PARTON DISTRIBUTIONS AT THE DAWN OF THE LHC*

Stefano Forte

Dipartimento di Fisica, Università di Milano and INFN, Sezione di Milano
Via Celoria 16, 20133 Milano, Italy

(Received November 22, 2010)

To the memory of Wu-Ki Tung

We review basic ideas and recent developments on the determination of the parton substructure of the nucleon in view of applications to precision hadron collider physics. We review the way information on PDFs is extracted from the data exploiting QCD factorisation, and discuss the current main two approaches to parton determination (Hessian and Monte Carlo) and their use in conjunction with different kinds of parton parameterisation. We summarise the way different physical processes can be used to constrain different aspects of PDFs. We discuss the meaning, determination and use of parton uncertainties. We briefly summarise the current state of the art on PDFs for LHC physics.

PACS numbers: 12.38.–t, 12.39.St

1. QCD in the LHC era

The theory and phenomenology of the strong interactions [1] have witnessed an impressive development in the last two decades, driven first by the availability of HERA [2] — a QCD machine — and then by the needs of present (Tevatron) and especially upcoming (LHC) hadron colliders [3]. The LHC will be looking for new physics in hadronic collisions.

The last time this happened was back in the early eighties, when the W and Z were discovered at the SPS collider [4] — and, of course, one may argue to which extent the W and Z then were genuinely "new" physics. At the time, QCD was at best a semi-quantitative theory: for example, in Ref. [5] a measured W cross-section of 0.63 ± 0.10 nb (at $\sqrt{s} = 630$ GeV) was described as "in agreement with the theoretical expectation" [6] of $0.47^{+0.14}_{-0.08}$ nb. One

* Lecture presented at the L Cracow School of Theoretical Physics "Particle Physics at the Dawn of the LHC", Zakopane, Poland, June 9–19, 2010.

reason why at that time a NLO calculation could not be expected to agree with the data to better than 20% is that the knowledge of nucleon structure was at the time extremely sketchy: a parton set consisted of three parton distributions (valence, quark sea, and gluon), differences at the 30% level between sets would be standard, and, of course, there would be no idea on the associate uncertainty (see Fig. 1).

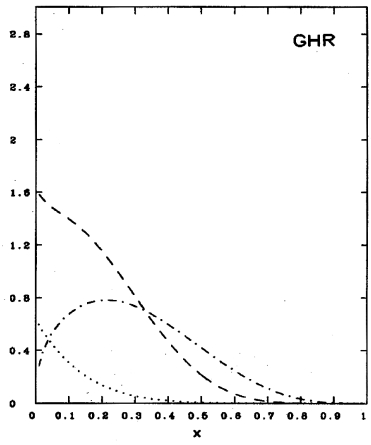

FIG. 25. Parton distributions of Glück, Hoffmann, and Reya (1982), at $Q^2=5$ GeV^2: valence quark distribution $x[u_v(x)+d_v(x)]$ (dotted-dashed line), $xG(x)$ (dashed line), and q_v (dotted line).

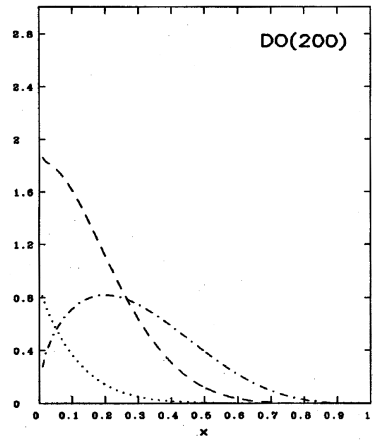

FIG. 27. "Soft-gluon" ($\Lambda=200$ MeV) parton distributions of Duke and Owens (1984) at $Q^2=5$ GeV^2: valence quark distribution $x[u_v(x)+d_v(x)]$ (dotted-dashed line), $xG(x)$ (dashed line), and $q_v(x)$ (dotted line).

Fig. 1. Comparison of two parton distribution sets [7, 8] from the early eighties (From Ref. [9]).

The evolution in time of parton distributions (see Fig. (2)) since then shows that it is only during the HERA age that predictions from different groups converged: this is both a consequence and a cause of the fact that perturbative QCD has now turned into a quantitative theory, which leads to predictions for hard processes with typical accuracies below 10%, and often of a few percent. Perturbative QCD today is an integral part of the Standard Model, and it is tested to an accuracy which is comparable to that of the electroweak sector: in fact, HERA has played for QCD a similar role as LEP for electroweak theory. In the last decade, theoretical and phenomenological progress has been impressive: at the LHC we can envisage quantitative control of QCD contribution to collider signal and background processes at the percent level, as will be necessary for discovery at the LHC [3].

Progress in QCD has taken place in (at least) five distinct directions, namely (listing from the bottom beam nucleons up to the final state): First, the understanding of the structure of the nucleon in terms of parton distributions has now become a quantitative science. Second, perturbative

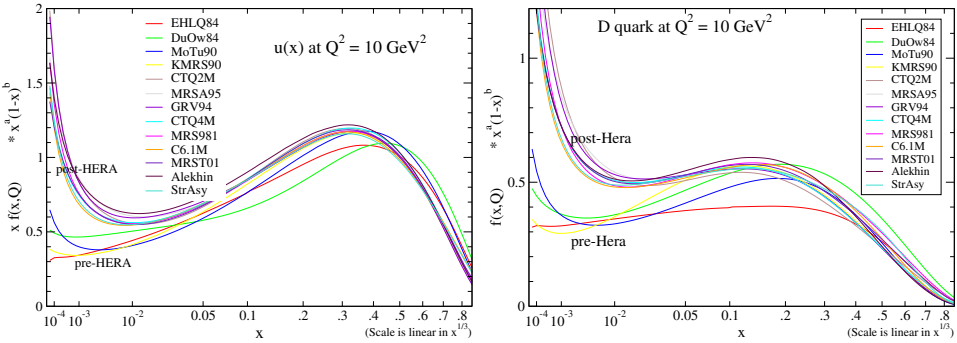

Fig. 2. Historical evolution of the up (left) and down (right) quark distributions (from Ref. [10]).

computations are being pushed to hard processes with increasingly high numbers of particles and at increasingly high orders, thanks to the development of a variety of techniques which include twistor methods, analyticity techniques, and the use of exact results from supersymmetric QCD and the AdS/CFT duality. Third, all-order resummation of perturbation theory is being extended in various kinematic regimes (small x and large x) to new classes of observables (typically less inclusive), to higher logarithmic orders, and it is being accomplished using perturbative renormalisation-group methods, path integral techniques, and effective field theory methods. Fourth, definitions of jet observables which are both consistent with perturbative factorisation to all orders and numerically efficient have been constructed theoretically and implemented in computer interfaces. Fifth, new collinear subtraction algorithms have been developed which make the development and implementation of next-to-leading order Monte Carlo codes possible.

The lectures at the Zakopane school on which this paper is based, ambitiously entitled "QCD at the dawn of the LHC", covered the first three of these topics: parton distributions (PDFs), perturbative computations, and resummation. Here we will concentrate on PDFs; recent good reviews of progress in perturbative computations are in Refs. [11, 12], while a comprehensive overview of resummation is unfortunately not available yet. At Zakopane, jets and Monte Carlos where discussed by other speakers; excellent recent reviews of these topics are in Refs. [13, 14] respectively.

The purpose of this overview of PDFs is both to provide an elementary introduction to the subject, and also a summary of recent developments, several of which are little known outside a small group of practitioners. Progress in this field has been largely driven by two series of HERA-LHC workshops 2004–2005 and 2006–2007, which have organised and stimulated the transfer of know-how from deep-inelastic scattering to hadron collider physics, and whose results are collected in the respective reports

Refs. [15,16]. Since 2007, the PDF4LHC working group has been formed [17] with a mandate from the CERN directorate to steer and coordinate research on PDFs for the LHC community: many of the more recent ideas discussed here were developed in the context of this working group.

This review is organised as follows. We will start with the more basic concepts, then work our way to somewhat more advanced developments. First, we will very briefly review some basic (mostly kinematic) facts on QCD factorisation. We will then present the two main existing approaches (Hessian and Monte Carlo) to the determination of PDFs and the way they are used in conjunction with various forms of parton parameterisation. Next, we will review standard ideas on how information on PDFs can be extracted from the data. We will then discuss in some detail the problem of PDF uncertainties — what they mean and how they are determined. In the final section, we will briefly summarise the state of the art: the role of theoretical uncertainties, and the current understanding of standard candle processes at the LHC.

These lectures are dedicated to the memory of Wu-Ki Tung, who pioneered this field, pursued it for more than 30 years, and shaped much of our current understanding of it. He encouraged me to keep pursuing it in an e-mail he sent to me on March 27, 2009.

2. Factorisation

Factorisation of cross-sections into hard (partonic) cross-sections and universal parton distributions is the basic property of QCD which makes it predictive in the perturbative regime, and which enables a determination of parton distributions. Here we only review some basic facts which will be useful for our subsequent discussion, while referring to standard textbooks [18] and recent reviews [1] for a detailed treatment.

2.1. Electro- and hadro-production kinematics

The basic factorisation for hadroproduction processes has the structure

$$\begin{aligned}\sigma_X\left(s,M_X^2\right) &= \sum_{a,b}\int_{x_{\min}}^1 dx_1dx_2 f_{a/h_1}\left(x_1,M_X^2\right)f_{b/h_2}\left(x_2,M_X^2\right)\hat{\sigma}_{ab\to X}\left(x_1x_2s,M_X^2\right)\\ &= \sigma_{ab}^0\sum_{a,b}\int_\tau^1\frac{dx_1}{x_1}\int_{\tau/x_1}^1\frac{dx_2}{x_2}f_{a/h_1}\left(x_1,M_X^2\right)f_{b/h_2}\left(x_2,M_X^2\right)C\left(\frac{\tau}{x_1x_2},\alpha_{\rm s}\left(M_X^2\right)\right)\\ &= \int_\tau^1\frac{dx}{x}\mathcal{L}\left(x\right)C\left(\frac{\tau}{x},\alpha_{\rm s}\left(M_X^2\right)\right),\end{aligned}\tag{1}$$

where $f_{a/h_i}(x_i)$ is the distribution of partons of type a in the ith incoming hadron; $\hat{\sigma}_{q_a q_b \to X}$ is the parton-level cross-section for the production of the desired (typically inclusive) final state X; the minimum value of x_i is clearly $x_{\min} = \tau$, with

$$\tau \equiv \frac{M_X^2}{s} \tag{2}$$

the scaling variable of the hadronic process, and in the last step we defined the parton luminosity

$$\begin{aligned} \mathcal{L}(x) \;&\equiv\; \int_x^1 \frac{dz}{z} f_{a/h_1}\left(z, M_X^2\right) f_{b/h_2}\left(\frac{x}{z}, M_X^2\right) \\ &=\; \int_x^1 \frac{dz}{z} f_{a/h_2}\left(z, M_X^2\right) f_{b/h_1}\left(\frac{x}{z}, M_X^2\right) . \end{aligned} \tag{3}$$

The hard coefficient function $C\left(\tau, \alpha_{\rm s}(M_X^2)\right)$ is defined by viewing the parton-level cross-section as a function of the hard scale M_X^2 and the dimensionless ratio of this scale to the center-of-mass energy $\hat{s}$ of the partonic subprocess

$$\frac{M_X^2}{\hat{s}} = \frac{\tau}{x_1 x_2} \tag{4}$$

in terms of the scaling variable. At the lowest order in the strong interaction, the partonic cross-section is then either zero (for partons that do not couple to the given final state at leading order), or else just a function fixed by dimensional analysis times a Dirac delta, and the hard coefficient function is thus defined as

$$\begin{aligned} \hat{\sigma}_{ab\to X}\left(s, M_X^2\right) \;&=\; \sigma_0 C_{ab}\left(\tau, \alpha_{\rm s}\left(M_X^2\right)\right) , \\ C_{ab}\left(x, \alpha_{\rm s}\left(M_X^2\right)\right) \;&=\; c_{ab}\delta(1-x) + O(\alpha_{\rm s}) , \end{aligned} \tag{5}$$

where c_{ab} is a matrix with non-vanishing entries only between quark and antiquark states, which will be discussed explicitly in Sec. 2.2 below (see in particular Eqs. (18) and (19). For example, for virtual photon (Drell–Yan) production c_{ab} is nonzero when ab is a pair of a quark and an antiquark of the same flavour, and in such case $\sigma_0 = \frac{4}{9}\pi\alpha\frac{1}{s}$.

The factorised result Eq. (1) holds both for inclusive cross-sections and for rapidity distributions

$$\frac{d\sigma}{dM_X^2 dY}\left(\tau, Y, M_X^2\right) = \sum_{i,j} \int_{x_1^0}^1 dx_1 \int_{x_2^0}^1 dx_2\, f_i^1\left(x_1, M_X^2\right) f_j^2\left(x_2, M_X^2\right) \times \frac{d\hat{\sigma}_{ij}}{dM_X^2 dy}\left(\frac{\tau}{x_1 x_2}, y, \alpha_{\rm s}\left(M_X^2\right)\right), \tag{6}$$

where the hadronic cross-section is differential with respect to the rapidity Y of the final state X, while the partonic cross-section is differential in partonic rapidity

$$y = Y - \frac{1}{2}\ln\frac{x_1}{x_2}: \tag{7}$$

the effect of parton emission from the incoming hadrons is to perform a Lorentz boost from the hadronic center-of-mass frame to a frame in which the energy of each of the two incoming hadrons are rescaled by x_1 and x_2, respectively. The lower limit of integration $x_{\rm min}$ is then fixed by requiring that the rapidity of the incoming partons be at least sufficient to yield the observed final state rapidity

$$x_1^0 = \sqrt{\tau} e^Y, \qquad x_2^0 = \sqrt{\tau} e^{-Y}. \tag{8}$$

At leading order, the two partons couple directly to the final state so $y = 0$ and

$$Y_{\rm LO} = -\frac{1}{2}\ln\frac{x_1}{x_2}. \tag{9}$$

Equation (1) is to be contrasted with the standard factorisation for the deep-inelastic structure functions $F_i(x, Q^2)$

$$F_i\left(x, Q^2\right) = x \sum_i \int_x^1 \frac{dz}{z} C_i\left(\frac{x}{z}, \alpha_{\rm s}\left(Q^2\right)\right) f_i\left(z, Q^2\right). \tag{10}$$

Here in the argument of the structure function $x = \frac{Q^2}{2p\cdot q}$ is the standard Bjorken variable and the hard coefficient function is the structure function computed with an incoming parton and $f_i(z, Q^2)$ is the distribution of the ith parton in the only incoming hadron. Also in this case at lowest $O(\alpha_{\rm s}^0)$ it is either zero (for incoming gluons) or a constant (an electroweak charge) times a Dirac delta.

Note that the structure functions are related to the cross-section which is actually measured in lepton–hadron scattering by

$$\frac{d^2\sigma^{\mathrm{NC},\ell^\pm}}{dxdQ^2}\left(x,y,Q^2\right) = \frac{2\pi\alpha^2}{xQ^4}\left[Y_+F_2^{\mathrm{NC}}\left(x,Q^2\right) \mp Y_-xF_3^{\mathrm{NC}}\left(x,Q^2\right) - y^2F_L^{\mathrm{NC}}\left(x,Q^2\right)\right] \quad (11)$$

for neutral-current charged lepton $\ell^\pm$ DIS, where the longitudinal structure function is defined as

$$F_{\mathrm{L}}\left(x,Q^2\right) \equiv F_2\left(x,Q^2\right) - 2xF_1\left(x,Q^2\right)\,, \quad (12)$$

and

$$Y_\pm \equiv 1 \pm (1-y)^2\,, \quad (13)$$

in terms of the electron momentum fraction

$$y \equiv \frac{p\,q}{p\,k} = \frac{Q^2}{xs}\,, \quad (14)$$

(not to be confused with the partonic rapidity Eq. (7)), where p and k are respectively the incoming proton and lepton momenta, q is the virtual photon momentum ($q^2 = -Q^2$) and in the last step, which holds neglecting the proton mass, s is the center-of-mass energy of the lepton–proton collision. Similar expressions hold for charged-current charged and neutral lepton scattering.

The set of values of y over which the PDF is probed is of course the same in the hadro- and lepto-production cases, and it ranges from the scaling variable of the hadronic process to one: $x \leq y \leq 1$ in Eq. (10), and $\tau \leq x_1, x_2 \leq 1$ in Eq. (1). The kinematic region which is typical of the collider (HERA) or fixed-target DIS experiments is compared in Fig. 3 to that of LHC processes, whose typical cross-sections are also shown.

There is an important kinematic difference when comparing the hadronic and deep-inelastic factorisation formulae, Eqs. (1) and (10), respectively. This is related to the fact that the leading order coefficient function is proportional to a Dirac delta. For DIS, this implies that at leading order, the value of the structure function at given x determines the quark distributions at the same value of x, and it is only at next-to-leading order, where the coefficient function has a nontrivial dependence on x, that the PDF is probed for all values $x \leq y \leq 1$. But for hadronic processes, because there are two partons in the initial state, even at leading order, for inclusive cross-sections the delta kills one but not both of the convolution integrals in Eq. (1), so all

values $\tau \leq x_i \leq 1$ are probed. However, for rapidity distributions because of the further kinematic constraint (Eq. (9)) the leading order kinematics is also fixed, and for given Y and M_X^2 the momentum fractions of both partons are fixed.

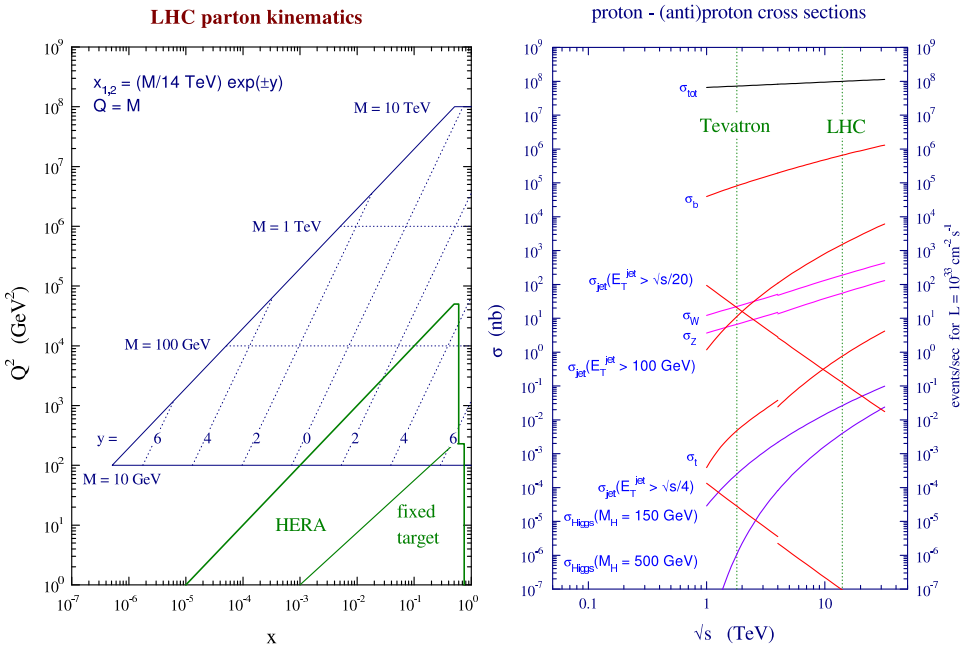

Fig. 3. LHC kinematics (left) and processes (right) (from Ref. [15]).

2.2. Constraints on PDFs

The kinematics of the factorised expressions Eqs. (1) and (10) immediately implies that, as discussed in Sec. 2.1, at leading order deep-inelastic structure functions and rapidity distributions provide a direct handle on individual quark and antiquark PDFs (DIS), or pairs of PDFs (Drell–Yan). It is possible to understand what is dominantly measured by each individual process by looking at the leading order expressions, bearing in mind that, of course, beyond leading order all other contributions turn on (and that NLO corrections can be quite large, in fact of the same order of magnitude as the LO for Drell–Yan).

The leading order contributions to the DIS structure functions F_1 and F_3 are the following (at leading order $F_2 = 2xF_1$)

$$
\begin{aligned}
&\text{NC} \qquad F_1{}^{\gamma} = \sum_i e_i^2 \left(q_i + \bar{q}_i\right) \,, \\
&\text{NC} \qquad F_1{}^{Z,\,\text{int.}} = \sum_i B_i \left(q_i + \bar{q}_i\right) \,, \\
&\text{NC} \qquad F_3{}^{Z,\,\text{int.}} = \sum_i D_i \left(q_i + \bar{q}_i\right) \,, \\
&\text{CC} \qquad F_1^{W^+} = \bar{u} + d + s + \bar{c} \,, \\
&\text{CC} \qquad -F_3^{W^+}/2 = \bar{u} - d - s + \bar{c} \,,
\end{aligned} \tag{15}
$$

where NC and CC denotes neutral or charged current scattering and we have lumped together the contributions coming from Z exchange and from γZ interference, with couplings given by

$$
B_q\left(M_X^2\right) = -2e_q V_\ell V_q P_Z + \left(V_\ell^2 + A_\ell^2\right)\left(V_q^2 + A_q^2\right) P_Z^2 \,, \tag{16}
$$

$$
D_q\left(M_X^2\right) = -2e_q A_\ell A_q P_Z + 4V_\ell A_\ell V_q A_q P_Z^2 \tag{17}
$$

in terms of the electroweak couplings of quarks and leptons listed in Table I and the propagator correction $P_Z = M_X^2/(M_X^2 + M_Z^2)$.

TABLE I

Electroweak couplings of fermions.

Fermions	e_f	V_f	A_f
u, c, t	$+2/3$	$\left(+1/2 - 4/3\sin^2\theta_W\right)$	$+1/2$
d, s, b	$-1/3$	$\left(-1/2 + 2/3\sin^2\theta_W\right)$	$-1/2$
ν_e, ν_μ, ν_τ	0	$+1/2$	$+1/2$
e, μ, τ	-1	$\left(-1/2 + 2\sin^2\theta_W\right)$	$-1/2$

The leading order contribution to Drell–Yan is given by

$$
\begin{aligned}
&\gamma \qquad \frac{d\sigma}{dM_X^2 dy}\left(M_X^2, y\right) = \frac{4\pi\alpha^2}{9M_X^2 s}\sum_i e_i^2 L^{ii}\left(x_1^0, x_2^0\right) \,, \\
&W \qquad \frac{d\sigma}{dy} = \frac{\pi G_{\text{F}} M_V^2\sqrt{2}}{3s}\sum_{i,j} |V_{ij}^{\text{CKM}}| L^{ij}\left(x_1^0, x_2^0\right) \,, \\
&Z \qquad \frac{d\sigma}{dy} = \frac{\pi G_{\text{F}} M_V^2\sqrt{2}}{3s}\sum_i \left(V_i^2 + A_i^2\right) L^{ii}\left(x_1^0, x_2^0\right) \,,
\end{aligned} \tag{18}
$$

in terms of the differential leading order parton luminosity

$$L^{ij}\left(x_1,x_2\right)\equiv q_i\left(x_1,M_X^2\right)\bar{q}_j\left(x_2,M_X^2\right)+q_i(x_2,M_X^2)\bar{q}_j\left(x_1,M_X^2\right) \tag{19}$$

and the CKM matrix elements V_{ij}, with x_i^0 given by Eq. (8). This shows explicitly that, as already mentioned, for a rapidity distribution the leading order parton kinematics (*i.e.* the values of x_i) is completely fixed by the hadronic kinematics (*i.e.* the values of y and M_X^2).

Note that while at a pp collider (or when a p beam collides with a p fixed target) such as the LHC it makes no difference whether the incoming quark and antiquark come from either of the initial-state hadrons, at a $p\bar{p}$ collider such as the Tevatron (or when a p beam collides with a deuterium fixed target) there are two different contributions, according to whether each of the incoming partons is extracted from either of the initial-state hadrons.

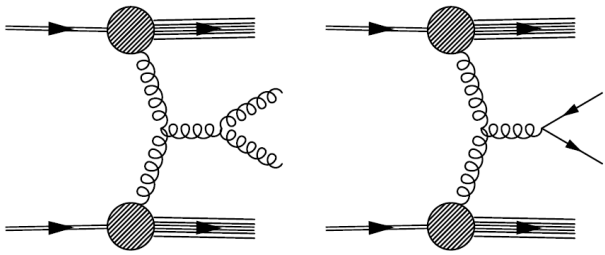

Fig. 4. Leading order diagrams for inclusive jet production from gluons.

Leading-order information on the gluon can be extracted from jet production (see Fig. 4), or from scaling violations, as measured for instance by the Q^2 dependence of deep-inelastic structure functions. The latter are coupled to the gluon even at leading order through the singlet QCD evolution equations, which in terms of Mellin moments

$$f_i\left(N,Q^2\right)\equiv\int_0^1 dx\,x^{N-1}f_i\left(x,Q^2\right) \tag{20}$$

of parton distributions take the form

$$\begin{aligned}\frac{d}{dt}\begin{pmatrix}\Sigma\left(N,Q^2\right)\\ g\left(N,Q^2\right)\end{pmatrix} &= \frac{\alpha_{\rm s}(t)}{2\pi}\begin{pmatrix}\gamma_{qq}^{\rm S}\left(N,\alpha_{\rm s}(t)\right) & 2n_f\gamma_{qg}^{\rm S}\left(N,\alpha_{\rm s}(t)\right)\\ \gamma_{gq}^{\rm S}\left(N,\alpha_{\rm s}(t)\right) & \gamma_{gg}^{\rm S}\left(N,\alpha_{\rm s}(t)\right)\end{pmatrix}\\ &\quad\times\begin{pmatrix}\Sigma\left(N,Q^2\right)\\ g\left(N,Q^2\right)\end{pmatrix},\end{aligned} \tag{21}$$

$$\frac{d}{dt}\,q_{ij}^{\rm NS}\left(N,Q^2\right) = \frac{\alpha_{\rm s}(t)}{2\pi}\,\gamma_{ij}^{\rm NS}(N,\alpha_{\rm s}(t))\,q_{ij}^{\rm NS}\left(N,Q^2\right), \tag{22}$$

where the singlet combination of quark distributions is defined as

$$\Sigma\left(x,Q^2\right) \equiv \sum_{i=1}^{n_f}\left(q_i\left(x,Q^2\right)+\bar{q}_i\left(x,Q^2\right)\right) , \tag{23}$$

and the remaining nonsinglet combinations can be taken as any linearly independent set of $2n_f - 1$ differences of quark and antiquark distributions, $q_{ij}^{\rm NS}(N,Q^2) = q_i^{\rm NS}(N,Q^2) - q_j^{\rm NS}(N,Q^2)$ which all evolve according to individual, decoupled equations.

The leading order anomalous dimensions are shown in Fig. 5, while at leading order all nonsinglet $\gamma_{ij}^{\rm NS}$ are equal to each other and are also equal to γ_{qq}. The qualitative behaviour of perturbative evolution is then deduced recalling that Mellin transformation maps the large (small) $x \to 1$ ($x \to 0$) region into the large (small) $N \to \infty$ ($N \to 0$) region. A first relevant feature is that as the scale increases all PDFs decrease at large x and increase at small x. A second important feature is that because the gluon has the rightmost singularity at small N it drives small x scaling violations, and

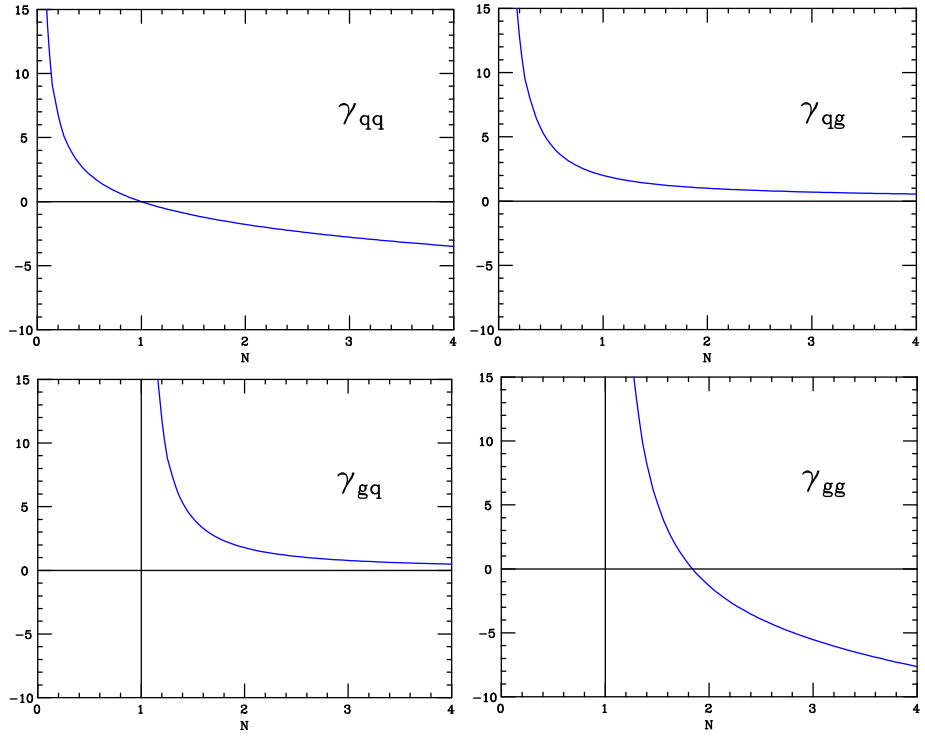

Fig. 5. The matrix of leading order anomalous dimensions shown as a function of the Mellin variable N (Eq. (20)).

thus in particular at sufficiently small x and large Q^2 all PDFs have the same shape, driven by the gluon. Finally, the evolution of the gluon (driven by γ_{gg}) is strongest at either large or small x but its coupling to the quark (driven by γ_{qg}) is only large at small x, so it is only at not too large x that scaling violations provide leading constraints on the gluon.

Finally, it is important to note that constraints on PDFs come from their cross-talk imposed by sum rules: specifically the conservation of baryon number

$$\int_0^1 dx \left(u^p\left(x,Q^2\right) - \bar{u}^p\left(x,Q^2\right)\right) = 2 = 2\int_0^1 dx \left(d^p\left(x,Q^2\right) - \bar{d}^p\left(x,Q^2\right)\right) \quad (24)$$

and the conservation of total energy-momentum

$$\int_0^1 dx x \left[\sum_{i=1}^{N_f}\left(q^i\left(x,Q^2\right) + \bar{q}_i\left(x,Q^2\right)\right) + g\left(x,Q^2\right)\right] = 1\,. \quad (25)$$

Clearly, these sum rules provide constraints on the behaviour of parton distributions even in the region where there are no data.

3. Statistics

A determination of parton distributions is a determination of at least seven independent functions: three light quark and antiquark distributions and the gluon at some initial scale, from which PDFs at all other scales can be obtained solving evolution equations. More functions must be determined if one wishes to keep open the possibility [19] that heavy quarks PDFs are at least in part of "intrinsic" nonperturbative origin, rather than being determined radiatively from gluons by QCD evolution. A determination of PDFs with uncertainties thus involves determining a probability distribution in a space of several independent functions. Because experimental data used for this determination will always be finite in number, this is in principle an ill-posed (or unsolvable) problem.

The time-honoured [20, 21] method to make this problem tractable is to assume a specific functional form for parton distributions, which projects the infinite dimensional problem onto a finite-dimensional parameter space. This method is justified because PDFs are expected to be smooth functions of the scaling variable x. Because $0 \leq x \leq 1$, a representation of these functions with finite accuracy must be possible on a finite basis of functions: hence, a representation of PDFs must be possible in terms of a finite number of parameters. The problem is then reduced to the choice of an optimal

parameterisation, namely, one that for given accuracy minimises the number of parameters without introducing a bias. We will discuss below two such parameterisations.

Whatever the parameterisation, determining a set of PDFs involves computing a number of physical processes with a given set of input PDFs, and extremising a suitable figure of merit, such as a χ^2 or likelihood function in order to determine a best-fit set of PDFs. Existing sets of parton distributions which are made available for the computation of LHC processes through standard interfaces are determined and delivered following two main strategies: a "Hessian" approach, in which the best-fit result is given in the form of an optimal set of parameters and an error matrix centered on this optimal fit to compute uncertainties, and a Monte Carlo approach, in which the best fit is determined from the Monte Carlo sample by averaging and uncertainties are obtained as variances of the sample.

It turns out that available Hessian PDF sets are mostly based on a "standard" parameterisation, inspired by various QCD arguments. On the other hand, the only available full Monte Carlo PDF set is based on a rather different form of parameterisation, which adopts neural networks as interpolating functions in an attempt to reduce the bias related to the choice of functional form [22]. However, Monte Carlo studies based on other standard [23] and non-standard [24] parameterisations have also been presented.

Here we will summarise the main features of both the Hessian and Monte Carlo approach, and in each case also discuss the parton parameterisation which is most commonly used with each approach, and the way the best fit is determined in each case — which in turn requires a peculiar algorithm within the neural network approach.

3.1. Hessian uncertainties and the "standard" approach

The standard approach to PDF determination is based on assuming for PDFs at some reference scale Q_0 a functional form inspired by counting rules [25], which suggest that PDFs should behave as $f_i(x) \underset{x\to 1}{\sim} (1-x)^{\beta_i}$, and Regge theory, which suggest [26] that they should behave as $f_i(x) \underset{x\to 0}{\sim} x^{\alpha_i}$. Note that these limiting behaviours are necessarily approximate, because even if they hold at some scale, at any other scale perturbative evolution will correct them by logarithmic terms which behave as $\ln(1-x)$ as $x \to 1$ and as $\ln x$ as $x \to 1$. Therefore, even if counting rules and Regge theory actually provide predictions for the values of the exponents β_i and α_i respectively (for given parton and parent hadron), they are taken as free fit parameters.

Based on this, typically PDFs are assumed to have the form

$$f_i\left(x, Q_0^2\right) = x^{\alpha_i}(1-x)^{\beta_i} g_i(x)\,, \tag{26}$$

where $g_i(x)$ tends to a constant for both $x \to 0$ and $x \to 1$. For instance, the CTEQ/TEA Collaboration assumes generally [27, 28]

$$xf\left(x, Q_0^2\right) = a_0 x^{a_1}(1-x)^{a_2} \exp\left(a_3 x + a_4 x^2 + a_5\sqrt{x} + a_6 x^{-a_7}\right) \tag{27}$$

with different parameters a_i for each PDF, but some parameters fixed or set to zero for some PDFs — for example, parameters a_6 and a_7 are nonzero only for the gluon distribution. Other groups assume that $g_i(x)$ is a polynomial in x or in $\sqrt{x}$: for instance HERAPDF [29] assumes $g_i(x) = 1 + \epsilon_i\sqrt{x} + D_i x + E_i x^2$.

Different choices are possible for the set of linearly independent combinations of PDFs for which the parameterisation Eq. (26) is adopted, and for the total number of free parameters to be used. For instance CTEQ/CT parameterises the "valence" light combinations $u_v = u - \bar{u}$, $d_v = d - \bar{d}$, the antiquark distributions $\bar{u}$ and $\bar{d}$, the two strangeness combinations $s^{\pm} = s \pm \bar{s}$ (but in the CTEQ6.6 [27] and CT10 [28] fits it is assumed that $s - \bar{s} = 0$) and the gluon, with 22 (CTEQ6.6) or 26 (CT10) free parameters; MSTW08 [30] parameterises also u_v, d_v, $s^{\pm}$ and the gluon, and then the two combinations $\bar{u} \pm \bar{d}$ with a total of 28 parameters, and so forth.

Given a parameterisation of PDFs, the problem is reduced to that of determining best fit values and uncertainty ranges for the parameters. In a Hessian approach, this is done by minimising a figure of merit such as

$$\chi^2(\vec{a}\,) = \frac{1}{N_{\rm dat}} \sum_{i,j} \left(d_i - \bar{d}_i(\vec{a}\,)\right) {\rm cov}_{ij} \left(d_j - \bar{d}_j(\vec{a}\,)\right) , \tag{28}$$

where the sum runs over all data points, d_i are experimental data with experimental covariance matrix ${\rm cov}_{ij}$ (including all correlated and uncorrelated statistical and systematic uncertainties), $\bar{d}_i(\vec{a}\,)$ are theoretical predictions which are obtained by evolving the starting PDFs at any scale Q^2 using the evolution equations Eq. (22), and then folding the result with known partonic cross-sections according to the factorisation theorems Eqs. (1), (6), (10), and $\vec{a}$ denotes the full set of parameters on which the PDFs at scale Q_0 depend, which we may view as a vector in parameter space (which is 26-dimensional for CT10, and so on). The χ^2 thus is a function of the $\vec{a}$ through the predictions $\bar{d}_i(\vec{a}\,)$.

Note that the χ^2 Eq. (28) is normalised to the number of data points: this is conventionally done in order to allow for approximate comparisons of fit quality between fits with different numbers of data points; in practice, this is likely to be close to the χ^2 per degree of freedom because typical datasets include thousands of data, while the total number of parameters needed to describe accurately all PDFs with functional forms like Eq. (26), though of

course unknown, is likely to be rather lower than a hundred. For the sake of future discussions it is convenient to also introduce an unnormalised

$$\bar{\chi}^2 = N_{\rm dat}\chi^2 \,. \tag{29}$$

It is important to note that there are subtleties in the definition of the χ^2, which may make the comparison of χ^2 values from different groups only qualitatively significant, because slightly different definitions are used. The main subtlety is related to the inclusion of normalisation uncertainties, which cannot be simply introduced in the covariance matrix, as this would bias the fit [31]: a full unbiased solution [32] requires an iterative construction of the covariance matrix, but other approximate solutions are also adopted.

Once the χ^2 is defined, for given data χ^2 is a function of the PDF parameters through the predictions $\bar{d}_i$ which in turn depend on the PDFs. Hence, the best fit set of parameters can be identified with the absolute minimum of the χ^2 in parameter space. Furthermore, the variance of any observable X which depends on parameters $\vec{a}$ (such as a physical cross-section, or indeed the PDFs themselves), if we assume linear error propagation $X(\vec{a}\,) \approx X_0 + a_i\partial_i X(\vec{z}\,)$, is given by

$$\sigma_X^2 = \sigma_{ij}\partial_i X \partial_j X \,. \tag{30}$$

Here σ_{ij} is the covariance matrix of the parameters which, in turn, assuming that the χ^2 is a quadratic function of the parameters in the vicinity of the minimum, is given by (see *e.g.* [33, 34])

$$\sigma_{ij} = \partial_i\partial_j\bar{\chi}^2|_{\rm min} \tag{31}$$

i.e. it is the (Hessian) matrix of second derivatives of the unnormalised $\bar{\chi}^2$ Eq. (29), evaluated at its minimum.

The Hessian method for the determination of uncertainties thus in particular implies that the one-σ (*i.e.* 68% confidence level) for the parameters themselves is the ellipsoid in parameter space which is fixed by the condition $\bar{\chi}^2 = \bar{\chi}^2_{\rm min} + 1$. As we will discuss in Sec. 5.1, in practice this argument may have to be modified in realistic cases, in order to account for various effects (such as incorrect estimation of the covariance matrix of the data).

However, for the time being let us stick to the textbook argument, and make a couple of observations on it. The first observation is that we are always free to adjust the parameterisation in such a way that all eigenvalues of the Hessian matrix σ_{ij} are equal to one, by simply diagonalising the matrix and rescaling the eigenvectors by the eigenvalues, *i.e.* by looking for new parameters $a'_j(a_i)$ such that

$$\sigma_{ij}\left(a_i - a_i^{\rm min}\right)\left(a_j - a_j^{\rm min}\right) = \sum_{i=1}^{N_{\rm par}}\left(a'_i\right)^2 \tag{32}$$

which immediately implies that

$$\sigma_X^2 = \left|\vec{\nabla}' X\right|^2 , \tag{33}$$

where the gradient is computed with respect to $\vec{a}\,'$. Equation (33) has the immediate interesting consequence that the total contribution to the uncertainty due to two different sources, being the length of a vector, is simply found by adding the components *i.e.* the different uncertainties in quadrature (even when the two uncertainties are correlated). This has been emphasised recently in Ref. [35], where it is shown explicitly that, contrary to what one may naively think, the total uncertainty due to PDF parameters and some other parameter (such as the value of the strong coupling constant) is simply found adding the two uncertainties in quadrature.

The second comment has to do with the fact that the one-σ interval in parameter space corresponds to the contour $\Delta N_{\rm dat}\chi^2 = 1$ about the minimum. This is identical to the statement that the Hessian Eq. (31) is the covariance matrix in parameter space. This simple fact is sometimes a source of confusion because it seems to contradict the observation that the standard deviation of the (unnormalised) χ^2 distribution with $N_{\rm dof}$ degrees of freedom is $\sqrt{2N_{\rm dof}}$: in fact, sometimes (see *e.g.* [36]) it is incorrectly stated that one-σ contours correspond to $\Delta\bar{\chi}^2 \sim N_{\rm dof}$. However, the contradiction is only apparent: $\Delta\bar{\chi}^2 \sim N_{\rm dof}$ sets a hypothesis-testing criterion [37], namely, it gives the size of fluctuations of $\Delta\bar{\chi}^2$ upon repetition of the experiment, and thus the range of $\bar{\chi}^2$ values away from the mean $\langle\bar{\chi}^2\rangle = N_{\rm dat}$ which are acceptable for a given theory (experiment). On the other hand, $\Delta\bar{\chi}^2 = 1$ provides a parameter-fitting criterion [37]: it gives the range of parameter values which are compatible at one sigma for a given experimental result (and theory).

A simple example may help in understanding the distinction. Consider the case of a simple linear fit, in which one has a set of data which are expected to satisfy a linear law $y = x + k$, with unknown intercept k that one wishes to determine by fitting to data (see Fig. 6). Define the deviation $\Delta_i \equiv d_i - \bar{d}_i(x_i)$ between the ith data point d_i and the linear prediction $\bar{d}_i = x_i + k$. If Δ_i are Gaussianly distributed with standard deviation σ about their true values, then clearly the average square deviation $\sigma_\Delta^2 = \langle\Delta^2\rangle = N_{\rm dat}\sigma^2$. This is the "hypothesis testing" fluctuation range of the $\bar{\chi}^2$. However, the best-fit intercept k is just the average deviation $k = \langle\Delta\rangle$, and the square uncertainty on it is $\sigma_k^2 = \frac{\sigma_\Delta^2}{N_{\rm dat}}$: so the "parameter fitting" range for k is indeed by a factor $N_{\rm dat}$ smaller than the expected total square fluctuation, because the best-fit value is determined as a mean, whose square fluctuation is by a factor $N_{\rm dat}$ smaller than the fluctuation of the individual data.

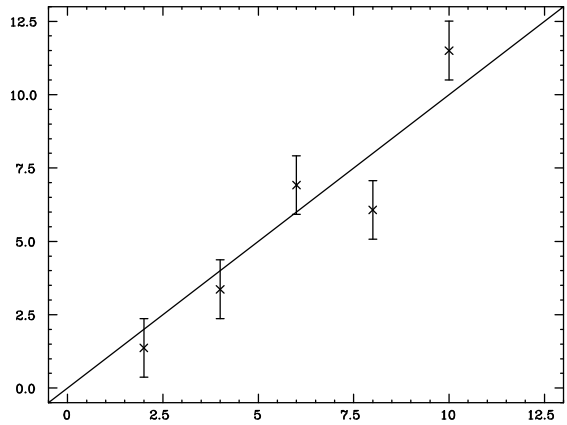

Fig. 6. Fit to data Gaussianly distributed about a linear law.

3.2. Monte Carlo uncertainties and the NNPDF approach

A Monte Carlo approach differs from the Hessian approach in the way the uncertainty on the observable is determined in terms of the uncertainties in parameter space: the distinction Hessian *versus* Monte Carlo thus has to do only with the way uncertainties are propagated from parameters to observables. However, the Monte Carlo way of propagating uncertainties is especially convenient when used together with a parameterisation whose functional form is less manageable, for instance because the number of parameters is particularly large, or because the functional form is less simple than that of Eq. (26), or, more, in general, whenever linear error propagation and the quadratic approximation to the χ^2 in parameter space are not advisable, for reasons of principle or of practice. Therefore, we will first discuss the distinction between Hessian and Monte Carlo *per se*, then turn a brief review of the way a Monte Carlo approach has been used by the NNPDF group together with a choice of basic underlying functional form for PDFs which differs from that of Eq. (26), and finally addresses some issues related to the determination of the best fit PDFs when such functional forms are adopted.

3.2.1. Monte Carlo uncertainties

Whereas in a Hessian approach parameters are assumed to be Gaussianly distributed with covariance matrix σ_{ij} given by the Hessian Eq. (31), in a Monte Carlo approach the probability distribution in parameter space is given by assigning a Monte Carlo sample of replicas of the total parameter set. For example, if one uses the parameterisation Eq. (26) one would then simply give a list of $N_{\rm rep}$ replica copies of the vector of parameters $\vec{s}$. Any

observable X is then computed by repeating its determination $N_{\rm rep}$ times, each time using a different parameter replica: the central value for X is the average of these $N_{\rm rep}$ results, the standard deviation is the variance, and in fact any moment of the probability distribution can be determined from the sample of $N_{\rm rep}$ values of X thus obtained.

Of course, this begs the question of how the distribution of parameter values, *i.e.* the distribution of parameter replicas is determined in the first place. In fact, this may look like a hopeless task: let us say that for each parameter the probability distribution in parameter space is given for each parameter as a histogram with three bins, one corresponding to the one-σ region about the central value of the given parameter, and two for the two outer regions. Then, for $N_{\rm par}$ parameters the total number of bins is equal to $3^{N_{\rm par}} \gtrsim 3 \times 10^9$ with $N_{\rm par} = 20$ parameters. Hence it looks like the total number of replicas must be hopelessly large in order to have sufficient statistics. This, however, is not necessarily the case, because it may well turn out that most of the bins are actually empty. To understand this, recall the Hessian computation of the uncertainty on X Eq. (33): it is clear that in order to determine the uncertainty on X, it is sufficient to know the distribution in parameter space along the direction of $\vec{\nabla}' X$. Hence, for this specific observable only one parameter is relevant. Even if one wants to determine the uncertainty on observables which probe any direction in parameter space, for any reasonably smooth function the number of bins which is needed in order to get an accurate representation of the probability distribution is likely to be much smaller than 10^9. This then again raises the question of how one should sample the replica distribution in parameter space.

The answer is found by noting that the maximum likelihood method gives a way of mapping the probability distribution in data space onto the probability distribution in parameter space. Namely, assume one has data d_i with covariance matrix cov_{ij}. Then, generate $N_{\rm rep}$ *data* replicas d_i^α, with $\alpha = 1, 2, \ldots, N_{\rm rep}$. For each value of α, *i.e.* for each replica, the whole set of data $i = 1, 2, \ldots N_{\rm dat}$ is replicated, in such a way that if one takes the average over the $N_{\rm rep}$ replicas d_n^α of the nth data point, then in the limit $N_{\rm rep} \to \infty$ this average tends to the original data value d_n; if one computes the variance of these $N_{\rm rep}$ values in the same limit it tends to the standard deviation of the data; and if one computes the covariance of the nth and mth data replicas it tends to the covariance matrix element cov_{nm}. Now, for each data replica, determine a best-fit parameter vector $\vec{a}^{\,\alpha}$ by minimising the χ^2 Eq. (28), but of the fit to the replica data d_i^α, rather than the original data. We end up with a Monte Carlo set of best-fit parameter vectors $\vec{a}^{\,\alpha}$: again, the average over these $\alpha + 1, 2, \ldots, N_{\rm rep}$ vectors $\vec{a}^{\,\alpha}$ gives us the best-fit parameters $\vec{a}^{\,\alpha}$, and the covariance of the nth and mth components of the

parameter vector gives us the covariance matrix σ_{mn}. In fact, it is easy to check (see *e.g.* [32]) that for Gaussianly distributed data the results coincides with the Hessian covariance matrix Eq. (31).

The procedure is summarised in Fig. 7: one starts with experimental data (denoted as F_i in the figure), generates data replicas (denoted as $F_i(1)\ldots F_i(N)$) and fits a set of PDFs to each data replica (denoted as $q_0(i)$). The PDFs can be parameterised in any desired way at some reference scale, and they are fitted to the data replicas in the way discussed in Sec. 3.1, namely by evolving them to the scale of the data, using them to compute observables, and minimising the χ^2 of the comparison to the data with respect to the parameters.

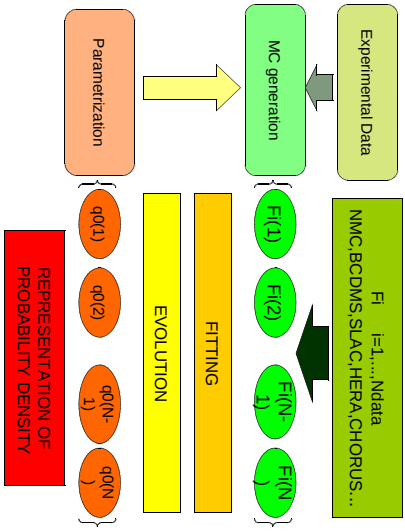

Fig. 7. Schematic representation of the construction of a Monte Carlo representation of parton distributions.

But then, the problem of constructing an adequate sampling of parameter space has been reduced to that of constructing an adequate Monte Carlo representation of the original data: *i.e.* the space of parameters is sampled in a way which is determined by the distribution of the data ("importance sampling"). Whether a given set of replicas provides an accurate enough representation of the data is then something that may be checked explicitly for a given sample, by comparing means, variances and covariances from the sample with the desired features of the data. For a typical set of data used in a parton fit the numbers of replicas required turn out to be surprisingly small: for instance, in Fig. 8 we show a scatter plot of the averages *versus* central values and variances *versus* standard deviations for the set of $N_{\rm dat} = 3372$ data points included in the NNPDF1.2 [38] parton fit, com-

puted using $N_{\rm rep} = 10$, 100, 1000 Monte Carlo Replicas. It is clear that the scatter plot deviates by just a few percent from a straight line already for $N_{\rm rep} = 10$ for central values, and for $N_{\rm rep} = 100$; $N_{\rm rep} = 1000$ replicas turn out to be only necessary in order to get percent accuracy on correlation coefficients.

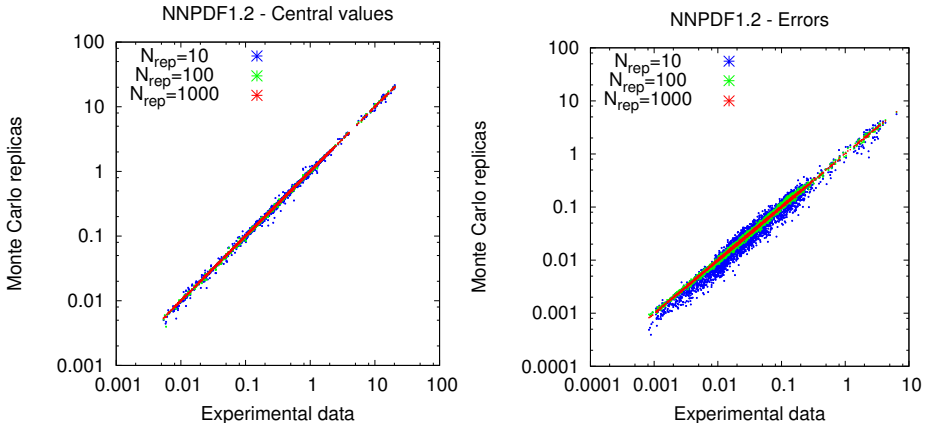

Fig. 8. Scatter plot of central values and uncertainties of a Monte Carlo sample compared to the data for the dataset of the NNPDF1.2 parton fit [38].

It should finally be mentioned that within a Monte Carlo approach it is possible to sidestep the problem of choosing an adequate parameterisation by using Bayesian inference [39]. Namely, one starts from some prior Monte Carlo representation of the probability distribution based on some initial subset of data, or even on assumptions. Then, the initial Monte Carlo set is updated by including the information contained in new data through Bayes' theorem. Without entering into details, it is clear that this can be done by changing the distribution of replicas: more or less copies are taken for those replicas which agree or respectively do not agree with the new data, in a way which is specified by Bayes' theorem. To the extent that results do not depend too much on the choice of prior, which is often the case if the information used through Bayes' theorem is sufficiently abundant, final results are then free of bias. Whereas the construction of a parton set fully based on this method has so far not been completed, preliminary results have been presented [40] on the inclusion of new data in an existing Monte Carlo fit using this methodology.

3.2.2. Neural network parameterisation and cross-validation

The Monte Carlo approach has been recently used for the determination of a PDF set in conjunction with the use of neural networks as a parton parameterisation. Neural networks are just another functional form. In

analogy to polynomial forms, they have the feature that any function (with suitable assumptions of continuity) may be fitted in the limit of infinite number of parameters; unlike polynomial forms they are nonlinear, and they are "unbiased" in that a finite-dimensional truncation of the neural network parameterisation is adequate to fit a very wide class of functions (for instance, both periodic and non periodic) without the need to adjust the form of the parameterisation to the desired problem.

A very simple example of neural network is the function

$$f(x) = \frac{1}{1 + e^{\theta_1^{(3)} - \frac{\omega_{11}^{(2)}}{1+e^{\theta_1^{(2)} - x\omega_{11}^{(1)}}} - \frac{\omega_{12}^{(2)}}{1+e^{\theta_2^{(2)} - x\omega_{21}^{(1)}}}}}, \tag{34}$$

where $\theta_n^{(i)}$ and $\omega_{nm}^{(i)}$ are free parameters. This is a 1-2-1 neural network, parameterised by six free parameters: 1-2-1 refers to the way the neural network is constructed, by iterating recursively the response function $g(x) = \frac{1}{1+e^{\theta-\beta x}}$ on nodes arranged in layers which feed forward to the next layer, with the first (last) layer containing the input (output) variables.

In Refs. [38–41] PDFs are parameterised using 2-5-3-1 neural networks, with 37 free parameters (the input has two variables because x and $\ln x$ are treated as two independent inputs, thereby increasing the redundancy of the parameterisation). The six light flavours and antiflavours are parameterised and the gluon are parameterised in this way, so that the total number of parameters is $37 \times 7 = 259$, thus rather larger than the typical numbers used when dealing with parameterisations of the form Eq. (26). Such a large number of parameters clearly reduces considerably the risk of a parameterisation bias, but it poses the problem that if the best fit is determined as the absolute minimum of the χ^2 one may end up fitting data fluctuations, which is clearly not desirable. Even if these fluctuations average out when averaging over Monte Carlo replicas this would be a very inefficient way of proceeding.

The advantage of a neural network parameterisation can be understood from Fig. 9, where a gluon distribution determined using neural networks is compared to the simplest version of parameterisation Eq. (26), and also to a very flexible parameterisation based on orthogonal polynomials. The neural network gluon distribution shown in Fig. 9 corresponds to $N_{\rm rep} = 25$ replicas from the Monte Carlo set of Ref. [41], and it is displayed along with the average and one-σ contour computed from the set. On the same plot a parameterisation of the form Eq. (26) is also shown, with typical values of the parameters α and β, and with $g(x) = 1$. It is compared to a set of Monte Carlo replicas of the gluon which were constructed in Ref. [24] by expanding the gluon distribution on a basis of 15 independent Chebyshev polynomials,

while also imposing an increasing penalty p to fits with large arc-length (and thus more oscillations). The fits based on orthogonal polynomials display large uncontrolled oscillations which are only tamed by appropriately tuning the length penalty. The fits based on neural networks, despite having a number of free parameters which is more than double than those using orthogonal polynomials, do not display a similarly unstable behaviour, even though they do show considerable flexibility, and in fact the ensuing one-σ band, though accounting well with its width for the functional freedom is actually quite stable.

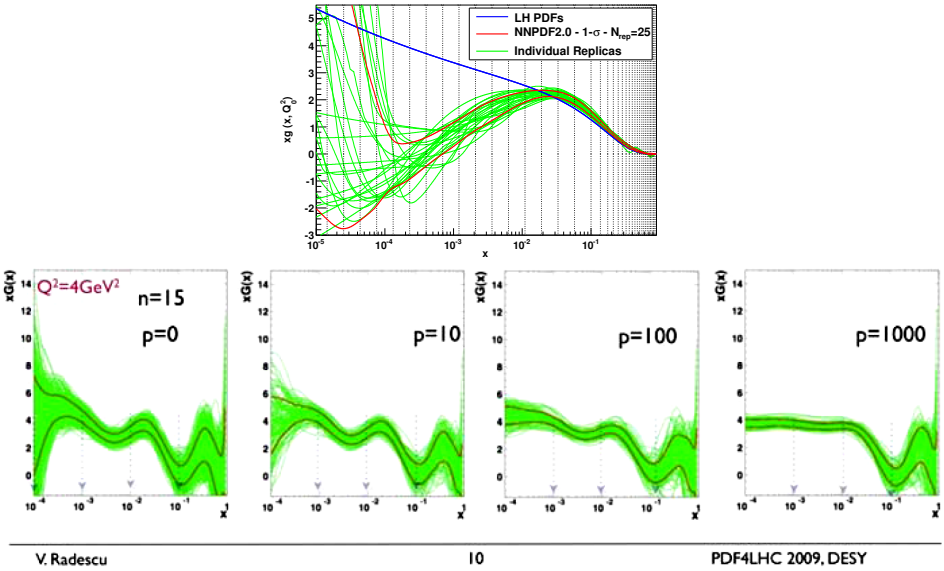

Fig. 9. Upper: 25 gluon replicas based on a neural network parameterisation, together with their one-σ range, compared to a toy gluon distribution of the form Eq. (26) with $g(x) = 1$ and typical values of the parameters α and β (from Ref. [41]). Lower: gluon replicas based on a parameterisation on a basis of Chebyshev polynomials together with their one-σ range; subsequent plots correspond to an increasingly high penalty proportional to the length of the fitted curve (from Ref. [24]).

The best fit is instead determined using a cross-validation method (see Fig. 10). Namely, the data are randomly divided into a training and a validation sample. The χ^2 is computed both for the data in the training sample and those in the validation sample. Only the training χ^2 is minimised, but the validation χ^2 is also monitored as the minimisation proceeds. The best fit is defined as the point at which the validation χ^2 stops improving even though the training χ^2 may keep improving: this is the point at which one is starting to fit the statistical noise of the training sample. In order to ensure

a lack of bias, the partitioning of the data is done randomly in a different way for each data replica. Also, in practice, in order to minimise the effect of random fluctuations in the data (or of the minimisation algorithm) the stopping criterion must be imposed after a suitable averaging, such as for instance the moving average of values of the χ^2 found in the last N_s iterations of the minimisation algorithm.

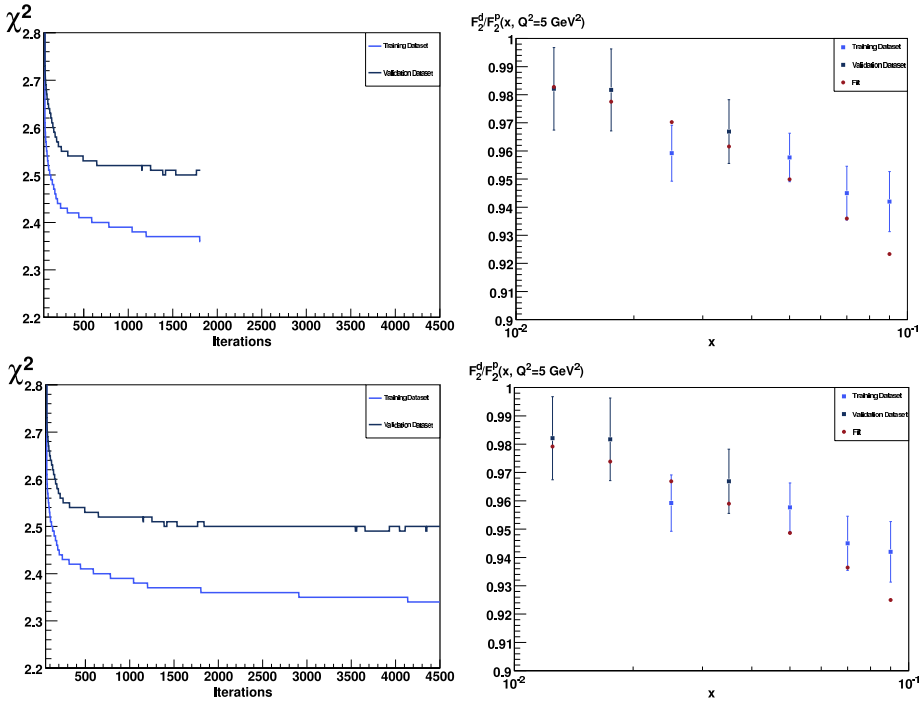

Fig. 10. Determination of the best-fit by cross-validation: the χ^2 of the fit to points in the training set (upper curve) and the validation set (lower curve) are shown as a function of the number of iterations of the minimization algorithm (left plots), but only the χ^2 for training data is minimized. The data are shown in the right plot, along with the best fit: the validation set (dark) includes the 1st, 2nd and 4th point, the training set (light) the other points, while the best fit points are those without error band. The upper plots show the best fit at stopping point (optimal fit) and the χ^2 up to this point, the lower plots show the χ^2 up to and the best fit at an "overlearning" point.

4. From data to PDFs

Once a parton parameterisation and a methodology have been chosen, the determination of a PDF set relies on the choice of a set of physical observables. The problem is that, even after projecting the problem on a

finite-dimensional parameter space, we must still determine seven independent PDFs, which means that we need seven linearly independent pieces of information at fixed scale for each value of x. For instance, a determination of deep-inelastic structure functions F_1 and F_3 for charged-current deep-inelastic scattering provides, according to Eq. (15) four independent linear combinations of quark distributions (if $W^{\pm}$ can be distinguished), with two more linear combinations provided by neutral current structure functions. All individual light quark and antiquark flavours then can be determined by linear combination. This situation would be realistic at a neutrino factory with both neutrino and antineutrino beams and the possibility of identifying the charge of the final state lepton on an event-by-event basis [42, 43].

Unfortunately, this theoretically and phenomenologically very clean option is at best far in the future, so at present the information on individual PDFs can only be achieved by combining information from different processes into so-called "global" fits. The idea is that, even though each electroproduction or hadroproduction observable depends on all PDFs through the factorisation formulae Eqs. (1), (10), inclusion of specific processes or combination of processes may give a specific handle on individual PDFs or combinations of PDFs on which it depends most strongly (typically, through its leading order form). We will now review one at a time each of these individual handles on PDF, then briefly discuss how they are combined in modern more or less global fits.

4.1. Isospin singlet and triplet

Neutral current deep-inelastic (DIS) structure function data only provide a determination of the charge-conjugation even combination $q_i + \bar{q}_i$ of quarks and antiquarks, for each quark flavour i. Specifically, photon DIS data only determine the fixed combination in which each flavour is weighted by the square of the electric charge, see Eq. (15). However, one may separate off the isospin triplet and singlet components by considering DIS on both proton and deuteron targets, assuming that the deuterium structure function is simply the incoherent sum of the proton and neutron ones $F_2^d = \frac{1}{2}(F_2^p + F_2^n)$ (up to small nuclear corrections which can be accounted for through models, such as that of Ref. [44]), and then using isospin symmetry to relate the quark and antiquark distributions of the proton and neutron

$$u^p\left(x,Q^2\right) = d^n\left(x,Q^2\right) , \qquad d^p\left(x,Q^2\right) = u^n\left(x,Q^2\right) . \tag{35}$$

One then has

$$F_2^p\left(x,Q^2\right) - F_2^d\left(x,Q^2\right) = \tfrac{1}{3}\Big[\left(u^p + \bar{u}^p\right) - \left(d^p + \bar{d}^p\right)\Big]\left[1 + O(\alpha_{\rm s})\right] \tag{36}$$

so that the difference of proton and deuteron structure functions provides a leading-order handle on the isospin triplet combination

$$T_3\left(x,Q^2\right) \equiv u\left(x,Q^2\right) + \bar{u}\left(x,Q^2\right) - \left[d\left(x,Q^2\right) + \bar{d}\left(x,Q^2\right)\right]. \tag{37}$$

Note that even beyond leading order $F_2^p - F_2^d$ only depends on T_3, which can thus be determined without further assumptions: a theoretically very clean, though necessarily not especially accurate determination [45].

4.2. Light quarks and antiquarks

Modern DIS data are available over a wide range of values of Q^2, extending well into the region where the CC contributions are sizable: in fact HERA-I data are available both for CC and NC scattering, both with electron and positron beams. Unfortunately, collider data only provide a fixed combination Eq. (11) of the structure functions F_1 and F_3, because for given x and Q^2 Eq. (11) implies that y can be varied only by changing the center-of-mass energy of the hadron–lepton collision. Hence, HERA data only provide three independent combinations of structure functions and thus of parton distributions (NC and CC with positively or negatively charged leptons). However, a fourth combination is provided because the Q^2 dependence of the γ^* and Z contributions to NC scattering is different (see Eq. (17). It follows that the very precise HERA data can determine four independent linear combinations of PDFs, which can be chosen as the two lightest flavours and antiflavours, with strangeness then determined by assumption.

Even without a neutrino factory, data on neutrino deep-inelastic scattering are available, but typically on approximately isoscalar nuclear targets. Because the energy of the neutrino beam typically has a (more or less broad) spectrum, the value of y Eq. (14) is not fixed, and the contributions of F_1 and F_3 to the cross-section can be disentangled. On an isoscalar target at leading order

$$\begin{aligned}
F_2^{\nu} &= x\left(u + \bar{u} + d + \bar{d} + 2s + 2\bar{c}\right) + O(\alpha_{\mathrm{s}}), \\
F_2^{\bar{\nu}} &= x\left(u + \bar{u} + d + \bar{d} + 2\bar{s} + 2c\right) + O(\alpha_{\mathrm{s}}), \\
F_3^{\nu} &= u - \bar{u} + d - \bar{d} + 2s - 2\bar{c} + O(\alpha_{\mathrm{s}}), \\
F_3^{\bar{\nu}} &= u - \bar{u} + d - \bar{d} - 2\bar{s} + 2c + O(\alpha_{\mathrm{s}})
\end{aligned} \tag{38}$$

so neutrino data provide an accurate handle on the total valence component

$$V\left(x,Q^2\right) = \sum_{i=}^{n_f} \left(q\left(x,Q^2\right) - \bar{q}_i\left(x,Q^2\right)\right). \tag{39}$$

A more direct determination of the light flavour decomposition can be obtained by exploiting the fact that the Drell–Yan cross-section probes various parton combinations, which can be selected by looking at different final states. In particular one can notice [46] that for neutral-current Drell–Yan if both data on proton and neutron (or deuteron) targets are available, using isospin Eq. (35) one gets at leading order

$$\frac{\sigma^{pn}}{\sigma^{pp}} \sim \frac{\frac{4}{9}u^p\bar{d}^p + \frac{1}{9}d^p\bar{u}^p}{\frac{4}{9}u^p\bar{u}^p + \frac{1}{9}d^p\bar{d}^p} + O(\alpha_{\rm s}) + \text{heavier quarks}\,, \tag{40}$$

where we have omitted the dependence on the kinematic variables, which at leading order is as in Eq. (18). As discussed there, if the rapidity distribution is measured, the leading order partonic kinematic is completely fixed: for given y and Q^2 only partons with x_1, x_2 given by Eq. (8) contribute. Here "heavier quarks" denote strange and heavier flavours, which give a smaller contribution at least in the region of $x \gtrsim 0.1$ in which most of the contribution to the sum rule integrals Eq. (25) is concentrated.

In particular, because of the sum rule Eq. (24), in the the region which gives the dominant contribution to the integral (the "valence" region $x \gtrsim 0.1$) the up distribution is roughly twice as large as the down distribution (assuming $\bar{u} \sim \bar{d}$) so the first term in both the numerator and the denominator of Eq. (40) gives the dominant contribution, and the ratio reduces to $\frac{\sigma^{pn}}{\sigma^{pp}} \approx \frac{\bar{d}^p}{\bar{u}^p}$. Hence this particular combination of cross-sections provides a sensitive probe of the $\bar{u}/\bar{d}$ ratio: indeed, it has been used to provide first evidence that this ratio, though of order one, deviates from unity [47, 48].

In the charged current case, one may exploit the fact that using charge-conjugation symmetry to relate the p and $\bar{p}$ PDFs

$$q_i^p = \bar{q}_i^{\bar{p}} \tag{41}$$

at leading order one gets

$$\begin{aligned}\frac{\sigma^{p\bar{p}}_{W^+}}{\sigma^{p\bar{p}}_{W^-}} \;=\; & \frac{u^p(x_1)d^p(x_2) + \bar{d}^p(x_1)\bar{u}^p(x_2)}{d^p(x_1)u^p(x_2) + \bar{u}^p(x_1)\bar{d}^p(x_2)} \\ & + O(\alpha_{\rm s}) + \text{Cabibbo suppressed} + \text{heavy quarks}\,,\end{aligned} \tag{42}$$

where heavy quarks denotes charm and heavier flavours. In writing Eq. (42) we have assumed that the cross-sections are differential in rapidity. If the kinematics is chosen in such a way that x_i are in the "valence" region, in which quark distributions are sizably larger than antiquark ones, the ratio Eq. (42) is mostly sensitive to the light quark ratio u/d [49, 50] and indeed it has been used to provide the first accurate determinations of it [51].

The sizable impact of Drell–Yan data on a PDF fit is demonstrated in Fig. 11, where we compare the value and uncertainty of PDF combinations which are sensitive to the light flavour decomposition before and after inclusion of Drell–Yan data in a DIS fit, namely the total valence Eq. (39) and the light sea asymmetry

$$\Delta_s\left(x,Q^2\right) \equiv \bar{d}(x,Q^2) - \bar{u}\left(x,Q^2\right) \ . \tag{43}$$

The DIS fit includes both the fixed-target proton and neutron data (which thus determine well the isotriplet component and give a handle on the singlet–triplet separation), the precise HERA data (which give a handle on each individual light flavour and antiflavour), and several neutrino data (which determine well the valence component). The Drell–Yan data included contain both proton and deuteron fixed target γ production data, and W production. It is apparent that the accuracy on the valence, which is already quite good in the DIS-only fit, is reduced by a large factor by the inclusion of Drell–Yan data, and the effect is even more impressive on the light antiquark asymmetry which, despite the accuracy of the HERA data, is only determined with large uncertainties by DIS data.

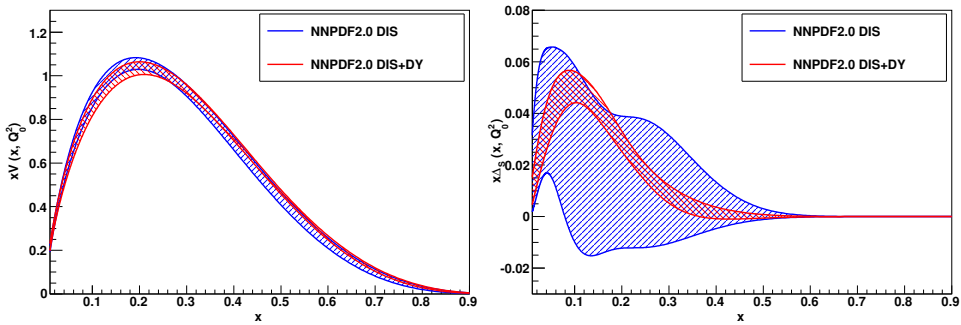

Fig. 11. Impact of Drell–Yan data on the NNPDF2.0 [41] parton fit. Left: the total valence distribution; right: the light antiquark asymmetry.

The strong impact of W and Z production data can be seen quantitatively by computing the correlation coefficient between the W and Z cross-section and individual parton distributions, which can be computed both in a Hessian approach using standard error propagation, or in a Monte Carlo approach from the covariance of the cross-section and the parton distribution over the Monte Carlo sample. Results obtained in the Hessian approach using CTEQ6.6 [27] are shown in Fig. 12: correlations are quite large, even though results shown here are obtained using the total cross-section, which is a much less sensitive probe of PDFs than the rapidity distributions discussed above.

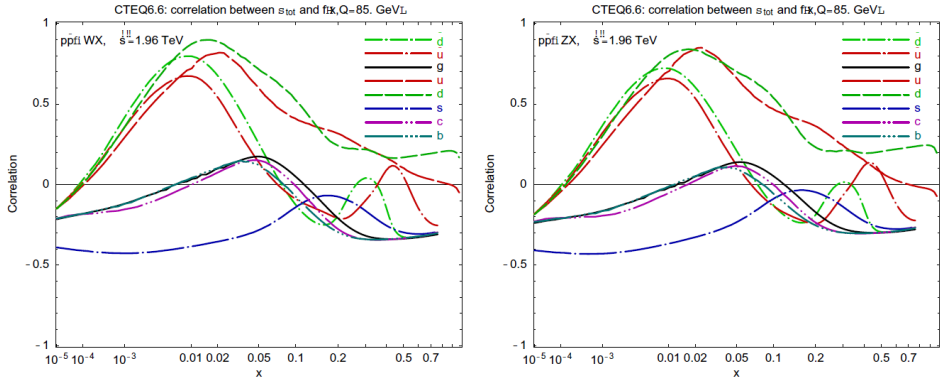

Fig. 12. Correlation between the total W and Z cross-section at the Tevatron and individual parton distributions (from Ref. [27]).

4.3. Strangeness

The determination of strangeness is nontrivial because, of course, it has the same electroweak couplings as the down distribution, while it is typically smaller than it (except at small x where all PDFs are the same size, as discussed in Sec. 2.2). The only way of determining it accurately from deep-inelastic scattering data is to include semi-inclusive information. A simple way of doing this is to use data for neutrino deep-inelastic charm production (known as dimuon production, because charm is tagged by the muon from its decay together with the muon due to the charged current neutrino interaction). At leading order the structure functions are then just

$$\begin{aligned} F_2^{\nu,p,c}\left(x,Q^2\right) &= xF_3^{\nu,p,c}\left(x,Q^2\right) \\ &= 2x\Big(|cd|^2\, d(x) + |V_{cs}|^2\, s(x) + |V_{cb}|^2\, b(x)\Big) + O\left(\alpha_{\rm s}^2\right), \\ F_2^{\bar\nu,p,c}\left(x,Q^2\right) &= -xF_3^{\bar\nu,p,c}\left(x,Q^2\right) \\ &= 2x\Big(|V_{cd}|^2\, \bar d(x) + |V_{cs}|^2\, \bar s(x) + |V_{cb}|^2\, \bar b(x)\Big) + O\left(\alpha_{\rm s}^2\right) \end{aligned} \tag{44}$$

so up to CKM suppressed terms they measure strangeness directly.

In Fig. 13 the behaviour upon inclusion of dimuon data of a fit to a set of DIS data which includes both neutrino and HERA data is shown: it is clear that before inclusion of the dimuon data the fit (NNPDF1.1 [52]) cannot determine either of the two strange combinations

$$s^{\pm}\left(x,Q^2\right) \equiv s^{+}\left(x,Q^2\right) \pm s^{-}\left(x,Q^2\right) \tag{45}$$

but after their inclusion it determines both, though with limited accuracy due to the limited accuracy and kinematic coverage of the available dimuon data. In this plot, we also show the result one obtains for strangeness if one simply assumes it to be proportional to the light quark sea, *i.e.* if by assumption one sets $s^- = 0$ and $s^+ = \frac{1}{2}(\bar{u} + \bar{d})$. This is often done in PDF determinations based on DIS data only: the result is then misleadingly accurate. This comparison should thus be taken as a warning that, when using PDF sets in which some PDFs are fixed by assumption, some uncertainties may be significantly underestimated.

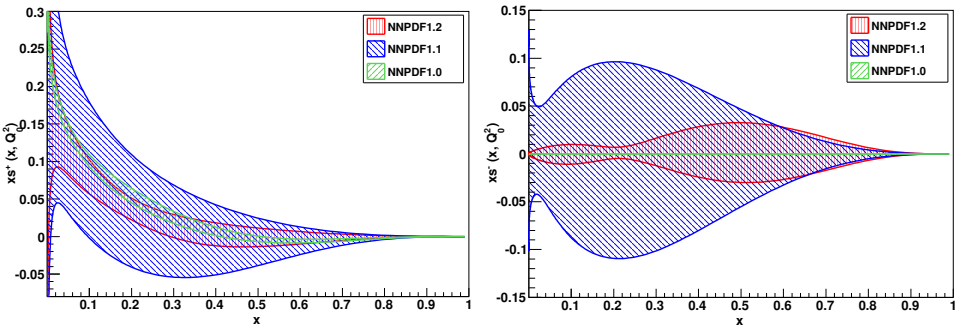

Fig. 13. The strange distributions $s^+ = s + \bar{s}$ (left) and $s^- = s - \bar{s}$ (right) determined in a fit to DIS including dimuon data (NNPDF1.2 [38]) and not including them (NNPDF1.1 [52]). A fit without dimuon data where strangeness is fixed by assumption (NNPDF1.0 [53]) is also shown.

Of course the Drell–Yan data discussed above also constrain strangeness. Specifically, the cross-section ratio Eq. (42) receives a contribution from strange and charm quarks which, up to CKM matrix elements, is identical to the contribution from down and up quarks respectively. Well above charm threshold this contribution is sizable, so comparing Drell–Yan data above and below charm threshold potentially leads to a rather accurate determination of strangeness. Indeed, in Fig. 14 we show the impact of including Drell–Yan data in a fit with DIS data only (same pair of fits already shown in Fig. 11). The DIS dataset contains dimuon data, and it is similar to the dataset on which the fit of Fig. 13 is based, from which it mostly differs because of improvements in the HERA data and in fit methodology; however, the Drell–Yan data have a visible impact on the total strangeness s^+, and lead to a very striking improvement in the determination of the strangeness asymmetry s^-.

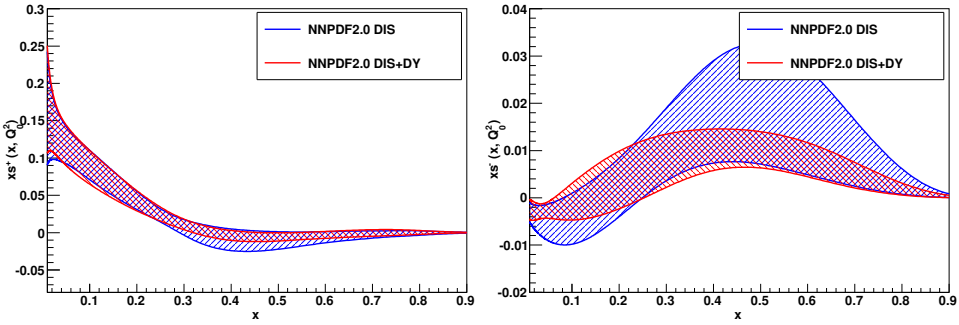

Fig. 14. Impact of Drell–Yan data on the strange distributions $s^+ = s + \bar{s}$ (left) and $s^- = s - \bar{s}$ (right) using the NNPDF2.0 [41] parton fit.

4.4. Gluons

The determination of the gluon distribution is nontrivial because the gluon does not couple to electroweak final states. It does, however, mix at leading order through perturbative evolution: so, even using parton-model (*i.e.* $\mathrm{O}(\alpha_{\mathrm{s}}^0)$) expressions for cross-sections and structure functions, the gluon does determine their scale dependence. Indeed

$$\frac{d}{dt}F_2^{\mathrm{s}}\left(N,Q^2\right) = \frac{\alpha_{\mathrm{s}}\left(Q^2\right)}{2\pi}\left[\gamma_{qq}(N)F_2^{\mathrm{s}}+2\,n_f\gamma_{qg}(N)g\left(N,Q^2\right)\right]+O\left(\alpha_{\mathrm{s}}^2\right)\ , \quad (46)$$

where by $F_2(N,Q^2)$ we denote the Mellin moments, Eq. (20), of the singlet component (defined as in Eq. (23)) of the F_2 structure function.

It follows that the gluon is mostly determined by scaling violations, or by its coupling to strongly-interacting final-states, *i.e.* jets. The main shortcoming of the determination from scaling violations is that, as already pointed out in Sec. 2.2, the gluon only couples strongly to other PDFs for sufficiently small x: for instance, Fig. 5 shows clearly that for $N > 2$ the γ_{qg} term rapidly becomes negligible in comparison to the γ_{gg} term. On the other hand, the gluon distribution is expected to be quite small at large x, and, as also discussed in Sec. 2.2, to further shift towards smaller x as the scale increases. Hence, the large x gluon is likely to be small and affected by large uncertainties, which can only be reduced by looking at hadronic (jet) final states.

Indeed, in Fig. 15 we show the effect of the inclusion of jet data in a PDF fit based on DIS data. At small x there is essentially no effect: scaling violations are sufficient to determine the gluon quite accurately. At large x, even though the determination of the gluon from scaling violations is reasonably accurate, its accuracy is still quite significantly improved by the inclusion of jet data. A feature of this plot which is worth noting is

the beautiful consistency of these two determinations. This is an extremely strong consistency check for the perturbative QCD framework: the gluon determined from scaling violation and evolved up to the much higher jet scale is in perfect agreement with the jet data, and indeed the best accuracy is obtained combining the two determinations.

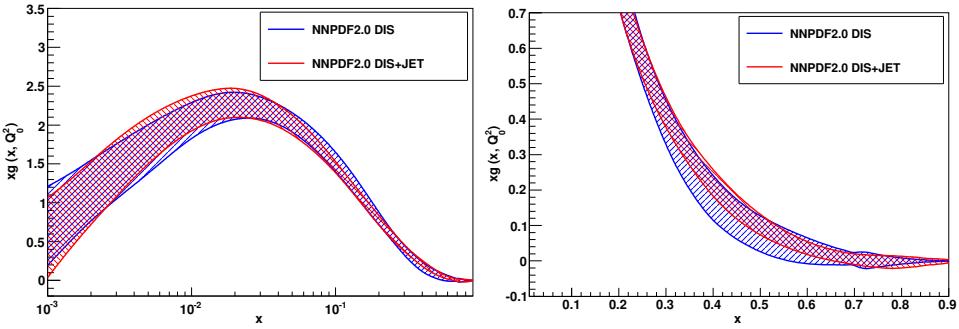

Fig. 15. Impact of jet data on the gluon distributions at small x (left) and large x (right) using the NNPDF2.0 [41] parton fit.

4.5. Global fits

It is clear that the wider the number of different processes, the greater the amount of information which is being used in the determination of PDFs. The price to pay for this, as we will discuss in Sec. 5, is that the determination of PDFs and especially their uncertainties from diverse and possibly inconsistent data might be nontrivial — at the very least, it is going to be computationally intensive. Current global fits use all the processed discussed so far in order to control as much as possible different aspects of PDFs.

The dataset used in one such fit (NNPDF2.0 [41]) is shown in Fig. 16. Different data in this set constrain different aspects of PDFs, along the lines of the preceding discussion, in a way which, referring to this specific dataset, can be summarised as follows:

- information on the overall shape of quarks and gluons at medium x as well as on the isosinglet–isotriplet separation come from fixed-target DIS data on proton and deuterium targets (dominated by γ^* exchange), denoted in the plot as NMC [54], NMCpd [55], SLAC [56] and BCDMS [57];
- an accurate determination of the behaviour of the gluon and quark at small x (where it is dominated by the singlet) and by individual light flavours at medium x (where CC and NC data play a role in separating individual flavours) is found from the very precise HERA CC and NC

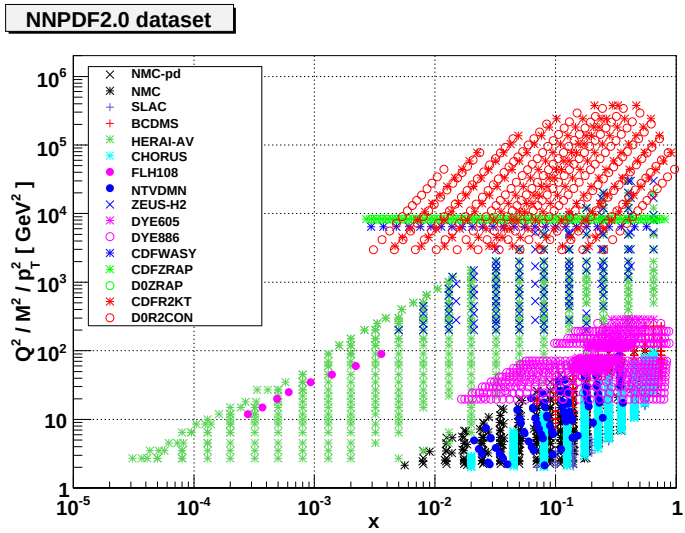

Fig. 16. The data set used in a typical global PDF determination (NNPDF2.0 [41]).

data denoted in the plot as HERAI-AV [58], which were obtained by combining the ZEUS and H1 data from the HERA-I run. More recent HERA-II ZEUS NC [59] and CC [60] data (ZEUS-H2) are also used;

- information on the flavour separation at small x comes from Tevatron Drell–Yan data (in particular the W asymmetry, as discussed above) denoted in the plot as CDFWASY [61], CDFZRAP [62], D0ZRAP [63];
- the flavour separation at medium x is mostly controlled by the Tevatron Drell–Yan data on fixed proton and nucleus target, DYE605 [64] and DYE866 [65–67] in the figure;
- the total valence component is constrained by the neutrino inclusive DIS data, denoted as CHORUS [68] in the plot;
- strangeness is controlled by neutrino dimuon data (NTVDMV [69,70]), as well as by the interplay of the W and Z production data with lower scale DIS and Drell–Yan data;
- the large x gluon, already determined by DIS scaling violations, is further constrained by Tevatron jet data (CDFR2KT [71], D0R2CON [72]).

Other global fits may differ in some detail, such as the specific choice of experiments or the addition or subtraction of some set of data, but are mostly based on datasets constructed on the basis of a similar logic. Smaller datasets, typically a subset of the above, are also considered.

Future improvements on some of these processes, in particular Drell–Yan (including W and Z) production and jet production will certainly come from the LHC, both because of the higher available center-of-mass energy (compare Fig. 3), and because of the higher statistics which will be accumulated once higher or design luminosity are reached. Some other processes which are likely to become important at the LHC are prompt photon and heavy quark production, as well as Higgs production (if the Higgs is found and understood), all of which are sensitive probes of the gluon distribution. We will briefly come back on these issues in Sec. 6.2, after discussing the current main difficulty in the understanding of PDFs, namely, the treatment of PDF uncertainties.

5. PDF uncertainties

The accurate determination of PDF uncertainties is clearly necessary if one wants to be able to obtain meaningful predictions from the factorised QCD expressions of Sec. 2. Because PDFs are determined by comparing QCD predictions to the data, as discussed in Sec. 3, any uncertainty in the theory used to obtain these predictions will propagate onto the PDFs themselves. Such uncertainties include genuine theoretical uncertainties, such as a lack of knowledge of higher-order perturbative corrections: these, generally, do not have a simple statistical interpretation (and in particular they are generally not Gaussian). They also include a lack of knowledge of parameters in the theory, in particular the value of the strong coupling constant $\alpha_{\rm s}$ and the heavy quark masses m_c and m_b, which generally do follow Gaussian statistics. The treatment of these uncertainties is in principle straightforward, in the sense that all one has to do is propagate them onto the PDFs — their effect on PDFs is no different from their effect on the calculation of a physical observable, and PDFs do not entail any new problem. For example, if it is agreed that higher order corrections on cross-sections can be conventionally estimated by varying renormalisation and factorisation scales in a certain range, to be interpreted, say, as a 90% confidence level with flat distribution, the associate PDF uncertainty is simply found by repeating the PDF determination while performing this variation. We will refer to these as "theoretical uncertainties", and come back to them in Sec. 6.1.

On top of these, however, PDFs are affected by statistical uncertainties which are related to the way the information contained in the data is propagated onto a PDF determination following the process summarised in Fig. 7. The determination of these uncertainties is highly nontrivial because, as discussed in Sec. 3, the desired final outcome of this process is the determination of a probability distribution in a space of functions: these uncertainties are supposed to behave as genuine statistical uncertainties, with a well-defined

probability distribution, and it is not obvious how to make sure, and then verify, that this is the case. These will be referred to as "PDF uncertainties" for short.

First attempts to determine PDF sets which include PDF uncertainties are only quite recent [73–75]; they immediately met with the difficulty that as soon as wide enough datasets (such as those discussed in Sec. 4.5) are fitted, a standard statistical approach does not seem to be adequate [76, 77]. Furthermore, results obtained for relevant LHC processes such as Higgs production using various different sets [78] do not always agree well with each other. On both of these issues, there has been considerable progress over the last several years. On the one hand, the understanding of statistical issues related to PDF uncertainties has advanced considerably, and it will be reviewed in the remainder of this section. On the other hand, existing PDF determination show a distinct convergence as various phenomenological and theoretical issues are addressed and understood, as we will see in Sec. 6.2.

5.1. Tolerance

Available fits to wide enough sets of data based on the Hessian approach and the "standard" parton parameterisation, Eq. (26), discussed in Sec. 3.1 run into the difficulty that the best-fit is not simultaneously a best-fit for individual datasets. Specifically, one can test for the possibility that the χ^2 of the fit to individual datasets entering the global fit may be improved by moving away from the global minimum by introducing Lagrange multipliers to select which dataset to minimise [37]. Results, shown in Fig. 17, are disquieting: not only the minima of individual experiments do not coincide with the global minimum but some of these minima seem to deviate much more than one might expect on the basis of statistical fluctuations, and there even seem to be runaway directions for some experiments.

This suggests that likelihood contours (for example one-σ) for the global fit can only be determined while simultaneously testing for the degree of agreement of individual experiments with it. The way this is done is by introducing the concept of "tolerance", defined as follows [76]. First, the Hessian matrix is diagonalised. Next, one moves the value of each eigenvector away from the minimum of the global fit in either direction, and one computes the χ^2 of each experiment. Then, for each experiment one determines both the position of the minimum of the χ^2 and the one-σ interval about it (corresponding to the $\Delta\bar{\chi}^2 = 1$ variation about the minimum), or equivalently the 90% confidence level (obtained by rescaling the former interval by the factor $C_{90} = 1.64485\ldots$ [34]). Finally, one takes the envelope of the error bands for individual experiments at the desired confidence level

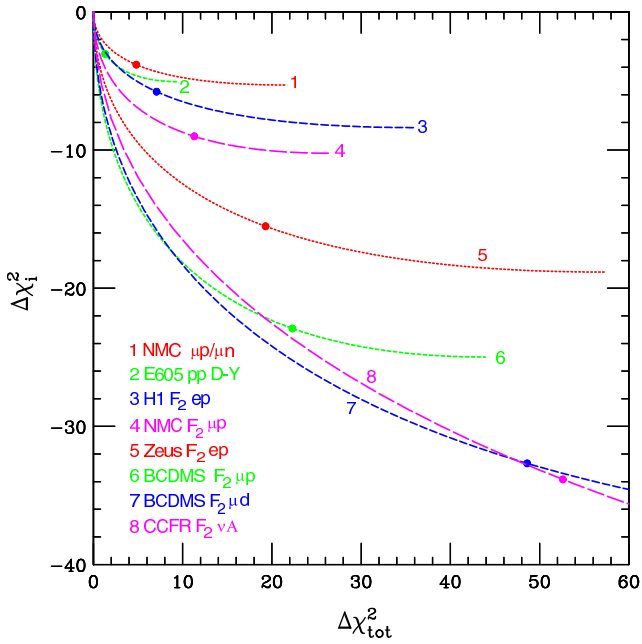

Fig. 17. Decrease of the χ^2 for various dataset entering a global fit plotted as a function of the increase in χ^2 of the global fit when moving away from the global minimum (from Ref. [37]).

(C.L., henceforth). For example, at the 90% C.L. one determines the range of variation in parameter space along this eigenvector about the minimum such that the 90% C.L. interval of each experiment overlaps with this range. This gives a tolerance interval for the given eigenvector. The width of this interval can be measured in units of the variation of the χ^2 of the global fit. This defines a tolerance: $T^2 = \Delta\chi^2$ is the width of the envelope (see Fig. 18).

The 90% C.L. is finally taken to be $\Delta\chi^2 = T^2$ instead of $\Delta\chi^2 = c_{60}^2$ (equivalently, the one σ contour is $\Delta\chi^2 = T^2/c_{60}^2$). The logic behind this is that PDFs should allow one to obtain predictions for new processes at the desired confidence level: for instance, the actual result for a new measurement should have a 68% chance of actually falling into the predicted one-σ band. If new experiments behave as the experiments which are already included in the fit do on average, then this will happen for the one-σ band defined in this way, while if the one-σ band were defined on the basis of standard statistics the chances of the measurements falling outside the band would be much higher. It should be stressed that therefore a tolerance analysis is required for a fit based on this methodology to be reliable (unless the dataset adopted is very small and/or consistent).

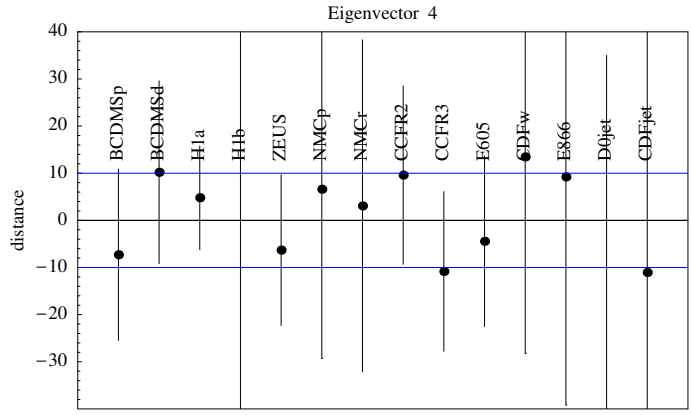

Fig. 18. Determination of the tolerance for a global fit (from Ref. [76]). The distance on the y axis is given in units of $\sqrt{\bar{\chi}^2}$ Eq. (29) of the global fit. The interval shown for each experiment corresponds to a 90% confidence level. The band shown, corresponding to $\Delta\bar{\chi}^2 = 100$ is just wide enough to accommodate the ZEUS (upper variation) and CCFR2 (lower variation) experiments.

In Ref. [76] it was found that in practice $T^2 = 100$ worked for all eigenvalues and experiments at 90% C.L. for the dataset and fit considered there, corresponding to $\Delta\chi^2 = T^2/c_{60}^2 \approx 37$ at one sigma. A similar analysis in Ref. [77] found instead $T^2 = 50$. Taken at face value, this would imply that all experimental uncertainties have been underestimated by a factor of about $T/c_{60} \approx 6$ (for $T^2 = 100$) or $T/c_{60} \approx 4$ (for $T^2 = 50$). While some uncertainty underestimation is possible, such a large factor is at best puzzling, and thus its origin deserves further investigation.

The concept of tolerance was subsequently refined, by suggesting that instead of a global tolerance value for all eigenvalues, a different tolerance value, determined as above, be adopted along each eigenvector direction. This is called "dynamical" tolerance [30]. Proceeding in this way, one finds a tolerance $T \lesssim 6.5$, with most values being in the range $2 < T < 5$, so the large tolerance problem is somewhat mitigated. Also, in this approach it is possible to trace which individual experiment is controlling the tolerance range for each eigenvalue. This, together with the expression of the eigenvector in terms of the original parameters, provides insight on the relation between data and PDF parameters and their mutual consistency. Such an analysis is displayed in Fig. 19, where both the tolerance analysis for one specific eigenvector, and then the experiments and corresponding band which control the tolerance interval for each eigenvector.

It is interesting to contrast this treatment of uncertainties in the Hessian approach with "standard" parameterisation with the Monte Carlo approach together with neural network parameterisation discussed in Sec. 3.2.2. In

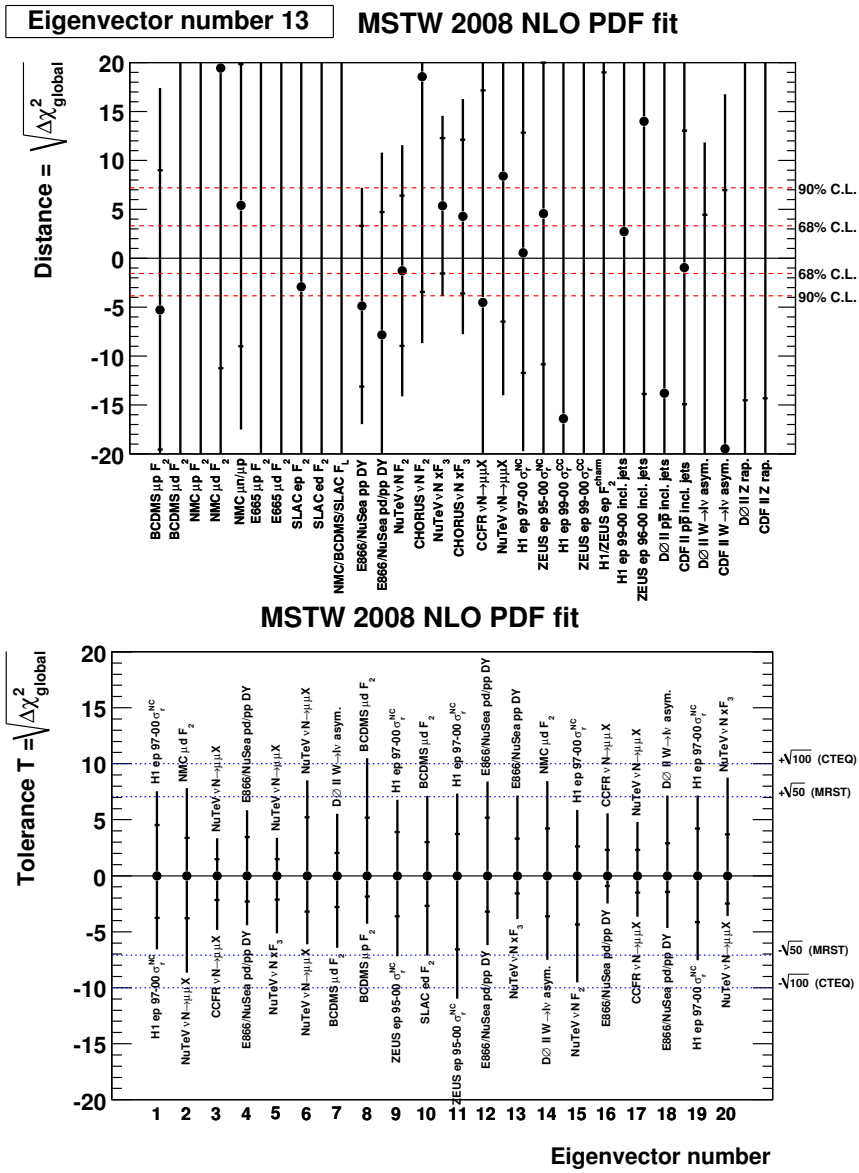

Fig. 19. Dynamical tolerance for the MSTW08 PDF fit. Upper: the tolerance interval for the 13^{th} eigenvector; the inner and outer uncertainty bands correspond for each experiment to the 68% C.L. and 90% C.L. ranges. Lower: the tolerance interval for each eigenvector, along with the experiments which determine it; the $T^2 = 50$ and $T^2 = 100$ previously [76, 77] adopted are also shown (from Ref. [30]).

that approach, uncertainty bands corresponding to any given confidence level can be computed directly from the Monte Carlo sample: the one-σ interval is just the standard deviation of the sample, and one may even check whether it indeed corresponds to the central 68% of the distributions of PDF replicas. This is shown in Fig. 20 for the gluon distribution (from Ref. [38]): in this case (and in fact [41] in most cases) the one-σ and 68% C.L. intervals coincide. In a Monte Carlo approach, whether or not the fits behave consistently when comparing fit results to new data, and then including these new data into the fit, can be verified *a posteriori* by performing statistical tests on the fit results. These tests were performed successfully for the fits of Refs. [38, 41, 53].

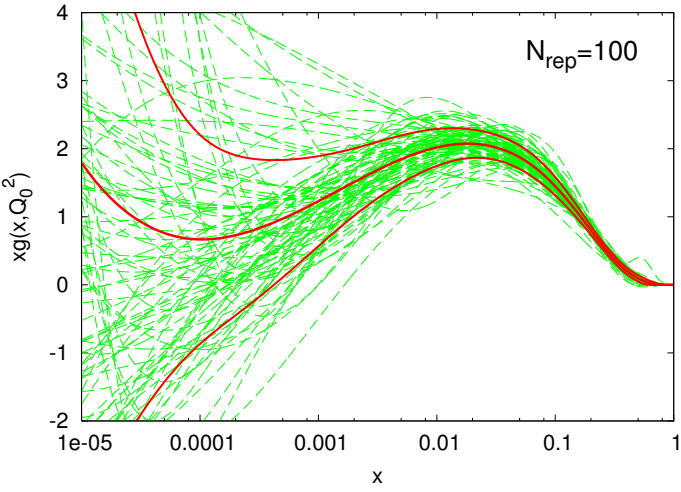

Fig. 20. One-σ interval computed from a distributions of 100 replicas of the gluon distribution.

The question of the appropriate range in global χ^2 which corresponds to one sigma is thus side-stepped. In principle, it can be answered *a posteriori*: in a Monte Carlo approach, the χ^2 of the mean is a property of the Monte Carlo sample, so one could compute the one-σ interval from the sample itself. In practice, it is nontrivial to do this accurately because, as explained in Sec. 31, the $\bar{\chi}^2$ has fluctuations of order $N_{\rm dat}$, and in a Monte Carlo approach these fluctuations take place replica by replica, so one needs a very large sample to determine the χ^2 accurately.

However, the issues which may be responsible for the large tolerances can be addressed both in a Hessian and in a Monte Carlo approach as we will now discuss.

5.2. Parametrisation bias and data incompatibility

The large tolerance values discussed in Sec. 5.1 are, by definition, a manifestation of the poor mutual compatibility of the experiments that go into the global fit. One possible explanation for this is that experiments are genuinely incompatible with each other within their stated uncertainties, *i.e.* that their published uncertainties are underestimated. We will refer to this possible explanation as "data incompatibility".

Another possible explanation is that the way uncertainties are propagated from experiments onto PDFs leads to underestimating the uncertainty in the latter. For example, assume that experiment A does not depend on some PDF parameter, and that one determines PDFs from this experiment, but instead of leaving the undetermined parameter free, one fixes it in some arbitrary way. If the ensuing PDF is then used to predict another experiment B which happens to depend on the undetermined parameter the likelihood of results being in agreement with the prediction will not depend on statistics, but rather in the arbitrary way the parameter has been fixed. We will refer to this as "parameterisation bias".

Of course, other options are possible: for example, that the theory which is being used is not adequate. In the latter case, however, one would have to find a convincing argument why this theoretical inadequacy has not been seen elsewhere.

Data incompatibility in the Hessian approach was recently studied in a quantitative way in Ref. [79], exploiting the observation [80] that once the χ^2 has been written in the form of Eq. (32) one can perform a further linear transformation of the parameters which preserves this form, while also diagonalising the contribution to the χ^2 from some specific subset of data. After this simultaneous diagonalisation, the χ^2 is written as the sum of a contribution from the data in the given subset and the rest: the distance of the minima of these two contributions to the χ^2 in units of the corresponding standard deviation measures the compatibility of the given subset of data with the rest of the global dataset. The idea is then to study the distribution of such distances, in all cases in which the experiment does contribute significantly to the global minimum. If experimental uncertainties are correctly estimated, they should be Gaussianly distributed. The results of this analysis, shown in Fig. 21, suggest that the distribution of discrepancies deviates significantly from a Gaussian distribution, and that if it is fitted to a Gaussian its uncertainty should be rescaled by about a factor 2. This suggests uncertainty underestimation by a similar factor, which corresponds to a value of the tolerance for 90% C.L. of order of $T^2 \sim 10$.

This suggest that data incompatibility can explain only in part the need for large tolerance. Further evidence that data incompatibility is at most moderate can be obtained in a Monte Carlo approach, by comparing the

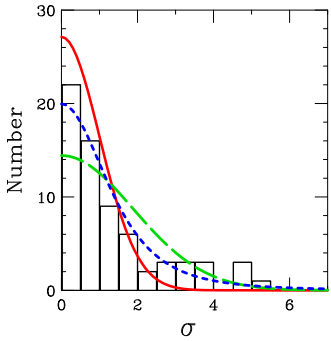

Fig. 21. Distribution of discrepancies between each experiment entering a global fit, and the global best fit. The solid (red) curve is the standard Gaussian distribution, the dashed (green) curve is a Gaussian distribution with σ rescaled by a factor 1.88, the dotted (blue) curve is a Lorentzian distribution (from Ref. [79]).

effect of the subsequent inclusion of different datasets into a fit. Indeed, if some datasets were incompatible with others, then the effect of their inclusion in the global fit would change according to whether the global fit already includes the data with which they are incompatible or not. Assume for example that the gluon determined from jets is compatible with that found exploiting scaling violations in DIS data, but less compatible with that found from scaling violations in Drell–Yan: then, inclusion of jet data in a pure DIS fit would have a different effect than their inclusion in a fit which contains both DIS and Drell–Yan. When such tests are performed [41] no evidence for data incompatibility is found, as demonstrated in Fig. 22.

Let us now turn to the possibility that parameterisation bias may be responsible for the effect. The way this could happen in a Hessian approach was recently exemplified in Ref. [81]. Assume a relevant parameter on which PDFs depend is not fitted, but rather fixed by assumption at a value which is away from its best-fit. Then, clearly (see Fig. 23) the one-σ range for the other parameters when this parameter is kept fixed corresponds to a variation of χ^2 which is greater than the standard $\Delta\chi^2$ found when moving away from the minimum. A first estimate of the possible size of this effect was also provided in Ref. [81] by simply repeating the PDF fit of Ref. [27], but with a much more general parameterisation, based on expanding the gluon on a basis of orthogonal polynomial, analogous to that of Ref. [24] shown in Fig. 9. With this more general parameterisation, fits whose χ^2 is similar to or better than the best-fit χ^2 of the more restrictive parameterisation are found to span a band which corresponds roughly to the $\Delta\bar{\chi}^2 = 10$ range for the restrictive parameterisation. So this suggests a tolerance of at least $T^2 = 10$ just to account for the bias on the gluon shape imposed by the parameterisation of Ref. [27].

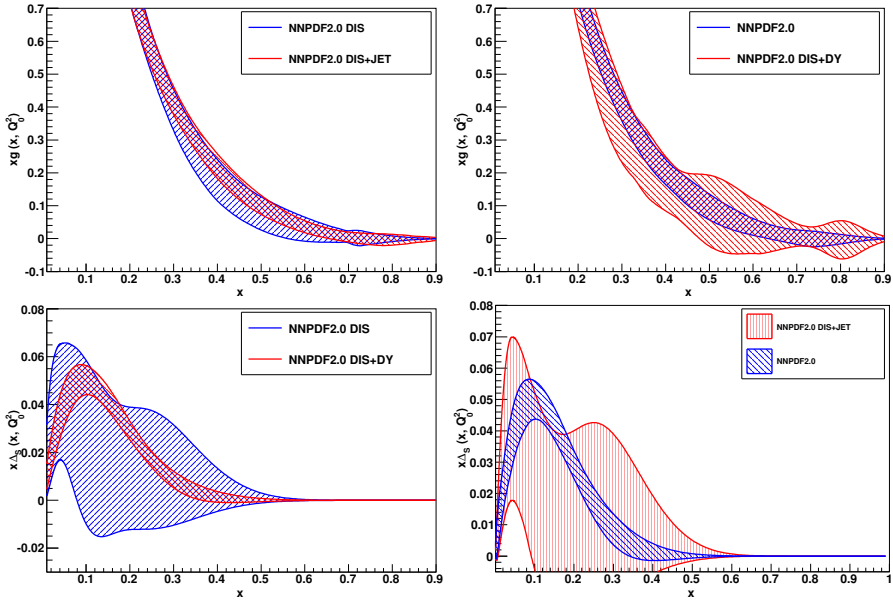

Fig. 22. Tests of data compatibility by changing the order of inclusion of data in fits with different datasets, based on the NNPDF2.0 [41] PDF determination. Top: effect on the gluon distribution of the inclusion of jet data in a fit to DIS data only (left) or on a fit to DIS+Drell–Yan data (right). Bottom: effect on the sea asymmetry Eq. (43) of the inclusion of Drell–Yan data in a fit to DIS data only (left) or on a fit to DIS+jet data (right).

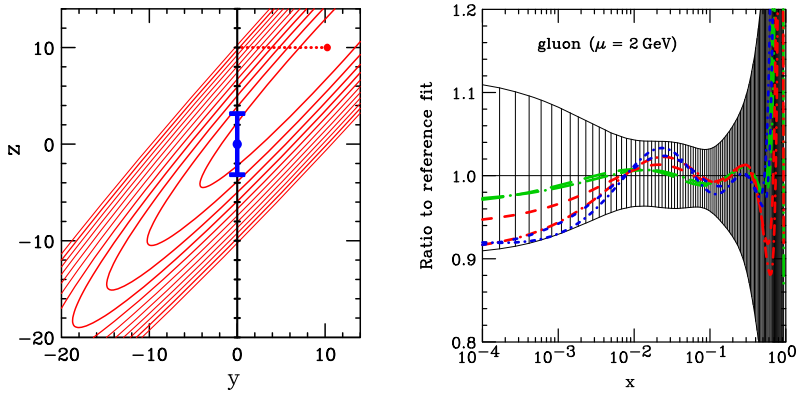

Fig. 23. Parametrisation bias. Left: the one-σ range for parameter z when parameter y is kept fixed away from the best-fit value corresponds to a range which is larger than $\Delta\chi^2 = 1$. Right: various gluons (dashed curves) based on a general parameterisation which lead to the same or better fit quality as the best-fit gluon of Ref. [27], compared to the $\Delta\chi^2 = 10$ band about the latter gluon (from Ref. [81]).

The issue of parton parameterisation and bias thus deserves further investigation. First, one may ask whether the Gaussian assumption is by itself a source of bias. This has been investigated in Ref. [23], using a Monte Carlo approach together with a standard parton parameterisation. Data replicas based on the HERA DIS data ($F_i(n)$ of Fig. 7) have been generated either using a Gaussian or a lognormal distribution (see Fig. 24), and used for a PDF fit based on the "standard" functional form Eq. (26). Results are shown in Fig. 24, and are seen to be essentially indistinguishable. The choice of the probability distribution of the data does not seem to play any major role, as one might have expected from the central limit theorem: with so many data, everything looks Gaussian. On the other hand, in the same figure we also show the gluon obtained in a fit to exactly the same data, but using the neural network functional form and associate cross-validation methodology of Ref. [53]. It is clear that the uncertainty is now much wider. This suggests that it is the form of parameterisation which plays a dominant role, rather than the form of the probability distribution.

The issue has been investigated further in the HERAPDF [29] PDF fits, where the standard $\Delta\chi^2 = 1$ PDF uncertainty based on a "standard" functional form Eq. (26) has been supplemented by a further parameterisation uncertainty, obtained by varying the assumed functional form (in particular, the large x behaviour, the number of terms in the polynomial Eq. (27), and the assumptions on strangeness which is not fitted). It is clear from Fig. 25 that this leads to a sizable enlargement of the PDF uncertainty band.

The Monte Carlo approach together with neural network offers an interesting way of searching for the origin of uncertainties, in that different sources of uncertainty can be switched on and off one at a time. In particular, one may perform the following exercise. First, one freezes the generation of data replicas, and one takes each replica dataset equal to the central values of the data. Recall from Sec. 3.2.2 that each replica is fitted to a different, randomly chosen subset of the data. Hence, all datasets F_i of Fig. 7 are now the same, and only the way they are partitioned in validation and training sets changes between replicas. Each PDF replica is thus obtained as the fit to a different partition of the central experimental data. The (square) fluctuation of the data are reduced by a factor two: instead of having replicas which fluctuate about experimental data which in turn fluctuate about their "true" values, one only has different subsets of central data fluctuating about their true values.

In Table II we compare some indicators of fit quality and results for a fit obtained in this way to those of the corresponding standard fit (using the fit of Ref. [38], based on DIS data). In particular, we compare the χ^2 of the best fit in either case: this is unchanged. However, the average χ^2 of each replica fit is smaller by a factor two. This is as it should be: in both cases the best-fit

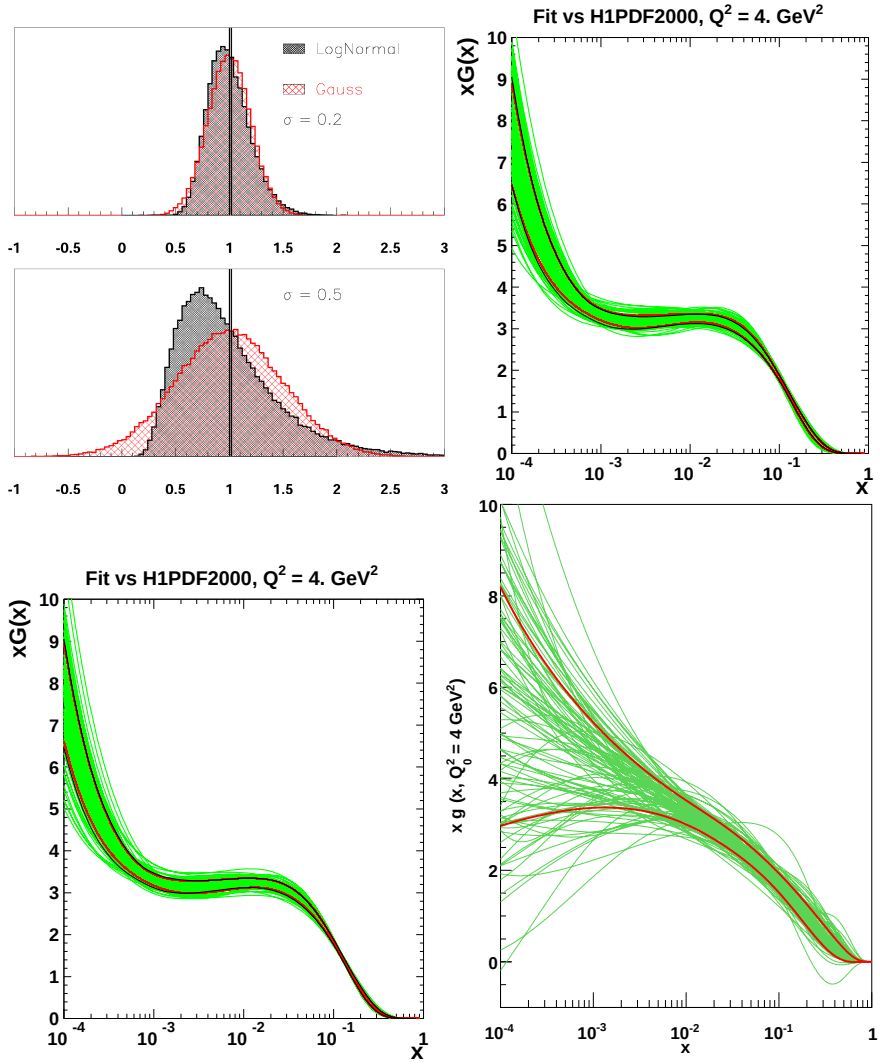

Fig. 24. Comparison of fits based on Gaussianly or log-normally distributed data. Upper left, upper right, lower left, lower right: Comparison of the Gaussian and lognormal distribution. Gluon determined from lognormal data. Gluon determined from Gaussian data (from Ref. [23]). Gluon determined from the same Gaussian data, but using the neural network parameterisation of Ref. [53] (from Ref. [82]).

is reproducing the same central best-fit value, but the fluctuation of replicas about it are now suppressed by a factor two, and thus the average χ^2 per replica is reduced by approximately the same factor. The surprising result however is found when one computes the average percentage uncertainty in the prediction obtained in either case. This is determined as the percentage

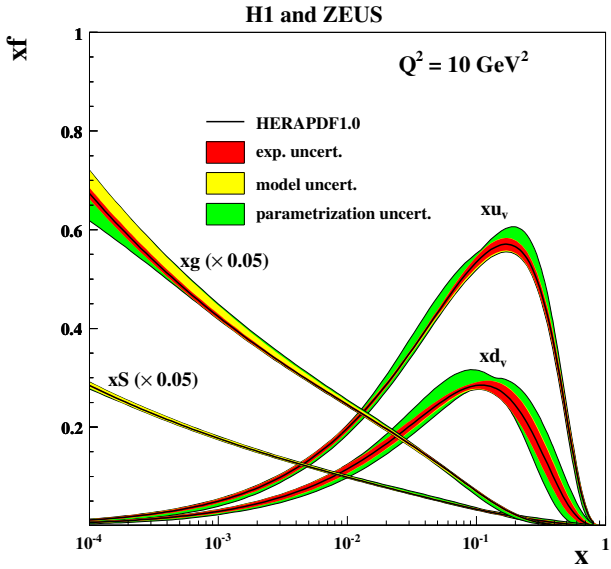

Fig. 25. Comparison of statistical and parameterisation uncertainties in the HERAPDF fit (model uncertainties denote what we call theoretical uncertainties), from Ref. [29].

uncertainty of the prediction obtained from the final replica PDF set, for all datapoints included in the fit, averaged over datapoints. One might expect that having halved the fluctuations, the average uncertainty should be reduced by a factor $\sqrt{2}$; in actual fact, it is reduced by a much smaller amount.

TABLE II

Values of the χ^2 for the best fit (first row), the average and standard deviation of the χ^2 of individual replicas (second row), and the percentage uncertainty of the prediction averaged over all data points (third row) for the PDF determination of Ref. [38] (first column); the same but with all PDF replicas fitted to different partitions of the experimental central values (second column); the same but with all data replicas fitted to the same partition of the experimental central values (in the latter case the process has been repeated with 5 different choice of fixed partition and averaged).

	Replicas	Central value	Fixed partition
χ^2	1.32	1.32	$\sim$1.3
$\langle\chi^2\rangle_{\rm rep}$	2.79 ± 0.24	1.65 ± 0.20	$\sim 1.6 \pm 0.2$
$\langle\sigma\rangle_{\rm dat}$	0.039	0.035	$\sim$0.03

The origin of this state of affairs can be understood by performing an even more extreme text: one simply produces 100 replicas fitted to exactly the same partition of the central data. In this case, all F_i contain the same data, partitioned in the same way into training and validation sets. Naively one may think that this may lead to simply repeating 100 times the same fit. This is not necessarily the case because each replica is determined by initialising the neural networks at random, and then minimising by means of an (equally random) genetic algorithm. Hence one starts each time from a different point in the very wide parameter space, and then the minimum is approached along a different path. Indeed, in Fig. 26 we show the χ^2 profiles along the minimisation, as a function of the number of iterations of the minimisation algorithm, for two individual replicas: it is clear that even though the final χ^2 values are quite similar, the number of iterations and profiles that take there are quite different, thereby showing that the minimum is approached along different paths.

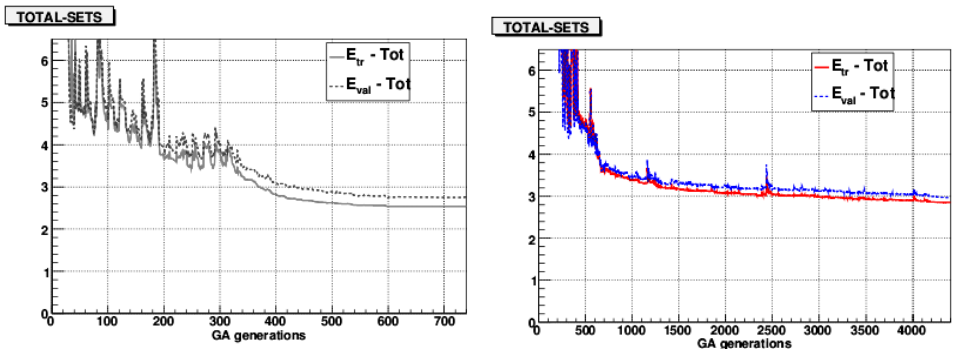

Fig. 26. The χ^2 for the training dataset and total dataset for two different replicas fitted to exactly the same training subset of the central values of the data of Ref. [38], shown as a function of the number of iterations of the minimisation algorithm.

In this case, in order to make sure that results do not depend on the particular partition that has been picked in the first place, the whole procedure is repeated five times with five different choice of starting partition and results are then averaged. Results are shown in Table II (they should be taken as indicative, because one should use rather more than five fixed partitions for accurate results). These results are quite surprising. The χ^2 of the best fit and the average over replicas are unchanged, and this is to be expected: it shows that indeed the five replicas chosen are not special. However, very surprisingly, the average uncertainty, which one might expect to be tiny, is more than 50% of that of the original fit. This is also seen by comparing results for PDFs (see Fig. 27: the uncertainty band is smaller, but of the

same order of magnitude as that of the full fit. The inevitable conclusion is that a large fraction of the uncertainty band, probably more than half, does not depend on the fluctuations in the data. Rather, it is a consequence of the fact that there is an infinity of functions that provide fits of comparable quantity to the data. Different minimisation profiles such as those shown in Fig. 26 land on somewhat different minima; the uncertainty of this fit is then a measure of the spread of this space of minima. Once understood that a sizable fraction is simply due to this "functional" uncertainty, it is clear why it is more difficult to capture with a fixed parameterisation, which then requires a suitable tolerance in order to mimic it.

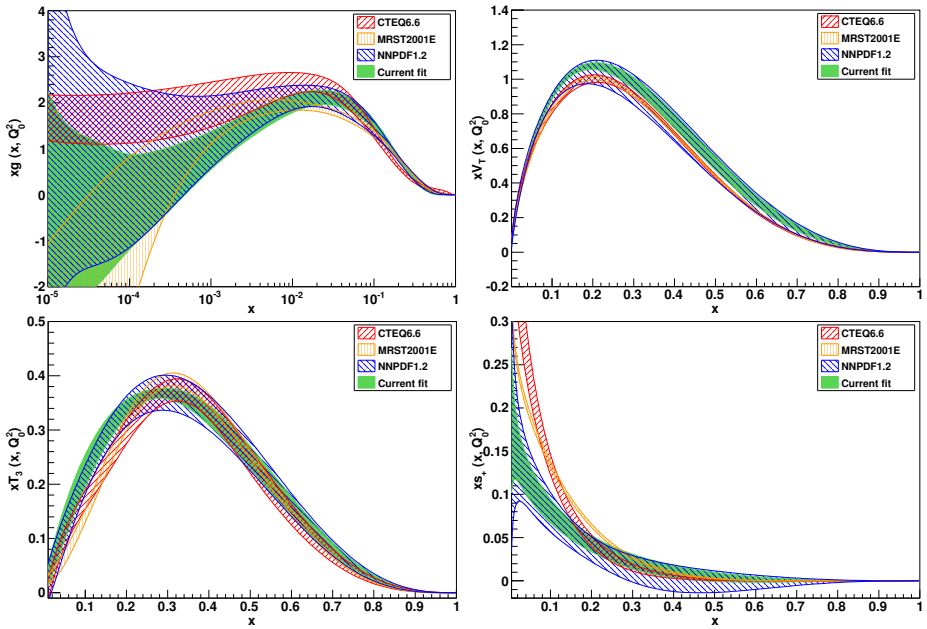

Fig. 27. Comparison of PDFs from the Monte Carlo set of Ref. [38] (NNPDF1.2, hatched slanted to left) to PDFs determined using the same data and procedure, but fitting all replicas to the central data (dark). The PDFs from Ref. [77] (MRST2001E, vertical hatched) and Ref. [27] (CTEQ6.6, hatched slanted to right) are also shown for comparison. The PDFs shown are the gluon (top left), total valence Eq. (39) (top right), triplet Eq. (37) (bottom left) and total strangeness s^+ Eq. (45) (bottom right).

6. Recent developments

The state of the art in PDF determination is moving very fast and thus any attempt to review it would necessarily become obsolete quite rapidly: a recent review of the status of the field as of November 2010 is in Ref. [83].

Here, in an attempt to discuss issues of somewhat less fleeting value, we will first briefly review theoretical uncertainties, which are the current frontier of PDF determination, then summarise what progress has been made and what remains to be done in the determination of PDFs for the LHC in the years to come.

6.1. Theoretical uncertainties

As already mentioned, the PDF uncertainties discussed in Sec. 5 are the result of propagating into the space of PDFs the uncertainty on the data on which the PDF determination is based. Most of the effort has gone so far in their determination and understanding because they are likely to be at present the dominant uncertainty. However, it has been recently realized that in many cases uncertainties related to the theory used to extract PDFs from the data may be larger than one may think. These, as already mentioned, include both uncertainties in the theory itself (such as higher order corrections) but also uncertainties in the knowledge of the free parameters on the theory.

The most obvious source of theoretical uncertainty is the value of the strong coupling α_s. The PDF which depends most strongly on it is the gluon distribution, which, as discussed in Sec. 4.4 is largely determined by scaling violations: the rather strong dependence of the gluon on the value of α_s is shown in Fig. 28 for various PDF sets. Note that even though, as discussed in Sec. 3.1 the total PDF $+ \alpha_s$ uncertainty can be obtained by determining these two uncertainties separately and adding results in quadrature [35], when determining the α_s uncertainty, the value of α_s in the factorisation formula Eq. (1) must be varied both in the PDFs and in the partonic cross-section $\hat{\sigma}$. This is especially important when dealing with processes, such as Higgs production in gluon–gluon fusion [83, 85] (or also top production) which depend on the gluon PDF and start at a high order in α_s. For this purpose, PDF sets corresponding to different values of α_s are necessary and have thus been produced by several group (using sets with PDFs given as a continuous function of α_s is in principle also possible, but practically more cumbersome and less accurate).

The only other free parameters in the QCD Lagrangian are the quark masses, *i.e.*, in the perturbative regime, the heavy quark masses. Dependence of PDFs on them are larger than one might naively expect, and has two different origins which we will now discuss in turn. The first, is simply the fact that even though all perturbative computation are done up to power-suppressed terms in Q^2, terms of order $\frac{m_c^2}{Q^2}$ and $\frac{m_b^2}{Q^2}$ may have a non-negligible impact on PDF fits. This was brought to general attention by the comparison [27] of two calculations of the W and Z production cross-sections

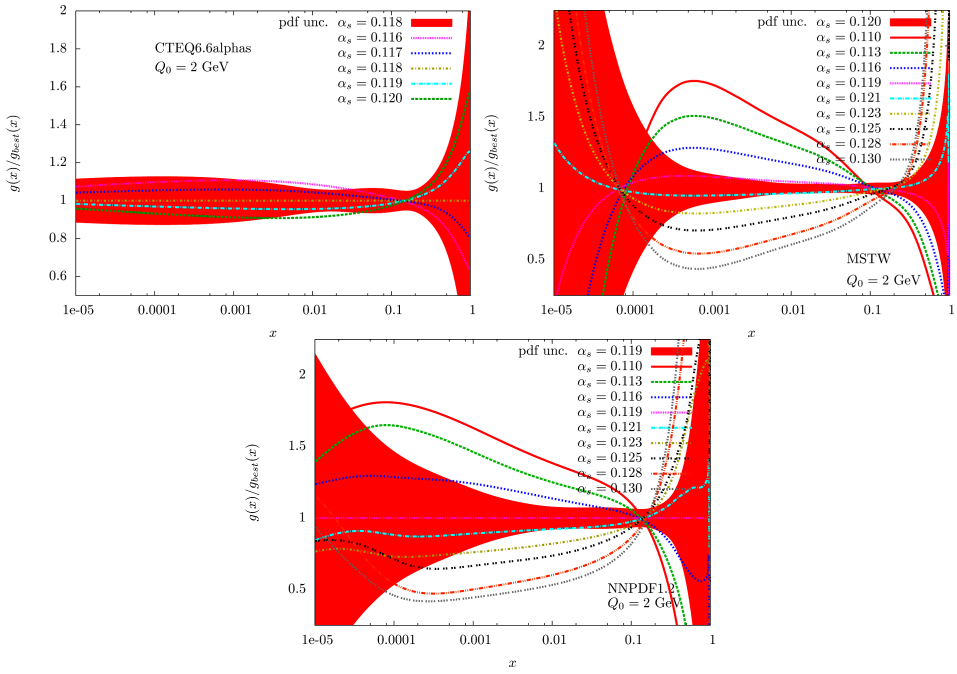

Fig. 28. Dependence on α_s of the gluon distribution determined in the fits of Ref. [27] (CTEQ6.6, upper left), Ref. [30,84] (MSTW08, upper right) and Ref. [38] (NNPDF1.2, lower) (from Ref. [85]).

based on PDF sets which differ mostly because one does include $\frac{m_c^2}{Q^2}$ corrections (CTEQ6.1 [86]) while the other (CTEQ6.6 [27]) does not. It turns out (see Fig. 29) that these corrections change the result by an amount which is almost twice the (statistical) PDF uncertainty (as defined in Sec. 5).

In order to understand what is going on here, one must recall the way heavy quark PDFs are defined, and heavy quarks are treated in perturbative QCD computations. Usually, perturbative QCD computations are performed in a decoupling renormalization scheme [87], in which heavy quarks decouple from perturbative Feynman diagrams for scales much lower than the quark mass. In such a scheme, below threshold the number of flavours in the evolution equation for the strong coupling and for parton distributions Eq. (22) is equal to the number of light flavours. So in particular whereas there may still exist a non-perturbative nonvanishing "intrinsic" heavy quark PDF below threshold [19], it will only start evolving and coupling to other PDFs through evolution equations above threshold: for all practical purposes, below charm threshold $n_f = 3$. However, for scales which are much larger that the heavy quark mass, there is no reason to treat the heavy quark on a different footing from any other light quark: for scales Q^2 much larger

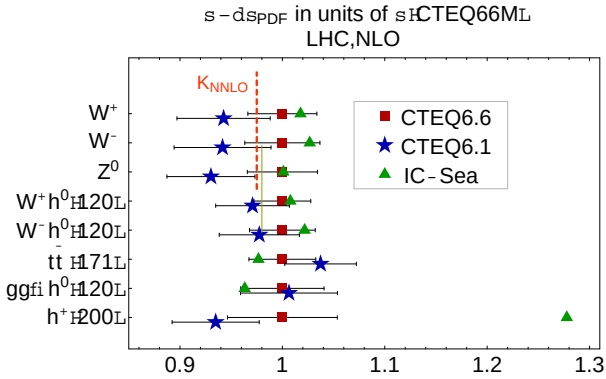

Fig. 29. The total W and Z cross-section at the LHC with $\sqrt{s} = 14$ TeV computed using PDFs determined including charm mass corrections (red squares [27]) and neglecting them (blue stars [86]). Other processes and PDF sets not discussed here are also shown (from Ref. [27]).

than the charm mass, all terms of the order of $\frac{m_c^2}{Q^2}$ can be neglected, and there are $n_f = 4$ massless flavours. The problem arises for scales which are higher but not much higher than the heavy quark mass: then, the heavy quark can be produced and it does not decouple, yet it is not necessarily a good approximation to treat it as massless, *i.e.* to neglect $\frac{m_c^2}{Q^2}$ corrections.

This situation is illustrated in Fig. 30, where we show the neutral-current photon-induced DIS F_2^c structure function (*i.e.* the contribution to F_2 in which the virtual photon couples to a charm quark) computed in various approximations. The two curves labelled ZM-VFN (purple, highest curves) are curves in which charm is simply treated as another massless flavour. The charm PDF is assumed to vanish below threshold, and above threshold a (massless) charm component is generated by perturbative evolution — note that even if an intrinsic component did exist, it should be small, and only non-negligible for very large x [19]. The two NLO and NNLO ZM-VFN (zero mass-variable flavour number) curves correspond to the case in which anomalous dimensions are computed up to $O(\alpha_s^2)$ and $O(\alpha_s^3)$ respectively: the solution to evolution equations then includes all contributions of the order of $\alpha_s^k \ln^n \frac{Q^2}{\mu^2}$ to all orders in α_s, with $n \geq k-1$ at NLO and with $n \geq k-2$ at NNLO. They are called "variable flavour number" curves because the number of active flavours is increased by one unit when each heavy quark threshold is crossed. For light quarks, $\mu^2 = Q_0^2$ — the starting scale of perturbative evolution — and for the charm contribution $\mu^2 = m_c^2$. Given that $Q_0 \sim m_c$, at high scale there is no reason not to include the charm contribution in evolution equations along with the light contributions.

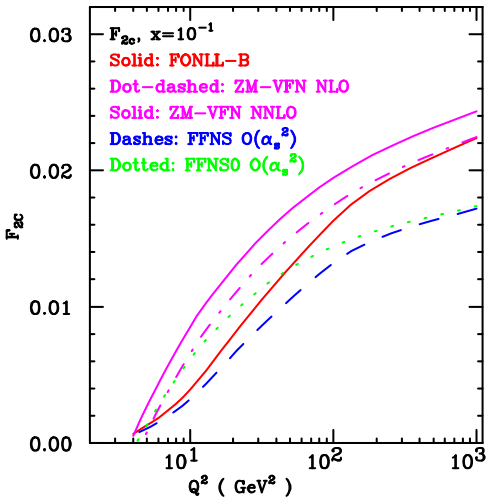

Fig. 30. The deep-inelastic charm structure function F_{2c} computed in various approximations (see text), plotted as a function of scale for fixed $x = 10^{-1}$. The same conventional PDFs are used for all plots.

However, the evolution equations neglect all quark mass effects. Indeed, the curve labelled FFN (fixed flavour number) $O(\alpha_s^2)$ shows the result of the computation fixed order in α_s, but now with the dependence on m_c fully included. This is a fixed-flavour number result because charm is included as a massive quark in partonic cross-sections, but only the lighter flavours contribute to evolution equations and the running of α_s. Its limit for $Q^2 \to \infty$, which coincides with the contributions to the VF-ZFN NNLO, but up to $O(\alpha_s^2)$ only, is labelled as FFN0. The FFN0 and FFN results are different, and their difference, which is sizable for $Q^2 \lesssim 50$ GeV^2 is a measure of the size of mass suppressed contributions (note however that all curves come together at threshold, where F_{2c} vanishes). In this region, $\ln \frac{Q^2}{m_c^2}$ is not large and indeed the ZM-VFN NNLO curve and the FFN0 are quite close: the inclusion of higher order powers of $\ln \frac{Q^2}{m_c^2}$ in the ZM-VFN NNLO has little effect. On the other hand, for $Q^2 \gtrsim 100$ GeV^2 the FFN0 and FFN curve become quite close — their difference are mass-suppressed contributions which here have little effect — but the ZM-VFN curve is much higher: here $\ln \frac{Q^2}{m_c^2}$ is large and its all-order inclusion in the ZM-VFN result is important. In fact, the difference between the FFN curve and the ZM-VFN curve is much larger than the difference between the NLO and NNLO ZM-VFN curves: it is important to resum $\ln \frac{Q^2}{m_c^2}$ to all orders, but whether this resummation is done to NLO or NNLO has a much smaller impact.

So summarizing: for low $Q^2 \lesssim 50$ GeV^2 it is important to include mass corrections and not important to resum $\ln \frac{Q^2}{m_c^2}$ to all orders, so the most accurate result is the FFN one. For $Q^2 \lesssim 50$ GeV^2 the converse is true and the most accurate result is the ZM-VFN one. The question is whether the two can be combined. That this is possible in principle to any perturbative order is a consequence of a factorisation theorem for massive quarks proven to all orders in Ref. [88]. A practical implementation was suggested and worked out up to NLO in Ref. [89] — the so-called ACOT method, which was used to produce the massive CTEQ6.6 result of Fig. 29. Other implementations were suggested in Refs. [90, 91] (TR method) for deep-inelastic scattering, and in Ref. [92] for hadroproduction (FONLL method, recently generalized to DIS and worked out up to NNLO in Ref. [93]). These various methods, which thus combine both a massive fixed-order FFN computation and a massless all-order ZM-VFN resummation, are often referred to as "GM-VFN" (general mass, variable flavour number) methods.

The FONLL curve is also shown in Fig. 30, in a version (called FONLL-B) which includes all terms which are contained in the ZM-VFN NLO and FFN $O(\alpha_s^2)$ calculations. It is clear that the FONLL-B curve nicely interpolates between the FFN curve, more accurate at low scale, and the ZM-VFN one, more accurate at high scale. Without entering a discussion of these various prescriptions, which would be quite technical, it is important to understand that (unless an error is made) all GM-VFN prescriptions include all the terms included in the ZM-VFN and FFN calculations: they may differ first, in the orders at which either of the ZM-VFN and FFN terms are included, and furthermore, because even with the same terms included, subleading terms are generally different, and may in practice have a non-negligible impact. The impact of these subleading terms can only be reduced by pushing to higher orders both the ZM-VFN and FFN contributions which are included in the GM result. This is illustrated in Fig. 31, where the FFN, ZM-VFN and GM-VFN results are compared. The GM-VFN are shown in the version adopted by CTEQ6.6, Ref. [27], of the ACOT method of Ref. [89] (only available at NLO), in the version adopted by MSTW08 [30] of the TR method of Refs. [90, 91], and with the FONLL method of Ref. [93]. It is clear that the differences between the various GM schemes are sizable, and only start decreasing at NNLO-$O(\alpha_s^2)$.

Considerable progress has been made recently in the benchmarking of all these GM schemes (see Ref. [94]), and the use of a GM-VFN scheme, which is highly desirable, has become more widespread. However, a systematic estimate of the uncertainties related to the choice of specific heavy quark scheme is not available in existing PDF sets.

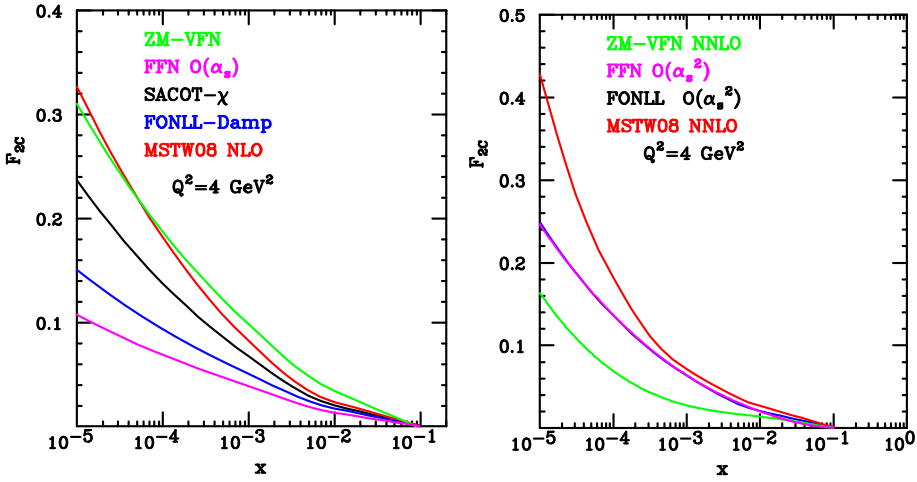

Fig. 31. Comparison of FFN, ZM-VFN and GM-VFN computations of the deep-inelastic charm structure function F_{2c}: left, FFN $O(\alpha_s)$ and ZM-VFN NLO; right, FFN $O(\alpha_s)$ and ZM-VFN NNLO. In each case the various GM-VFN combine all terms in the FFN and ZM-VFN results, and only differ by subleading terms. Results are shown as a function of x, at a scale little higher than the threshold. The same conventional PDFs are used for all plots.

So far we have only discussed one of the two ambiguities related to the treatment of heavy quarks. The other one has simply to do with the value of the heavy quark mass. This has a considerable impact because, as we have seen, apart from possible intrinsic contributions, heavy quark PDFs are obtained by assuming them to vanish at threshold, and then to be generated by perturbative evolution. But changing the mass changes the position of the threshold, and thus the amount of evolution. This has been very recently argued to have a potentially non-negligible impact on phenomenology [96]. A first systematic study has been performed in Ref. [95] within the MSTW08 PDF fit. Some representative results are shown in Fig. 32: when the heavy quark masses are varied in a range which is representative of their uncertainty, the heavy PDFs vary by an amount (more than 5%) which is of the same order or larger than the PDF uncertainty. This variation then propagates onto all other PDFs and especially the gluon, both due to mixing upon perturbative evolution, and to sum rules.

It seems clear that for accurate phenomenology the values of the heavy quark masses will have to be treated analogously to what is now being done for the strong coupling: PDF sets with varying heavy quark masses will have to be provided, and the value of the masses will have to be varied simultaneously in the partonic cross-section and in the PDF sets.

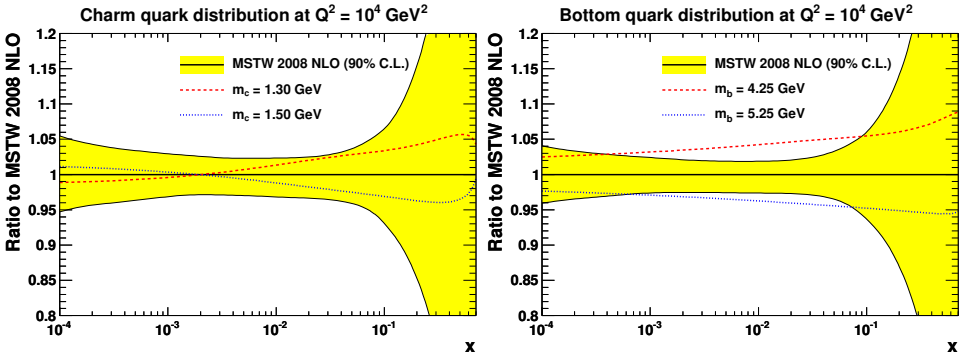

Fig. 32. Dependence of the heavy quark PDFs at a typical electroweak scale on the heavy quark mass: left, charm; right, bottom (from Ref. [95]).

Finally, it should be recalled that there is at present no reliable estimate of the effect on PDFs of the uncertainty due to the truncation of the perturbative expansion. This is possibly a relatively smaller effect in comparison to those we discussed so far, but this too will have to be included in PDF sets, for example by providing sets which correspond to different values of the renormalization and factorisation scales, whose variation is a way of estimating unknown higher order corrections. All in all, a proper treatment of theoretical uncertainty is the current frontier in PDF uncertainties.

6.2. PDFs for the LHC

An ideal PDF determination should include all of the following features:

- It should be based on a dataset which is as wide as possible in order to ensure that all relevant experimental information is retained; in particular, all processes discussed in Sec. 4 should be used.

- It should be based on a sufficiently general and unbiased parton parametrization and/or it should include a careful estimate of the effect of varying the parton parameterisation.

- It should provide PDF uncertainty bands which have been either *a priori* (tolerance) or a posterior (Monte Carlo) checked to provide consistently-sized confidence levels for individual experiments.

- It should include heavy quark mass effects through a GM-VFN scheme, and provide an estimate of the uncertainties due to subleading terms not included in the scheme which has been adopted.

- It should provide PDFs for a variety of values of α_s, reasonably thinly spaced and in a range which is representative of the uncertainty on this parameter.

- *Ditto* for the values of heavy quark masses.

- It should be based on computations performed at the highest available perturbative order, namely NNLO for evolution equations and for most or all of the processes used for PDF determination.

- It should include an estimate of uncertainties related to the truncation of the perturbative expansion.

At present, there exists no PDF determination which has all these features simultaneously. Which of these features is most important for accurate results it will be possible to say with certainty only after such a PDF determination is constructed. However, the various features have been listed in the approximate (decreasing) likely order of importance, at least in the opinion of this author, based on the arguments presented so far.

Therefore, existing sets satisfy only some of these requirements. Current PDF sets are provides through a standard interface, LHAPDF [98], which is regularly updated for the inclusion of new sets and updates. Current PDF sets and their salient features include the following (listed in order of decreasing number of datasets included):

- **MSTW08** [30, 84, 95] Latest in the MRS-MRST-MSTW series of fits (Ref. [99] and subsequent papers). All data of Sec. 4, plus HERA DIS jets. Hessian approach with parameterisation Eq. (26) for seven independent PDFs (three lightest flavour and antiflavours and the gluon); 28 free parameters, 8 of which are held fixed in the determination of uncertainties; dynamical tolerance uncertainties. GM-VFN scheme. Results available for various values of α_s, m_b and m_c. NNLO perturbative order (whenever available).

- **CT10** [28] Latest in the CTEQ series of fits (Ref. [100] and subsequent papers, see also [101]). All data of Sec. 4. Hessian approach with parameterisation Eq. (26) for six independent PDFs (two lightest flavour and antiflavours, total strangeness and the gluon); 26 free parameters; dynamical tolerance uncertainties. GM-VFN scheme. Results available for various values of α_s. NLO perturbative order.

- **NNPDF2.0** [41] Latest in the NNPDF series of fits (Ref. [45] and subsequent papers). All data of Sec. 4. Monte Carlo approach with neural network parameterisation for seven independent PDFs (the three

lightest flavours and antiflavours, total strangeness and the gluon); 259 (37×7) free parameters; cross-validation uncertainties. ZM-VFN scheme. Results available for various values of α_s. NLO perturbative order.

- **JR** [102] Latest in the GR-GRV-GJR series of fits (Ref. [103] and subsequent papers). All data of Sec. 4 except W and Z production. Hessian approach with parameterisation Eq. (26) for five independent PDFs (two lightest flavour and antiflavours and the gluon); 20 free parameters; fixed tolerance uncertainties. Results available for single value of α_s, but Hessian includes α_s. FFN scheme. NNLO perturbative order.

- **ABKM** [104] Latest in the Alekhin series of fits (Ref. [73] and subsequent papers). All DIS data of Sec. 4 and fixed-target virtual photon Drell–Yan production. Hessian approach with parameterisation Eq. (26) for six independent PDFs (two lightest flavour and antiflavours, total strangeness and the gluon); 21 free parameters; no tolerance. Results available for single value of α_s, but Hessian includes α_s, and also heavy quark masses. FFN scheme. NNLO perturbative order.

- **HERAPDF1.0** [29] HERA only DIS data Hessian approach with para-metrization Eq. (26) for five independent PDFs (two lightest flavour and antiflavours, and the gluon); 10 free parameters; no tolerance but inclusion of parameterisation uncertainties. GM-VFN scheme. Results available for various values of α_s. NNLO perturbative order.

Detailed comparisons of these PDF sets is the subject of ongoing benchmarking exercises, with the aim of arriving at the most accurate common determination of PDFs. The successfulness of the enterprise can be gauged by putting side by side (see Fig. 33) predictions for Higgs production via gluon–gluon fusion at the LHC with different PDF sets obtained as the first PDF with uncertainties were published [78], with those obtained more recently as first collisions were taking place at the LHC [85] (see Fig. 33). The improvement is quite clear, and convergence between PDF sets has further improved since.

The status of the computation of the simplest LHC processes (which have been suggested as "standard candles", *e.g.* as a means to measure the machine luminosity [106]) is summarized in the plots of Fig. 34, where predictions for $W^{\pm}$, Z and top total cross-sections obtained at NLO using the ABKM09 [104], CTEQ6.6 [27], HERAPDF1.0 [29], GJR08 [107], MSTW08 [30] and NNPDF2.0 [41] sets are plotted as a function of α_s. The

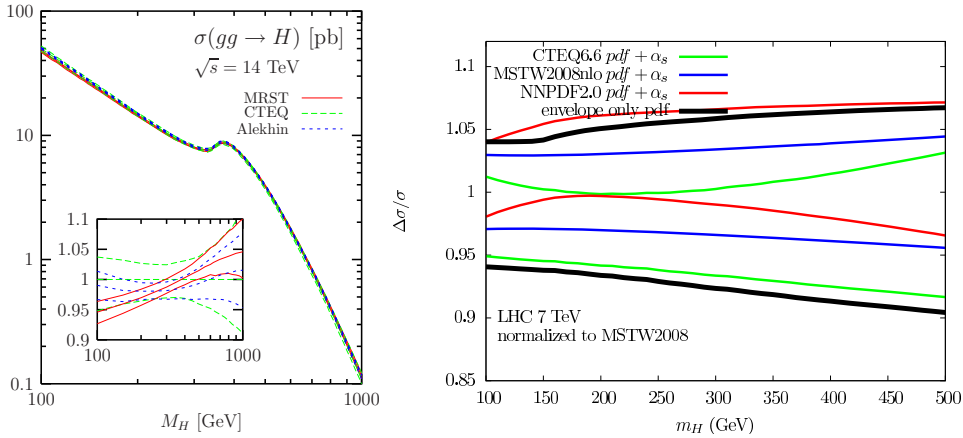

Fig. 33. Comparison of the NLO Higgs cross-section at the LHC as a function of the Higgs mass computed using various PDF sets. Left: status 2004 (from Ref. [78]) using Alekhin2002 [97], CTEQ6 [76] and MRST2001E [77] PDFs ($\sqrt{s} = 14$ TeV). Right: status 2010 (from Ref. [85]) using CTEQ6.6 [27], MSTW2008 [30] and NNPDF2.0 [53] PDFs ($\sqrt{s} = 7$ TeV).

dotted lines show how each prediction can be extrapolated to different values of α_s (the extrapolation is not available for the ABKM and GJR sets since these are only published for a single value of α_s, though in principle the information on the α_s dependence is contained in their covariance matrix). It is clear that first, once brought to a common value of α_s the various sets are in fair agreement, and second, the agreement further improves when restricted to PDF sets that are based on more similar assumptions (such as common dataset, number of independent PDFs *etc.*).

An interesting question that remains is how one can proceed when some disagreement remains, and there is no clear reason to favor one set over the other. In this case, Bayesian statistics provides an answer [108]: in Bayesian terms, the probability distribution for PDFs is a distribution of true values, *i.e.* $P(f)$ expresses the degree of belief that the true value is indeed f. Then, given two different, but *a priori* equally reliable determinations $P_1(f)$ and $P_2(f)$ of the probability distribution, the combined probability is just $P(f) = \frac{1}{2}(P_1(f) + P_2(f))$, with obvious generalizations if the determinations of the probability distribution are more than two or not all equally likely. In a Monte Carlo approach this is especially easy to implement: the combined probability distribution is obtained by simply taking a Monte Carlo sample in which half of the replicas come from either distribution. A 68% C.L. of the combined probability is then simply the region which contains the central 68% of all the given distributions, *i.e.* 68% of the combined replica set.

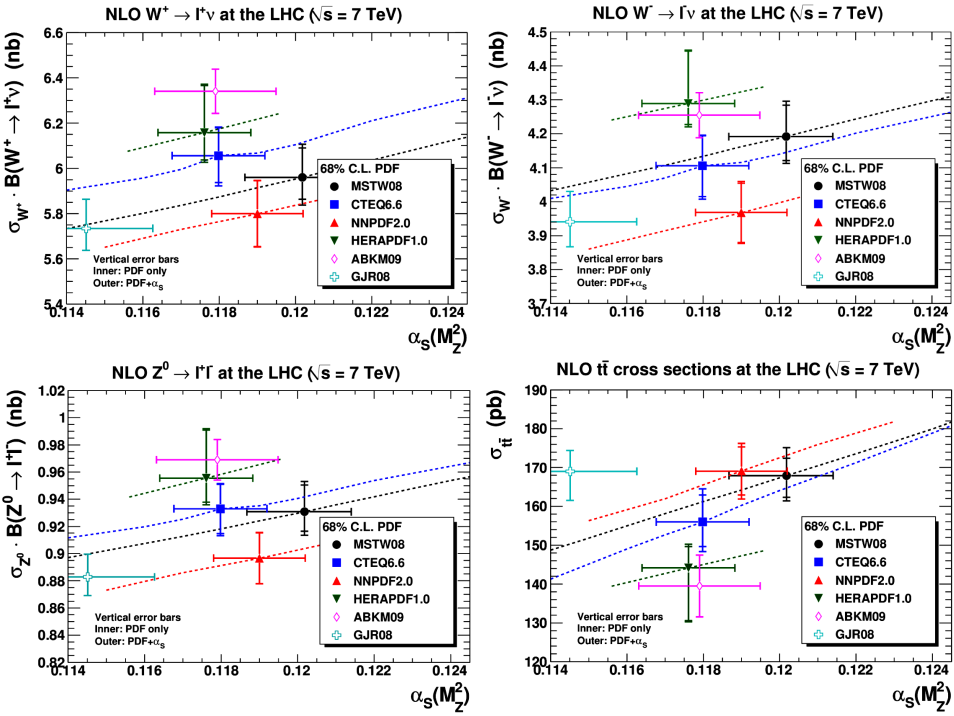

Fig. 34. LHC "standard candles" computed using various PDF sets: total cross-section for the production of W^+ (top left), W^- (top right), Z (bottom left) $t\bar{t}$ (bottom right) (from Ref. [105]).

In practice, a reasonable approximation to the Bayesian estimate may well consist of simply taking the envelope (*i.e.* the union) of the 68% intervals of the probability distributions which are being combined: this typically leads to a slight overestimate of the error band, because some of the replicas in the outer 32% of one of the distribution may fall within the central 68% of another distribution, but this in practice is often a small correction, hence the envelope prescription can be a simple and effective way of combining probability distributions.

7. Conclusion

The physics of parton distributions is now close to a beginning, rather than a conclusion: LHC data for many of the processes discussed in Sec. 4 are being collected and will soon be published. The availability of LHC data is likely to change significantly our perspective on the subject. The kinematic range of "old" processes such as Drell–Yan will be extended and their

accuracy will improve. Processes which are at present not competitive will become important (such as perhaps prompt photon production). Entirely new processes will play a role, such as Higgs production. Hopefully, classes of new physical processes will be discovered: whatever their nature, they will be observed in proton collisions, and thus pose new challenges to our understanding of the nucleon. It may turn out that an interplay with new machines, such as an electron–proton LHeC collider [109], or perhaps even a neutrino factory [42] may be necessary in order to exploit fully the LHC potential. A review of this subject written ten years from now is likely to be quite different from the present one, and possibly much more interesting.

I am grateful to the organizers of the Zakopane School and especially Michał Praszałowicz for giving me the privilege to present these lectures, and I thank all participants and lecturers, in particular Paul Hoyer, for their interest and critical input. My understanding of this subject was shaped by many discussions, especially within the PDF4LHC workshop [17]: I would like to thank in particular A. de Roeck, A. Glazov, J. Huston, P. Nadolsky, J. Pumplin, R. Thorne. I also thank all the members of the NNPDF Collaboration and especially J. Rojo for innumerable discussions on the subject of these lectures. This work was partly supported by the European network HEPTOOLS under contract MRTN-CT-2006-035505.

REFERENCES

[1] G. Altarelli, *QCD: The Theory of Strong Iinteractions*, in *Landolt–Boernstein I 21A: Elementary Particles 4*.

[2] M. Klein, R. Yoshida, *Prog. Part. Nucl. Phys.* **61**, 343 (2008) [arXiv:0805.3334 [hep-ex]].

[3] M.L. Mangano, *Int. J. Mod. Phys.* **A23**, 3833 (2008) [arXiv:0802.0026 [hep-ph]].

[4] G. Arnison *et al.* [UA1 Collaboration], *Phys. Lett.* **B122**, 103 (1983); G. Arnison *et al.* [UA1 Collaboration], *Phys. Lett.* **B126**, 398 (1983); M. Banner *et al.* [UA2 Collaboration], *Phys. Lett.* **B122**, 476 (1983); P. Bagnaia *et al.* [UA2 Collaboration], *Phys. Lett.* **B129**, 130 (1983).

[5] G. Arnison *et al.* [UA1 Collaboration], *Lett. Nuovo Cim.* **44**, 1 (1985).

[6] G. Altarelli, R.K. Ellis, M. Greco, G. Martinelli, *Nucl. Phys.* **B246**, 12 (1984); G. Altarelli, R.K. Ellis, G. Martinelli, *Z. Phys.* **C27**, 617 (1985).

[7] M. Gluck, E. Hoffmann, E. Reya, *Z. Phys.* **C13**, 119 (1982).

[8] D.W. Duke, J.F. Owens, *Phys. Rev.* **D30**, 49 (1984).

[9] E. Eichten, I. Hinchliffe, K.D. Lane, C. Quigg, *Rev. Mod. Phys.* **56**, 579 (1984) [Addendum *Rev. Mod. Phys.* **58**, 1065 (1986)].

[10] W.K. Tung, arXiv:hep-ph/0409145.

[11] C.F. Berger, D. Forde, arXiv:0912.3534 [hep-ph].

[12] S. Weinzierl, arXiv:1005.1855 [hep-ph].

[13] G.P. Salam, *Eur. Phys. J.* **C67**, 637 (2010) [arXiv:0906.1833 [hep-ph]].

[14] P. Nason, *PoS* **RADCOR2009**, 018 (2010) [arXiv:1001.2747 [hep-ph]].

[15] S. Alekhin *et al.*, arXiv:hep-ph/0601012; arXiv:hep-ph/0601013.

[16] Z.J. Ajaltouni *et al.*, arXiv:0903.3861 [hep-ph].

[17] A. de Roeck, "The PDF4LHC Initiative", Sec. 6 in Ref. [82]; http://www.hep.ucl.ac.uk/pdf4lhc

[18] R.K. Ellis, W.J. Stirling, B.R. Webber, *Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol.* **8**, 1 (1996).

[19] S.J. Brodsky, P. Hoyer, C. Peterson, N. Sakai, *Phys. Lett.* **B93**, 451 (1980).

[20] M. Gluck, E. Reya, *Phys. Rev.* **D14**, 3034 (1976).

[21] P.W. Johnson, W.K. Tung, *Nucl. Phys.* **B121**, 270 (1977).

[22] S. Forte, L. Garrido, J.I. Latorre, A. Piccione, *J. High Energy Phys.* **0205**, 062 (2002) [arXiv:hep-ph/0204232].

[23] J. Feltesse, A. Glazov, V. Radescu, "Experimental Error Propagation", Sec. 3.2 in Ref. [82].

[24] V. Radescu, talk at "PDF4LHC", DESY, November 2009; A. Glazov, S. Moch, V. Radescu, arXiv:1009.6170 [hep-ph].

[25] S.J. Brodsky, G.R. Farrar, *Phys. Rev. Lett.* **31**, 1153 (1973).

[26] H.D.I. Abarbanel, M.L. Goldberger, S.B. Treiman, *Phys. Rev. Lett.* **22**, 500 (1969).

[27] P.M. Nadolsky *et al.*, *Phys. Rev.* **D78**, 013004 (2008) [arXiv:0802.0007 [hep-ph]].

[28] H.L. Lai *et al.*, *Phys. Rev.* **D82**, 074024 (2010) [arXiv:1007.2241 [hep-ph]].

[29] F.D. Aaron *et al.* [H1 and ZEUS Collaboration], *J. High Energy Phys.* **1001**, 109 (2010) [arXiv:0911.0884 [hep-ex]].

[30] A.D. Martin, W.J. Stirling, R.S. Thorne, G. Watt, *Eur. Phys. J.* **C63**, 189 (2009) [arXiv:0901.0002 [hep-ph]].

[31] G.D'Agostini, *Nucl. Instrum. Methods* **A346**, 306 (1994).

[32] R.D. Ball *et al.* [NNPDF Collaboration], *J. High Energy Phys.* **1005**, 075 (2010) [arXiv:0912.2276 [hep-ph]].

[33] G. Cowan, *Statistical Data Analysis*, Oxford, UK: Clarendon, 1998.

[34] Sec. 33 in K. Nakamura [Particle Data Group], *J. Phys. G* **37**, 075021 (2010).

[35] H.L. Lai *et al.*, *Phys. Rev.* **D82**, 054021 (2010) [arXiv:1004.4624 [hep-ph]].

[36] M. Hirai *et al.* [Asymmetry Analysis Collaboration], *Int. J. Mod. Phys.* **A18**, 1203 (2003).

[37] J.C. Collins, J. Pumplin, arXiv:hep-ph/0105207.

[38] R.D. Ball *et al.* [NNPDF Collaboration], *Nucl. Phys.* **B823**, 195 (2009) [arXiv:0906.1958 [hep-ph]].

[39] W.T. Giele, S.A. Keller, D.A. Kosower, arXiv:hep-ph/0104052.

[40] M. Ubiali, talk at "QCD at the LHC", Trento 2010; http://indico.cern.ch/materialDisplay.py?contribId=49 &sessionId=19&materialId=slides&confId=93790

[41] R.D. Ball *et al.*, *Nucl. Phys.* **B838**, 136 (2010) [arXiv:1002.4407 [hep-ph]].

[42] M.L. Mangano *et al.*, arXiv:hep-ph/0105155.

[43] S. Forte, M.L. Mangano, G. Ridolfi, *Nucl. Phys.* **B602**, 585 (2001) [arXiv:hep-ph/0101192].

[44] S.A. Kulagin, R. Petti, *Nucl. Phys.* **A765**, 126 (2006) [arXiv:hep-ph/0412425].

[45] L. Del Debbio *et al.* [NNPDF Collaboration], *J. High Energy Phys.* **0703**, 039 (2007) [arXiv:hep-ph/0701127].

[46] S.D. Ellis, W.J. Stirling, *Phys. Lett.* **B256**, 258 (1991).

[47] A. Baldit *et al.* [NA51 Collaboration], *Phys. Lett.* **B332**, 244 (1994).

[48] E.A. Hawker *et al.* [FNAL E866/NuSea Collaboration], *Phys. Rev. Lett.* **80**, 3715 (1998) [arXiv:hep-ex/9803011].

[49] E.L. Berger, F. Halzen, C.S. Kim, S. Willenbrock, *Phys. Rev.* **D40**, 83 (1989) [Erratum *Phys. Rev.* **D40**, 3789 (1989)].

[50] A.D. Martin, R.G. Roberts, W.J. Stirling, *Mod. Phys. Lett.* **A4**, 1135 (1989).

[51] F. Abe *et al.* [CDF Collaboration], *Phys. Rev. Lett.* **74**, 850 (1995) [arXiv:hep-ex/9501008].

[52] J. Rojo *et al.* [NNPDF Collaboration], arXiv:0811.2288 [hep-ph].

[53] R.D. Ball *et al.* [NNPDF Collaboration], *Nucl. Phys.* **B809**, 1 (2009) [Erratum *Nucl. Phys.* **B816**, 293 (2009)] [arXiv:0808.1231 [hep-ph]].

[54] M. Arneodo *et al.* [New Muon Collaboration], *Nucl. Phys.* **B483**, 3 (1997) [arXiv:hep-ph/9610231].

[55] M. Arneodo *et al.* [New Muon Collaboration], *Nucl. Phys.* **B487**, 3 (1997) [arXiv:hep-ex/9611022].

[56] L.W. Whitlow *et al.*, *Phys. Lett.* **B282**, 475 (1992).

[57] A.C. Benvenuti *et al.* [BCDMS Collaboration], *Phys. Lett.* **B223**, 485 (1989).

[58] F.D. Aaron *et al.* [H1 Collaboration], *Phys. Lett.* **B665**, 139 (2008) [arXiv:0805.2809 [hep-ex]].

[59] S. Chekanov *et al.* [ZEUS Collaboration], *Eur. Phys. J.* **C62**, 625 (2009) [arXiv:0901.2385 [hep-ex]].

[60] S. Chekanov *et al.* [ZEUS Collaboration], *Eur. Phys. J.* **C61**, 223 (2009) [arXiv:0812.4620 [hep-ex]].

[61] T. Aaltonen *et al.* [CDF Collaboration], *Phys. Rev. Lett.* **102**, 181801 (2009) [arXiv:0901.2169 [hep-ex]].

[62] T. Aaltonen *et al.* [CDF Collaboration], *Phys. Lett.* **B692**, 232 (2010) [arXiv:0908.3914 [hep-ex]].

[63] V.M. Abazov *et al.* [D0 Collaboration], *Phys. Rev.* **D76**, 012003 (2007) [arXiv:hep-ex/0702025].

[64] G. Moreno *et al.*, *Phys. Rev.* **D43**, 2815 (1991).

[65] J.C. Webb *et al.* [NuSea Collaboration], arXiv:hep-ex/0302019.

[66] J.C. Webb, arXiv:hep-ex/0301031.

[67] R.S. Towell *et al.* [FNAL E866/NuSea Collaboration], *Phys. Rev.* **D64**, 052002 (2001) [arXiv:hep-ex/0103030].

[68] G. Onengut *et al.* [CHORUS Collaboration], *Phys. Lett.* **B632**, 65 (2006).

[69] M. Goncharov *et al.* [NuTeV Collaboration], *Phys. Rev.* **D64**, 112006 (2001) [arXiv:hep-ex/0102049].

[70] D.A. Mason, "Measurement of the Strange–Antistrange Asymmetry at NLO in QCD from NuTeV Dimuon Data," FERMILAB-THESIS-2006-1.

[71] A. Abulencia *et al.* [CDF-Run II Collaboration], *Phys. Rev.* **D75**, 092006 (2007) [Erratum *Phys. Rev.* **D75**, 119901 (2007)] [arXiv:hep-ex/0701051].

[72] V.M. Abazov *et al.* [D0 Collaboration], *Phys. Rev. Lett.* **101**, 062001 (2008) [arXiv:0802.2400 [hep-ex]].

[73] S. Alekhin, *Eur. Phys. J.* **C10**, 395 (1999) [arXiv:hep-ph/9611213].

[74] V. Barone, C. Pascaud, F. Zomer, *Eur. Phys. J.* **C12**, 243 (2000) [arXiv:hep-ph/9907512].

[75] M. Botje, *Eur. Phys. J.* **C14**, 285 (2000) [arXiv:hep-ph/9912439].

[76] J. Pumplin *et al.*, *J. High Energy Phys.* **0207**, 012 (2002) [arXiv:hep-ph/0201195].

[77] A.D. Martin, R.G. Roberts, W.J. Stirling, R.S. Thorne, *Eur. Phys. J.* **C28**, 455 (2003) [arXiv:hep-ph/0211080].

[78] A. Djouadi, S. Ferrag, *Phys. Lett.* **B586**, 345 (2004) [arXiv:hep-ph/0310209].

[79] J. Pumplin, *Phys. Rev.* **D81**, 074010 (2010) [arXiv:0909.0268 [hep-ph]].

[80] J. Pumplin, *Phys. Rev.* **D80**, 034002 (2009) [arXiv:0904.2425 [hep-ph]].

[81] J. Pumplin, arXiv:0909.5176 [hep-ph].

[82] M. Dittmar *et al.*, arXiv:0901.2504 [hep-ph].

[83] S. Forte *et al.*, "Parton Distribution Functions", Sec. 8 in S. Dittmaier, C. Mariotti, G. Passarino, R. Tanaka, eds. *Handbook of LHC Higgs Cross-Sections*; CERN Yellow Report (in preparation).

[84] A.D. Martin, W.J. Stirling, R.S. Thorne, G. Watt, *Eur. Phys. J.* **C64**, 653 (2009) [arXiv:0905.3531 [hep-ph]].

[85] F. Demartin *et al.*, *Phys. Rev.* **D82**, 014002 (2010) [arXiv:1004.0962 [hep-ph]].

[86] D. Stump *et al.*, *J. High Energy Phys.* **0310**, 046 (2003) [arXiv:hep-ph/0303013].

[87] J.C. Collins, F. Wilczek, A. Zee, *Phys. Rev.* **D18**, 242 (1978).

[88] J.C. Collins, *Phys. Rev.* **D58**, 094002 (1998) [arXiv:hep-ph/9806259].

[89] M.A.G. Aivazis, J.C. Collins, F.I. Olness, W.K. Tung, *Phys. Rev.* **D50**, 3102 (1994) [arXiv:hep-ph/9312319].

[90] R.S. Thorne, R.G. Roberts, *Phys. Rev.* **D57**, 6871 (1998) [arXiv:hep-ph/9709442].

[91] R.S. Thorne, *Phys. Rev.* **D73**, 054019 (2006) [arXiv:hep-ph/0601245].

[92] M. Cacciari, M. Greco, P. Nason, *J. High Energy Phys.* **9805**, 007 (1998) [arXiv:hep-ph/9803400].

[93] S. Forte, E. Laenen, P. Nason, J. Rojo, *Nucl. Phys.* **B834**, 116 (2010) [arXiv:1001.2312 [hep-ph]].

[94] Sec. 22 in J.R. Andersen *et al.* [SM and NLO Multileg Working Group], arXiv:1003.1241 [hep-ph].

[95] A.D. Martin, W.J. Stirling, R.S. Thorne, G. Watt, *Eur. Phys. J.* **C70**, 51 (2010) [arXiv:1007.2624 [hep-ph]].

[96] A.M. Cooper-Sarkar, *PoS* **DIS2010**, 023 (2010) arXiv:1006.4471 [hep-ph].

[97] S. Alekhin, *Phys. Rev.* **D68**, 014002 (2003) [arXiv:hep-ph/0211096].

[98] http://projects.hepforge.org/lhapdf/; see also D. Bourilkov, R.C. Group, M.R. Whalley, arXiv:hep-ph/0605240.

[99] A.D. Martin, R.G. Roberts, W.J. Stirling, *Phys. Rev.* **D37**, 1161 (1988).

[100] J. Botts *et al.* [CTEQ Collaboration], *Phys. Lett.* **B304**, 159 (1993) [arXiv:hep-ph/9303255].

[101] J.G. Morfin, W.K. Tung, *Z. Phys.* **C52**, 13 (1991).

[102] P. Jimenez-Delgado, E. Reya, *Phys. Rev.* **D79**, 074023 (2009) [arXiv:0810.4274 [hep-ph]].

[103] M. Glück, E. Reya, *Nucl. Phys.* **B130**, 76 (1977).

[104] S. Alekhin, J. Blumlein, S. Klein, S. Moch, *Phys. Rev.* **D81**, 014032 (2010) [arXiv:0908.2766 [hep-ph]].

[105] G. Watt, presented at the "PDF4LHC" meeting, March 26, 2010 (CERN). Plots from http://projects.hepforge.org/mstwpdf/pdf4lhc/

[106] J. Anderson *et al.*, "Proton–Proton Luminosity, Standard Candles and PDFs at the LHC", Sec. 5 in Ref. [82].

[107] M. Gluck, P. Jimenez-Delgado, E. Reya, *Eur. Phys. J.* **C53**, 355 (2008) [arXiv:0709.0614 [hep-ph]].

[108] G. D'Agostini, arXiv:hep-ph/9512295.

[109] P. Newman, *Nucl. Phys. Proc. Suppl.* **191**, 307 (2009) [arXiv:0902.2292 [hep-ex]].