2.5 Bayesian fits and Colibri

Priors, nested sampling and Hamiltonian Monte Carlo, evidence, and closure tests.

1. The posterior for the parameters of a fit

Chapter 2.3 introduced three ways to attach an uncertainty to a fitted parton distribution function (PDF). This chapter develops the third. A Bayesian fit assigns the parameters θ of the parametrisation a probability density and computes that density given the data. Bands on $f(x, Q^2)$ and on cross-sections follow by pushing it through the model.

The product rule for a joint density can be written in two orders, $P(\theta, D) = P(D|\theta)\,P(\theta) = P(\theta|D)\,P(D)$. Dividing by $P(D)$, and writing the posterior $P(\theta|D)$ as $\rho(\theta|D)$, gives Bayes' theorem:

$$\rho(\theta|D) = \frac{\mathcal{L}(\theta)\,\pi(\theta)}{Z}, \qquad Z = \int d^p\theta\; \mathcal{L}(\theta)\,\pi(\theta). \tag{2.5.1}$$

The symbols are those of chapter 2.3: p is the number of parameters and $\rho(\theta|D)$ the posterior density. $P(\cdot)$ stands for any other probability density. The *prior* $\pi(\theta)$ is the density assigned before the data are used. The *likelihood* $\mathcal{L}(\theta) = P(D|\theta)$ is the probability of the observed data D if the parameters were θ. The *posterior* $\rho(\theta|D)$ is the result. Its normalisation Z is the *evidence*: the probability the model as a whole assigned to the data.

For data with Gaussian errors of covariance C and theory predictions $T(\theta)$, the likelihood is the χ^2 of chapter 2.2 in an exponent:

$$\begin{aligned} \mathcal{L}(\theta) &= \mathcal{L}_{\text{norm}}\, e^{-\chi^2(\theta)/2}, \qquad \mathcal{L}_{\text{norm}} = \frac{1}{\sqrt{(2\pi)^{N_{\text{dat}}}\,\det C}}, \\ \chi^2(\theta) &= \big(D - T(\theta)\big)^{\top} C^{-1} \big(D - T(\theta)\big). \end{aligned} \tag{2.5.2}$$

The constant $\mathcal{L}_{\text{norm}}$ does not depend on θ. It cancels in the posterior, so parameter inference needs only differences of χ^2. It does not cancel in Z, so evidences may be compared only between models fitted to the same data with the same covariance.

No approximation has been made. The Hessian method replaces χ^2 by its quadratic expansion, and the replica method equals the posterior only when T is linear in θ (chapter 2.3). The cost of exactness is computational. The posterior can only be evaluated point by point, and in $p = 52$ dimensions a grid is impossible. What can be had is a set of points $\theta^{(s)}$ with weights w_s distributed according to it. Any quantity $g(\theta)$, such as $x f(x, Q_0^2)$ at one x, then has

$$\begin{aligned} \mathbb{E}[g \mid D] &= \int d^p\theta\; g(\theta)\,\rho(\theta|D) \;\approx\; \sum_s w_s\, g\big(\theta^{(s)}\big), \\ P(D'|D) &= \int d^p\theta\; P(D'|\theta)\,\rho(\theta|D). \end{aligned} \tag{2.5.3}$$

The first relation needs no linearisation of g; percentiles of $g(\theta^{(s)})$ give the band. The second is the *posterior predictive* density for a future measurement D', in which its experimental error is folded with the parameter uncertainty.

Sections 3 to 7 are about producing the samples and Z; sections 9 and 10 report what happened when this project did so.

2. Priors and bounds

The prior used in all our fits, one of the two kinds Colibri offers, is flat inside a box: $\pi(\theta) = 1/V_\pi$ for θ_i between a lower and an upper bound a distance W_i apart, zero outside, with $V_\pi = \prod_i W_i$. What this does depends on whether the data constrain the direction in question.

Take one direction along which the likelihood is Gaussian with width σ, centred in a prior of width W. The posterior is that Gaussian cut off at $\pm W/2$. With $a = W/2\sigma$, and ϕ, Φ the density and cumulative distribution of the unit Gaussian, its variance is

$$\mathrm{Var}(\theta) = \sigma^2 \left[1 - \frac{2a\,\phi(a)}{2\Phi(a) - 1} \right] \longrightarrow \begin{cases} \sigma^2, & W \gg \sigma, \\ W^2/12, & W \ll \sigma. \end{cases} \tag{2.5.4}$$

The two limits are two regimes.

- **A constrained direction ignores the prior.** For $W = 10\sigma$ the bracket differs from 1 by 1.5×10^{-5}. The correction falls like $e^{-a^2/2}$.
- **A flat direction returns the prior.** When the data do not distinguish values of θ across the box, the posterior is uniform and its variance is $W^2/12$, that of the prior. The quoted uncertainty is then a statement about the bounds, and doubles if they are moved twice as far apart.

A fit can contain both kinds of direction. In our 52-parameter fit to 648 deep inelastic points the data constrain about 14 combinations of parameters (section 8).

Two further properties of a flat prior matter in practice.

- **Flat is a statement about coordinates.** Densities transform with a Jacobian, $P(u) = P(\theta)\,|d\theta/du|$. A prior flat in a normalisation A on $[0.01, 1]$ puts $0.9/0.99 = 91\%$ of its probability in the top decade; a prior flat in $\ln A$ puts 50% there. The choice must be declared before the fit.
- **Flat in the parameters is not flat in the function.** For $xf = Ax^\delta(1-x)^\eta$, a box of width $\Delta\delta$ becomes a spread $\Delta\delta\,\ln(1/x)$ in $\ln xf$: nothing at $x \to 1$, a factor $10^{3\Delta\delta}$ at $x = 10^{-3}$. The prior on the function is widest where data are absent.

Bounds and the sampler

Gradient-based samplers work on an unbounded space. The standard device, used by the numpyro code in this project, maps each bounded parameter to an unbounded one:

$$\theta = \theta_{\mathrm{lo}} + W\,u, \qquad u = \frac{1}{1+e^{-z}}, \qquad \frac{d\theta}{dz} = W\,u(1-u), \qquad \pi(z) = u(1-u). \tag{2.5.5}$$

The flat prior on θ becomes the logistic density in z, of variance $\pi^2/3$. Along a flat direction the sampler therefore sees a potential of unit scale in z, whatever W is. The corresponding precision, $3/\pi^2 = 0.304$, reappears in section 5.

3. Metropolis sampling and its cost

A Markov chain is a rule $T(\theta \to \theta')$ for moving from the current point to the next. If the rule satisfies *detailed balance* with respect to the posterior,

$$\rho(\theta|D)\,T(\theta\to\theta') = \rho(\theta'|D)\,T(\theta'\to\theta), \tag{2.5.6}$$

then integrating over θ shows that a population distributed as ρ is still distributed as ρ after one move. The chain's points, after an initial transient, are draws from the posterior.

The Metropolis algorithm [2] builds such a rule from any symmetric proposal. Propose $\theta' = \theta + \ell\,\xi$ with ξ a vector of independent unit Gaussians and ℓ a step length, and accept with probability

$$\alpha = \min\!\left(1, \frac{\rho(\theta'|D)}{\rho(\theta|D)}\right) = \min\!\left(1,\; e^{-[\chi^2(\theta')-\chi^2(\theta)]/2}\,\frac{\pi(\theta')}{\pi(\theta)}\right); \tag{2.5.7}$$

otherwise stay at θ. Only a ratio enters, so Z is never needed. Detailed balance holds because for $\rho(\theta') < \rho(\theta)$ the flux $\theta\to\theta'$ is $\rho(\theta)\cdot\rho(\theta')/\rho(\theta)$ and the reverse flux is $\rho(\theta')\cdot 1$.

The difficulty is the step length. For a Gaussian posterior with standard deviations σ_i along its principal axes, the mean acceptance in the limit of many dimensions is (derivation D1)

$$\bar\alpha = 2\,\Phi\!\left(-\frac{s}{2}\right), \qquad s^2 = \ell^2 \sum_{i=1}^{p} \frac{1}{\sigma_i^2}. \tag{2.5.8}$$

Every direction contributes to s, and the narrowest contribute most. To keep the acceptance finite the step must shrink like $\ell \sim \sigma_{\min}/\sqrt{p}$. The optimum is $s = 2.38$, where $\bar\alpha = 0.234$ [3]. With such a step the chain diffuses. Along the widest direction it needs a number of steps

$$\tau \;\simeq\; \frac{4\,\sigma_{\max}^2}{\bar\alpha\,\ell^2} \;=\; 3.0 \sum_{i=1}^{p} \frac{\sigma_{\max}^2}{\sigma_i^2} \;\le\; 3\,p\,\kappa, \qquad \kappa = \frac{\sigma_{\max}^2}{\sigma_{\min}^2}, \tag{2.5.9}$$

to produce one independent draw. κ is the condition number of the posterior covariance. In the check script a 20-dimensional Gaussian with $\kappa = 100$ gives $\tau = 1500$ likelihood evaluations against 1380 from Eq. (2.5.9). For our fits the ledgers record condition numbers of the curvature matrix between 10^{11} and 10^{13}, according to the matrix measured: 1.6–1.8×10^{11} at the minimum of the fit to real deep inelastic data (PPDF-36), and 2.8×10^{13} in the Phase A measurement at the planted truth of a closure test (PPDF-19). Chapter 2.3 sets the measurements side by side. The samplers below exist to avoid the random walk.

---

4. Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) [4, 5] replaces the random step by a long deterministic trajectory. It needs the gradient of χ^2, which automatic differentiation supplies at the cost of a few likelihood evaluations.

The construction

Define the *potential* $U(\theta) = \tfrac12\chi^2(\theta) - \ln\pi(\theta)$, so that the posterior is e^{-U}/Z'. Introduce one auxiliary momentum r_i per parameter and the Hamiltonian

$$\mathcal{H}(\theta, r) = U(\theta) + \tfrac12\, r^{\top} M^{-1} r. \tag{2.5.10}$$

M is a positive-definite *mass matrix*, a free choice. The joint density $e^{-\mathcal{H}}$ factorises into the posterior for θ times a Gaussian of covariance M for r. Sampling the pair and discarding r therefore samples the posterior.

Hamilton's equations are

$$\frac{d\theta}{dt} = \frac{\partial \mathcal{H}}{\partial r} = M^{-1} r, \qquad \frac{dr}{dt} = -\frac{\partial \mathcal{H}}{\partial \theta} = -\nabla U(\theta). \tag{2.5.11}$$

The flow conserves $\mathcal{H}$; it preserves volume in (θ, r), since the divergence $\partial_\theta \cdot (M^{-1} r) + \partial_r \cdot (-\nabla U)$ vanishes; and it is reversed by flipping the sign of r. Together these make it a move that leaves $e^{-\mathcal{H}}$ invariant. One HMC step is: draw $r \sim \mathcal{N}(0, M)$, follow the flow for some time, keep the new θ. The fresh r changes the energy, and the flow carries the point far across the level set of that energy.

Leapfrog and the acceptance step

The flow must be integrated numerically. The *leapfrog* scheme with step size ε is

$$\begin{aligned} r_{1/2} &= r_0 - \tfrac{\varepsilon}{2} \nabla U(\theta_0), \\ \theta_1 &= \theta_0 + \varepsilon\, M^{-1} r_{1/2}, \\ r_1 &= r_{1/2} - \tfrac{\varepsilon}{2} \nabla U(\theta_1), \end{aligned} \tag{2.5.12}$$

repeated n_{lf} times. Each line changes one of θ, r by an amount depending only on the other, so leapfrog preserves volume and is reversible, both exactly. It does not conserve $\mathcal{H}$. The error $\Delta\mathcal{H} = \mathcal{H}(\theta_{\mathrm{end}}, r_{\mathrm{end}}) - \mathcal{H}(\theta_0, r_0)$ is corrected by a Metropolis test: accept the end point with probability

$$\alpha = \min\left(1,\ e^{-\Delta\mathcal{H}}\right). \tag{2.5.13}$$

With this test the algorithm samples the posterior exactly for any ε and n_{lf} (derivation D2). Integration error costs efficiency, not correctness.

Step size and trajectory length

For a Gaussian posterior and $M = 1$ each principal direction is a harmonic oscillator of frequency $\omega_i = 1/\sigma_i$. Leapfrog is stable only below a critical step, and to cross the widest direction the trajectory must last a quarter of the longest period (derivation D3):

$$\varepsilon < \frac{2}{\omega_{\max}} = 2\,\sigma_{\min}, \qquad n_{\mathrm{lf}} \simeq \frac{\pi}{2}\,\frac{\sigma_{\max}}{\varepsilon} \ > \ \frac{\pi}{4}\sqrt{\kappa}. \tag{2.5.14}$$

The cost per independent draw grows as $\sqrt{\kappa}$ gradients, against $p\,\kappa$ likelihoods for Metropolis. Dimension enters only through the energy error. For a p-dimensional unit Gaussian and a quarter-period trajectory,

$$\mathrm{std}(\Delta\mathcal{H}) = \frac{\varepsilon^2 \sqrt{p}}{4}, \tag{2.5.15}$$

so holding the acceptance fixed requires $\varepsilon \propto p^{-1/4}$, a far weaker dependence than the $p^{-1/2}$ of Metropolis.

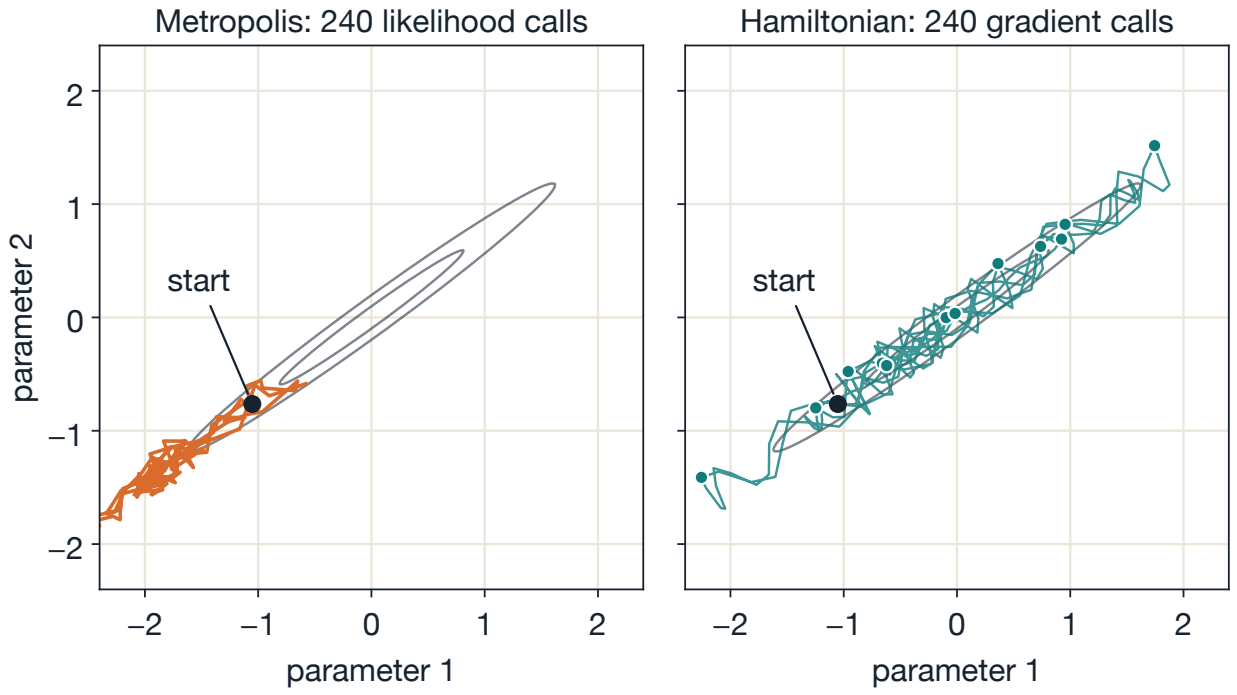

Figure 1. The same budget spent two ways on a two-dimensional Gaussian with $\sigma_{\max}/\sigma_{\min} = 12.5$ (ellipses: one and two sigma). Left: 240 Metropolis steps, limited by the narrow direction; the chain has covered one end of the distribution. Right: 12 Hamiltonian trajectories of 20 leapfrog steps; the dots are the 12 draws. Each trajectory oscillates across the narrow direction while travelling along the wide one. Computed by the check script; illustrative.

Where the curvature of U is locally much larger than the step size was tuned for, leapfrog goes unstable, the energy error grows without bound and the step is rejected. Samplers flag this as a *divergence*. More than a few per cent of divergences means regions of high curvature are under-sampled, which biases the result.

---

5. NUTS and the mass matrix

Choosing the trajectory length automatically

Equation (2.5.14) asks for a trajectory about a quarter period long. Shorter, and the sampler reverts to a random walk; longer, and the trajectory turns back and wastes gradients. The No-U-Turn Sampler (NUTS) [6] removes the choice. It extends the trajectory forwards and backwards in time, doubling its length at each stage, and stops when the two ends (θ^-, r^-) and (θ^+, r^+) start to approach each other:

$$(\theta^+ - \theta^-) \cdot M^{-1} r^- < 0 \qquad \text{or} \qquad (\theta^+ - \theta^-) \cdot M^{-1} r^+ < 0. \tag{2.5.16}$$

The next point is drawn from the states along the trajectory in a way that preserves detailed balance. A cap on the number of doublings j, the *maximum tree depth*, bounds the cost. In our runs it was 8, at most $2^8 - 1 = 255$ leapfrog steps per draw. The step size is tuned during a discarded *warm-up* phase to reach a target mean acceptance, 0.9 in our final runs.

Preconditioning

The factor $\sqrt{\kappa}$ in Eq. (2.5.14) depends on M. Write the Fisher matrix of the fit as $F_{ij} = \frac{1}{2}\,\partial_i\partial_j\chi^2$ at the minimum $\hat{\theta}$, so that the Gaussian approximation to the posterior has covariance $\Sigma = F^{-1}$. In the notation of chapter 2.3 this F is $\frac{1}{2}H$, half the Hessian of χ^2, and Σ is the parameter covariance $V = 2H^{-1}$. It coincides with the Fisher matrix $J^\top C^{-1} J$ of that chapter exactly for a model linear in θ, and otherwise up to the residual term of the Hessian. For $U = \frac{1}{2}\delta^\top F \delta$ with $\delta = \theta - \hat{\theta}$, Eq. (2.5.11) gives $\ddot{\delta} = -M^{-1}F\,\delta$: the squared frequencies are the eigenvalues of $M^{-1}F$. So

$$M = F = \Sigma^{-1} \quad\Longrightarrow\quad \ddot{\delta} = -\delta : \qquad \text{every direction has } \omega = 1, \quad \kappa_{\text{eff}} = 1. \tag{2.5.17}$$

In the 20-dimensional test of section 3 the cost per independent draw falls from 1500 likelihoods (Metropolis) to 21 gradients (HMC, $M = 1$) to 4.3 gradients (HMC, $M = \Sigma^{-1}$).

Usual practice estimates Σ from the warm-up draws, by default only its diagonal. That fails when the warm-up cannot explore. Phase B of this project used the Fisher matrix itself, computed by finite differences of the gradient before any sampling. In the coordinates of Eq. (2.5.5) the ledger records the transformation as

$$\begin{aligned} &F_z = D^\top F_\theta\, D, \qquad D_{ii} = W_i\, u_i(1-u_i), \\ &M = F_z \ \text{ with its eigenvalues floored at } 3/\pi^2. \end{aligned} \tag{2.5.18}$$

The floor is the prior precision in z: a direction the data do not constrain is given the scale of the prior.

What preconditioning cannot do

A constant M removes a linear distortion of the posterior and nothing else. Four situations defeat it.

- **Curvature that changes with position.** In a curved valley, F at one point is the wrong metric elsewhere. The step size must suit the worst region.
- **A flat direction bounded by the prior.** There the posterior is a box, not a Gaussian.
- **A Fisher matrix evaluated away from the minimum.** At a minimum F is positive semi-definite. Elsewhere it can have negative eigenvalues, which say nothing about widths. The ledger records 7 to 9 of them at the starting point of each of four Phase B runs: the optimiser had not converged, and those directions received the floor.
- **Several separated modes.** No metric connects them.

> **What sets the cost of a Markov chain.** Random-walk Metropolis pays of order $p\,\kappa$ likelihoods per independent draw. HMC pays $\sqrt{\kappa}$ gradients. HMC with $M = \Sigma^{-1}$ pays a number of order one, to the extent that the posterior is Gaussian. The difficulty of a real fit is in that last clause.

---

6. Diagnostics: effective sample size and R-hat

A chain always returns numbers. Two statistics, defined in chapter 2.3 and derived here, decide whether they are usable.

Effective sample size

Successive draws are correlated. Let a chain of n draws of some scalar have variance σ^2 and autocorrelation ρ_t at lag t. The variance of its mean is (derivation D4)

$$\mathrm{Var}(\bar{\theta}) = \frac{\sigma^2}{n}\left[1 + 2\sum_{t=1}^{n-1}\left(1 - \frac{t}{n}\right)\rho_t\right]. \tag{2.5.19}$$

For n much longer than the correlation length this is the variance of the mean of n_{eff} independent draws, with

$$n_{\mathrm{eff}} = \frac{n}{\tau}, \qquad \tau = 1 + 2\sum_{t=1}^{\infty}\rho_t. \tag{2.5.20}$$

n_{eff} is the *effective sample size* (ESS) and τ the integrated autocorrelation time. A chain in which each draw remembers a fraction φ of the previous one, $\theta_{t+1} = \varphi\,\theta_t + \text{noise}$, has $\rho_t = \varphi^t$ and

$$\tau = \frac{1+\varphi}{1-\varphi}. \tag{2.5.21}$$

At $\varphi = 0.9$, 19 draws are worth one. Three points govern the use of ESS.

- **It is defined per quantity.** Each parameter has its own τ, and a fit is as converged as its worst parameter.
- **It adds over chains.** Independent chains on the same posterior give $n_{\mathrm{eff}} = \sum_c n_{\mathrm{eff},c}$ for each parameter. The sum is taken per parameter first and the minimum afterwards. Summing each chain's own minimum mixes different parameters and gives only a lower bound.
- **The estimate presupposes stationarity.** ρ_t is estimated from the chain itself. A chain that has explored only a corner of the posterior reports the autocorrelation of its motion within that corner.

An ESS of 100 fixes the posterior mean to 0.1σ, which is why 100 is a common threshold.

R-hat

The third point above is what a second statistic addresses [7]. Run m chains of n draws from dispersed starting points. Let W be the average of the within-chain variances and B/n the variance of the m chain means. Then

$$\widehat{\mathrm{var}}^{+} = \frac{n-1}{n}\,W + \frac{B}{n}, \qquad \hat{R} = \sqrt{\frac{\widehat{\mathrm{var}}^{+}}{W}}. \tag{2.5.22}$$

If all chains sample the same distribution, $B/n \simeq \sigma^2\tau/n$ is negligible and $\hat{R} \to 1$. If the chains sit in different places, B/n stays finite as n grows, and $\hat{R}^2 - 1$ is the between-chain variance in units of the within-chain variance. *Split* $\hat{R}$ applies the formula to the two halves of one chain. It detects drift, and cannot detect a chain that is stationary in the wrong place.

$\hat{R}$ has no meaning unless the chains target one distribution. Suppose each of m chains is given its own noise realisation of the data and samples its own posterior perfectly. In a linear Gaussian problem the best fit moves between realisations with covariance Σ, the posterior covariance itself (chapter 2.3). So $B/n \to W$ and

$$\mathbb{E}\big[\hat{R}^2\big] \;\longrightarrow\; 2, \qquad \hat{R} \approx \sqrt{2} = 1.41, \tag{2.5.23}$$

for every data-constrained parameter, with no sampling fault at all. Section 10 returns to this.

7. Nested sampling

Markov chains give posterior draws and no evidence. Nested sampling [8] computes Z and yields the posterior as a by-product, using likelihood values only.

The evidence as a one-dimensional integral

For each likelihood level λ define the fraction of prior probability lying above it,

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} d^p\theta \; \pi(\theta). \tag{2.5.24}$$

X falls from 1 at $\lambda = 0$ to 0 at $\lambda = \mathcal{L}_{\max}$. Inverting it gives a function $\mathcal{L}(X)$: the likelihood on the contour that encloses prior volume X. In terms of it (derivation D5),

$$Z = \int_0^1 dX \; \mathcal{L}(X). \tag{2.5.25}$$

The p-dimensional integral has become the area under a decreasing curve. The map from θ to X cannot be computed, but it can be sampled.

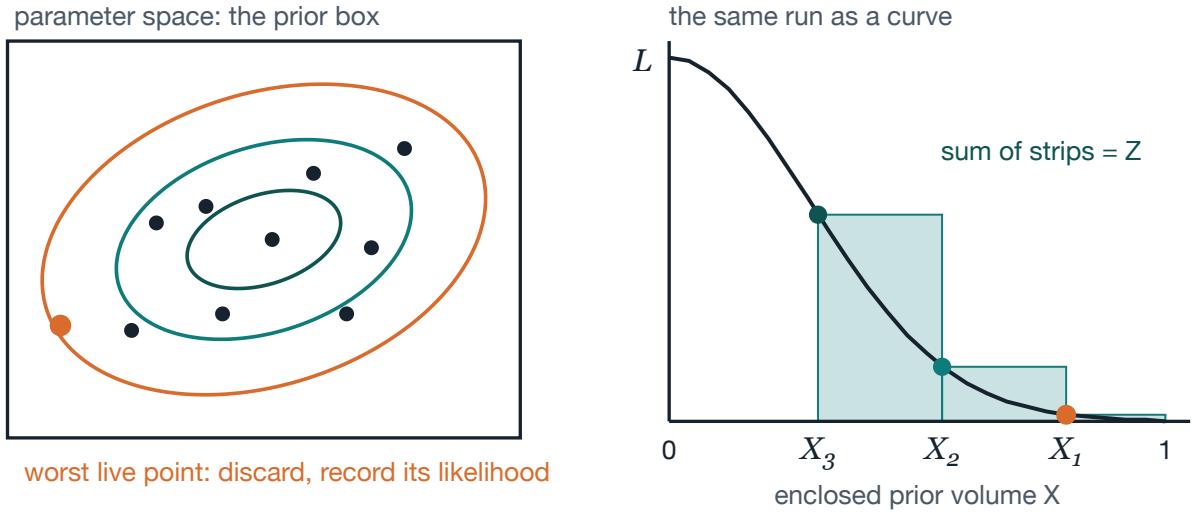

Figure 2. Nested sampling. Left: the live points are spread uniformly over the part of the prior inside the current likelihood contour; the point of lowest likelihood defines the next contour. Right: each discarded point contributes a strip of height $\mathcal{L}_k$ and width $X_{k-1} - X_k$. The heights are measured. The widths are known only statistically.

Shrinkage

Keep N_{live} *live points* distributed according to the prior, restricted to $\mathcal{L} > \lambda$. Remove the one with the lowest likelihood, $\mathcal{L}_k$, and raise the threshold to it. The volume enclosed by the new contour is a fraction t of the old one, and t is the largest of N_{live} independent uniform numbers (derivation D5). So

$$\mathbb{E}[\ln t] = -\frac{1}{N_{\text{live}}}, \qquad \mathrm{Var}[\ln t] = \frac{1}{N_{\text{live}}^2}, \qquad \ln X_k = -\frac{k}{N_{\text{live}}} \pm \frac{\sqrt{k}}{N_{\text{live}}}. \tag{2.5.26}$$

Then replace the removed point by a new one drawn from the prior inside the new contour, and repeat. The volume contracts geometrically at a known average rate, whatever the dimension and the shape of the likelihood.

The algorithm and its by-products

With $X_k = e^{-k/N_{\rm live}}$ assigned to the k-th discarded point,

$$Z \approx \sum_k \mathcal{L}_k \, (X_{k-1} - X_k), \qquad w_k = \frac{\mathcal{L}_k \, (X_{k-1} - X_k)}{Z}. \tag{2.5.27}$$

The discarded points with weights w_k sample the posterior, as Eq. (2.5.3) requires. The run stops when the most the remaining volume could contribute, $\mathcal{L}^{\rm live}_{\rm max} X_k$, is below a set fraction of the accumulated Z.

The length of a run is set by the *information*, written H in the nested-sampling literature and $\mathcal{I}$ here to keep H for the Hessian:

$$\mathcal{I} = \int d^p\theta \; \rho(\theta|D) \; \ln \frac{\rho(\theta|D)}{\pi(\theta)} = \big\langle \ln \mathcal{L} \big\rangle_{\rm post} - \ln Z. \tag{2.5.28}$$

$e^{-\mathcal{I}}$ is roughly the fraction of the prior volume that the posterior occupies. The posterior mass sits near $\ln X \approx -\mathcal{I}$. Reaching it takes about $N_{\rm live}\mathcal{I}$ iterations, by which time the random walk of $\ln X_k$ in Eq. (2.5.26) has accumulated the uncertainty

$$\sigma(\ln Z) \approx \sqrt{\frac{\mathcal{I}}{N_{\rm live}}}, \qquad n_{\rm iter} \gtrsim N_{\rm live}\, \mathcal{I}. \tag{2.5.29}$$

Halving the error on $\ln Z$ costs four times the live points and four times the run.

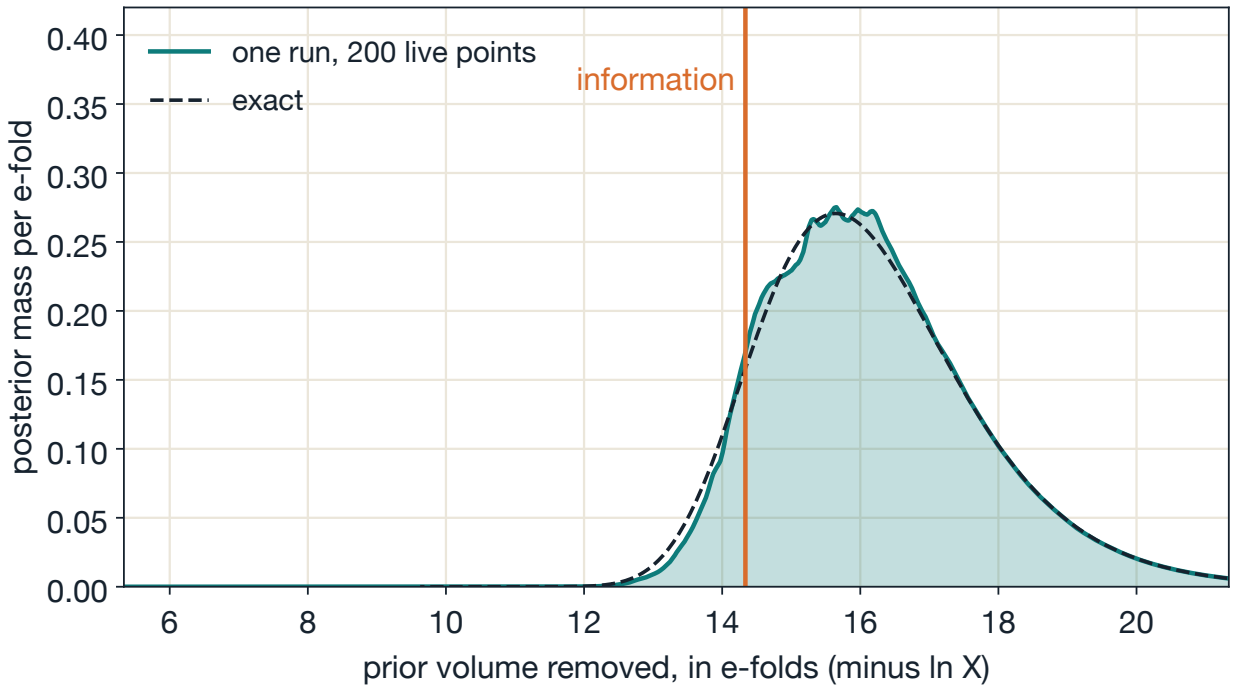

Figure 3. A nested-sampling run on a four-parameter Gaussian toy with a flat prior, from the check script. For the first 10 e-folds the run crosses prior volume where the likelihood is negligible. The posterior mass arrives in a bump just beyond $-\ln X = \mathcal{I} = 14.3$. The run of 200 live points took 5,109 iterations and returned $\ln Z = -16.33$ against the exact -16.34. Over 60 runs with 100 live points the scatter of $\ln Z$ was 0.40, against $\sqrt{\mathcal{I}/N_{\text{live}}} = 0.38$.

Drawing inside the contour

The hard step is the replacement: a draw from the prior restricted to $\mathcal{L} > \mathcal{L}_k$, a region of unknown shape whose volume is $e^{-k/N_{\text{live}}}$ of the box. Drawing from the whole prior and rejecting has acceptance X_k. Drawing from a region built around the live points fails in the same way, more slowly: if the region is too generous by a factor $1+\delta$ per axis the acceptance is $(1+\delta)^{-p}$, which for $\delta = 0.3$ and $p = 52$ is 10^{-6}.

UltraNest [9], the nested sampler that Colibri calls, by default draws from such a region and rejects. Alternatively it is given a *step sampler*: start from a live point and make `nsteps` one-dimensional slice moves inside the contour. The number must be large enough that the new point is independent of its starting point, since the shrinkage law assumes independent draws. The project's rule (section 10) is to configure the step sampler whenever $p \gtrsim 10$. UltraNest's *reactive* mode also adds live points during the run when its error estimates ask for them.

Step sampling inside a contour is a random walk, and inherits the problem of section 3. When the contour is hundreds of thousands of times longer in one direction than another, a fixed number of slice moves does not decorrelate the new point, the contours stop shrinking at the assumed rate, and the run stalls or returns a biased Z without warning.

8. Evidence, Occam factor and complexity

Bayes factors

Apply Bayes' theorem to two models $\mathcal{M}_1, \mathcal{M}_2$ fitted to the same data:

$$\frac{P(\mathcal{M}_1|D)}{P(\mathcal{M}_2|D)} = \frac{Z_1}{Z_2}\,\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}, \qquad \ln B_{12} = \ln Z_1 - \ln Z_2. \tag{2.5.30}$$

The *Bayes factor* B_{12} converts prior odds into posterior odds. Z is the likelihood averaged over the prior: a model is rewarded for the probability it gave the data with all the prior volume it committed, not for the best point it contains.

The Occam factor

Expand χ^2 to second order about its minimum $\hat{\theta}$, with F as in section 5, and take a flat prior of volume V_π that contains the peak. The Gaussian integral gives (derivation D6)

$$\ln Z \simeq \ln \mathcal{L}_{\rm norm} - \frac{\chi^2_{\rm min}}{2} + \ln \frac{(2\pi)^{p/2}\,(\det F)^{-1/2}}{V_\pi}. \tag{2.5.31}$$

The first two terms are the best-fit likelihood. The third is the logarithm of the *Occam factor*: the volume the posterior occupies divided by the volume the prior offered. It is negative. A direction whose width the data reduce from W to σ costs $\ln(W/\sqrt{2\pi}\sigma)$, and an extra parameter is worth having only if it lowers $\chi^2_{\rm min}/2$ by more than that.

The expansion overstates the cost of a direction the data do not constrain. In one dimension the exact result for a Gaussian likelihood centred in the box is

$$Z = \mathcal{L}_{\rm norm}\, e^{-\chi^2_{\rm min}/2}\,\frac{\sqrt{2\pi}\,\sigma}{W}\,\big[2\Phi(a) - 1\big], \qquad a = \frac{W}{2\sigma}. \tag{2.5.32}$$

For $W \gg \sigma$ the bracket is 1 and the Occam factor is $\sqrt{2\pi}\sigma/W$. For $W \ll \sigma$ the product tends to 1: a parameter the data cannot see costs nothing. The evidence penalises prior volume that the data ruled out, not the number of parameters.

A prior width enters $\ln Z$ as $-\ln W$ per constrained direction while leaving the posterior unchanged (section 2). A Bayes factor is therefore only as meaningful as the priors. A verdict is robust when the gap in $\chi^2_{\rm min}/2$ is much larger than p times the logarithm of any defensible change in widths.

Bayesian complexity

The number of directions the data constrain can be read from the samples. Define the *Bayesian complexity* [10] as the posterior mean of χ^2 above its minimum. For a Gaussian posterior (derivation D7)

$$C_b = \big\langle \chi^2 \big\rangle_{\rm post} - \chi^2_{\rm min} = p, \tag{2.5.33}$$

and for a Gaussian likelihood in a box aligned with its principal axes, direction by direction,

$$C_b = \sum_{i=1}^{p} \left[1 - \frac{2a_i\,\phi(a_i)}{2\Phi(a_i) - 1}\right], \qquad a_i = \frac{W_i}{2\sigma_i}. \tag{2.5.34}$$

Each summand is the bracket of Eq. (2.5.4): 1 for a direction the data constrain, $W_i^2/12\sigma_i^2 \to 0$ for one they do not. C_b counts the effective number of constrained parameters, $p_{\rm eff}$.

For the 52-parameter MSHT20 form on 648 real deep inelastic points the ledger gives $\chi^2_{\rm min} = 649.5$, complexity 13.73 and $\ln Z = -381.9 \pm 0.29$. Against the 13-parameter Les Houches form on the same data ($\chi^2_{\rm min} = 3626.5$, $\ln Z = -1835.1 \pm 0.16$) the difference is $\Delta \ln Z = +1453$. Half the gap in $\chi^2_{\rm min}$ is 1488, so the Occam terms account for -35: the verdict is set by fit quality, and no choice of prior widths reverses it. The comparison of a 36-parameter

grid model with the Les Houches form, $\Delta \ln Z = -108.8$, is of the other kind. The ledger notes that the grid's prior was a narrow band around an existing global fit and that "evidence comparisons are prior-sensitive by construction".

9. Colibri and closure tests

The code

Colibri [1] is a PDF-fitting framework from Maria Ubiali's group in Cambridge (the HEP-PBSP project). It releases no PDF set. It is written in Python on JAX, which provides automatic differentiation and compilation for CPU or GPU, and takes its data, covariance matrices and FK tables from the public NNPDF code. The user supplies a model: a list of parameter names and a JAX function returning the distributions on the x grid in the evolution basis (chapter 2.1). One model and one dataset can then be put through each uncertainty method with nothing else changed.

METHOD	WHAT COLIBRI DOES	EXACT WHEN
Analytic	closed-form Gaussian posterior	predictions linear in the parameters
Hessian	gradient descent, then the curvature at the minimum; optional tolerance	the posterior is Gaussian
Monte Carlo	one fit per pseudo-data replica, one replica per process	linear model, Gaussian errors
Bayesian	nested sampling with UltraNest, flat or Gaussian priors	the sampler has converged

The analytic route refuses hadronic data, whose predictions are quadratic in the distributions. The Bayesian route returns weighted posterior samples, $\ln Z$ with its error, the minimum and posterior-mean χ^2, and the complexity. Colibri has no Hamiltonian sampler: its second Bayesian action, through the blackjax library, is also a nested sampler (Phase B ledger). The NUTS runs in this project call numpyro on Colibri's likelihood through its standalone interface.

Closure tests

A *closure test* (chapter 2.3) replaces the data by pseudo-data generated from chosen true parameters θ^* and runs the unchanged pipeline. Colibri implements two levels.

- **Level 0**: $D = T(\theta^*)$, no noise. The minimum of χ^2 is exactly zero at θ^*.
- **Level 1**: $D = T(\theta^*) + \eta$, with one noise vector η drawn from the experimental covariance.

At level 1 the fit statistics have known expectations (chapter 2.3; derivation D7 adds the posterior mean). For a model that is linear over the region the posterior occupies,

$$\mathbb{E}\left[\chi^2_{\min}\right] = N_{\mathrm{dat}} - p_{\mathrm{eff}}, \qquad \mathbb{E}\left[\langle \chi^2 \rangle_{\mathrm{post}}\right] = N_{\mathrm{dat}}, \qquad \mathrm{Var}\left[\chi^2_{\min}\right] = 2\,(N_{\mathrm{dat}} - p_{\mathrm{eff}}). \tag{2.5.35}$$

The best fit absorbs one unit of noise per constrained direction, and the posterior average gives it back. A nested-sampling run at level 0 does not report $\chi^2_{\min} = 0$, because its best point is only the best of a finite number of posterior draws.

Our reproductions of Colibri's closure tests, from the Phase A ledger:

MODEL AND DATA	LEVEL	MINIMUM CHI-SQUARED PER POINT	POSTERIOR-MEAN CHI-SQUARED PER POINT	COMPLEXITY	LOG EVIDENCE
Les Houches, 13 parameters, 648 points	0	1.96×10^{-4}	4.248 in total	4.121	-7.257 ± 0.13
the same	1	0.971	0.977	3.88	-321.8 ± 0.15
MSHT20, 52 parameters, 648 points	0	0.00741	15.19 in total	10.39	-54.65 ± 0.32
the same	1	0.961	0.979	11.6	-363.7 ± 0.26
MSHT20, 52 parameters, 3,089 points	1	0.9934	0.9991	17.64	-1639.0

The level-1 rows agree with Eq. (2.5.35): the scatter of χ^2/N_{dat} for 648 points is $\sqrt{2/648} = 0.056$, and in the last row the expected minimum is $(3089 - 17.64)/3089 = 0.9943$.

The last row also carries a warning. Its χ^2 statistics are correct, and the ledger first recorded it as a pass. The check of the bands against the planted truth then found pointwise one-sigma coverage of 31 to 40%, and a Mahalanobis χ^2 of the truth of 4429 where about 52 was expected. The sample had "collapsed onto part of" a nearly degenerate set of equal χ^2, and the ledger's final verdict was that the 52-parameter form is degenerate on deep inelastic data alone, with no certifiable Bayesian result. The statistics of Eq. (2.5.35) test the likelihood and the minimum. Only coverage tests the sample.

10. The Bayesian arm of this project

The project's fits use the MSHT20 form with 52 parameters, NNPDF theory 40000000 (NNLO, $\alpha_S(M_Z) = 0.118$, $Q_0 = 1.65$ GeV) and Colibri. What follows is in the ledgers' own numbers [11].

Nested sampling stopped scaling

On 648 points nested sampling worked (section 9), once a step sampler was configured: a 24-parameter run without one fell to an efficiency of 0.003%, and with 100 slice steps ran at 0.87%. On the 19 deep inelastic datasets the level-0 run escalated its live points "300→600→1200→2400 over ~18h without converging". With Drell–Yan and W, Z data added, the ESS was "stuck at 1 for 5 days while live points escalated 600→1200→2400→4800". In Phase B a nested run made "1,084,065 evaluations at 0.68% efficiency with log Z never leaving 0.00%". The cause is the one at the end of section 7.

NUTS with a Fisher mass matrix worked, at small scale

In Phase A, NUTS with a dense mass matrix built from the Fisher geometry reached "ESS~1000, 0 divergences" in about 80 minutes on the problem where nested sampling had sat at ESS 1 for five days. A first attempt with a diagonal mass matrix "was slow for the same reason" nested sampling was. On real deep inelastic data the whitened sampler reached a median ESS of 181 (minimum 55), with 80 divergences in 3,000 draws.

A converged sampler is not a calibrated result. Over five noise draws of a level-1 closure test the ledger records "converged mean coverage ~58%" against a nominal 68%, and a mean of 38% for the combination T_8, which carries the strange distribution.

At 3,089 points it did not converge

Phase B pre-registered a gate before any run: "ESS ≥ 100 median, R̂ < 1.05, divergences < 5%. Coverage is not reported below this bar." 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.

RUN	DRAWS KEPT	MINIMUM ESS	R-HAT	DIVERGENCES
one chain	1,200	3.2	2.4545 (split)	0.25%
one chain, longer	37,500	3.0	3.0830 (split)	2 of 37,500
three chains, one dataset	3 × 6,000	30.8 (summed)	1.1185 (three-chain)	4.30%, 0.45%, 0.40%

- **More draws did not help.** Thirty-one times the draws moved the minimum ESS from 3.2 to 3.0. The chain had adapted to a step size of 8.4×10^{-4} and was not exploring. The ledger: "Low divergences plus low ESS is the signature of a stuck chain, not a healthy one". ESS grows with n only for a chain that mixes.
- **More chains did.** Per-chain minima of 11.2, 12.0 and 6.3 combine, parameter by parameter, to 30.8. The ledger estimates that the gate needs about 10 chains at about 4.2 hours each. Compiling the sampler for this likelihood took 200 to 400 GB of memory, so chains ran one at a time.
- **The gate was not met.** On the three-chain run the medians over the 52 parameters were 222 for the combined ESS and 1.0296 for $\hat{R}$, against 30.8 and 1.1185 for the worst parameter. The gate fails on $\hat{R}$ under either reading of the ESS criterion, and on ESS as well under the ledger's. Under the pre-registered rule no posterior band is quoted. The parameters with the lowest ESS were in the `rho` sector (the ratio $\bar{d}/\bar{u}$ of chapter 2.1), those with the largest $\hat{R}$ in up-valence and `rho`.

A retracted result

Before the three-chain run, the ledger had recorded a four-chain run with combined ESS 111.2 and $\hat{R} = 1.894$, with a physical reading: the chains had settled in different regions because the sea and strange directions are flat. Both numbers were withdrawn the next day.

The script's seed argument selected the level-1 noise realisation, not the sampler's random stream, which was fixed. The four runs were one chain on each of four pseudo-datasets. The proof uses a quantity that depends only on the noise draw: $\chi^2(\theta^*)$ was 3047.977, 3038.504, 3086.695 and 3072.615 in the four runs. Pooled ESS across different posteriors is meaningless, and $\hat{R}$ across them measures how far the best fit moves between noise draws, as in Eq. (2.5.23). The ledger's check agrees: in units of each direction's width, the four centres differed by a median of 1.95 along the 45 data-constrained directions and 0.042 along the 7 prior-dominated ones.

The corrected run fixed the data seed, varied the chain seed, and asserted that all chains report the same $\chi^2(\theta^*)$ (3087.661). With the chains sharing a dataset, $\hat{R}$ fell from 1.894 to 1.1185. Neither retracted number had been published.

> **Replicates must be shown to be replicates.** $\hat{R}$ and pooled ESS assume that every chain targets the same posterior. In a level-1 closure test, χ^2 at the planted truth depends only on the noise draw, so it must be identical across chains. Check that first.

Derivations

D1. The acceptance rate and cost of random-walk Metropolis

Derives (2.5.8), (2.5.9)

The acceptance. Work in the principal axes of a Gaussian posterior, $-\ln\rho = \sum_i \theta_i^2/2\sigma_i^2$, with the chain in equilibrium, so $\theta_i = \sigma_i\zeta_i$ with ζ_i unit Gaussians. Propose $\theta_i' = \theta_i + \ell\xi_i$. The logarithm of the ratio in Eq. (2.5.7) is

$$\Delta = -\sum_i \frac{(\theta_i + \ell\xi_i)^2 - \theta_i^2}{2\sigma_i^2} = -\sum_i \frac{\ell}{\sigma_i}\,\zeta_i\xi_i - \frac{1}{2}\sum_i \frac{\ell^2}{\sigma_i^2}\,\xi_i^2 .$$

For large p the second sum tends to its mean, $-s^2/2$ with $s^2 = \ell^2\sum_i \sigma_i^{-2}$. The first is a sum of many independent terms of zero mean and total variance s^2. So $\Delta \sim \mathcal{N}(-s^2/2,\ s^2)$.

The average. For a Gaussian with mean $-s^2/2$ and variance s^2, $\mathbb{E}[e^\Delta;\Delta<0] = \mathbb{P}[\Delta>0] = \Phi(-s/2)$, because multiplying the density by e^Δ shifts its mean to $+s^2/2$. Hence $\bar\alpha = \mathbb{E}[\min(1,e^\Delta)] = 2\Phi(-s/2)$, Eq. (2.5.8).

The optimum. Each accepted step moves coordinate i by variance ℓ^2, so the mean squared displacement per step is $\bar\alpha\,\ell^2 \propto s^2\,\Phi(-s/2)$. This is largest at $s = 2.38$, where $\bar\alpha = 0.234$.

The cost. A coordinate that diffuses with variance $h = \bar\alpha\ell^2$ per step inside a Gaussian of width σ has autocorrelation $\rho_t = e^{-ht/2\sigma^2}$, as for the chain of Eq. (2.5.21) with $\varphi = 1 - h/2\sigma^2$. Its autocorrelation time is $\tau \simeq 4\sigma^2/h$. Inserting $\ell^2 = 2.38^2/\sum_i\sigma_i^{-2}$ and $\bar\alpha = 0.234$ gives the coefficient $4/(0.234\times 5.66) = 3.0$ of Eq. (2.5.9).

D2. Hamiltonian Monte Carlo leaves the posterior invariant

Derives (2.5.13)

Write $y = (\theta, r)$ and let Ψ be the map "apply $n_{\rm lf}$ leapfrog steps, then flip the sign of r".

Ψ preserves volume. Each line of Eq. (2.5.12) has the form $r \to r + a(\theta)$ or $\theta \to \theta + b(r)$. The Jacobian matrix of such a map is triangular with unit diagonal, so its determinant is 1. The flip has determinant ±1. A composition of such maps has $|\det \partial\Psi/\partial y| = 1$.

Ψ is its own inverse. Apply one leapfrog step to $(\theta_1, -r_1)$. The first line gives $-r_1 - \frac{\varepsilon}{2}\nabla U(\theta_1) = -r_{1/2}$. The second gives $\theta_1 - \varepsilon M^{-1} r_{1/2} = \theta_0$. The third gives $-r_{1/2} - \frac{\varepsilon}{2}\nabla U(\theta_0) = -r_0$. The step has been retraced, and a final flip returns (θ_0, r_0). So $\Psi(\Psi(y)) = y$.

Detailed balance. The proposal is deterministic: from y the chain can go only to $y' = \Psi(y)$. The probability flux from a small region dy around y to its image is $e^{-\mathcal{H}(y)}\,dy\ \min(1, e^{-\mathcal{H}(y')+\mathcal{H}(y)}) = \min\big(e^{-\mathcal{H}(y)}, e^{-\mathcal{H}(y')}\big)\,dy$. The reverse flux, from the image region dy' back to dy, is $\min\big(e^{-\mathcal{H}(y')}, e^{-\mathcal{H}(y)}\big)\,dy'$. The two are equal because $dy' = dy$. This is Eq. (2.5.6) for the density $e^{-\mathcal{H}}$, with the acceptance of Eq. (2.5.13).

The momentum. $\mathcal{H}$ is even in r, so the flip can be dropped. Redrawing r from $\mathcal{N}(0, M)$ is an exact draw from its conditional density and also leaves $e^{-\mathcal{H}}$ invariant. The whole step therefore has $e^{-\mathcal{H}}$ as its stationary density, whose marginal in θ is the posterior. Nothing in the argument used the size of ε.

D3. Leapfrog on a Gaussian: stability and energy error

Derives (2.5.14), (2.5.15)

One step. Take one direction with $U = \frac{1}{2}\omega^2\theta^2$ and $M = 1$. Substituting into Eq. (2.5.12) gives a linear map:

$$\begin{pmatrix}\theta_1\\ r_1\end{pmatrix} = \begin{pmatrix}1 - \frac{1}{2}\varepsilon^2\omega^2 & \varepsilon\\ -\varepsilon\omega^2\left(1 - \frac{1}{4}\varepsilon^2\omega^2\right) & 1 - \frac{1}{2}\varepsilon^2\omega^2\end{pmatrix}\begin{pmatrix}\theta_0\\ r_0\end{pmatrix}. \tag{2.5.36}$$

Stability. The determinant is $(1 - \frac{1}{2}\varepsilon^2\omega^2)^2 + \varepsilon^2\omega^2(1 - \frac{1}{4}\varepsilon^2\omega^2) = 1$, so the eigenvalues multiply to 1 and sum to the trace, $2 - \varepsilon^2\omega^2$. They lie on the unit circle, and the motion is bounded, if and only if the trace is between -2 and 2: $\varepsilon\omega < 2$. Beyond that one eigenvalue exceeds 1 in magnitude and the trajectory grows exponentially. The largest frequency decides: $\varepsilon < 2/\omega_{\max} = 2\sigma_{\min}$, the first part of Eq. (2.5.14). For $\varepsilon\omega \ll 1$ the map is a rotation by $\varepsilon\omega$ per step, and the slowest direction needs a quarter turn, $n_{\mathrm{lf}}\,\varepsilon\,\omega_{\min} = \pi/2$, the second part.

A conserved quantity. Direct substitution of Eq. (2.5.36) shows that the map conserves

$$\tilde{\mathcal{H}} = \tfrac{1}{2}r^2 + \tfrac{1}{2}\omega^2\theta^2\left(1 - \tfrac{1}{4}\varepsilon^2\omega^2\right) \tag{2.5.37}$$

exactly. Leapfrog solves a slightly different oscillator without error. Since $\mathcal{H} = \tilde{\mathcal{H}} + \frac{1}{8}\varepsilon^2\omega^4\theta^2$, the error in the true energy after any number of steps is

$$\Delta\mathcal{H} = \frac{\varepsilon^2\omega^4}{8}\left(\theta_{\mathrm{end}}^2 - \theta_0^2\right). \tag{2.5.38}$$

It is of order ε^2 and does not grow with the length of the trajectory.

Many dimensions. For p directions with $\omega = 1$, start in equilibrium: θ_0 and r_0 are independent unit Gaussians. After a quarter period $\theta_{\mathrm{end}} \simeq r_0$. Then $\Delta\mathcal{H} = \frac{1}{8}\varepsilon^2\sum_i(r_{0,i}^2 - \theta_{0,i}^2)$, a sum of p independent terms each of variance $2 + 2 = 4$. Its standard deviation is $\frac{1}{8}\varepsilon^2\sqrt{4p} = \frac{1}{4}\varepsilon^2\sqrt{p}$, Eq. (2.5.15).

D4. Effective sample size and R-hat

Derives (2.5.19), (2.5.20), (2.5.21), (2.5.23)

The variance of a correlated mean. For a stationary chain, $\mathrm{Cov}(\theta_s, \theta_t) = \sigma^2\rho_{|s-t|}$. Then

$$\mathrm{Var}(\bar{\theta}) = \frac{1}{n^2}\sum_{s,t=1}^{n}\sigma^2\rho_{|s-t|} = \frac{\sigma^2}{n^2}\Big[n + 2\sum_{t=1}^{n-1}(n-t)\,\rho_t\Big],$$

since the lag t occurs for $n - t$ ordered pairs, twice. This is Eq. (2.5.19). When ρ_t has decayed long before $t \sim n$, the factor $1 - t/n$ is 1 wherever ρ_t is not negligible, and the bracket becomes τ of Eq. (2.5.20).

The geometric case. For $\rho_t = \varphi^t$, $\tau = 1 + 2\varphi/(1-\varphi) = (1+\varphi)/(1-\varphi)$, Eq. (2.5.21).

R-hat for chains on different data. Let chain c sample, without error, a Gaussian posterior of variance σ^2 centred at $\hat\theta_c$. Its mean is $\hat\theta_c$ plus sampling noise of variance $\sigma^2\tau/n$, and $W \to \sigma^2$. If all chains share one dataset the $\hat\theta_c$ are equal, $\mathbb{E}[B/n] = \sigma^2\tau/n \to 0$ and $\hat R \to 1$. If each has its own noise realisation, the $\hat\theta_c$ are independent draws from the sampling distribution of the best fit, of variance σ^2. Then $\mathbb{E}[B/n] \to \sigma^2$, $\widehat{\mathrm{var}}^{+} \to 2\sigma^2$ and $\hat R^2 \to 2$ on average, Eq. (2.5.23). With $m = 4$ the variance of four centres fluctuates widely, so individual parameters scatter around this value. A parameter fixed by the prior has the same centre in every realisation and gives $\hat R = 1$.

D5. Nested sampling: the evidence integral, shrinkage and the error

Derives (2.5.25), (2.5.26), (2.5.29)

The one-dimensional integral. Write the likelihood at each point as an integral over levels, $\mathcal{L}(\theta) = \int_0^\infty d\lambda\ \Theta\big(\mathcal{L}(\theta) - \lambda\big)$, with Θ the step function. Insert this in Z and exchange the order of integration:

$$Z = \int_0^\infty d\lambda \int d^p\theta\ \pi(\theta)\,\Theta\big(\mathcal{L}(\theta) - \lambda\big) = \int_0^\infty d\lambda\ X(\lambda).$$

This is the area under the decreasing curve $X(\lambda)$, between $\lambda = 0$ where $X = 1$ and $\lambda = \mathcal{L}_{\max}$ where $X = 0$. The same area, sliced along the other axis, is $\int_0^1 dX\ \mathcal{L}(X)$, Eq. (2.5.25).

One shrinkage. Let the current contour enclose volume X^*, and let θ be drawn from the prior restricted to its interior. The volume enclosed by the contour through θ is $X(\mathcal{L}(\theta))$. For any $x < X^*$, the probability that this is less than x is the prior mass with $X < x$ divided by the prior mass inside the contour, x/X^*. So $u = X(\mathcal{L}(\theta))/X^*$ is uniform on $[0, 1]$, whatever the likelihood. The live point with the lowest likelihood has the largest u. The largest of N independent uniforms is below t with probability t^N, so it has density Nt^{N-1} and

$$\mathbb{E}[\ln t] = \int_0^1 dt\ Nt^{N-1} \ln t = -\frac{1}{N}, \qquad \mathbb{E}[(\ln t)^2] = \frac{2}{N^2}, \qquad \mathrm{Var}[\ln t] = \frac{1}{N^2}.$$

Many shrinkages. Successive factors are independent, so $\ln X_k = \sum_{j\le k} \ln t_j$ has mean $-k/N$ and variance k/N^2, Eq. (2.5.26).

The information. From Eq. (2.5.1), $\ln(\rho/\pi) = \ln\mathcal{L} - \ln Z$, and averaging over the posterior gives the second form of Eq. (2.5.28). For a Gaussian posterior of covariance Σ well inside a flat prior, $\mathcal{I} = \ln V_\pi - \frac{1}{2}\ln\det(2\pi e\,\Sigma)$: the logarithm of the ratio of prior volume to posterior volume.

The error. The posterior mass per unit $\ln X$ is $\mathcal{L}(X)\,X/Z$. It is concentrated in a bump around $\ln X \approx -\mathcal{I}$ (Figure 3; for a Gaussian the peak is at $-\mathcal{I} - \frac{1}{2}\ln(\pi p)$ at large p). The run reaches the bump after $k \approx N\mathcal{I}$ iterations. The likelihoods $\mathcal{L}_k$ are measured exactly, but the volume assigned to the whole bump is uncertain by the accumulated scatter of $\ln X_k$, which is $\sqrt{k}/N = \sqrt{\mathcal{I}/N}$. Since $Z \approx \mathcal{L} \times X$ over the bump, $\ln Z$ inherits that uncertainty. This is Eq. (2.5.29).

D6. The Laplace evidence and the Occam factor

Derives (2.5.31), (2.5.32)

The expansion. At the minimum the gradient of χ^2 vanishes, so $\chi^2(\theta) \simeq \chi^2_{\min} + \delta^\top F\,\delta$ with $\delta = \theta - \hat\theta$ and $F_{ij} = \frac{1}{2}\partial_i\partial_j\chi^2$. With a flat prior, Eqs. (2.5.1) and (2.5.2) give

$$Z \simeq \frac{\mathcal{L}_{\mathrm{norm}}\, e^{-\chi^2_{\min}/2}}{V_\pi} \int d^p\delta\ e^{-\frac{1}{2}\delta^\top F\delta}.$$

The integral. Rotate to the eigenvectors of F, with eigenvalues $1/\sigma_i^2$. If the box extends many σ_i from the peak in every direction, the limits can be taken to infinity and each direction gives $\sqrt{2\pi}\,\sigma_i$. The product is $(2\pi)^{p/2}\prod_i \sigma_i = (2\pi)^{p/2}(\det F)^{-1/2}$. Taking the logarithm gives Eq. (2.5.31). For a model linear in θ the expansion is exact.

One direction with a finite box. Keep the limits. $\int_{-W/2}^{W/2} d\delta\, e^{-\delta^2/2\sigma^2} = \sqrt{2\pi}\,\sigma\,[\Phi(a) - \Phi(-a)] = \sqrt{2\pi}\,\sigma\,[2\Phi(a) - 1]$ with $a = W/2\sigma$. Dividing by W gives Eq. (2.5.32). For small a, $2\Phi(a) - 1 \simeq 2a\,\phi(0) = W/\sqrt{2\pi}\sigma$, and the Occam factor tends to 1.

D7. Bayesian complexity and the closure expectations

Derives (2.5.4), (2.5.33), (2.5.34), (2.5.35)

Gaussian posterior. With $\chi^2 - \chi^2_{\min} = \delta^\top F \delta$ and posterior covariance $\langle \delta\delta^\top \rangle = F^{-1}$,

$$C_b = \left\langle \delta^\top F\, \delta \right\rangle = \mathrm{tr}\big(F\,\langle \delta\delta^\top \rangle\big) = \mathrm{tr}\, 1 = p.$$

This is Eq. (2.5.33).

With a box. In the eigenbasis of F, $\delta^\top F \delta = \sum_i \delta_i^2/\sigma_i^2$. If the box is aligned with that basis the posterior factorises, and each δ_i is a Gaussian of width σ_i cut off at $\pm W_i/2$. The variance of a unit Gaussian restricted to $[-a, a]$ follows from integrating $t^2\phi(t)$ by parts, $\int_{-a}^{a} t^2\phi\, dt = [2\Phi(a) - 1] - 2a\phi(a)$, and dividing by the normalisation $2\Phi(a) - 1$. This gives Eq. (2.5.4) and, summed over directions, Eq. (2.5.34). Expanding for small a, the bracket is $a^2/3 = W^2/12\sigma^2$.

Level-1 closure. Chapter 2.3 derives the distribution of the minimum: for a model that is linear near the truth, $\chi^2_{\min}$ of a level-1 fit follows a χ^2 distribution with $N_{\mathrm{dat}} - p_{\mathrm{eff}}$ degrees of freedom, where p_{eff} is the number of directions the data constrain. Its mean is $N_{\mathrm{dat}} - p_{\mathrm{eff}}$ and its variance $2(N_{\mathrm{dat}} - p_{\mathrm{eff}})$. The posterior mean of χ^2 lies above the minimum by the complexity, and $C_b = p_{\mathrm{eff}}$ by Eq. (2.5.34). Adding the two gives $\mathbb{E}\langle\chi^2\rangle_{\mathrm{post}} = N_{\mathrm{dat}}$. This is Eq. (2.5.35).

References

1. 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; code at github.com/HEP-PBSP/colibri. Sections 9 and 10 describe the code as used in this project.
2. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, J. Chem. Phys. 21 (1953) 1087; W. K. Hastings, Biometrika 57 (1970) 97.
3. G. O. Roberts, A. Gelman and W. R. Gilks, Ann. Appl. Probab. 7 (1997) 110, for the optimal acceptance rate.
4. S. Duane, A. D. Kennedy, B. J. Pendleton and D. Roweth, “Hybrid Monte Carlo”, Phys. Lett. B195 (1987) 216.
5. R. M. Neal, “MCMC using Hamiltonian dynamics”, in *Handbook of Markov Chain Monte Carlo* (2011), arXiv:1206.1901. Sections 4 and 5 follow its line of argument.
6. M. D. Hoffman and A. Gelman, “The No-U-Turn Sampler”, J. Mach. Learn. Res. 15 (2014) 1593, arXiv:1111.4246.
7. A. Gelman and D. B. Rubin, Statist. Sci. 7 (1992) 457; A. Vehtari, A. Gelman, D. Simpson, B. Carpenter and P.-C. Bürkner, Bayesian Anal. 16 (2021) 667, arXiv:1903.08008, for the split and rank-normalised versions.

8. J. Skilling, “Nested sampling for general Bayesian computation”, Bayesian Anal. 1 (2006) 833.
9. J. Buchner, “UltraNest: a robust, general purpose Bayesian inference engine”, J. Open Source Softw. 6 (2021) 3001.
10. D. J. Spiegelhalter, N. G. Best, B. P. Carlin and A. van der Linde, “Bayesian measures of model complexity and fit”, J. R. Statist. Soc. B 64 (2002) 583.
11. This project’s ledgers: `context/RESULTS_LEDGER.md` (Phase A: PPDF-1, 3, 7, 8, 9, 11, 12, 16, 18, 19, 21, 25, 28), `context/PHASE_B_LEDGER.md` (PPDF-33, PPDF-36) and `context/PHASE_B_PREREG.md` (Phase B). Every number about our own runs is quoted from them.