3.4 Three methods on one fit

This project's comparison of Hessian, Monte Carlo and Bayesian uncertainties on the same 52-parameter fit.

1. The question

This chapter reports what happened when the three constructions of chapter 2.3 were applied to one fit. The specification came from Maria Ubiali in Cambridge, in July 2026: put the MSHT20 parametrisation, with its 52 free parameters, into her group's Colibri framework, fit it to the data used by NNPDF, and compare Hessian, Monte Carlo and Bayesian uncertainties on the same fit [1].

Chapter 2.3 says what to expect. For a model that is linear across the width of the likelihood, with Gaussian errors and no bound in reach, the three constructions return one distribution. A disagreement identifies a premise that has failed. Here the first one fails: the data do not determine all 52 parameters of the parton distribution functions (PDFs).

The results, in the order the chapter derives them:

- **The form fits.** On 82 data sets and 4,615 points the minimum is $\chi^2/N_{\text{dat}} = 1.2380$ (PPDF-35).
- **Some directions are flat, and more data stop removing them.** The number of unconstrained directions among the 52 falls $12 \to 9 \to 8 \to 8$ as data are added in four steps (PPDF-32, corrected value).
- **On real deep-inelastic data the Monte Carlo and Bayesian bands differ by factors of 12 to 144** (PPDF-23, PPDF-28), and the Hessian cannot be inverted to a band: its condition number is $1.6\text{–}1.8 \times 10^{11}$ (PPDF-36).
- **On closure tests the Monte Carlo band over-covers**, 75.0% and 82.7% against a nominal 68.3% (PPDF-33). No Hessian coverage could be claimed. The Bayesian sampler did not meet its convergence gate at 3,089 points, and no posterior band is quoted (PPDF-33, PPDF-34).

Every number below is quoted from the project's ledgers [1, 2, 3] with its result ID. Phase A is PPDF-1 to PPDF-28, on deep-inelastic scattering (DIS) data, with a few entries on a 73-data-set selection that adds Drell–Yan, W and Z data. Phase B starts at PPDF-30, on the global data set, with gates written down before any run [3]. PPDF-17, 27 and 29 were reserved and never run.

The ledgers are append-only: a wrong entry is corrected by a later one, not edited. The state file counts three previously published numbers as corrected, on PPDF-32, 33 and 36 [1]. The ledger also corrects a published count of missing FK tables (PPDF-37) and retracts one internal result (PPDF-33) [2]. This chapter gives the current value and, where the correction teaches something about a diagnostic, the history. Where two entries disagree and no later entry settles it, the text says so.

> **Two kinds of number.** A *closure* number comes from pseudo-data generated from known parameters, so coverage of the truth can be measured. A *real-data* number has no truth behind it. Only the first kind can support a statement about calibration.

2. The set-up

Nothing in this section changes between the three methods.

Model. The MSHT20 form of chapter 2.1, ported to JAX as a Colibri model with "52 free params, 4 analytic sum rules" (PPDF-10). The port passed 96 checks against the reference implementation, with a worst deviation $\lesssim 10^{-9}$.

Theory. NNPDF theory 40000000, read from its theory card [3]: NNLO QCD, the FONLL-C heavy-quark scheme, $Q_0 = 1.65$ GeV, $\alpha_S = 0.118$ at 91.2 GeV, pole masses $m_c = 1.51$, $m_b = 4.92$ and $m_t = 172.5$ GeV, target-mass corrections on, no photon distribution. The card allows a fitted charm distribution, and the port sets charm to zero at Q_0. A check of this mismatch was pre-registered (PPDF-30b) and not run: dropped from the job queue by oversight according to the ledger, deferred because it needs a model change according to the state file. The effect is unmeasured.

Data. Phase A began with four fixed-target data sets, 648 points after cuts (PPDF-1), and moved to 19 DIS data sets. Phase B added the rest of the NNPDF4.0-like selection in three steps, called rungs (PPDF-30):

RUNG	ADDED	DATA SETS	POINTS	HADRONIC TABLES
R0	all deep-inelastic data	19	3,089	0
R1	Drell–Yan, W and Z	54	3,981	59
R2	jets, dijets, direct photon	59	4,534	64
R3	top pairs and single top	82	4,615	101

The published NNPDF4.0 fit has 4,618 points at this order [5]. Theory, model and cuts are the same on every rung, so a change from one rung to the next is attributable to the data. The pre-registration's own table places the top data in R2; the runs used the split shown, which is the one in the project's code.

Loss. The fits minimise

$$L(\theta) = \chi^2(\theta) + P(\theta), \qquad \chi^2(\theta) = \big(D - T(\theta)\big)^T C_{t_0}^{-1} \big(D - T(\theta)\big), \tag{3.4.1}$$

where D is the data vector, $T(\theta)$ the prediction, C_{t_0} the t_0 covariance matrix of chapter 2.2 and P the positivity and integrability penalty. Phase B uses NNPDF's 17 positivity and 2 integrability constraints. For Phase A the record is not uniform: the Phase B ledger says in one place that its fits ran with the penalty off and in another, like the pre-registration, that they used about five constraints. The ledger does not give the split between the two terms at the minima of PPDF-31 and PPDF-35; it labels the PPDF-31 fits "penalties ON".

Prior. Each parameter has a flat prior on an interval declared in Phase A before any fit, centred on the publication values of the reference code (PPDF-12). Phase B kept the intervals.

Closure tests. Pseudo-data are generated from planted parameters θ^*:

$$\text{level 0:} \quad D = T(\theta^*), \qquad\qquad \text{level 1:} \quad D = T(\theta^*) + \eta, \quad \eta \sim \mathcal{N}(0, C). \tag{3.4.2}$$

3. What counts as a flat direction

The central quantity of the project is a count: how many of the 52 directions the data leave unconstrained. The definition needs a scale, and the scale is the prior.

Let $H_{ab} = \partial^2\chi^2/\partial\theta_a\partial\theta_b$, in the convention of chapter 2.3 where the Gaussian covariance is $2H^{-1}$. Let w_a be half the width of the prior interval of parameter a, and measure displacements from a reference point θ^0 in units of each parameter's own half-width, $\theta_a = \theta^0_a + w_a\,t_a$. The curvature in those units is

$$H'_{ab} = w_a\,H_{ab}\,w_b, \qquad H'\,v_k = \lambda'_k\,v_k. \tag{3.4.3}$$

H' is dimensionless, and its eigenvalues do not change if a parameter and its bounds are rescaled together. A step of length t along the unit eigenvector v_k costs (derivation D1)

$$\Delta\chi^2 = \tfrac{1}{2}\,\lambda'_k\,t^2. \tag{3.4.4}$$

The one-standard-deviation width that the data alone give along v_k is the step at which $\Delta\chi^2 = 1$. The project calls a direction flat when that width exceeds one prior half-width:

$$\sigma'_k = \sqrt{\frac{2}{\lambda'_k}} > 1 \quad\Longleftrightarrow\quad \lambda'_k < 2. \tag{3.4.5}$$

In a flat direction the prior box is narrower than the data's own constraint. What a Bayesian fit returns there follows from the same two numbers. For a direction aligned with the box, the posterior is a Gaussian of width σ' cut off at ±1, with standard deviation (D1)

$$s(\lambda') = \sigma'\left[1 - \frac{2a\,\phi(a)}{2\Phi(a) - 1}\right]^{1/2}, \qquad a = \frac{1}{\sigma'} = \sqrt{\frac{\lambda'}{2}}, \tag{3.4.6}$$

where ϕ and Φ are the unit Gaussian density and its integral. For $\lambda' \gg 2$ this is σ': the data decide. For $\lambda' \to 0$ it is $1/\sqrt{3} = 0.577$: the prior decides, as in chapter 2.3. At $\lambda' = 2$ it is 0.540, within 7% of the prior value. The threshold is a convention tied to the box, and the ledger declines to widen a box for that reason: the counts would no longer be comparable.

Two further facts are used below. At the minimum of a level-0 fit the residuals vanish and

$$H = 2\,J^T C^{-1} J, \qquad J_{ia} = \frac{\partial T_i}{\partial\theta_a}, \tag{3.4.7}$$

which is twice the Fisher information and has no negative eigenvalue. The ill-conditioning is measured by

$$\kappa' = \frac{\lambda'_{\max}}{\lambda'_{\min}}. \tag{3.4.8}$$

4. Validation before any comparison

A comparison of three methods is also a comparison of three pieces of code. These checks separate the two.

- **Colibri's own closure tests reproduce.** The level-0 and level-1 tests of the Colibri paper [4], for all three routes, on 648 points with a 13-parameter model: "6/6 cells, all consistent" (PPDF-1 to PPDF-6).
- **The three implementations agree where they must.** On a two-parameter problem with an exact posterior from a grid, the widths were 0.99 (Hessian), 1.00 (Monte Carlo) and 0.97 (Bayesian) of the exact one. Made extremely degenerate, it reproduced each failure with the truth known: an indefinite Hessian, a Monte Carlo band 1.32 times too wide with 81% of replicas at the prior bound, and a Bayesian band at 0.19 of the exact width from a chain with an effective sample size (ESS) near 22, against 0.93 at ESS near 221 (PPDF-24).
- **The 52-parameter model closes on 648 points.** At level 1 the minimum was $\chi^2/N_{\mathrm{dat}} = 0.961$ and the posterior mean 0.979. The Bayesian complexity of chapter 2.5 was 10 to 12 of 52 (PPDF-11): four DIS data sets constrain about a fifth of the form.
- **The fit has an external anchor.** On real data the minimum on R3 is $\chi^2 = 5707.67$, $\chi^2/N_{\mathrm{dat}} = 1.2368$ (PPDF-31). A second run gave 5713.228 and 1.2380 (PPDF-35), 0.097% apart. NNPDF4.0's published value is 1.17, a ratio of 1.058, and the pre-registered pass mark was 1.35. The state file warns against an efficiency claim: NNPDF's 763 parameters are regularised network weights, and the project's own effective number of degrees of freedom is about 10 to 14.
- **Reproducibility has a measured limit.** One closure card run on 64 and on 128 processor cores returned minima of 3068.8573 and 3090.4757, 0.7% apart. The ledger traces this to the order of floating-point sums, amplified by the condition number. Values of χ^2 from different core counts should not be differenced.

5. Phase A: deep-inelastic data

Phase A established the disagreement and measured its cause. Several of its first statements were later corrected.

648 points

The Bayesian fit to real data reached $\chi^2/N_{\mathrm{dat}} = 1.002$ with complexity 13.73 (PPDF-12). Gradient descent found 0.9845, and a finite-difference Hessian there had 11 negative eigenvalues, recorded as "consistent with zero" at the resolution of the differencing (PPDF-14). Where the data constrain the fit, the gluon at intermediate x, the Bayesian and Monte Carlo widths agreed to a median ratio of 1.8. Elsewhere the ratio ran from 0.6 to 10, and the Hessian was recorded as unusable (PPDF-15).

All deep-inelastic data

With 19 data sets the level-1 closure test returned correct χ^2 statistics and a pointwise coverage of 31 to 40% (PPDF-16a; its point count was first recorded as about 3,287 and later corrected to 3,089). The verdict was that the form is under-constrained by DIS data (PPDF-16); chapter 2.5 gives the sampler's side.

The degeneracy was then measured without a sampler, as the spectrum of Eq. (3.4.7) at the planted truth (PPDF-19). The record of that count is inconsistent.

SOURCE	DIS ONLY	WITH DRELL–YAN, W AND Z
Phase A ledger, PPDF-19	16 flat	12 flat (73 data sets)
Phase B documents, quoting Phase A	14 flat	12 flat
Phase B, central differences (section 6)	12 flat	9 flat (54 data sets)

The Phase A entry counts eigenvalues "below prior-constraint floor" and quotes condition numbers of 2.8×10^{13} and 2.7×10^{13}; the Phase B documents use Eq. (3.4.5) and quote about 10^{11}. The ledgers do not reconcile 16 with 14. The last row is current. All three support the conclusion drawn at the time: Drell–Yan data remove some flat directions, not all.

The eigenvectors named the sector (PPDF-20). The flattest directions were dominated by the parameters `rho_*`, then `sp_*` (the strange sea), then large-x down valence. The ledger glosses `rho_*` as the $s - \bar{s}$ asymmetry. In the port's code, as read for chapter 2.1, the eight `rho` parameters describe the ratio $\bar{d}/\bar{u}$, and the asymmetry has the two parameters `sm_*`. This chapter uses the parameter names. With 15 high-order Chebyshev coefficients held fixed, the "severe" flat directions, those with $\sigma > 10$, fell from 5 to 0.

Real data, and two corrections

On real DIS data the first three-way comparison found no positive-definite Hessian covariance, a Bayesian band of 2 to 17% of the central value from a chain with ESS near 24, and a Monte Carlo band of 2 to 5 times the value: a ratio of 12 to 144 (PPDF-23). Two later entries revised the reading in opposite directions.

- **PPDF-25** repeated the closure test with five noise draws and converged chains. Mean coverage was about 58%, and the ledger records the earlier collapse as a sampling artefact of an unconverged chain. Coverage on a single draw swung between 47% and 76%. The robust residue was T_8, the combination that carries the strange distribution, at a mean of 38%.
- **PPDF-28** converged the real-data chain (median ESS 181, minimum 55) under a metric fixed in advance. The relative widths, 0.021 for the singlet, 0.038 for the gluon, 0.149 for valence and 0.169 for T_8, were identical to those of the ESS-24 chain. The narrow real-data band is the posterior and not an artefact.

So only the collapse in the closure test was an artefact: "the 3-method divergence is real". The Phase B ledger carries the spread forward by flavour as 144 (singlet), 80 (gluon), 12 (valence) and 33 (T_8), and the converged closure coverage as 62, 56, 76 and 38%, mean 57.9%. Whether more data cure the T_8 under-coverage stayed open: the run that would have answered it ended at ESS 30 and no claim was made (PPDF-26).

---

6. Phase B: the ladder of flat directions

Phase B asked the open question in a form that needs no sampler: count the flat directions of Eq. (3.4.5) at the planted truth of a level-0 fit, on each rung, with the penalty switched off. The reading was fixed in advance. Eight or fewer flat directions on R3 would mean the global data cure the degeneracy; 12 or more would mean it is structural to the form [3].

The first measurement and its flaw

The first ladder read 14, 12, 10, 10, with 7, 6, 6 and 6 negative eigenvalues (PPDF-32, entry of 5 August). It was published. The matrix was built by forward differences of the gradient with a step of 10^{-4} prior half-widths, a choice the ledger says "nobody tested".

The test is to vary the step. At a level-0 minimum Eq. (3.4.7) holds, so a negative eigenvalue cannot be curvature, and a real eigenvalue cannot move with the step. The scan on R0 gave:

STEP / PRIOR HALF-WIDTH	FORWARD: NEGATIVE	FORWARD: FLAT	CENTRAL: NEGATIVE	CENTRAL: FLAT
10^{-2}	13	17	4	12
10^{-3}	10	15	0	11
10^{-4}	7	14	0	12
10^{-5}	5	13	0	12
10^{-6}	2	12	not run	not run

Nothing in the forward columns has converged. The published 14 is the value at one step.

Why forward differences did this

Chapter 2.3 derives the leading errors: a forward difference with step h returns $H + E$ with E of order h, built from third derivatives of χ^2; a central difference has an error of order h^2. What matters here is the effect of E on a *group* of small eigenvalues. Let V_s hold the eigenvectors of the soft eigenvalues Λ_s, those far below the rest of the spectrum. To first order in E the estimated soft eigenvalues are the eigenvalues of (D2)

$$S = \Lambda_s + V_s^T E\, V_s, \qquad E_{ab} = \frac{h}{4}\left(\frac{\partial^3\chi^2}{\partial\theta_a\,\partial\theta_b^2} + \frac{\partial^3\chi^2}{\partial\theta_b\,\partial\theta_a^2}\right). \tag{3.4.9}$$

The second term need be neither small compared with Λ_s nor positive. Each estimated soft eigenvalue therefore behaves as

$$\lambda_k^{\mathrm{est}}(h) = \lambda_k + c_k\, h \quad \text{(forward)}, \qquad \lambda_k^{\mathrm{est}}(h) = \lambda_k + d_k\, h^2 \quad \text{(central)}, \tag{3.4.10}$$

with c_k and d_k set by higher derivatives. The scan confirms two consequences.

- **Negative eigenvalues proportional to the step.** The smallest forward eigenvalue was -1.354×10^{2}, -1.281×10^{1} and -1.210 at the last three steps, a factor of 10.6 per decade. The number of negative eigenvalues falls with h because fewer of the small positive λ_k are overtaken by $|c_k|h$.
- **An inflated flat count.** An eigenvalue slightly above 2 is counted as flat when $c_k h$ pulls it below 2. The count falls from 17 to 12 as the step shrinks.

With central differences the smallest eigenvalue is $+3.695 \times 10^{-6}$ at 10^{-4} and $+2.583 \times 10^{-6}$ at 10^{-5}, and the flat count is 12 at both. The cost is 104 gradient evaluations, about a second.

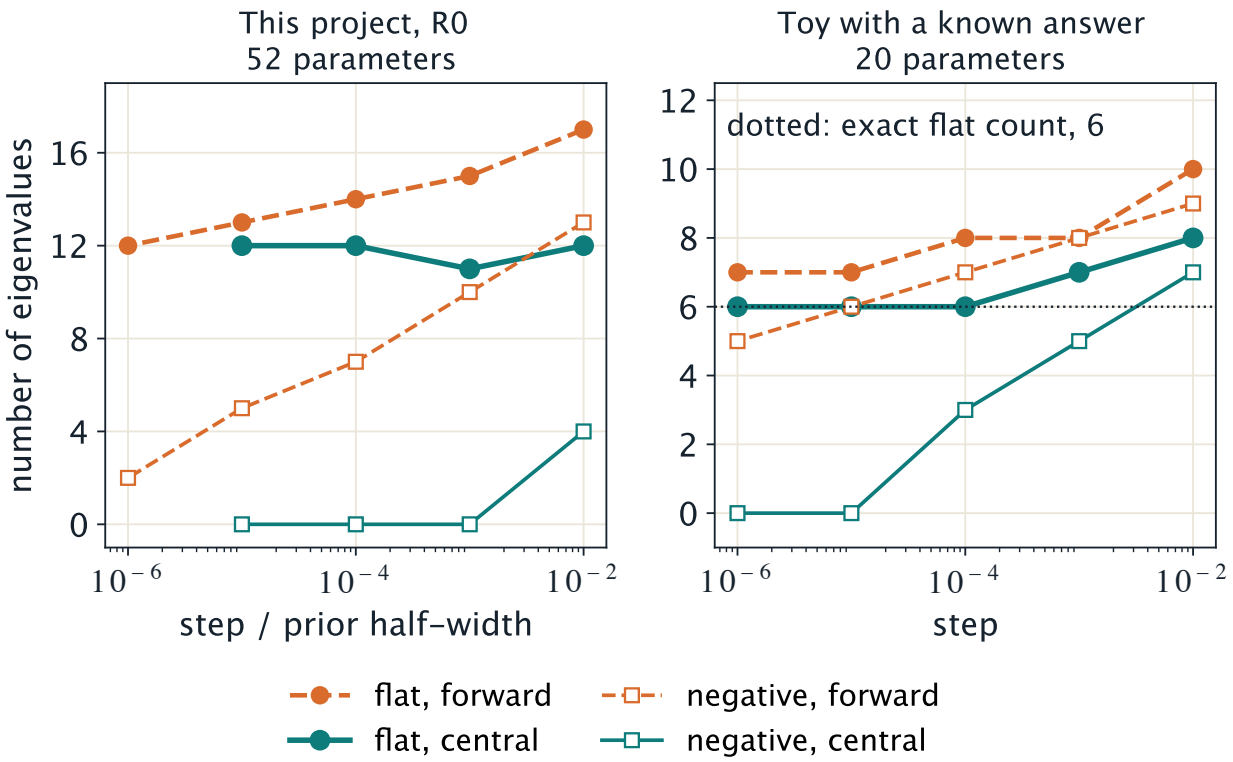

Figure 1. Counts of flat eigenvalues (filled circles) and negative eigenvalues (open squares) against the differencing step; orange dashed for forward differences, teal for central. Left: the project's scan on R0, from the ledger. Right: a toy with 20 parameters, a positive-definite Hessian of condition number 10^{11} and exactly 6 flat directions, from the check script.

The corrected ladder

All four rungs were measured again by central differences, each at two steps that agree (PPDF-32, entry of 6 August):

RUNG	DATA SETS	FORWARD (SUPERSEDED)	CENTRAL	NEGATIVE, CENTRAL
R0	19	14	12	0
R1	54	12	9	0
R2	59	10	8	0
R3	82	10	8	0

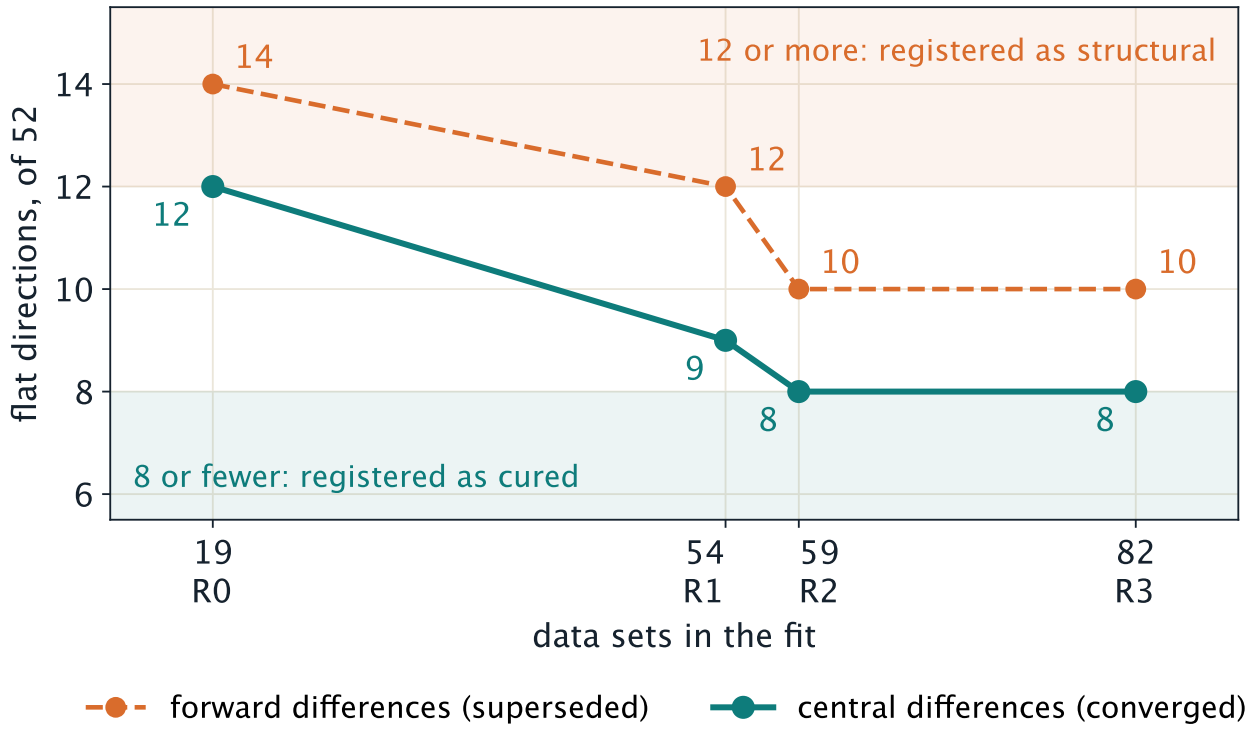

Figure 2. The ladder of flat directions (PPDF-32). Teal: the converged central-difference counts. Orange: the forward-difference counts they replace. The shaded bands are the thresholds registered before the measurement.

What the correction changed and what it left:

- **The saturation survives.** R2 and R3 agree at 8: the 23 top-quark data sets remove no flat direction, in either version. This was the claim.
- **The registered verdict changed.** At 10 the result read as partial. At 8 it lies exactly on the "cured" threshold. The ledger reports it as a boundary result and not a pass: 8 of 52 directions remain unconstrained.
- **A secondary claim was wrong by a factor of three.** Per data set added, jets and photon data had been called seven times as effective as Drell–Yan data (0.400 against 0.057). The corrected figures are $1/5 = 0.200$ and $3/35 = 0.086$, a ratio of 2.3.
- **A Phase A headline changed.** "14 of 52 flat on DIS" became 12.

An addendum to the pre-registration records the correction and leaves the gates as written [3].

The real-data Hessian

The Phase B ledger attributes the negative eigenvalues reported in Phase A for real data (10 and 11 of 52) to forward differencing. The Phase A entry for the 11 (PPDF-14) records central differences of likelihood values. Central differences of the gradient at a real-data minimum on R0 gave a smallest scaled eigenvalue of -7.089×10^3, -4.724×10^4 and -3.054×10^3 at three steps: a swing of 15, not monotonic (PPDF-36). Equation (3.4.10) omits rounding. A gradient is a sum of large cancelling terms, each known to relative precision ϵ, so a difference quotient carries an error of order

$$\delta\lambda \sim \frac{\epsilon\,|G|_{\text{terms}}}{h}, \tag{3.4.11}$$

where $|G|_{\text{terms}}$ is the size of those terms. It grows as the step shrinks. At a level-0 minimum the terms vanish with the residuals; at a real-data minimum they do not. The ledger concludes that the signs "can be neither confirmed nor refuted", and that the case against a Hessian band rests on what is stable: $\kappa' = 1.6\text{–}1.8 \times 10^{11}$ and 11 to 12 flat directions at all three steps.

Counts that share a name

The ledgers hold other counts on the same 19 data sets: 7 flat and 45 stiff in the analysis of the retracted run of section 9, 14 to 16 directions floored at the prior in section 8, and 14, 15 and 20 eigenvalues below the prior precision at three candidate truths. They use other points, losses or normalisations, and one entry compares its 14 with the superseded 14 of the ladder. Only the ladder's counts are converged and comparable across rungs.

---

7. Which data hold the flat directions down

The ladder adds data in a fixed order. A second measurement removes one class of data from R3 and recounts (PPDF-38).

CLASS REMOVED FROM R3	DATA SETS	FLAT OF 52	CHANGE	PER DATA SET
neutrino DIS (CHORUS, NuTeV)	4	9	+1	0.250
jets and dijets	4	9	+1	0.250
charged-lepton DIS	15	10	+2	0.133
Drell–Yan, W and Z	33	10	+2	0.061
top and single top	23	8	0	0.000

All five counts are converged. The classes add up to 79 data sets, not 82: the Drell–Yan and jet classes have 33 and 4 members where the rungs add 35 and 5, and the ledger does not say where the other three sets went. In the project's code the fifth set of the jet rung is the direct-photon measurement.

Incremental and marginal counts

Information adds over independent data: by Eq. (3.4.7) the level-0 Hessian of a union of classes is

$$H = \sum_c H_c, \qquad H_c = 2\, J_c^T C_c^{-1} J_c. \tag{3.4.12}$$

Flat counts do not add. Write $n(\mathcal{S})$ for the number of flat directions when the set of classes $\mathcal{S}$ is fitted. The ladder measures an *incremental* gain and the leave-one-out scan a *marginal* one:

$$\Delta_c^{\text{inc}} = n(\mathcal{S}_{<c}) - n(\mathcal{S}_{<c} \cup c), \qquad \Delta_c^{\text{mar}} = n(\mathcal{A} \setminus c) - n(\mathcal{A}), \tag{3.4.13}$$

where $\mathcal{S}_{<c}$ is what was present when c was added and $\mathcal{A}$ is everything. If "flat" meant an exactly zero eigenvalue, then for any class (D3)

$$\Delta_c^{\mathrm{mar}} \;\le\; \Delta_c^{\mathrm{inc}}. \tag{3.4.14}$$

A class gains least when everything else is already there. The ledger's numbers obey this: Drell–Yan data are worth 3 incrementally and 2 marginally, jets 1 and 1, top 0 and 0. Only the top class is the same set of data in both; the other two differ by the three unassigned sets. A class whose information is duplicated elsewhere scores high in the ladder and low in the scan, so the ledger forbids tabulating the two together.

The ledger treats the top row as a null control that "passes exactly", and calls the ladder and the scan two independent methods. Equation (3.4.13) qualifies that. For the last class added, $\mathcal{S}_{<c} = \mathcal{A} \setminus c$ and the two quantities are one number: by the rung definitions in the project's code, R3 without its 23 top data sets is R2. The ledger's entry names the same script, `fisher_eps_scan.py`, for both, with a separate generator for the run cards of the scan. The agreement shows that the two ways of building the run card give the same input. It is not a second measurement.

Neutrino DIS, the only class sensitive to the strange distribution, and jets each hold down one direction that nothing else constrains, about four times the Drell–Yan figure per data set. The ledger cautions that one direction is a small effect. The five marginal changes sum to 6, against the $52 - 8 = 44$ directions the full data set constrains (our arithmetic): most of the constraint is held redundantly.

The experiment that could not be run

The surviving flat directions sit in the strange and sea sector, and the standard direct probe of strangeness, W production with charm, is not among the 82 data sets. A fifth rung with five such data sets was specified (PPDF-37). It cannot be evaluated. Theory 40000000 ships no FK tables for these measurements; tables exist only in theories of the NNPDF3.1 era, which Colibri's prediction code cannot contract. New grids must be generated. Whether the residual degeneracy is a data deficit or a property of the form is therefore open.

One number from this search stood wrong for several days. The ledger first recorded that 88 of the 131 data-set definitions in the theory lacked FK tables. Counted properly, 7 of 131 lack them; the first count had looked at only one of the two file formats in the directory. (An earlier table gives 132 definitions.) Missing grids are not why the fit has 82 data sets.

---

8. Monte Carlo and Hessian bands on the closure test

Phase B compared the methods on level-1 closure data. This section gives the two arms that produced numbers and the diagnostics that decided which could be kept.

A truth that the fit can live with

The first planted truth was the published MSHT20 parameter point. On R0 its data term at level 1 was 3,078.65 and its penalty 70,729.11, a ratio of 22.97. A posterior penalised that strongly away from the truth cannot cover it. Three candidates were evaluated on real R0 data (PPDF-33):

CANDIDATE TRUTH	FIT QUALITY ON REAL DATA	PENALTY
published MSHT20 point	14.4	70,729.11
projected: least penalty, close to the published point	51.4	0.084
penalised best fit, clipped to the prior box	1.143	9.014

The second row moved the parameters by a median of 1.8% of a prior half-width and made the description of real data much worse, because its one large move was along a stiff direction. With $\kappa' \sim 10^{11}$, distance in parameter space says nothing about distance in predictions. The third candidate was planted. Unclipped, it had `g_etam`, the power of $(1-x)$ in the second term of the gluon, outside its prior; holding it inside cost $\Delta\chi^2 = 0.256$ for a move of 0.85 half-widths, which by Eq. (3.4.4) is a flat direction with $\lambda' \approx 0.7$. The R3 truth has two parameters on the clip margin, a caveat recorded before the R3 coverage existed.

The Monte Carlo arm

The R0 ensemble has 39 usable replicas of 40, the R3 ensemble 20 of 20 (PPDF-33).

QUANTITY	R0	R3
parameters whose one-sigma interval contains the truth	39 of 52, 75.0%	43 of 52, 82.7%
Monte Carlo width over prior width, median	0.646	0.600
parameters with Monte Carlo width above the prior width	6	1

A coverage estimate from n independent trials has the binomial error of chapter 2.3,

$$\hat{c} = \frac{n_{\mathrm{in}}}{n}, \qquad \sigma_{\hat{c}} = \sqrt{\frac{c\,(1-c)}{n}} = 6.5 \text{ points for } n = 52,\ c = 0.6827. \tag{3.4.15}$$

The two results are 1.0 and 2.2 of these errors above nominal. That arithmetic is ours, not the ledger's, and it overstates the significance: 52 correlated parameters are fewer than 52 trials, and each number comes from one noise draw. The ledger states that the band over-covers, as in Phase A, and that the R0 value is stable against ensemble size (73.1% at 29 replicas). It also says the two clip-margin parameters are excluded from coverage statements, yet 43 of 52 has all 52 in its denominator.

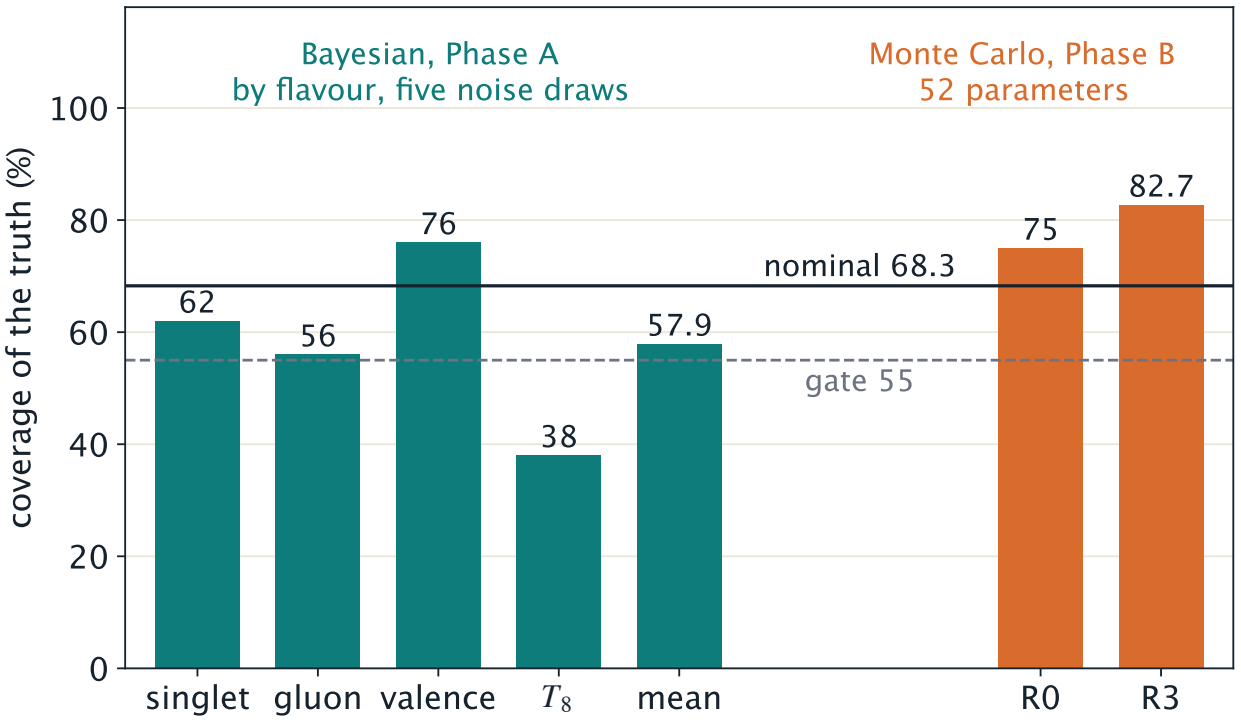

Figure 3. The coverage numbers that survive in the ledgers. Teal: converged Bayesian sampling on DIS closure data in Phase A (PPDF-25, as carried forward in the Phase B ledger). Orange: Monte Carlo replicas in Phase B (PPDF-33). The two groups use different metrics. No Hessian bar and no Phase B Bayesian bar exist.

The Hessian arm and the pull test

Colibri's own Hessian action ran out of memory at this scale, so the band was built from the central-difference Hessian at the best fit $\hat{\theta}$, with any direction softer than the prior given the prior width. A first coverage figure from three noise draws, 78.8% on average, did not survive two checks.

The first split the directions. The data-constrained ones covered the truth 16.0% of the time and the prior-floored ones 93.8%. The healthy average came from directions that had been handed the prior width.

The second explains the 16.0%. For a level-1 fit of a correct model the best fit scatters about the truth with the covariance the Hessian reports (chapter 2.3), so the pull of the truth along eigen-direction k,

$$z_k = \frac{v_k \cdot (\theta^* - \hat{\theta})}{\sigma_k}, \tag{3.4.16}$$

with $\sigma_k = \sqrt{2/\lambda_k}$ the Hessian width, is a unit Gaussian. If the expansion point is wrong by b standard deviations, the interval $\pm r\sigma$ about it covers the truth with probability (D4)

$$c(b, r) = \Phi(r - b) - \Phi(-r - b), \tag{3.4.17}$$

which for $r = 1$ is 68.3% at $b = 0$, 47.7% at $b = 1$ and 15.7% at $b = 2$. And the largest of 52 unit Gaussian pulls is small:

$$P\Big(\max_k |z_k| \geq a\Big) = 1 - \big[2\Phi(a) - 1\big]^{52}, \tag{3.4.18}$$

with a median of 2.5 and a probability of 3×10^{-5} of exceeding 5. The largest pulls on the three draws were 72.1, 117.8 and 102.9. The optimiser had stopped short of the minimum, and in a stiff direction, where σ_k is tiny, a small shortfall is an enormous pull. The 16.0% measured the optimiser.

The analysis therefore refuses a Hessian coverage statement when $\max |z| \geq 5$. The largest pull was 5.3 on R0 and 57.7 on R3, so none is made. Widths are quoted for scale only. On R0 the ratio of Monte Carlo to Hessian width over the 52 parameters has median 1.09, minimum 0.44 and maximum 31.5 (`sea_c6`), and four of the six largest ratios are sea or strange parameters. On R3 the ledger withdraws the ratio: its maximum of 767 came from the unconverged expansion point.

The guard itself failed once. Four probe runs had been centred on the truth, so their pulls were exactly zero, they passed a test that asked only for small pulls, and they reported 100% coverage, which an interval centred on the truth has by construction. The test now also requires a non-zero distance from the truth, and the usable expansion points went from 4 of 11 to none. Each wrong number, the ledger notes, was caught by a consistency argument and never by an error message.

---

9. The Bayesian arm and its gate

Chapter 2.5 tells the story of the sampler. This section records the gate and its outcome, and derives the two checks that separated a valid result from a void one.

The gate. The pre-registration reads: “ESS ≥ 100 median, R̂ < 1.05, divergences < 5%. Coverage is not reported below this bar” [3], where $\hat{R}$ is the between-chain statistic of chapter 2.5. The ledger applies the ESS threshold to the worst parameter and states that the gate is unchanged. The two readings differ. The outcome below fails on $\hat{R}$ under either.

One chain. On the R0 closure data a chain of 1,200 draws, with the NUTS sampler of chapter 2.5 preconditioned by the Hessian at the best fit, had 0.25% divergences, a smallest ESS of 3.2 and a split $\hat{R}$ of 2.4545. A chain of 37,500 draws gave 3.0 and 3.0830. The ledger sets the 3.2 beside an ESS of 226 from Phase A and reads it as the cost of going from 648 to 3,089 points. That needs two qualifications. The Phase B ledger says that Phase A’s whitened sampler was preconditioned differently, from a pilot nested-sampling run, and the Phase A ledger does not say which sampler produced the ESS of 226. Nor does it give the point count of that run, while it describes its converged real-data chain (PPDF-28) as on full DIS data.

Four chains that were four data sets

A run of four chains followed, with seeds 202, 303, 404 and 505. It reported a combined ESS of 111.2, meeting the gate, and a four-chain $\hat{R}$ of 1.894, failing it, and a reading: the chains had settled in different regions because the sea directions are flat. All of it was withdrawn the next day.

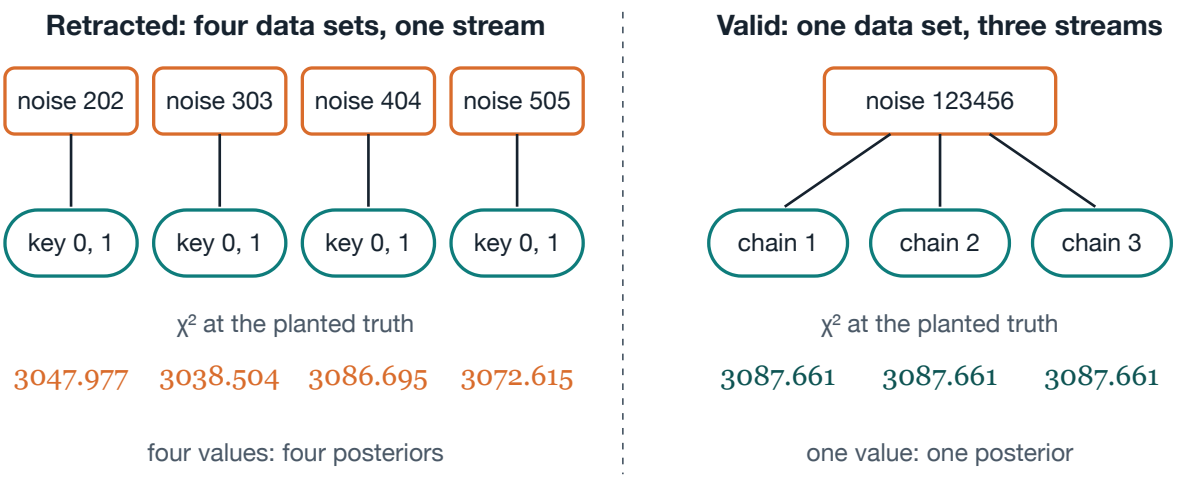

Figure 4. The retracted and the valid multi-chain runs (PPDF-33). Orange boxes are pseudo-data sets, labelled by the seed of the noise draw; teal boxes are Markov chains, labelled by the seed of the sampler. The value of χ^2 at the planted truth, printed by every run, distinguishes the two designs.

The proof uses a quantity that depends on the noise and on nothing else. At level 1, by Eq. (3.4.2), the residual at the planted truth is the noise itself:

$$\chi^2(\theta^*) = \eta^T C^{-1} \eta. \tag{3.4.19}$$

The model, the truth and the sampler have dropped out. Over noise draws this is a χ^2 variable with N_{dat} degrees of freedom (D5),

$$\langle \chi^2(\theta^*) \rangle = N_{\mathrm{dat}}, \qquad \mathrm{sd} = \sqrt{2N_{\mathrm{dat}}} = 78.6 \ \text{ for } N_{\mathrm{dat}} = 3089, \tag{3.4.20}$$

so chains that share a data set print the same value to every digit, and chains on different data sets print values tens of units apart. The four runs printed 3047.977, 3038.504, 3086.695 and 3072.615. The script's seed argument had set the noise draw, and the sampler's keys were fixed in the code. Pooled ESS over four posteriors has no meaning, and $\hat{R}$ across them measures how far the best fit moves between noise draws; chapter 2.5 derives $\hat{R} \approx \sqrt{2}$ for that case.

The invariant had already served once: when the planted truth was changed, the data term at the truth stayed at 3078.647. The valid run's 3087.661 equals that number plus the penalty of 9.014; the ledger does not state the decomposition.

The run had also been combined wrongly. The first figure, 71.9, summed each chain's smallest ESS. For chains c and parameters a (D5),

$$\min_a \sum_c n_{\mathrm{eff},c}(a) \ \geq \ \sum_c \min_a n_{\mathrm{eff},c}(a), \tag{3.4.21}$$

where $n_{\mathrm{eff},c}(a)$ is the ESS of parameter a in chain c. The worst parameter differs from chain to chain, and the left side is the correct combined figure.

The valid test

The corrected run fixed the noise seed, varied the chain seed, asserted equality of $\chi^2(\theta^*)$ across chains, and ran three chains of 6,000 draws (PPDF-33, 8 August).

CHAIN SEED	SMALLEST ESS	SPLIT R-HAT	DIVERGENCES
1	11.2	1.2495	4.30%
2	12.0	1.2300	0.45%
3	6.3	1.3207	0.40%

Combined by the left side of Eq. (3.4.21) the ESS is 30.8. The three-chain $\hat{R}$ has a maximum of 1.1185 over the 52 parameters and a median of 1.0296; the pooled median ESS is 222. The gate is not met. From the definition in chapter 2.5,

$$\hat{R}^2 - 1 \simeq \frac{B/n}{W}, \tag{3.4.22}$$

the variance of the chain means in units of the within-chain variance. At 1.1185 it is 0.25: the chain means differ by half a standard deviation in the worst parameter. With nothing changed except that the chains share a data set, the excess of $\hat{R}$ over one fell from 0.894 to 0.119. The retracted number had measured the noise draws.

ESS adds over chains on one posterior. The ledger estimates that the gate needs about 10 chains at about 4.2 hours each; in parallel that is 1,280 processor cores against a quota of 200. Under the pre-registered rule no posterior band is quoted, and the five-draw coverage test, PPDF-34, was not attempted.

Which parameters. The retraction also removed a physical claim, that between-chain disagreement sits in the sea and strange sector with the flat directions. In the valid run the smallest ESS belongs to `rho_c2` (30.8), `rho_c3` and `rho_c1`, and the largest $\hat{R}$ to `uv_c2`, `uv_c3` and `rho_c4`. The ledger's last word is that flatness and poor mixing are separate problems in separate sectors. The state file's summary still names up-valence alone for the sampler, from the single chain; its table gives the later result.

10. What is established, what is not, and what is open

The table sorts the claims by their standing in the ledgers.

STANDING	CLAIM	RESULT
Established	the 52-parameter form fits the global data, 1.2380 per point	PPDF-31, 35
Established	flat directions 12, 9, 8, 8; no gain from top data	PPDF-32
Established	no negative eigenvalue at a level-0 minimum	PPDF-32
Established	real-data Hessian: condition number 1.6–1.8×10^{11}, no usable band	PPDF-36
Established	Monte Carlo band over-covers on closure data, 75.0% and 82.7%	PPDF-33
Established	Bayesian gate not met at 3,089 points: ESS 30.8, R-hat 1.1185	PPDF-33
Established	on real DIS data, Bayesian and Monte Carlo widths differ by 12 to 144	PPDF-23, 28
Not established	signs of the smallest real-data eigenvalues	PPDF-36
Not established	any Hessian coverage; any width ratio on R3	PPDF-33
Not established	any Bayesian band or coverage in Phase B	PPDF-33, 34
Not established	whether method disagreement grows or shrinks with global data	PPDF-33
Not established	whether the 8 residual directions are a data deficit or structural	PPDF-37, 38
Not established	the effect of the charm mismatch	PPDF-30b
Retracted	four-chain ESS 111.2 and R-hat 1.894, and the mechanism built on them	PPDF-33
Corrected	ladder 14, 12, 10, 10; “seven times”; 88 of 131 missing tables	PPDF-32, 37

The state file lists three experiments that need no outside input [1]. None has been run.

Do the flat directions move the distributions? A flat direction matters only if a physical quantity changes along it. For a quantity f, such as a distribution at given x and Q^2, with gradient $g_a = w_a\,\partial f/\partial\theta_a$ in scaled units, the variance separates over the eigenvectors of Eq. (3.4.3) (D6):

$$\operatorname{var} f = \sum_k \left(g \cdot v_k\right)^2 s_k^2, \qquad s_k^2 = \begin{cases} 2/\lambda'_k & \text{constrained} \\ \leq 1 & \text{flat, bounded by the box.} \end{cases} \tag{3.4.23}$$

The experiment evaluates $\delta f_k = |g \cdot v_k|\, s_k$ for the 8 flat directions across x. If every δf_k is negligible, the degeneracy is internal to the parametrisation and the methods should agree on f even where they disagree on θ. If

not, Eq. (3.4.23) says where in x and in which flavour they must disagree.

Freeze and refit. Fixing m parameters deletes m rows and columns of H'. The eigenvalues μ_i of what remains interlace those of the full matrix (D6):

$$\lambda'_i \le \mu_i \le \lambda'_{i+m} \qquad \Longrightarrow \qquad n - m \;\le\; n_{\text{frozen}} \;\le\; n, \tag{3.4.24}$$

where n is the flat count before. Freezing never adds a flat direction, and removing all 8 requires freezing at least 8 parameters, chosen to span the flat eigenvectors. The planned experiment freezes 8, refits 44 and asks whether the Hessian becomes invertible, whether the sampler meets its gate and whether the fit quality survives. MSHT do the same when they fix parameters before making eigenvector sets [6], at the same price: the frozen directions contribute nothing to the band.

Measure the tolerance. Suppose a Hessian band at a converged minimum covers a planted truth with frequency c_1 at $T = 1$. If the pulls are Gaussian with a common inflated width, the factor that restores nominal coverage is (D6)

$$T = \frac{1}{\sqrt{2}\,\operatorname{erf}^{-1}(c_1)}. \tag{3.4.25}$$

A coverage of 26.1% would give $T^2 = 9$. Closure data are consistent by construction, so a T^2 above 1 there would measure the failure of the quadratic approximation alone, apart from the data inconsistency behind the tolerances of chapter 2.4. The experiment needs an expansion point that passes the pull test of Eq. (3.4.18). No run has yet produced one.

Two further items need people, not compute: W-plus-charm grids at theory 40000000, and a larger processor quota for parallel chains.

Derivations

D1. The scaled curvature and the width of a bounded direction

Derives (3.4.4), (3.4.5), (3.4.6)

The quadratic form. Near a stationary point, $\Delta\chi^2 = \tfrac{1}{2}\,\delta^T H\,\delta$ with $\delta = \theta - \hat\theta$. Write $\delta_a = w_a\,t_a$. Then $\Delta\chi^2 = \tfrac{1}{2}\sum_{ab} t_a\,(w_a H_{ab} w_b)\,t_b = \tfrac{1}{2}\,t^T H' t$.

Along an eigenvector. For $t = t\,v_k$ with $|v_k| = 1$, $\Delta\chi^2 = \tfrac{1}{2}\lambda'_k t^2$, Eq. (3.4.4). Setting $\Delta\chi^2 = 1$ gives $t = \sqrt{2/\lambda'_k}$, Eq. (3.4.5). The same follows from the covariance $2H'^{-1}$, whose projection on v_k is $2/\lambda'_k$.

Invariance. Replace θ_a by θ_a/s_a. Then $H_{ab} \to s_a H_{ab} s_b$ and $w_a \to w_a/s_a$, so H' is unchanged.

The bounded posterior. With a flat prior on $|t| \le 1$, the posterior density is proportional to $e^{-t^2/2\sigma'^2}$ on that interval. Put $u = t/\sigma'$ and $a = 1/\sigma'$. Integrating by parts, $\int_{-a}^{a} u^2\phi\,du = [2\Phi(a) - 1] - 2a\phi(a)$. Dividing by the normalisation $2\Phi(a) - 1$ and multiplying by σ'^2 gives the variance, whose square root is Eq. (3.4.6). For small a the bracket tends to $a^2/3$, so $s \to 1/\sqrt{3}$. The check script confirms the formula by quadrature.

An example from the ledger. For `g_etam`, Eq. (3.4.4) gives $\lambda' = 2 \times 0.256/0.85^2 = 0.71$. The ledger quotes 0.7.

D2. Forward differences and a cluster of soft eigenvalues

Derives (3.4.9), (3.4.10), (3.4.11)

The error matrix. From chapter 2.3, the forward-difference estimate of column b of the Hessian is $H_{ab} + \frac{h}{2}\,\partial^3\chi^2/\partial\theta_a\partial\theta_b^2$ plus terms of order h^2. Symmetrising gives $H + E$ with E as in Eq. (3.4.9).

Why the one-vector formula fails. For an isolated eigenvalue, chapter 2.3 gives $\delta\lambda_k = v_k^T E v_k$. That assumes the elements $v_j^T E v_k$ are small compared with the spacing $\lambda_j - \lambda_k$. Soft eigenvalues have small spacings, and the assumption fails.

The soft block. Split the eigenvectors into a soft set V_s and a stiff set with eigenvalues of order Λ, far above the soft ones. In that basis $H + E$ has a soft block $\Lambda_s + V_s^T E V_s$, a stiff block, and off-diagonal blocks of order $|E|$. Eliminating the stiff block changes the soft block by a term of order $|E|^2/\Lambda$, which is second order. To first order the soft eigenvalues are those of S in Eq. (3.4.9), which must be diagonalised exactly because its two terms are of the same size.

Scaling. $E \propto h$, so when the second term of S dominates, every estimated soft eigenvalue is proportional to h, with the sign of the corresponding eigenvalue of $V_s^T E V_s$. For central differences the error matrix is of order h^2. This is Eq. (3.4.10).

Check. The toy of Figure 1 has $\chi^2 = |A\,g(\theta) - d|^2 + |B(\theta - \theta^*)|^2$ with $g_j = e^{\theta_j}\theta_{j+10}^2$, noise-free data and an analytic gradient. Its exact Hessian has ten eigenvalues below 33 and the next at 1.7×10^3. At $h = 10^{-6}$ the lowest forward-difference eigenvalue is -80.96; the soft-block formula gives -80.96 and the one-vector formula -3.75. In this toy every eigenvalue of the soft block of E is negative. The formula does not fix the sign in general.

Rounding. A gradient computed in floating point carries an absolute error ϵ times the size of the terms summed in it, and dividing a difference of two gradients by h gives Eq. (3.4.11). Truncation error falls with h and rounding error rises, so a soft eigenvalue is resolved only in a window of steps, which at a real-data minimum can close.

D3. Incremental and marginal counts

Derives (3.4.14)

Exact flat directions. Take "flat" to mean an exactly zero eigenvalue. By Eq. (3.4.12), $v^T H v = \sum_c 2\,|C_c^{-1/2} J_c v|^2$, which vanishes only if $J_c v = 0$ for every class. So $n(\mathcal{S}) = p - \dim U_{\mathcal{S}}$, where $U_{\mathcal{S}}$ is the span of the rows of all J_c in $\mathcal{S}$ and $p = 52$.

The gain from one class. Adding a class with row space U_c to a set with row space U lowers n by

$$\dim(U + U_c) - \dim U = \dim U_c - \dim(U_c \cap U).$$

Monotonicity. If $U \subseteq U'$ then $U_c \cap U \subseteq U_c \cap U'$, so the gain against the larger set is no greater. The incremental gain is measured against $\mathcal{S}_{<c}$ and the marginal one against $\mathcal{A} \setminus c$, which contains it. This is Eq. (3.4.14). If c is the last class added, the two sets are the same and so are the gains.

Limits. With a threshold at $\lambda' = 2$ in place of exact zeros the inequality is not a theorem, because eigenvalues can cross the threshold in either order as classes are combined. The check script verifies the exact statement on 300 random cases and on a four-class example with incremental gains (4, 2, 1, 0) and marginal gains (0, 1, 1, 0).

D4. Coverage with a displaced centre, and the largest pull

Derives (3.4.17), (3.4.18)

Displaced centre. Let $\hat\theta = \theta^* + b\sigma + \sigma\xi$ with ξ a unit Gaussian and b a fixed offset. The interval $\hat\theta \pm r\sigma$ contains θ^* when $-r - b < \xi < r - b$, with probability $\Phi(r-b) - \Phi(-r-b)$. This is Eq. (3.4.17).

Centre at the truth. If $\hat\theta = \theta^*$ identically, the coverage is 100% and carries no information.

The largest of 52 pulls. For independent unit Gaussians, $P(|z_k| < a) = 2\Phi(a) - 1$ for each, and $P(\max_k |z_k| < a)$ is the product, Eq. (3.4.18). Setting the product to one half gives a median of 2.48. At $a = 5$ the probability of exceeding is $52 \times 5.7 \times 10^{-7} = 3 \times 10^{-5}$.

Unit pulls. For a linear model, $\hat\theta - \theta^* = (J^T C^{-1} J)^{-1} J^T C^{-1} \eta$ has covariance $(J^T C^{-1} J)^{-1} = 2H^{-1}$. Projected on eigenvector k and divided by $\sigma_k = \sqrt{2/\lambda_k}$ it has unit variance. The check script confirms Eq. (3.4.17) by simulation, and a parameter coverage of 68.3% with unit pulls for a linear fit over 20,000 noise draws.

D5. The replicate test and the combination of chains

Derives (3.4.19), (3.4.20), (3.4.21), (3.4.22)

The invariant. At level 1, $D - T(\theta^*) = \eta$, and Eq. (3.4.1) gives $\chi^2(\theta^*) = \eta^T C^{-1} \eta$, Eq. (3.4.19). Neither the form of T nor the value of θ^* appears. The check script confirms this for a non-linear model with two different truths and one noise seed.

Its distribution. Write $\eta = C^{1/2} e$ with e a vector of $N_{\rm dat}$ independent unit Gaussians. Then $\chi^2(\theta^*) = |e|^2$, with mean $N_{\rm dat}$ and variance $2N_{\rm dat}$, Eq. (3.4.20). The four retracted values have a range of 48.2. Four draws do not test the width; the point is that they differ at all.

Combining ESS. For each parameter a and chain c, $n_{{\rm eff},c}(a) \geq \min_{a'} n_{{\rm eff},c}(a')$. Summing over chains gives $\sum_c n_{{\rm eff},c}(a) \geq \sum_c \min_{a'} n_{{\rm eff},c}(a')$ for every a, including the one that minimises the left side. This is Eq. (3.4.21). The recorded pairs obey it: 30.8 against $11.2 + 12.0 + 6.3 = 29.5$, and 111.2 against $8.1 + 13.2 + 7.1 + 43.5 = 71.9$.

The meaning of R-hat. From chapter 2.5, $\hat R^2 = \frac{n-1}{n} + \frac{B/n}{W}$. For long chains the first term is 1, which gives Eq. (3.4.22). $\hat R = 1.1185$ gives 0.251 and a spread of chain means of $\sqrt{0.251} = 0.50$ within-chain standard deviations. $\hat R = 1.894$ gives 2.59.

D6. The three open experiments

Derives (3.4.23), (3.4.24), (3.4.25)

Variance along eigenvectors. To first order $\delta f = g \cdot t$ in scaled coordinates, so var $f = g^T \Sigma' g$ with Σ' the covariance of t. In the Gaussian approximation $\Sigma' = 2H'^{-1} = \sum_k (2/\lambda'_k)\, v_k v_k^T$, the first case of Eq. (3.4.23). For a flat direction the box limits the variance, and $s_k \leq 1$ is the half-width itself. The separation is approximate for the bounded part, because the box is not aligned with the eigenvectors.

Interlacing. Let H' have eigenvalues $\lambda'_1 \leq \cdots \leq \lambda'_p$, and let M, with eigenvalues $\mu_1 \leq \cdots \leq \mu_{p-m}$, be what is left after deleting m rows and the same columns. By the min–max characterisation, μ_i is the minimum over i-dimensional subspaces of the retained coordinates of the largest value of the quadratic form on the unit sphere of the subspace. Those subspaces are among the i-dimensional subspaces of the full space, so $\mu_i \geq \lambda'_i$. The same argument for $-H'$ gives $\mu_i \leq \lambda'_{i+m}$. Counting eigenvalues below 2, the first inequality gives $n_{\rm frozen} \leq n$ and the

second $n_{\text{frozen}} \geq n - m$. This is Eq. (3.4.24). The check script confirms both on 400 random matrices, and shows a case in which freezing two stiff parameters leaves both flat directions in place.

Tolerance from coverage. Let the pulls be Gaussian with standard deviation q instead of 1. A band of half-width $T\sigma$ covers with probability $\operatorname{erf}\left(T/(q\sqrt{2})\right)$. At $T = 1$ this is c_1, so $q = 1/\left(\sqrt{2}\,\operatorname{erf}^{-1}(c_1)\right)$, and nominal coverage is restored at $T = q$, Eq. (3.4.25). For $q = 3$, $c_1 = 26.1\%$ and $T^2 = 9$. If the pulls are not Gaussian, or their width differs between directions, no single factor restores coverage everywhere.

References

1. This project, state file (`context/STATE.md`): the specification, the corrections and the open experiments.
2. This project, Phase A results ledger (`context/RESULTS_LEDGER.md`), PPDF-1 to PPDF-28, and Phase B results ledger (`context/PHASE_B_LEDGER.md`), PPDF-30 to PPDF-38, with the dated corrections and the retraction.
3. This project, Phase B pre-registration (`context/PHASE_B_PREREG.md`) and its addendum of 6 August 2026: the theory card, the rungs, the gates and the falsifiers.
4. M. N. Costantini, L. Mantani, J. M. Moore, V. Schütze Sánchez and M. Ubiali, “Colibri: A new tool for fast-flying PDF fits”, Eur. Phys. J. C 86 (2026) 22, arXiv:2510.03391.
5. R. D. Ball et al. (NNPDF), “The path to proton structure at 1% accuracy”, Eur. Phys. J. C 82 (2022) 428, arXiv:2109.02653. The point count and fit quality quoted beside ours are the values recorded in the ledgers.
6. 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.
7. J. M. Moore, *Parton Distributions in Beyond the Standard Model Theories*, PhD thesis, University of Cambridge; M. Ubiali, *A new approach in the determination of Parton Distribution Functions*, PhD thesis (2010); S. Amoroso et al., “Snowmass 2021 whitepaper: Proton structure at the precision frontier”, arXiv:2203.13923. The methods and closure tests, as developed in chapters 2.3 and 2.5.