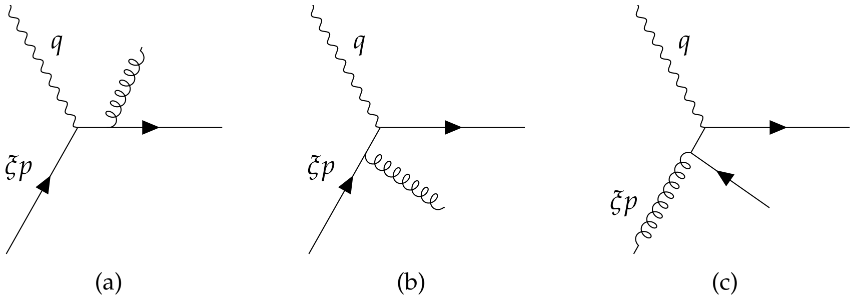


Figure 5: The NLO QCD's real emissions contributing to the scattering process. Final (5a) and initial (5b) state gluon emissions are displayed alongside the gluon-initiated (5c) real quark emission.

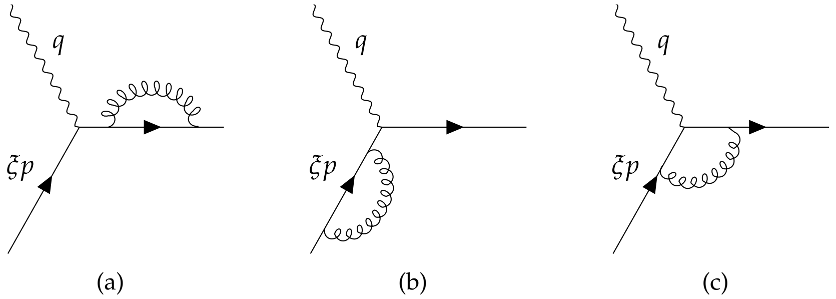


Figure 6: The NLO QCD's virtual contributions to partonic scattering for a quark initiated process. Initial (6a) and final (6b) state quark self-energy are displayed alongside the vertex correction (6c).

for these contributions depends on the integration over the momentum of the virtual/real emitted parton,

$$d\Phi \propto \int \frac{d^3k}{E} \propto \log^2 \Lambda, \tag{21}$$

which can be divided into transverse and longitudinal contributions that are both logarithmically divergent.

In particular, by writing $d^3k \propto dk_t^2\, dE$, we can distinguish between IR singularities, arising in the $E \to 0$ region, and collinear singularities, coming from $k_t \to 0$. The former are commonly referred to as soft singularities, since the emitted parton has zero energy in the limit. Soft diagrams are safe, meaning that divergent contributions coming from loop integrals cancel against the soft divergence of Equation 21 [9, 10].

On the other hand, collinear divergences coming from Figure 5b and Figure 5c must be regularized with a cutoff Q_{cut}. For instance, the quark-initiated initial state emission yields the following structure function:

$$\widehat{F}_2^i = e_i^2 \left[\delta(\xi - x) + \frac{x\alpha_S}{2\pi\xi^2}\left(P_{qq}(z)\log\frac{Q^2}{Q^2_{\text{cut}}} + \text{finite terms}\right) + \mathcal{O}(\alpha_S^2)\right], \tag{22}$$

with $z = x/\xi$ and

$$P_{qq}(z) = \frac{4}{3}\frac{1+z^2}{1-z}. \tag{23}$$

This function, which describes the $q \to q$ parton splitting, is known as the quark splitting function. Splitting functions are perturbative objects that can be written for any kind of parton splitting, e. g. P_{qg}, P_{gq}, P_{gg} and their anti-quark corrispectives. As we shall discuss later in great detail, collinear emission diagrams contribute to the DGLAP evolution of PDFs: in this light, the splitting functions can be seen as kernels of the evolution operators.

Eventually, the substitution $Q_{\text{cut}} \to 0$ in Equation 22 yields unphysical results related to a collinear initial-state gluon emission. The situation belongs to the long range regime of strong interactions and thus cannot be treated within perturbative QCD. Hence, a suitable scale λ_F^2 – the aforementioned factorization scale – is identified to separate hard perturbative contributions from soft ones in the logarithms by means of the following factorization:

$$\log \frac{Q^2}{Q_{\text{cut}}^2} = \log \frac{Q^2}{\lambda_F^2} + \log \frac{\lambda_F^2}{Q_{\text{cut}}^2}. \tag{24}$$

When computing the total structure functions, divergent behaviors are subtracted inside the PDFs. If one accounts for the gluon-initiated process as well, which is described by the P_{gq} splitting function, they will find that the total DIS F_2 structure function reads

$$F_2(x, Q^2) = \sum_i \hat{F}_2^i \left(x, \frac{Q^2}{\lambda_F^2} \right) \otimes f_i(x, \lambda_F^2). \tag{25}$$

This formula generalizes Equation 14 and Equation 15 at higher orders in perturbation theory and cures divergent behaviors with the introduction of an explicit λ_F^2 dependence in the PDFs. Since the factorization scale is unobservable, such dependence should cancel against the one acquired by the hard coefficients $\hat{F}_a$. In analogy with UV renormalization, the outcome of the factorization of collinear singularities depends on the treatment of finite terms and the regularization adopted before the subtraction of divergent contributions into the kernel. Following the analogy, an RGE describing the PDF behavior with the scale can be written as discussed in Section 1.3.

Upon renormalization and subtraction of collinear singularities of the process-dependent and scheme-dependent structure function $\hat{F}_2$, Equation 25 reads

$$F_2(x, Q^2) = \sum_i \hat{F}_2^i \left(x, \alpha_S(Q^2), \frac{Q^2}{\lambda_F^2} \right) \otimes f_i(x, \lambda_F^2) + \mathcal{O}\left(\frac{\Lambda_{QCD}^2}{Q^2} \right), \tag{26}$$

The $\hat{F}$ functions, that can be computed order by order in perturbation theory, are sometimes indicated as scheme-dependent Wilson's coefficients C_2^i. The origin of such nomenclature lies in the fact that the first discussion of Equation 26 was delivered in Mellin space – see below, e. g. Equation 37 – where the factorization theorem reads

$$F_2(n, Q^2) = \sum_i C_2^i \left(n, \alpha_S(Q^2), \frac{Q^2}{\lambda_F^2} \right) f_i(n, \lambda_F^2) + \mathcal{O}\left(\frac{\Lambda^2}{Q^2} \right) \tag{27}$$

and it can be seen as an application of Wilson's operator product expansion.

To summarize, the NLO QCD contributions to scattering amplitudes show IR divergent behaviors due to the emission of soft and collinear partons in the initial

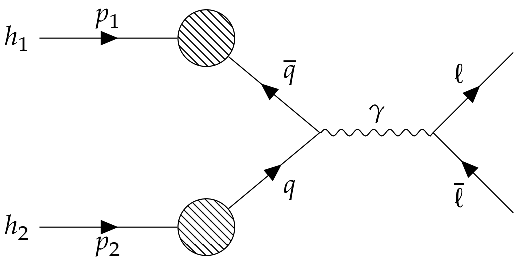


Figure 7: A pictorial diagram of a DY process. The quark anti-quark annihilation that produces a lepton anti-lepton pair is initiated inside the blobs with QCD perturbative corrections.

state. The integration over small angles, or small k_t, can be regularized with a cutoff and separated from the perturbative scale by means of logarithm properties of Equation 24. This yields a factorization of physical cross sections into perturbative and non-perturbative contributions. As it happens with renormalization, the price for removing the divergences is the dependence of the observables from the factorization scale, that must come with a RGE as described in Section 1.3.

1.2.3 *Hadronic processes*

Before moving onto the λ_F^2 scale dependency of the PDFs, we shall give some insights into the application of factorization to more complicated strong processes. Indeed, DIS only involves scattering of a lepton with a hadronic target – we call it leptonic process – and therefore can be solved with the computation of a single set of PDFs. Every time collisions happen between two hadronic states, such as the so-called DY process shown in Figure 7, we say that the process is hadronic.

The study of hadronic processes plays a key role in experimental physics. From the standpoint of PDF determination, data coming from hadronic processes, such as high-energy hadron-hadron collisions or hadronization of soft QCD radiation, aim at extending the kinematic coverage provided by DIS. A general hadron-hadron collision can be described as

$$h_1(p_1) + h_2(p_2) \to W(Q) + X, \tag{28}$$

where the incoming hadrons produce a final state composed by en exclusive part W with invariant mass Q^2 and an inclusive part X. In Equation 28, the W state produced can generally be a non-strongly interacting state – such as a weak boson or the Higgs boson – or a strongly interacting heavy quark pair or jet.

During a hadronic process, both hadrons contribute with their own PDFs to the non-perturbative region of the scattering amplitude and the hard cross section can be computed at fixed order in α_S from QCD corrections to parton-parton scattering. This means that there exist two sets of PDFs, f_{i/h_1} and f_{j/h_2}, respectively describing the partons of the first and the second hadron that enter the process. It can be shown that factorization accounts for both families, and therefore the observables are computed with the following formula:

$$\begin{aligned}\sigma(p_1,p_2,Q) = \sum_{i,j}\int_{Q^2/s}^{1} dx_1\, dx_2 \\ f_{i/h_1}(x_1,\lambda_F^2)\,\widehat{\sigma}_{ij}\left(\alpha_S(Q^2),\frac{Q^2}{sx_1x_2},\frac{Q^2}{\lambda_F^2}\right) f_{j/h_2}(x_2,\lambda_F^2) + \mathcal{O}\left(\frac{\Lambda^2}{Q^2}\right)\end{aligned} \tag{29}$$

where $s = (p_1+p_2)^2$ is the centre-of-mass energy.

In conclusion, we can say that factorization for hadronic processes consists in computing physical observables upon convolution of PDFs with the hard partonic cross sections $\widehat{\sigma}_{ij}$. This can be schematically written as

$$\sigma = \sum_{i,j} f_{i/h_1} \otimes \widehat{\sigma}_{ij} \otimes f_{j/h_2}, \tag{30}$$

and it generalizes to all orders in perturbation theory.

Calculations such as the one in Equation 30 are implemented during computational approaches to the problem of PDF determination. As we shall discuss at the end of the next section, factorization can be exploited by using pre-computed perturbative informations regarding both the hard cross sections $\widehat{\sigma}_{ij}$ and the evolution of the PDFs with the factorization scale.

1.3 PDF EVOLUTION

The factorization scale dependence of the PDFs is encoded in the DGLAP evolution equations [11, 12, 13]. Following the notation introduced in Equation 13, the DGLAP equations for the i-th parton distribution follow from the RGE and read

$$\lambda_F^2 \frac{\partial^2}{\partial \lambda_F^2} f_i(x,\lambda_F^2) = \frac{\alpha_S(\lambda_F^2)}{2\pi} \sum_j P_{ij}(x,\alpha_S) \otimes f_j(x,\lambda_F^2). \tag{31}$$

They state that the evolution of PDFs with the factorization scale is determined by convolution of a matrix P_{ij}, whose entries are the hard splitting functions, with the PDFs themselves within the ordinary linear product in Mellin space. Currently, the splitting functions have been computed up to three-loops in perturbation theory [14, 15].

1.3.1 *Solution to the DGLAP equations*

We give examples of how Equation 31 can be decoupled and solved in both analytical and numerical fashion. Needless to say, the indices i and j of Equation 31 run over quark's flavours and gluon, and therefore the matrix P_{ij} can be thought in terms of blocks that possess different symmetrical features. These can be investigated by imposing physical requirements such as charge conjugation invariance and flavour symmetry, leading to the following identifications:

$$P_{qq} = P_{\overline{q}\overline{q}}, \quad P_{\overline{q}q} = P_{q\overline{q}}, \quad P_{qg} = P_{\overline{q}g}, \quad P_{gq} = P_{g\overline{q}}. \tag{32}$$

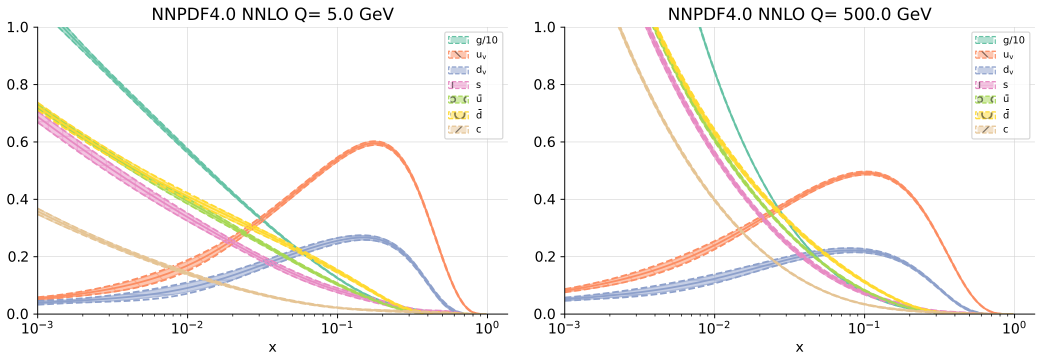


Figure 8: The NNPDF4.0 determination at $Q = 5$ and $Q = 500$ GeV, with $\alpha_S(m_Z) = 0.118$. PDFs were determined at $Q_0 = 1.65$ GeV and evolved through the APFEL framework.

A further simplification comes from expressing the matrix in a maximally diagonal basis, called evolution basis. Since the statements of Equation 32 lead to the conclusion that the rank of P_{ij} is not maximal, we cannot expect the equations to decouple completely. Indeed, it can be shown that almost every rotated flavour decouples, except from two.

If we denote the PDFs of the up (u), down (d), strange (s), charm (c), bottom (b) and top (t) quarks by f_i, we can look at the combination $f_i^{\pm} = f_i \pm \bar{f}_i$. Eleven – out of thirteen – PDFs can be obtained with suitable arrangements of the functions $f_i^{\pm}$. These are the non-singlet combinations, composed by $N_{\text{flav}} = 6$ valences $V_i = f_i^{-}$, and the five triplet distributions

$$\begin{aligned} T_3 &= u^+ - d^+ \\ T_8 &= u^+ + d^+ - 2s^+ \\ T_{15} &= u^+ + d^+ + s^+ - 3c^+ \\ T_{24} &= u^+ + d^+ + s^+ + c^+ - 4b^+ \\ T_{35} &= u^+ + d^+ + s^+ + c^+ + b^+ - 5t^+. \end{aligned} \tag{33}$$

The non-singlet distributions f_{NS} satisfy the decoupled DGLAP equations

$$\lambda_F^2 \frac{\partial^2}{\partial \lambda_F^2} f_{NS}(x, \lambda_F^2) = \frac{\alpha_S(\lambda_F^2)}{2\pi} P_{NS}(x, \alpha_S) \otimes f_{NS}(x, \lambda_F^2), \tag{34}$$

which feature the P_{NS} splitting functions, given by analogue rotations $P^{\pm}$ of the splitting functions, where P^{-} and P^{+} are used for the valences and triplets respectively.

It might seem at first glance that such change of basis, altough facilitating computations, obscures the physical meaning of the evolved PDFs. In truth, several arguments seem to agree upon the fact that the evolution basis encodes physical information that cannot be deduced from the flavour basis. For instance, valences are useful when it comes to the determination of the intrinsic composition of hadronic matter. Indeed, the peaks of the up and down valences in Figure 8 reflect the well-known fact that two u-valence quarks and one d-valence quark carry the entire proton electric charge and baryon number.

The remaining two PDF degrees of freedom, that do not decouple, can be introduced with the following argument. Equation 32 states that the gq and qg splitting functions are flavour independent, therefore any difference of quark and anti-quark distributions (f_i^-) decouples from the gluon PDF. The only combination that remains, and thus pairs with $g(x, \lambda_F^2)$, is the so-called singlet quark distribution

$$\Sigma(x, \lambda_F^2) = \sum_i f_i^+(x, \lambda_F^2), \tag{35}$$

which evolves with the gluon according to the following coupled system:

$$\lambda_F^2 \frac{\partial^2}{\partial \lambda_F^2} \begin{pmatrix} \Sigma(x, \lambda_F^2) \\ g(x, \lambda_F^2) \end{pmatrix} = \frac{\alpha_S(\lambda_F^2)}{2\pi} \begin{pmatrix} P_{\Sigma\Sigma}(x, \alpha_S) & P_{\Sigma g}(x, \alpha_S) \\ P_{g\Sigma}(x, \alpha_S) & P_{gg}(x, \alpha_S) \end{pmatrix} \otimes \begin{pmatrix} \Sigma(x, \lambda_F^2) \\ g(x, \lambda_F^2) \end{pmatrix}. \tag{36}$$

The DGLAP equations can be solved perturbatively by computing an evolution kernel which gives the PDFs at a final scale Q upon convolution with the distributions at the reference scale Q_0. It is convenient to solve the equations in Mellin space, i. e. switching to the n-moments of the distributions. Recalling the expression anticipated in Equation 27, the moments read

$$f_i(n, \lambda_F^2) = \int_0^1 dx\, x^{n-1} f_i(x, \lambda_F^2). \tag{37}$$

where $N \in \mathbb{C}$. The advantage of this transformations is that convolutions are turned into simple product in Mellin space: however, the complex n-dependence makes it difficult to transform the solutions back into x-space.

Practically, solutions to the DGLAP equations are found with numerical methods either by direct integration in x-space or by transformation into Mellin space. In the context of PDF determination, the former is adopted by the NNPDF collaboration[1] through the APFEL package [16], able to perform DGLAP evolution up to Next-to-Next-to-Leading Order (NNLO) in QCD and to LO in QED by means of higher order interpolations and Runge-Kutta techniques.

1.3.2 *Fast Kernel interface*

This final part aims at presenting how theoretical calculations based on factorization and DGLAP evolution are performed in the NNPDF fitting methodology. What follows represents a bridge that connects the final parts of the present chapter to what is discussed in the next one. It can either be seen as a conclusion of the theoretical discussion of the DGLAP equations, or a first glance at the way PDFs are fitted within the NNPDF methodology.

We shall restrict ourselves to the numerical implementation of the calculations of the DIS structure functions, since more complicated hadronic processes follow the same logic. Within the collinear QCD factorization framework, the DIS F_2 structure function can be decomposed following the general approach of Equation 30:

$$F_2(x, Q^2) = \sum_i C_i(x, Q^2) \otimes f_i(x, Q^2). \tag{38}$$

where C_i are the process-dependent Wilson coefficients and f_i are the PDFs.

[1] As presented in the abstract and extensively discussed in the next chapter, the NNPDF collaboration determines the structure of the proton using contemporary methods of artificial intelligence. This thesis is the result of work performed on the NNPDF code within the Milan group of NNPDF.

The DGLAP equations provide an operator $\Gamma(Q^2, Q_0^2)$ that evolves the PDFs from the initial parameterization scale Q_0 into the hard-scattering scale Q. In this light, Equation 38 can be rewritten as

$$F_2(x, Q^2) = \sum_i C_i(x, Q^2) \otimes \Gamma_{ij}(Q^2, Q_0^2) \otimes f_j(x, Q_0^2). \tag{39}$$

This is useful when computing theoretical predictions for datasets coming from different experiments, since the PDFs are fitted at a common scale Q_0 even if data are provided at different hard scales. The latter fact is itself the reason why the direct calculation of Equation 39 during a PDF fit is not practical. Indeed, it would require solving the DGLAP equations for each new boundary condition that comes with specific hard scales, and then convoluting with the process-dependent coefficients at the hard scale.

In order to increase computational efficiency, all the perturbative information stored inside C_i and Γ_{ij} can be pre-computed with a suitable interpolation basis. The NNPDF methodology exploits the `APFELgrid` environment to compute such information. Within this approach, the dependence on the PDFs at the input scale Q_0 is factorized as follows. First, the PDFs are expanded over a set of interpolating functions I_α that span the (x, Q^2) kinematic region, thus giving a collection of grid-valued input scale PDFs $f_i(x_\alpha, Q_0^2)$. Then, evolution is performed on such grid PDFs to give their dependence at the hard scale. This part is encoded in the composition of Mellin convolutions that are collectively stored inside the `FK Tables`, i. e. a set of functions FK_i^α that operate on the grid-valued input scale PDFs to give the structure function:

$$F_2(x, Q^2) = \sum_i \sum_\alpha FK_i^\alpha(x, x_\alpha, Q^2, Q_0^2) f_i(x_\alpha, Q_0^2). \tag{40}$$

All the information about the partonic cross sections and the DGLAP evolution is then encoded inside the `FK Tables`. Hence, the `APFELgrid` method guarantees that series of convolutions can be expressed and pre-computed in matrix multicplications, thus increasing the efficiency of DIS structure function calculations by several orders of magnitude.

1.4 QUARK MASSES

We make some remarks on the treatment of quark masses. As discussed above, quarks are confined inside hadrons and are not observed directly as physical particles. For this reason, quark masses must be determined through their influence on hadronic properties, depending upon some theoretical framework. In this light, the m_f terms in Equation 3 represent bare parameters that, based on the renormalization scheme adopted, will contribute to theoretical predictions in different ways.

We can make a coarse distinction of quarks into two families depending on their mass: light quarks and heavy quarks. The former are the up, down and strange quarks. Measures of their mass indicate that they are non-perturbative objects, in the sense that their production threshold is considerably smaller than Λ_{QCD}. On the other hand, heavy quarks – the charm, bottom and top quarks – have masses that are higher – or comparable, as it happens for the charm – than the

QCD reference scale. For this reason, depending on the scale Q of a process, each flavour of quark contributes to QCD features in different ways. For example, it is generally true that the three light quarks can be considered massless since we assume $m^2 \ll Q^2$. They emerge in loops and in real emission diagrams, therefore contributing to the running of α_S and the DGLAP evolution. On the other hand, heavy quarks can be produced in the final state only for certain scales $Q^2 \geqslant m^2$, and can be treated as massless in the limit $Q^2 \gg m^2$.

A general treatment of heavy quarks is subject to Appelquist and Carazzone's decoupling theorem [17]. The theorem states that, if a QFT features some heavy fields whose masses are very large compared to the other fields in the Lagrangian, then the Green functions for processes at energies $Q \ll M$ are the same as those obtained by simply omitting the heavy field in the QFT, up to corrections of inverse powers of the heavy mass M. As a consequence, the heavy fields decouple at low momenta except for their contribution to renormalization effects, such as the calculation of the β function. Effective theories are used to make the decoupling explicit. One example is given by the Variable Flavour Number Scheme (VFNS), which is a description of the running of α_S where the number of considered flavours, called active flavours, varies with the renormalization scale λ_R. Starting from a given number N_{flav} of flavours, whenever λ_R increases such as to cross the production threshold of the $N_{\text{flav}} + 1$-th flavour, the RGE is computed switching to $N_{\text{flav}} + 1$ active flavours.

2
PDF DETERMINATION

Knowledge of PDFs is crucial in order to make theoretical predictions of SM processes at hadron colliders. Altough PDF universality allows to use their functional form to describe all kind of strong interactions, it is not possible to determine such form from first principles since the PDF x-dependence belongs to the non-perturbative regime of QCD. Therefore, PDFs are determined by fits to experimental data.

Among the many obstacles that this approach must overcome, the most important lies in the fact that PDFs are continuous functions and, in principle, they cannot be determined from a discrete set of data. In this sense it can be said that the problem of PDF determination is somewhat ill-posed, since the PDF space has an infinite number of dimensions contrary to data space. For this reason, a particular functional form for the x-dependence of the PDFs must be chosen in terms of a set of free parameters, tuned with experimental data. This choice clearly represents a bias introduced by human prejudice that cannot be removed from the methodology, if not for the unrealistic case of an inifnite-dimensional parameter space.

Another complication is represented by the fact that experimental outputs used by fitting methodologies are measures of observable quantities – e. g. cross sections and rapidity distributions – rather than of unobservable parton distributions, and PDFs are obtained from those through factorization theorems. This gives rise to many sources of error coming from Missing Higher Order Uncertainty (MHOU) in perturbative series, as well as finite approximations made by algorithms employed to compute the DGLAP evolution from different experimental scales to the common parametrization scale.

Moreover, in order for a PDF set to be exploited in high precision physics, its determination must be delivered within some faithful representation of the uncertainties. Indeed PDFs represent one of the main sources of uncertainty in Higgs physics and in precision measurements such as the determination of the W boson mass [18]. An appropriate treatment of correlations between points coming from different datasets, as well as a correct estimation of systematic uncertainties, is essential to a well performing fit. For instance, systematic errors are determined within experimental setups and can be subject to under/over-estimations, or inconsistencies[1] between different experiments, that can bias a specific kinematic region or a PDF feature constrained by such inconsistent datasets.

In this chapter, we discuss how the methodological issues mentioned above can be overcomed within the NNPDF approach to the determination of parton distributions. As anticipated at the end of the Chapter 1, the NNPDF collaboration is one of the active groups that extract PDFs from experimental data through the exploitation of state-of-the-art computational techniques that belong to the wide family

[1] The formal definition of inconsistency between datapoints will be delivered in Chapter 4, Section 4.3.1.

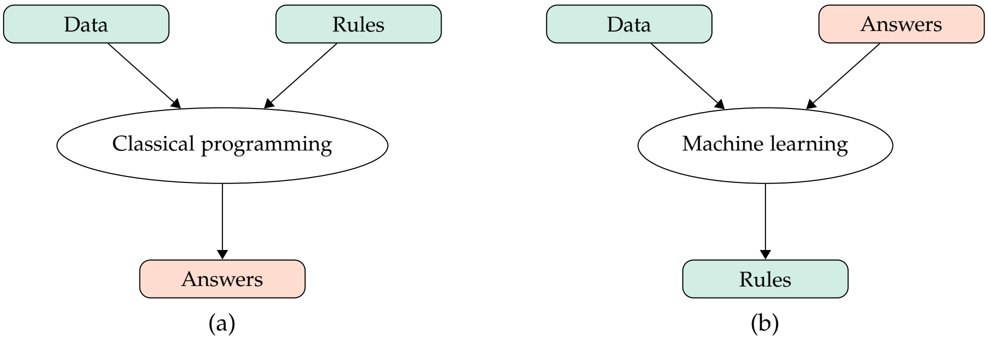


Figure 9: The schematic approach of classical programming in Physics (9a) alongside the machine learning approach (9b). Figure inspired by [23].

of machine learning. Other collaborations that provide results on this subject are CTEQ [19], MSHT [20], HERAPDF [21] and ABM [22].

The chapter is structured as follows: in Section 2.1, we deliver an introduction to machine learning focusing on deep learning and neural networks, which are specific machine learning methodologies adopted by NNPDF for the determination of proton's PDF. Altough we do not have the ambition to cover a wide subject such as machine learning in a single section, we shall provide the reader with the informations needed to understand how the NNPDF methodology works, which is explained in Section 2.2 and Section 2.3. While the former will focus on the theoretical constraints that are imposed on the PDFs by the fitting framework, the latter will deliver a detailed description of its architecture.

In the end, even if the subject is a part of the PDF fitting methodology, we shall discuss how PDF uncertainties are estimated by NNPDF and highlight the differences with the methods employed by other collaborations. We reserve Section 2.4 to this aim.

2.1 NEURAL NETWORKS

Physicists exploit computer science for a wide variety of tasks. For instance, they can determine the evolution of a dynamical system by sampling probability density distributions of its Hamiltonian through Monte Carlo techniques. The same methods are used in lattice QCD for a non-perturbative approach to strong interactions. Every time a simulation is performed, computers are provided with known theoretical rules and a set of data to be processed according to such rules, and outputs are specific answers that depend on the input data. The assets of these methodologies can be easily identified: provided that the underlying rules are correct and correctly implemented, the outcomes guarantee the entire knowledge of specific states of a system and can be used to design and test physical models.

The problem of PDFs determination, however, does not belong to such paradigm. Indeed, PDFs themselves are the rules according to which experimental data are generated by Nature and their shape has to be be deduced through machine learning methods.

2.1.1 *Machine learning*

Machine learning arises from the necessity to interpret experimental data and find statistical structures in them to eventually learn the underlying set of rules that govern the measured data. One can think of it as the opposite of simulation algorithms, as pictured in Figure 9.

A typical machine learning system is trained, rather than explicitly programmed, on a set of input data points and on examples of the expected outputs. For instance, the input data could be images of pets that need to be classified into dogs, cats and iguanas. Likewise, inputs may be profiles, i. e. a collection of parameters yielding a given rating in the real subset $[0, 1]$. The former examples falls under the category of classification problems, while the latter is a typical example of regression.

In both situations, the input dataset shows the same features: it is a set of pairs (x, y) describing the input x and the expected output y that reads

$$\mathcal{D} = \{(x, y) \,|\, x \in X,\, y \in Y\}, \tag{41}$$

where $Y \sim [0, 1]$ in a regression problem and $Y = \{1, 2, \dots, N_{\text{class}}\}$ for classification. The machine learning model is trained on the input data and produces an output, i. e. a prediction $\widehat{y}(X, \theta)$, that depends on a set of paramters θ which are recurrently optimized by the model itself. The optimal set of parameters is found in a way such that it corresponds to an output $\widehat{y}$ that represents the data provided. This involves defining a function $\mathcal{L}$, called loss or cost function, and finding the set of parameters that minimize it:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\widehat{y}(X, \theta), y). \tag{42}$$

The nature of the cost function depends on the specific task that a machine learning system must achieve. Maximum likelihood methods are usually compatible with the least squares approach, where

$$\mathcal{L}(\widehat{y}(X, \theta), y) = (\widehat{y}(X, \theta) - y)^2, \tag{43}$$

but classification algorithms are likely to adopt some characteristic function such as the Hinge loss

$$\mathcal{L}(\widehat{y}(X, \theta), y) = \max(0, 1 - y\widehat{y}(X, \theta)). \tag{44}$$

The minimization of the cost function is typically achieved through numerical routines, mainly using gradient-based or Genetic Algorithms (GAs). GAs import natural adaptation phenomena into computer science following simple evolution rules: populations, i. e. several instances of fitting parameters, evolve by means of random variations such mutation and recombination, followed by natural selection of the fittest as pictured in Figure 10. Such search problems can often benefit from an effective use of parallelism and do not require any assumption on the cost function used to determine the evolution.

GAs have been exploited by the NNPDF collaboration over the past decades and they have now been replaced by gradient-based search methods. The choice is dictated by the quick developement of minimization algorithms within `Python` machine learning libraries such as `TensorFlow` [24]. Gradient-based methods, contrary to GAs, require that the cost function is differentiable as a function of the model parameters.

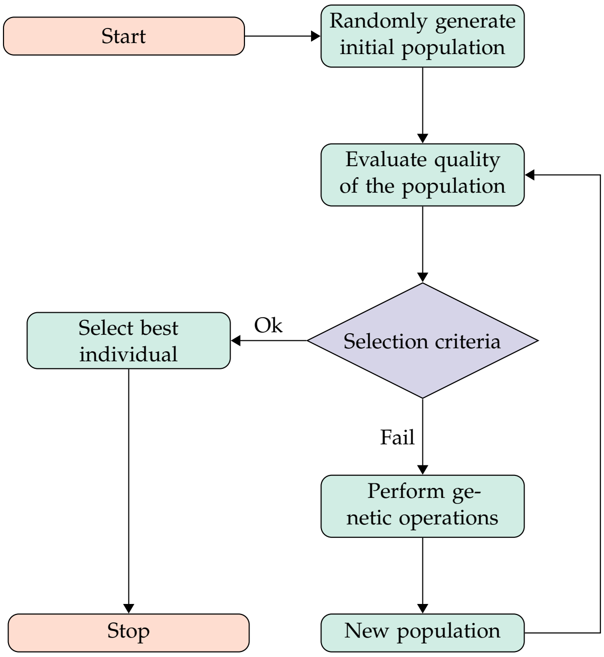

Start
Randomly generate initial population
Evaluate quality of the population
Selection criteria
Ok
Select best individual
Fail
Perform genetic operations
New population
Stop

Figure 10: A typical GA flow chart. Populations evolve according to random mutations of the fittest individuals and, when the stopping criterion is reached, the best individual is chosen.

The simplest of these algorithms is Gradient Descent (GD), where the parameters are updated at each step t, called epoch, according to the direction of the gradient evaluated on the previous epoch's configuration:

$$\begin{aligned} g_t &\leftarrow \nabla_{\theta_{t-1}} \mathcal{L}(\widehat{y}(X, \theta^{t-1}), y) \\ \theta_t &\leftarrow \theta_{t-1} - \eta g_{t-1}, \end{aligned} \tag{45}$$

where η is a learning rate. Small learning rates correspond to – sometimes excessively – slow learning that is guaranteed to converge to a minimum of the loss function. On the other hand, bigger learning rates may not be such that the algorithm ever converges to a minimum. Moreover, GD algorithms as defined above are usually found to remain stuck on saddle points, thereby returning a set of parameters which do not correspond to the global minimum of the cost function.

A wide class of more efficient gradient-based minimization algorithms that aim at improving GD goes under the name of Stochastic Gradient Descent (SGD). These methods apply classical GD to small subsets of the training set, called batches, that are sampled from the entire dataset with stochastic techniques. The parameter update is identical to the one given in Equation 45, except from the fact that the cost function is now evaluated separately for each one of the N_{batch} batches:

$$g_{t-1} \leftarrow \nabla_\theta \sum_{k=1}^{N_{\text{batch}}} \mathcal{L}_k(\widehat{y}(X_k, \theta_{t-1}), y_k). \tag{46}$$

For this reason, the parameters are updated N_{batch} times every epoch.

Variants of this algorithm aim at improving efficiency with the introduction of momentum terms, or real-time adaptation of the learning rate. For instance,

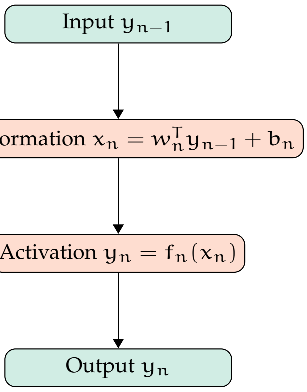

Input y_{n-1}
Transformation $x_n = w_n^T y_{n-1} + b_n$
Activation $y_n = f_n(x_n)$
Output y_n

Figure 11: Schematic description of the n-th layer of a neural network. The information is recieved from the previous layer and rotated according to the layer's weigths and biases. In the end, the output is carried to the next layer after passing through a non-linear activation function.

the latest NNPDF exploits the Nestorov momentum algorithm, or `Nadam`, with the following parameter update:

$$\begin{aligned} g_t &\leftarrow \nabla_{\theta_{t-1}} \mathcal{L}(\widehat{y}(X, \theta^{t-1}), y) \\ m_t &\leftarrow \mu m_{t-1} + (1-\mu) g_t \\ n_t &\leftarrow \nu n_{t-1} + (1-\nu) g_t^2 \\ \theta^t &\leftarrow \theta^{t-1} - \eta \frac{m_t}{\sqrt{n_t + \epsilon}}. \end{aligned} \tag{47}$$

The choice of a specific optimizing algorithm – usually called optimizer – has an impact on the machine learning model used for PDF determination. The choice of the NNPDF methodology to use `Nadam` is the consequence of a specific tuning of the parameters that define the architecture of the methodology, as we will discuss in Section 2.1.4.

2.1.2 *Deep learning*

Deep learning is a subfield of machine learning. It is a mathematical framework where multiple layers of representations of data are successively fitted according to classical machine learning methods, such as the minimization of a suitable cost function. The layered structure, from which the word deep comes from, is called neural network.

Neural networks function as in Figure 9: they are fed with some input data and produce predictions that are compared with the measured results, hence updating their parameters to find the best representation of the given data. The parameters, collectively indicated as θ in our general introduction of the previous section, are divided in two classes: the weights w – and biases b – and the thresholds. These parameters are linked to the successive linear transformations that the network performs on the input data as the information travels through the layered structure, as we shall discuss below.

In this section we shall describe the structure of a fully-connected Feed Forward Neural Network (FFNN), as implemented by the NNPDF collaboration within

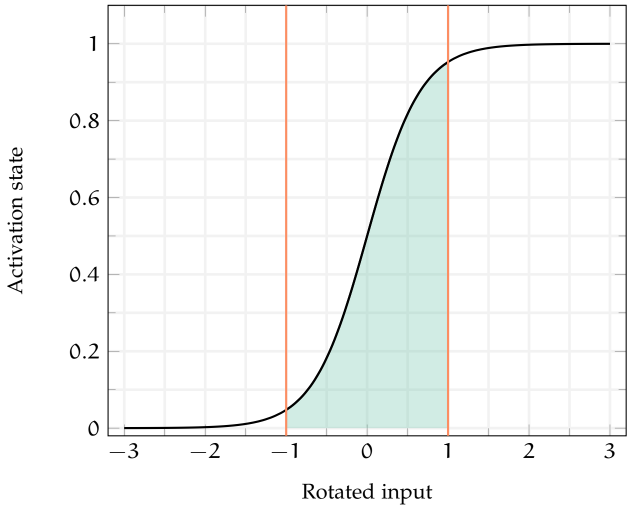


Figure 12: The sigmoid activation function. The region where the function is close to zero, at the left of the activation zone, represents the inactive state.

the problem of determination of PDFs. These represent the simplest forms of neural network and their functionality is identical to more complex models, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs). In general, the choice of a specific network is subject to its task: for instance RNNs are useful to analyse time series, i. e. data that depend on past instances of themselves, while CNNs are adopted in image classification or pattern recognition tasks.

As mentioned before, neural networks are built upon multiple layers stacked on top of each other. A layer can be identified with a set of weigths that implement a linear transformation on the input. In this context, the model learns by finding a set of values for each layer's weigths such that input data are correctly mapped to their target values by the network. The first layer of a neural network, called the input layer, implements the first of the sequence of linear transformations, which are carried on by the middle layers, also called hidden layers. Each hidden layer receives the information from the previous one and, after applying a linear transformation, it filters the rotated input with a non-linear activation function described by a set of parameters called thresholds, as pictured in Figure 11.

With the notation of the figure, we can give a formal definition of the action of a neural network on the input data. We define a FFNN of depth d as a sequence of layers $\ell_1, \dots, \ell_d$, each one endowed with a set of weigths w, biases b and an activation function f, whose parameters are the thresholds. The information travels through the network with iterative applications of the activation functions: if x_0 is the input dataset, then the prediction $\hat{y}$ of the network is

$$\hat{y} = f_d(b_d + w_d^T f_{d-1}(\dots f_2(b_2 + w_2^T f_1(b_1 + w_1^T x_0)))). \tag{48}$$

The choice of the activation function can be optimized for a specific task. In general, the activation function's domain is characterized by two regions corresponding to a binary activation state, i. e. on and off. However, as pictured in Figure 12, usual activation functions provide a third region which is used to continuously

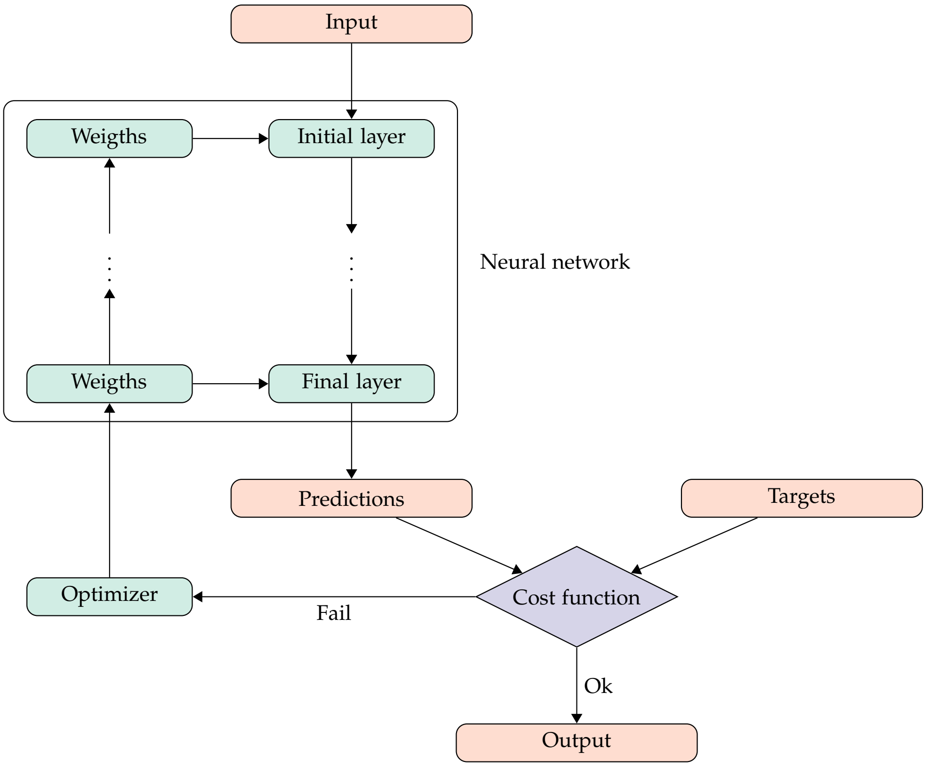


Figure 13: The backpropagation algorithm flow chart adopted by neural networks. In the first run, weigths are randomly generated and the stopping criterion is bypassed.

separate the two binary states, commonly named activation region. The activation can be non-linear, as it happens for the sigmoid

$$f_d(x) = \frac{1}{1+e^{-x}}, \tag{49}$$

or rectified as a ReLU

$$\mathrm{ReLU}(x) = \max\{0, x\}. \tag{50}$$

Once the output is produced according to Equation 48, the network is trained by comparing such output to the targets associated with the input data. This involves defining a cost function and a method that searches its minima through the update of the parameters. However, since a neural network is characterized by multiple layers of parameters – weigths and thresholds – the optimization algorithm must propagate the information on the updates backwards through the network. Specifically, gradient-based minimization algorithms such as the ones described in Section 2.1.1 are required to compute derivatives of the composition of several functions, as in Equation 48. This is done via the chain rule by computing the contribution to the cost function from the final layers back to the starting ones, with a procedure called backpropagation. State of the art backpropagation algorithms that exploit symbolic differentiation are implemented by machine learning `Python` libraries, including `TensorFlow`, and their effeiciency is one of the main reasons for the code migration implemented within the latest NNPDF release. A schematic representation of how the algorithm functions within a neural network is pictured in Figure 13.

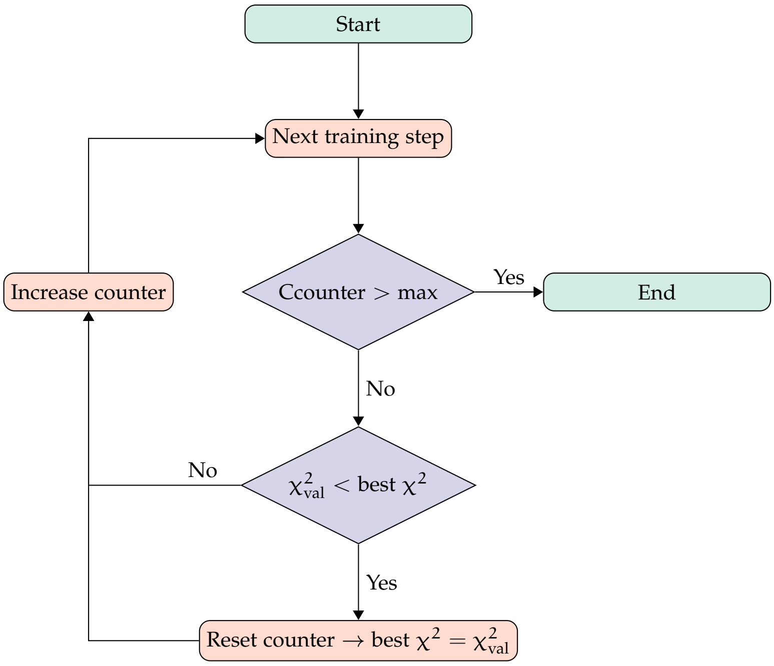

Start
Next training step
Ccounter > max
Yes
End
No
Increase counter
No
χ^2_{val} < best χ^2
Yes
Reset counter → best $\chi^2 = \chi^2_{\text{val}}$

Figure 14: Flowchart describing the patience algorithm used to determine the optimal length of the training procedure. Inspired by [25].

2.1.3 *Early stopping*

A fundamental part of the optimization strategy deployed by a machine learning method is represented by the way it avoids overfitting. Also referred to as overlearning, overfitting is the phenomenon by which a machine learning method delivers wrong predictions when presented with new data, i. e. data that have not been used for its training. There exist a family of criteria, that can be implemented during the minimization of the cost function, which aim at avoiding this phenomenon.

Frequently in machine learning, overfitting is avoided by performing the cost function minimization on a randomly selected subset of the input dataset, called training dataset and indicated with $\mathcal{D}_{\text{tr}}$. The remaining fraction of data, which belongs to the validation dataset $\mathcal{D}_{\text{val}}$, plays the role of a control sample and is used to monitor the training process. Let us indicate with $\mathcal{L}_{\text{tr}}$ and $\mathcal{L}_{\text{val}}$ the cost function evaluated respectively on the training and validation datasets.

While $\mathcal{L}_{\text{tr}}$ is generally decreasing as the number iterations grows, the usual trend of $\mathcal{L}_{\text{val}}$ shows the presence of a minimum reached at a specific iteration of the minimization algorithm. This suggests that, starting from the minimum of the validation cost function, the methodology has started learning the noise in its training data. Therefore, the presence of a minimum in $\mathcal{L}_{\text{val}}$ suggests that the optimization algorithm should be stopped at that point. When the optimal stopping point is defined as the global minimum of $\mathcal{L}_{\text{val}}$, computed over a large fixed number of iterations, the strategy is called look-back. Instead, if the minimization is stopped when the validation loss no longer improves for a defined number of iterations, we say that a patience algorithm has been deployed. The latter, which is adopted by the latest NNPDF determination, is schematized in Figure 14.

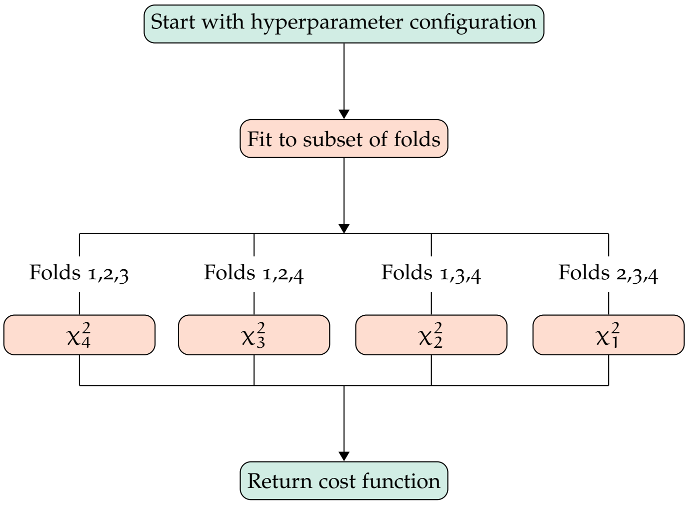


Figure 15: Diagrammatic representation of the k-fold algorithm used for the hyperparameter optimization inspired by [25].

We point out here that ulterior checks are performed within the NNPDF minimization strategy, corresponding to supplemental decision node – purple diamond-shaped ones in Figure 14. We shall postpone the discussion on such additional requirements to the section devoted to NNPDF architecture.

2.1.4 *Hyperparameter optimization*

The architecture of a machine learning system is specfied by a set of features such as the learning rate and batch size of a gradient-based optimization, or the number and size of hidden layers in a neural network. In order to differentiate them from the parameters that are optimized during the learning process, i. e. the weights and thresholds that are collectively indicated as θ, such features are called hyperparameters and indicated with capital letters Θ. The choice of hyperparameters is crucial for the determination of the best model that fits the input data: for instance, we already mentioned that the choice of the final layer's activation function is subject to the shape of the target data.

Hyperparameter setups usually depend on human prejudice and experience and this represents a potential source of bias: for this reason, it is ideal to determine them using an automated and consistent methodology, that is, a hyperparameter scan. The basic idea behind these scans is to generate different instances of the hyperparameter set, choose a suitable figure of merit and then train models with different hyperparameters looping over such instances. The hyperparameters of the model which best performs on the chosen metric are then selected as input parameters to build the architecture of the machine learning model.

Among the motivations for choosing an automated hyperparameter optimization procedure is the fact that there is usually a considerable degree of correlation between hyperparameters and therefore one cannot be determined independently from the others. For this reason, methods such grid searches in hyperparameter space are preferred to single tuning of each subspace, even if the computational cost can significantly increase. Regardless of the way the hyperscan is performed, the metric adopted should be such that it does not lead to over-trained hyperparameters. The training cost function $\mathcal{L}(\widehat{y}(X,\theta), y)$ is the best metric for the task, since it validates the model architecture in the same way the model is trained. However, it is subject to over-learning problems that can be avoided with cross-validation methods, such as the ones introduced in Section 2.1.3.

The hyperoptimization of the NNPDF methodology adopts a k-folding cross-validation algorithm, which is schematized in Figure 15. A k-folding algorithm generates a large number of hyperparameter instances Θ and partitions the input dataset into N_{folds} distinct subsets $\mathcal{D}_k \subset \mathcal{D}$. The basic idea is to produce N_{folds} fits for each hyperparameter configuration: in each of the fits, one fold is left out, and the remaining folds are combined into a dataset which is then separated in training and validation subsets. For each hyperparameter configuration, N_{folds} cost functions $\mathcal{L}_k$ are computed for each fold that has been left out of the fitting procedure.

The overall cost function $\mathcal{L}(\Theta)$ is then computed as the mean of each $\mathcal{L}_k$ cost function coming from the N_{folds} fits:

$$\mathcal{L}(\Theta) = \frac{1}{N_{\text{folds}}} \sum_{k=1}^{N_{\text{folds}}} \mathcal{L}_k. \tag{51}$$

The best hyperparameter configuration is then selected by minimization of the overall cost function:

$$\Theta_{\text{best}} = \arg\min_{\Theta} \mathcal{L}(\Theta). \tag{52}$$

The NNPDF methodology exploits the `hyperopt` [26] library to perform the hyperparameter scan using a Bayesian optimization algorithm. Contrary to a blind grid search, a Bayesian optimization algorithm updates a prior probability distribution of the score given the configuration for each hyperparameter as models are trained. The possibility to perform such automated search of the hyperparameter space is a consequence of the improved computational performance of NNPDF4.0, mostly due to the significant changes in the architecture of the neural network and the optimization strategy, as explained in detail in the following sections.

2.2 THEORETICAL CONSTRAINTS

We review the general structure of the PDF parametrization adopted by the NNPDF fitting methodology, and the theoretical constraints imposed upon it. Precisely, we discuss the parametrization basis, sum rules and positivity and integrability of the fitted PDFs.

2.2.1 *Parametrization basis*

A PDF analysis requires a choice of basis, i. e. a set of linearly independent flavour combinations that are parametrized at a input scale Q_0. A priori, the number of independent PDFs is 13. In truth, under the hypothesis that heavy quarks are generated by the perturbative evolution, one can reduce themselves to fitting a smaller number of independent PDFs.

As it is stated in Chapter 1, a possible way to treat quark masses is to consider different numbers of active flavours depending on the kinematic region. In the following, we assume that the three light quarks always contribute to the DGLAP evolution, therefore their PDFs are independently parametrized. Heavy quarks could then be parametrized with the introduction of PDFs that are set to zero below the mass threshold and evolve according to the DGLAP equations in the asymptotic sector. For instance, the NNPDF determination parametrizes the charm PDF just above its threshold and do not conisder the top and the bottom since it is assumed that their non-perturbative component is negligible.

The input scale adopted by the NNPDF collaboration in the latest releases [27, 25] is chosen at $Q_0 = 1.65$ GeV. The gluon and the three lightest quarks are fitted independently with their antiparticles, including the total charm PDF for an overall eight – out of thirteen – independent PDFs. Explicitly, the NNPDF3.1 determination adopted seven flavours from the evolution basis, including an independently parametrized total charm PDF c^+, in the assumption that the charm valence c^- would vanish at the input scale:

$$\mathcal{B}_{\text{NNPDF}} = \{g, c^+, \Sigma, T_3, T_8, V, V_3, V_8\} \tag{53}$$

where the triplet distributions are written in Equation 33, and the valences combinations are

$$\begin{aligned} V &= u^- + d^- + s^+ \\ V_3 &= u^- - d^- \\ V_8 &= u^- + d^+ - 2s^- . \end{aligned} \tag{54}$$

The latest release NNPDF4.0 adopts T_{15} instead of c^+ as supplement to the evolution basis: as one can see from Equation 33, this choice is completely consistent with the assumption that that the charm valence vanishes at Q_0.

2.2.2 *Sum rules*

The NNPDF collaboration aims at determining proton's parton distributions. The proton is composed by two up-valence quarks and one down-valence quark which carry the entire proton electric charge and baryon number. Nevertheless, all types of quark flavours can be found inside the proton as quantum effects coming from loop corrections in perturbation theory giving rise to $q\bar{q}$ pairs.

The following sum rules translate the previous sentences into a constraint on the first momenta of the valences:

$$\int_0^1 dx\, x u^-(x, Q) = 2, \quad \int_0^1 dx\, x d^-(x, Q) = 1, \tag{55}$$

and

$$\int_0^1 dx\, x s^-(x, Q), \int_0^1 dx\, x c^-(x, Q), \int_0^1 dx\, x b^-(x, Q), \int_0^1 dx\, x t^-(x, Q) = 0. \tag{56}$$

Provided that these relations hold at the input parametrization scale Q_0, DGLAP evolution equations ensure that they will hold at any scale. Valence sum rules can also be implemented in the evolution basis, where

$$\int_0^1 dx\, V(x, Q) = \int_0^1 dx\, V_8(x, Q) = 3, \quad \int_0^1 dx\, V_3(x, Q) = 1, \tag{57}$$

and the others momenta vanish.

An additional degree of freedom is fixed imposing the so-called momentum sum rule, i. e. requiring that the total momentum of the proton is realized summing over the partons momenta:

$$\int_0^1 dx\, x\Sigma(x,Q) = 1 - \int_0^1 dx\, xg(x,Q). \tag{58}$$

This agrees with the assumption that PDFs should continuously decrease towards zero when $x \to 1$, which is a consequence of their continuity at $x = 1$. Additionally, non perturbative arguments suggest that the behavior at $x = 1$ should be a power law. Therefore, a suitable parametrization of the small and large x regions yields

$$f(x) = N\, x^{\alpha}(1-x)^{\beta} p(x), \tag{59}$$

where N is a normalization that can be determined from the sum rules and $p(x)$ carries all the unknown dependence of the PDFs on the momentum fraction x.

Equation 59 represents a preprocessing tool that should speed up the training of the fitting framework, provided that it does not bias the results. To this aim, the exponents α and β are iteratively determined in a self-consistent way, as explained in [28].

2.2.3 *Positivity*

Ulterior theoretical restrictions can be imposed upon the PDF fitting framework by means of positivity checks. Specifically, two related quantities are involved in positivity constraints: observables such as hadron-level cross sections and PDFs themselves.

The former are non-negative quantities because they are probability distributions, and should remain positive at any given order in perturbation theory. For this reason, indirect positivity limitations can be imposed that penalize the PDFs which yield negative values for physical observables such as the DIS structure functions F_2^u, F_2^d, F_2^s and F_L, or the flavour-diagonal DY rapidity distributions $\sigma_{u\bar{u}}$, $\sigma_{d\bar{d}}$ and $\sigma_{s\bar{s}}$. This is done in the latest NNPDF release for the massless quarks, which means that a further contraint on F_2^c on the charm quark can be imposed in certain schemes.

The penalization is implemented during the cost minimization using a Lagrange multiplier, which strongly diminishes the weight of those PDF configurations leading to negative observables by adding the following term to the cost function:

$$\mathcal{L} \mapsto \mathcal{L} + \sum_{k=1}^{N_{\text{obs}}} \Lambda_k \sum_{i=1}^{N_{\text{pts}}} \text{Elu}_{\alpha}(-\sigma_k(f(x_i, Q^2 = 5\ \text{GeV}^2))), \tag{60}$$

with

$$\text{Elu}_{\alpha}(t) = \begin{cases} t & \text{when } t > 0 \\ \alpha(e^t - 1) & \text{when } t < 0 \end{cases} \tag{61}$$

and α is a suitable parameter. In Equation 60, N_{obs} is the number of observables upon which positivity is required and N_{pts} is the number of pseudodata points used to implement the positivity constraints. The values of x_i are given by 10 points logarithmically spaced in the small-x region ($10^{-7} \div 10^{-1}$) and 10 points linearly spaced between 0.1 and 0.9. The Lagrange multiplier Λ_k increases exponentially during the minimization and its maximum value must be chosen in a way such that the effectiveness of the constraints is sufficiently accurate. For this reason, the NNPDF methodology adopts $\Lambda_{max} = 10^{10}$ for the DY observables and $\Lambda_{max} = 10^6$ for all other positivity observables included in the analysis. The starting value of the Lagrange multiplier is not imposed, but determined during the hyperparameter optimization algorithm described in Section 2.1.4.

The choice of $Q^2 = 5$ GeV in Equation 60 is the result of fine tuning. In general, positivity is violated al small scales. For this reason, if positivity is enforced in such kinematic region, the DGLAP evolution guarantees that it will be preserved at all scales [29].

We now turn to direct positivity constraints that can be imposed on PDFs. The LO pQCD treatment of hadron processes interprets PDFs as probability distributions and, therefore, it might seem that they cannot be negative. However, beyond LO, PDFs are defined through collinear subtractions on parton-level cross sections that depend on the factorization scheme adopted and thus they may be negative for specific forms of the subtraction. There exist some schemes where PDF positivity is guaranteed at all orders, e. g. the physical scheme, and it was shown [30] that transforming such positive schemes into the $\overline{\text{MS}}$ factorization scheme preserves PDF positivity. For this reason, the latest NNPDF release features additional positivity constraints that apply on the massless quark and gluon PDFs directly through a penalty cost function similar to the one in Equation 60.

2.2.4 *Integrability*

The last theoretical constraint that can be imposed independently of the fitting framework is PDF integrability. The topic was already indirectly discussed in Section 2.2.2 and is a direct consequence of the sum rules imposed on the PDFs. Precisely, the valence sum rules of Equation 55 restrict the small-x behavior of the valences to

$$\lim_{x\to 0} xV(x,Q) = \lim_{x\to 0} xV_3(x,Q) = \lim_{x\to 0} xV_8(x,Q) = 0 \qquad \forall Q, \tag{62}$$

where we restricted ourselves to the distributions that enter the fitting basis in Equation 53.

Moreover, the momentum sum rule imposed on the singlet and gluon distributions implies that

$$\lim_{x\to 0} x^2\Sigma(x,Q) = \lim_{x\to 0} x^2 g(x,Q) = 0 \qquad \forall Q. \tag{63}$$

A further constraint such as the ones in Equation 62 can be imposed on the T_3 and T_8 parton distributions following some perturbative arguments, for a total of seven independent rules. However, this number is reduced to five in the implementation of the integrability restrictions, since it turns out that Equation 63 is automatically satisfied when fitting to the experimental data.

BASIS	PDFS	GRID POINTS
Evolution	$f_k = T_3, T_8$	$x_i = \{10^{-9}\}$
Flavour	$f_k = T_3, T_8, V, V_3, V_8$	$x_i = \{10^{-9}, 10^{-8}, 10^{-7}\}$

Table 1: The integrability constraint specifics adopted by the latest NNPDF in evolution and flavour basis.

The PDF integrability constraints are imposed through Lagrange multipliers in the same way as positivity's. The total cost function is supplemented with the following contribution:

$$\mathcal{L} \mapsto \mathcal{L} + \sum_k \Lambda_k^{(\text{int})} \sum_{i=1}^{N_{\text{pts}}} \left[x f_k(x_i, Q_i^2) \right]^2 , \tag{64}$$

where this time, since the limit $x \to 0$ is discussed, the points x_i are all taken in the small-x region. The remaining parameters of Equation 64 depend upon the choice of basis adopted and are listed in Table 1.

As it happens for the positivity constraint, the Lagrange multipliers grow exponentially during the minimization, with maximum value $\Lambda^{(\text{int})} = 100$.

2.3 THE NNPDF4.0 NEURAL NETWORK

Since the beginning of the NNPDF collaboration, PDF determination has increased in precision and reliability thanks to recent discoveries in the machine learning field and the availability of experimental measurements covering wider kinematic regions. As stated in the introduction of this chapter, several collaborations are actively delivering their own PDF sets. At the very bottom, the main difference between them stands in how they fit the unknown x-dependence of Equation 59, and how they propagate data space uncertainties into PDF space.

This section's aim is to present the NNPDF approach to the former problem. This is done through the description of the architecture of the neural network adopted by NNPDF4.0 and its main differences with previous releases. A bayesian derivation of the fit's cost function χ^2 is then presented, along with a description of the fitting framework and stopping criteria.

2.3.1 *Hyperparameters*

The NNPDF4.0 methodology adopts a single neural network to fit the unknown x-dependence of the PDFs. Its hyperparameters have been determined with a k-folding hyperoptimization, as described in Section 2.1.4. Among others, this includes the network architecture, the activation functions, the optimizer, the learning rates, and the initial values of the Lagrange multipliers.

Table 2 displays a comparison between the main hyperparameters in the last two NNPDF releases. As already mentioned, the latest NNPDF exploits the brand-new gradient-based optimizers provided by `Python` libraries such as `TensorFlow`. The main difference, however, is in the introduction of a flexible architecture that fits all the eight PDF functional forms with a single densely connected network,

PARAMETER	NNPDF4.0	NNPDF3.1
Architecture	2-25-20-8	2-5-3-1
Activation function	Hyperbolic tangent	Sigmoid
Optimizer	`Nadam`	Genetic algorithm
Learning rate	2.6×10^{-3}	–
Free parameters	763	296
Max epochs	17×10^{3}	80 (max generations)

Table 2: The principal hyperparameters of the NNPDF4.0 and NNPDF3.1 releases compared. Differences arise in every field and, as a consequence of that, the total number of free parameters is almost triplicated in the latest version.

rather than using a neural net per flavour. This increases sensitivity to cross-correlations between different PDFs.

Figure 16 shows the NNPDF4.0 neural network's architecture, with its revisited final eight-dimensional layer. The network performs fits to experimental data in order to determine the x-dependence of Equation 59. In this framework, the relation between the PDFs and the network ouptut is

$$xf_k(x, Q_0; \theta) = A_k x^{1-\alpha_k}(1-x)^{\beta_k} \mathrm{NN}_k(x, \theta), \qquad k = 1, \ldots, 8 \tag{65}$$

where NN_k denotes the activation state of the k-th neuron in the final layer.

2.3.2 *Cost function*

The cost function adopted by the NNPDF4.0 methodology is the χ^2 computed with the published experimental covariance matrix C. Specifically, it is the sum across all datasets of the Gaussian likelihood normalized by the number of datapoints

$$\chi^2 = \frac{1}{N_{\text{data}}} \sum_{i,j=1}^{N_{\text{data}}} (D-T)_i C^{-1}_{ij} (D-T)_j, \tag{66}$$

where T are the theoretical predictions computed from the neural network's output using a suitable factorization framework, and D are the experimental datapoints.

We briefly discuss how the covariance matrix is built from experimental uncertainties. These are usually encoded in three main contributions: uncorrelated errors σ^{uncorr}, correlated additive systematics σ^{add} and correlated multiplicative systematics σ^{mult}. The first ones are constructed by sum in quadrature of the published statistical errors with the uncorrelated systematics and contribute to the diagonal of the covariance matrix. The last two systematics are delivered in matrices, σ^{add}_{ik} and $\sigma^{\text{mult}}_{ik}$, where the index i refers to the experimental point and k referes to the source of systematic uncertainty detected within the experiment.

The number of additive and multiplicative systematics may depend on several aspects of the experimental measurement performed for the dataset determination,

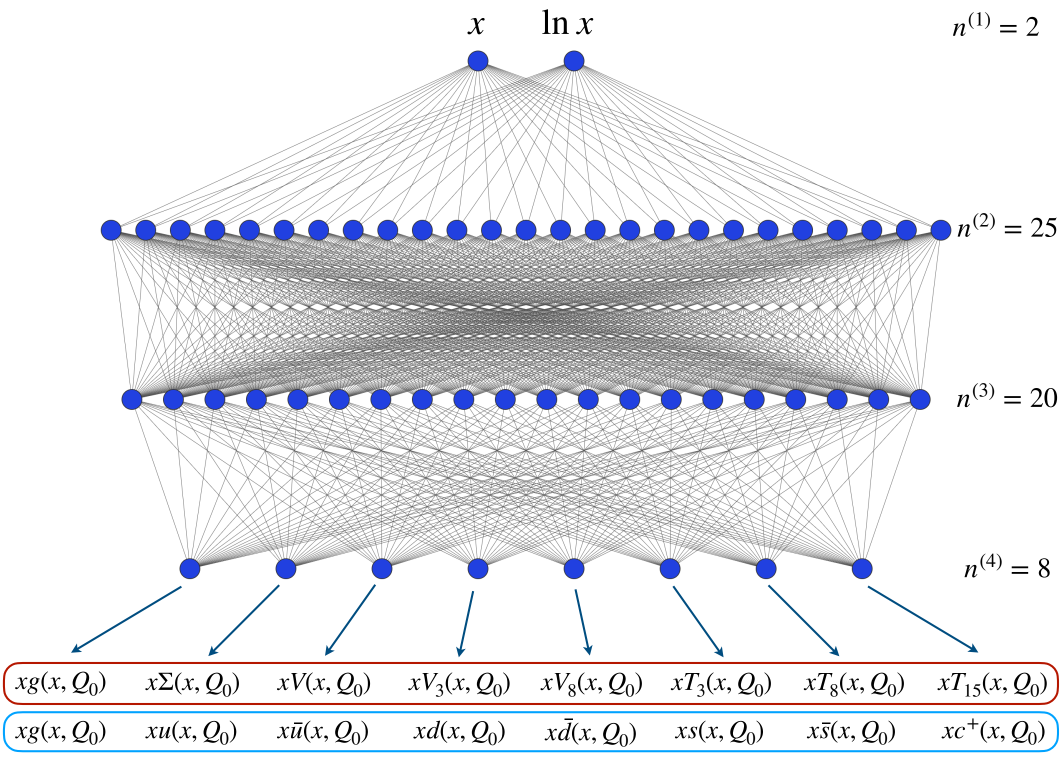

x
$\ln x$
$n^{(1)} = 2$
$n^{(2)} = 25$
$n^{(3)} = 20$
$n^{(4)} = 8$
$xg(x,Q_0)$ $x\Sigma(x,Q_0)$ $xV(x,Q_0)$ $xV_3(x,Q_0)$ $xV_8(x,Q_0)$ $xT_3(x,Q_0)$ $xT_8(x,Q_0)$ $xT_{15}(x,Q_0)$
$xg(x,Q_0)$ $xu(x,Q_0)$ $x\bar{u}(x,Q_0)$ $xd(x,Q_0)$ $x\bar{d}(x,Q_0)$ $xs(x,Q_0)$ $x\bar{s}(x,Q_0)$ $xc^+(x,Q_0)$

Figure 16: The NNPDF4.0 neural network architecture. A single network is adopted, whose outputs are the PDFs in the evolution (red box) or flavour (blue box) basis. Taken from [25].

as we shall discuss in Chapter 4. The NNPDF4.0 covariance matrix is built in the following way:

$$C_{ij} = \delta_{ij}\sigma_i^{\text{uncorr}}\sigma_j^{\text{uncorr}} + \sum_{k=1}^{N_{\text{add}}} \sigma_{ik}^{\text{add}}\sigma_{jk}^{\text{add}} + \widehat{y}_i\widehat{y}_j \sum_{k=1}^{N_{\text{mult}}} \sigma_{ik}^{\text{mult}}\sigma_{jk}^{\text{mult}}, \qquad (67)$$

where $\widehat{y}_i$ are the theoretical predictions for the observables measured by experiments. Theoretical predictions are preferred to experimental central values in the calculation of the covariance matrix in order to avoid the so-called D'Agostini bias [31]. For this reason, the expression of Equation 67 is sometimes referred to as t_0-covariance matrix in literature.

2.3.3 *Bayesian derivation of the likelihood*

We present a Bayesian argument for the specific choice of Equation 66 for the measure of the fit quality. We warn the reader that we shall make strong use of the notation introduced in this section for the rest of this work.

Factorization theorems state that there exists a forward map ϕ that, given a set of PDFs describing the non-perturbative QCD region of a strong process, computes observable quantities by means of convolutions with Wilson coefficients. We can schematize the statement in the following way. Let M be the model space, i. e. PDF space, and O the space of observables: the forward map ϕ is defined as

$$\begin{aligned} \phi &: M \to O \\ f(x, Q) &\mapsto \sigma(x, Q) = \sum_i f_i \otimes \hat{\sigma}_i . \end{aligned} \tag{68}$$

In order to give a formal definition of the forward map, we require that M and P are infinite dimensional Banach spaces.

The purpose of PDF fitting methodologies is to determine the inverse map ϕ^{-1} from experimental measurements of instances of such observables. Therefore, we must give a definition of a second map π that projects infinite-dimensional observable quantities into a finite-dimensional data space D:

$$\begin{aligned} \pi &: O \to D \\ \sigma(x, Q) &\mapsto y = \{y_i \,|\, \forall i = 1, \dots, N_{\text{data}}\}. \end{aligned} \tag{69}$$

Fitting methodologies are based on the fact that experimental values are representations of the underlying PDFs through the composition $\pi \circ \phi$. Hence, we can determine the the inverse map $(\pi \circ \phi)^{-1}$ by sampling the probability distribution in model space. This can be achieved within Bayesian statistics through maximization of an estimator that identifies the model $f^* \in M$ which is most likely to yield the observed data-space distribution. The procedure is usually referred to as Maximum A Posteriori (MAP) estimator computation and its complete analysis within the NNPDF approach to PDF determination is fully delivered in [32]. Here, we outline its main features and derive the figures of merit adopted by the latest NNPDF releases.

Experimental observations are published as a set of central values and uncertainties, reflecting the several sources of noise in measurements such as finite precision of experimental apparati. Assuming gaussian noise the prior probability distribution for the measure y of a certain observable σ, given its central value y_0 and data space covariance matrix C, is:

$$p_0(y) \propto \exp\left[-\frac{1}{2}\sum_{i,j}(y - y_0)_i C^{-1}_{ij}(y - y_0)_j\right]. \tag{70}$$

In a similar way, we can write down a prior distribution of the input model given a central value f_0 and a model space covariance matrix C'. In the gaussian assumption, this reads

$$p_0(f) \propto \exp\left[-\frac{1}{2}\sum_{i,j}(f - f_0)_i C'^{-1}_{ij}(f - f_0)_j\right]. \tag{71}$$

The cost function can be derived in this framework with a maximum likelihood approach. The PDF model f^* that is most likely to give a measured set of observables is obtained through maximization of the probability distribution in model space, marginalized over the data:

$$f^* = \arg\max_{f \in M} \int dy\, p(f \,|\, y) p_0(y). \tag{72}$$

The data prior distribution is given in Equation 70, while the probability of the model given the data can be written in terms of known quantities by means of Bayes theorem:

$$p(f\,|\,y) = \frac{p(y\,|\,f)\,p_0(f)}{\int df'\,p(y\,|\,f')\,p_0(f')}. \tag{73}$$

Discarding the normalization and the model prior, the likelihood $p(y\,|\,f)$ is given by the forward map $\pi \circ \phi(f)$ and, generally, is a distribution $\rho(y - \pi \circ \phi(f))$.

Such distribution describes the correlations between the input model and the observables induced by the forward map. Uncertainties can arise when comparing measured data with fixed-order theoretical predictions every time a methodology updates its parameters in the optimization algorithm. Specifically, theory uncertainties are results of MHOU in perturbative QCD calculations: this problem can be addressed [33] in the assumption that there exists a theory covariance matrix S such that the correlations induced by the forward map are distributed gaussianly according to

$$\rho(y - \pi \circ \phi(f)) \propto \exp\left[-\frac{1}{2}\sum_{i,j}(y - \pi \circ \phi(f))_i S^{-1}_{ij}(y - \pi \circ \phi(f))_j\right]. \tag{74}$$

We can conclude that the likelihood used as a cost function during a PDF fit generally depends upon three different covariance matrices: the data space covariance C, the theory covariance matrix S and the covariance matrix that describes the model priors C′. In this picture, Equation 73 reads

$$p(f\,|\,y) = \exp\left[-\frac{1}{2}\big|f - f_0\big|_{C'} - \frac{1}{2}\big|y - \pi \circ \phi(f)\big|_S\right], \tag{75}$$

where we introduced the compact notation

$$\big|x\big|_C = \sum_{i,j} x_i C^{-1}_{ij} x_j. \tag{76}$$

The likelihood is then given by Equation 72 through the following integration

$$f^* = \arg\max_{f\in M} \int dy \exp\left[-\frac{1}{2}\big|y - y_0\big|_C - \frac{1}{2}\big|f - f_0\big|_{C'} - \frac{1}{2}\big|y - \pi \circ \phi(f)\big|_S\right] \tag{77}$$

It can be shown [33] that, after the transformation $\Delta = y - \pi \circ \phi(f)$ and integration over $d\Delta$, the terms can be re-arranged into the following expression:

$$f^* = \arg\max_{f\in M} \exp\left[-\frac{1}{2}\big|f - f_0\big|_{C'} - \frac{1}{2}\big|\pi \circ \phi(f) - y_0\big|_{C''}\right], \tag{78}$$

with $(C'')^{-1} = C^{-1} - C^{-1}(C^{-1} + S^{-1})^{-1}C^{-1}$. This expression leads to the identification of C'' with $C + S$. In order to prove it, one can start by showing that $C^{-1} + S^{-1}$ is equivalent to $C^{-1}(C + S)S^{-1}$, and therefore its inverse must be

$$(C^{-1} + S^{-1})^{-1} = (C^{-1}(C + S)S^{-1})^{-1} = S(C + S)^{-1}C. \tag{79}$$

Then, plugging Equation 79 into the expression of C'', one finds

$$\begin{aligned}(C'')^{-1} &= C^{-1} - C^{-1}S(C + S)^{-1} = \\ &= (C^{-1}(C + S) - C^{-1}S)(C + S)^{-1} = (C + S)^{-1},\end{aligned} \tag{80}$$

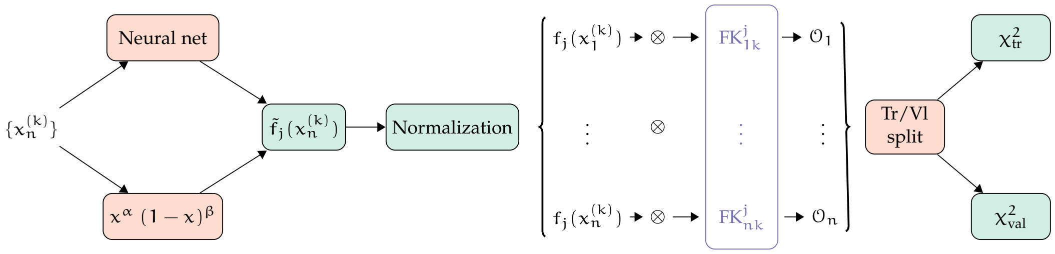


Figure 17: Diagrammatic representation of the NNPDF4.0 fitting framework and the calculation of the cost function χ^2. Each block represents an independent component of the code. Figure taken from [25].

whence $C'' = C + S$.

Through this very powerful argument, one can incorporate theory uncertainties inside the determination of the likelihood by summing the theory covariance matrix with the experimental one. The MAP PDF model f^* is then given by maximizing the logarithm of Equation 77. Since the aim of this work is not to investigate the impact of MHOUs, we shall however restrict ourselves to the situation where $S = 0$. This gives

$$f^* = \arg\min_{f \in M} \left[\left|\pi \circ \phi(f) - y_0\right|_C + \left|f - f_0\right|_{C'}. \right] \tag{81}$$

Observe that the first part that is minimized in Equation 81 is nothing else than Equation 66, with the obvious identification of $\pi \circ \phi(f)$ with the theoretical predictions made from the network's output, and y_0 with the experimental data. The presence of the second term acts as a regulator that represents all the assumptions that are made on the prior probability distributions of the PDFs by the fitting methodology. From a technical standpoint, such regulator is not implemented directly with a covariance matrix, but its effect is represented by the requirements discussed in Section 2.2.

2.3.4 *Fitting framework*

Up to this point, we described the principal components of the latest NNPDF fitting framework, i. e. the network's architecture, the PDF basis chosen, the hyperparameter selected through hyperscan optimization, the preprocessing factors and the cost function adopted. Here we show how these ingredients enter the fitting procedure and the method adopted to avoid over-learning.

The NNPDF4.0 code is characterized by a modular structure, as pictured in Figure 17. The figure shows how the fitting code evaluates the physical observables in terms of the input fitted PDFs. Starting from a matrix of $x_n^{(k)}$ Bjorken variables, where n labels the experimental dataset and k the node in the x-grid, the code evaluates the neural network and the preprocessing factors to construct a PDF which is subsequently normalized according to the sum rules presented in Section 2.2.2. This produces the PDFs at the input scale, that are convoluted with the FK tables presented in Section 1.3.2 to give the physical observables that enter the calculation of the cost function.

As we already mentioned in our introduction to machine learning, the χ^2 is computed for two separate subsets of the entire dataset, that is, training and validation sets. This division advantages the implementation of a cross-validation method to avoid overfitting through the patience algorithm, presented in Section 2.1.3. The algorithm is itself subject to the hyperoptimization procedure, with a hyperoptimized patience of 1/10-th of the maximum number of epochs delivered in Table 2.

Once the fit quality is determined to be satisfactory, the fitted PDFs are subject to further post-fit selection checks. According to these checks, the PDF ensamble is pruned and biased PDF are discarded. This ensures that the final delivery of the fit satisfies all the known theoretical constraints on PDFs that, even though already penalized in the fit's cost function, might have had statistical weight during the fitting procedure.

2.4 ERROR PROPAGATION

As discussed in the introduction of Section 2.3, what defines a PDF fitting methodology is how it provides the unknown x-dependence of Equation 59, and how it consistently determines PDF uncertainties. This final section is devoted to the discussion of the latter topic.

It is commonly agreed upon that today's PDF determinations should be able to provide percent-level accuracy in order to be best exploited in frontier high-energy physics. We can summarize the problem within Bayesian statistics as follows. With the notation introduced in Section 2.3.3, an estimation of the central value and uncertainty of a generic observable $\sigma \in \mathcal{O}$ is given by the posterior model distribution $p(f\,|\,y)$:

$$E[\sigma] = \int df\, \sigma[f]\, p(f\,|\,y), \quad \text{Var}[\sigma] = \int df\, (\sigma[f] - E[\sigma])\, p(f\,|\,y). \tag{82}$$

From a practical standpoint, the computation of such integrals is far from trivial. Indeed, knowledge of a closed analytic form is subject to the choice of a specific observable and specific implementations should therefore be incorporated within the methodology. Moreover, we have seen that giving an explicit form of the probability $p(f\,|\,y)$ is hindered by the difficulty to understand the shape of the model prior. For this reason, estimates of PDF uncertainties are provided in different ways.

2.4.1 *The Hessian method*

The Hessian method is adopted by several PDF determinations for error propagation. Recalling its main features here does not have the sole purpose of highlighting its differences with the NNPDF approach. Indeed, the understanding of the Hessian method is necessary to the developement of the arguments that motivate the results presented in this thesis.

The Hessian method estimates uncertainties in the model's parameters space in terms of displacements from the optimal parameters that induce fluctuations in the fit's cost function. In other words, it computes the uncertainty of a quantity – such as a PDF at a given x and Q^2, or an observable – as the linear propagation of the parameter space shifts in the model/data space respectively.

In order to deliver a formal description of the method, we make the following definitions. Let Ω be the parameter space and $\chi^2 \colon \Omega \to \mathbb{R}$ the cost function, such that its global minimum is reached at $\theta_0 \in \Omega$. Assuming that such function is quadratic in a neighborhood $U(\theta_0) \subset \Omega$, we can write

$$\Delta\chi^2 = \chi^2(\theta) - \chi^2(\theta_0) = \frac{1}{2}\sum_{i,j}(\theta-\theta_0)_i H_{ij}(\theta-\theta_0)_j. \tag{83}$$

Optimal parameters θ_0 yield an optimal representation of a PDF set $f_0 = f(\theta_0)$ through a function $\Omega \to M$ – where M is defined in Section 2.3.3 – whose restriction on $U(\theta_0)$ is bijective. One can invert such function and, upon composition with χ^2, find a map $M \to \mathbb{R}$. Then, they can ask how errors are propagated from the fluctuations of the cost function to fluctuations of PDFs by mapping hypersurfaces with the inverse function $\mathbb{R} \to M$. However, since the cost function is quadratic by hypothesis in $U(\theta_0)$, the inverse only exists upon restriction of the domain: as we shall discuss below, this corresponds to finding pairs of eigenvector PDF sets $f_0^{\pm}$ for each eigenvector of the parameter-space covariance matrix.

Formally, the Hessian approach can be derived in $U(\theta_0)$, which can be spanned with the eigenvectors of the covariance matrix $C = H^{-1}$, that is clearly symmetric upon requirement that χ^2 be at least C^2 in $U(\theta_0)$. Provided Ω with an inner product,

$$\langle\theta_1, \theta_2\rangle = \sum_i (\theta_1)_i(\theta_2)_i, \tag{84}$$

the spectral theorem states that there exists an orthonormal basis of the parameter space consisting of eigenvectors v_n of C, i. e. $\langle v_n, v_m\rangle = \delta_{mn}$. Such eigenvectors can be rescaled with the square root of their eigenvalue, $v_n \mapsto v_n/\sqrt{\lambda_n}$, so that their quadratic functions will have a simple normalization.

Every $\theta \in U(\theta_0)$ can be decomposed on the diagonal basis with suitable coefficients $\alpha_n = \langle\theta, v_n\rangle$ to find the following expression for Equation 83:

$$\Delta\chi^2 = \frac{1}{2}\sum_{n,m}\sum_{i,j}(\alpha_n v_n)_i H_{ij}(\alpha_n v_n)_j = \frac{1}{2}\sum_n \alpha_n^2. \tag{85}$$

This is the equation of a hypersphere in $U(\theta_0)$ with radius $\Delta\chi^2\sqrt{2}$ that defines the allowed parameters displacements yielding tolerated fluctuations in the cost function. The border parameters $\theta_n^{\pm} = \theta_0 \pm t v_n$ define pairs of eigenvector PDF sets $f_n^{\pm}$ through the aforementioned bijective map $U(\theta_0) \to M$.

Values of t can increase until they reach the value of a "tolerance" T, which defines the region of acceptable fits, and is greater than one every time the quadratic hypothesis fails to describe the behavior of the cost function about its global minimum. That being the case, iterative adjustments of t are performed [34].

Once the border is defined in PDF space, any quantity $F(f)$ – included PDFs themselves – can be provided with symmetric uncertainties given by

$$\sigma_F^2 = \sum_n \left[F(f_n^+) - F(f_n^-)\right]^2. \tag{86}$$

Therefore, the PDF uncertainty estimation is a direct application of Equation 86 to the fitted functional forms.

The Hessian method relies on a consistent choice of the tolerance T. In the ideal Gaussian case, the 68-th percentile of the quantity $F(f)$ is given by the value that induces a unit variation in the χ^2, i. e. $T^2 = 1$. However, it is commonly agreed upon that such values lead to an overall underestimation of the PDF uncertainties. Therefore, larger tolerances are usually adopted, such as $T^2 = 100$ instead of $T^2 = 2.7$ for the 90-th percentile. As we shall discuss later in great detail, this tension can be caused by the presence of inconsistencies between different datasets, by MHOUs or by the parametrization choices made by the fitting framework.

In order to avoid underestimations of the PDF uncertainties, it was suggested [35] that the tolerance for each independent direction in $U(\theta_0)$ migth be determined separately, following the condition that all datasets should be described within their 90-th percentile. This concept, called "dynamical tolerance", produces uncertainties that are difficult to examine in a statistical sense, due to the large deviation from the expected value $T^2 = 1$ of the tolerance.

2.4.2 *Monte Carlo estimation*

The NNPDF approach to inverse problems and to the computation of Equation 82 is based on a Monte Carlo estimation. While the formalism laid out in Section 2.3.3 gives a quantitative description of how data space priors propagate the information on the uncertainties into model space, the Monte Carlo method aims at sampling directly from the model posterior distribution. The posterior $p(f|y)$ is described in this context by an ensamble of fit results, obtained from i.i.d. random artificial input data drawn from a multivariate distribution governed by the experimental covariance matrix and the published central values. By means of the Central Limit Theorem (CLT), the mean and standard deviation of the fit results are expected to propagate correctly the data space covariance and central values into PDF space.

During a NNPDF fit, an ensamble of N_{rep} artificial data is generated for each experimental point according to a multigaussian distribution given by the experimental covariance matrix:

$$y_k^{\text{art}} = y^{\text{exp}} + \eta_k, \tag{87}$$

where $\eta_k \in \mathcal{N}(0, C)$ for each replica $k = 1, \ldots, N_{\text{rep}}$. Then, N_{rep} independent fits are performed on the artificial datasets, yielding a Monte Carlo ensamble $\{f_k\}$ of PDFs replicas which provide a faithful description of the posterior model distribution with uncertainties propagated from the input data space. The CLT guarantees that the central value and standard deviation of the model replicas are represented by

$$E[f] = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} f_k, \quad \text{Var}[f] = \frac{1}{N_{\text{rep}} - 1} \sum_{k=1}^{N_{\text{rep}}} (f_k - E[f])^2. \tag{88}$$

This approach follows the observation made in Section 2.3.3 that the NNPDF MAP estimation is performed without assuming a specific model prior, bur rather by regulating the likelihood with additional constraints. Contrary to the Hessian method, it requires no assumption on the nature of the cost function at its minima, nor the choice of a tolerance. As we have discussed, the tuning of tolerances can be far from trivial and the discussion on the correct way to set its value is still an open matter.

The advantages of choosing a Monte Carlo approach to the problem of PDF uncertainties are however subject to two main constraints. Firstly, the computational cost of constructing a sufficiently copious ensemble of fit results may be daunting for a wide class of PDF fitting methodologies. Within the latest NNPDF release, the code has been greatly optimized thanks to the many updates discussed throughout this chapter, and it is now possible to perform Monte Carlo analyses in relatively short times.

It is also desirable that the Monte Carlo approach gives predictions for PDF uncertainties that are comparable to the ones determined by the Hessian method. In this light, a confirmation on the equivalence of the two methodologies has been shown [34] fitting a Monte Carlo ensemble with a Gaussian distribution, which is always assumed in a Hessian fit. The NNPDF4.0 methodology features the possibility to produce Hessian sets from Monte Carlo via the `MC2Hessian` algorithm [36, 37].

Part II

STATE OF THE ART

This part is devoted to the statement of the purpose of the thesis and to the presentation of the state-of-the-art methods adopted to achieve it.

3 STATEMENT OF THE PROBLEM

Among the new data introduced in the latest NNPDF release, most of them come from high precision collider experiments, such as HERA and the LHC. Many of these datasets are mostly affected by systematic – rather than statistical – uncertainties. As a consequence, their χ^2 often becomes highly sensitive to the model assumed for the treatment of published experimental systematic errors.

It is not clear whether inaccuracies in the estimate of systematic uncertainties can be the cause of poor χ^2 observed in the latest PDF determinations. The χ^2 of the latest NNLO PDF sets from the NNPDF and CTEQ collaborations were determined respectively with values of $\chi^2 = 1.16$ [25] and 1.17 (1.19) [19]. However consistent with the previous NNPDF [27] – and consistently decreasing with growing perturbative orders – these values cannot be the result of pure statistical fluctuations. For instance, the NNPDF4.0 χ^2 determined with $N_{\text{data}} = 4618$ experimental points is 5.44-σ away from the expected value of $\chi^2 = 1$.

Since the beginning of the past decade, it was suggested that the tension between theory predictions and experimental data could be solved by rescaling the χ^2 of a suitable factor, at least for methodologies based on the Hessian approach that use fixed PDF functional forms. This led to the introduction of the tolerance method in the context of Hessian determination of uncertainties, as described in Section 2.4.1. The large tolerances adopted can therefore be understood as a manifestation of an incompatibility between the experimental datasets included in a PDF fit. This claim seems to be supported by the fact that fits to specific sets of data, such as the ones provided by HERAPDF, still adopt a tolerance of $T = 1$ for the 68-th percentile estimation in their Hessian framework. Quantitative studies [38] suggest that the effect of inconsistencies among datasets in the Hessian approach is larger than what is predicted by Gaussian statistics, thereby validating the $T > 1$ approach.

It is worth comparing this treatment of uncertainties with the Monte Carlo approach, together with neural network parametrization adopted by the NNPDF collaboration, which provides uncertainty bands by direct computation from the Monte Carlo sample. It is not clear whether the necessity of a tolerance $T > 1$ is a consequence of the Hessian approach – i. e. of the quadratic expansion about the minimum of the χ^2 – or of the fixed functional form parametrized by the methodologies that adopt such error propagation. In this light, studies performed on the NNPDF methodology should clarify if large values of the χ^2 are a consequence of inconsistencies among experimental data.

This thesis aims at answering the questions above by determining whether inconsistencies between experimental datasets can be detected within the NNPDF approach and, if not, how would such inconsistencies corrupt the results for the global uncertainties estimated by the fitting framework. It was recently shown [39] with indirect studies on the NNPDF4.0 global dataset that, regardless of the reason why large values of χ^2 are obtained, the χ^2s found do not imply an underestimation of uncertainties. A direct analysis, which is not yet been carried out, should provide more accurate insights on how the methodology responds to inconsistent data.

This can be easily achieved with the employement of a powerful tool known as the closure test. The method, extensively described in the next chapter, consists in performing standard NNPDF fits to fake experimental data that are generated from theoretical predictions of a known underlying set of PDFs. For this reason, a closure test is by definition free of experimental inconsistencies and should be delivered with the correct $\chi^2 = 1$. It is however possible to simulate their presence through manipulations of the experimental data. In this way, thanks to the introduction of ad-hoc statistical estimators, one can exploit the knowledge of all prior probability distributions and quantitatively determine the methodology's response to inconsistent data in terms of the faithfulness of its uncertainties.

4 CLOSURE TESTS

This chapter is devoted to the presentation of how closure tests are implemented in the NNPDF methodology. Their definition is given in Section 4.1, while Section 4.2 introduces the main statistical estimators that can be computed from a closure test in order to make statements about the fitting methodology. In the context of what anticipated in Chapter 3, we shall discuss how to introduce artificial inconsistencies in a closure test in Section 4.3.

4.1 BASICS OF CLOSURE TESTS

Validation techniques have been adopted by the NNPDF collaboration since the previous releases [40] in the form of closure tests. The basic idea of a closure test is quite easy to understand in the notation of Section 2.3.3: it is designed to study how the methodology fits the inverse map $(\pi \circ \phi)^{-1}$ for a suitable set of datapoints artificially generated with a guessed forward map $\pi \circ \phi$.

A closure test setup consists of two parts: first the information contained in its input PDFs is carried from PDF space to data space and, second, fake data are generated with some noise. The former step requires a precise theoretical framework – e. g. NNLO pQCD, as in the latest NNPDF released fits to real data – and is implemented through computation of theoretical predictions encoded in the FK tables. With these informations, one can generate a set of observables $\sigma = \phi(f)$ from a guessed input PDF $f \in M$, with known but realistic statistical properties. Then, fake exprimental central values $z = \pi(\sigma)$ are generated from the projection of σ onto a finite-dimensional data space, by means of the map π introduced in Equation 69.

In a closure test, the map is constructed following the reasonable assumption that experimental data should be distributed Gaussianly around the value of the observable σ, with some level of uncertainty given by the experimental covariance matrix C. This leads to

$$z = \pi(\sigma) = \sigma + \eta \tag{89}$$

where η is sampled from a multivariate normal distribution $\mathcal{N}(0, C)$.

We believe that Equation 89 represents the best way to artificially reproduce the outcome of an experimental measurement. Indeed, experimental outputs consist of a central value z and a set of uncertainties encoded into the covariance matrix C such that the pair (z, C) represents the measured observable. Signals of new Physics are discovered whenever sets of data points are way distant from the expected values that statistical fluctuations and systematic uncertainties cannot account for them.

In this picture, by generating fake data z directly from the experimental covariance matrix C and the observable σ, the pair (z, C) is by construction a representation of σ without signals of new Physics involved. This is what we mean when we say that fake data have no internal inconsistencies and are also entirely consistent with the theoretical model adopted to produce them.

If one performs a fit to fake data, they should reproduce the assumed underlying PDF within the correct uncertainties and, by manually setting the noise incorporated within the fake data, one can investigate further the impact of statistical inconsistencies on the methodology. Fits are performed within the Monte Carlo replica method adopted for error propagation in the NNPDF methodology. The data $y^{(k)}$ used for the k-th replica in the PDF fit are produced by adding a further layer of fluctuations sampled from the same multivariate normal distribution of Equation 89. In formula, this reads

$$y^{(k)} = \sigma + \eta + \epsilon^{(k)} \qquad \forall k = 1, \dots, N_{\text{reps}} \tag{90}$$

where each Monte Carlo replica is generated by sampling an independent noise vector $\epsilon^{(k)}$ from the same multivariate normal distribution $\mathcal{N}(0, C)$.

For the sake of consistency with previous works on this subject, we shall henceforth refer to the artificially generated data z as level-one data and to y as level-two data. In this context, the true values of the observables, σ, are called level-zero data. Observe that level-two data are nothing more than the data replicas adopted by a standard NNPDF fit, while the feature introduced by a closure test is the level-zero and level-one data.

4.2 STATISTICAL ESTIMATORS

A successful closure test must be such that the resulting PDF fit yields a faithful statistical description of the known underlying law. In order to assess quantitatively the degree of success of a closure test, we define in this section a set of statistical estimators to measure relevant features such as deviations from Gaussianity and under/over learning of the neural network.

The construction of these estimators should be guided by the fact that their main purpose is to compare two quantities – e. g. model predictions, underlying data or level-one data – and determine the presence of biases in accordance to their uncertainties. If we define such quantities as q_1 and q_2, a suitable statistical estimator will therefore be the squared norm of the difference vector $\delta = q_1 - q_2$, i. e. its inner product with itself (δ, δ), in a space where the metric is given by the covariance matrix. Thus, the form of our estimators will schematically be $\delta^T C^{-1} \delta$, up to some arbitrary normalization.

4.2.1 *Exploiting the* χ^2

It is easy to understand that the figure of merit adopted during fitting, and defined in Equation 66, it not a good closure test estimator by itself since it does not exploit the informations about the underlying PDF and the true values of the physical observables. However, two statistical estimators can be computed from it that give informations about overfitting and uncertainties.

The first one, called Δ_{χ^2}, is a rather coarse estimator. It is evaluated from comparing the expectation value of the model predictions and the level-one data z and the χ^2 evaluated between the underlying true values σ and the same level-one data. For this task, we introduce the following notation: we use the symbol $\widehat{f}$

for the fitted PDFs, and therefore the theoretical predictions of the model will be $\widehat{\sigma} = \phi(\widehat{f})$. Then, the $\Delta\chi^2$ reads

$$\Delta_{\chi^2} = \frac{\chi^2(E_\epsilon[\widehat{\sigma}], z) - \chi^2(\sigma, z)}{\chi^2(\sigma, z)}. \tag{91}$$

Since both χ^2s if Equation 91 are computed with respect to the same level-one dataset, this estimator determines whether the fit is more distant from the underlying truth that the level-one data. If all data were scorrelated, one would expect it to be zero in a successful closure test. Deviations from the expected value are however seen even if the replica distribution is sampled perfectly form the posterior distribution [41] as a consequence of the fact that expeirmental data are indeed correlated. Therefore, values of $\Delta_{\chi^2} < 0$ remain acceptable since they indicate that the model predictions slightly overfit the underlying data.

A similar estimator, indicated as φ, can be computed from the level-zero data instead of level-one:

$$\varphi = \sqrt{E_\epsilon\left[\chi^2(\widehat{\sigma}, \sigma)\right] - \chi^2(E_\epsilon[\widehat{\sigma}], \sigma)} \tag{92}$$

Given that level-zero data do not fluctuate, one expects the uncertainty on the predicted value to decrease towards zero in the limit.

4.2.2 *Bias and variance*

We introduce here a pair of statistical estimators that play a key role in the quantification of the successful outcome of closure tests: the bias and the variance. Bias and variance measure two different sources of error expected to contribute to the χ^2 evaluated on test datasets.

The bias is the mean squared error of the theory predictions from the underlying truth. In other words, it is the effective uncertainty predicted by the fit in units of the experimental covariance matrix. In this light, the variance is the nominal error, i. e. the error computed as the mean square distance of the replica predictions from their central value. With this terminology, one expects that the two quantities are comparable every time the closure test has delivered a faithful representation of the true PDFs.

In the context of a closure test, knowledge of prior model distributions capacitate the quantification of the bias and variance of a PDF fit. The PDF space estimation of such indicators is extremely facilitated by the Monte Carlo approach to uncertainties determination adopted by the NNPDF methodology. Indeed, the variance is nothing more than what already presented in Equation 88, while the bias can be easily identified with the following expression:

$$\text{Bias} = \frac{1}{N_{\text{rep}}} \sum_{k}^{N_{\text{rep}}} (f - \widehat{f}^{(k)})^2. \tag{93}$$

The two quantities can be sampled from a number N_x of points in the x-grid and an estimation can be given by mediating over the number of points chosen. The value of N_x must of course be such that it yields fairly stable outcomes, if compared with the fluctuations over the grid. It turns out that, since PDFs are continuous functions, sampling more than a few number of points in PDF space corresponds to using highly correlated data points. As a consequence, the PDF covariance matrix becomes near-singular and numerical issues arise during its inversion.

The reasons expalined above suggest that the bias and variance should be calculated in data space, where the covariance matrix is given by the experimental covariance matrix. The data space bias is defined as the difference between the central value of the model replica predictions $E_\epsilon[\widehat{\sigma}]$ and the observables σ themselves, normalized by the number of data points:

$$\text{Bias} = \frac{1}{N_{\text{data}}} \sum_{ij} (E_\epsilon[\widehat{\sigma}] - \sigma)_i C^{-1}_{ij} (E_\epsilon[\widehat{\sigma}] - \sigma)_j . \tag{94}$$

It is worth noting that the definitions of data space and PDF bias are quite similar. Indeed, the latter can be written in the former's fashion for $C^{-1} = \mathbb{1}$, which is entirely consistent with the assumption made in Section 2.3.3 of trivial prior distributions in model space.

The data space definition of the variance is the expectation value over the replicas of the difference between the expectation value of all replica predictions $E_\epsilon[\widehat{\sigma}]$ and that specific replica prediction $\widehat{\sigma}^{(k)}$. In formula,

$$\text{Variance} = \frac{1}{N_{\text{data}}} E_\epsilon \left[\sum_{ij} (E_\epsilon[\widehat{\sigma}] - \widehat{\sigma}^{(k)})_i C^{-1}_{ij} (E_\epsilon[\widehat{\sigma}] - \widehat{\sigma}^{(k)})_j \right], \tag{95}$$

whence, again, the PDF space variance can be retrieved with $C^{-1} = \mathbb{1}$.

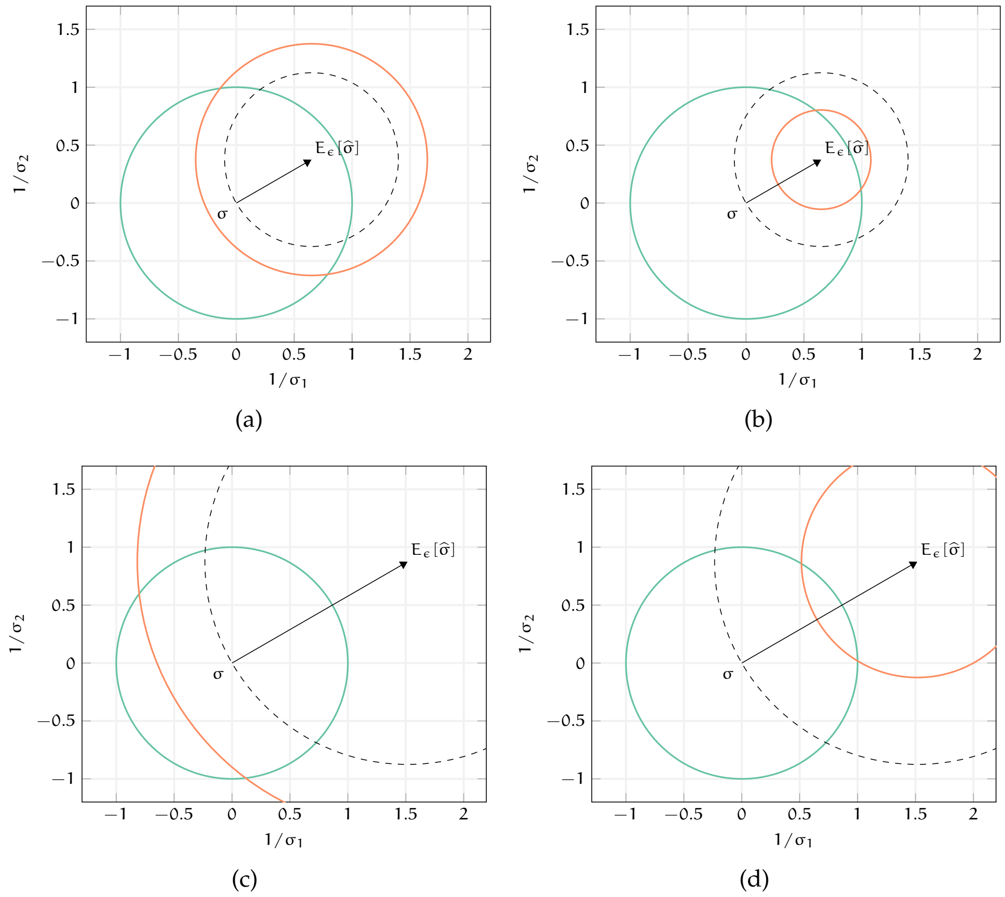
The interplay between bias and variance is fundamental in order to distinguish "good-looking" results from truly good ones in a closure test. As we already mentioned, the bias represents the true error made by the methodology, while the variance is the nominal error stated in a PDF fit. In order to better understand this, we can distinguish between four situations that can happen in the analysis of a closure test output, as in Figure 18.

Generally, fit predictions $E_\epsilon[\widehat{\sigma}]$ are shifted away from the origin by the bias, and have their own statistical properties, i. e. the variance. Figure 18a represents the optimal situation where the bias is within the one-sigma confidence level of fluctuations of the true observables, and it is comparable with the variance of the model's prediction. Of course, we only expect the bias to be inside the blue circle in the 68% of the situations, and therefore this result is particularly good.

On the other hand, even when small biases are found, the closure test might be unsuccesful if those are underestimated by the variance, as in Figure 18b. In this sense, the PDFs still deliver a successful representation of the underlying law, but the nominal error is underestimated.

The outcome of a PDF fit represented by Figure 18c is the opposite situation. The prediction for the PDFs is corrupted, but still acceptable since the distance is less than two-sigmas. The real uncertainty of the fit is however correctly accounted for by its nominal value.

Finally, what is displayed in Figure 18d represents the worst possible outcome of a closure test that does not fit the data correctly and does not even account for this in its uncertainty bands.



(a) (b)

(c) (d)

Figure 18: Geometric interpretation of the statistical estimators in data space. The bias is represented by the arrow (or by the corresponding dashed circle), while the variance is the orange circle. The origin is the true value of the observable, σ, and the blue unit circle around it represents its observational noise.

4.2.3 *Quantile statistics in PDF and data space*

We present another family of estimators that determine if – and in what measure – the fitting methodology gives a faithful representation of PDF uncertainties. This is achieved through a quantitative estimation of how the posterior distributions fitted by the neural network deviate from the Gaussian hypothesis imposed on their priors. As it happens for bias and variance, we prefer sampling the posterior distributions in data space in order to overcome numerical issues that arise from highly correlated datapoints.

Quantiles are n-sigma characteristic functions. We thus define:

$$\xi_{n\sigma} = \frac{1}{N_{\text{data}}} \sum_{i=1}^{N_{\text{data}}} I_{A_i}\big(E_\epsilon[\sigma_i] - \widehat{\sigma}_i\big), \tag{96}$$

where A_i is the n-sigma interval for the theory prediction of the i-th observable, and $I_{A_i}(x)$ is its characteristic function:

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{elsewhere.} \end{cases} \tag{97}$$

4.2.4 *Multiclosure tests*

When PDF fits are performed to real data, the level-one data z of Equation 89 are fixed at the published central value. On the other hand, in the context of a closure test fit the fake central values z are viewed as stochastic variables due to their dependence on the shift vector η. Hence, the statistical estimators presented in Section 4.2 are also stochastic variables and, in order to best characterize the behavior of a methodology, we need to understand their probability distribution rather than computing a single occurrence of them.

This can be achieved by running more than one closure test fit, generating several instances of the random shifts η. We refer to this family of closure tests as multiclosure tests. In principle, a multiclosure test features more than one dataset describing the measurement of the same observable and showing different – but consistent – central values and correlations. A situation like this is impossible to happen in the real world, since experimental outputs only have one central value. Thus, when one performs multiclosure tests, they truly are generating several "runs of the Universe" [25] itself.

Studies of multiclosure tests have only been made possible by the computational speed-up from deployment of best-performing machine learning algorithms featured by the latest release of NNPDF. Such tests are designed for the introduction of new statistical estimators that determine the expectation values of the aforementioned probability distributions of old indicators. This is done by taking expectation values across different instances of the shift η: for example, the bias reads

$$E_\eta[\text{bias}] = \frac{1}{N_{\text{data}}} E_\eta \left[\sum_{ij} (E_\epsilon[\hat{\sigma}] - \sigma)_i C^{-1}_{ij} (E_\epsilon[\hat{\sigma}] - \sigma)_j \right]. \tag{98}$$

The expectation value of the bias across closure fits represents the expected distance between the central predictions and the true values in units of the covariance matrix averaged across all data. Of course, as it has been pointed out in Section 4.2.2, one should measure deviations from the expected bias in terms of the variance of the model's predictions. We thus define the analogue expectation value of the estimator in Equation 95 as

$$E_\eta[\text{variance}] = \frac{1}{N_{\text{data}}} E_\eta \left[E_\epsilon \left[\sum_{ij} (E_\epsilon[\hat{\sigma}] - \hat{\sigma}^{(k)})_i C^{-1}_{ij} (E_\epsilon[\hat{\sigma}] - \hat{\sigma}^{(k)})_j \right] \right]. \tag{99}$$

We can interpret the expectation value of the variance as the uncertainty of the predictions propagated from PDFs when averaged across all data in units of the experimental covariance matrix. If the uncertainty associated to the PDF replicas is faithful, we then expect to find $E_\eta[\text{bias}] = E_\eta[\text{variance}]$ in the limit of large replicas and large number of fits, as pointed out in Section 4.2.2.

In the context of a multiclosure test, we can give a quantitative description of what was only qualitatively described in Figure 18. Since both bias and variance are squared quantities, we can look at

$$\sqrt{\mathcal{R}_{bv}} = \sqrt{\frac{E_\eta[\text{bias}]}{E_\eta[\text{variance}]}} \tag{100}$$

as a measure of how much the uncertainty has been over or under estimated by the methodology. If $\sqrt{\mathcal{R}_{bv}} < 1$, replica predictions are fluctuating more than central

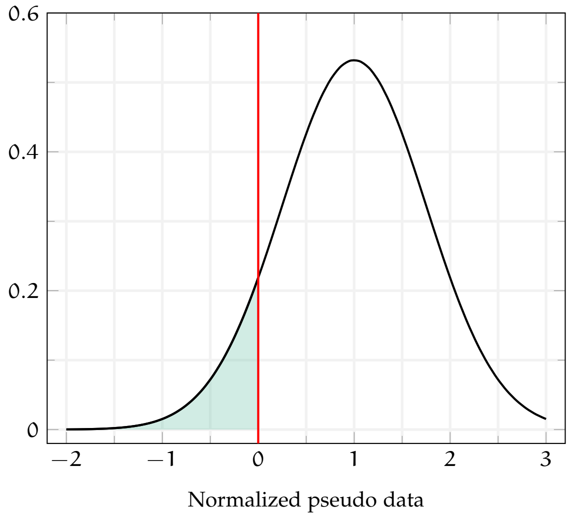


Figure 19: Graphical representation of pseudo data generation normalized by the experimental central value. All data sampled in the blue region must be rejected by the methodology since observables cannot be negative. This introduces sources of non-Gaussianity in the pseudo data generation procedure.

predictions, thus the methodology estimate of the PDF uncertainties is too generous. On the other hand, $\sqrt{\mathcal{R}_{bv}} > 1$ corresponds to a situation where the PDF fitting methodology has delivered underestimated uncertainties. Should one obtain deviations from the value $\sqrt{\mathcal{R}_{bv}} = 1$ in a multiclosure fit, their interpretation of the outcome may vary with the statement that they are trying to make. For instance, the ability of NNPDF to deliver PDF fits at the percent-level has been put in doubt in recent works [42] and multiclosure tests performed on NNPDF4.0 can determine whether the methodology addresses inconsistencies in the correct way.

Before moving onto the discussion of inconsistent data, we give the definition of the quantiles for multiclosure tests. The expectation value across N_{fits} is simply taken by the sampled average, yielding

$$\xi_{n\sigma} = \frac{1}{N_{\text{data}}} \frac{1}{N_{\text{fits}}} \sum_{i=1}^{N_{\text{data}}} \sum_{\ell=1}^{N_{\text{fits}}} I_{A_i^\ell}\left(E_\epsilon[\sigma_\ell^i] - \hat{\sigma}^i\right). \tag{101}$$

The expected quantile estimators can also be computed within the Gaussianity assumption from the bias-to-variance ratio $\mathcal{R}_{bv}$. This is possible in the basis which diagonalizes the experimental covariance matrix, where the sum over datapoints is actually a sum over the eigenvectors of the covariance. In this basis, the expected $\xi_{n\sigma}$ is nothing more than the error function

$$\xi_{n\sigma} \simeq \text{erf}\left(\frac{n\mathcal{R}_{bv}}{\sqrt{2}}\right). \tag{102}$$

Agreement between the computation of $\xi_{n\sigma}$ from Equation 101 and from the error function of Equation 102 indicates that Gaussianity is overall preserved by the fits. Sources of non-Gaussianity are however present in every PDF fit every time data are stochastically generated. This is due to the positivity constraints on experimental central values, that force the introduction of cutoffs in the tails of the distributions as displayed in Figure 19.

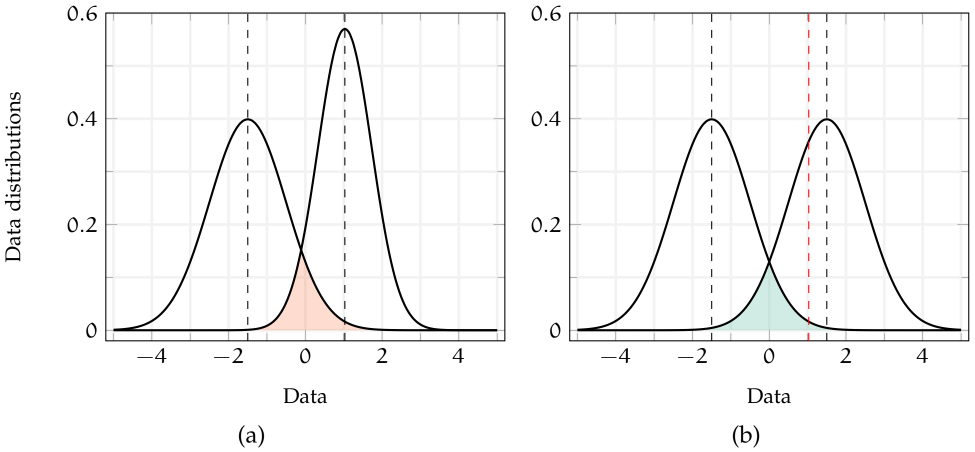


Figure 20: Compatibility of gaussian distributions with different central values and uncertainties. The pink area is equal to the blue area.

4.3 INCONSISTENT DATA

It has been suggested that poor fitting χ^2 results, such as the ones presented in Chapter 3, may be consequence of inconsistencies inside the experimental datasets. It is quite trivial to understand the reasons that support the statement with the following example.

Consider fitting some noisy points in $\mathbb{R}^2$ without assuming a particular functional form. If the cost function only measures the distance of points from the fitted curve, the fitting methodology will try to reach all the points at the same time, and such massive overfitting is only avoided with cross validation and patience algorithms. If however the cost is weighted by the covariance matrix, as it happens for NNPDF's χ^2, the fitted curve will tend to satisfy the constraints imposed by statistically heavier points. In this light, the χ^2 of the fit is clearly corrupted if there is a subset of the training dataset whose central values or uncertainties are inconsistent with the other points.

We aim at determining whether a methodology can simultaneously avoid over fitting and correctly fit inconsistent data. We present in this section a method that can be adopted to answer the question.

4.3.1 *Definition of inconsistency*

By inconsistency, we mean a situation where the nominal uncertainty on a datapoint is smaller than its real uncertainty. Since we always have to do with data that are distributed according to a standard deviation, this is equivalent of saying that inconsistencies arise whenever two datapoints are incompatible within their uncertainties.

The statement above can be understood by looking at Figure 20. The compatibility of two measurements can be estimated by looking at the intersection of the probability distributions of the data. Figure 20a represents a situation where the distributions of two observables with central values indicated by the dashed black lines are compatible in the pink area. On the other hand, the two observables are also compatible in Figure 20b even though their central values have incremented their distance, since the distribution at the right also incremented its width. When we say that shifts to central values are somewhat equivalent to changes in the uncertainties, we mean that for a given shift of the central value there always exists a corresponding shift of the standard deviation such that the pink area is equivalent to the blue one.

Therefore, manipulations of the covariance matrix are sufficient to produce inconsistencies in experimental datasets. These manipulations should be performed such as to reproduce an inconsistency by means of the definition given above, i. e. in a way that the nominal uncertainty is smaller than the real one. This can be done within a closure test in the following way:

1. generate the level-one fake experimental central values $z = \pi(\sigma)$ according to Equation 89 with the correct experimental uncertainties, i. e. $\eta \in \mathcal{N}(0, C)$;
2. produce the Monte Carlo fits ensemble and calculate their figures of merit using a covariance matrix C', which underestimates the uncertainties of C.

In this way, the closure test is performed adopting C' in every step that is common to a standard PDF fit, while the real experimental covariance matrix is only used for the additional generation of fake data.

4.3.2 *Underestimation of systematic errors*

In order to reproduce a realistic experimental situation, the covariance matrix should be manipulated at the level of systematic uncertainties. The main difference between systematic and statistical errors is that, unlike the latter, systematic uncertainties cannot be arbitrarily reduced in magnitude by increasing the number of measurements of a specific observable. Rather, systematic uncertainties are intrinsic features of the measurement system. Since it is usually challenging to precisely determine their magnitude or functional dependence, inconsistent systematic biases are likely to appear in a PDF fit.

For this reason, the principles introduced in Section 4.3.1 should not be applied to the total covariance matrix of Equation 67, but only to the σ_{ik}^{add} and $\sigma_{ik}^{\text{mult}}$ correlated matrices. If we call S_{ij} the matrix obtained considering only the last two addenda of Equation 67, the best way to produce a covariance matrix S' that underestimates the uncertainties of S is to determine S' by changing one eigenvalue of S in the subspace corresponding to the inconsistent dataset.

Since S is symmetric by construction, there exists an orthogonal transformation V such that S can be described in terms of a diagonal matrix D

$$S = VDV^T, \tag{103}$$

and the diagonal elements of D are the positive eigenvalues $\lambda_\alpha \in \mathbb{R}^+$ of S, obtained through the eigenvalue equation $Sv_\alpha = \lambda_\alpha v_\alpha$, with eigenvectors v_α. Consider a set of indices $A \subseteq \{1, \dots, N_{\text{data}}\}$: with the operation

$$\lambda_\alpha \mapsto \lambda'_\alpha < \lambda_\alpha \qquad \forall \alpha \in A \tag{104}$$

the inconsistencies are introduced in the diagonal basis on D, thereby yielding the manipulated diagonal covariance $D'(A)$. Such matrix can then be transformed back into the experimental basis with the same orthogonal matrix V, to give the inconsistent covariance matrix

$$S'(A) = VD'(A)V^T. \tag{105}$$

The substitution of Equation 104 has an important physical meaning. It corresponds to a situation where experimental collaborations have given nominal uncertainties smaller than the actual ones and, in the limit $\lambda'_\alpha = 0$, the uncertainty on the datapoints has been completely forgotten. Even though the latter is a rather unrealistic situation, it will be considered in this thesis in order to provide a solid baseline for further studies on this subject.

4.3.3 *An equivalent way to introduce inconsistencies*

The method introduced in the previous section aims at simulating a situation where uncertainties are underestimated by changing a subset of the covariance matrix eigenvalues.

Suppose that we change only one eigenvalue by setting it to zero. In that case, we can state that an equivalent way to manipulate the covariance matrix is to remove a systematic uncertainty from the determination of S. With the notation of Equation 67 and of Section 4.3.2, we can formalize the statement as follows. The matrix S_{ij} of Section 4.3.2 is constructed by a product $\widehat{S}^T\widehat{S}$, where $\widehat{S}_{ik}$ encodes the value of the k-th systematic uncertainty on the i-th datapoint. For a given systematic uncertainty β, the substitution $\widehat{S}_{i\beta} \to 0$ implies that the matrix S computed from $\widehat{S}$ has a vanishing eigenvalue.

It is indeed quite trivial to show that a matrix with a vanishing column must have an eigenvalue which is zero. For simplicity, consider the matrix $M \in \mathbb{R}^{p\times q}$ with elements of the first column equal to zero: $m_{i1} = 0$ for $i = 1, \dots, p$. The eigenvalue equation is found through the computation of the determinant

$$\det(M - \lambda\mathbb{1}) = \begin{vmatrix} -\lambda & m_{12} & \dots & m_{1q} \\ \vdots & \ddots & \ddots & \vdots \\ -\lambda & m_{p2} & \dots & m_{pq} \end{vmatrix} = 0. \tag{106}$$

By means of the Laplace expansion about the first column of M, the determinant is

$$\begin{aligned} \det(M - \lambda\mathbb{1}) &= \sum_{i=1}^{p} (m_{i1} - \lambda)(-1)^{i+1}\mathrm{cof}(m_{i1} - \lambda) = \\ &= -\lambda \sum_{i=1}^{p} (-1)^{i+1}\mathrm{cof}(m_{i1} - \lambda), \end{aligned} \tag{107}$$

and therefore $\lambda = 0$ is always a solution of Equation 106. The statement is therefore demonstrated by arguing that, if $\widehat{S}$ has a vanishing eigenvalue, there exists a vector v such that $\widehat{S}v = 0$, whence $\widehat{S}^T\widehat{S}v = 0$ and the matrix $S = \widehat{S}^T\widehat{S}$ has a null eigenvalue as well.

The fact that inconsistencies can be equivalently introduced by removing a systematic uncertainty from the construction of the covariance matrix is important for two reasons. First of all, this eases the implementation and the numerical efficiency of the manipulations, since inversion in the dataset subspace is avoided and the inconsistency is introduced during the construction of the covariance matrix. Secondly, the fact that the user can choose a specific systematic uncertainty to remove from the covariance matrix makes the manipulation more transparent and calibrable.

4.4 EXPERIMENTAL AND THEORETICAL INPUTS

Before delivering the results of this thesis, we shall describe the dataset used to produce them. Experimental inputs are needed for the construction of the covariance matrix, used to produce level-one and level-two data, compute the fitting χ^2 and derive the closure test's statistical estimators. In what follows, soon after a brief discussion on how to divide them in order to compute the closure test estimators, we present the datasets adopted by this work.

4.4.1 *The in-sample and out-of-sample division*

It is a standard procedure to compute data space estimators for a closure test with data that have been left out of the fitting sample, in the same manner as fitting methodologies perform the training/validation splitting to avoid over learning.

The in-sample and out-of-sample dataset division for closure test estimators is however not as essential as the latter, since post-fit analyses do not risk to over fit informations. Above all, such division represents a way to understand if the fitting methodology is able to deliver correct predictions out of the training data range. For this reason, normal closure test analyses such as the one delivered in [25] do not consider in-sample statistics. It is assumed that the methodology does not fail to mimic the input training data, and therefore any in-sample statistical estimator is by definition equivalent to the global χ^2 of the fit.

However, this changes when inconsistencies are introduced. To the best of our knowledge, the impact of inconsistent data on PDF fitting methodologies is unknown. Because of the fact that the forward map is corrupted at some point during a closure test fit, it cannot be stated that in-sample statistic is equivalent to the figures of merit adopted during the fit. When fitting artificially generated inconsistent data, not only the assumption of a-priori well-performing in-sample statistic fails, but it is not even clear whether it should in princple outperform the out-of-sample counterpart. In other words, one must not discard the possibility to find worse in-sample indicators than the out-of-sample. Indeed, the latter investigate how results are generalized by the methodology while the former validate it, and there is no overlap between the two since the partition on datapoints is exclusive.

For such reasons, we shall adopt a in-sample and out-of-sample division in the analyses delivered in this thesis, focusing with the same effort on the results given by both.

4.4.2 *Closure test datsets*

The analyses performed in this thesis almost exploit the entire NNPDF4.0 baseline dataset. It consists of 4004 in-sample and 606 out-of-samle experimental points coming from the main international experiments and covering a wide kinematic

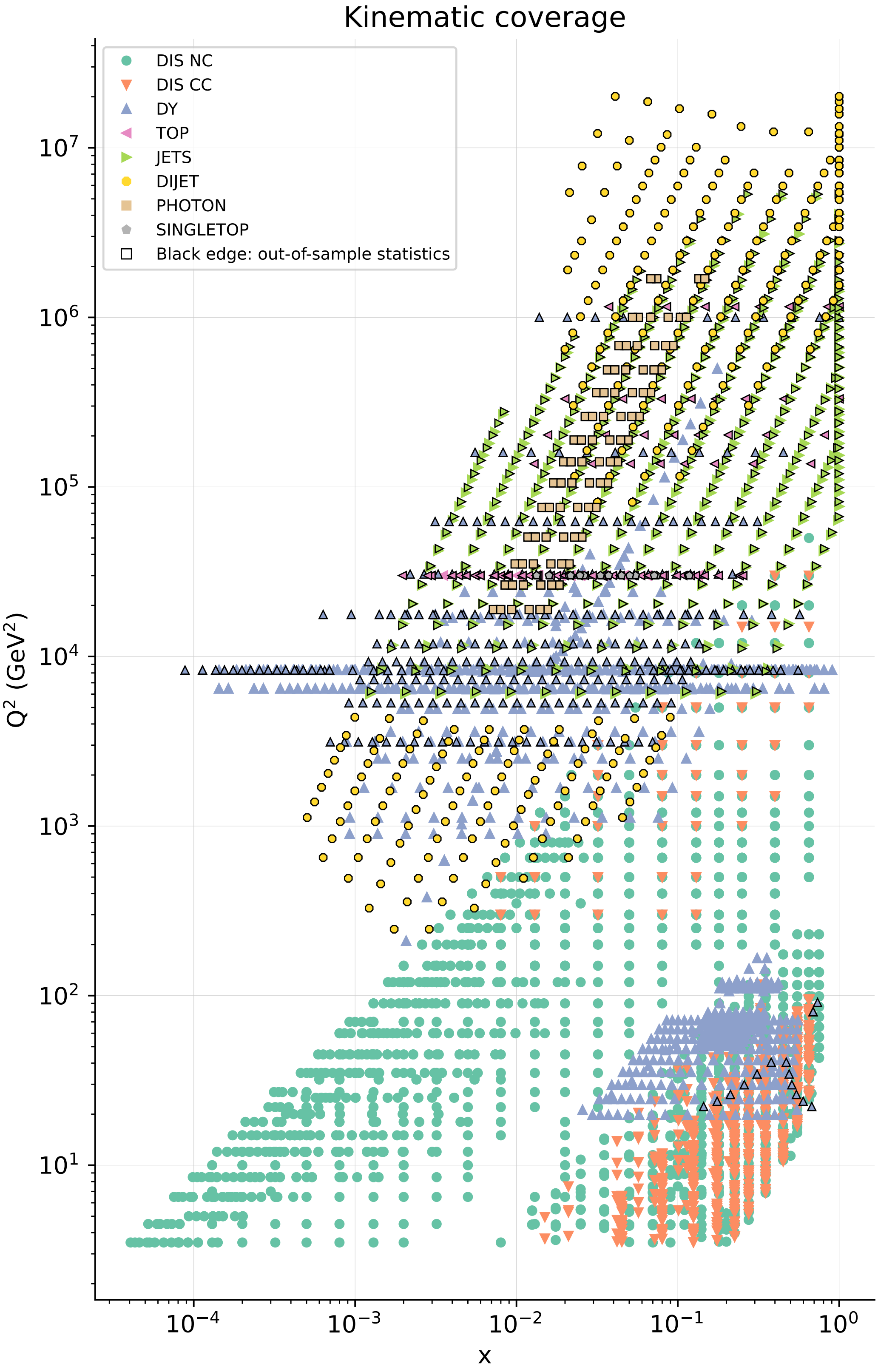


Figure 21: The (x, Q^2) kinematic coverage of the datasets used in this thesis. Points have been distinguished according to the process that produced them.

region both in x (down to 10^{-4}) and Q^2 (up to over 10^7 GeV^2), as displayed in Figure 21. In order to ensure continuity in the results provided, we adopt the dataset division already used in [25], which has an historical interest since it consists in performing the closure tests with data from the NNPDF3.1 determination and use all the new data introduced in NNPDF4.0 for out-of-sample statistics.

The main processes considered are fixed-target and collider DIS, DY and jet production, electroweak boson production and top quark pair production. DIS data provide almost 75% of the experimental points for the in-sample dataset, with 2100 measurements of neutral current processes and 989 of charged current from the NMC, SLAC, BCDMS and HERA combination. In particular, neutral current DIS from a proton target is used to probe the quark sea distribution $f_i + \bar{f}_i$ and flavour separation. Additional information on the valence quark distributions comes from neutrino scattering on nuclear targets, provided by the experiments CHORUS and NuTeV.

The remaining fraction of in-sample datapoints comes, for instance, from fixed-target DY measurements from the Fermilab experiments, which constrain the u/d combination, or from the earliest measurements of $p\bar{p}$ collisions recorded by the Tevatron colliders, which provide information on the quark flavour separation at large-x. The same collaborations also provide single-inclusive jet production cross sections, which are of great importance for the determination of the gluon PDF alongside data for the top pair production cross sections recorded at the LHC. For additional informations on the in-sample data we refer to [27], while the out-of-sample dataset is described in [25].

4.4.3 *Choice of the underlying PDFs*

We discuss the underlying PDFs that have been used in this thesis.

The choice of an input PDF is the foundation of a closure test and, in order for the test to be nontrivial, it is necessary to select it as sufficiently complex as possible so as to stress the methodology. We choose to perform the validation by assuming that the input PDF is a single specific replica selected out of the 1000 replicas of the NNPDF4.0 NNLO global determination. We refer to such PDF set as the underlying set, while the selected replica is the underlying PDF.

One could argue that the choice would compromise the consistency of the closure test, since the attempt at validating the methodology must be independent from its output. It is easy to find several reasons why it is not the case. First of all, since the closure test is performed within the NNPDF methodology, sampling from its replica ensemble guarantees that the underlying PDF satisfies known theoretical constraints and that the test performance is the highest possible thanks to the optimized hyperparameters. Secondly – and definitely most importantly – a closure test is not to determine whether the methodology returns the real physical PDFs: it rather endorses the absence of biases and inconsistencies in it by comparing the central values and uncertainties of its predictions to the underlying inputs.

The foremost criterion that can be used to determine the underlying PDF is what better guarantees model-independency and consists in picking a random replica from the underlying set. Nevertheless, there are ways of better stressing the closure tests within the adopted set by introducing some bias in the choice of the input PDF. This produces closure tests outputs that can be compared to the ones coming from the unbiased underlying law in an effort to determine whether the methodology can offset fluctuations in level-one data. Such alternative input PDFs can be determined either by looking at the most fluctuating replica of the underlying set, or by sampling the farthest from the central value.

Analyses on the performance of closure tests with different underlying PDFs have been carried out in this thesis and are presented in Section 5.1.2

Part III

FUGUE

We deliver the results of this thesis and derive from them a number of conclusions regarding the problems stated in the previous parts.

5 RESULTS

In this chapter, we present the results of closure test analyses performed on the NNPDF4.0 fitting methodology with inconsistent data. The aim of the analyses is to determine the impact of inconsistencies in the fitting framework by performing the manipulations described in Chapter 4 on four different datasets that are included in the training closure test sample, as described in Section 4.4.2. Results are presented in four different sections, starting from Section 5.2.

An introduction is delivered in Section 5.1 and aims at providing the results of preliminary analyses made on the closure test framework in order to determine the number of fits and replicas to be used in the fits, the underlying PDFs and the systematic uncertainties to be removed in inconsistent datasets. Conclusions are left to Section 5.6.

5.1 INTRODUCTION

We describe the methods and results used to tune the choice of inconsistent datasets, number of fits and replicas used.

5.1.1 *Inconsistent datasets*

The processes studied are single inclusive jet production, neutral current DIS and DY electroweak boson production. Both jet and electroweak boson productions are measured at the LHC by the ATLAS and the LHCb detectors, while the DIS measurements come from the combination of the HERA and H1 experiments. These datasets belong to different kinematic regions: as one can see from Figure 22, DIS data cover the small-Q region with a sufficient number of measurements both in the small-x and large-x limits, while DY and jet rapidity measurements provide kinematic coverage of $Q^2 > 10^4$ GeV. Different kinematic regions are linked by the DGLAP evolution, and we expect that the inconsistencies will propagate and have effects outside of their region. This means that, in principle, results can be dependent on the constraints given by a specific dataset on its kinematic region.

In particular, we can study different situations. Firstly, we can introduce inconsistencies in a kinematic region covered by a single group of data. Considering Figure 21, we see that the large-x and large-Q^2 region is mainly covered by jet data and therefore this situation corresponds to the closure test performed with inconsistencies in the single inclusive jet production. Another region which is constrained by a single group of data is the small-x region at $Q^2 = 10^4$ GeV, which is covered by LHCb data. However, the region is linked to points at smaller Q^2 and larger x by the DGLAP evolution and we should not discard the possibility that the behavior of LHCb data will differ from the one obtained from jet data.

On the other hand, inconsistencies can be introduced in a kinematic region covered by two sets of data, corresponding to different processes. This is the case of DIS and DY data from the ATLAS detector. We see from Figure 21 and from Fig-

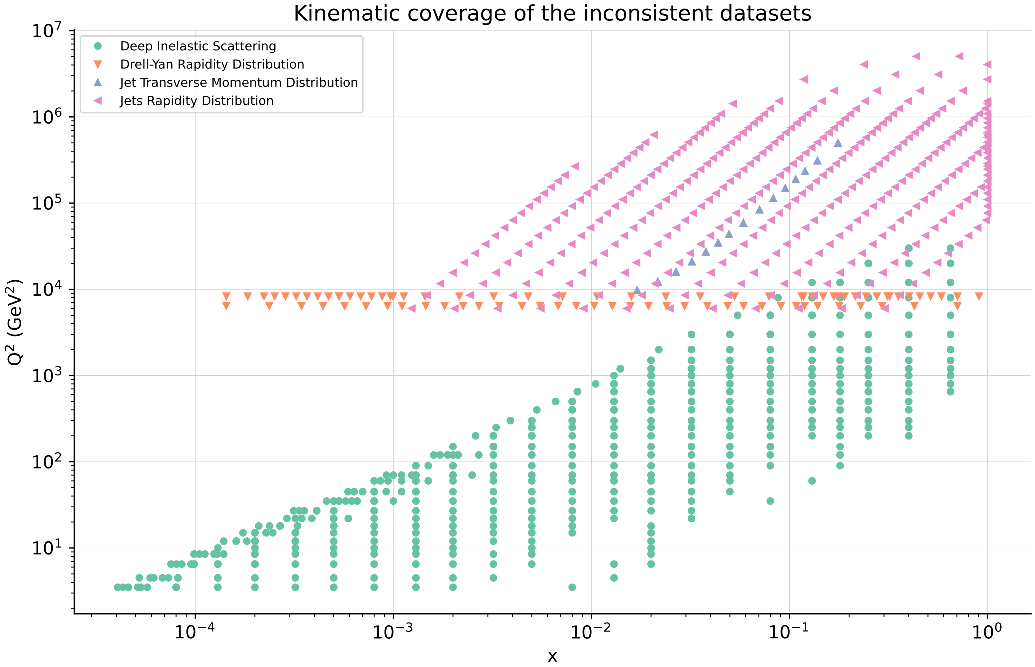


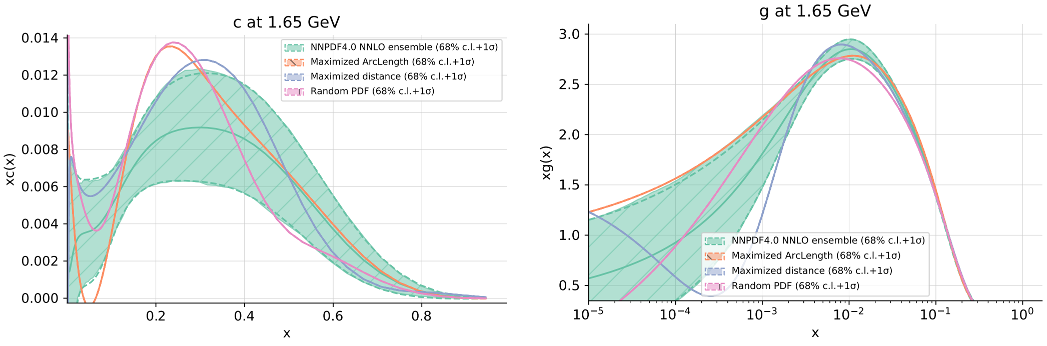
Figure 22: Kinematic coverage of the inconsistent data used in the analyses presented in this chapter.

ure 22 that the two groups of data are linked to each other by DGLAP evolution. It can be possible that, in such cases, the presence of inconsistent data has a different impact on the fitting methodology, if compared to the previous situation. Indeed, two datasets that constrain the same features of the PDFs are incompatible whenever one of them is inconsistent: we expect that the PDF fit will "trust" either the inconsistent data or the consistent ones. In this light, we cannot discard the possibility that the fit does not recognize the inconsistency, thereby following the trend of the heavier set of data.

In Table 3, we show the systematic uncertainties removed for each closure test. The last column shows the impact on the trace of the covariance matrix – which

DATASETS	POINTS	SYSTEMATIC	IMPACT
ATLAS jets	201	Jet flav. comp.	$\sim 10^{-5}\%$
HERA DIS NC	447	δ_{rel}	$\sim 10^{-11}\%$
		$\delta_{\gamma p}$	$\sim 10^{-11}\%$
		δ_{had}	$\sim 10^{-12}\%$
ATLAS DY	46	Luminosity	$\sim 36\%$
LHCb DY	30	Luminosity	$\sim 0.12\%$
		Beam	$\sim 0.08\%$

Table 3: The systematic uncertainties chosen to perform the inconsistent closure tests, along with the impact of their absence on the covariance matrix trace.



(a) (b)

Figure 23: The central value and 68% confidence band for charm (23a) and gluon (23b) PDF at the input parametrization scale Q = 1.65 GeV from the NNPDF4.0 NNLO PDF set.

is the sum of its eigenvalues – obtained removing such systematic uncertianties. For each group of closure tests, the systematic uncertainties chosen were the ones which had greater impact in the inconsistent data.

In all this cases, we do expect to see that the inconsistent closure test estimators – or, at least, one of them – differ from the ones computed for a standard NNPDF4.0 closure test performed on the same input data and underlying PDF. For this reason, we shall provide our results as comparisons with the values obtained for such closure test.

5.1.2 *Effects of different underlying PDFs*

As anticipated in Section 4.4.3, there is no guarantee that the choice of the underlying PDF for a closure test fit will not have an impact on the methodology's response. We investigate here the outcomes of closure tests performed using three different underlying PDFs from the latest NNPDF4.0 determination. As already discussed in the previous chapter, there is no point in using particularly inaccurate shapes for the underlying PDF and therefore we do not explore the possibility of using a non-NNPDF input.

We perform here some preliminary analyses aimed at determining whether the choice of a particular underlying PDF from the NNPDF4.0 set has an impact on the outcomes of a closure test. The NNPDF4.0 NNLO ensemble features 1000 Monte Carlo replicas from the latest NNPDF determination: its central replica represents the methodology's best guess for the true PDFs. We determine the underlying PDF for three different closure tests by sampling from the aforementioned ensemble:

1. a random replica;
2. the most distant replica from the central value of the set;
3. the replica which fluctuates the most, i. e. whose arc-length is maximal in the set.

REPLICA	ARC LENGTH	L1 DISTANCE	KL DIVERGENCE
Maximized arc length	8.507	56.31	−21.35
Maximized distance	8.505	109.1	119.3
Random PDF	8.486	51.46	3.731

Table 4: Values of arc length, distance and KL divergence for the three replicas chosen as underlying PDFs for the preliminary studies on the NNPDF4.0 closure tests.

While the first criterion is what better guarantees the flexibility of a closure test, the second an third ones aim at stressing the methodology in order to see if non-negligible changes in the closure test outputs are produced.

The most distant replica can be found in two ways that turn out to be equivalent for the cases studied, i. e. through maximization of the L1 distance or the Kullback-Leibler (KL) divergence between replicas and central values. Both L1 and KL are distances in PDF space, defined as follows. If $c^{(\alpha)}(x)$ is the central value for the α-th flavour and $q_k^{(\alpha)}(x)$ is the k-th replica for the same flavour α, the L1 distance is

$$d_k^{(\alpha)} = \int_0^1 \left| c^{(\alpha)}(x) - q_k^{(\alpha)}(x) \right| dx\,, \tag{108}$$

while the KL distance is

$$d_k^{(\alpha)} = \int_0^1 q_k^{(\alpha)}(x) \log \frac{c^{(\alpha)}(x)}{q_k^{(\alpha)}(x)}\, dx\,. \tag{109}$$

We maximize the set of distances over the replica index k, and find a finite set containing the indices of the most distant replicas for each flavour α:

$$\left\{ \arg\max_k d_k^{(\alpha)} \quad \forall \alpha \in \text{flavs} \right\}. \tag{110}$$

The index with the highest number of occurences is then chosen as input PDF for the closure tests. If there is more than one replica with such features, the choice is random within the subset. As anticipated, the replica which maximizes the L1 distance turned out to be the one with the maximum KL divergence as well.

The third input replica is chosen according to the arc-length of the PDFs, i. e. through maximization of the following collection of integrals:

$$\ell_k^{(\alpha)} = \int_0^1 \sqrt{1 + \left[\frac{dq_k^{(\alpha)}}{dx} \right]^2}\, dx \tag{111}$$

over the k-th Monte Carlo replica and the α-th flavour of the NNPDF4.0 NNLO set, as done in Equation 110. The three underlying PDFs produced are shown in Figure 23 for the charm and gluon, while the values of distance and arc length are listed in Table 4. From a posterior analysis, we see that the arc length is a rather coarse indicator since all replicas show similar values for $\ell_k^{(\alpha)}$. For this reason, we discard the arc length criterion in this preliminary tuning analyses.

The results of such analyses are delivered here in Table 5 for the statistical estimators computed from the χ^2, and in the following studies on finite size effects, Section 5.1.3, for the $\sqrt{\mathcal{R}_{bv}}$ ratio estimator. Both outcomes emphasyze the fact that choosing a random PDF is equivalent to selecting the most distant of the NNPDF4.0 set.

GROUP	DATA	$E_\eta[\chi^2]$		$E_\eta[\Delta_{\chi^2}]$		$E_\eta[\varphi]$	
		RAND	DIST	RAND	DIST	RAND	DIST
DIS NC	2100	0.986	0.988	−0.008	−0.004	0.123	0.145
DIS CC	989	0.956	0.967	−0.008	−0.009	0.123	0.131
DY	731	0.839	0.849	−0.019	−0.015	0.199	0.225
Top	13	0.976	0.989	−0.011	−0.045	0.303	0.274
Jets	171	1.011	1.077	−0.026	0.005	0.151	0.127
Total	4004	0.953	0.966	−0.011	−0.007	/	/

Table 5: First moments of the χ^2, Δ_{χ^2} and φ distributions for the closure tests performed with random (rand) and maximally distant (dist) underlying PDFs.

5.1.3 *Finite size effects*

The estimators presented in Chapter 4 are expected to deliver a faithful description of the outcome of a closure test in a sufficiently large limit of fits and replicas adopted. For instance, in this thesis we performed 25 fits of 50 replicas for each family of closure test considered. When calculations are carried out with a finite number of samples, one expects their results to be different from the true value of such indicators, and that this difference will decrease when additional samples are considered. We refer to this situation as a finite size effect.

Naturally, finite size effects cannot be measured directly since the true value of a statistical estimator is not known a priori. However, one can perform indirect measures of these effects by looking at the trend of the computed indicators with increasing number of fits and replicas used to produced them. Such indirect search methods have been used in this thesis as a preliminary study on the performance of NNPDF4.0 closure tests, aimed at determining the minimum amount of fits and replicas for which finite size effects can be discarded from the conclusions of this work. The results of the study are given here.

Figure 24 shows the trend of the $\sqrt{\mathcal{R}_{b\nu}}$ indicator for increasing number of fits, computed within a standard NNPDF4.0 closure test with a random underlying PDF from the latest NNPDF determination. The upper diagram shows the values of the ratio alongside their uncertainty. The green curve is a simple 5-points moving average, which represents for each $N_{\text{fits}} \geqslant 5$ the average of the indicator computed using $N_{\text{fits}} - 4, \ldots, N_{\text{fits}}$ fits. The moving average gives sufficient informations about the trend of $\sqrt{\mathcal{R}_{b\nu}}$ and, therefore, convergence can be measured by taking derivatives of the green curve. This is done in the second panel of Figure 24, where the first and second derivative of the moving average are shown as a function of N_{fits}. We can see that, as N_{fits} inreases towards $N_{\text{fits}} = 25$, both curves tend to zero. Additionally, one can observe that the two derivatives have opposite signs almost everywhere, thereby indicating that convergence of the moving average is taking place as a dampened oscillation.

The bottom panel of Figure 24 displays a second method used to determine whether convergence is reached for $\sqrt{\mathcal{R}_{b\nu}}$ with $N_{\text{fits}} \leqslant 25$. The 10-points and 15-points variance are displayed. For each point in the N_{fits} axis, the curves in-

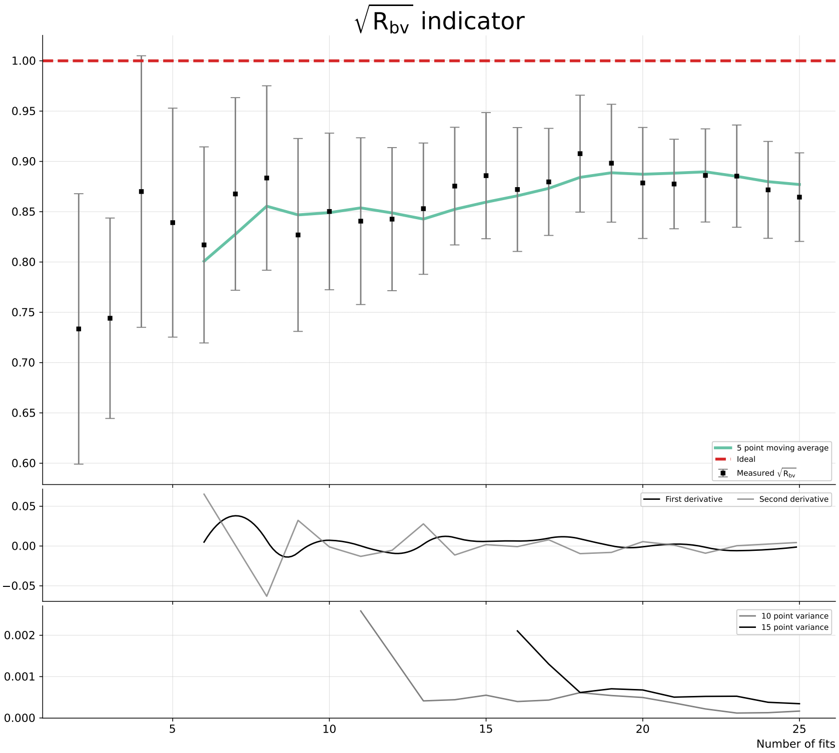


Figure 24: Trend of $\sqrt{\mathcal{R}_{bv}}$ for a standard closure test, performed with increasing number of fits. The 5-points moving average is displayed in green, and its first and second derivative are shown in the second panel. The third panel shows the 10-points and 15-points variance of the data.

dicate the variance of the values of $\sqrt{\mathcal{R}_{bv}}$ computed using $N_{\text{fits}} - k + 1, \ldots, N_{\text{fits}}$ fits, where $k = 10$ and $k = 15$ respectively. We can see that the two variance curves quickly decrease towards zero in the first steps and then maintain the value until $N_{\text{fits}} = 25$ is reached.

The analysis carried out leads to the following conclusions. Given a fixed number of 50 replicas for each fit, a closure test performed with 25 fits trained on different stochastic instances of the level-one data is sufficient to determine the distribution of the statistical estimators. The decision to study the case of fixed number of replicas and variable number of fits comes from the following consideration. The replica number determines how precise will be the uncertainties delivered by the fitting methodology within the Monte Carlo approach, while the number of fits determines how many samples are drawned from the distributions of the statistical estimators. Clearly, a closure test performed using a small number of fits yields inaccurate predictions for the distributions of statistical estimators. At a fixed value of the computational power used by closure tests, preference must be given to the number of fits since we are particularly interested in seeing how such distributions change when inconsistencies are present. By doing so, we do not discard the possibility that the number of replicas is small enough to yield some small fluctuations in the uncertainties predicted by the methodology. Even if such situation was present in the analyses carried out, it would not bias the results obtained comparing the standard closure test outputs with the inconsistent ones as long as both are perforemd with the same number of replicas.

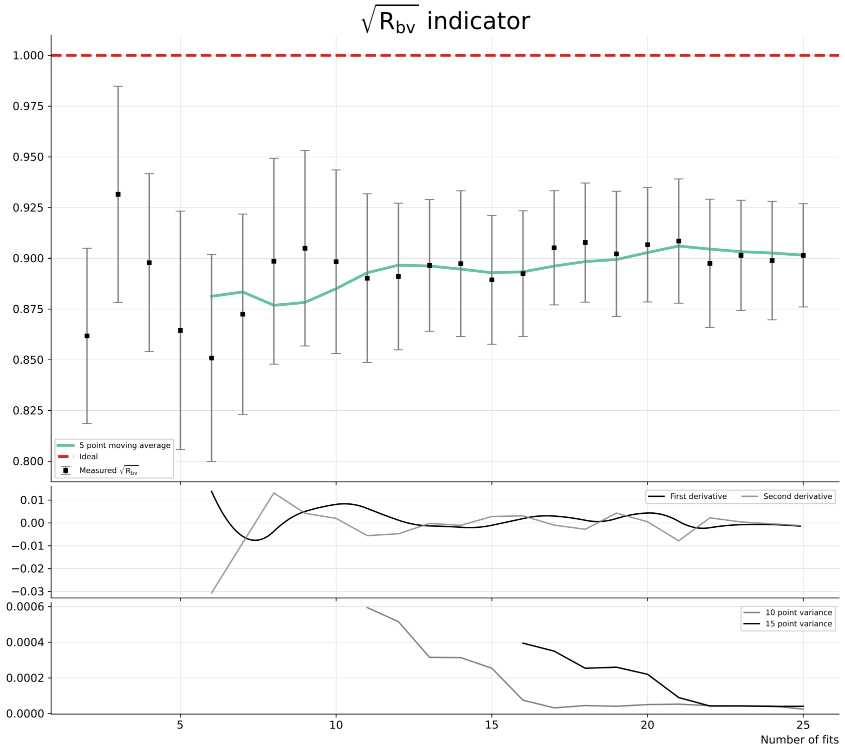


Figure 25: Trend of $\sqrt{\mathcal{R}_{bv}}$ for a standard closure test, performed with increasing number of fits, and with underlying PDF chosen according to the distance criterion.

In conclusion, we show in Figure 25 the same convergence plot of Figure 24 for a closure test performed with the underlying PDF which is most distant from the central value of the NNPDF4.0 NNLO determination. As we can see, results are consistent with what found for a random input PDF. Therefore, the following analyses will be performed with the random PDF.

GROUP	$E_\eta[\chi^2]$	Δ FROM REFERENCE CT		
		$\Delta E_\eta[\chi^2]$	$\Delta E_\eta[\Delta_{\chi^2}]$	$\Delta E_\eta[\varphi]$
DIS NC	0.991	0.594σ	< 0.001	< 0.001
DIS CC	0.979	1.812σ	-0.002	-0.002
DY	0.854	0.976σ	-0.003	-0.001
Top	1.091	1.037σ	0.052	-0.003
Jets	0.980	-1.004σ	0.018	-0.007
Total	0.963	1.585σ	< 0.001	/

Table 6: In-sample values of χ^2, and differences of χ^2, Δ_{χ^2} and φ between an inconsistent closure test obtained removing jet systematic uncertainty and the reference closure test.

5.2 SINGLE INCLUSIVE JETS

We now present the results of this thesis, starting from the analyses performed on the closure tests obtained with inconsistent systematic uncertainties correlated through the following datasets:

1. exclusive ATLAS measurements for $W^\pm$ production associated with $N_{\text{jets}} > 1$ jets of light quarks at $\sqrt{s} = 8$ TeV [43], for a total of $N_{\text{data}} = 30$ datapoints;
2. single-inclusive ATLAS jet production at $\sqrt{s} = 8$ TeV [44], with $N_{\text{data}} = 171$.

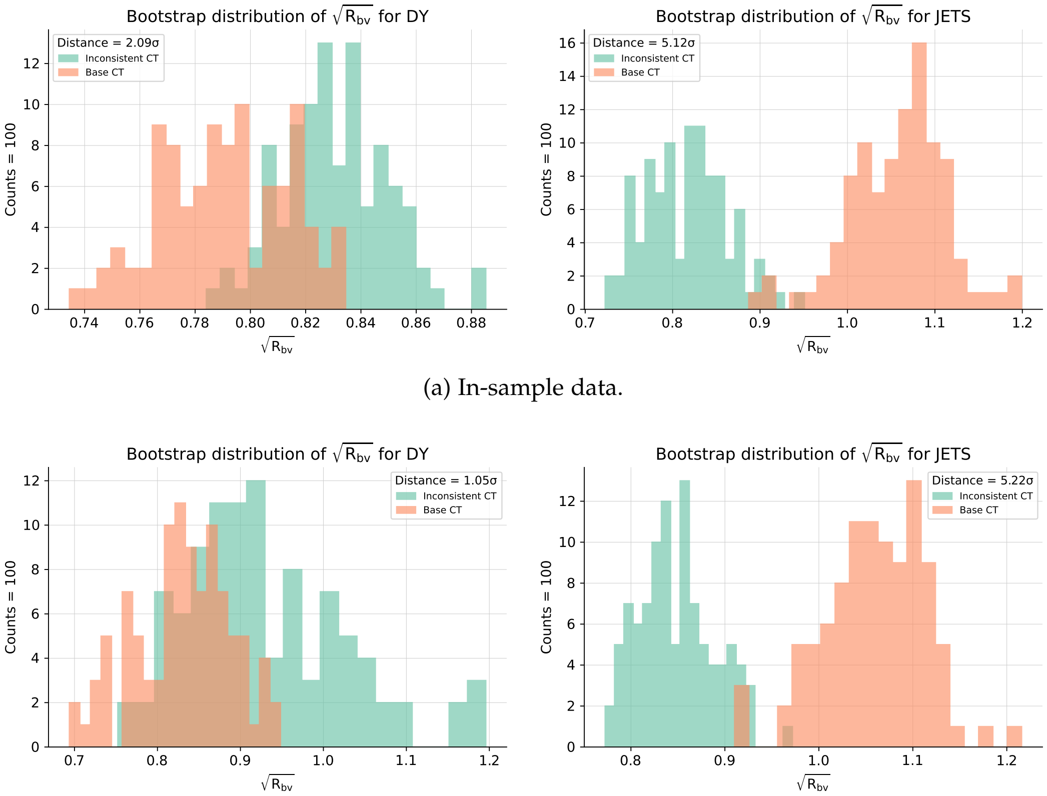
The measurements are recorded by the ATLAS experiment at the LHC at CERN for jets defined by the anti-k_t jet clustering algorithm, using data corresponding to an integrated luminosity of 20.2 fb^{-1}.

The information on experimental uncertainties is retrieved from the corresponding HEPDATA entry. We treat the jet flavour composition uncertainty as fully correlated between the $W^\pm +$ jet data and the inclusive jet data, even if the jet radius parameter used in the clustering algorithm by the former ($R = 0.4$) differs from the latter ($R = 0.6$). This is supported by the fact that mild differences in PDF fits arise when the uncertainties are treated as uncorrelated [45].

5.2.1 χ^2 *estimators*

The first statistical estimators that can measure the presence of inconsistencies in jet data are the ones derived from the χ^2. Table 6 displays the difference between the values of $E_\eta[\chi^2]$, $E_\eta[\Delta_{\chi^2}]$ and $E_\eta[\varphi]$ calculated from a reference closure test and the closure tests with inconsistent jet data. We can see that changes in the χ^2 for the inconsistent closure test are at the level of statistical fluctuations. Here, the χ^2 standard deviation is computed considering that the total degrees of freedom is equal to the number of datapoints times the number of fits performed:

$$\sigma_{\chi^2} = \frac{2}{\sqrt{N_{\text{fits}} N_{\text{data}}}}. \tag{112}$$



(a) In-sample data.

(b) Out-of-sample data.

Figure 26: The $\sqrt{\mathcal{R}_{b\nu}}$ bootstrap distributions for jets and DY data, compared to the reference closure test.

The only process that shows a relevant improvement in the χ^2 is the inconsistent one, i. e. the jets. If comparable changes were also seen for other process, this would suggest that the methodology has followed the trend of the new uncertainties in jet data, penalizing every other dataset. The impact on non-jet processes is however small and cannot lead to such conclusion. Additionally, the impact on Δ_{χ^2} and φ of the inconsistencies is negligible and values obtained are almost equal to the ones of the reference closure test. The only exception is represented by top data, which however consist of only 13 datapoints in sample and, therefore, their impact is negligible.

5.2.2 *Bias and variance*

We turn now to the statistical indicator used to determine whether the closure test delivered faithful uncertainties. As already mentioned in Chapter 4, the main interest in performing more than one closure test fit is in determining the distribution of statistical indicators, rather than sampling a single occurrence. For this reason, distributions of $\sqrt{\mathcal{R}_{b\nu}}$ are generated through resampling techniques – using bootstrap resampling [46] – starting from the 25 values obtained for each fit performed. Such distributions are shown for each main process considered in the in-sample and out-of-sample data in Appendix A, Figure 37 and Figure 38 respectively. The plots display the $\sqrt{\mathcal{R}_{b\nu}}$ distributions for the consistent and inconsistent closure

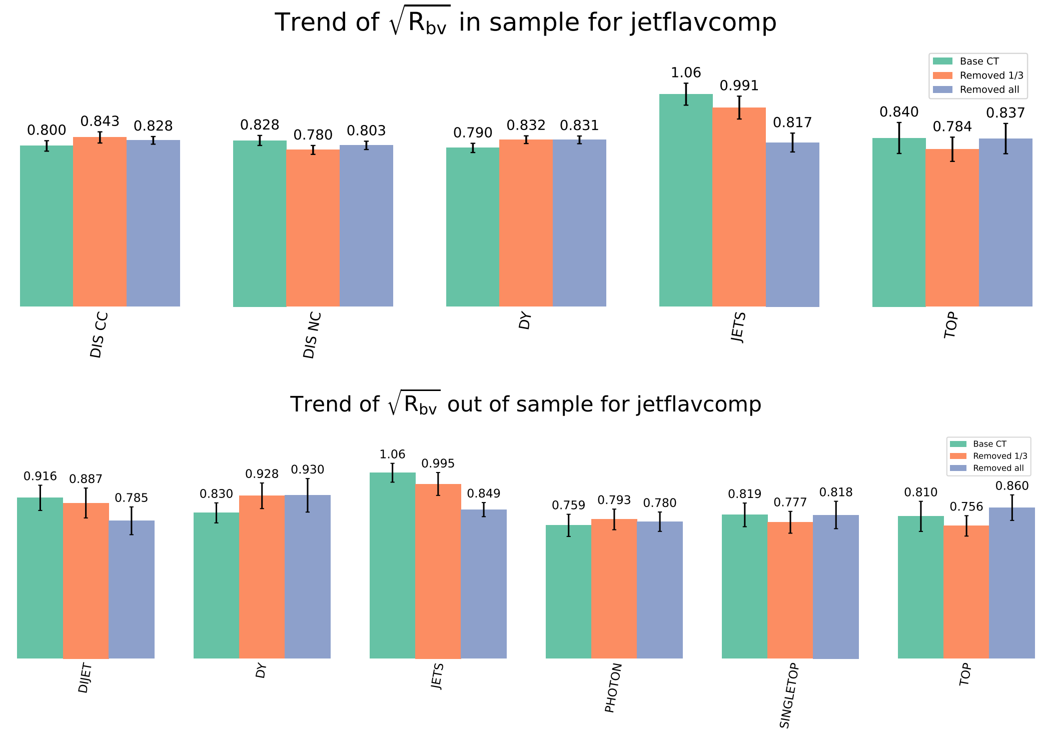


Figure 27: Trend of $\sqrt{\mathcal{R}_{bv}}$ for different levels of inconsistency incorporated within the jets systematic uncertainty.

tests and the distance between the two is expressed in units of the variance of the inconsistent distribution.

In order to facilitate the discussion, we report in Figure 26 the DY and jet data distributions for the in-sample and out-of-sample division. The plots suggest that the bias and variance ellipsoids in data space have indeed modified their shapes as a consequence of the methodology's reaction to inconsistent data. In particular, there is a strong indication that the uncertainties for the jet data have been overestimated with respect to the consistent closure test, both in-sample and out-of-sample. This is implied by the fact that the two distributions are separated by a distance over 5σ. An additional 2σ tension is seen for the in-sample DY distributions, suggesting a slight underestimation of the uncertainties in that data region, which is reduced out of sample.

In order to determine whether the displacements of the distributions for inconsistent data do indicate an overfitting of the uncertainties or they are simple fluctuations, we perform a third closure test retaining the 33% of the systematic. The trend is shown in Figure 27 and it confirms the observations made above about jet data in and out of sample, with smaller effects propagated into the out-of-sample dijet region. The behavior of $\mathcal{R}_{bv}$ on in-sample DY data suggests that the effect on such process is negligible, since we do not find a visible trend in the values of the estimator.

GROUP	$\xi_{1\sigma}$		$\Delta\xi_{1\sigma}$	
	INCONSISTENT	REFERENCE	INCONSISTENT	REFERENCE
DY	0.712	0.715	-0.169σ	-1.788σ
Top	0.759	0.772	0.745σ	0.776σ
Dijet	0.778	0.726	-0.434σ	-0.012σ
Photon	0.760	0.801	-1.260σ	-0.321σ
Singletop	0.784	0.789	0.102σ	0.285σ
Jets	0.771	0.716	0.451σ	2.444σ
Total	0.755	0.731	0.560σ	-0.107σ

Table 7: Out-of-sample measured $\xi_{1\sigma}$ and deviation from the expected value in terms of the bootstrap error for inconsistencies in jet data and reference fit.

5.2.3 *Sources of non-gaussianity*

We now turn to non-Gaussianity, which can be determined by looking at the values of $\xi_{1\sigma}$ and, in particular, at the difference $\Delta\xi_{1\sigma}$ between the expected value of the indicator – see Equation 102 – and the computed one. In the second cand third column of Table 7 one finds measures of $\xi_{1\sigma}$ for the inconsistent closure test and the reference fits which suggest that the uncertainties have been globally overestimated in both closure tests and in the same amount. The distance $\Delta\xi_{1\sigma}$ between the measured $\xi_{1\sigma}$ and its expected value computed from $\sqrt{\mathcal{R}_{b\nu}}$ is given in the last two columns in units of the bootstrap error. We see that differences are at the level of statistical fluctuations and, as expected, the global amount of non-Gaussianity is slightly increased in the inconsistent fit: however, the effects can be explained as a consequence of statistical noise and we do not make assumptions based on such values of $\Delta\xi_{1\sigma}$.

5.2.4 *Concluding remarks*

The analyses performed suggest that the impact of inconsistent jet data on the NNPDF4.0 methodology is measurable in the context of a closure test, but not from standard fit quality estimators such as the χ^2.

The negligible deterioration of the χ^2 suggests that this indicatore is incapable of distinguishing whether a PDF fit was performed on inconsistent data. The values of χ^2 also suggest that the replica central values have been correctly fitted by the methodology, even in presence of inconsistent training data in the jet region. What did change are the uncertainties delivered, as was seen from the bias-to-variance ratio distributions. This suggests that the PDF replica distribution yield a similar central value, but its spread grows in a inconsistent fit and therefore uncertainties are overestimated in the jet data region. We are led to conclude that the fit detected the inconsistency and, in order to deliver a result which is consistent with the χ^2 computed, overfitted the nominal uncertainty of the inconsistent data.

GROUP	$E_\eta[\chi^2]$	Δ FROM REFERENCE CT		
		$\Delta E_\eta[\chi^2]$	$\Delta E_\eta[\Delta_{\chi^2}]$	$\Delta E_\eta[\varphi]$
DIS NC	0.989	0.389σ	< 0.001	-0.001
DIS CC	0.963	0.521σ	< 0.001	-0.001
DY	0.843	0.270σ	-0.002	< 0.001
Top	0.966	-0.091σ	0.045	-0.010
Jets	1.047	1.183σ	0.018	-0.014
Total	0.959	0.858σ	< 0.001	/

Table 8: In-sample values of χ^2, and differences of χ^2, Δ_{χ^2} and φ between a closure test obtained removing the δ_{rel} systematic uncertainty and the reference closure test.

5.3 NEUTRAL CURRENT HERA COMBINED

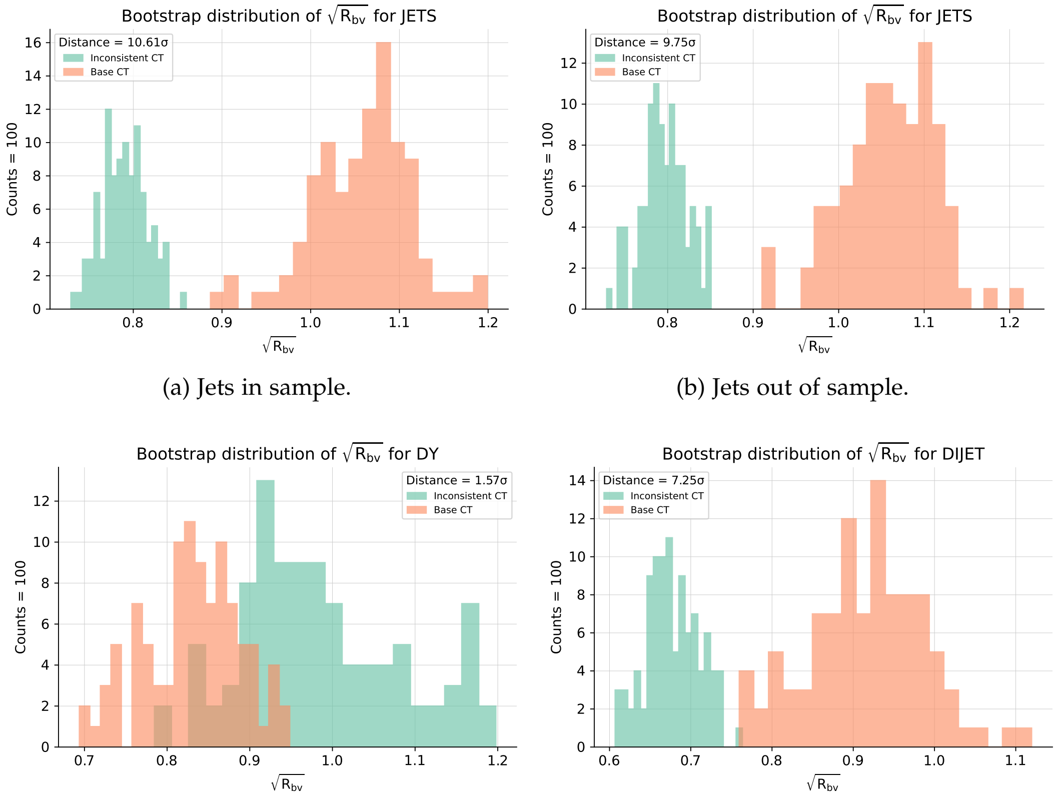
The section is devoted to the investigation of the impact of inconsistencies in the collider neutral current cross-section data from the HERA measurement combination, which was already included in the NNPDF3.1 PDF determination. Specifically, we introduce artificial inconsistencies in the following datasets [21]:

1. inclusive DIS $e^{\pm}p$ scattering at $\sqrt{s} = 820$ GeV, with $N_{\text{data}} = 70$;
2. inclusive DIS $e^{\pm}p$ scattering at $\sqrt{s} = 920$ GeV featuring a total of $N_{\text{data}} = 377$ datapoints.

We remove the correlated sources of systematic uncertainties arising from the combination procedure. In particular, we perform two closure tests by removing the δ_{rel} systematic uncertainty in the first one, and additionally removing the $\delta_{\gamma p}$ and δ_{had} errors in the second one. The former uncertainty is introduced as the difference between the χ^2 computed by treating all uncertainties as multiplicative, and the χ^2 obtained with all additive uncertainties except from the normalization uncertainty. On the other hand, the latter are defined as the differences between the nominal combination and the combinations in which systematic uncertainties associated with the photoproduction background and hadronic energy scale were taken as correlated across dataset.

As one can see from Table 8, the global χ^2 shows a slight improvement with respect to the consistent closure test, mainly due to the diminished jet χ^2. Changes are however at the level of statistical fluctuations and suggest that there is no correlation between them and the presence of inconsistent data, as indicated by the differences in Δ_{χ^2} and φ.

The in-sample and out-of-sample $\sqrt{\mathcal{R}_{b\nu}}$ distributions for each process considered are shown in Appendix A for the reference closure test and the one obtained removing all three systematic uncertainties. Here, we display the most interesting distributions in Figure 28 and the trend of the indicator with increasing inconsistencies in Figure 29. We can appreciate how the distributions for jet data are separated, thereby indicating that the effect of inconsistent data is clearly not a statistical fluctuation. This happens both in sample and out of sample, with values that are reduced in the inconsistent closure test.



(a) Jets in sample.

(b) Jets out of sample.

(c) Out-of-sample dijet and DY.

Figure 28: The $\sqrt{\mathcal{R}_{bv}}$ bootstrap distributions for jets, dijets and DY data, compared to the reference closure test.

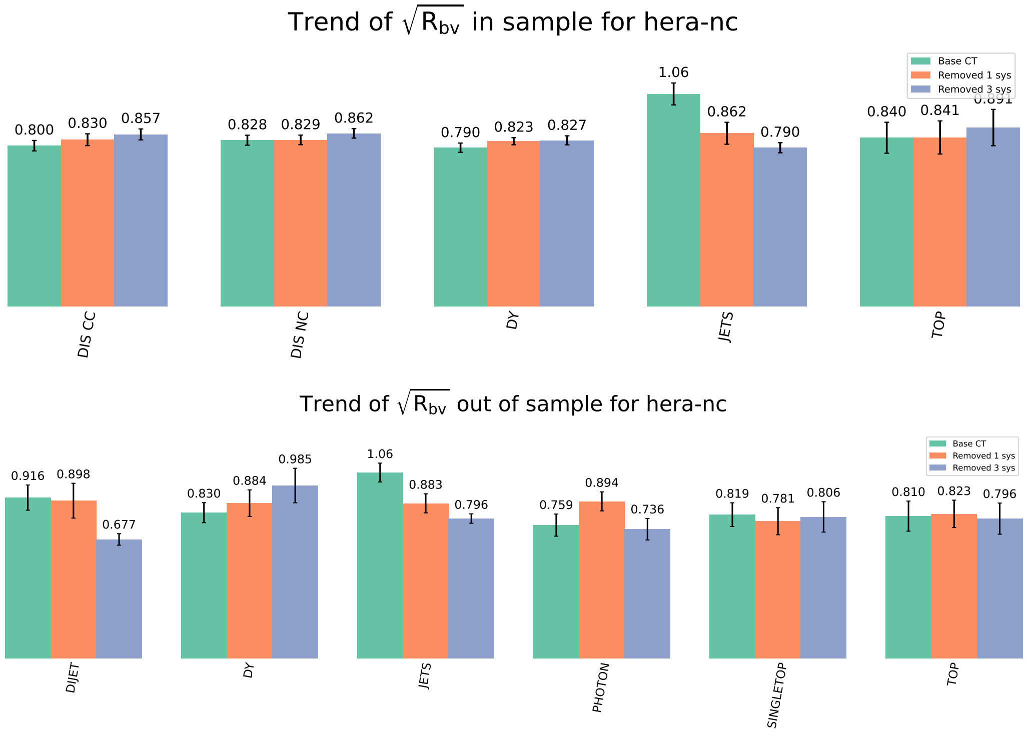


Figure 29: Trend of $\sqrt{\mathcal{R}_{bv}}$ for different levels of inconsistency incorporated within HERA systematic uncertainties.

GROUP	$\xi_{1\sigma}$		$\Delta\xi_{1\sigma}$	
	INCONSISTENT	REFERENCE	INCONSISTENT	REFERENCE
DY	0.727	0.715	-0.386σ	-1.788σ
Top	0.764	0.772	0.207σ	0.776σ
Dijet	0.742	0.726	0.108σ	-0.012σ
Photon	0.691	0.801	-1.576σ	-0.321σ
Singletop	0.767	0.789	-0.756σ	0.285σ
Jets	0.745	0.716	0.073σ	2.444σ
Total	0.736	0.731	-0.244σ	-0.107σ

Table 9: Out-of-sample measured $\xi_{1\sigma}$ and deviation from the expected value in terms of the bootstrap error. Results are provided for the inconsistent fit without the δ_{rel} uncertainty alongside the reference closure test.

Again, the out-of-sample data region sees a decrease of $\sqrt{\mathcal{R}_{b\nu}}$ for dijet data, which are correlated with jets. The opposite happens for DY data, even though we clearly see from Figure 28c that the two distributions have still a non-negligible region of overlap.

The trend of $\sqrt{\mathcal{R}_{b\nu}}$ as the inconsistency increases enforces the hypotesis that DY and jet data have an opposite behavior. It is interesting seeing that there is no trend of $\sqrt{\mathcal{R}_{b\nu}}$ for neutral current DIS. This suggests that the methodology response is different from the previous situation.

In conclusion, we can deduce the impact of inconsistent data on sources of non-Gaussianity in Table 9. Results for $\Delta\xi_{1\sigma}$ are again similar for the inconsistent fits compared to the refernce closure test, thereby confirming the negligible impact of inconsistent data on the Gaussian assumptions. Values of $\xi_{1\sigma}$ are also similar for the two closure tests, both indicating a small overestimation of the uncertainties.

5.3.1 *Concluding remarks*

The analyses performed on the inconsistent closure tests suggest that the fit quality cannot measure the impact of inconsistent data on the methodology, as happened for jet data in Section 5.2.

Among the closure test estimators, we can draw a large amount of information from the bias-to-variance ratio. Contrary to what happened for jet data, the PDF fits delivered the correct uncertainties for DIS data, i. e. where the inconsistencies were introduced. We note that uncertainties have been overestimated for jet data and in the correlated dijet subset, while the opposite happened for DY data.

We conclude that, given the weight of DIS datapoints in the training dataset, the methodology learned the inconsistencies in the data and delivered biased predictions for the PDFs. In particular, we observe from Figure 30 that the small-x kinematic region is covered mostly by DIS data from HERA and LHCb DY data. Since the methodology was trained on inconsistent data, it delivered biased small-x dependences of PDFs and, as a consequence, the $\sqrt{\mathcal{R}_{b\nu}}$ indicator for out-of-sample DY

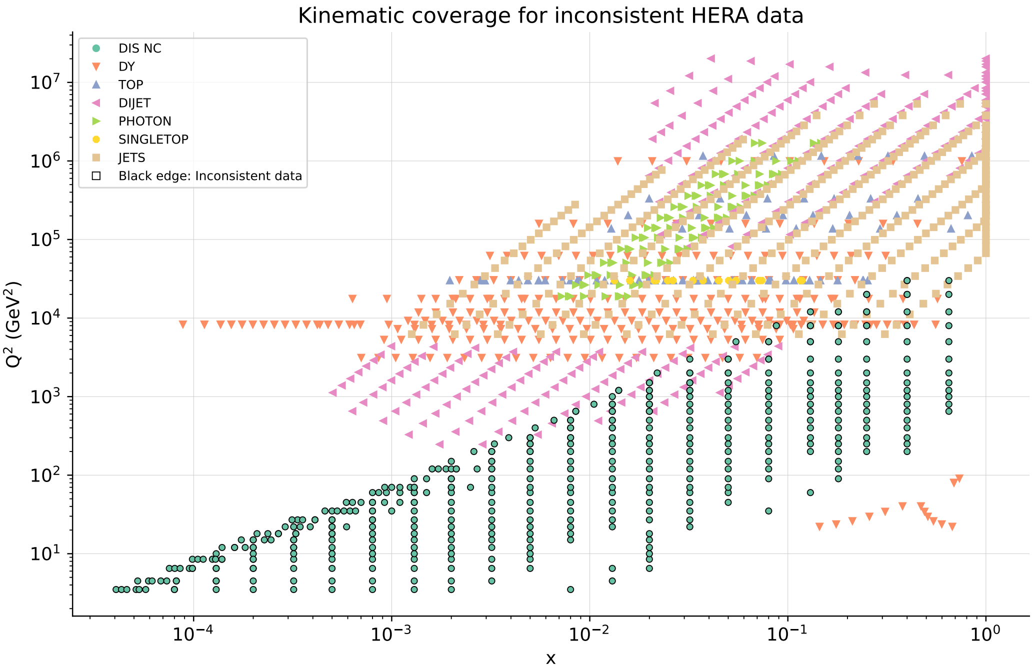


Figure 30: Kinematic coverage of the out-of-sample data compared to the in-sample inconsistent HERA datasets.

data has increased. This is most likely due to underestimation of LHCb uncertainties in that region.

GROUP	$E_\eta[\chi^2]$	Δ FROM REFERENCE CT		
		$\Delta E_\eta[\chi^2]$	$\Delta E_\eta[\Delta_{\chi^2}]$	$\Delta E_\eta[\varphi]$
DIS NC	0.981	-0.569σ	0.001	0.002
DIS CC	0.971	1.198σ	-0.001	0.001
DY	0.854	0.978σ	-0.003	0.005
Top	1.128	1.368σ	0.003	-0.005
Jets	0.997	-0.441σ	0.017	-0.006
Total	0.957	0.600σ	0.001	/

Table 10: In-sample values of χ^2, and differences of χ^2, Δ_{χ^2} and φ between a closure test obtained removing the ATLAS luminosity systematic uncertainty and the reference closure test.

5.4 ATLAS ELECTROWEAK BOSON PRODUCTION

We investigate the impact of inconsistencies in measurements of the electron and muon decay channels of inclusive DY gauge boson production. We consider the combination of measurements of production cross sections for the inclusive DY processes $W^\pm \to \ell\nu$ and $Z/\gamma^* \to \ell\ell$, with $\ell = e, \mu$, performed in proton-proton collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. The analysis uses data taken in the year 2010 [47] with proton beam energies of 3.5 TeV. For the electron channels the luminosity is 36.2 pb^{-1}, while a smaller value is used for the muon channel, 32.6 pb^{-1}.

Since the electron and muon $W^\pm$ and Z cross sections are combined to form a single joint measurement, systematic uncertainties have been correlated between the various datasets by the experimental collaboration. In the notation of Equation 67, it means that the sum over the systematic errors k of the σ_{ik} matrices has already been performed and we cannot distinguish between different sources of systematic uncertainty. The only source of systematic uncertainty which has not been correlated is the one on the luminosity of the ATLAS detector. Therefore, we perform a closure test removing such systematic and a second one also rescaling the correlated uncertainties by a factor 2.

The χ^2, Δ_{χ^2} and φ estimators are shown in Table 10. We can see how the changes in the χ^2 are at the level of statistical fluctuations. The same conclusions can be drawned for the Δ_χ^2 and φ estimators, suggesting once again that inconsistencies were not detected by the fit quality.

The $\sqrt{\mathcal{R}_{b\nu}}$ distributions are shown in Appendix A, Figure 41 and Figure 42, where one can find the complete set of bootstrap distributions of the indicator for in-sample and out-of-sample data respectively. We display here the trend of $\sqrt{\mathcal{R}_{b\nu}}$ for increasing level of inconsistency in the data. As we can see from Figure 31, the impact of inconsistencies is higher than what was seen in the previous cases. Indeed, contrary to the previous situations, this time almost every correlated systematic uncertainty was artificially modified. In the in-sample and out-of-sample data region, Figure 31 does not show clearly the impact on DY: for this reason, we

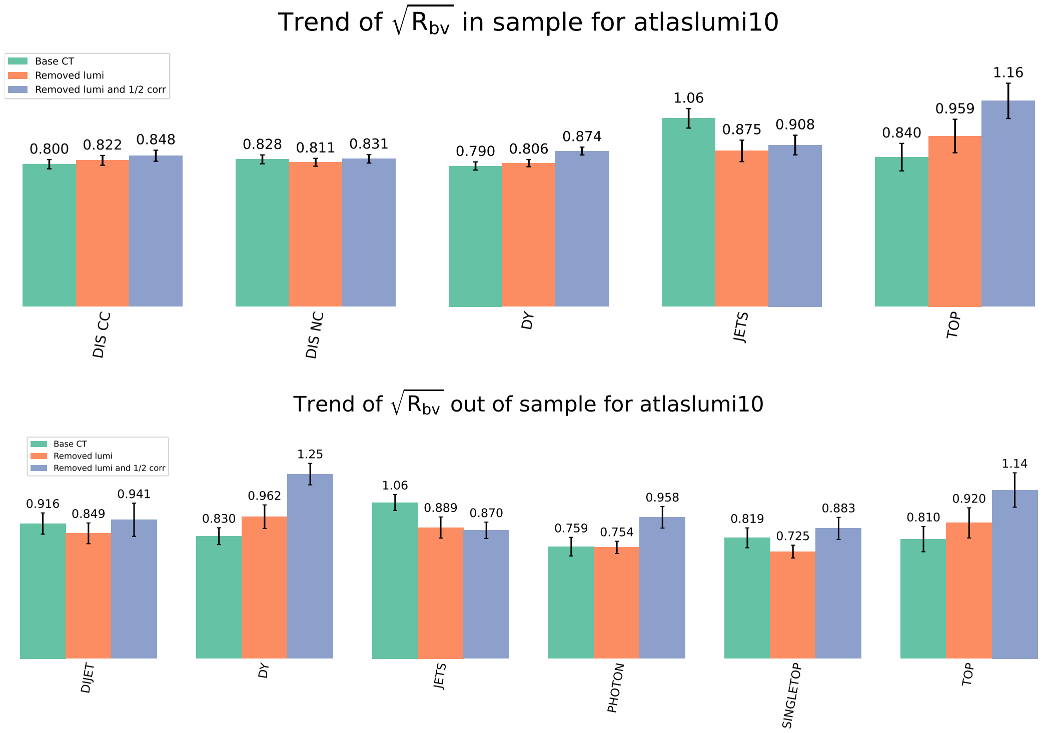
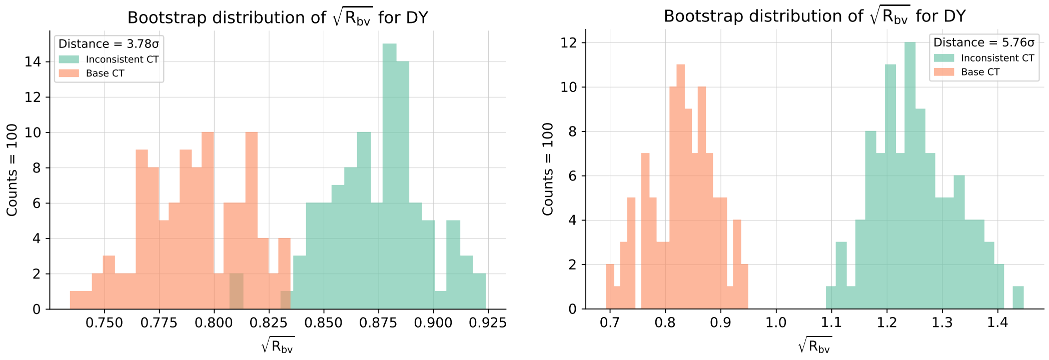


Figure 31: Trend of $\sqrt{\mathcal{R}_{bv}}$ for different levels of inconsistency incorporated within ATLAS luminosity systematic uncertainty.

report here the complete distributions in Figure 32 in order to highlight the fact that uncertainties have been overestimated for the inconsistent data region.



(a) DY in sample.

(b) DY out of sample.

Figure 32: The $\sqrt{\mathcal{R}_{bv}}$ bootstrap distributions for DY data, compared to the reference closure test.

In conclusion, we can see from Table 11 that the measured values of $\xi_{1\sigma}$ are similar for the inconsistent closure test and the reference one, as happened in the previous cases. On the other hand, we can observe that the $\Delta\xi_{1\sigma}$ indicator has increased in the photon data region for the inconsistent fits, thereby indicating that sources of non-Gaussianity might have arised in the region constrained by such data. As a consequence, the overall $\Delta\xi_{1\sigma}$ increased but, since it remains in

GROUP	$\xi_{1\sigma}$		$\Delta\xi_{1\sigma}$	
	INCONSISTENT	REFERENCE	INCONSISTENT	REFERENCE
DY	0.682	0.715	-0.535σ	-1.788σ
Top	0.731	0.772	0.515σ	0.776σ
Dijet	0.746	0.726	-0.411σ	-0.012σ
Photon	0.752	0.801	-2.624σ	-0.321σ
Singletop	0.804	0.789	-1.098σ	0.285σ
Jets	0.773	0.716	0.854σ	2.444σ
Total	0.734	0.731	0.611σ	-0.107σ

Table 11: Out-of-sample measured $\xi_{1\sigma}$ and deviation from the expected value in terms of the bootstrap error. Results are provided for the inconsistent fit without the ATLAS luminosity uncertainty alongside the reference closure test.

the 1σ confidence band, we cannot conclude that inconsistent data had an impact on non-Gaussinaity in this closure test.

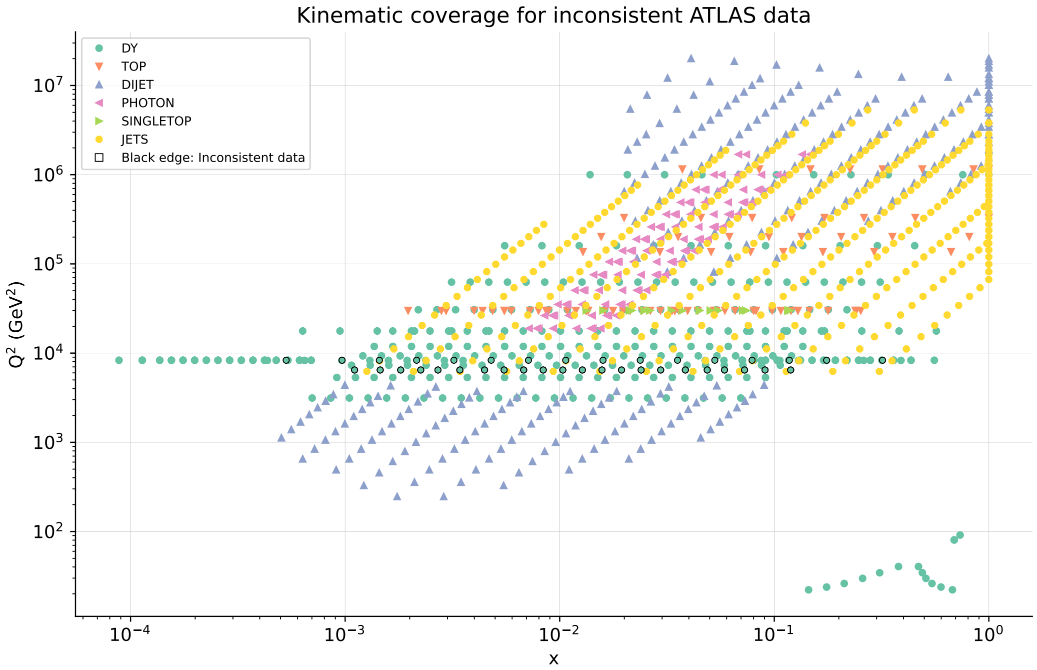


Figure 33: Kinematic coverage of the out-of-sample data compared to the in-sample inconsistent ATLAS datasets.

5.4.1 *Concluding remarks*

We make some final remarks on the results of the closure tests performed on inconsistent training data from the ATLAS DY weak boson production. As happened for jet data and for neutral current DIS in the previous sections, the fit quality cannot determine whether inconsistencies had an impact on the PDF fit.

Informations from the bootstrap distributions of $\sqrt{\mathcal{R}_{b\nu}}$ suggest that uncertainties have been wrongly delivered for DY processes – which include the inconsistent data – in and out of sample, as pictured in Figure 32. On the other hand, top data seem to have been delivered with underestimated uncertainties both in sample and out of sample.

Note that, even though the in-sample DY blue and pink distributions of Figure 32a are separated from each other, their width is particularly small if compared to the expected value of $\sqrt{\mathcal{R}_{b\nu}} = 1$. Therefore, such distributions are not as different has in the out-of-sample region: this leads to the conclusion that the methodology made wrong deductions from the training dataset, i. e. that it predicted similar – although slightly different – uncertainties for the inconsistent data. From Figure 33, we see that the inconsistent x-region is shared by DY with top and jet data mainly. As expected, the effects of the inconsistencies introduced can be seen in the fact that the uncertainty on such datasets has been wrongly predicted, i. e. underestimated for the former and overestimated for the latter.

GROUP	$E_\eta[\chi^2]$	Δ FROM REFERENCE CT		
		$\Delta E_\eta[\chi^2]$	$\Delta E_\eta[\Delta_{\chi^2}]$	$\Delta E_\eta[\varphi]$
DIS NC	0.999	1.533σ	< 0.001	< 0.001
DIS CC	0.959	0.225σ	-0.003	< 0.001
DY	0.851	0.808σ	0.003	0.003
Top	1.044	-0.036σ	0.025	-0.008
Jets	1.007	-0.139σ	0.010	-0.008
Total	0.963	1.565σ	< 0.001	/

Table 12: In-sample values of χ^2, and differences of χ^2, Δ_{χ^2} and φ between a closure test obtained removing the LHCb luminosity systematic uncertainty and the reference closure test.

5.5 LHCB ELECTROWEAK BOSON PRODUCTION

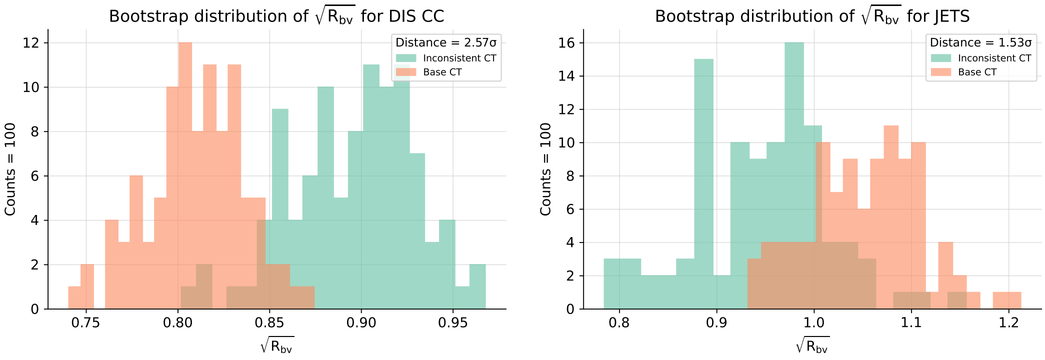


Figure 34: The in-sample $\sqrt{\mathcal{R}_{b\nu}}$ bootstrap distributions for jets and charged current DIS data, compared to the reference closure test.

We investigate the impact of inconsistencies in LHCb [48] measurements of electroweak boson production at $\sqrt{s} = 8$ TeV. Contrary to what has been presented in Section 5.4 for the ATLAS experiment, LHCb data only come from the muon channel at a luminosity of 2.0 fb^{-1}. As stated in the paper, sources of uncertainty that come from external input, such as the beam energy and luminosity determinations, are delivered as separate from other contributions. For this reason, we focus on that kind of systematic errors to perform the manipulation of the experimental covariance matrix, first removing the luminosity and then also the beam energy.

As expected, we see in Table 12 that changes in the χ^2-based estimators are at the level of statistical fluctuations, which is consistent with the observations made in the previous sections.

The $\sqrt{\mathcal{R}_{b\nu}}$ distributions are shown in Appendix A, Figure 43 and Figure 44, as bootstrap distributions of the indicator for in-sample and out-of-sample data respectively. We display here the trend of $\sqrt{\mathcal{R}_{b\nu}}$ for increasing level of inconsistency in the data and some interesting distributions for the in-sample dataset. Figure 34

GROUP	$\xi_{1\sigma}$		$\Delta\xi_{1\sigma}$	
	INCONSISTENT	REFERENCE	INCONSISTENT	REFERENCE
DY	0.704	0.715	-1.245σ	-1.788σ
Top	0.812	0.772	-1.337σ	0.776σ
Dijet	0.777	0.726	-1.431σ	-0.012σ
Photon	0.714	0.801	-0.618σ	-0.321σ
Singletop	0.876	0.789	-0.127σ	0.285σ
Jets	0.768	0.716	1.608σ	2.444σ
Total	0.756	0.731	0.045σ	-0.107σ

Table 13: Out-of-sample measured $\xi_{1\sigma}$ and deviation from the expected value in terms of the bootstrap error. Results are provided for the inconsistent fit without the LHCb luminosity uncertainty alongside the reference closure test.

shows the bootstrap distributions of jet data and charged current DIS in sample. We can see a strong increase of $\sqrt{\mathcal{R}_{b\nu}}$ in the DIS region, while the value of the estimator has dropped for jet data less than what was seen in the previous cases studied. From Figure 35, we see that the trend of the first moments of the $\sqrt{\mathcal{R}_{b\nu}}$ distributions for in-sample DIS and jet data suggests that their uncertainties have been underfitted and overfitted respectively, while fluctuations are observed for the value of $\sqrt{\mathcal{R}_{b\nu}}$ for other processes. This is enforced in the out-of-sample data region, where jet data seem to be delivered with overestimated uncertainties.

In conclusion, we understand from Table 13 that sources of non-Gaussianity are present in the same amount for the reference and the inconsistent closure tests.

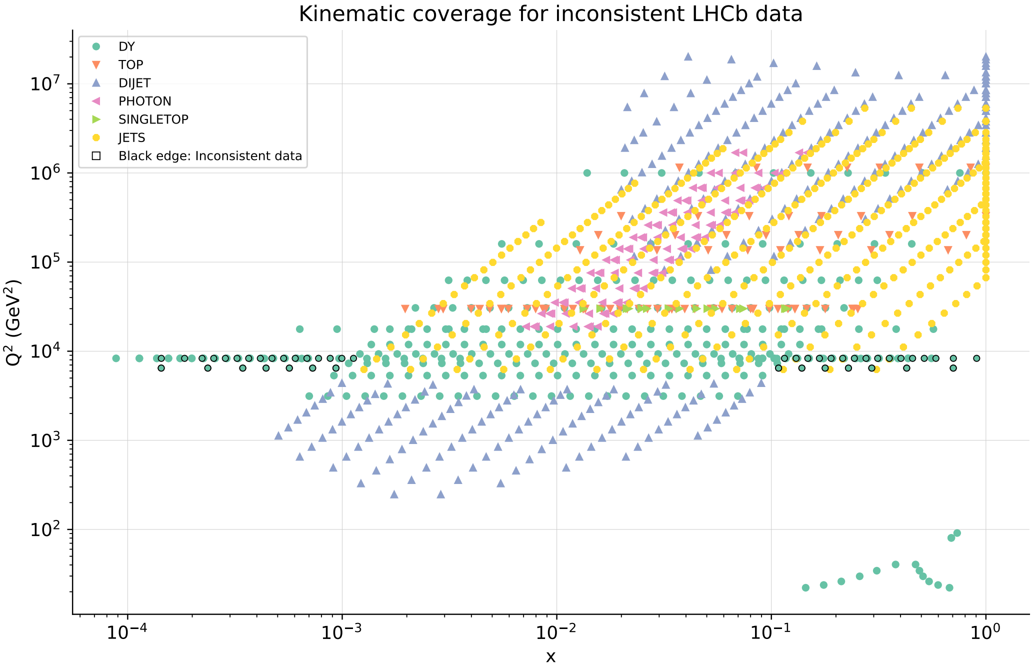


Figure 36: Kinematic coverage of the out-of-sample data compared to the in-sample inconsistent LHCb datasets.

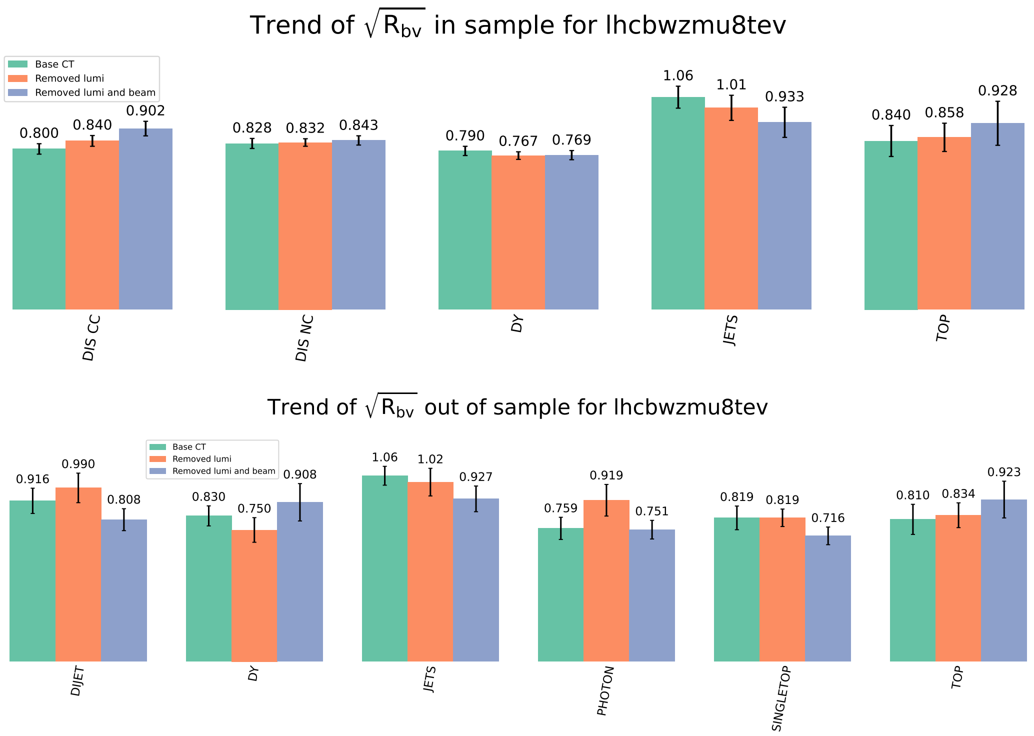


Figure 35: Trend of $\sqrt{\mathcal{R}_{bv}}$ for different levels of inconsistency incorporated within LHCb systematic uncertainties.

5.5.1 *Conclusive remarks*

First of all, we conclude that the χ^2 and the related statistical estimators are incapable of determining whether the PDF fit was performed on inconsistent data. This agrees with the observations made in the previous sections, where inconsistencies were introduced in different kinematic regions and processes.

The most interesting conclusion can be made from the in-sample DIS bootstrap distribution of the bias-variance ratio. Indeed, as one can see from Figure 36 and Figure 21, LHCb data and charged current DIS data are sensitive to the large-x behavior of PDFs. We see in Figure 34 that, as a consequence of the introduction of inconsistencies at large-x, in-sample DIS data's uncertainties were underfitted. The opposite happened at small-x where, as already observed in Section 5.3, the neutral current DIS data dominate over the DY data among which LHCb contributes. As expected, the neutral current DIS uncertainties were correctly predicted by the closure test.

5.6 FINAL REMARKS

We recall here the results obtained in the four closure tests and the conclusions anticipated for each of them. As explained in Section 5.1, inconsistencies have been introduced in four different datasets, corresponding to different kinematic region and number of datapoints in the training set. We found that the four closure test led to as many different outcomes and, even though it is difficult to generalize without performing a larger number of analyses for each situation, we can make the following reasonable conclusions.

The impact of inconsistent data on a PDF fit can be seen in the way it determines the uncertainties, but not in the central values. This is suggested by the fact that the differences in the χ^2 for the reference closure test and all four inconsistent fits is at the level of statistical fluctuations. From the distributions of the bias-to-variance indicator, we conclude that the impact of inconsistent data on uncertainties depends on the dataset in which they are introduced. We can distinguish two situations:

1. the inconsistent data cover a wide kinematic region without being constrained by PDFs determined from other data and evolved through the DGLAP equations;

2. several groups of data constrain a kinematic region, and inconsistencies are introduced in one of them.

The first situation happens for jet data in Section 5.2 and for LHCb data in Section 5.5. In the former case, almost every uncertainty has been correctly delivered by the closure test except from the one of jet data, i. e. where the inconsistency was placed. In the latter, uncertainties on DY data – among which were the inconsistencies – have not been corrupted. On the other hand, the charged current DIS sector was fitted with underestimated unceratinties, opposite to jet data. This seems to suggest that LHCb data are constrained by a PDF feature which is determined in the charged current DIS region, where the PDFs were parametrized, and evolved throug the DGLAP equations into the inconsistent data region. If such is the case, the LHCb closure tests are a good example of how the DGLAP causality domain can influence the outcomes of PDF fits with inconsistent data.

The second situation, where several groups of data constrain a kinematic region, happens in closure tests performed with inconsistent HERA and ATLAS data. The former fits, presented in Section 5.3, are trained on inconsistent data that are part of a process which covers more than the 75% of the training data. For this reason, the methodology most likely produced a set of PDFs that are constrained by such inconsistencies. Indeed, we see that uncertainties have been correctly predicted for the DIS data region and, on the contrary, DY data have been delivered with underfitted uncertainties both in and out of sample. This suggests that the weight of DIS was too large that the methodology could not consistently minimize the χ^2 without learning the inconsistency in the data, thereby penalizing the other datasets in the same kineamtic region.

Confirmations on the statement made above come from the analysis of the results of Section 5.4. Here, the inconsistent DY data could not be fitted with correct uncertainties due to the presence of a heavier dataset, DIS data, that led the methodology detect the inconsistency.

With the results gathered in this thesis, we can assume that the presence of inconsistencies in experimental dataset has an impact on the fitting methodology that entirely depends on the inconsistent data. Additional tests can be made in order to confirm the hypotesis. First of all, one can choose a different in-sample and out-of-sample dataset division. For instance, one can make folds of experimental datasets in order to use half of their datapoints in sample and the other half out of sample. Such dataset division would shed more light on the behavior of DIS data out of sample, or jet data in sample.

Another possible follow-up analysis could be performed on the same dataset division used here, by manually setting the weigths of a specific dataset in the calculation of the χ^2, so as to simulate what happened here with DIS data. In this way, the observations made in this thesis as a consequence of the four trends measured can be stressed with the presence of a wider set of situations.

6 CONCLUSIONS

In this work, we performed an explicit measurement of the impact of inconsistent data on the NNPDF4.0 fitting methodology. The main motivation for this research can be found in the fact that inconsistencies are one of the possible explainations for large values of χ^2 found in the latest PDF determinations.

We compared the results of a standard NNPDF4.0 closure test, which is by definition free of inconsistencies, with four different closure tests performed on inconsistent data. Inconsistencies arise whenever the nominal uncertainty on a certain set of experimental data is smaller than its actual one. For this reason, the inconsistent closure test are produce by fitting artificially generated data with an underestimated covariance matrix, obtained through manipulations of the experimental covariance matrix eigenvalues and columns.

The results, which have been layed out in Chapter 5, lead to the following conclusions. First of all, we can state that the presence of inconsistencies in experimental data has a minimal impact on the fit χ^2. This is observed for all four cases studied in this thesis in the fact that changes in the global χ^2 are at the level of statistical fluctuations, as well as the χ^2 evaluated for each single process. Moreover, the closure test estimators that depend explicitly on the fit quality, i. e. Δ_{χ^2} and φ, are indeed unchanged when comparing the inconsistent closure tests to the reference one. We conclude that, using standard fit quality parameters that are computed also in a PDF fit to experimental data, the impact of inconsistent data cannot be measured.

It is easy to identify two possible reasons why this happens: either the χ^2 is unable to measure the fit response to inconsistencies, or inconsistent data do not affect a PDF fit at all. In order to determine which assumption is correct, we turn to closure test estimators. By looking at the changes in the bias-variance ellipsoids in data space, given by the ratio $\sqrt{\mathcal{R}_{b\nu}}$, we can state that the impact of inconsistent data on a PDF fit is not negligible and can be measured in a closure test. The PDF fitting methodology responds to the presence of inconsistent data by delivering uncertainties that are over/under estimated depending on the kinematic region and weight of the inconsistent dataset.

The results of this work seem to agree with the hypotesis that inconsistencies are learned by the methodology if the inconsistent dataset has a sufficiently large impact in the minimization strategy. On this behalf, two opposite situations happen for jets and DIS data. When inconsistencies are introduced in the former, based on the impact of such dataset on the overall training dataset, the fit compensates for inconsistent data by overfitting their uncertainties. On the other hand, when trained on inconsistent DIS data, the fit learns the inconsistencies and delivers the uncertainties for such process in the same way as the reference unbiased closure test. As a consequence of the fact that inconsistent data have not been detected and properly treated, in the latter case the PDF determination delivers wrong predictions for consistent data that cover the same kinematic region of inconsistent ones.

The analyses performed in this thesis indicate that the impact of inconsistentcies on the NNPDF fitting framework depends on the dataset upon which they were included. This opens the possibility for future work on the subject. First of all, one can perform the same analyses with a different in-sample and out-of-sample dataset division, in order to have the same amount of data for a single process in and out of sample. Secondly, one can use the division adopted here and manually set the weigths of single datasets in the calculation of the fit quality, in order to simulate what happened naturally here for DIS data.

Part IV

APPENDIX

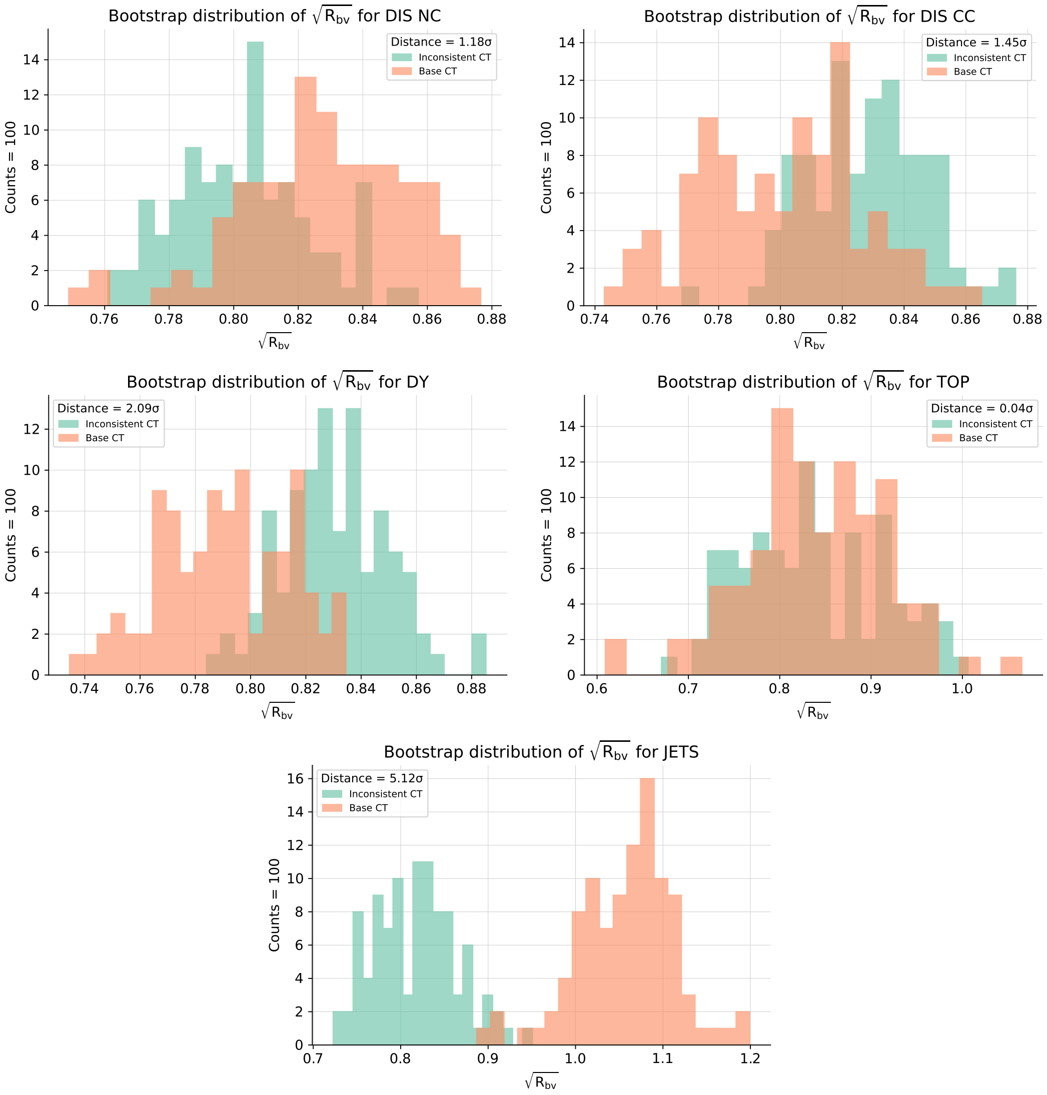
A

BIAS-VARIANCE BOOTSTRAP DISTRIBUTIONS

We display the entire set of figures showing the bootstrap distribution of the $\sqrt{\mathcal{R}_{\mathrm{b}\nu}}$ estimator, divided by process, for the results delivered in Chapter 5.

SINGLE INCLUSIVE JETS

The following plots integrate the results of Section 5.2.



Bootstrap distribution of $\sqrt{R_{bv}}$ for DIS NC

Bootstrap distribution of $\sqrt{R_{bv}}$ for DIS CC

Bootstrap distribution of $\sqrt{R_{bv}}$ for DY

Bootstrap distribution of $\sqrt{R_{bv}}$ for TOP

Bootstrap distribution of $\sqrt{R_{bv}}$ for JETS

Figure 37: In-sample $\sqrt{\mathcal{R}_{\mathrm{b}\nu}}$ bootstrap distribution. Inconsistencies in jet flavour composition are compared with standard closure test.

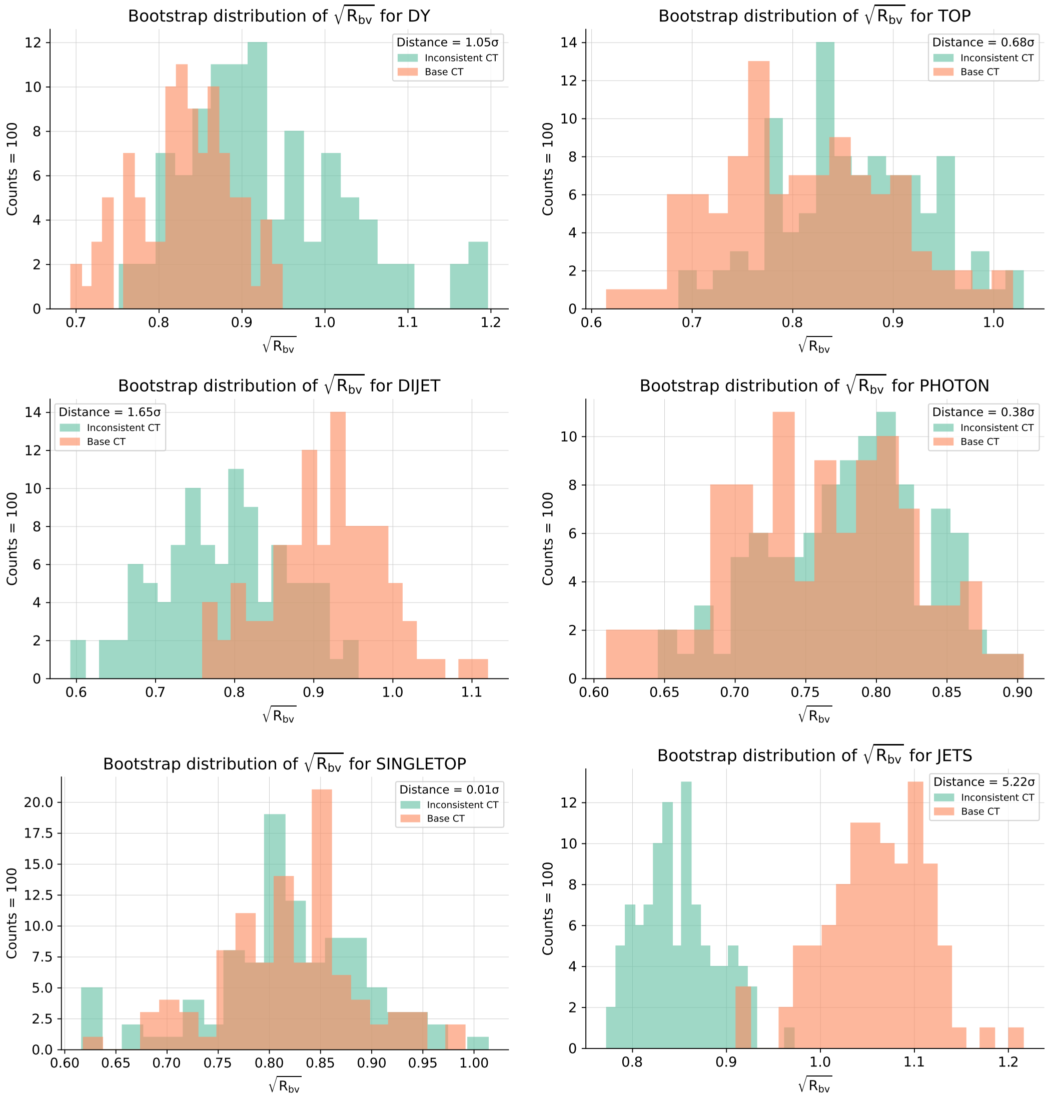


Figure 38: Out-of-sample $\sqrt{\mathcal{R}_{bv}}$ bootstrap distribution. Inconsistencies in jet flavour composition are compared with standard closure test.

NEUTRAL CURRENT HERA COMBINED

The following plots integrate the results of Section 5.3.

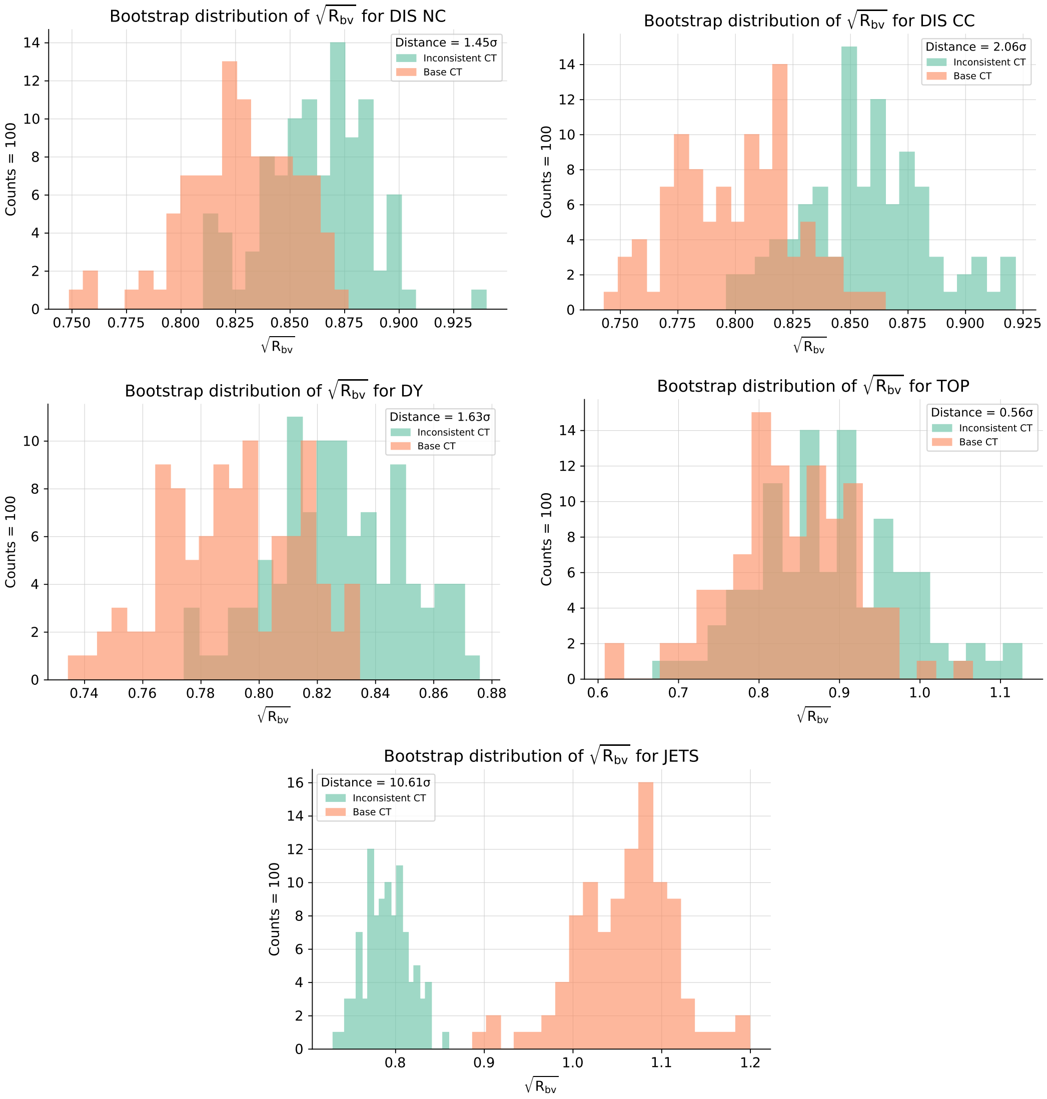


Figure 39: In-sample $\sqrt{\mathcal{R}_{bv}}$ bootstrap distribution. Inconsistencies in neutral current DIS are compared with standard closure test.

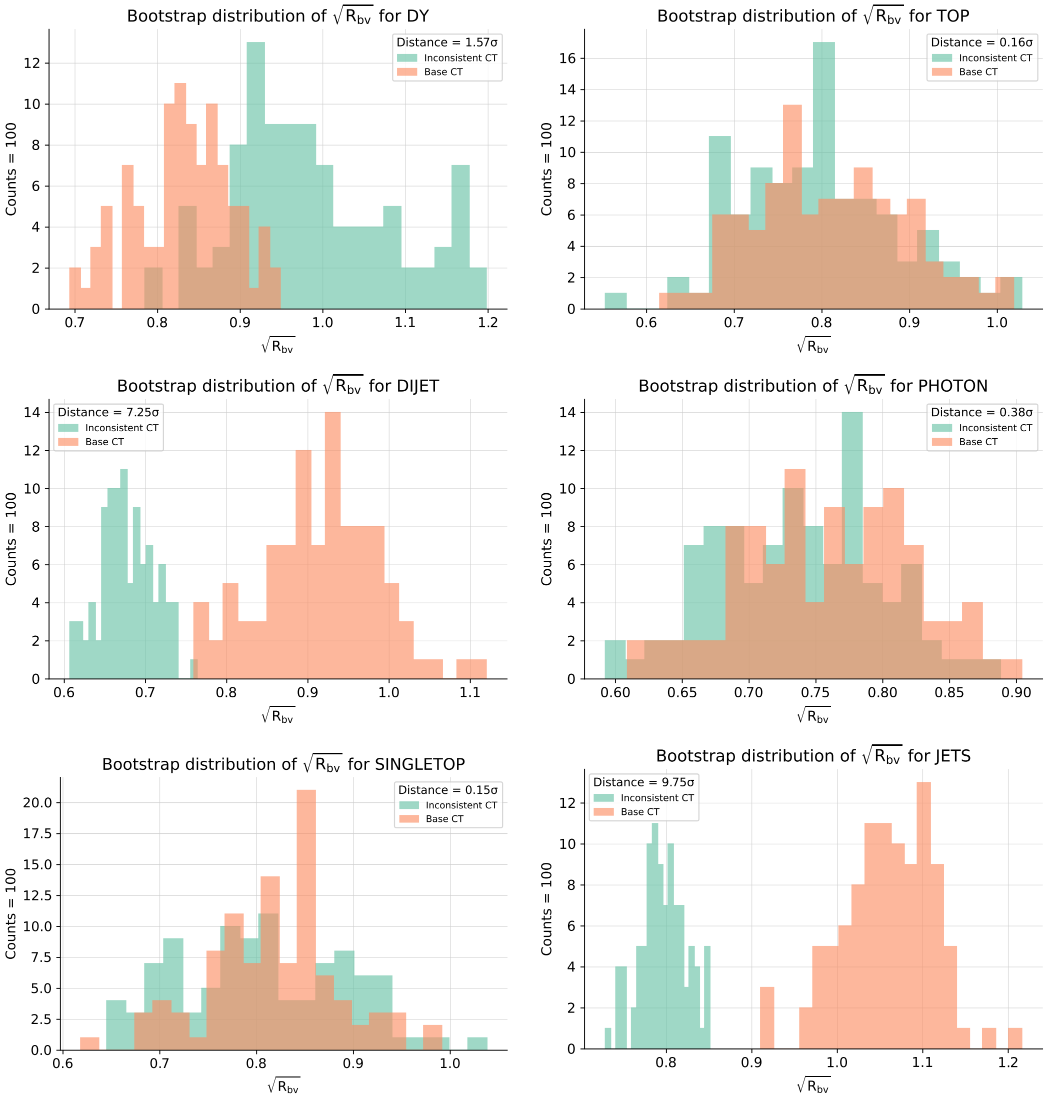


Figure 40: Out-of-sample $\sqrt{\mathcal{R}_{b\nu}}$ bootstrap distribution. Inconsistencies in neutral current DIS are compared with standard closure test.

ATLAS ELECTROWEAK BOSON PRODUCTION

The following plots integrate the results of Section 5.4.

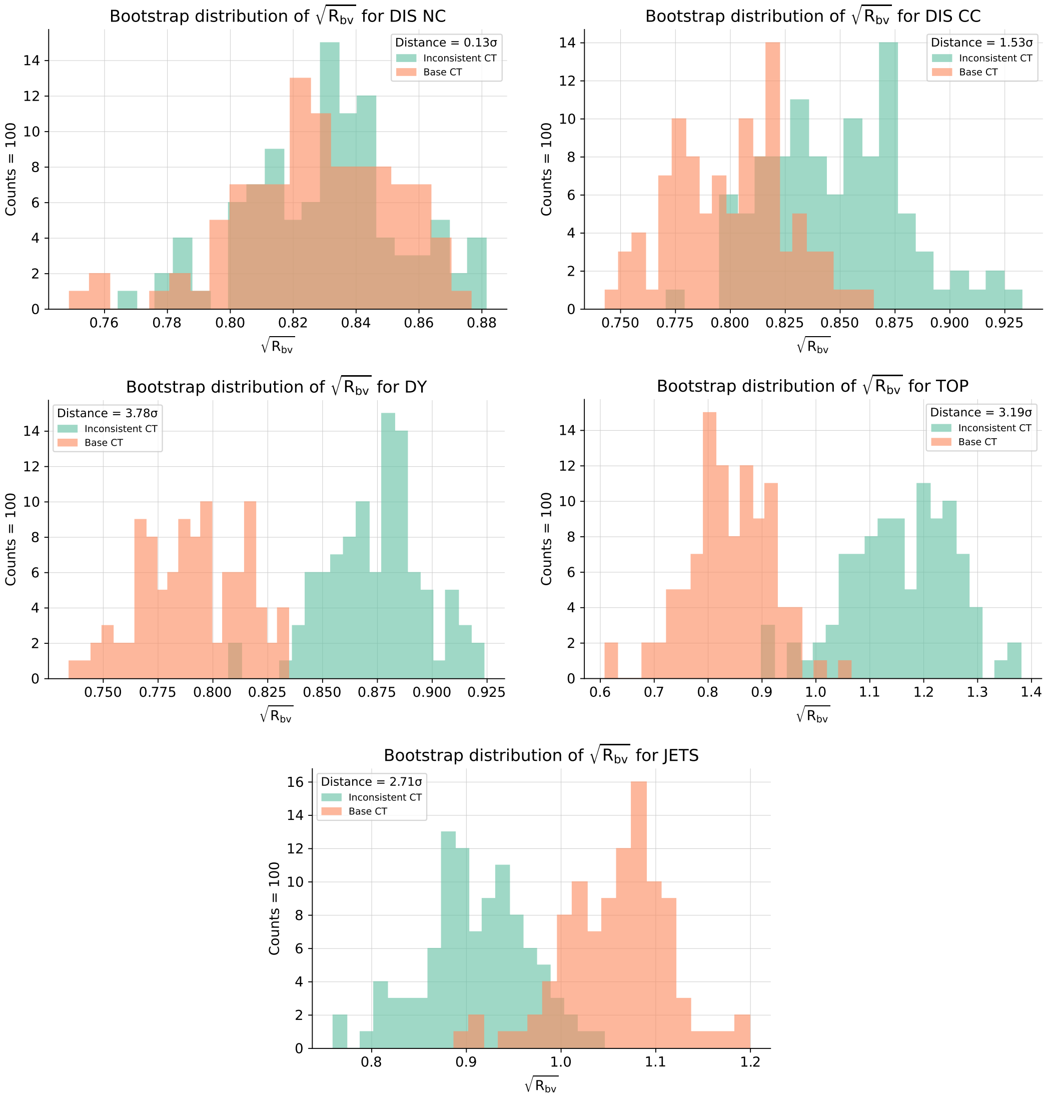


Figure 41: In-sample $\sqrt{\mathcal{R}_{b\nu}}$ bootstrap distribution. Inconsistencies in ATLAS DY data are compared with standard closure test.

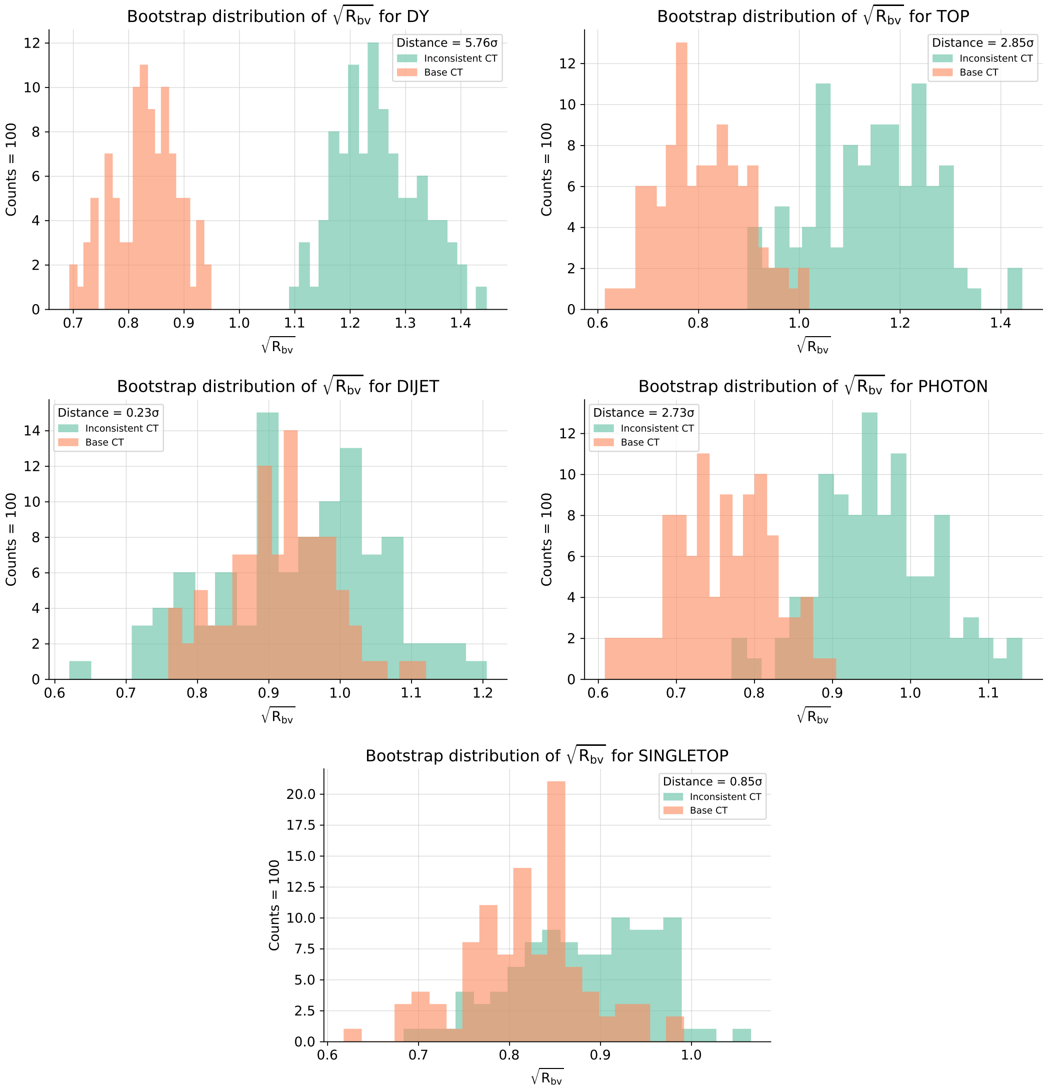


Figure 42: Out-of-sample $\sqrt{\mathcal{R}_{bv}}$ bootstrap distribution. Inconsistencies in ATLAS DY data are compared with standard closure test.

LHCb electroweak boson production

The following plots integrate the results of Section 5.5.

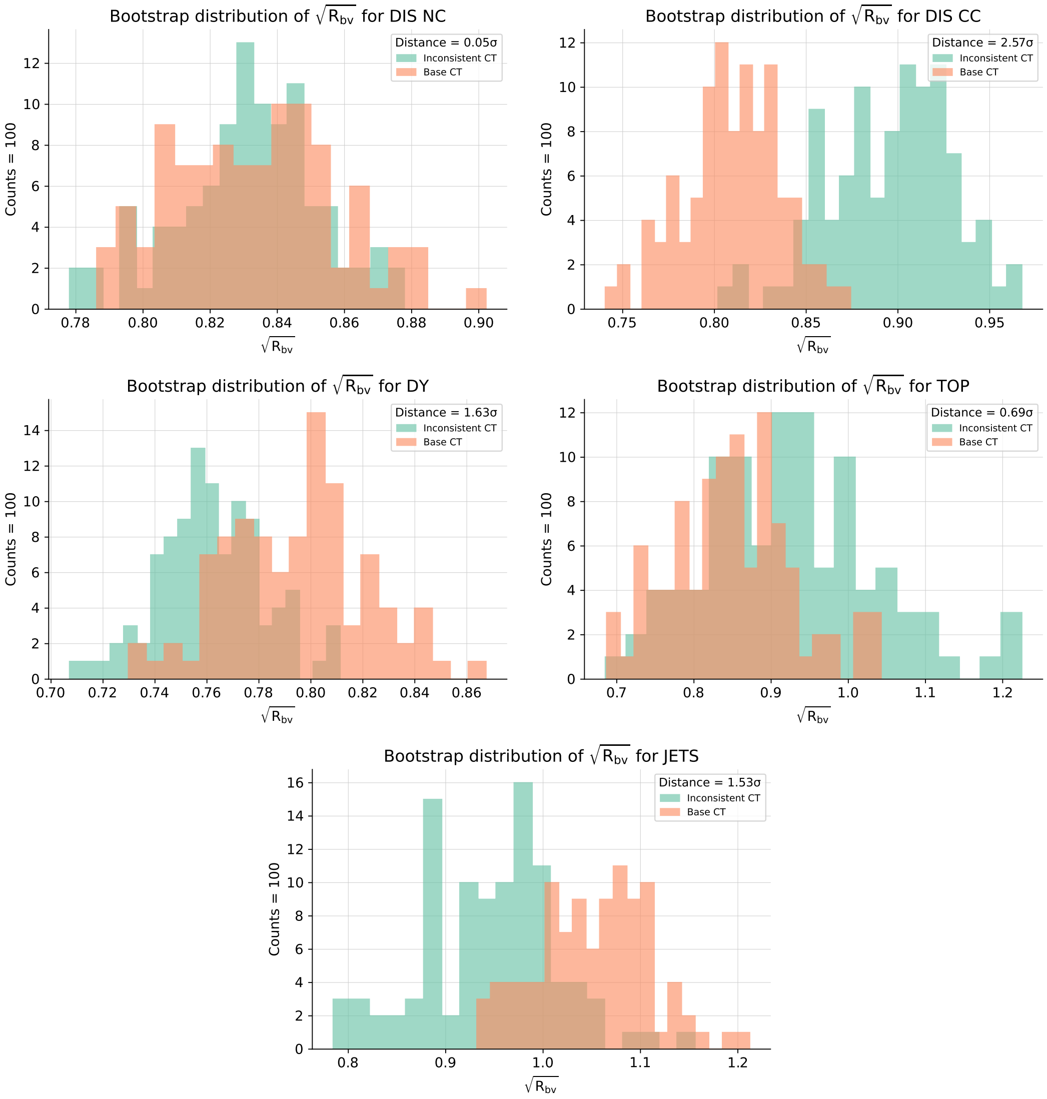


Figure 43: In-sample $\sqrt{\mathcal{R}_{b\nu}}$ bootstrap distribution. Inconsistencies in LHCb DY data are compared with standard closure test.

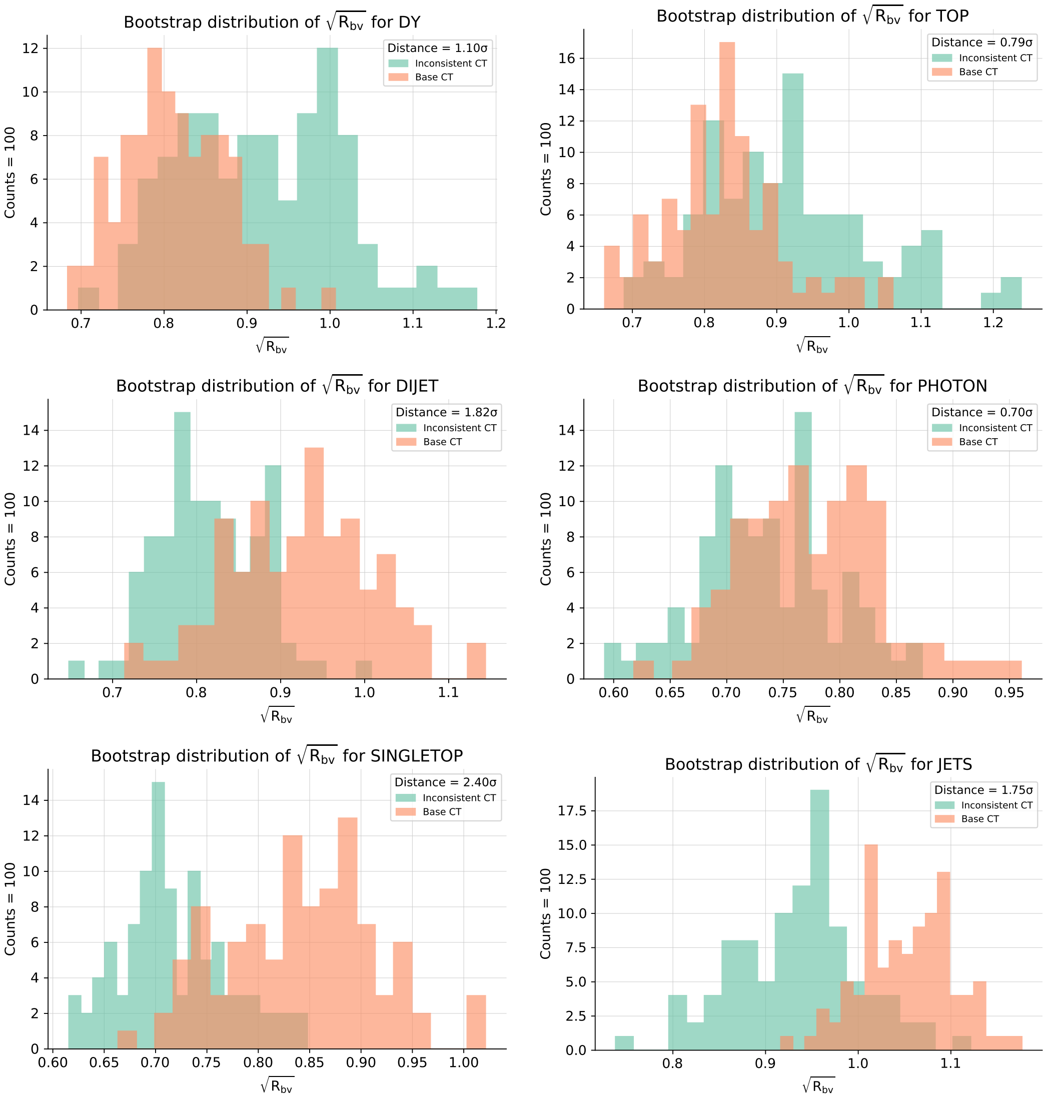


Figure 44: Out-of-sample $\sqrt{\mathcal{R}_{bv}}$ bootstrap distribution. Inconsistencies in LHCb DY data are compared with standard closure test.

BIBLIOGRAPHY

[1] Michael Edward Peskin et al. *An Introduction to Quantum Field Theory*. Reading, USA: Addison-Wesley (1995) 842 p. Westview Press, 1995 (cit. on p. 3).

[2] Murray Gell-Mann. "Symmetries of baryons and mesons." In: *Phys. Rev.* 125 (1962), pp. 1067–1084. DOI: 10.1103/PhysRev.125.1067 (cit. on p. 3).

[3] G. Zweig. "An SU(3) model for strong interaction symmetry and its breaking. Version 2." In: *DEVELOPMENTS IN THE QUARK THEORY OF HADRONS. VOL. 1. 1964 - 1978*. Ed. by D. B. Lichtenberg et al. Feb. 1964, pp. 22–101 (cit. on p. 3).

[4] David J. Gross et al. "Ultraviolet Behavior of Nonabelian Gauge Theories." In: *Phys. Rev. Lett.* 30 (1973), pp. 1343–1346. DOI: 10.1103/PhysRevLett.30.1343 (cit. on p. 4).

[5] H. David Politzer. "Reliable Perturbative Results for Strong Interactions?" In: *Phys. Rev. Lett.* 30 (1973), pp. 1346–1349. DOI: 10.1103/PhysRevLett.30.1346 (cit. on p. 4).

[6] F. Herzog et al. "The five-loop beta function of Yang-Mills theory with fermions." In: *JHEP* 02 (2017), p. 090. DOI: 10.1007/JHEP02(2017)090 (cit. on p. 6).

[7] David D'Enterria et al. *High-precision α_s measurements from LHC to FCC-ee*. 2015. DOI: 10.48550/ARXIV.1512.05194 (cit. on p. 7).

[8] John C. Collins et al. "Factorization of Hard Processes in QCD." In: (2004). DOI: 10.48550/ARXIV.HEP-PH/0409313 (cit. on p. 8).

[9] T. D. Lee et al. "Degenerate Systems and Mass Singularities." In: *Phys. Rev.* 133 (1964), B1549–B1562. DOI: 10.1103/PhysRev.133.B1549 (cit. on p. 11).

[10] T. Kinoshita. "Mass singularities of Feynman amplitudes." In: *J. Math. Phys.* 3 (1962), pp. 650–677. DOI: 10.1063/1.1724268 (cit. on p. 11).

[11] Guido Altarelli et al. "Asymptotic Freedom in Parton Language." In: *Nucl. Phys. B* 126 (1977), pp. 298–318. DOI: 10.1016/0550-3213(77)90384-4 (cit. on p. 14).

[12] Yuri L. Dokshitzer. "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics." In: *Sov. Phys. JETP* 46 (1977), pp. 641–653 (cit. on p. 14).

[13] V. N. Gribov et al. "Deep inelastic e p scattering in perturbation theory." In: *Sov. J. Nucl. Phys.* 15 (1972), pp. 438–450 (cit. on p. 14).

[14] A. Vogt et al. "The Three-loop splitting functions in QCD: The Singlet case." In: *Nucl. Phys. B* 691 (2004), pp. 129–181. DOI: 10.1016/j.nuclphysb.2004.04.024 (cit. on p. 14).

[15] S. Moch et al. "The Three loop splitting functions in QCD: The Nonsinglet case." In: *Nucl. Phys. B* 688 (2004), pp. 101–134. DOI: 10.1016/j.nuclphysb.2004.03.030 (cit. on p. 14).

[16] Valerio Bertone et al. "APFEL: A PDF evolution library with QED corrections." In: *Computer Physics Communications* 185.26 (2014), pp. 1647–1668. DOI: 10.1016/j.cpc.2014.03.007 (cit. on p. 16).

[17] Thomas Appelquist et al. "Infrared singularities and massive fields." In: *Phys. Rev. D* 11 (10 1975), pp. 2856–2861. DOI: 10.1103/PhysRevD.11.2856 (cit. on p. 18).

[18] S. Amoroso et al. "Les Houches 2019: Physics at TeV Colliders: Standard Model Working Group Report." In: *11th Les Houches Workshop on Physics at TeV Colliders: PhysTeV Les Houches*. 2020 (cit. on p. 19).

[19] Tie-Jiun Hou et al. "New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC." In: *Phys. Rev. D* 103.1 (2021), p. 014013. DOI: 10.1103/PhysRevD.103.014013 (cit. on pp. 20, 45).

[20] S. Bailey et al. "Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs." In: *Eur. Phys. J. C* 81.4 (2021), p. 341. DOI: 10.1140/epjc/s10052-021-09057-0 (cit. on p. 20).

[21] H. Abramowicz et al. "Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data." In: *Eur. Phys. J. C* 75.12 (2015), p. 580. DOI: 10.1140/epjc/s10052-015-3710-4 (cit. on pp. 20, 74).

[22] S. Alekhin et al. "The ABM parton distributions tuned to LHC data." In: *Phys. Rev. D* 89.5 (2014), p. 054028. DOI: 10.1103/PhysRevD.89.054028 (cit. on p. 20).

[23] F. Chollet. *Deep Learning with Python*. Manning Publications Company, 2017. ISBN: 9781617294433 (cit. on p. 20).

[24] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015 (cit. on p. 21).

[25] Richard D. Ball et al. "The path to proton structure at 1% accuracy." In: *Eur. Phys. J. C* 82.5 (2022), p. 428. DOI: 10.1140/epjc/s10052-022-10328-7 (cit. on pp. 26, 27, 29, 34, 37, 45, 52, 57, 59).

[26] J. Bergstra et al. *Making a Science of Model Search*. 2012. DOI: 10.48550/ARXIV.1209.5111 (cit. on p. 28).

[27] Richard D. Ball et al. "Parton distributions from high-precision collider data." In: *Eur. Phys. J. C* 77.10 (2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5 (cit. on pp. 29, 45, 59).

[28] Richard D. Ball et al. "Parton distributions for the LHC Run II." In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040 (cit. on p. 30).

[29] Richard D. Ball et al. "A first unbiased global NLO determination of parton distributions and their uncertainties." In: *Nucl. Phys. B* 838 (2010), pp. 136–206. DOI: 10.1016/j.nuclphysb.2010.05.008 (cit. on p. 31).

[30] Alessandro Candido et al. "Can $\overline{\mathrm{MS}}$ parton distributions be negative?" In: *JHEP* 11 (2020), p. 129. DOI: 10.1007/JHEP11(2020)129 (cit. on p. 31).

[31] G. D'Agostini. "On the use of the covariance matrix to fit correlated data." In: *Nucl. Instrum. Meth. A* 346 (1994), pp. 306–311. DOI: 10.1016/0168-9002(94)90719-6 (cit. on p. 34).

[32] Luigi Del Debbio et al. "Bayesian approach to inverse problems: an application to NNPDF closure testing." In: *Eur. Phys. J. C* 82.4 (2022), p. 330. DOI: 10.1140/epjc/s10052-022-10297-x (cit. on p. 35).

[33] Rabah Abdul Khalek et al. "Parton Distributions with Theory Uncertainties: General Formalism and First Phenomenological Studies." In: *Eur. Phys. J. C* 79.11 (2019), p. 931. DOI: 10.1140/epjc/s10052-019-7401-4 (cit. on p. 36).

[34] G. Watt et al. "Study of Monte Carlo approach to experimental uncertainty propagation with MSTW 2008 PDFs." In: *JHEP* 08 (2012), p. 052. DOI: 10.1007/JHEP08(2012)052 (cit. on pp. 39, 41).

[35] A. D. Martin et al. "Parton distributions for the LHC." In: *Eur. Phys. J. C* 63 (2009), pp. 189–285. DOI: 10.1140/epjc/s10052-009-1072-5 (cit. on p. 40).

[36] Stefano Carrazza et al. "An Unbiased Hessian Representation for Monte Carlo PDFs." In: *Eur. Phys. J. C* 75.8 (2015), p. 369. DOI: 10.1140/epjc/s10052-015-3590-7 (cit. on p. 41).

[37] Stefano Carrazza et al. "Specialized minimal PDFs for optimized LHC calculations." In: *Eur. Phys. J. C* 76.4 (2016), p. 205. DOI: 10.1140/epjc/s10052-016-4042-8 (cit. on p. 41).

[38] Jon Pumplin. "Experimental consistency in parton distribution fitting." In: *Phys. Rev. D* 81 (2010), p. 074010. DOI: 10.1103/PhysRevD.81.074010 (cit. on p. 45).

[39] Roy Stegeman. "The negligible impact of experimental inconsistencies in the NNPDF4.0 global dataset." In: *PoS* ICHEP2022 (2022), p. 787. DOI: 10.22323/1.414.0787 (cit. on p. 45).

[40] Richard D. Ball et al. "Parton distributions for the LHC Run II." In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040 (cit. on p. 47).

[41] Luigi Del Debbio et al. "Bayesian approach to inverse problems: an application to NNPDF closure testing." In: *Eur. Phys. J. C* 82.4 (2022), p. 330. DOI: 10.1140/epjc/s10052-022-10297-x (cit. on p. 49).

[42] Aurore Courtoy et al. "Parton distributions need representative sampling." In: (May 2022) (cit. on p. 53).

[43] Morad Aaboud et al. "Measurement of differential cross sections and W^+/W^- cross-section ratios for W boson production in association with jets at $\sqrt{s} = 8$ TeV with the ATLAS detector." In: *JHEP* 05 (2018). [Erratum: JHEP 10, 048 (2020)], p. 077. DOI: 10.1007/JHEP05(2018)077 (cit. on p. 70).

[44] Morad Aaboud et al. "Measurement of the inclusive jet cross-sections in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector." In: *JHEP* 09 (2017), p. 020. DOI: 10.1007/JHEP09(2017)020 (cit. on p. 70).

[45] Georges Aad et al. "Determination of the parton distribution functions of the proton using diverse ATLAS data from pp collisions at $\sqrt{s} = 7$, 8 and 13 TeV." In: *Eur. Phys. J. C* 82.5 (2022), p. 438. DOI: 10.1140/epjc/s10052-022-10217-z (cit. on p. 70).

[46] B. Efron. “Bootstrap Methods: Another Look at the Jackknife.” In: *The Annals of Statistics* 7.1 (1979), pp. 1 –26. DOI: 10.1214/aos/1176344552 (cit. on p. 71).

[47] Georges Aad et al. “Measurement of the inclusive $W^{\pm}$ and Z/gamma cross sections in the electron and muon decay channels in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector.” In: *Phys. Rev. D* 85 (2012), p. 072004. DOI: 10.1103/PhysRevD.85.072004 (cit. on p. 78).

[48] Roel Aaij et al. “Measurement of forward W and Z boson production in pp collisions at $\sqrt{s} = 8$ TeV.” In: *JHEP* 01 (2016), p. 155. DOI: 10.1007/JHEP01(2016)155 (cit. on p. 82).