2.3 Three ways to an uncertainty

Hessian, Monte Carlo replicas and Bayesian sampling: what each assumes and when they must agree.

1. What an uncertainty on a function means

A fit returns one set of parton distribution functions (PDFs). This chapter establishes what it means to attach an error band to that set, derives the three constructions in use (the Hessian method, Monte Carlo replicas and Bayesian sampling), and derives when they must agree and how they come apart.

The starting point is fixed by chapters 2.1 and 2.2. The distributions at the input scale Q_0 are a family $f(x, Q_0^2; \theta)$ with p parameters θ_a. The data are a vector D_i, $i = 1 \dots N_{\mathrm{dat}}$, with covariance matrix C, and the theory prediction for point i is $T_i(\theta)$. For Gaussian errors the probability of the data given the parameters, the likelihood $\mathcal{L}$, is

$$\mathcal{L}(D \,|\, \theta) \propto \exp\Big[-\tfrac{1}{2}\chi^2(\theta) \Big], \qquad \chi^2(\theta) = \big(D - T(\theta)\big)^T C^{-1} \big(D - T(\theta)\big). \tag{2.3.1}$$

An uncertainty on a function is not a new kind of object. It is a probability distribution $\rho(\theta)$ on the p-dimensional parameter space, carried forward to whatever is computed from the parameters. Any quantity $\mathcal{O}(\theta)$ that depends on the distributions, such as the value $f(x, Q^2; \theta)$ at one point or a cross-section, inherits a mean and a variance:

$$\langle \mathcal{O} \rangle = \int d^p\theta \; \rho(\theta)\, \mathcal{O}(\theta), \qquad \sigma_{\mathcal{O}}^2 = \big\langle \mathcal{O}^2 \big\rangle - \langle \mathcal{O} \rangle^2. \tag{2.3.2}$$

The error band drawn around a distribution is $\sigma_{\mathcal{O}}$ for $\mathcal{O} = f(x, Q^2)$, evaluated at each x in turn. The full information, including correlations between points, is $\rho(\theta)$.

When ρ is narrow, $\mathcal{O}$ can be expanded to first order about the centre $\hat{\theta}$ of the distribution. Write $g_a = \partial \mathcal{O} / \partial \theta_a$ at $\hat{\theta}$ and let $V_{ab} = \langle \delta\theta_a \delta\theta_b \rangle$ be the covariance matrix of the parameters, with $\delta\theta = \theta - \hat{\theta}$. Then $\mathcal{O} - \mathcal{O}(\hat{\theta}) = g^T \delta\theta$ and

$$\sigma_{\mathcal{O}}^2 = g^T V \, g. \tag{2.3.3}$$

This is linear error propagation. It needs only V, not the whole of ρ.

The three methods differ in how they obtain ρ:

- The **Hessian method** assumes ρ is a Gaussian, finds its covariance from the curvature of χ^2 at the minimum, and uses Eq. (2.3.3).
- The **Monte Carlo method** produces a sample from ρ by refitting fluctuated copies of the data, and evaluates Eq. (2.3.2) as a sample average.
- **Bayesian sampling** defines ρ as the posterior, likelihood times prior, and draws a sample from it.

Sections 2 to 5 treat the Hessian method, section 6 the replicas, section 7 the Bayesian route. Section 8 proves when they coincide, section 9 runs all three on a toy with one nearly flat direction, and section 10 describes the closure test that calibrates any of them. Comparing the three on one 52-parameter fit is the purpose of this project; its results are in chapter 3.4.

2. The quadratic approximation and the Hessian

This section derives the parameter covariance from the shape of χ^2 near its minimum.

Let $\hat{\theta}$ be the minimum of χ^2. The gradient vanishes there, so the expansion in $\delta\theta = \theta - \hat{\theta}$ begins at second order:

$$\Delta\chi^2 \equiv \chi^2(\theta) - \chi^2_{\min} = \tfrac{1}{2}\,\delta\theta^T H\,\delta\theta + \mathcal{O}(\delta\theta^3), \qquad H_{ab} = \left.\frac{\partial^2\chi^2}{\partial\theta_a\,\partial\theta_b}\right|_{\hat{\theta}}. \tag{2.3.4}$$

H is the Hessian matrix. Conventions differ by a factor of two in the literature: some authors define H as half of this, so that $\Delta\chi^2 = \delta\theta^T H\,\delta\theta$. Here H is always the matrix of second derivatives of χ^2.

Differentiating Eq. (2.3.1) twice gives the Hessian in terms of the Jacobian $J_{ia} = \partial T_i/\partial\theta_a$ and the residuals $r_i = D_i - T_i$:

$$H_{ab} = 2\left(J^T C^{-1} J\right)_{ab} - 2\sum_{i,j} r_i\,(C^{-1})_{ij}\,\frac{\partial^2 T_j}{\partial\theta_a\,\partial\theta_b}. \tag{2.3.5}$$

The first term is twice the Fisher information matrix $F = J^T C^{-1} J$. It is built from first derivatives only, and it is positive semi-definite because $v^T F v = |C^{-1/2} J v|^2 \geq 0$. The second term couples the residuals to the curvature of the model. It vanishes when the model is linear in the parameters, and when the fit is exact. At a real minimum it is small when the residuals are noise-like, because their signs then alternate and the sum averages down. Dropping it is the Gauss–Newton approximation $H \simeq 2F$.

Inserting Eq. (2.3.4) into Eq. (2.3.1), the likelihood near the minimum is $\exp[-\tfrac{1}{4}\delta\theta^T H \delta\theta]$. Read as a probability density in θ, this is a Gaussian $\exp[-\tfrac{1}{2}\delta\theta^T V^{-1}\delta\theta]$ with

$$V = 2\,H^{-1} \;\simeq\; F^{-1}. \tag{2.3.6}$$

The factor 2 is there because $\chi^2 = -2\ln\mathcal{L}$. Reading the likelihood as a density in θ is a step that the three methods justify differently; section 8 returns to it. Equation (2.3.6) requires H to be positive definite. Section 5 examines what happens when it is not.

The meaning of a unit change in χ²

Equations (2.3.3) and (2.3.6) give the standard deviation of any observable. The same number has a geometrical reading. Ask for the smallest $\Delta\chi^2$ at which an observable can be moved by a given amount $\Delta\mathcal{O}$. The answer, derived in D1, is $\Delta\chi^2 = \Delta\mathcal{O}^2/\sigma^2_{\mathcal{O}}$. So the one-standard-deviation range of any single observable is the range it explores on the surface $\Delta\chi^2 = 1$, whatever the number of parameters. For a Gaussian this range holds 68.27% of the probability.

The statement concerns one observable at a time. The region that holds 68.27% of the joint probability of all parameters is larger: $\Delta\chi^2 = 2.30$ for two parameters and 56.3 for 52. A PDF error band is a statement about one quantity, so the relevant value is 1.

3. Eigenvector sets and the master formulas

This section turns the covariance matrix into a finite list of PDF sets from which a user can compute the uncertainty on anything. The construction is due to the CTEQ group [4] and is described in Ubiali's thesis, section 2.1 [2].

H is real and symmetric, so it has orthonormal eigenvectors v_k with eigenvalues λ_k, $k = 1 \dots p$. Define rescaled coordinates along them:

$$z_k = \sqrt{\frac{\lambda_k}{2}}\, v_k^T \delta\theta \qquad \Longrightarrow \qquad \Delta\chi^2 = \tfrac{1}{2}\sum_k \lambda_k \,(v_k^T \delta\theta)^2 = \sum_k z_k^2. \tag{2.3.7}$$

In the z coordinates the surfaces of constant $\Delta\chi^2$ are spheres, and the Gaussian of Eq. (2.3.6) has unit variance in every direction. The eigenvector sets are the $2p$ points at distance T from the minimum along each axis:

$$S_k^{\pm}: \qquad \theta = \hat{\theta} \pm T\sqrt{\frac{2}{\lambda_k}}\, v_k, \qquad \Delta\chi^2 = T^2. \tag{2.3.8}$$

T is the tolerance; $T = 1$ is the value derived in section 2, and section 4 discusses other choices. (The tolerance is a single number and carries no index, which distinguishes it from the theory vector $T_i(\theta)$.) Each $S_k^{\pm}$ is a complete set of distributions, which is evolved and tabulated like the central one.

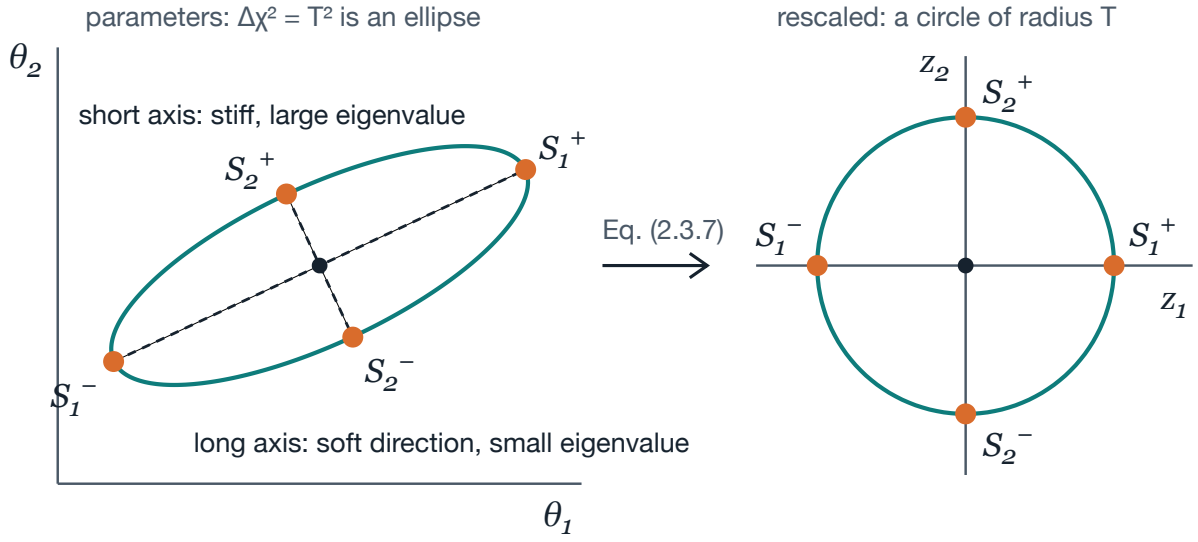

Figure 1. The Hessian construction with two parameters. The eigenvectors of H are the axes of the ellipse $\Delta\chi^2 = T^2$, and the eigenvector sets $S_k^{\pm}$ lie where the axes meet it, at distance $T\sqrt{2/\lambda_k}$ from the minimum. A small eigenvalue gives a long axis. As $\lambda_1 \to 0$ the axis, and every error band that depends on it, grows without limit.

Now take an observable and evaluate it on every set. If $\mathcal{O}$ is linear in θ over the range of the sets, the differences $\mathcal{O}(S_k^+) - \mathcal{O}(S_k^-)$ measure the components of its gradient along the eigenvectors, and Eq. (2.3.3) becomes (D2)

$$\Delta\mathcal{O} = \frac{1}{2}\sqrt{\sum_{k=1}^{p}\Big[\mathcal{O}(S_k^+) - \mathcal{O}(S_k^-)\Big]^2} \;=\; T\,\sqrt{g^T V\, g}. \tag{2.3.9}$$

This is the symmetric master formula. The user needs no derivatives and no knowledge of the parametrisation: $2p + 1$ evaluations of the observable are enough. When $\mathcal{O}$ is visibly non-linear, so that the two sets of a pair do not move it by equal and opposite amounts, the usual prescription keeps for each direction the larger excursion on each side:

$$\Delta\mathcal{O}_{\pm} = \sqrt{\sum_{k=1}^{p}\Big[\max\big(\pm(\mathcal{O}(S_k^+) - \mathcal{O}_0),\; \pm(\mathcal{O}(S_k^-) - \mathcal{O}_0),\; 0\big)\Big]^2}, \qquad \mathcal{O}_0 = \mathcal{O}(\hat{\theta}). \tag{2.3.10}$$

Equation (2.3.10) reduces to Eq. (2.3.9) for a linear observable. For a curved one it is a prescription and not a theorem: the χ^2 is still treated as quadratic. In the check script, an observable with a mild quadratic term has exact extreme values $+2.654$ and -2.513 on the surface $\Delta\chi^2 = 1$; Eq. (2.3.10) gives $+2.672$ and -2.476.

---

4. The tolerance

The groups that use the Hessian method in global fits do not use $T = 1$. This section states what they do and what it is equivalent to.

Because every set in Eq. (2.3.8) is displaced in proportion to T, choosing $\Delta\chi^2 = T^2$ is the same as rescaling the covariance matrix:

$$\Delta\chi^2 = T^2 \qquad \Longleftrightarrow \qquad V \;\to\; T^2\,V = 2\,T^2 H^{-1}. \tag{2.3.11}$$

The argument for $T > 1$, as summarised by Ubiali [2], is that the conditions under which $T = 1$ was derived do not hold in a global fit. Data sets from different experiments pull the same parameters in incompatible directions, theoretical uncertainties are absent from C, and the experimental errors need not be Gaussian. A band built with $\Delta\chi^2 = 1$ around the compromise minimum then fails to cover the values individual experiments prefer.

A minimal model shows what is being corrected. Take N measurements d_i of a single quantity θ, each with quoted error σ, whose true scatter is $s\sigma$ with $s > 1$. The fit is the mean $\hat\theta = \sum_i d_i/N$. Its Hessian error is $\sigma/\sqrt{N}$ by Eq. (2.3.6), but its true scatter is $s\sigma/\sqrt{N}$. Since $\sum_i (d_i - \hat\theta)^2$ has expectation $(N-1)s^2\sigma^2$,

$$\frac{\text{true error}}{\text{Hessian error}} = s, \qquad \frac{\langle \chi^2_{\min} \rangle}{N-1} = s^2. \tag{2.3.12}$$

The band is restored by $T = s$, and the fit quality reveals s. This is the scale factor the Particle Data Group applies to discrepant measurements, and Ubiali notes that a tolerance above one is equivalent to it [2]. A simulation with $N = 20$ and $s = 2$ returns 2.002 and 4.000.

A global fit is not this model. The inconsistencies are between data sets and along particular directions, so the tolerance cannot be read off the total χ^2. Two procedures are in use [2]:

- **A fixed tolerance.** The CTEQ group examined how far the fit could move along each eigenvector before some data set was no longer acceptably described at 90% confidence, and averaged over directions. The result was $T^2 \sim 100$; the MRST group used $T^2 \sim 50$.
- **A dynamic tolerance.** The MSTW group, and MSHT after it, set a separate T_k for each eigenvector direction, by requiring that every data set remains within its own confidence limit as the fit moves along that direction. Ubiali quotes $T \sim 3$ for 68% bands and $T \sim 5$ for 90% from the MSTW analysis. Chapter 2.4 derives the criterion.

Neither rule follows from the likelihood of Eq. (2.3.1), and Ubiali's assessment is that $T > 1$ has no rigorous statistical justification [2]. A tolerance is best read as a declared correction for a likelihood known to be incomplete. The Monte Carlo and Bayesian methods as usually run contain no such factor, which is one of the mechanisms of section 8.

---

5. Flat directions, conditioning and numerical Hessians

The Hessian method fails in one place: directions in parameter space that the data constrain weakly or not at all.

Flat directions

A direction v is flat if moving along it changes no prediction, $Jv = 0$. Then $Fv = 0$, and by Eq. (2.3.5) the Hessian eigenvalue is zero up to the residual term. Equation (2.3.8) puts the corresponding eigenvector sets at $\sqrt{2/\lambda} \to \infty$. The method is correct to report an infinite uncertainty on a combination the data do not determine. The difficulty is practical. A nearly flat direction has a small eigenvalue whose value, and even whose sign, is set by numerical error and by the residual term, and the quadratic expansion is least reliable there, because the fit moves far before $\Delta\chi^2$ reaches one.

Flat directions are the normal condition of a flexible parametrisation. In this project's 52-parameter fit, 12 directions are flat on deep-inelastic data alone and 8 remain flat on the full global data set (PPDF-32 [9]). These are counts of eigenvalues of the curvature matrix, at the planted truth of a noise-free closure test, that fall below a threshold set by the prior width. A first count from a forward-difference matrix read 14 and 10; the ledger corrected it with central differences, for the reason derived below.

The condition number

The severity is measured by the condition number $\kappa = \lambda_{\max}/\lambda_{\min}$. If H is known only up to an error δH, the error in its inverse obeys, to first order in δH,

$$\frac{\|\delta(H^{-1})\|}{\|H^{-1}\|} \;\le\; \kappa\,\frac{\|\delta H\|}{\|H\|}, \qquad \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}, \tag{2.3.13}$$

where $\|\cdot\|$ is the largest singular value. The bound follows from $\delta(H^{-1}) = -H^{-1}\,\delta H\,H^{-1}$ with $\|H^{-1}\| = 1/\lambda_{\min}$ and $\|H\| = \lambda_{\max}$. If the entries of H carry a relative error of 10^{-6} from finite differencing and $\kappa \sim 10^{11}$, the smallest eigenvalues, and with them the largest variances, are undetermined. The Hessian of this project's fit to real deep-inelastic data, computed by central differences at the minimum in parameters scaled by their prior widths, has a condition number of 1.6–1.8×10^{11}, stable under changes of the differencing step (PPDF-36 [9]). No band can be computed from it. Other condition numbers in the ledgers belong to other matrices: 1.2×10^{11} is the curvature at the planted truth of the closure test on the same data sets, from the first, forward-difference estimate (PPDF-32 [9]), and 2.8×10^{13} is the Phase A measurement at its planted truth (PPDF-19 [8]).

Numerical Hessians

The Hessian of a real fit is rarely available in closed form. It is built by differencing the gradient $G_a = \partial\chi^2/\partial\theta_a$, which automatic differentiation supplies. With step h along parameter b, and e_b the unit vector in that direction, the two standard formulas and their leading errors are (D3)

$$\begin{aligned} \text{forward:}\quad & \frac{G_a(\hat\theta + h\,e_b) - G_a(\hat\theta)}{h} = H_{ab} + \frac{h}{2}\,\frac{\partial^3\chi^2}{\partial\theta_a\,\partial\theta_b^2} + \mathcal{O}(h^2), \\ \text{central:}\quad & \frac{G_a(\hat\theta + h\,e_b) - G_a(\hat\theta - h\,e_b)}{2h} = H_{ab} + \frac{h^2}{6}\,\frac{\partial^4\chi^2}{\partial\theta_a\,\partial\theta_b^3} + \mathcal{O}(h^4). \end{aligned} \tag{2.3.14}$$

The forward formula costs p extra gradients and the central one $2p$. Either estimate differs from the true matrix by an error matrix E, and the effect on an eigenvalue is given by first-order perturbation theory:

$$\lambda_k^{\text{est}} = \lambda_k + v_k^T E\, v_k, \qquad E_{ab}^{\text{forward}} = \frac{h}{4}\left(\frac{\partial^3\chi^2}{\partial\theta_a\,\partial\theta_b^2} + \frac{\partial^3\chi^2}{\partial\theta_b\,\partial\theta_a^2}\right). \tag{2.3.15}$$

The shift has no fixed sign, and its size is set by third derivatives that have nothing to do with λ_k. A stiff eigenvalue is unaffected. A flat one is replaced by $v_k^T E v_k$: with forward differences, a number proportional to h that is as often

negative as positive. A negative eigenvalue appears to say that the fit sits on a saddle, when the estimate simply has no resolution there.

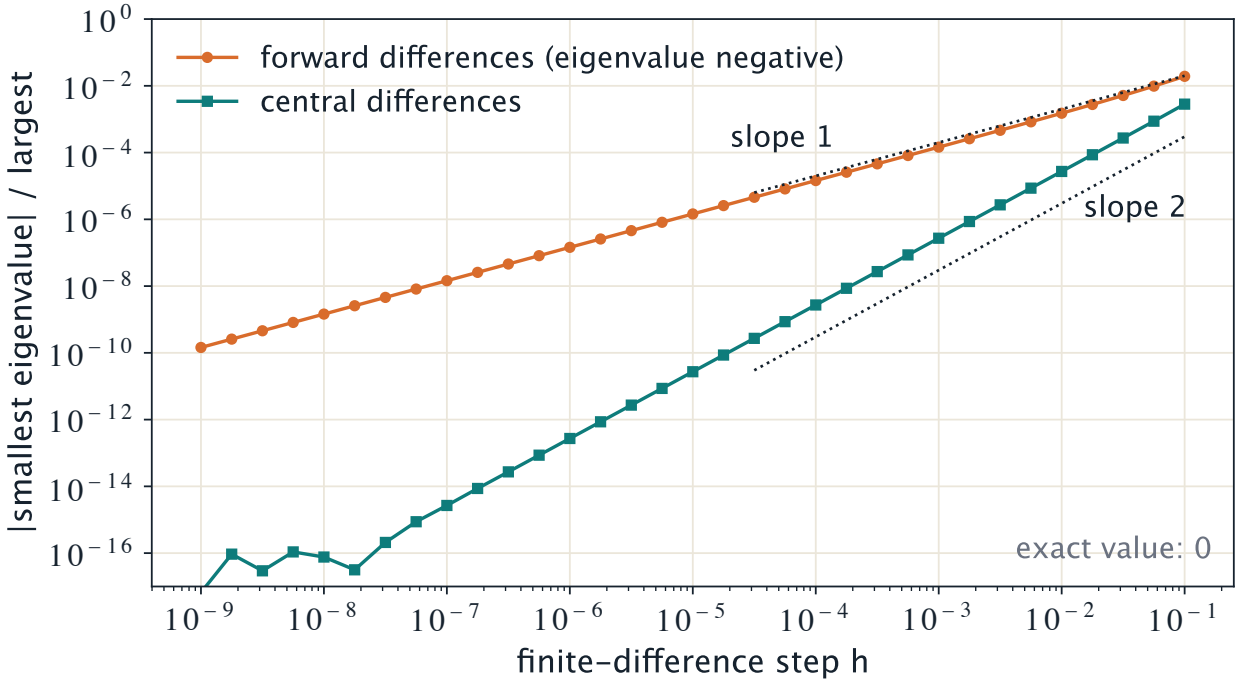

Figure 2. The estimated smallest eigenvalue of a Hessian whose true smallest eigenvalue is exactly zero. The model is $f = e^{\theta_1}\theta_3^2(1-x)^{\theta_2}$ fitted to noise-free data, in which only the product $e^{\theta_1}\theta_3^2$ is determined. Forward differences of the gradient return a negative eigenvalue proportional to the step. Central differences return an error proportional to the square of the step, smaller by a factor of 5,000 at $h = 10^{-4}$. Computed by the check script.

This project met exactly this. At the exact minimum of a noise-free closure fit, where the Hessian is positive semi-definite by Eq. (2.3.5), a forward-difference Hessian was scanned over the step, and then replaced by central differences [9]:

STEP / PRIOR HALF-WIDTH	FORWARD: NEGATIVE EIGENVALUES	FORWARD: SMALLEST SCALED EIGENVALUE	CENTRAL: NEGATIVE EIGENVALUES	CENTRAL: SMALLEST SCALED EIGENVALUE
10^{-2}	13	-7.945×10^{4}	4	-4.183×10^{1}
10^{-3}	10	-3.704×10^{3}	0	$+8.357 \times 10^{-5}$
10^{-4}	7	-1.354×10^{2}	0	$+3.695 \times 10^{-6}$
10^{-5}	5	-1.281×10^{1}	0	$+2.583 \times 10^{-6}$
10^{-6}	2	-1.210×10^{0}	not run	not run

The forward column falls by a factor of ten per decade of step, which is the first-order error of Eq. (2.3.15) and not curvature. With central differences the negative eigenvalues are gone for steps of 10^{-3} and below. The ledger's conclusion is that every negative-eigenvalue count from these scans was an artefact of forward differencing [9].

Reducing h does not help indefinitely. The gradient is known to a relative precision ϵ of about 10^{-16} at best, so differencing it adds a rounding error of order $\epsilon|G|/h$. At the minimum of a fit to real data the individual terms of the gradient are large and cancel, and the rounding error can exceed the small eigenvalues at every step. In this project's real-data Hessian the smallest eigenvalue changed by a factor of 15, non-monotonically, across three steps, while the condition number stayed fixed (PPDF-36 [9]). The signs of the smallest eigenvalues were unresolved; the conclusion that no band exists rested on the condition number.

---

6. Monte Carlo replicas

This section derives what the replica ensemble is a sample of. The method was introduced for PDF fits by the NNPDF collaboration [5, 6], and the description follows Moore's thesis, section 1.3.4 [1].

The method

Generate N_{rep} artificial data sets by adding to the measured data noise drawn from the experimental covariance matrix, and fit each one:

$$\begin{aligned} D^{(k)} &= D + \eta^{(k)}, \qquad \eta^{(k)} \sim \mathcal{N}(0, C), \\ \theta^{(k)} &= \arg\min_{\theta} \, \big(D^{(k)} - T(\theta)\big)^{T} C^{-1} \big(D^{(k)} - T(\theta)\big). \end{aligned} \tag{2.3.16}$$

Each $\theta^{(k)}$ defines one PDF replica. The ensemble is treated as a sample from $\rho(\theta)$, so that Eq. (2.3.2) becomes

$$\langle \mathcal{O} \rangle = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} \mathcal{O}\big(\theta^{(k)}\big), \qquad \sigma_{\mathcal{O}}^{2} = \frac{1}{N_{\text{rep}} - 1} \sum_{k=1}^{N_{\text{rep}}} \Big[\mathcal{O}\big(\theta^{(k)}\big) - \langle \mathcal{O} \rangle\Big]^{2}. \tag{2.3.17}$$

No derivative of the observable and no assumption of linearity is needed. The divisor $N_{\text{rep}} - 1$ makes the variance estimate unbiased, and is the convention of this guide; some sources divide by N_{rep} (chapter 2.4). The relative error of $\sigma_{\mathcal{O}}$ from a finite sample is $1/\sqrt{2(N_{\text{rep}} - 1)}$, which is 7% for 100 replicas (derived in chapter 2.4).

The linear case

Suppose the model is linear, $T(\theta) = X\theta$ with a fixed matrix X, and write $F = X^{T} C^{-1} X$. Then the fit is a linear map from data to parameters, and a Gaussian fluctuation of the data produces a Gaussian fluctuation of the fit. The result (D4) is

$$\theta^{(k)} \sim \mathcal{N}\big(\hat{\theta},\; F^{-1}\big), \qquad F^{-1} = 2H^{-1}. \tag{2.3.18}$$

The replicas are distributed about the fit to the central data with exactly the Hessian covariance at $T = 1$.

One point is easily misread. The replicas scatter about $\hat{\theta}$, not about the true parameters θ^{*}. The measured data are themselves one fluctuation about the truth, so $\hat{\theta}$ scatters about θ^{*} with covariance F^{-1}, and a replica carries both fluctuations:

$$\big\langle (\theta^{(k)} - \hat{\theta})(\theta^{(k)} - \hat{\theta})^{T} \big\rangle = F^{-1}, \qquad \big\langle (\theta^{(k)} - \theta^{*})(\theta^{(k)} - \theta^{*})^{T} \big\rangle = 2F^{-1}. \tag{2.3.19}$$

The first is the uncertainty of the fit. The second is what is seen when single replicas are compared with a known truth.

Beyond the linear case

For a non-linear model the fit is a non-linear map of the noise, and three things change.

The ensemble is not the posterior. The cleanest statement is for a model with as many parameters as data points and an invertible map $T(\theta)$, so that every replica is fitted exactly, $T(\theta^{(k)}) = D^{(k)}$. Changing variables from $D^{(k)}$ to $\theta^{(k)}$ gives the replica density (D5)

$$\rho_{\mathrm{MC}}(\theta) = \mathcal{L}(D \,|\, \theta) \left| \det \frac{\partial T}{\partial \theta} \right| . \tag{2.3.20}$$

The replicas sample the likelihood multiplied by the Jacobian of the model. A Bayesian posterior with a flat prior is the likelihood alone. The two agree when the Jacobian is constant, which is the linear case. Moore's thesis obtains the one-parameter form of this result for a quadratic model: the replica density is the Bayesian posterior rescaled by the derivative of the theory, plus a spike at the turning point of the parabola [1, section 4.7]. For $T = e^{\theta}$ and one datum $D = 1 \pm 0.15$, computed numerically, the replica mean is -0.0117 and the flat-prior posterior mean is -0.0365; the Hessian centre is 0. The two differ by 0.17 of the width, an effect of second order in the error.

The replica is not at the minimum. Fits with many parameters are stopped before convergence, by cross-validation in the NNPDF methodology [1]. Stopping matters most in the directions the data constrain least. Model the fit as gradient descent, $d\theta/dt = -\partial\chi^2/\partial\theta$, started from initial values with variance s_0^2 that are independent of the data, and stopped at time t. Along an eigenvector of H with eigenvalue λ_k the ensemble variance is (D5)

$$\sigma_k^2(t) = \left(1 - e^{-\lambda_k t}\right)^2 \frac{2}{\lambda_k} + e^{-2\lambda_k t}\, s_0^2 . \tag{2.3.21}$$

For a stiff direction, $\lambda_k t \gg 1$, this is the Hessian variance $2/\lambda_k$. For a flat direction, $\lambda_k t \ll 1$, it is s_0^2: the spread of the starting points. The ensemble is finite where the Hessian diverges, and its width there is a property of the initialisation and the stopping rule, not of the data.

There may be several minima. Each replica converges to the minimum whose basin contains its starting point. The ensemble then weights the minima by the size of their basins under the initialisation, which is not their posterior probability.

> **What a replica ensemble is.** It is the distribution of outcomes of a fitting procedure when the data fluctuate. For a linear model fitted to convergence it is exactly the Gaussian $\mathcal{N}(\hat{\theta}, 2H^{-1})$. Otherwise it depends on the Jacobian of the model, the starting points, the stopping rule and any bounds on the parameters, and none of these appear in the likelihood.

---

7. Bayesian sampling

This section defines the third construction and the diagnostics that say whether a sample from it can be trusted. Chapter 2.5 derives the algorithms and both diagnostics.

Bayes' theorem makes the step from likelihood to a density in θ explicit. With a prior density $\pi(\theta)$, which states what is assumed about the parameters before the data are used, the posterior is

$$\rho(\theta \mid D) = \frac{\mathcal{L}(D \mid \theta)\,\pi(\theta)}{Z}, \qquad Z = \int d^{p}\theta\; \mathcal{L}(D \mid \theta)\,\pi(\theta). \tag{2.3.22}$$

Z is the evidence, a normalisation that matters only for comparing models. In PDF fits the prior is usually flat inside a box of bounds on each parameter. The posterior cannot be written in closed form for a non-linear model, so it is represented by a sample, to which Eq. (2.3.17) is applied.

Markov-chain Monte Carlo builds a random walk in θ whose stationary distribution is the posterior. The Metropolis rule proposes a step and accepts it with probability $\min(1, \rho_{\text{new}}/\rho_{\text{old}})$, so Z is never needed. Successive points are correlated, and in many dimensions the walk explores a long, narrow posterior slowly.

Hamiltonian Monte Carlo adds a momentum variable and moves along a trajectory of fictitious dynamics in which $-\ln\rho$ is the potential energy, so that distant proposals are accepted with high probability. The No-U-Turn Sampler (NUTS) chooses the trajectory length automatically. The method needs gradients, which Colibri's JAX implementation provides, and it is sensitive to conditioning: the step size is set by the stiffest direction and the trajectory length by the softest.

Nested sampling computes Z by keeping a population of points drawn from the prior and repeatedly replacing the point of lowest likelihood by one of higher likelihood. The discarded points, suitably weighted, are a sample from the posterior. It handles separated modes, and its cost grows with the ratio of prior to posterior volume.

Diagnostics

A chain of n correlated draws is worth fewer than n independent ones. If ρ_t is the autocorrelation of a quantity at lag t along the chain, the variance of its sample mean is σ^2/n_{eff} with

$$n_{\text{eff}} = \frac{n}{\tau}, \qquad \tau = 1 + 2\sum_{t=1}^{\infty} \rho_t. \tag{2.3.23}$$

n_{eff} is the effective sample size (ESS) and τ the integrated autocorrelation time. For a chain with $\rho_t = 0.9^t$, $\tau = 19$.

The second diagnostic compares independent chains. Run m chains of length n. Let W be the average of the variances within each chain, and B/n the variance of the m chain means. Then

$$\hat{R} = \sqrt{\frac{\frac{n-1}{n}W + \frac{B}{n}}{W}}. \tag{2.3.24}$$

If all chains sample the same distribution, the chain means agree to within $\sqrt{W/n_{\text{eff}}}$ and $\hat{R} \to 1$. If the chains have not yet found the same region, the spread of their means stays finite and $\hat{R}^2 - 1$ is that spread in units of W.

Neither number can prove convergence; both can disprove it. This project fixed its criteria in advance. The pre-registration reads "ESS ≥ 100 median, R̂ < 1.05, divergences < 5%". The ledger applies the ESS threshold to the worst of the 52 parameters and the $\hat{R}$ threshold to the largest value, which is the stricter reading [9]. On the closure test with full deep-inelastic data, three valid chains gave for the worst parameter a combined ESS of 30.8 and an $\hat{R}$ of 1.1185; the medians over the 52 parameters were 222 and 1.0296 (PPDF-33 [9]). By Eq. (2.3.24) the chain means still differ by half a within-chain standard deviation in the worst parameter. The gate was not met: on $\hat{R}$ under either reading of the ESS criterion, and on ESS as well under the ledger's. No Bayesian band is quoted for that fit.

8. When the three agree, and how they come apart

This section collects the results into one theorem and one list.

The linear-Gaussian theorem

Take a model linear in its parameters, Gaussian data errors with a covariance matrix that does not depend on the parameters, a flat prior with no bounds in reach, fits run to convergence, and $T = 1$. Then (D4)

$$\underbrace{\rho(\theta \mid D)}_{\text{Bayesian posterior}} = \underbrace{\mathcal{N}\big(\hat{\theta},\, 2H^{-1}\big)}_{\text{Hessian}} = \underbrace{\text{distribution of } \theta^{(k)}}_{\text{Monte Carlo replicas}}. \tag{2.3.25}$$

The three methods compute one object. The equality is exact, not approximate, and it holds for every observable.

It extends approximately to non-linear models whenever the model is linear across the width of the likelihood. Where that holds, agreement of the three methods is a prediction. Where the methods disagree, one of the premises has failed, and the pattern of disagreement identifies which.

The mechanisms of disagreement

MECHANISM	HESSIAN	MONTE CARLO REPLICAS	BAYESIAN POSTERIOR
Flat direction	variance $2/\lambda \to \infty$; sign of λ numerically unresolved	spread of the starting points, Eq. (2.3.21)	the prior, Eq. (2.3.26)
Non-Gaussian likelihood	symmetric Gaussian by construction	likelihood times Jacobian, Eq. (2.3.20)	the likelihood itself
Bounds on parameters	ignored; the band can cross them	replicas accumulate on the bound	probability is cut off at the bound
Several minima	describes one	weights them by basin size	weights them by probability, if the sampler finds them
Tolerance	$V \to T^2 V$	none	none
Unconverged computation	finite-difference error, Eq. (2.3.15)	early stopping, too few replicas	low ESS, $\hat{R} > 1$

The first row is the one that matters most for PDFs, and the Bayesian entry is a short calculation. Let the parameters split as $\theta = (u, w)$, where the predictions depend on u only, and let the prior factorise, $\pi = \pi_u(u)\,\pi_w(w)$. Then the likelihood is $\mathcal{L}(D|u)$ and Eq. (2.3.22) gives

$$\rho(u, w \mid D) = \frac{\mathcal{L}(D|u)\,\pi_u(u)}{\int du'\,\mathcal{L}(D|u')\,\pi_u(u')} \;\times\; \pi_w(w). \tag{2.3.26}$$

The posterior of w is its prior. For a flat prior of half-width L the standard deviation is $L/\sqrt{3}$. The Bayesian answer on a flat direction is a statement of the assumed bounds, and it is recognisable as one: a posterior width equal to the prior width marks a direction the data did not measure. The Hessian and replica bands carry no such marker.

The last row is a property of execution, not of the methods, and it can imitate any of the others. In this project's first closure comparisons the Bayesian band appeared to collapse; that was an unconverged sample, identified after the

three implementations had been validated on a two-parameter problem with a known answer (PPDF-24, PPDF-25 [8]).

9. A worked toy with one nearly flat direction

This section runs the three methods on a model in which a direction can be made flat by removing data.

The model

The model is one PDF-like function with a normalisation and a large-x exponent:

$$f(x;\theta) = \theta_1\,(1-x)^{\theta_2}, \qquad \theta^* = (1,\,3), \qquad 0 \le \theta_2 \le 10. \tag{2.3.27}$$

The data are 12 points equally spaced in x up to a largest value $x_{\max}$, each with an uncorrelated error of 10% of the true value. The observable is f at $x = 0.5$. Since $\ln f = \ln\theta_1 + \theta_2 \ln(1-x)$ and $\ln(1-x) \simeq -x$ at small x, data confined to small x determine the normalisation and say little about the exponent. Reducing $x_{\max}$ makes θ_2 nearly flat, as a real fit loses its hold on large-x exponents where the data stop.

The Hessian band uses the exact Hessian of Eq. (2.3.5) at the best fit and Eq. (2.3.3). The Monte Carlo band uses 200 replicas, each fitted to its global minimum inside the bounds. The Bayesian band uses a flat prior on θ_1 and on θ_2 within its bounds; the posterior is integrated numerically, not sampled, so sampler convergence is not at issue. Everything is repeated for 3,000 independent data sets at each $x_{\max}$.

The results

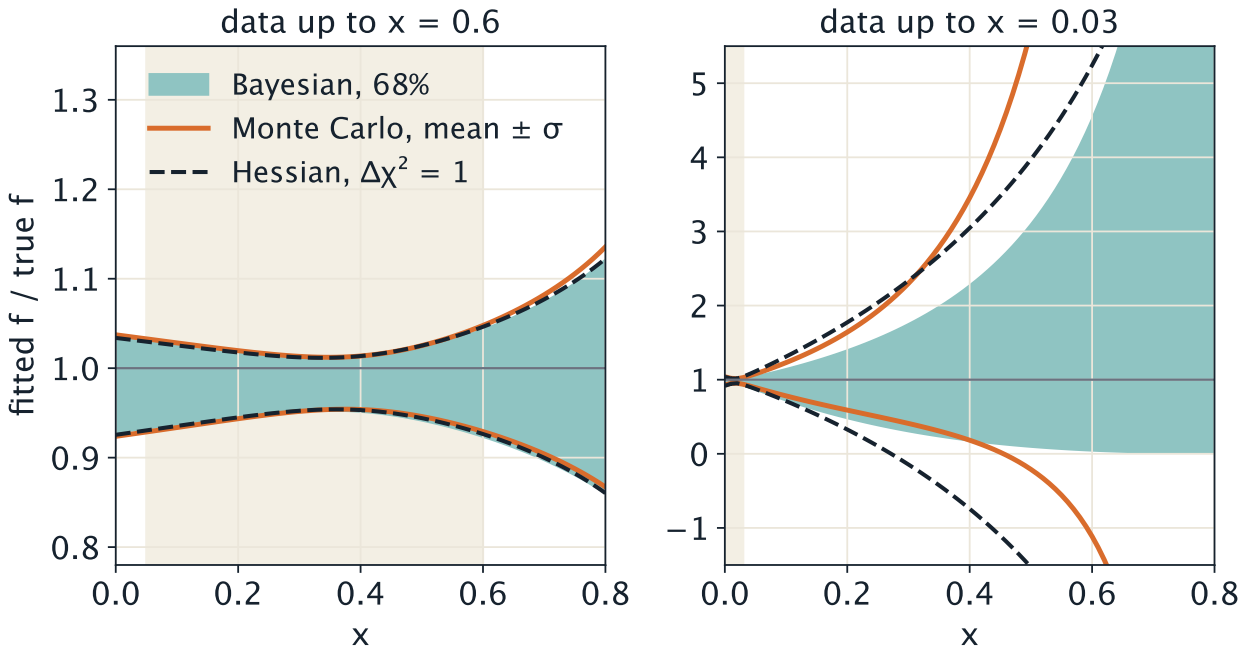

Figure 3. The three bands for one data set, as the ratio of the fitted to the true function. The shaded strip marks where the data are. Left: data up to $x = 0.6$; the three bands coincide. Right: data up to $x = 0.03$; the bands agree inside the data region and separate outside it. The Hessian and Monte Carlo bands cross zero; the Bayesian band is limited by the prior. Toy model, computed by the check script.

LARGEST DATA POINT	ERROR ON THE EXPONENT	CONDITION NUMBER	HESSIAN WIDTH	MONTE CARLO WIDTH	BAYESIAN WIDTH	REPLICAS AT A BOUND
0.6	0.11	20	0.041	0.041	0.041	0%
0.12	0.78	740	0.49	0.58	0.57	0.3%
0.03	3.3	1.3×10^4	2.2	2.6	1.7	33%
0.015	6.6	5.3×10^4	4.9	3.0	1.8	61%

The second column is the standard deviation of θ_2 from the Fisher matrix at the true parameters, to be compared with the prior range of 10. The widths are medians over data sets of the one-standard-deviation band on $f(0.5)$, in units of the true value.

With data up to $x = 0.6$ the three widths agree to better than 1%, as Eq. (2.3.25) requires. As the data retreat they separate, in the way the table of section 8 predicts.

The Bayesian width saturates. It stops growing at about 1.8. The spread of $(1 - x)^{\theta_2}$ at $x = 0.5$ when θ_2 is uniform on its prior range is 1.81 in the same units. This is Eq. (2.3.26): the posterior has become the prior.

The Monte Carlo ensemble meets the bounds. With data up to $x = 0.015$, 61% of all replicas have their best fit at $\theta_2 = 0$ or $\theta_2 = 10$, and the width of the ensemble is set by the distance between the bounds. This project recorded the same behaviour on its own two-parameter validation problem, where 81% of replicas ended at the prior bound in the degenerate case (PPDF-24 [8]).

The Hessian width is unstable. Over data sets, the central 80% of Hessian widths on the last row runs from 0.038 to 36, against 1.4 to 3.6 for Monte Carlo and 1.2 to 2.2 for the Bayesian band. Linear propagation about the best fit gives a band proportional to the best-fit value of $f(0.5)$. When the best fit lands at $\theta_2 = 10$, that value is nearly zero and so is the band; at $\theta_2 = 0$ the band is enormous. The Hessian is positive definite in every one of these fits. The method fails here through the expansion point, not through the sign of an eigenvalue.

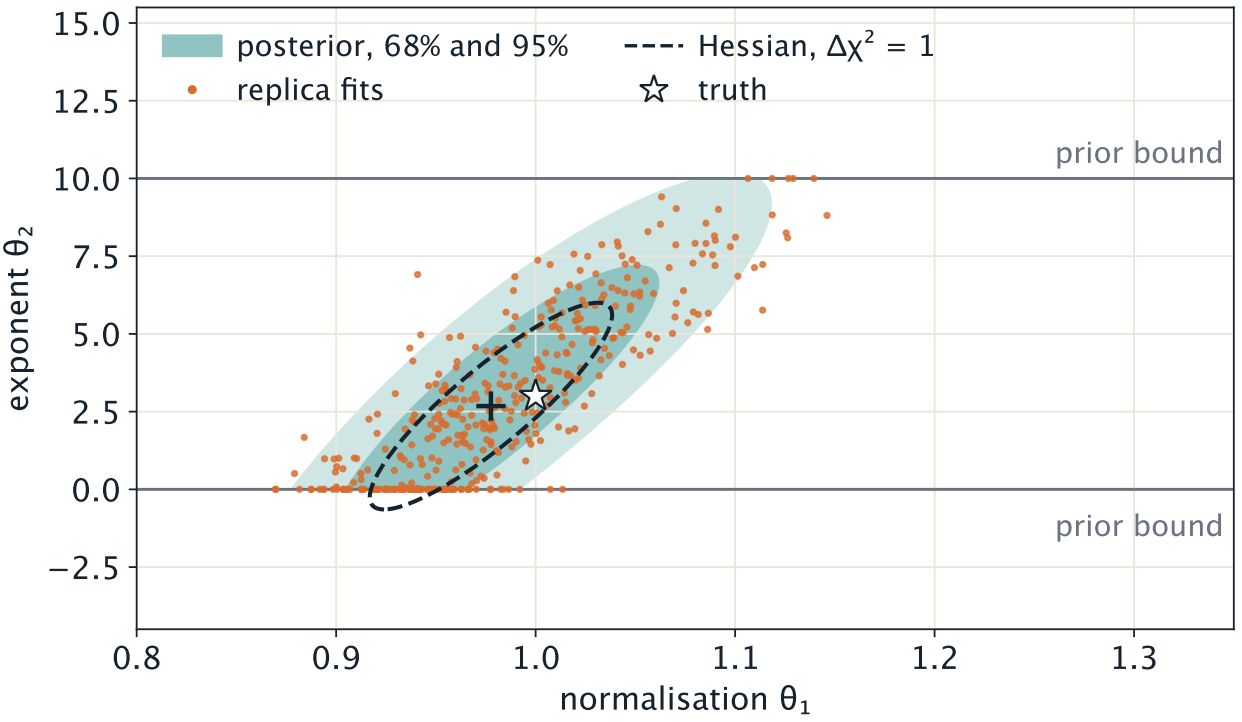

Figure 4. The same data set as the right panel of Figure 3, in the parameter plane. The posterior (shaded) is cut off by the prior bounds. The replica fits (dots) follow the same valley, and 23% of them sit on a bound. The Hessian contour (dashed) is an ellipse about the best fit (cross) that takes no account of the bounds. The Hessian error on the exponent is 3.3; the posterior standard deviation is 2.4 and the prior's is 2.9.

A real fit has 52 parameters and several flat directions, but the mechanisms are the same. In this project's closure test on deep-inelastic data, the ratio of the Monte Carlo to the Hessian error over the 52 parameters has a median of 1.09 and a maximum of 31.5, and for 6 of the 52 the Monte Carlo error is wider than the prior (PPDF-33 [9]). The methods agree for most parameters and differ by large factors for a few.

---

10. Closure tests: calibrating any method

The comparison of section 9 was possible because the truth was known. A closure test applies the same idea to a real fitting framework. It has been part of the NNPDF methodology since NNPDF3.0 [3, 7]: distributions fitted to pseudo-data generated from a known underlying law must reproduce the statistical distributions expected from the assumed experimental uncertainties [3].

The three levels

Choose true parameters θ^* inside the model. Three kinds of pseudo-data are distinguished [1, section 5.1]:

$$\begin{aligned} &\text{level 0:} \quad D = T(\theta^*), \\ &\text{level 1:} \quad D = T(\theta^*) + \eta, \\ &\text{level 2:} \quad D^{(k)} = T(\theta^*) + \eta + \epsilon^{(k)}, \end{aligned} \tag{2.3.28}$$

with η and $\epsilon^{(k)}$ independent draws from $\mathcal{N}(0, C)$. A fit to level-0 data tests the minimiser and the flexibility of the model: χ^2 must approach zero. Level 1 is one simulated experiment. Level 2 is the set of Monte Carlo replicas

generated from a level-1 data set, as in Eq. (2.3.16).

What a correct method must return

For a level-1 fit of a model that is linear near the minimum, with p_{eff} directions that the data constrain, the minimum χ^2 is distributed as χ^2 with $N_{\text{dat}} - p_{\text{eff}}$ degrees of freedom (D6):

$$\left\langle \chi^2_{\min} \right\rangle = N_{\text{dat}} - p_{\text{eff}}, \qquad \text{var}\left(\chi^2_{\min}\right) = 2\left(N_{\text{dat}} - p_{\text{eff}}\right). \tag{2.3.29}$$

A value well above this means the fit did not find the planted truth; a value well below means it has absorbed noise. Flat directions do not count towards p_{eff}.

The sharper test is coverage. A 68.27% band should contain the truth in 68.27% of cases. Measure the fraction $\hat{c}$ of n independent cases in which it does. If the method is calibrated, each case is an independent trial with success probability $c = 0.6827$, and

$$\hat{c} = c \pm \sqrt{\frac{c\,(1-c)}{n}}. \tag{2.3.30}$$

A coverage significantly below nominal means the bands are too narrow. The cases must be independent for Eq. (2.3.30) to apply; values of f at neighbouring x, or correlated parameters, are not, and the true error on $\hat{c}$ is then larger. Coverage also varies from one level-1 draw to the next: this project found single-draw coverages between 47% and 76% and quotes only averages over draws [9].

In the toy of section 9, the coverage on the first row is 68.9% for all three methods, against an expected $68.3 \pm 0.8\%$ for 3,000 trials. On the last row it is 65% for the Hessian band, 63% for Monte Carlo and 89% for the Bayesian band. The Bayesian over-coverage is not a fault of the calculation. A credible interval is calibrated on average over truths drawn from the prior. Here the truth is fixed at $\theta_2 = 3$, near the middle of a wide prior, and is covered more often than 68%; a truth near the edge would be covered less often. In a prior-dominated direction, coverage measures the relation between the truth and the prior.

In this project

The fits use the MSHT20 parametrisation with 52 free parameters in the Colibri framework, with NNLO theory (NNPDF theory 40000000), $\alpha_S(M_Z) = 0.118$ and $Q_0 = 1.65$ GeV. The closure results that bear on this chapter are these [8, 9]:

- **Monte Carlo.** On the level-1 closure test with deep-inelastic data, the one-standard-deviation replica interval contains the true value for 39 of the 52 parameters, 75.0%; on the global data set the figure is 82.7% (PPDF-33). The ledger records this as over-coverage. By Eq. (2.3.30) the binomial error for 52 independent trials is 6.5 percentage points, so 75.0% is 1.0 standard deviation above nominal. That comparison is our arithmetic, not the ledger's, and the parameters are not independent.
- **Bayesian.** With converged sampling on deep-inelastic data, the mean coverage over five noise draws was 57.9% against a nominal 68% (PPDF-25). On the Phase B closure test the convergence gate of section 7 was not met and no coverage is quoted (PPDF-33, PPDF-34).
- **Hessian.** No coverage is claimed. The best fit used as expansion point was more than five standard deviations from the truth in the worst parameter, and the analysis refuses a coverage statement in that case (PPDF-33).

One open experiment would connect these results to section 4: measuring the tolerance T^2 at which the Hessian bands cover a planted truth. Chapter 3.4 gives the full account.

Derivations

D1. The covariance from the Hessian, and the unit change in χ²

Derives (2.3.6)

The covariance. With Eq. (2.3.4), the density proportional to the likelihood is $\exp[-\frac{1}{4}\delta\theta^T H\delta\theta]$. A Gaussian with covariance V is $\exp[-\frac{1}{2}\delta\theta^T V^{-1}\delta\theta]$. Matching the exponents, $V^{-1} = \frac{1}{2}H$, which is Eq. (2.3.6). The check script integrates a two-parameter example directly and recovers $2H^{-1}$.

The constrained minimum. Fix a change $\Delta\mathcal{O} = g^T\delta\theta$ of a linear observable and minimise $\Delta\chi^2 = \frac{1}{2}\delta\theta^T H\delta\theta$ subject to it. With a Lagrange multiplier μ, the function to make stationary is $\frac{1}{2}\delta\theta^T H\delta\theta - \mu\,(g^T\delta\theta - \Delta\mathcal{O})$. Differentiating with respect to $\delta\theta$,

$$H\,\delta\theta = \mu\, g \qquad\Longrightarrow\qquad \delta\theta = \mu\, H^{-1} g = \tfrac{1}{2}\mu\, V\, g.$$

Fixing the multiplier. Insert this into the constraint: $\Delta\mathcal{O} = \frac{1}{2}\mu\, g^T V g$, so $\mu = 2\Delta\mathcal{O}/(g^T V g)$ and $\delta\theta = V\, g\, \Delta\mathcal{O}/(g^T V g)$.

The value at the minimum. Using $H = 2V^{-1}$,

$$\Delta\chi^2 = \tfrac{1}{2}\,\delta\theta^T H\,\delta\theta = \delta\theta^T V^{-1}\delta\theta = \frac{\Delta\mathcal{O}^2\; g^T V\, V^{-1} V\, g}{(g^T V g)^2} = \frac{\Delta\mathcal{O}^2}{g^T V g} = \frac{\Delta\mathcal{O}^2}{\sigma_{\mathcal{O}}^2}.$$

The stationary point is a minimum because H is positive definite. Setting $\Delta\chi^2 = 1$ gives $|\Delta\mathcal{O}| = \sigma_{\mathcal{O}}$: the largest change of $\mathcal{O}$ on the surface $\Delta\chi^2 = 1$ is one standard deviation, in any number of dimensions.

The probability content. A linear function of Gaussian variables is Gaussian, so $\mathcal{O}$ has a one-dimensional Gaussian distribution of width $\sigma_{\mathcal{O}}$, and the interval $\pm\sigma_{\mathcal{O}}$ holds $\mathrm{erf}(1/\sqrt{2}) = 68.27\%$.

D2. The master formulas

Derives (2.3.9), (2.3.10)

The gradient in the eigenbasis. Expand the gradient of the observable on the eigenvectors of H: $g = \sum_k g_k v_k$ with $g_k = v_k^T g$. Since $V = 2H^{-1} = \sum_k (2/\lambda_k)\, v_k v_k^T$, Eq. (2.3.3) reads

$$\sigma_{\mathcal{O}}^2 = g^T V g = \sum_k \frac{2}{\lambda_k}\, g_k^2.$$

The sets measure the same components. For a linear observable, Eq. (2.3.8) gives

$$\mathcal{O}(S_k^{\pm}) - \mathcal{O}_0 = g^T\Big(\pm T\sqrt{\tfrac{2}{\lambda_k}}\, v_k\Big) = \pm T\sqrt{\tfrac{2}{\lambda_k}}\, g_k, \qquad \mathcal{O}(S_k^{+}) - \mathcal{O}(S_k^{-}) = 2T\sqrt{\tfrac{2}{\lambda_k}}\, g_k.$$

Summing. Square, sum over k and divide by 4:

$$\frac{1}{4}\sum_k \left[\mathcal{O}(S_k^{+}) - \mathcal{O}(S_k^{-})\right]^2 = T^2 \sum_k \frac{2}{\lambda_k} g_k^2 = T^2\, g^T V g,$$

which is Eq. (2.3.9). At $T = 1$ it is the standard deviation. For other T it is the standard deviation of the rescaled Gaussian of Eq. (2.3.11).

D3. Errors of finite-difference Hessians

Derives (2.3.14), (2.3.15)

Taylor expansion. Expand the gradient component G_a along parameter b about $\hat{\theta}$, writing primes for derivatives with respect to θ_b:

$$G_a(\hat{\theta} \pm h\, e_b) = G_a \pm h\, G'_a + \frac{h^2}{2} G''_a \pm \frac{h^3}{6} G'''_a + \mathcal{O}(h^4), \qquad G'_a = H_{ab}.$$

Forward. Subtract G_a and divide by h: the result is $H_{ab} + \frac{h}{2} G''_a + \mathcal{O}(h^2)$, and $G''_a = \partial^3 \chi^2 / \partial\theta_a \partial\theta_b^2$.

Central. Subtract the two expansions and divide by $2h$. The even terms cancel, leaving $H_{ab} + \frac{h^2}{6} G'''_a + \mathcal{O}(h^4)$. These are the two lines of Eq. (2.3.14).

The error matrix. The forward estimate is not symmetric in a and b, because the correction involves two derivatives along b and one along a. Symmetrising it, as is always done before diagonalising, gives the matrix E^{forward} of Eq. (2.3.15).

The eigenvalue shift. For a symmetric matrix $H + E$ with E small, write the perturbed eigenvector as $v_k + \delta v_k$ and the eigenvalue as $\lambda_k + \delta\lambda_k$. To first order, $(H + E)(v_k + \delta v_k) = (\lambda_k + \delta\lambda_k)(v_k + \delta v_k)$ gives $H\delta v_k + E v_k = \lambda_k \delta v_k + \delta\lambda_k v_k$. Multiply on the left by v_k^T and use $v_k^T H = \lambda_k v_k^T$ and $v_k^T v_k = 1$. The terms in δv_k cancel and $\delta\lambda_k = v_k^T E v_k$.

The consequence. For forward differences $\delta\lambda_k = h\, c_k$, where c_k is a combination of third derivatives with no fixed sign. The estimated eigenvalue is negative when $c_k < 0$ and $\lambda_k < h|c_k|$. For the model of Figure 2 the check script finds $\delta\lambda = -2587\, h$ against a largest eigenvalue of 1.78×10^4, in agreement with Eq. (2.3.15) to better than 0.1%.

D4. The linear-Gaussian theorem

Derives (2.3.18), (2.3.19), (2.3.25)

The fit as a linear map. For $T(\theta) = X\theta$, setting the gradient of $\chi^2 = (D - X\theta)^T C^{-1} (D - X\theta)$ to zero gives the normal equations $X^T C^{-1} (D - X\theta) = 0$. With $F = X^T C^{-1} X$, assumed invertible,

$$\hat{\theta} = A\, D, \qquad A = F^{-1} X^T C^{-1}.$$

The Hessian. χ^2 is exactly quadratic, with $H = 2X^T C^{-1} X = 2F$. So $2H^{-1} = F^{-1}$, and the Hessian method assigns $\mathcal{N}(\hat{\theta}, F^{-1})$.

The replicas. The same map applied to $D^{(k)} = D + \eta^{(k)}$ gives $\theta^{(k)} = \hat{\theta} + A\,\eta^{(k)}$. A linear function of a Gaussian vector is Gaussian, with mean $\hat{\theta}$ and covariance

$$\left\langle A\,\eta\,\eta^T A^T \right\rangle = A\, C\, A^T = F^{-1} X^T C^{-1}\; C\; C^{-1} X\, F^{-1} = F^{-1} F\, F^{-1} = F^{-1}.$$

This is Eq. (2.3.18).

The posterior. Write $\theta = \hat{\theta} + \delta$. Then $D - X\theta = (D - X\hat{\theta}) - X\delta$, and

$$\chi^2(\theta) = \chi^2_{\min} - 2\,\delta^T X^T C^{-1}(D - X\hat\theta) + \delta^T X^T C^{-1} X\,\delta = \chi^2_{\min} + \delta^T F\,\delta.$$

The cross term vanishes by the normal equations. With a flat prior, Eq. (2.3.22) gives $\rho(\theta|D) \propto \exp[-\frac{1}{2}\delta^T F \delta]$, which is $\mathcal{N}(\hat\theta, F^{-1})$. The three distributions are identical: Eq. (2.3.25).

Two layers. If the data are $D = X\theta^* + \eta$, then $\hat\theta = \theta^* + A\eta$ and $\theta^{(k)} = \theta^* + A\eta + A\eta^{(k)}$. The two noise vectors are independent, so the covariance of $\theta^{(k)}$ about θ^* is $F^{-1} + F^{-1}$, which is Eq. (2.3.19).

D5. Replicas beyond the linear case

Derives (2.3.20), (2.3.21)

An invertible model. Let $p = N_{\text{dat}}$ and let $\theta \mapsto T(\theta)$ be one-to-one onto the range of the data. Then the minimum of the replica χ^2 is zero, reached at $T(\theta^{(k)}) = D^{(k)}$. The pseudo-data have density $\mathcal{N}(D^{(k)}; D, C)$. Changing variables from $D^{(k)}$ to $\theta^{(k)}$ multiplies the density by the Jacobian determinant:

$$\rho_{\text{MC}}(\theta) = \mathcal{N}\big(T(\theta); D, C\big)\left|\det\frac{\partial T}{\partial\theta}\right| \propto \exp\Big[-\tfrac{1}{2}\chi^2(\theta)\Big]\left|\det\frac{\partial T}{\partial\theta}\right|.$$

The first factor is the likelihood of Eq. (2.3.1) evaluated on the real data. This is Eq. (2.3.20).

Stopped gradient descent. Take a linear model, so that the χ^2 of replica k is quadratic about its own minimum $\theta^{(k)}_{\min}$ with Hessian H. Project onto an eigenvector of H with eigenvalue λ, and call the component a. Then $\partial\chi^2/\partial a = \lambda\,(a - a_{\min})$ and the descent equation $da/dt = -\partial\chi^2/\partial a$ has the solution

$$a(t) = a_{\min} + \big(a(0) - a_{\min}\big)\,e^{-\lambda t} = \big(1 - e^{-\lambda t}\big)\,a_{\min} + e^{-\lambda t}\,a(0).$$

The variance. By Eq. (2.3.18), $a_{\min}$ varies over replicas with variance $2/\lambda$. The starting value $a(0)$ has variance s_0^2 and is independent of the noise. Adding the two contributions gives Eq. (2.3.21). As $\lambda \to 0$ the first term behaves as $(\lambda t)^2(2/\lambda) = 2\lambda t^2 \to 0$ and the second tends to s_0^2. The check script integrates the descent for 400,000 replicas with eigenvalues 40 and 0.05 up to $t = 2$: the flat direction has variance 0.763, against a starting variance of 0.49 and a Hessian variance of 40.

D6. The minimum χ² of a level-1 fit

Derives (2.3.29)

The projection. Let $D = T(\theta^*) + \eta$ and linearise the model about the truth, $T(\theta) = T(\theta^*) + J\,\delta$. Chapter 2.2 (its derivation D5) shows that in whitened variables, $e = C^{-1/2}\eta$ and $\tilde J = C^{-1/2}J$, the minimum is $\chi^2_{\min} = |(1 - P)\,e|^2$, where e is a vector of N_{dat} independent unit Gaussians and P is the orthogonal projector onto the column space of $\tilde J$. That derivation assumes $\tilde J$ has full rank p. The only change needed here is the dimension of the column space.

Counting dimensions. The column space has dimension $p_{\text{eff}} = \operatorname{rank}\tilde J$, the number of independent directions in parameter space that change the predictions. In an orthonormal basis adapted to it, $(1 - P)e$ has $N_{\text{dat}} - p_{\text{eff}}$ non-zero components, each an independent unit Gaussian. The sum of their squares has the χ^2 distribution with $N_{\text{dat}} - p_{\text{eff}}$ degrees of freedom, whose mean and variance are those of Eq. (2.3.29).

Flat directions. A flat direction has $\tilde J v = 0$ and does not enlarge the column space. A fit with 52 parameters of which 12 are flat removes 40 units of χ^2, not 52.

References

1. J. M. Moore, *Parton Distributions in Beyond the Standard Model Theories*, PhD thesis, University of Cambridge. Sections 1.3.4 (the replica method), 4.7 (its pitfalls for a quadratic model) and 5.1 (level 0, 1 and 2 pseudo-data).
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 for the Hessian method and the tolerance.
3. S. Amoroso et al., "Snowmass 2021 whitepaper: Proton structure at the precision frontier", arXiv:2203.13923.
4. J. Pumplin, D. Stump, R. Brock, D. Casey, J. Huston, J. Kalk, H. L. Lai and W. K. Tung, Phys. Rev. D 65 (2001) 014013. The Hessian eigenvector method.
5. W. T. Giele, S. A. Keller and D. A. Kosower, arXiv:hep-ph/0104052. Monte Carlo ensembles of distributions.
6. L. Del Debbio, S. Forte, J. I. Latorre, A. Piccione and J. Rojo, "Unbiased determination of the proton structure function F_2^p with faithful uncertainty estimation", JHEP 03 (2005) 080.
7. R. D. Ball et al. (NNPDF), JHEP 04 (2015) 040, arXiv:1410.8849; L. Del Debbio, T. Giani and M. Wilson, Eur. Phys. J. C 82 (2022) 330, arXiv:2111.05787. Closure tests.
8. This project, Phase A results ledger (`context/RESULTS_LEDGER.md`): PPDF-19, PPDF-24, PPDF-25.
9. This project, Phase B results ledger (`context/PHASE_B_LEDGER.md`): PPDF-32, PPDF-33, PPDF-34, PPDF-36; and the pre-registration of its gates (`context/PHASE_B_PREREG.md`).