2.4 The global fits: MSHT, CT and NNPDF

Data, methodology and tolerance in the three main determinations, and where they differ.

1. What a global fit is

A global fit determines all the parton distribution functions (PDFs) of the proton at once, from every kind of hard-scattering measurement for which a reliable calculation exists. Chapters 2.1 to 2.3 built the parts: a parametrisation at a starting scale Q_0, a prediction $T_i(\theta)$ for each data point D_i, a χ^2 with covariance matrix C, and three ways to an uncertainty. This chapter describes how three collaborations assemble them, and how to use what they publish.

The fit has to be global because a single process measures one or two combinations of the distributions. Deep inelastic scattering (DIS) of electrons on protons measures the charge-weighted sum of quarks and antiquarks, and sees the gluon only through scaling violations. Separating u from d, quarks from antiquarks, and strange from the rest needs other targets, beams and final states.

The loss function is a sum over data sets n, each with N_n points,

$$\chi^2(\theta) = \sum_n \chi^2_n(\theta), \qquad N_{\mathrm{dat}} = \sum_n N_n, \tag{2.4.1}$$

when errors are uncorrelated between data sets. Three questions then have no unique answer, and the groups answer them differently:

1. Which data sets enter, and with what cuts.
2. What family of functions the distributions are drawn from.
3. How far the parameters may move from the best fit before the fit is called unacceptable.

Textbook statistics answers the third with $\Delta\chi^2 = 1$. Two of the three groups use a much larger number, the tolerance, and section 4 derives where a number of that size comes from.

The facts about the groups are taken from the Snowmass 2021 report [1], from Ubiali's thesis [2], from Moore's thesis [3], and from the first edition of this guide. Where those sources are silent, so is this chapter. Every formula and worked number is checked by the script linked at the end.

2. The data and what each kind constrains

The table lists the classes of measurement in a global fit. The older experiments and their roles are those of Table 2.1 of Ubiali's thesis [2]. The LHC processes are those of Table I of the Snowmass report [1].

PROCESS	EXPERIMENTS NAMED IN THE SOURCES	WHAT IT MAINLY CONSTRAINS
DIS of electrons and muons on fixed targets	SLAC, BCDMS, NMC, E665	quarks and gluon at medium and large x; proton and deuteron targets together separate u from d
DIS of neutrinos	CCFR, CHORUS, NuTeV	quarks against antiquarks; the dimuon data constrain the strange sea
DIS at the HERA collider	H1 and ZEUS, combined	quarks at small x; the gluon at small and medium x; charm and bottom through F_2^c, F_2^b
Drell–Yan pairs on fixed targets	E605, E772, E866	antiquarks; the ratio of deuteron to proton cross-sections gives $\bar{d}/\bar{u}$
W and Z production at colliders	CDF, D0; the LHC at 7, 8 and 13 TeV	u/d at large x; the flavour decomposition of the sea, including strangeness
Jets	CDF, D0, HERA; LHC inclusive jets and dijets	the gluon at medium and large x
Top quarks	LHC top-pair production	the gluon at large x

The Snowmass table adds single-top production, Z bosons at high transverse momentum, W with charm, a photon or vector boson with a jet, and Drell–Yan pairs of low mass [1].

Three remarks. First, a global fit of 2010 had about 2000 data points [2]. The project ledger records 4,618 for the published NNPDF4.0 fit at next-to-next-to-leading order (N2LO). Second, data are cut where fixed-order perturbation theory is unreliable. A typical cut removes DIS points with $Q^2 < 2$ GeV^2 or with hadronic invariant mass squared $W^2 \lesssim 10$ GeV^2 [2]; our fits use $Q^2 > 3.49$ GeV^2 and $W^2 > 12.5$ GeV^2 (chapter 1.6). Third, the data sets do not fully agree. The Snowmass report notes that χ^2/N_{dat} for the LHC data and for the full data set is above the statistical expectation in all three fits, by a similar amount, and that the disagreements between the most precise experiments are sometimes stronger than random fluctuations would give [1]. The tolerance exists because of this fact.

3. The three groups side by side

The Snowmass report compares the N2LO sets MSHT20, CT18 and NNPDF4.0, with the smaller analyses ABMP16 and ATLASpdf21 [1]. The table collects what our sources say about the first three.

	MSHT	CT	NNPDF
Set compared in [1]	MSHT20	CT18, with variants CT18A, CT18X, CT18Z	NNPDF4.0
Earlier sets	MRS, MRST, MSTW08, MMHT14	CTEQ6, CTEQ6.6, CT10, CT14	NNPDF1.0, 2.0, 3.0, 3.1
Perturbative order	N2LO	N2LO	N2LO
Heavy quarks	Thorne–Roberts general-mass scheme	S-ACOT general-mass scheme	FONLL general-mass scheme; charm fitted in NNPDF4.0
Functional form	powers of x and $1-x$ times Chebyshev polynomials in $1-2\sqrt{x}$	powers of x and $1-x$ times Bernstein polynomials in $\sqrt{x}$	neural network times powers of x and $1-x$
Free parameters	52 in the form this project uses	29 in CT18	259 in NNPDF2.0; 763 recorded for NNPDF4.0
Uncertainty	Hessian eigenvector sets	Hessian eigenvector sets	Monte Carlo replicas
Tolerance	dynamic: one per eigenvector and per sign	global, $\Delta\chi^2 \approx 100$ at 90% confidence, with a limit on each experiment	none
Minimisation	run to the minimum	run to the minimum	gradient descent, stopped early

The order and the sets are from [1]. The heavy-quark schemes and the forms are from the first edition [4], with FONLL from [3] and fitted charm from [1]. The CT18 parameter count is from the PDF4LHC21 report [8]. The tolerances are from [2] and [4]. The other parameter counts are from chapter 2.1, from [2] and from the project ledger. Chapter 1.6 treats the general-mass schemes.

Two smaller analyses mark the other end of the tolerance question. ABMP16 and the HERAPDF sets define their error bands by $\Delta\chi^2 = 1$, which the Snowmass report gives as a possible reason for the smaller ABMP16 uncertainties [1].

4. MSHT: fixed forms and the dynamic tolerance

The group and its form

MSHT20 is the current set of a line of fits that began with Martin, Roberts and Stirling and passed through MRST, MSTW08 and MMHT14. Its method is a fixed functional form fitted by minimising χ^2, with Hessian uncertainties. The form is that of chapter 2.1: for each distribution a normalisation, two powers, and a series of Chebyshev polynomials in $y = 1 - 2\sqrt{x}$, adopted in MMHT14. The Snowmass report singles out the more flexible MSHT20 parametrisation as a development since 2013, and notes that it changed the down valence distribution relative to earlier fits [1]. This project uses that form with 52 free parameters. The published MSHT20 uncertainty set is built from a subset of them, formed into 32 eigenvectors, which gives 64 members [7].

The Hessian in eigenvector coordinates

Chapter 2.3 expands χ^2 about its minimum $\hat{\theta}$. With $H_{ab} = \partial^2\chi^2/\partial\theta_a\partial\theta_b$, eigenvalues λ_k and orthonormal eigenvectors v_k,

$$\Delta\chi^2 \equiv \chi^2(\theta) - \chi^2(\hat\theta) = \tfrac{1}{2}\,\delta\theta^T H\,\delta\theta = \sum_k z_k^2, \qquad z_k = \sqrt{\lambda_k/2}\;\, v_k\cdot\delta\theta. \tag{2.4.2}$$

Each z_k measures distance along one eigenvector in units in which the global χ^2 rises by z_k^2. A Hessian set consists of the best fit and, for each k, two members displaced to $z_k = +T_k^+$ and $z_k = -T_k^-$. The numbers $T_k^\pm$ are the tolerances. If the data were mutually consistent, with Gaussian errors and an adequate parametrisation, the one-standard-deviation choice would be $T = 1$.

MRST used a single $T = \sqrt{50}$ for a 90% confidence band. MSTW08 replaced it by a procedure that sets $T_k^\pm$ separately for each eigenvector from the data themselves, and the group has used that procedure since [2, 4].

One data set along one eigenvector

Fix an eigenvector k and write t for z_k, with all other z zero. The global χ^2 is $\chi_0^2 + t^2$. An individual data set does not have its minimum at $t = 0$. To second order,

$$\chi_n^2(t) = \chi_{n,0}^2 + g_n\,t + \tfrac{1}{2}h_n\,t^2. \tag{2.4.3}$$

Here $\chi_{n,0}^2$ is the contribution of data set n at the global best fit, g_n is its slope there, and h_n its curvature. Summing Eq. (2.4.3) over data sets must reproduce $\chi_0^2 + t^2$ for every t, so

$$\sum_n g_n = 0, \qquad \sum_n h_n = 2. \tag{2.4.4}$$

The quantity $h_n/2$ is the share of the constraint on this direction that data set n supplies. Data set n alone would prefer the point

$$t_n^* = -\frac{g_n}{h_n}, \qquad \sum_n h_n\,t_n^* = 0. \tag{2.4.5}$$

The global best fit is the curvature-weighted mean of the points the data sets prefer. Consistent data sets would scatter about zero by at most about $\sqrt{2/h_n}$, one standard deviation of data set n on its own. Tension means a larger scatter.

The criterion

MSTW08 asks a different question from "by how much has the global χ^2 risen". It asks whether each data set is still described acceptably [2]. This is a hypothesis test. If the theory is right and the errors Gaussian, χ_n^2 follows the χ^2 distribution with N_n degrees of freedom. Write ξ_q for the point below which a fraction q of that distribution lies. A data set is described at 68% confidence if $\chi_n^2 < \xi_{68}$.

Applied literally, the test fails at the best fit for any data set whose $\chi_{n,0}^2$ is already above ξ_{68}, and section 2 said that many are. MSTW08 therefore rescales, so that the value at the global minimum plays the role of the centre of the distribution [2]. With the median ξ_{50} as the centre, the condition is

$$\chi_n^2(t) \;\le\; \frac{\chi_{n,0}^2}{\xi_{50}}\,\xi_{68}. \tag{2.4.6}$$

The rescaling is the same as enlarging all the errors of data set n by the factor $\sqrt{\chi_{n,0}^2/\xi_{50}}$, which makes its best-fit χ^2 equal to the median. Equation (2.4.6) allows the data set a rise of

$$\Delta_n = \chi^2_{n,0}\Big(\frac{\xi_{68}}{\xi_{50}} - 1\Big). \tag{2.4.7}$$

For a 90% band, ξ_{90} replaces ξ_{68}. Equation (2.4.6) is Eq. (58) of the MSTW08 paper [6], which states it for the 90% region and applies it in the same way at 68%. The paper calls $\xi_{50} \simeq N$ the most probable value; the value it quotes, 162.3 for a data set of 163 points, is the 50th percentile used here.

The tolerance

Setting $\chi^2_n(t) - \chi^2_{n,0} = \Delta_n$ in Eq. (2.4.3) gives a quadratic equation for t, with one root on each side:

$$t_n^{\pm} = \frac{-g_n \pm \sqrt{g_n^2 + 2h_n\Delta_n}}{h_n}. \tag{2.4.8}$$

Data set n accepts the fit for $t_n^- < t < t_n^+$. The tolerance in each sense is set by the first data set to reach its limit:

$$T_k^+ = \min_n\, t_n^+, \qquad T_k^- = \min_n\, \big|t_n^-\big|. \tag{2.4.9}$$

This is the dynamic tolerance. It differs between eigenvectors because different data sets constrain them, and between the two senses because a data set that prefers positive t objects sooner to negative t.

> **Two criteria.** $\Delta\chi^2 = 1$ answers a parameter-fitting question: given that the model and the errors are right, how well do the data determine this parameter. The dynamic tolerance answers a hypothesis-testing question: how far can the parameter move before some data set is no longer described. The first allows a rise of order 1 in χ^2. The second allows a rise of order $\sqrt{N_n}$, the width of the χ^2 distribution. That is the origin of tolerances well above 1.

How large it comes out

The size follows from the width of the χ^2 distribution. For large N the distribution is nearly Gaussian with mean N and standard deviation $\sqrt{2N}$, so $\xi_q \approx N + u_q\sqrt{2N}$, where u_q is the point below which a fraction q of a standard Gaussian lies: $u_{50} = 0$, $u_{68} = 0.468$, $u_{90} = 1.282$. Take a direction determined by a single data set that is well fitted: $h_n = 2$, $g_n = 0$, $\chi^2_{n,0} = \xi_{50}$. Then Eqs. (2.4.7) and (2.4.8) give

$$T = \sqrt{\xi_q - \xi_{50}} \;\approx\; \sqrt{u_q}\,(2N_n)^{1/4}. \tag{2.4.10}$$

With exact percentiles, $N_n = 100$ gives $T = 2.6$ at 68% and 4.4 at 90%; $N_n = 500$ gives 3.9 and 6.4. Ubiali's thesis quotes $T \sim 3$ for one-standard-deviation and $T \sim 5$ for 90% bands from the procedure in the MSTW analysis [2]. Equation (2.4.10) accounts for that size. A data set that pulls against the rest, $g_n \neq 0$, reaches its limit sooner on one side. A data set that supplies only part of the curvature, $h_n < 2$, reaches it later.

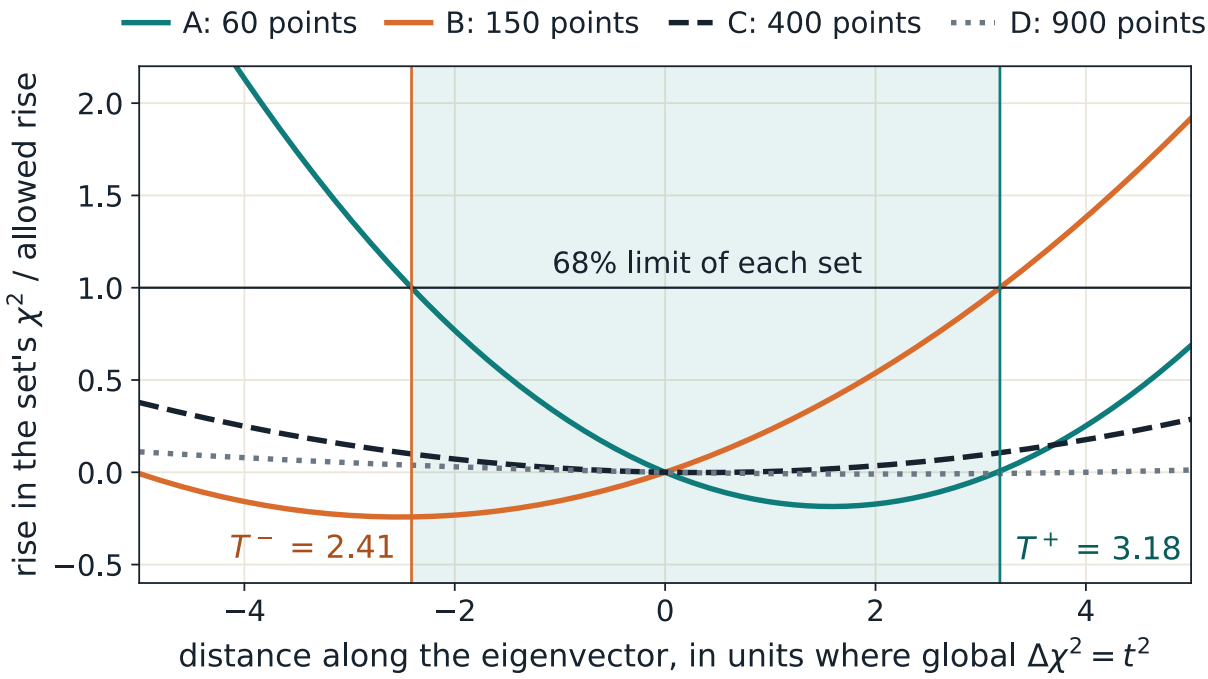

Figure 1. The dynamic tolerance in a toy with four data sets and one parameter direction (illustrative; simulated in the check script). Each curve is the rise of one data set's χ^2 divided by the rise it is allowed by Eq. (2.4.7), so every data set reaches its limit at height 1. Sets A and B pull in opposite directions. B, which prefers $t = -2.52$, is the first to object to positive t and sets $T^+ = 3.18$. A, which prefers $t = +1.58$, sets $T^- = 2.41$.

In Figure 1, set B has 150 points, $\chi^2_{B,0} = 156.7$, slope $g_B = +1.66$ and curvature $h_B = 0.660$. Its allowed rise is $\Delta_B = 8.63$, so $t_B^+ = (-1.66 + \sqrt{1.66^2 + 2 \times 0.660 \times 8.63})/0.660 = 3.18$. Set A has 60 points, $\chi^2_{A,0} = 63.9$, $g_A = -1.33$, $h_A = 0.840$ and $\Delta_A = 5.64$, which gives $t_A^- = -2.41$. Derivation D1 gives the general statement and D2 the percentiles.

5. CT: a global tolerance, two tiers and Lagrange multipliers

The group and its form

CT18 is the current set of the CTEQ line, which runs through CTEQ6, CTEQ6.6, CT10 and CT14. CTEQ6 was the first of these with published Hessian error sets, and the general-mass treatment of heavy quarks became the default with CTEQ6.5 and CTEQ6.6. The form is again a small-x power and a large-x power times a polynomial, here built from Bernstein polynomials $B_{k,n}(y) = \binom{n}{k} y^k (1-y)^{n-k}$ in $y = \sqrt{x}$.

CT addresses the dependence on the form directly. The CT18 analysis studied more than 250 alternative functional forms to estimate the part of the uncertainty that comes from the parametrisation [1]. It is published as a family. CT18 is the default. CT18A adds the ATLAS 7 TeV W and Z data, which the default fit leaves out because every recent fit describes them with too high a χ^2. CT18X uses a factorisation scale that depends on x in DIS. CT18Z combines these changes [1].

The global tolerance

The CTEQ study of the tolerance is described in Ubiali's thesis [2]. A fit with about 2000 points from many sources has theoretical uncertainties that are hard to quantify and experimental errors that may not be Gaussian. Along each eigenvector the group estimated the range of global χ^2 over which every data set is still fitted at 90% confidence, and averaged the ranges over eigenvectors. The result was $T \approx 10$, that is $\Delta\chi^2 \approx 100$, for a 90% band.

A tolerance can be read as an enlargement of the experimental errors. If every error is multiplied by s, χ^2 is divided by s^2. A one-parameter 90% interval is $\Delta\chi^2 = 1.645^2 = 2.71$ in the rescaled χ^2, which is $\Delta\chi^2 = 2.71\,s^2$ in the original one. So

$$T^2 = 2.71\,s^2, \qquad s = \frac{T}{1.645}. \tag{2.4.11}$$

$T = 10$ corresponds to $s = 6.1$ and $T = \sqrt{50}$ to $s = 4.3$. This is the statement in [2] that the tolerance is equivalent to multiplying experimental errors by a factor between four and six. The same thesis records the objection: a factor that large may inflate the uncertainty in regions where the data do determine the distributions, and no rigorous statistical argument fixes it.

The second tier

A global $\Delta\chi^2$ can be reached by spoiling the fit to one experiment while the others barely change. CT therefore adds a second condition. With P_n a penalty that is zero while data set n is inside its own 90% limit and grows quickly outside it, the error sets satisfy

$$\Delta\chi^2(\theta) + \sum_n P_n(\theta) \;\le\; T^2. \tag{2.4.12}$$

Tier one is the global ball of radius T in the coordinates of Eq. (2.4.2). Tier two stops the displacement along an eigenvector as soon as one experiment leaves its limit. Whichever binds first sets the member. The second tier is close to Eq. (2.4.9). The difference is that CT keeps a global upper bound and MSHT has none.

Lagrange-multiplier scans

The Hessian method assumes that χ^2 is quadratic and the observable linear out to the tolerance. The Lagrange-multiplier method, developed by CTEQ [4], removes both assumptions for one chosen observable $X(\theta)$, for example a cross-section. Minimise

$$\Psi(\theta;\lambda) = \chi^2(\theta) + \lambda\,X(\theta), \qquad \text{so that} \qquad \nabla\chi^2 = -\lambda\,\nabla X \tag{2.4.13}$$

at the minimum. That condition says that χ^2 is stationary on the surface of constant X: the fit found is the best one with X held at whatever value λ drives it to. Repeating for a ladder of λ traces the curve of χ^2 against X with all other directions re-fitted at every point. The uncertainty on X is the range over which that curve stays within the tolerance.

In the quadratic and linear approximation the scan can be done in closed form. Write $X = X_0 + \sum_k X_k z_k$ with $X_k = \partial X/\partial z_k$. Then (derivation D3)

$$z_k = -\tfrac{1}{2}\lambda X_k, \qquad \Delta X = -\tfrac{1}{2}\lambda \sum_k X_k^2, \qquad \Delta\chi^2 = \frac{(\Delta X)^2}{\sum_k X_k^2}. \tag{2.4.14}$$

The profile is a parabola, and at $\Delta\chi^2 = T^2$

$$\Delta X = T\Big(\sum_k X_k^2\Big)^{1/2} = \tfrac{1}{2}\Big(\sum_k \big[X(S_k^+) - X(S_k^-)\big]^2\Big)^{1/2}, \tag{2.4.15}$$

where $S_k^\pm$ are the Hessian members at $z_k = \pm T$. The scan and the Hessian formula agree exactly when the approximation holds. A scan is therefore a test of the Hessian set.

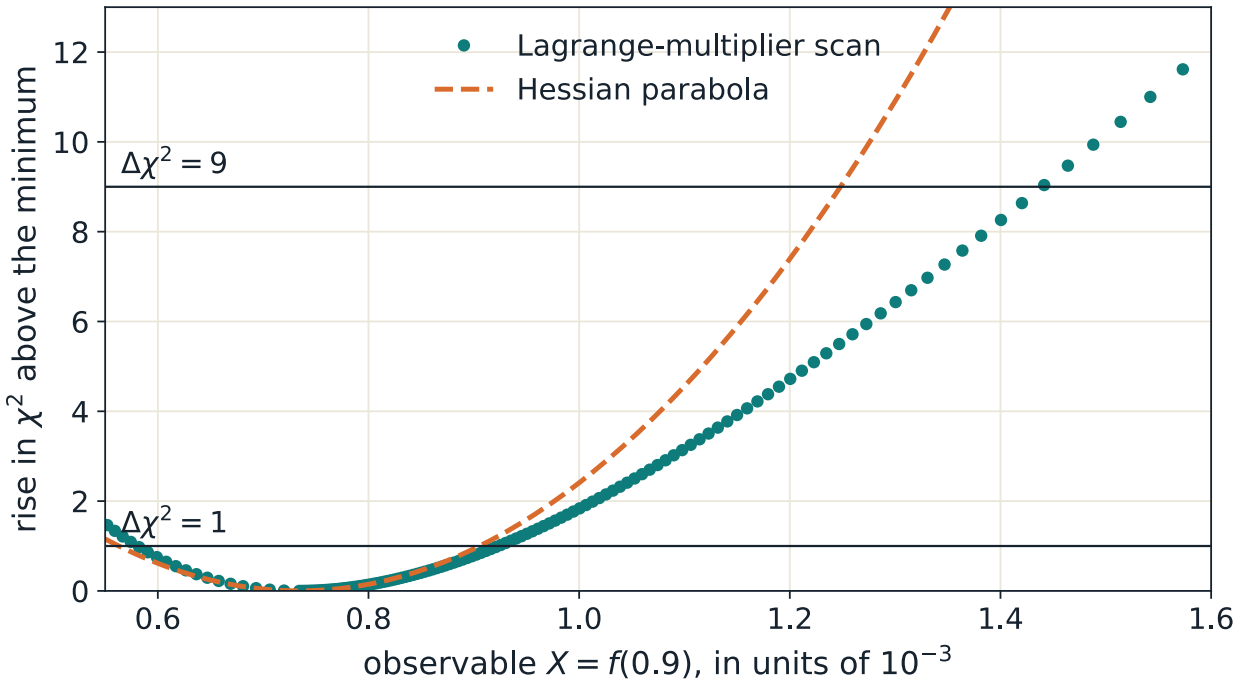

Figure 2. A Lagrange-multiplier scan in a toy (illustrative; computed in the check script). A form $A(1-x)^\eta$ is fitted to ten pseudo-data points at $x \le 0.6$ with 10% errors, and the observable is its extrapolation to $x = 0.9$. Points: the scan of Eq. (2.4.13). Dashed: the parabola of Eq. (2.4.14) from the Hessian. They agree near the minimum and separate as the tolerance grows.

In the toy of Figure 2 the best fit gives $X = 0.734 \times 10^{-3}$. At $\Delta\chi^2 = 1$ the symmetric Hessian error is ± 0.171 and the scan gives $+0.189,\ -0.154$, in the same units of 10^{-3}. At $\Delta\chi^2 = 9$ the Hessian error is ± 0.514 and the scan gives $+0.705,\ -0.379$. The extrapolated value depends exponentially on η, so equal steps in the parameter are unequal steps in X. A larger tolerance probes more of that non-linearity. A scan needs a set of full fits for each observable, and its result cannot be handed to a user as a set of members [2], so Hessian sets are what is published and scans are the check on them.

6. NNPDF: networks, replicas and stopping

What is different

NNPDF changes the answers to questions 2 and 3 of section 1 together. The distributions at Q_0 are a neural network times powers of x and $1-x$ (chapter 2.1). In NNPDF2.0 there were seven networks of architecture 2–5–3–1 with 37 parameters each, 259 in all [2]. NNPDF4.0 uses one network with an output for each distribution [1], and unlike the other two groups it fits the charm distribution at Q_0 as well as generating charm from gluon radiation [1]. The Hessian method is not used. The uncertainty is carried by Monte Carlo replicas (chapter 2.3):

$$D^{(k)} \sim \mathcal{N}(D, C), \qquad f^{(k)} = \text{the fit to } D^{(k)}, \qquad k = 1, \ldots, N_{\text{rep}}. \tag{2.4.16}$$

Each replica of the data is fitted separately, and the ensemble of fitted distributions is the result. No tolerance is introduced [2].

Why the minimum is not the answer

A form with a few tens of parameters cannot follow the noise in several thousand points. A network with hundreds can. Its best fit should have the χ^2 expected from the fluctuations of the data; below that, χ^2 improves only because noise is being fitted [2].

NNPDF finds that point by cross-validation [2]. For each replica the data are divided at random into a training set and a validation set. In NNPDF2.0 the division was half and half, made separately within each data set so that every kinematic region appears in both, and made afresh for each replica. The minimiser sees only the training χ^2. The validation χ^2 is monitored. While the fit is learning the underlying function both fall. When the fit starts to learn the fluctuations of the training points, the validation χ^2 stops falling and rises, and the fit is stopped there.

Stopping as a filter

The effect of stopping can be derived for a model that is linear in its parameters. Let $\chi^2_{\text{tr}} = |J\theta - d|^2$, where d holds the training data divided by their errors and J the corresponding derivatives of the predictions, and let λ_i be the eigenvalues of $J^T J$. In the notation of chapter 2.3, $J^T J$ is the Fisher matrix F of the training set and the Hessian is $2F$. Gradient descent from $\theta = 0$ with step $\theta \to \theta - \frac{1}{2}\eta \nabla \chi^2_{\text{tr}}$ gives, after t steps and along eigenvector i (derivation D4),

$$\theta_i(t) = \hat{\theta}_i \left[1 - (1 - \eta\lambda_i)^t\right], \tag{2.4.17}$$

where $\hat{\theta}_i$ is the component of the least-squares solution. Directions with $\eta\lambda_i t \gg 1$ have reached their fitted values, and those with $\eta\lambda_i t \ll 1$ have not moved. Stopping at step t fits the combinations that the data determine well and leaves the rest at their starting values. The number of directions fitted is

$$N_{\text{eff}}(t) = \sum_i \left[1 - (1 - \eta\lambda_i)^t\right], \tag{2.4.18}$$

which grows from zero to the number of parameters. The stopping point, not the architecture, sets the effective number of parameters. Chapter 2.3 gives the continuous-time form of this result and the variance of a replica ensemble stopped in this way.

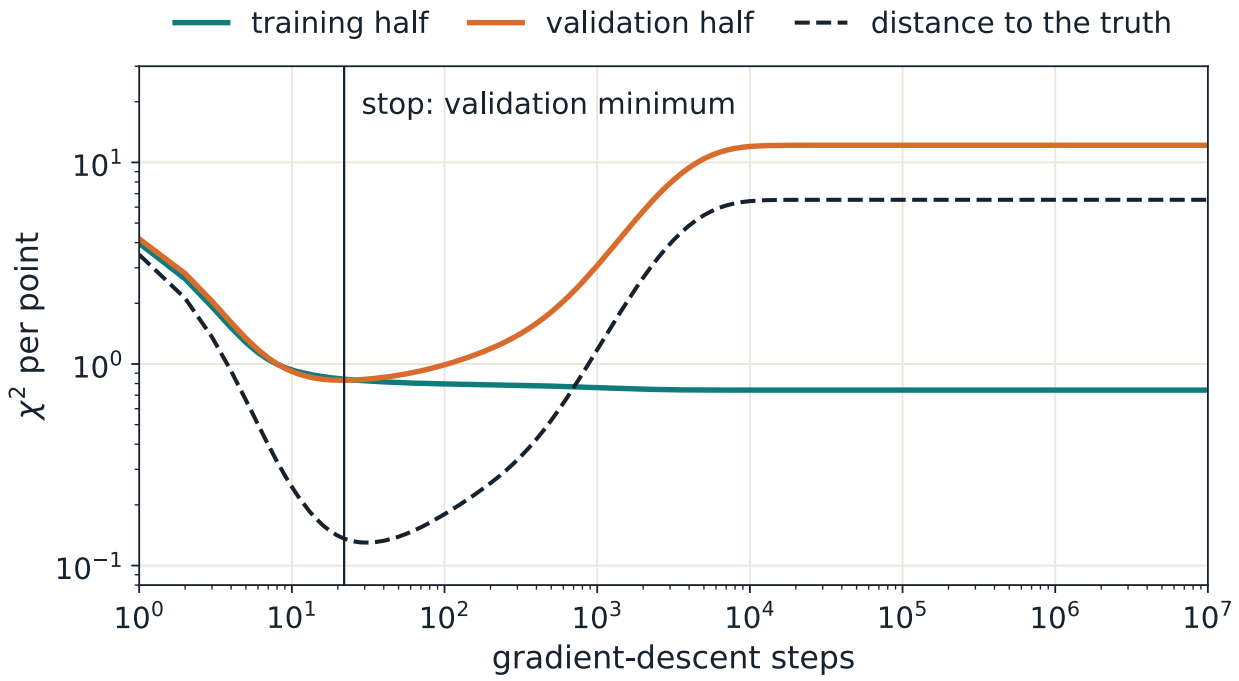

Figure 3. Early stopping in a toy (illustrative; computed in the check script). A 16-parameter polynomial is fitted by gradient descent to 40 training points; 40 more are held back. The training χ^2 per point only falls. The validation χ^2 has a minimum at step 22 and then rises. Dashed: the mean squared distance from the true function in units of the data error, which a real fit cannot know; its minimum lies close to the validation minimum.

In the toy of Figure 3 the validation minimum is at step 22, where χ^2 per point is 0.84 on the training half and 0.83 on the validation half, and $N_{\text{eff}} = 10.7$ of 16. Run to convergence, the training value falls to 0.74 and the validation value rises to 12.2. The distance to the truth is 0.136 at the stopping point and 6.5 at convergence.

Until NNPDF3.1 the minimiser was a genetic algorithm, and the stopping rule used moving averages of the two curves [2]. NNPDF4.0 uses gradient descent [1], with the same cross-validation [3].

Hyperoptimisation and closure tests

The architecture, the optimiser and its settings, and the stopping rule are choices. In earlier releases they were made by trial and error. NNPDF4.0 selects them by an automatic scan [1]. The method as a whole is tested by closure tests, used since NNPDF3.0 [1]. Pseudo-data are generated from a known set of distributions, the full procedure is run on them, and the output is compared with the known input. At level 0 the pseudo-data have no noise, and the fit should reproduce the input. At level 1 one layer of noise is added and a single fit is made. At level 2 the replica procedure is run on the level-1 data, and the test is whether the spread of the replicas covers the truth as often as it claims to. Chapters 2.3 and 2.5 give the estimators.

The NNPDF fitting code, data and theory tables are public [1]. This project runs on them.

7. Where the sets agree and where they differ

This section summarises the comparison of N2LO sets at $Q = 100$ GeV in the Snowmass report [1].

Agreement. The parton luminosities for producing a state of mass m_X at 14 TeV agree within uncertainties near $m_X \sim 100$ GeV, the region of W, Z and Higgs production. The quark–quark and quark–antiquark luminosities agree within two standard deviations over most of the mass range.

Strangeness. MSHT20 and NNPDF4.0 agree with each other and lie above CT18. The cause is a choice of data: the ATLAS 7 TeV W and Z measurement raises the strange distribution at intermediate and large x, and CT18 excludes it. CT18A and CT18Z, which include it, have larger strangeness.

Charm. NNPDF4.0 differs from the others, which agree among themselves. The cause is a choice of parametrisation: the fitted charm distribution, which in dedicated fits carries between 0 and 1% of the proton's momentum, raises charm at large x.

Valence quarks. The up valence distributions agree between $x \sim 10^{-2}$ and 0.5. Below that, where data are weak, the shapes differ. For down valence, CT18 has a different shape from MSHT20 and NNPDF4.0, connected in part to its lower strangeness.

The gluon at large mass. Near $m_X \sim 1$ TeV the gluon–gluon and gluon–quark luminosities of NNPDF4.0 are below those of MSHT20 and CT18. The report names both data and method as possible causes. NNPDF4.0 includes dijet data at 7 TeV and differential top-pair distributions from LHC Run II that the others do not.

Size of the uncertainty. NNPDF4.0 generally has the smallest uncertainty and CT18 the largest of the three. The benchmark exercise behind the PDF4LHC21 combination fitted a common reduced data set with each group's method. The differences that remained, from the parametrisation, the tolerance and the fitting method, were in some cases as large as the uncertainties themselves, or larger [1].

Each difference has one of three origins: data, form, or the definition of the uncertainty. Between published sets all three change at once.

---

8. Combining sets: PDF4LHC

Why and what

The PDF4LHC working group, established in 2008, issues recommendations and combined sets [1]. PDF4LHC15 combined CT14, MMHT2014 and NNPDF3.0. PDF4LHC21 combines CT18, MSHT20 and NNPDF3.1, after the benchmark comparison of section 7. For the most precise measurements the report still asks for comparison with the individual sets.

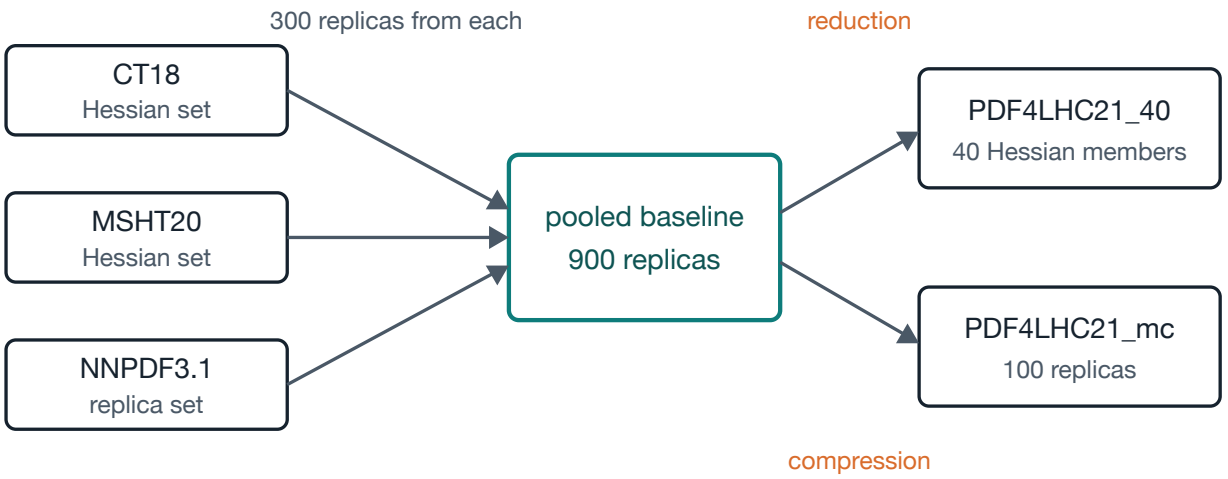

Figure 4. The PDF4LHC21 combination as described in the Snowmass report [1]. The two Hessian sets are converted to replicas by Eq. (2.4.19). Three hundred replicas from each group are pooled, Eq. (2.4.21). The pool of 900 is then reduced in two ways: to a Hessian set of 40 members, Eq. (2.4.24), and to a compressed set of 100 replicas.

From a Hessian set to replicas

A Hessian set states a Gaussian distribution in the coordinates z_k. To sample it, draw independent standard Gaussian numbers r_{jk} and form

$$f^{(k)}(x) = f_0(x) + \sum_j \tfrac{1}{2}\left[f_j^+(x) - f_j^-(x)\right] r_{jk}. \tag{2.4.19}$$

Here f_0 is the central member and $f_j^\pm$ the pair for eigenvector j at one standard deviation. The replicas have mean f_0 and covariance (derivation D5)

$$\left\langle \delta f(x)\, \delta f(x') \right\rangle = \sum_j \tfrac{1}{4}\left[f_j^+(x) - f_j^-(x)\right]\left[f_j^+(x') - f_j^-(x')\right], \tag{2.4.20}$$

which is the covariance the Hessian master formula of section 9 assigns. The tolerance is carried along: it is built into the positions of $f_j^\pm$.

Pooling

With every set in replica form, the combination is the union of equal numbers of replicas from each of G groups. Let group g have mean μ_g and standard deviation σ_g for some quantity. The pooled mean and variance are (derivation D5)

$$\mu = \frac{1}{G}\sum_g \mu_g, \qquad \sigma^2 = \frac{1}{G}\sum_g \sigma_g^2 + \frac{1}{G}\sum_g (\mu_g - \mu)^2. \tag{2.4.21}$$

The pooled variance is the average of the groups' variances plus the variance of their central values. Disagreement between groups becomes part of the uncertainty. For example, three groups with 1.000 ± 0.020, 1.030 ± 0.015 and 0.985 ± 0.025 pool to 1.0050 ± 0.0277: the root-mean-square of the three uncertainties is 0.0204 and the spread of the central values adds 0.0187 in quadrature. The pooled distribution is a mixture and is not Gaussian. In this example 66.1% of it lies within one standard deviation of the mean.

PDF4LHC21 took $N_{\rm rep} = 300$ from each group, 900 in all [1].

From replicas back to a Hessian set

Nine hundred members are too many for routine use. Sample the replicas on a grid of points x_i, and collect the deviations from the mean in a matrix with one column per replica:

$$\mathcal{D}_{ik} = \frac{f^{(k)}(x_i) - \langle f(x_i)\rangle}{\sqrt{N_{\rm rep}-1}}, \qquad \mathrm{cov}\big(f(x_i), f(x_{i'})\big) = \big(\mathcal{D}\mathcal{D}^T\big)_{ii'}. \tag{2.4.22}$$

The singular value decomposition $\mathcal{D} = USV^T$ writes the covariance as US^2U^T. Define one member per singular value s_j:

$$f_j(x_i) = \langle f(x_i)\rangle + s_j\, U_{ij} = \langle f(x_i)\rangle + \sum_k \mathcal{D}_{ik} V_{kj}. \tag{2.4.23}$$

The second form shows that each member is the mean plus a linear combination of the original replicas. For any quantity X linear in f, keeping the $N_{\rm eig}$ largest singular values (derivation D6),

$$\Delta X = \Big(\sum_{j=1}^{N_{\rm eig}} \big[X(f_j) - X(\langle f\rangle)\big]^2\Big)^{1/2}, \qquad \text{missing variance} = \sum_{j>N_{\rm eig}} s_j^2. \tag{2.4.24}$$

This is the idea of the mc2hessian method. The other method considered for PDF4LHC21, META-PDF, first refits every replica with one common functional form built on Bernstein polynomials and then applies the same principal-component reduction to the parameters of that form. The published Hessian set, PDF4LHC21_40, has 40 members and comes from the META-PDF route [1].

In a toy pool of three groups of 300 replicas on 40 grid points (check script), five members keep 99.8% of the total variance but give only 51% of the uncertainty at the worst grid point. Ten give 91% there and twenty give all of it. The total is dominated by the directions in which the groups differ, so a reduction has to be judged point by point.

The second published set, PDF4LHC21_mc, has 100 replicas. It is made by a compression algorithm that selects the subset of replicas whose statistical properties are closest to those of the full 900 [1].

How many replicas

The estimates from $N_{\rm rep}$ replicas have sampling errors of their own. For a Gaussian quantity with standard deviation σ (derivation D6),

$$\delta\langle X\rangle = \frac{\sigma}{\sqrt{N_{\rm rep}}}, \qquad \frac{\delta\sigma}{\sigma} \approx \frac{1}{\sqrt{2(N_{\rm rep}-1)}}. \tag{2.4.25}$$

With 100 replicas the central value is known to a tenth of the uncertainty and the uncertainty itself to 7%. With 900 the figures are 3.3% and 2.4%.

---

9. How to read a published set

Sets are distributed through the LHAPDF library, in version 6 since 2013 and holding more than 1150 sets [1]. A set is a list of members. Each member is a complete set of distributions tabulated in x and Q^2. Member 0 is the central

one. What the others mean depends on the kind of set.

The rule common to all kinds: compute the observable separately with each member, and combine the results only afterwards. Combining the members first and computing once gives a different answer for anything not linear in the distributions, and no uncertainty.

Hessian sets

Members come in pairs $S_k^{\pm}$, one pair per eigenvector. Chapter 2.3 derives the uncertainty on an observable X with central value X_0:

$$\Delta X = \tfrac{1}{2}\Big(\sum_k \big[X(S_k^+) - X(S_k^-)\big]^2\Big)^{1/2}. \tag{2.4.26}$$

When the pair is not symmetric about X_0, the asymmetric form adds, for each eigenvector, the larger displacement in each sense:

$$\begin{aligned} \Delta X^+ &= \Big(\sum_k \big[\max\big(X_k^+ - X_0,\; X_k^- - X_0,\; 0\big)\big]^2\Big)^{1/2}, \\ \Delta X^- &= \Big(\sum_k \big[\max\big(X_0 - X_k^+,\; X_0 - X_k^-,\; 0\big)\big]^2\Big)^{1/2}, \end{aligned} \tag{2.4.27}$$

with $X_k^{\pm} = X(S_k^{\pm})$. Take four eigenvectors with $X_0 = 50.0$ and pairs $(50.9, 49.2)$, $(49.6, 50.5)$, $(50.3, 50.1)$, $(50.05, 49.90)$. Equation (2.4.26) gives $\Delta X = 0.970$. Equation (2.4.27) gives $+1.074,\ -0.900$. In the third pair both members lie above X_0: the larger counts upwards and the pair contributes nothing downwards.

The correlation between two observables follows from the same members:

$$\cos\varphi = \frac{1}{4\,\Delta X\,\Delta Y}\sum_k \big[X_k^+ - X_k^-\big]\big[Y_k^+ - Y_k^-\big]. \tag{2.4.28}$$

It matters for ratios. Add $Y_0 = 20.0$ with pairs $(20.5, 19.6)$, $(19.9, 20.2)$, $(20.1, 19.95)$, $(20.0, 20.02)$, so $\Delta Y = 0.480$ and $\cos\varphi = 0.980$. Computing X/Y member by member gives an uncertainty of 0.0156 on a ratio of 2.5. Combining the two relative errors as if independent would give 0.0772, five times too large.

A set produced by a reduction like Eq. (2.4.23) has one member per eigenvector, and Eq. (2.4.24) replaces Eq. (2.4.26).

Monte Carlo sets

Members 1 to N_{rep} are replicas, and member 0 is their average. The central value and uncertainty are the mean and standard deviation over replicas:

$$\langle X\rangle = \frac{1}{N_{\mathrm{rep}}}\sum_{k=1}^{N_{\mathrm{rep}}} X\big(f^{(k)}\big), \qquad \Delta X = \Big(\frac{1}{N_{\mathrm{rep}}-1}\sum_{k=1}^{N_{\mathrm{rep}}}\big[X\big(f^{(k)}\big) - \langle X\rangle\big]^2\Big)^{1/2}. \tag{2.4.29}$$

Ten replica values 50.4, 49.1, 50.9, 49.8, 50.2, 51.3, 49.5, 50.0, 48.9, 50.6 give $\langle X\rangle = 50.07$ and $\Delta X = 0.769$. Dividing by N_{rep} instead, as in [3], gives 0.729; the difference is below the sampling error of Eq. (2.4.25). When the distribution of X is not Gaussian, the interval between the 16th and 84th percentiles of the replica values is the better statement of a 68% band [3].

Confidence levels

A set must be read at the confidence level it was built for. Sets built at 90% are converted to one standard deviation by the Gaussian factor of Eq. (2.4.11):

$$\Delta X_{68} = \frac{\Delta X_{90}}{1.645}. \tag{2.4.30}$$

If the example above were a 90% set, the 68% uncertainty would be $0.970/1.645 = 0.590$.

All sets propagate the errors of the fitted data. Some also include contributions from the choice of data, method or theory [1]. The spread between groups is in no single set, which is the reason for section 8.

---

10. In this project

Of the three origins of difference in section 7, this project holds data and form fixed and varies the definition of the uncertainty.

The data and theory are NNPDF's. The project ledger [5] records a global selection of 82 data sets and 4,615 points, three fewer than the 4,618 of the published NNPDF4.0 N2LO fit. The form is MSHT20's, with 52 parameters, at $Q_0 = 1.65$ GeV, $\alpha_S(M_Z) = 0.118$ and NNPDF theory 40000000 (chapters 1.6 and 2.1). The minimum of that fit is $\chi^2/N_{\mathrm{dat}} = 1.2368$ (result PPDF-31), which the ledger sets beside the published NNPDF4.0 value of 1.17. A repeat of the fit on different hardware gave 1.2380 (PPDF-35). One mismatch between theory and form is deliberate and recorded in the ledger: the theory tables allow a fitted charm distribution, as in NNPDF4.0, while the port sets charm to zero at Q_0 (chapter 2.1).

On that single fit the uncertainty is then computed by the Hessian method, by Monte Carlo replicas and by Bayesian sampling, all in the Colibri code (chapter 2.5). Chapter 3.4 reports the comparison.

A Hessian band with $T = 1$ and a replica band agree when the model is linear and the data consistent (chapter 2.3). The global data are not fully consistent, and sections 4 and 5 showed how two groups respond. The project's state file lists as an open experiment the measurement of the tolerance T^2 that makes Hessian bands cover a known input in a closure test. Until that is done, no tolerance of our own is quoted.

Derivations

D1. A data set along an eigenvector

Derives (2.4.4), (2.4.5), (2.4.8)

The expansion. Along eigenvector k the parameters are $\theta(t) = \hat{\theta} + t\sqrt{2/\lambda_k}\, v_k$, from Eq. (2.4.2). Each χ^2_n is a smooth function of θ, so $\chi^2_n(t) = \chi^2_{n,0} + g_n t + \tfrac{1}{2} h_n t^2 + O(t^3)$ with $g_n = d\chi^2_n/dt$ and $h_n = d^2\chi^2_n/dt^2$ at $t = 0$.

The sums. By Eq. (2.4.1), $\sum_n \chi^2_n(t) = \chi^2(\theta(t)) = \chi^2_0 + t^2$. Comparing powers of t gives $\sum_n g_n = 0$ and $\tfrac{1}{2}\sum_n h_n = 1$, which is Eq. (2.4.4). The first is the statement that the global χ^2 is stationary at $\hat{\theta}$; the second fixes the unit of t.

The preferred points. $\chi^2_n(t)$ is smallest at $t^*_n = -g_n/h_n$. Then $\sum_n h_n t^*_n = -\sum_n g_n = 0$, Eq. (2.4.5).

The crossing. The condition $\chi^2_n(t) - \chi^2_{n,0} = \Delta_n$ is $\tfrac{1}{2}h_n t^2 + g_n t - \Delta_n = 0$. Its roots are Eq. (2.4.8). Since $h_n > 0$ and $\Delta_n > 0$ the square root exceeds $|g_n|$, so one root is positive and one negative. In terms of the preferred point, $t^{\pm}_n =$

$t_n^* \pm \sqrt{t_n^{*2} + 2\Delta_n/h_n}$: the allowed interval is centred on t_n^*, so a data set with $t_n^* < 0$ has $t_n^+ < |t_n^-|$ and objects first to positive t.

D2. Percentiles of the χ^2 distribution and the size of the tolerance

Derives (2.4.10), (2.4.11)

The distribution. A sum of squares of N independent standard Gaussian numbers has mean N and variance $2N$. For large N it is nearly Gaussian by the central limit theorem, so the point below which a fraction q lies is $\xi_q \approx N + u_q\sqrt{2N}$, with u_q the corresponding point of the standard Gaussian.

The single-set tolerance. With $h_n = 2$ and $g_n = 0$, Eq. (2.4.8) is $t^\pm = \pm\sqrt{\Delta_n}$. With $\chi^2_{n,0} = \xi_{50}$, Eq. (2.4.7) is $\Delta_n = \xi_q - \xi_{50} \approx u_q\sqrt{2N}$. So $T = \sqrt{u_q}\,(2N)^{1/4}$, Eq. (2.4.10). It grows as the fourth root of the number of points.

Numbers. For $N = 100$ the exact values are $\xi_{50} = 99.33$, $\xi_{68} = 106.07$ and $\xi_{90} = 118.50$, giving $T = 2.60$ and 4.38; the Gaussian estimate gives 2.57 and 4.26.

The error-inflation reading. Multiply every experimental error by s. The covariance matrix scales by s^2 and χ^2 by $1/s^2$, so a rise $\Delta\chi^2$ in the original χ^2 is a rise $\Delta\chi^2/s^2$ in the rescaled one. For one parameter with Gaussian errors, a two-sided 90% interval is ± 1.645 standard deviations, that is $\Delta\chi^2 = 1.645^2 = 2.71$ in the rescaled χ^2. Setting $T^2/s^2 = 2.71$ gives Eq. (2.4.11).

D3. The Lagrange-multiplier scan in the quadratic approximation

Derives (2.4.14), (2.4.15)

Set-up. In the coordinates of Eq. (2.4.2), $\chi^2 = \chi_0^2 + \sum_k z_k^2$. Let $X = X_0 + \sum_k X_k z_k$ be linear.

Minimise. $\Psi = \chi_0^2 + \sum_k z_k^2 + \lambda X_0 + \lambda \sum_k X_k z_k$. Setting $\partial\Psi/\partial z_k = 0$ gives $2z_k + \lambda X_k = 0$, so $z_k = -\tfrac{1}{2}\lambda X_k$. The fit moves along the gradient of X in these coordinates.

Read off. $\Delta X = \sum_k X_k z_k = -\tfrac{1}{2}\lambda \sum_k X_k^2$ and $\Delta\chi^2 = \sum_k z_k^2 = \tfrac{1}{4}\lambda^2 \sum_k X_k^2$. Eliminating λ gives $\Delta\chi^2 = (\Delta X)^2 / \sum_k X_k^2$, Eq. (2.4.14).

Compare with the members. The Hessian members are at $z_k = \pm T$ with the other coordinates zero, so $X(S_k^\pm) = X_0 \pm T X_k$ and $X(S_k^+) - X(S_k^-) = 2T X_k$. Then $\tfrac{1}{2}\big(\sum_k 4T^2 X_k^2\big)^{1/2} = T\big(\sum_k X_k^2\big)^{1/2}$, which is the value of ΔX at $\Delta\chi^2 = T^2$. This is Eq. (2.4.15).

The check script confirms each derivation in this section numerically: the sums and crossings of D1 on the toy of Figure 1, the relations of D3 on a random six-parameter quadratic, Eq. (2.4.17) against explicit gradient descent, and the identities of D5 and D6 by simulation.

D4. Gradient descent as a filter

Derives (2.4.17), (2.4.18)

The step. $\chi^2_{\rm tr} = |J\theta - d|^2$ has gradient $2J^T(J\theta - d)$. One step is $\theta \to \theta - \eta\,(J^T J\,\theta - J^T d)$.

Diagonalise. Let $J^T J\, w_i = \lambda_i w_i$ and write $\theta = \sum_i \theta_i w_i$, $J^T d = \sum_i c_i w_i$. The components decouple: $\theta_i \to (1 - \eta\lambda_i)\theta_i + \eta c_i$. The fixed point is $\hat\theta_i = c_i/\lambda_i$, the least-squares solution.

Solve. The distance from the fixed point obeys $\theta_i - \hat\theta_i \to (1 - \eta\lambda_i)(\theta_i - \hat\theta_i)$. Starting from $\theta_i = 0$, after t steps $\theta_i(t) - \hat\theta_i = -(1 - \eta\lambda_i)^t\,\hat\theta_i$, which is Eq. (2.4.17). The iteration converges if $\eta\lambda_i < 2$ for every i.

Count. The factor $1-(1-\eta\lambda_i)^t$ is the fraction of direction i that has been fitted. Summing over i gives Eq. (2.4.18). A small eigenvalue λ_i means that the data determine that combination poorly: its fitted value $\hat{\theta}_i = c_i/\lambda_i$ is the noise in c_i divided by a small number. Those are the components that stopping withholds.

D5. Converting and pooling sets

Derives (2.4.20), (2.4.21)

Conversion. In Eq. (2.4.19) write $\delta_j(x) = \frac{1}{2}[f_j^+(x) - f_j^-(x)]$. The r_{jk} have zero mean, so $\langle f^{(k)}\rangle = f_0$. They are independent with unit variance, $\langle r_{jk}r_{j'k}\rangle = \delta_{jj'}$, so $\langle \delta f(x)\delta f(x')\rangle = \sum_{jj'}\delta_j(x)\delta_{j'}(x')\langle r_j r_{j'}\rangle = \sum_j \delta_j(x)\delta_j(x')$. This is Eq. (2.4.20). For a quantity linear in f the variance is $\sum_j \frac{1}{4}[X_j^+ - X_j^-]^2$, the square of Eq. (2.4.26).

Pooling. Take n replicas from each of G groups. The mean of all Gn values is the mean of the group means, $\mu = \frac{1}{G}\sum_g \mu_g$. For the variance, write each value as $X-\mu = (X-\mu_g)+(\mu_g-\mu)$ and square. Within group g the cross term averages to zero because $\langle X-\mu_g\rangle_g = 0$, leaving $\sigma_g^2 + (\mu_g-\mu)^2$. Averaging over groups gives Eq. (2.4.21). This is the law of total variance.

D6. Replicas to a Hessian set, and sampling errors

Derives (2.4.23), (2.4.24), (2.4.25)

Covariance. By definition the sample covariance is $\frac{1}{N_{\text{rep}}-1}\sum_k \delta f^{(k)}(x_i)\,\delta f^{(k)}(x_{i'}) = (\mathcal{D}\mathcal{D}^T)_{ii'}$, Eq. (2.4.22).

Decompose. With $\mathcal{D} = USV^T$, where $U^TU = V^TV = 1$ and S is diagonal, $\mathcal{D}\mathcal{D}^T = US^2U^T = \sum_j s_j^2\, u_j u_j^T$, where u_j is column j of U. Multiplying $\mathcal{D} = USV^T$ on the right by V gives $\mathcal{D}V = US$, the second form of Eq. (2.4.23).

The formula. Let $X = \sum_i a_i f(x_i)$. Its variance is $a^T\mathcal{D}\mathcal{D}^T a = \sum_j s_j^2 (a\cdot u_j)^2$. The member f_j gives $X(f_j) - X(\langle f\rangle) = s_j\, a\cdot u_j$. So the sum of squares over all members is the variance, and truncating at N_{eig} omits $\sum_{j>N_{\text{eig}}} s_j^2 (a\cdot u_j)^2$. Summed over a complete set of observables, one for each grid point, the omitted variance is $\sum_{j>N_{\text{eig}}} s_j^2$. This is Eq. (2.4.24).

Sampling errors. The mean of N_{rep} independent values of variance σ^2 has variance σ^2/N_{rep}. For Gaussian values, $(N_{\text{rep}}-1)\,s^2/\sigma^2$ follows the χ^2 distribution with $N_{\text{rep}}-1$ degrees of freedom, whose variance is $2(N_{\text{rep}}-1)$. So the sample variance s^2 has relative error $\sqrt{2/(N_{\text{rep}}-1)}$, and its square root s has half that: Eq. (2.4.25).

References

1. S. Amoroso et al., “Snowmass 2021 whitepaper: Proton structure at the precision frontier”, arXiv:2203.13923. Section II A (comparison of sets), Table I, section V C (neural networks, hyperoptimisation, closure tests), section V D (delivery of sets), section IX A (LHAPDF) and section X (PDF4LHC21). The original papers for MSHT20, CT18, NNPDF4.0 and PDF4LHC21 are its references [9], [7], [10] and [26].
2. M. Ubiali, *A new approach in the determination of Parton Distribution Functions. Application to processes with heavy quarks in the initial states*, PhD thesis, Université catholique de Louvain and University of Edinburgh (2010). Section 2.1 (data, Table 2.1; the Hessian method, the tolerance in the CTEQ, MRST and MSTW analyses, and the Lagrange-multiplier method) and section 3.1 (the NNPDF method: parametrisation, genetic minimisation, cross-validation and stopping).
3. J. M. Moore, *Parton Distributions in Beyond the Standard Model Theories*, PhD thesis, University of Cambridge. Section 1.3 (functional forms, gradient descent, cross-validation, the Monte Carlo replica

method) and section 1.4.

4. The first edition of this guide, sections 3.5 to 3.8, for the lineage of the three groups, the heavy-quark schemes, the Bernstein form, the two-tier criterion and the levels of closure test.
5. Project records: `context/STATE.md` and `context/PHASE_B_LEDGER.md` (fixed setup and known seams, PPDF-30, PPDF-31, PPDF-35).
6. A. D. Martin, W. J. Stirling, R. S. Thorne and G. Watt, “Parton distributions for the LHC”, Eur. Phys. J. C 63 (2009) 189, arXiv:0901.0002. Section 6.2 for the dynamic tolerance.
7. S. Bailey, T. Cridge, L. A. Harland-Lang, A. D. Martin and R. S. Thorne, “Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs”, Eur. Phys. J. C 81 (2021) 341, arXiv:2012.04684.
8. R. D. Ball et al. (PDF4LHC Working Group), “The PDF4LHC21 combination of global PDF fits for the LHC Run III”, J. Phys. G 49 (2022) 080501, arXiv:2203.05506.