Approximation and quadrature

Reconstruct functions from samples or projections, derive deterministic and Monte Carlo quadrature rules, and test whether their error estimates are credible.

Read the subject in depth

Each volume develops its arguments in sequence, with worked problems and links to experiments.

- **Function approximation and numerical integration**
  Construct approximations from the information actually available. Derive their error and observe what changes when the function, data or sampling rule violates an assumption.

Advanced study across the subject

Each sequence connects the course to detailed chapters in neighbouring subjects. The complete course PDF includes their derivations, examples and exercises.

1. **Approximating a measured field**
   Reconstruct a potential from weighted modes. Compare orthogonality in a rectangle, cylinder and sphere, and determine which coefficients can be recovered from noisy measurements.
2. **Sampling, propagation and unresolved scales**
   Apply quadrature and spectral approximation to a real aperture, a decaying field and a radiated-power measurement. Look for an observed convergence rate and explain any plateau.

How the course fits together

1. **Chapters 1–4**
   Connect samples locally, choose polynomial nodes, then impose smooth joins to obtain splines. Integrating a reconstructed curve produces a quadrature rule with an error inherited from the approximation.
2. **Chapters 5–8**
   Choose nodes to integrate moment conditions exactly or adapt intervals where estimates disagree. Divided differences make incremental interpolation possible, while minimax approximation controls the worst pointwise error.
3. **Chapters 9–12**
   Represent functions by global frequencies or orthogonal projections. Endpoint regularity can limit convergence. Random quadrature replaces geometrical spacing with sampling and a statistical error estimate.

Contents

1. Chapter 1 · Piecewise interpolation and error bounds
2. Chapter 2 · Polynomial interpolation and node placement
3. Problem sheet 1 · Nodes and interpolation
4. Chapter 3 · Hermite pieces and cubic splines
5. Chapter 4 · Composite quadrature and cancellation
6. Problem sheet 2 · Smooth pieces and integrated loads
7. Chapter 5 · Gaussian quadrature and exactness
8. Chapter 6 · Adaptive integration and error estimates
9. Problem sheet 3 · Accuracy with few evaluations
10. Chapter 7 · Divided differences and derivative data
11. Chapter 8 · Uniform approximation and alternation
12. Problem sheet 4 · Data and uniform error
13. Chapter 9 · Fourier projection and sampled frequencies
14. Chapter 10 · Weighted projections and Legendre polynomials
15. Problem sheet 5 · Spectral measures of error
16. Chapter 11 · Quadrature near singularities and oscillations
17. Chapter 12 · Monte Carlo integration and stratification
18. Problem sheet 6 · Choosing where to sample
19. Chapter 13 · Evaluating an interpolant without rebuilding its polynomial
20. Chapter 14 · When a trapezoidal sum counts the wrong frequency perfectly
21. Chapter 15 · Deriving Romberg integration and recognising a broken error model
22. Chapter 16 · Summing a Chebyshev series by working backwards
23. Chapter 17 · Where should a random quadrature spend its samples?
24. Chapter 18 · Integrating an infinite endpoint with a finite table of samples
25. Problem sheet 7 · Project · How many sensors measure the mean around a ring?
26. Problem sheet 8 · Integrate a rapidly oscillating signal without resolving every crest
27. Function approximation and numerical integration
28. Approximating a measured field
29. Sampling, propagation and unresolved scales
30. Applied study
31. Notation and assumptions
32. Further reading

Chapter 1 · Piecewise interpolation and error bounds

Construct a continuous profile from sensor values, prove a local interpolation bound and distinguish sampling error from measurement uncertainty.

FROM FIRST PRINCIPLES

Build a line from two measured endpoints

Suppose values y_0, y_1 are known at $x_0 < x_1$ and no slopes are available. The simplest model reproducing both data is a line of constant slope. Let t denote the fraction of the panel traversed at the query point x. Starting with the slope of the chord, rewrite the same line in terms of its endpoint weights:

$$\begin{aligned} t &= \frac{x - x_0}{x_1 - x_0}, \\ I(x) &= y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0} \\ &= (1 - t)y_0 + t y_1. \end{aligned}$$

Its weights are nonnegative on the panel and add to one, so constants and endpoint values are reproduced exactly. When the data come from a twice differentiable function with $|f''| \le M$, curvature controls departure from the line. The two-node remainder gives

$$|f(x) - I(x)| \le \frac{M}{2}(x - x_0)(x_1 - x) \le \frac{M(x_1 - x_0)^2}{8}.$$

The last bound occurs at the midpoint. Halving panel width reduces this deterministic bound by a factor of four. Measurement errors in the endpoint values still enter through the interpolation weights and must be budgeted separately.

DERIVATION AND APPLICATIONS

Balance curvature bias against uncertainty in interpolated readings

Consider a panel of length $h = b - a$, with unknown smooth signal f and two independent noisy observations $y_a = f(a) + \epsilon_a,\ y_b = f(b) + \epsilon_b$. Assume each error has zero mean and a known variance, and that $|f''| \le M$ throughout the panel. For $t = (x - a)/h$, linear interpolation gives the estimator $\widehat{f} = (1 - t)y_a + t y_b$. Its expectation follows simply by taking the expected value of each measured endpoint,

$$\mathrm{E}[\widehat{f}] - f(x) = (1 - t)f(a) + t f(b) - f(x), \qquad |\,\text{bias}\,| \le \tfrac{1}{2} M h^2 t(1 - t).$$

Independence eliminates the mixed error term when the noise contribution is squared. Consequently the noise variance depends on the interpolation location even on a flat signal,

$$\operatorname{var}(\widehat{f}) = (1-t)^2 s_a^2 + t^2 s_b^2.$$

Expanding the mean squared error about the estimator's own expectation yields a sum of variance and squared bias. For equal endpoint standard deviations s, the midpoint therefore satisfies

$$\mathrm{E}[(\widehat{f} - f)^2] \le \frac{s^2}{2} + \left(\frac{M h^2}{8}\right)^2.$$

Grid refinement changes the second term, whereas independent repeat measurements at each endpoint reduce the first. Common-mode drift adds a covariance contribution that survives averaging between the two endpoints. In particular, an identical unknown offset passes unchanged through the interpolation because the weights sum to one.

WORKED EXAMPLE

Determine which uncertainty dominates a midpoint

At two stations $0.40\,\mathrm{m}$ apart the reading errors have $s = 0.030\,\mathrm{K}$, while $M = 0.50\,\mathrm{K/m^2}$. The maximum midpoint bias is $0.010\,\mathrm{K}$ and the noise deviation is $0.0212\,\mathrm{K}$. Their root-mean-square budget is

$$\sqrt{0.030^2/2 + 0.010^2} = 0.0235\,\mathrm{K}.$$

It is already dominated by reading noise.

SECOND APPLICATION

Select a spacing from a root-mean-square target

For another instrument let $s = 0.020$, $M = 2.00$, and target midpoint root-mean-square error 0.030, in consistent units. The permitted bias magnitude is

$$\sqrt{0.030^2 - 0.020^2/2} = 0.0265.$$

Thus

$$h \le \sqrt{8(0.0265)/2.00} = 0.325.$$

If the noise were doubled at fixed error target, the remaining squared bias allowance would shrink to 0.0001.

Start from this chapter. For a convex quadratic, the chord lies above the curve. Predict the sign and size change of the midpoint error when each source interval is halved.

OneSimulation laboratory · Measure what interpolation does to a heat integral

Reconstruct from the information available

A line of thermometers supplies values y_i at increasing coordinates x_i. It does not determine a unique continuous function: any function vanishing at every sensor can be added to a candidate profile. Choosing piecewise straight segments asserts that the unresolved curvature between consecutive readings is small enough for the intended measurement. This assumption may be excellent for predicting a mean temperature and poor for estimating a local heat flux, since differentiation increases the influence of measurement noise.

For $x \in [x_i, x_{i+1}]$, let $s = (x - x_i)/h_i$, where $h_i = x_{i+1} - x_i > 0$. The unique affine function matching both readings is $I_h y = (1 - s)y_i + s y_{i+1}$. It follows by writing an unknown intercept and slope, imposing the two endpoint values and eliminating both unknown coefficients. The weights are nonnegative and sum to one. The interpolated value lies between adjacent readings, constants are reproduced, and bounded additive errors of size ϵ at the two sensors change the interpolant by at most ϵ. These properties can fail when extrapolating a considerable distance outside the measured interval.

Why the error is quadratic

Suppose f is twice continuously differentiable on a panel. Fix an interior point t, and choose a coefficient c so that $g(x) = f(x) - I_h f(x) - c(x - x_i)(x - x_{i+1})$ also vanishes at t. It already vanishes at both panel endpoints. Rolle's theorem applied twice gives a point ξ where $g''(\xi) = 0$. The interpolant has zero second derivative there, so $c = f''(\xi)/2$. Evaluate the identity at t to obtain

$$f(t) - I_h f(t) = \frac{f''(\xi)}{2}(t - x_i)(t - x_{i+1}).$$

The product is nonpositive within the interval and has greatest magnitude $h_i^2/4$ at its midpoint. Thus the uniform error is at most $h_i^2 \max |f''|/8$. If every panel width is halved, an upper bound based on the same curvature falls by four. When curvature is positive, the segment overestimates the function. Knowledge of the curvature sign supplies information that an absolute error bound would discard.

Background · Can refining noisy samples make interpolation less useful?

Take the midpoint of two independent sensor readings with the same random error variance σ^2. The midpoint interpolant averages those readings. Its variance is $\sigma^2/2$, regardless of how close together the sensors stand. Meanwhile the size of the curvature bias is approximately $|f''|h^2/8$. For a twice differentiable field with nearly constant curvature and unbiased readings, the midpoint mean squared error is approximately

$$\mathbb{E}[(\hat{f} - f)^2] \simeq \frac{(f''h^2)^2}{64} + \frac{\sigma^2}{2}.$$

Refining the sampling interval reduces the first term but leaves the second term. Once the two terms are comparable, repeat readings or a calibrated smoothing fit can improve a value estimate more efficiently. If both readings share the same offset, taking their average preserves that offset in full.

Turn the bound into a sampling decision

On the unit interval consider a temperature $T = 290 + 8x^2$ K, with x given numerically in metres. On the first panel of length 0.25 m the endpoint readings are 290 and 290.5 K. Their average is 290.25 K. The actual midpoint is 290.125 K, an overestimate of 0.125 K. The bound $16(0.25)^2/8 = 0.125$ K is attained because the curvature is constant. A second panel of the same width has the same maximum interpolation error even though its temperature change is different.

Now allow an uncertainty of 0.03 K at every source node. A conservative total bound is $2h^2 + 0.03$ K for h in metres. Achieving 0.05 K requires $h \le 0.10$ m, or at least 11 measurements including both ends. No amount of interpolation onto a denser display grid changes this source requirement. For the slope, two readings with opposite errors can change the estimated derivative by $2\epsilon/h$, which grows when h shrinks. Choosing sensor spacing therefore depends on the quantity to be estimated. A spacing suitable for accurate point values may require additional smoothing before slopes inferred from noisy data are useful.

Interpret a finite sensor footprint

Sometimes the observation is already a mean over a finite sensor footprint. For a symmetric footprint of width w centred at x, the mean of a smooth field is $f(x) + f''(x)w^2/24 + O(w^4)$, by Taylor expansion and symmetry. For the quadratic temperature above this correction is exactly $16w^2/24$ K when w is measured in metres. With a 0.10 m footprint it is about 0.0067 K above the point value. Interpreting a cell mean as a point reading therefore adds another positive bias before the chord interpolation even begins. This distinction becomes important when comparing a temperature probe with simulation values stored as cell averages.

Distribute measurements according to the curvature they must

resolve

An equally spaced grid spends the same sampling budget in every part of a record. Suppose instead that the trajectory is twice continuously differentiable on each of several known regions, with a certified bound M_i for $|f''|$ on region i. Over a panel of width h_i the largest linear interpolation error is at most $M_i h_i^2/8$. Thus keeping every local bound below a specified tolerance ϵ requires $h_i \le \sqrt{8\epsilon/M_i}$ whenever $M_i > 0$. A region with four times the curvature bound needs twice as many panels per unit length.

For a fixed total number of panels the reverse question is also helpful. In a region of length L_i with n_i equal panels its bound is $M_i L_i^2/(8n_i^2)$. To minimise the largest such bound over positive continuous allocations, equalise those dominant local bounds. Otherwise moving some resolution from a region with a smaller bound to the largest can improve the maximum. The resulting allocation satisfies

$$n_i \propto L_i \sqrt{M_i}.$$

Counts must then be rounded and their actual bounds checked. Known transitions in curvature should occur at grid nodes if each local bound is to apply to its entire panel.

Place nine panels across a changing acceleration

A displacement history lasts three seconds. In the first two seconds its acceleration magnitude is bounded by 2 m/s², and in the last second by 18 m/s². The relative panel allocations are

$$2\sqrt{2} : \sqrt{18} = 2 : 3.$$

Nine panels cannot achieve exactly that ratio. Try four panels in the first interval and five in the second. Their widths are 0.5 and 0.2 s, giving error bounds $2(0.5)^2/8 = 0.0625$ m and $18(0.2)^2/8 = 0.0900$ m. Moving one panel to the late region instead gives bounds $2(2/3)^2/8 = 0.1111$ m and $18(1/6)^2/8 = 0.0625$ m. The four–five allocation has the smaller worst guarantee.

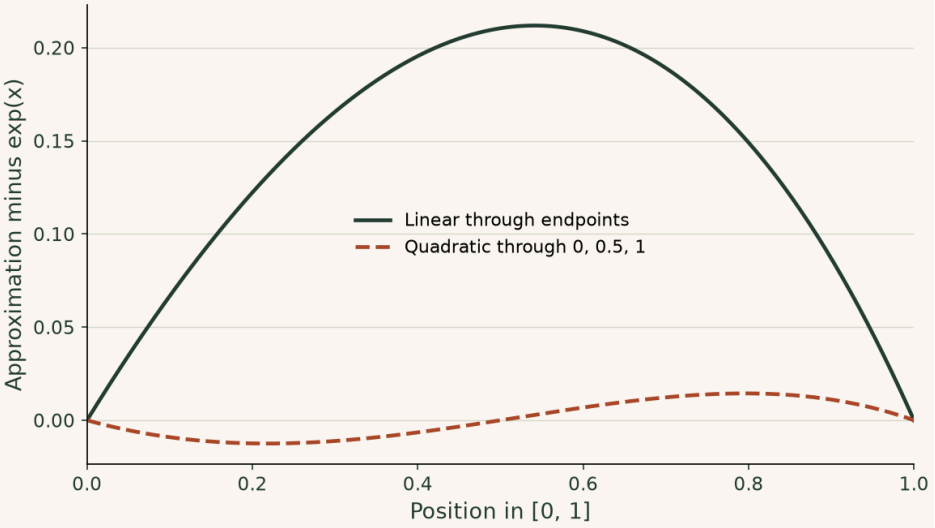
Nine equally spaced panels have width one third second throughout and include the transition at two seconds. Their late bound is 0.25 m, almost three times the nonuniform worst bound. This comparison is about a guaranteed maximum between observations, not only a discrepancy at selected test locations. Piecewise quadratic motions with constant accelerations attain the midpoint bound on each panel; slopes and positions can be matched at the transition without altering either local error.

The reasoning assumes acceleration bounds independent of the sampled readings. Estimating them from those same coarse readings can miss a short acceleration pulse entirely. If endpoint measurements also contain error bounded by 0.03 m, the convex interpolation weights add at most 0.03 m to each panel bound. The chosen allocation then certifies 0.12 m. Refining the grid further can reduce curvature error but cannot eliminate the measurement floor without better data.

If the allowable total error were 0.10 m, leaving 0.03 m for sensor uncertainty would allow at most 0.07 m for curvature. Four early panels already meet that condition. The late section would require at least six panels because $18/(8 \cdot 6^2) = 0.0625$, giving a ten-panel acquisition plan.

Where does a trustworthy interpolant cease to be accurate?

Interpolate the same smooth function on one interval with two and then three nodes. For the line the error changes with curvature, while for the quadratic the node product changes sign at the middle point. Inspect the signed error between observations.



The line lies above the convex exponential between its endpoints. Adding a midpoint makes the quadratic exact there but creates error lobes of opposite sign in the two halves. Agreement at the nodes alone tells nothing about the interpolant between them.

▶ **Read values from the plot**

Change parameters · Where does a trustworthy interpolant cease to be accurate?

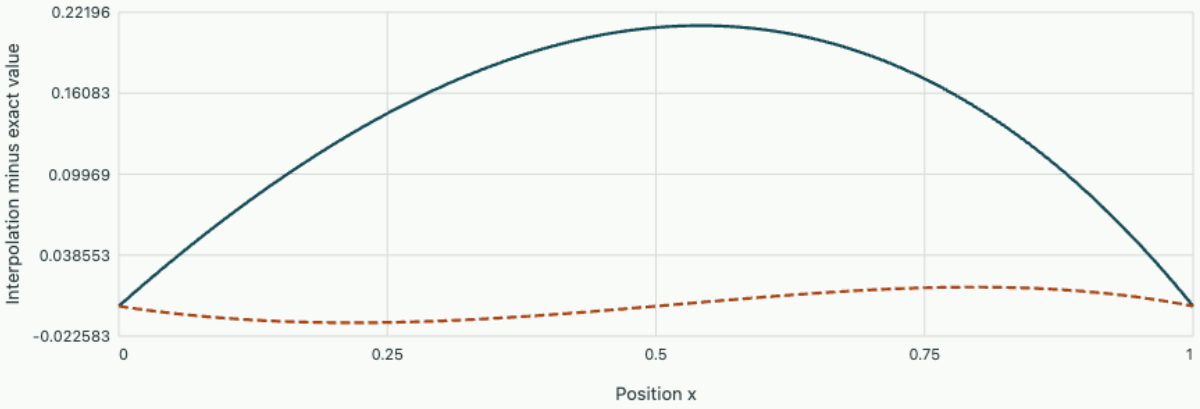
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Interval length 1

Exponential growth rate 1



Two endpoint values Endpoints and midpoint

All nodal errors vanish. Halve the interval length at fixed rate to compare the interior error of linear and quadratic interpolation.

Study the full derivation

Approximating a resonance with a rational function · Function approximation and numerical integration

A smooth response on a measured interval can be governed by a nearby pole. Derive a low-order Padé fit, construct a barycentric rational approximant from samples and check whether a predicted pole survives independent data.

Chapter 2 · Polynomial interpolation and node placement

Derive the Lagrange interpolant, diagnose amplification of sample noise and use cosine-distributed nodes to control a polynomial error factor.

FROM FIRST PRINCIPLES

Construct the unique interpolating polynomial

A polynomial of degree at most n has $n+1$ coefficients. To interpolate $n+1$ distinct abscissae x_i, build one basis polynomial for each measurement. The factors $x - x_j$ make it vanish at every other node; dividing by its value at x_i gives unit height there,

$$\ell_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j},$$

$$p(x) = \sum_{i=0}^{n} y_i \ell_i(x).$$

At $x = x_i$ all terms except y_i disappear. Uniqueness follows because the difference of two such fits would have degree at most n and at least $n+1$ distinct roots. The same representation also exposes sensitivity to equal bounded reading errors,

$$|\delta p(x)| \leq \left(\sum_{i=0}^{n} |\ell_i(x)| \right) \max_i |\delta y_i|.$$

The sum depends on the node placement and the evaluation location. Distinctness ensures existence and uniqueness but gives no guarantee of uniformly small amplification at high degree.

Background · Interpolation condition

$$p(x_i) = f_i$$

The polynomial matches the tabulated values at distinct nodes. Small nodal discrepancies do not guarantee a small error between nodes.

DERIVATION AND APPLICATIONS

Derive barycentric weights from the node polynomial

Every cardinal basis polynomial contains nearly the same product of factors. Factoring out the common part leads to a practical evaluation formula. Assume $n+1$ distinct real nodes and denote their monic

node polynomial by $L(x) = \prod_{i=0}^{n}(x - x_i)$. At a root x_i only one term in the derivative survives. Define a weight using that derivative,

$$L'(x_i) = \prod_{j \neq i}(x_i - x_j),$$

$$w_i = \frac{1}{L'(x_i)},$$

$$\ell_i(x) = L(x)\frac{w_i}{x - x_i}.$$

For x away from the nodes, express the interpolant as $L(x)$ times a rational sum. Apply the same formula to a polynomial identically equal to one, using the fact that interpolation reproduces constants. The two identities have the same factor $L(x)$,

$$p(x) = L(x)\sum_i \frac{w_i y_i}{x - x_i},$$

$$1 = L(x)\sum_i \frac{w_i}{x - x_i}.$$

Their ratio cancels the product and gives the barycentric form

$$p(x) = \frac{\sum_i w_i y_i/(x - x_i)}{\sum_i w_i/(x - x_i)}.$$

A common factor in the weights also cancels, allowing rescaling. At a node the individual terms have poles but the limit equals its stored datum; an implementation should return that datum directly on an exact match. The representation improves evaluation without repairing intrinsically poor node placement. If nodes are fixed and many observation sets arrive, their weights can be reused.

WORKED EXAMPLE

Reconstruct a calibration at a new input

Use nodes $0, 1, 3$, data $2, 1, 5$, and common-scaled weights $(2, -3, 1)$. At $x = 2$, the three terms in the denominator are $1, -3, -1$; their data-weighted sum is $2 - 3 - 5 = -6$. Thus

$$p(2) = (-6)/(-3) = 2.$$

Direct construction gives $p(x) = x^2 - 2x + 2$, confirming the reconstructed value.

SECOND APPLICATION

Derive reusable weights for a different measurement grid

For nodes $-1, 0, 1, 2$, the four products $L'(x_i)$ are $-6, 2, -2, 6$. Multiplying all reciprocals by six gives weights $(-1, 3, -3, 1)$. At $x = 1/2$, the normalized interpolation weights reduce to $(-1, 9, 9, -1)/16$. An

isolated unit error at the right endpoint changes this midpoint prediction by $-1/16$, while a uniform shift of all readings by one moves the result by exactly one.

One polynomial through all measurements

Given $n+1$ distinct nodes x_j, seek one polynomial p of degree at most n matching y_j at each node. Define $\ell_j(x) = \prod_{i \neq j} (x - x_i)/(x_j - x_i)$. This basis has value one at its own node and zero at all the others, so $p = \sum_j y_j \ell_j$ satisfies the observations. If another such polynomial existed, the difference would have $n+1$ separate zeros despite degree at most n. It must vanish identically. Uniqueness concerns exact arithmetic and says little by itself about sensitivity.

For the nodes $-1, 0, 1$, the three basis functions are $x(x-1)/2,\ 1-x^2,\ x(x+1)/2$. At $x = 1/2$ their values are $-1/8, 3/4, 3/8$. They sum to one but are not all positive. Consider samples $2, 1, 2$: the predicted value is 1.25, consistent with the quadratic $1 + x^2$. If the three readings instead carry independent signed perturbations bounded by ϵ, the possible prediction error at this point reaches

$$(1/8 + 3/4 + 3/8)\epsilon = 1.25\epsilon.$$

The worst perturbation takes signs matching those of the basis values.

Separate noise and approximation

The maximum of $\sum_j |\ell_j(x)|$ over the interval, the Lebesgue constant, bounds the operator's amplification of data errors. A different quantity controls the error of interpolation from exact samples. For sufficiently smooth f, repeating the Rolle argument with $n+2$ zeros gives

$$f(x) - p(x) = f^{(n+1)}(\xi) \prod_j (x - x_j)/(n+1)!.$$

This formula compares the exact function with the selected polynomial. It contains both a geometric factor controlled by the nodes and a higher derivative that may grow quickly with degree. A small geometric factor cannot by itself establish convergence for an arbitrary function.

On $[-1, 1]$ let $T_m(x) = \cos(m \arccos x)$. From the cosine addition formula, $T_{m+1} = 2xT_m - T_{m-1}$ with $T_0 = 1, T_1 = x$. Thus T_m is a polynomial of degree m with leading coefficient 2^{m-1} for $m \geq 1$. Its m zeros lie at $\cos[(2k+1)\pi/(2m)]$. Choosing them as interpolation nodes makes the monic geometric factor $2^{1-m}T_m(x)$, whose maximum magnitude is exactly 2^{1-m}.

Why the alternating extrema matter

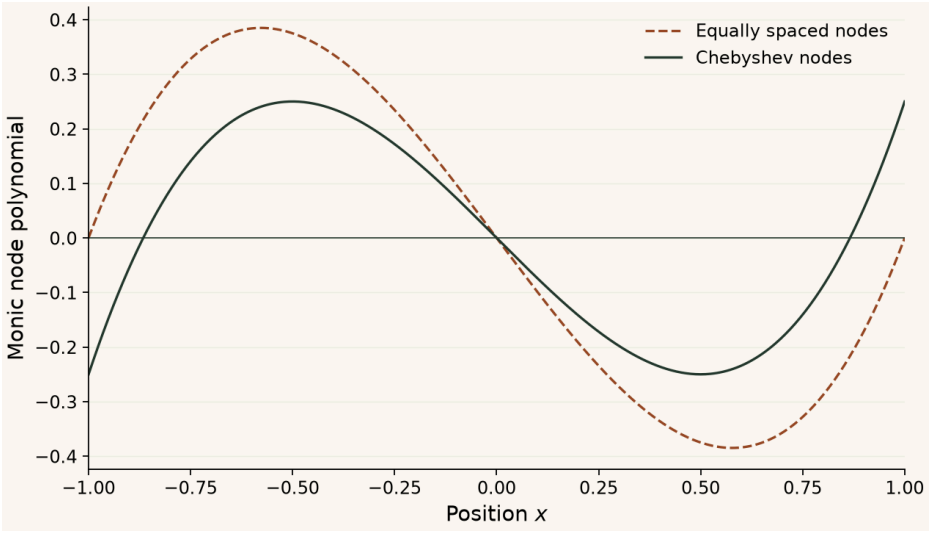
At the $m+1$ points $\cos(k\pi/m)$, this monic factor alternates between its positive and negative extreme. Suppose another monic degree-m polynomial had a strictly smaller maximum magnitude. Subtract it from the Chebyshev factor. The leading powers cancel, leaving degree at most $m-1$. At the successive extrema the difference must alternate in sign, forcing at least m distinct zeros between them. That

contradicts its degree. This minimax argument concerns the node polynomial, and explains the deliberate concentration of nodes near the endpoints.

For $m = 3$ the recommended points are $0, \pm\sqrt{3}/2$; their monic factor is $x^3 - 3x/4$ with magnitude at most 0.25. The equally spaced alternative $-1, 0, 1$ has factor $x^3 - x$, whose largest magnitude is

$$2/(3\sqrt{3}) = 0.385.$$

If data already exist at fixed locations, moving them is seldom an option. Subdividing the interval and fitting lower-degree polynomials locally instead limits long-range coupling. For evaluation at many points, barycentric interpolation can reuse stored node weights while avoiding explicit construction of a large monomial coefficient vector.



Three-node interpolation. The largest absolute value of the monic node factor is smaller for the Chebyshev roots, even though the endpoints are not interpolation nodes.

Predict the degree from the nearest complex singularity

Well-chosen nodes control amplification of data errors. The rate at which a smooth target itself can be approximated also depends on what happens away from the real interval. Write $x = (z + z^{-1})/2$. The unit circle maps to $[-1, 1]$, and the circle $|z| = \rho > 1$ maps to an ellipse with foci at the endpoints. The Chebyshev polynomial obeys $T_k(x) = (z^k + z^{-k})/2$. Consequently the Laurent coefficients of $f((z + z^{-1})/2)$ determine the Chebyshev coefficients of f. If the function extends analytically through the closed ellipse and has magnitude at most M there, the Cauchy coefficient bound gives $|a_k| \le 2M\rho^{-k}$ for $k \ge 1$. Summing this geometric tail bounds the error of the degree-n truncated series by $2M\rho^{-n}/(\rho - 1)$ on the interval. A real-axis corner invalidates this analytic argument.

A rational response with an exactly measurable tail

For $0 < q < 1$, take $f(x) = (1 - q^2)/(1 - 2qx + q^2)$. With $x = \cos\theta$, summing the two geometric series in $qe^{\pm i\theta}$ gives

$$f(x) = 1 + 2\sum_{k\geq 1} q^k T_k(x).$$

Its pole at $x = (q + q^{-1})/2$ lies on the ellipse with parameter $1/q$. The exact coefficients decay like q^k. At $x = 1$ all omitted terms have the same positive sign. Since $|T_k| \leq 1$, the uniform truncation error equals $2q^{n+1}/(1 - q)$.

For $q = 1/3$ the pole is at $5/3$, and the error equals 3^{-n}. Degree six leaves error $1/729 \approx 0.00137$. Degree seven reduces it to $1/2187 \approx 0.000457$. This establishes a degree for a specified error before constructing a node grid. These exact errors concern the truncated series. An interpolant fitted on Chebyshev nodes combines unresolved modes into lower coefficients and need not have precisely the same error, although analyticity still gives geometric convergence.

Practise on sheet 1 · How close can a pole move before the degree budget fails?. Further reading · University of Oxford · Chebyshev Series, Nick Trefethen · Lecture 3, Chebyshev series, part 2, Smoothness and convergence theorems.

Evaluate without forming monomial coefficients

Define barycentric weights $w_j = 1/\prod_{i\neq j}(x_j - x_i)$. Dividing the Lagrange interpolation formula and the identity $\sum \ell_j = 1$ by the common nodal polynomial gives the evaluation ratio $\sum_j w_j y_j/(x - x_j)$ divided by $\sum_j w_j/(x - x_j)$. At an exact interpolation node, return its stored value directly. Multiplying all weights by the same constant has no effect, which is useful for numerical scaling.

For nodes $-1, 0, 1$ the weights are proportional to $1, -2, 1$. At $x = 1/2$ with data $2, 1, 2$, the weighted numerator is $4/3 - 4 - 4 = -20/3$, and the denominator is

$$2/3 - 4 - 2 = -16/3.$$

Their ratio gives 1.25. This evaluates the very same polynomial as the basis construction. Better arithmetic for evaluation cannot remove an intrinsically large Lebesgue constant associated with a poor set of nodes.

Determine exactly how node errors can reinforce

For fixed distinct interpolation nodes let $L_j(x)$ denote the cardinal polynomials. Changing the readings by errors e_j changes a reconstructed value by $\sum_j e_j L_j(x)$. With independent bounds $|e_j| \leq \delta$, taking absolute values gives an amplification no larger than $\delta \sum_j |L_j(x)|$. At any fixed evaluation point this bound can be attained by choosing the sign of e_j to agree with the sign of $L_j(x)$. The sum of absolute cardinal values is therefore the exact pointwise error amplification for this particular error model.

If all reading errors are a single unknown constant instead, the amplification is exactly one because the cardinal polynomials sum to one. The same nodes have different noise behaviour for coherent offsets and independent bounded perturbations. Statistical independence with equal variances provides a third measure: the output variance is $\sigma^2 \sum_j L_j(x)^2$. The worst pointwise absolute sum, its maximum across

the interval, and this root-sum-of-squares factor should not be interchanged. They correspond to questions about adversarial signs, location and probabilistic noise, respectively.

Check a three-station calibration inside and beyond its range

Take stations at $-1, 0, 1$. Their cardinal polynomials are $x(x-1)/2,\ 1-x^2,\ x(x+1)/2$. At $x = 1/2$ they take values $(-1/8, 3/4, 3/8)$. The weights sum to one, but their absolute values sum to $5/4$. If each observation is accurate to within 0.02 units, the interpolated calibration can be wrong by 0.025 from observations alone. The specific error pattern $(-0.02, 0.02, 0.02)$ attains that value. Polynomial exactness for noise-free quadratic inputs provides no protection against that sign pattern.

With independent standard deviations 0.02, the same evaluation instead has standard deviation $0.02\sqrt{1/64 + 9/16 + 9/64} = 0.02\sqrt{46}/8 \approx 0.01696$. The smaller value follows from the assumed random cancellation and is not a guaranteed bound. At $x = 2$, beyond the calibration interval, the same weights become $1, -3, 3$. The worst bounded measurement error is now $7(0.02) = 0.14$. Every tabulated station can pass a calibration check while extrapolation is seven times more sensitive than an original reading.

Inside $[0, 1]$ the absolute sum simplifies to $1 + x - x^2$; its maximum occurs at one half and equals $5/4$. Symmetry gives the same maximum on the other half. For quadratic interpolation this gives a complete guarantee throughout the stated interval, with construction of the worst perturbation. A graph of the fitted polynomial itself would reveal none of these conditioning facts. A sensor specification therefore needs both an admissible evaluation range and a description of how the raw station errors may vary together.

The maximum amplification also gives a way to compare retrospective calibrations. If a process must meet a total error of 0.03 units and each raw persistent bounded error may be as large as 0.02, the three-point design leaves at most $0.03 - 1.25(0.02) = 0.005$ units for polynomial approximation bias throughout the interval. A check at the three stations themselves would assign a larger remaining allowance and would miss the interior maximum.

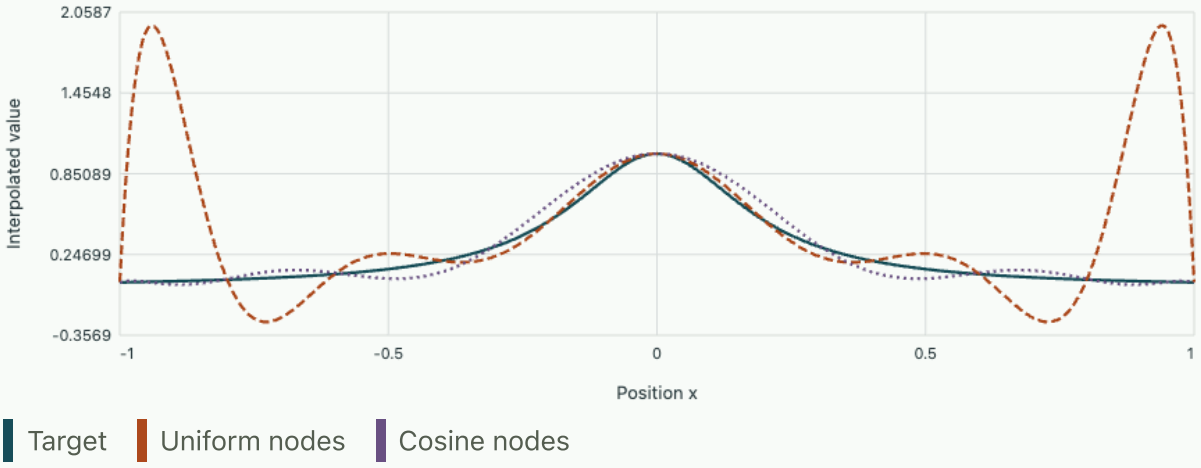
Change parameters · Polynomial interpolation and node placement

Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Polynomial degree **10**



Maximum observed error on 501 test points: uniform 1.9156, cosine 0.13217.

Study the full derivation

Integrating cosine interpolants on nested grids · Function approximation and numerical integration

Chebyshev points turn polynomial interpolation into a cosine series. Derive the exact moments of each basis polynomial, explain aliasing and find a case where two consecutive quadrature grids confidently agree on the wrong answer.

Problem sheet 1 · Nodes and interpolation

Analyse local interpolation errors, demonstrate noise amplification under extrapolation and compare two sets of global interpolation nodes.

Estimate a probe requirement

A smooth pressure profile on a 2 m pipe satisfies $|p''| \le 24$ kPa/m². Each probe has absolute uncertainty 0.01 kPa. Bound the total error of piecewise linear interpolation. Find the largest equal spacing consistent with 0.04 kPa total error and the smallest number of probes including the two endpoints.

▼ **Worked solution**

Between measured nodes the interpolation error of exact samples is bounded by $24h^2/8 = 3h^2$ kPa. The two sample errors enter with nonnegative weights summing to one, adding at most 0.01 kPa. The requirement $3h^2 + 0.01 \le 0.04$ gives $h \le 0.10$ m. Twenty intervals on a 2 m pipe require 21 probes.

This calculation controls worst-case pressure values and assumes the curvature estimate holds on every panel. A single unnoticed discontinuity would invalidate its use on that panel.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Predict beyond the sensors

Two sensors at positions 0 and 1 m each have error at most 0.5 K. Compute the interpolation weights at 0.4 m and the extrapolation weights at 1.5 m. Find the largest change to each predicted temperature and provide a sign pattern attaining each bound.

▼ **Worked solution**

At 0.4 the weights are 0.6 and 0.4. Equal positive errors produce the worst-case rise of 0.5 K. At 1.5 the weights are -0.5 and 1.5. Taking the first error negative and the second positive yields

$$0.25 + 0.75 = 1.0\,\mathrm{K}.$$

Even for a perfectly affine true profile the second estimate is more sensitive to data noise. If the true profile is curved, extrapolation adds an entirely separate approximation error.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Construct a minimax nodal factor

Choose two symmetric nodes $\pm c$ on $[-1, 1]$. Find the maximum magnitude of their monic nodal polynomial and minimise it over $0 \le c \le 1$. Use the result to bound the interpolation error when $|f''| \le 4$.

▼ **Worked solution**

The nodal polynomial is $x^2 - c^2$; its negative extreme is $-c^2$ at zero and positive extremes occur at the endpoints with value $1 - c^2$. Thus its norm is $\max(c^2, 1 - c^2)$, minimised at $c^2 = 1/2$. The nodes are $\pm 1/\sqrt{2}$ and the best monic norm is $1/2$.

The Lagrange remainder yields maximum error at most $4(1/2)/2 = 1$. For endpoints as nodes the same reasoning gives a bound of 2. The change improves this worst-case bound even though the endpoint readings themselves are no longer sampled.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

How close can a pole move before the degree budget fails?

For the same rational family, a changed physical parameter gives $q = 1/2$. Locate the pole and find the smallest truncation degree giving uniform error at most 0.002. Compare with $q = 1/3$ at that tolerance. Explain why increasing the density of a plot cannot establish a smaller worst-case error than the derived tail.

▼ **Worked solution**

The pole is $(1/2 + 2)/2 = 1.25$, closer to the interval than $5/3$. The exact tail is 2^{1-n}. At degree nine it is $1/256 \approx 0.00391$. At ten it is $1/512 \approx 0.00195$. Thus ten is the first sufficient degree.

For $q = 1/3$, degrees five and six give $1/243$ and $1/729$, so six suffices. The four extra degrees are explained by the slower coefficient decay. The endpoint $x = 1$ attains the tail exactly, regardless of how densely interior points are plotted.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Revisit chapter 2 · Predict the degree from the nearest complex singularity.

Chapter 3 · Hermite pieces and cubic splines

Build a cubic from endpoint values and slopes, then determine missing curvature information by joining adjacent pieces smoothly.

FROM FIRST PRINCIPLES

Determine a cubic from endpoint values and slopes

On one panel $[a, b]$, approximate a smooth curve with a cubic and use both endpoint heights and slopes. Let $h = b - a > 0$, $t = (x - a)/h$, endpoint heights y_0, y_1, and derivatives d_0, d_1 with respect to x. The conditions at the left endpoint immediately fix two coefficients,

$$p(t) = y_0 + h d_0 t + A t^2 + B t^3.$$

The factor h converts a derivative per unit x into one per unit t. Write $\Delta = y_1 - y_0$. Imposing the right endpoint value and derivative gives two linear equations,

$$A + B = \Delta - h d_0,$$
$$2A + 3B = h(d_1 - d_0).$$

Solving gives

$$A = 3\Delta - h(2d_0 + d_1),$$
$$B = -2\Delta + h(d_0 + d_1).$$

Thus four measurements determine four coefficients uniquely. Adjacent panels sharing heights and slopes join with a continuous first derivative. If only heights are measured, a spline construction supplies the missing information through smoothness conditions between panels and boundary conditions at the ends.

DERIVATION AND APPLICATIONS

Check whether a cubic panel preserves a rising signal

A cubic matching two heights and slopes can overshoot even when both endpoint heights increase. Consider a Hermite panel with $y_1 > y_0$, length h, slope of the chord $m = (y_1 - y_0)/h$, and endpoint derivatives d_0, d_1. The derivative of the cubic is a quadratic. Express that derivative using the nonnegative Bernstein basis on $0 \le t \le 1$,

$$p'(x) = d_0(1 - t)^2 + 2b\,t(1 - t) + d_1 t^2.$$

Here p' is with respect to the physical coordinate x. The first and last coefficients enforce the endpoint derivatives. Each of the three basis functions has integral $1/3$ over t. Since integrating $p'(x)$ across the physical panel must return the specified height gain, the middle coefficient is fixed,

$$y_1 - y_0 = \frac{h}{3}(d_0 + b + d_1),$$
$$b = 3m - d_0 - d_1.$$

Nonnegative coefficients provide a simple sufficient condition for a nondecreasing panel. The basis functions also sum to one, so they immediately bound the derivative when the condition holds,

$$d_0, d_1 \geq 0, \quad d_0 + d_1 \leq 3m \quad \Longrightarrow \quad 0 \leq p'(x) \leq \max(d_0, b, d_1).$$

A negative middle coefficient alone does not prove reversal, since the three basis functions overlap. If the sufficient test fails, evaluate the minimum of the quadratic on the panel, including its endpoints and any interior stationary point. Local monotonicity matters when interpolating cumulative counts or instrument calibrations that must be invertible.

WORKED EXAMPLE

Certify a rising calibration between samples

A signal rises from 2 to 5 over $h = 2$, so $m = 1.5$. Reported slopes are $d_0 = 1$ and $d_1 = 3$, giving $b = 0.5$. The derivative is nonnegative everywhere. At the midpoint it equals

$$0.25(1) + 0.5(0.5) + 0.25(3) = 1.25.$$

No hidden local maximum can lie inside this panel.

SECOND APPLICATION

Find a reversal that endpoint slopes conceal

On a separate unit panel take a rise of one and equal endpoint slopes of four. Then $m = 1$, $b = 3 - 8 = -5$, and the midpoint derivative is

$$1 - 2.5 + 1 = -0.5.$$

Both measured endpoint derivatives are positive, yet the fitted curve locally decreases. Replacing both slopes by 1.5 gives $b = 0$ and restores a certifiably nondecreasing interpolation while preserving the two measured heights.

Four conditions on a single panel

Line segments reproduce values but change slope abruptly at a sensor. A trajectory planner may require a smooth velocity as well as an accurate position. On a panel of width h let $s = (x - x_i)/h$. A cubic polynomial has four coefficients. Specifying y_i, y_{i+1} and slopes d_i, d_{i+1} determines it uniquely. A convenient basis writes the interpolant as

$$H = (2s^3 - 3s^2 + 1)y_i + (-2s^3 + 3s^2)y_{i+1} + h(s^3 - 2s^2 + s)d_i + h(s^3 - s^2)d_{i+1}.$$

To reconstruct these terms, seek a cubic multiplying y_i with value one at $s = 0$, zero at $s = 1$, and zero slope at both ends. The four linear conditions give $2s^3 - 3s^2 + 1$. The other basis functions follow by imposing a unit on their own datum and zero on the remaining three. The factors of h convert derivatives with respect to physical x into derivatives with respect to dimensionless s. Dropping these factors would make the interpolant change shape when switching between metres and centimetres.

For positions $y_0 = 0, y_1 = 1$ m over one second with both endpoint velocities zero, $H = 3s^2 - 2s^3$ m. At half time the position is 0.5 m but the speed is 1.5 m/s, greater than the mean speed of 1 m/s. Imposing rest at both ends forces the intervening speed to exceed the mean somewhere. The piece alone leaves endpoint accelerations $+6$ and -6 m/s². A smooth join to a subsequent segment therefore needs more than identical endpoint positions.

Recover slopes when only values are measured

A cubic spline fits one cubic on every interval, requiring continuity of both first and second derivatives at interior nodes. Let $M_i = S''(x_i)$. Since a cubic has a linear second derivative, on panel i interpolate M_i and M_{i+1} linearly and integrate twice. Use the two measured endpoint values to fix the integration constants. On equal panels the resulting right-hand slope at the left endpoint is $(y_{i+1} - y_i)/h - h(2M_i + M_{i+1})/6$. The left-hand slope at the right endpoint is $(y_{i+1} - y_i)/h + h(M_i + 2M_{i+1})/6$.

Equating these expressions at an interior join gives

$$M_{i-1} + 4M_i + M_{i+1} = 6(y_{i-1} - 2y_i + y_{i+1})/h^2.$$

This tridiagonal system relates each undetermined curvature to the neighbouring data. Two additional end conditions are needed. A natural spline imposes zero curvature at both endpoints. A clamped spline uses given endpoint slopes instead. Choose those conditions from what the experiment actually controls.

Examine the influence of endpoint assumptions

Fit a natural spline at $x = 0, 1, 2$ to $y = 0, 1, 0$. The end curvatures are zero and the interior equation gives $4M_1 = 6(0 - 2 + 0) = -12$, hence $M_1 = -3$. On the first panel $S'' = -3x$. Two integrations give

$$S = -x^3/2 + ax + b.$$

The endpoint values set $b = 0, a = 1.5$, yielding $S(0.5) = 0.6875$ compared with 0.5 from the straight chord. Symmetry gives zero slope at the summit and a reflected cubic on the second interval.

Smoothness alone cannot establish the true midpoint value. A sensor measuring a sharp corner would support a piecewise linear description, and monotone data can require a slope limiter to avoid artificial

extrema in a cubic fit. Likewise, setting endpoint curvature to zero in a mechanical fit asserts a moment condition only under the appropriate beam model. Report the end conditions with the fitted curve, especially when derivatives near the ends are used as physical observables.

Use independent endpoint information as a check

Suppose the three displacement values in the example come with an independent measurement of the initial slope, $S'(0) = 1$ rather than 1.5. Imposing a natural left endpoint as well as this slope would overdetermine that end. One reasonable choice is a mixed spline, clamped on the left and natural on the right. For $h = 1$, the slope formula at the left reads $1 - (2M_0 + M_1)/6 = 1$, giving $2M_0 + M_1 = 0$. The interior equation gives $M_0 + 4M_1 = -12$, since $M_2 = 0$.

Solving yields $M_1 = -24/7$ and $M_0 = 12/7$. A measured slope at one end has changed curvatures inside the interval. That coupling across segments distinguishes a global spline fit from interpolating each segment independently with prescribed slopes at every node.

Why the natural spline minimises bending

Consider all twice continuously differentiable interpolants of the same tabulated values on an interval. For a natural cubic spline S, each panel is cubic, S' and S'' agree across knots, and $S'' = 0$ at both endpoints. Let a competing interpolant be $S + w$. The perturbation w vanishes at every knot including the endpoints. Expanding the squared curvature integral gives

$$\int (S'' + w'')^2 = \int (S'')^2 + 2\int S''w'' + \int (w'')^2.$$

To prove that S minimises this integral, account for every boundary term in the cross product.

Integrate the cross product by parts twice on each panel. The interior integral containing $S''''w$ vanishes because the panel is cubic. The term $S'''w$ vanishes at every panel endpoint since $w = 0$ at each knot. The remaining interface contributions $S''w'$ cancel between adjacent panels because both factors are continuous. At the outer boundaries $S'' = 0$. Thus the cross term vanishes exactly. Equality of the objectives requires $w'' = 0$ throughout; a globally affine perturbation vanishing at two distinct knots is zero. This proves uniqueness within the admissible class and explains the natural endpoint condition as part of a variational statement.

Calculate the bending energy on an uneven grid

Suppose the interpolation points are $(0, 0), (1, 2), (3, 2)$. The panel widths are one and two, and the secant slopes are two and zero. For a natural cubic write $M_i = S''(x_i)$, so

$$M_0 = M_2 = 0.$$

Continuity of first derivative at the central knot gives $2(1+2)M_1 = 6(0-2)$, hence $M_1 = -2$. The bending moment changes linearly from zero to −2 on the first panel and from −2 to zero on the second. The longer panel contains twice the squared curvature integral because it sweeps the same range of curvature over twice the distance.

Specifically the first contribution is

$$\int_0^1 (-2x)^2 dx = 4/3.$$

On the second panel $S'' = x - 3$, giving $\int_1^3 (x-3)^2 dx = 8/3$. Total energy is four. Integrating the first curvature twice and fitting its endpoints gives

$$S = 7x/3 - x^3/3;$$

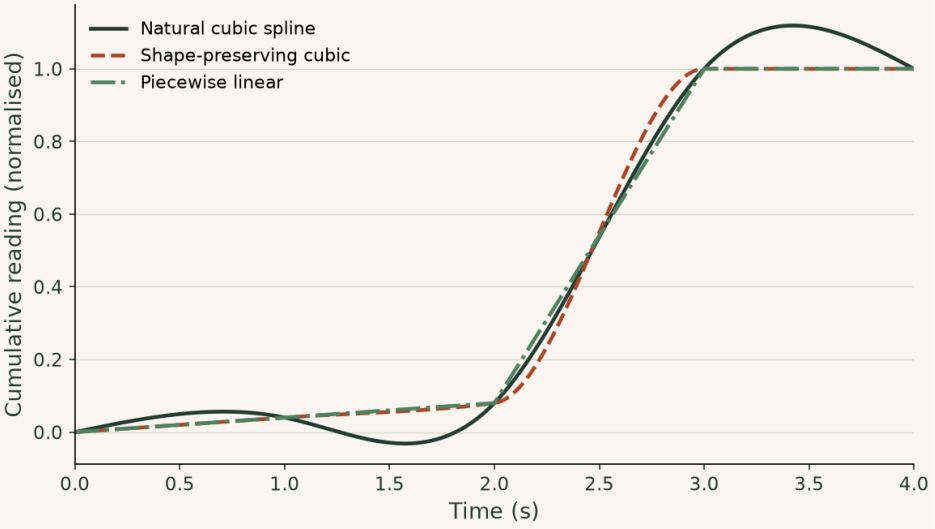
at the central knot the slope is $4/3$, neither of the adjacent straight-line secant slopes. On the second panel write $t = x - 1$. Then $S = 2 + 4t/3 - t^2 + t^3/6$, which returns to two at $t = 2$. The curve rises above the equal second and third data values in between.

Minimum bending therefore does not guarantee monotonicity between nearly flat measurements or an absence of overshoot. If flat endpoints describe a known plateau rather than merely two equal observations, a shape constraint changes the admissible functions. The preceding proof still explains why the natural interpolant won its own curvature competition, but gives no reason to prefer it after physical constraints have changed.

For these data the overshoot can even be located analytically. On the second panel the derivative vanishes when $4/3 - 2t + t^2/2 = 0$, giving the interior root $t = 2 - 2/\sqrt{3}$, about 0.845. A measured value taken near $x = 1.845$ would therefore help discriminate the minimum-bending curve from a physical plateau. Taking another measurement at either already constrained endpoint would do less to resolve that ambiguity.

Can an interpolated cumulative reading decrease?

A monotonically increasing quantity is sampled at integer times 0 through 4 with values 0, 0.04, 0.08, 1, 1. A natural spline enforces continuous second derivative and zero end curvature. A shape-preserving cubic constrains the node slopes.



All three curves interpolate every measurement. The natural spline overshoots the final plateau and must subsequently fall; differentiating it would imply a negative rate. Shape constraints matter when a fitted curve must obey a balance law.

▶ **Read values from the plot**

Change parameters · Can an interpolated cumulative reading decrease?

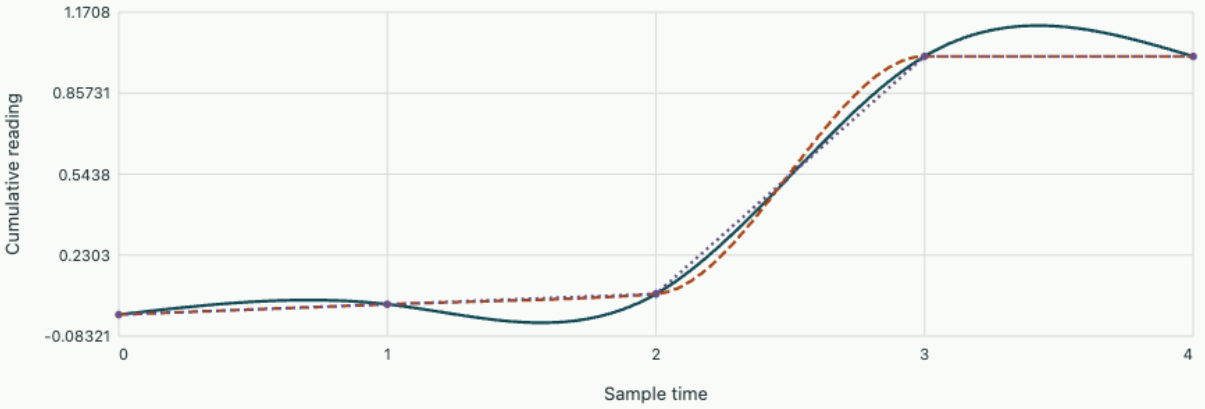
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Small initial increments **0.04**

Final plateau **1**



Both curves interpolate the same five monotone observations. Inspect the final plateau for an overshoot and a subsequent negative slope.

Chapter 4 · Composite quadrature and cancellation

Derive midpoint, trapezoidal and Simpson formulas, and use the sign of curvature to interpret a pair of numerical integral estimates.

FROM FIRST PRINCIPLES

Integrate the line joining two function values

To estimate accumulated area from sampled values, specify how the function is approximated between samples. On one panel of width h, connect $f(a)$ and $f(a+h)$ by a straight line. Its integral is its mean height times its width,

$$T_{\text{panel}} = \int_0^h [(1 - s/h)f(a) + (s/h)f(a+h)]ds = \frac{h}{2}[f(a) + f(a+h)].$$

Summing panels on a uniform mesh counts every interior value twice and each boundary value once. If $|f''| \le M$, the linear interpolation remainder at distance s from the left edge is bounded by $Ms(h-s)/2$. Integrating that bound and adding all panels gives

$$|I - T_h| \le \frac{(b-a)Mh^2}{12}.$$

Halving the step therefore quarters the upper bound for a fixed interval. The actual error can be smaller if curvature changes sign and panel errors cancel. Endpoint singularities or jumps can violate the smoothness assumption and require separate treatment.

DERIVATION AND APPLICATIONS

Explain why periodic trapezoidal quadrature can be unusually accurate

For a periodic smooth signal, the nominal second-order trapezoidal bound can substantially underestimate the observed accuracy. Consider one full period $[0, 2\pi]$ and N equally spaced samples $x_j = 2\pi j/N$. Since the first and final endpoints agree, their half weights combine. For one complex Fourier mode indexed by integer k, the sample mean is a geometric series,

$$\frac{1}{N}\sum_{j=0}^{N-1} e^{ikx_j} = \begin{cases} 1, & k \text{ is a multiple of } N, \\ 0, & \text{otherwise.} \end{cases}$$

For the second case the ratio is a nontrivial Nth root of unity; the numerator of the finite sum vanishes. The exact period integral of a nonconstant Fourier mode also vanishes. Expand a signal with absolutely summable Fourier coefficients so summation and sampling can be interchanged. The grid integrates exactly those nonconstant modes which do not alias into the constant mode,

$$f(x) = \sum_{k\in\mathbb{Z}} c_k e^{ikx},$$

$$T_N = 2\pi \sum_{\ell\in\mathbb{Z}} c_{\ell N},$$

$$I = 2\pi c_0.$$

The full quadrature error is therefore determined by the sampled multiples of N. For a signal with $|c_k| \le Cq^{|k|}$, $0 < q < 1$, summing the two geometric tails gives

$$|T_N - I| \le 4\pi C \frac{q^N}{1 - q^N}.$$

The conclusion depends on covering a complete period with the matched endpoints. A smooth nonperiodic signal generally acquires a discontinuity or derivative mismatch under periodic extension, changing the coefficient decay.

WORKED EXAMPLE

Detect the only troublesome harmonic

Integrate $f(x) = 2 + \cos(3x) + 0.1\cos(8x)$ with eight equally spaced points around a full period. The true integral is 4π. The third harmonic cancels across the grid, but $\cos(8x_j) = 1$ at every sample. Thus $T_8 = 4.2\pi$, an excess of 0.2π. With nine equally spaced points neither harmonic aliases into zero frequency, so the integral becomes exact.

SECOND APPLICATION

Select a point count from measured spectral decay

For a different periodic signal assume the coefficient envelope is verified with $C = 1$, $q = 1/2$. To keep absolute integral error below 0.010, require $4\pi/(2^N - 1) < 0.010$. Ten points give a bound 0.0123, while eleven give 0.00614. The latter is the smallest count certified by this envelope; the actual signal may converge faster through cancellations.

> **Start from this chapter.** Integrate each linear interpolant and add the cell contributions. Compare the resulting trapezoidal bias for a quadratic temperature law and for a linear law on identical nodes.
>
> OneSimulation laboratory · Measure what interpolation does to a heat integral

Measure an integral with finite samples

The energy delivered by a time-varying power input is the area under its history. If readings are available only at a fixed interval, integration must proceed from those samples or from additional physical assumptions. On one panel $[a, a + h]$, the midpoint rule uses $hf(a + h/2)$, while integration of the chord

between the endpoints gives $h[f(a) + f(a+h)]/2$. Both are exact for affine f. They may nonetheless give opposite errors on the same curved signal, which makes their comparison informative.

Expand about the midpoint m, using $s = x - m$. Odd powers of s integrate to zero on $[-h/2, h/2]$. The leading correction to $hf(m)$ is $f''(m)h^3/24$. Averaging the endpoint expansions instead gives $hf(m) + f''(m)h^3/8 + O(h^5)$. Subtracting the integral, the leading trapezoidal overestimate is $+f''(m)h^3/12$, while the midpoint estimate has leading error $-f''(m)h^3/24$. This expansion requires a smooth integrand. At a known abrupt switch in power, align a panel edge with the switch before applying it.

OneSimulation · Move one unresolved corner between quadrature points

A one-second power trace has the form $P(t) = a|t - c|$, with slope magnitude $a = 1$ W/s. Every panel away from the corner contains a straight line and contributes its exact energy. On the remaining panel put the corner a fraction θ along a width h. Splitting its area into two triangles and subtracting their sum from the midpoint approximation gives

$$Q_{\mathrm{mid}} - E = -ah^2 \min(\theta, 1-\theta)^2.$$

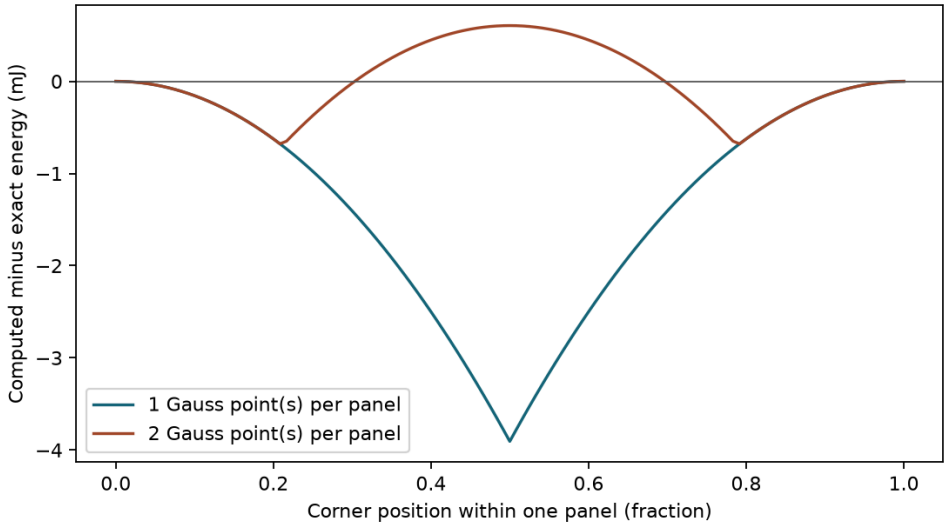
The sign follows directly from convexity. For eight panels the largest deficit is 3.90625 mJ. OneSimulation supplies the physical Gauss points and weights for each line element. The second sweep uses two Gauss points on exactly the same mesh. When the corner is central this rule overshoots by 0.6043 mJ. Its cubic exactness applies on a single smooth polynomial panel; the corner joins two different polynomials. Splitting the panel at the known corner makes either rule exact.

```
"""Slide a V-shaped power trace across an unchanged quadrature mesh."""
import matplotlib.pyplot as plt
import numpy as np
from onesim.cad.d1 import Line
from onesim.meshing import StructuredLineMesher
from onesim.numerics.fem.backends.spaces.line_lagrange import LineLagrangeSpace

cells = 8
line = Line(start=(0, 0, 0), end=(1, 0, 0))  # time interval, seconds
mesh = StructuredLineMesher(cells=cells).generate(line)
space = LineLagrangeSpace(mesh, order=1)
width = 1 / cells
fraction = np.linspace(0, 1, 121)
corner = (3 + fraction) * width  # position in fourth panel (s)
exact = (corner**2 + (1 - corner)**2) / 2  # J at 1 W/s

fig, ax = plt.subplots(figsize=(6.8, 3.8), constrained_layout=True)
for count, colour in [(1, "#176879"), (2, "#a4492b")]:
    energy = np.zeros_like(corner)
    for cell in range(cells):
        points, weights = space.quadrature(cell, order=count)
        power = np.abs(points[:, 0, None] - corner)  # W at 1 W/s
        energy += weights @ power
    error = energy - exact
    if count == 1:
        predicted = -width**2 * np.minimum(fraction, 1 - fraction)**2
        assert np.allclose(error, predicted)
    assert np.allclose(error[[0, -1]], 0)
```

```
    ax.plot(fraction, 1000 * error, color=colour, label=f"{count} Gauss point(s) per panel")
    print(f"{count} point(s): centre error {1000*error[len(error)//2]:+.4f} mJ")
ax.axhline(0, color="#555", lw=0.8)
ax.set(xlabel="Corner position within one panel (fraction)", ylabel="Computed minus exact energy
(mJ)")
ax.legend()
fig.savefig("approximation-and-quadrature.png", dpi=160)
```



Errors vanish when the corner lands at either panel boundary. Near the panel centre the midpoint rule underestimates the energy while the two-point rule overestimates it.

Download the Python example

Increase the order by cancelling a term

Cover a fixed interval with equal panels. There are $O(1/h)$ local errors of order h^3, giving global second-order accuracy for both rules. Now combine two parts midpoint and one part trapezoid. Their leading second-derivative terms cancel. On a panel of width H, the result is $H[f(a) + 4f(a + H/2) + f(a + H)]/6$, Simpson's rule. It integrates all cubics exactly: the quadratic error cancels and odd terms about the midpoint cancel by symmetry. For sufficiently smooth data the panel error is $O(H^5)$ and the composite error is $O(H^4)$.

Simpson's formula needs midpoint readings. With an even number of equal subintervals, join consecutive pairs and obtain the familiar alternating weights $1, 4, 2, 4, \ldots, 4, 1$. Applying those integer weights to unequally spaced observations loses the derivation. In that setting use correctly derived nonuniform weights or the trapezoidal rule panel by panel.

Check the signs with a quartic pulse

Let $P(t) = t^4$ W with t measured numerically in seconds over $[0, 1]$. The exact energy is $1/5$ J. With one whole panel, the midpoint estimate is $1/16 = 0.0625$ J and the trapezoidal estimate is 0.5 J. Convexity explains why the true integral falls between them. Simpson gives

$$(0 + 4/16 + 1)/6 = 5/24 = 0.2083333\,\mathrm{J}.$$

Its positive error of $1/120$ J is much smaller, though the quartic term prevents exactness.

Now split into two Simpson panels. The fourth derivative is constant, so halving panel width reduces the global error by exactly 16. The new estimate is $0.2 + (1/120)/16 = 0.20052083$ J. In a real measurement, noise can dominate such a small discretisation improvement. A common positive bias of 0.01 W integrates to an extra 0.01 J regardless of panel count. Distinguish uncertainty on the measurements from the discretisation of a known smooth curve. The rod experiment provides a direct check of the trapezoidal error sign and its reduction on mesh refinement. Measure the area under a sampled profile.

Build a Romberg table from nested panels

On a smooth interval the composite trapezoid often admits the even-power expansion $T(h) = I + c_2 h^2 + c_4 h^4 + \cdots$. Combining two grid estimates can remove the first term. Romberg integration repeats that cancellation on successive columns. Set $R_{j,0} = T(h_0/2^j)$. If column $m-1$ has leading error proportional to h^{2m}, halving h divides that term by 4^m. Solving the two leading error equations for the common limit gives

$$R_{j,m} = R_{j,m-1} + \frac{R_{j,m-1} - R_{j-1,m-1}}{4^m - 1}, \qquad j \ge m.$$

The second extrapolation therefore uses 15 in its denominator, where the first extrapolation used three. Existing readings can be reused: halving the step retains the old trapezoid with half its old weight and adds the new midpoint samples with the fine spacing. The jth row thus needs only the newly introduced points. Each extra cancellation requires enough smoothness for another term in the expansion. A jump, an endpoint singularity or noisy samples may destroy the expected pattern.

To integrate $1/(1+x)$ on $[0,1]$, the trapezoids with one, two and four panels give $3/4,\ 17/24,\ 1171/1680$. The two first extrapolations are $R_{1,1} = 25/36 \approx 0.6944444$ and $R_{2,1} = 1747/2520 \approx 0.6932540$. Using the second denominator yields

$$R_{2,2} = 4367/6300 \approx 0.6931746.$$

The exact answer is $\log 2 \approx 0.6931472$, so the error falls from about 0.003877 in the four-panel trapezoid to 0.0000274 after the second cancellation. Altogether only five distinct function locations were used. Compare columns of the same order before interpreting changes along the diagonal as a convergence measurement.

Practise on sheet 2 · Does three-level extrapolation integrate every smooth input exactly?. Further reading · University of California, Berkeley · Math 128A Numerical Analysis, lecture notes · Chapter 4, Numerical Differentiation and Integration, §4.5 Romberg Integration.

Handle a jump with one-sided values

Consider a heater that supplies 4 W until $t = 0.3$ s and 1 W from then to one second. The energy is

$$4(0.3) + 1(0.7) = 1.9\,\mathrm{J}.$$

A two-panel trapezoidal rule reading at 0, 0.5 and 1 s sees the values 4, 1 and 1. It predicts

$$0.5(4 + 1)/2 + 0.5(1 + 1)/2 = 1.75\,\mathrm{J}.$$

The piecewise linear representation invents a sloping transition lasting half a second.

If the switch time is known, split there and use the left and right limiting powers separately. Each constant segment then integrates exactly, including the correct jump without treating an isolated endpoint as carrying energy. If the switch time itself has uncertainty 0.01 s, the resulting energy uncertainty is 0.03 J because the power level changes by 3 W. Refining a numerical integration grid cannot resolve that timing uncertainty by itself.

Turn known endpoint slopes into a quadrature correction

For a uniform composite trapezoid rule on $[a, b]$, examine a single panel by expanding the integrand symmetrically around its midpoint m. The panel trapezoid value is $hf(m) + h^3 f''(m)/8 + O(h^5)$; integrating that same expansion gives $hf(m) + h^3 f''(m)/24 + O(h^5)$. The excess of the panel trapezoid is therefore $h^3 f''(m)/12 + O(h^5)$. Summing over the panels brings in the midpoint approximation to the integral of f''. Its leading continuous value is $f'(b) - f'(a)$, which telescopes across the entire interval.

Under a bounded fourth derivative this gives

$$T_h - I = h^2[f'(b) - f'(a)]/12 + O(h^4).$$

Subtracting that boundary term gives a fourth-order corrected rule with exactly the same function samples, provided the endpoint slopes are known accurately. This argument also explains why the uncorrected trapezoid can be unusually accurate for a smooth periodic integrand: matching endpoint slopes sets the leading correction to zero without a special cancellation at each interior point. Matching values alone does not suffice if the periodic extension has a slope jump.

Account for both integration error and slope error

Integrate $f(x) = x^4$ from zero to one with two equal panels. The three values are $0, 1/16, 1$, so

$$T_{1/2} = (1/2)(1/16 + 1/2) = 9/32 = 0.28125.$$

The true integral is 0.2 and the endpoint derivative difference is four. Subtracting $(1/2)^2(4)/12 = 1/12$ gives $19/96 \approx 0.1979167$. The corrected error is $-1/480$. Its sign is now opposite to the original error;

removing a leading term need not provide an upper bound for the remaining integral.

With four panels the trapezoid value becomes $113/512 \approx 0.2207031$. Subtracting $1/48$ gives about 0.1998698 and error $-1/7680$. The error magnitude falls by sixteen, exactly the fourth-order factor in this polynomial example. The fourth derivative is constant, so the next terms that could disturb this ratio vanish. For a generic smooth function one would expect only an asymptotic approach toward sixteen.

Now suppose a noisy slope measurement makes the right endpoint derivative too large by 0.12 and the left one is exact. On the four-panel grid the corrected integral is displaced by an additional $-h^2(0.12)/12 = -0.000625$, larger than the discretisation error that remained with exact slopes. On a refinement series with that same systematic slope error, the ultimate error becomes second order once more. The theoretical gain therefore depends on the source and accuracy of the endpoint derivatives. Comparing the two error terms before commissioning derivative measurements gives a concrete sampling decision.

A separate possibility is to estimate the endpoint derivatives from nearby samples. To preserve a fourth-order global correction, the error in their difference should be of order h^2 or smaller. A basic one-sided first difference normally has only first-order accuracy and can introduce an order-three term once multiplied by h^2. A suitable higher-order one-sided approximation may repair that deterministic issue, but measurement noise is amplified when nearby noisy readings are subtracted and divided by their short spacing.

Change parameters · Composite quadrature and cancellation

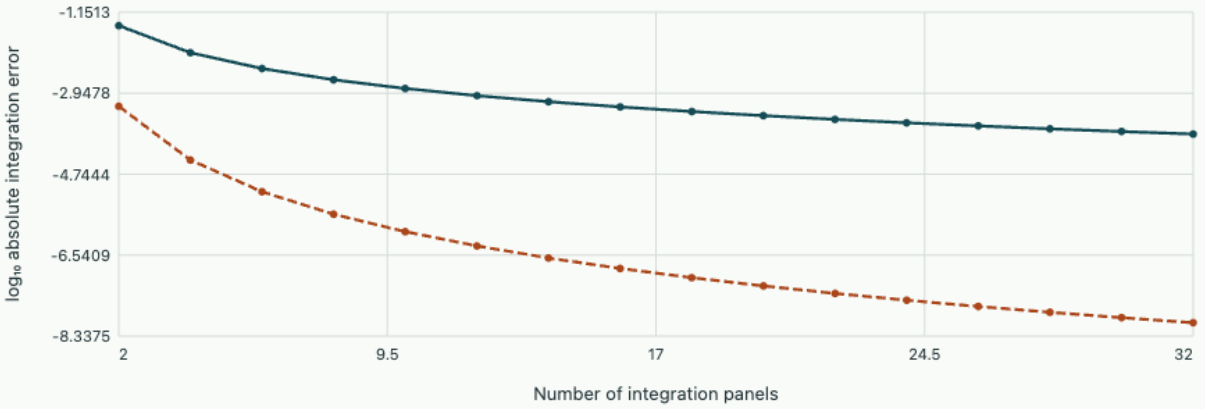
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Exponential rate **1**

Maximum panels **32**



Exact integral 1.7183. The even panel counts allow both methods on the same nodes; compare second and fourth order under refinement.

Study the full derivation

Integrate the oscillations analytically and sample the envelope · Function approximation and numerical integration

A sensor model asks for many Fourier integrals of a slowly changing envelope. Direct trapezoids can return a full-scale signal where the true integral nearly cancels. Derive a Filon rule from exact oscillatory moments and find where it needs a different local model.

Problem sheet 2 · Smooth pieces and integrated loads

Build a constrained trajectory, infer a spline curvature, and check quadrature by energy conservation over a measured interval.

Connect two resting positions

A carriage moves from 0 to 4 m in two seconds, starting and finishing at rest. Find the unique cubic trajectory satisfying these four endpoint data. Determine midpoint speed and the two endpoint accelerations. Can this cubic join a stationary segment with continuous acceleration?

▼ **Worked solution**

With $s = t/2$ the trajectory is $4(3s^2 - 2s^3)$. Differentiating with respect to time gives $v = 12s(1 - s)$ m/s, so $v(1\ \mathrm{s}) = 3$ m/s. Acceleration is $6(1 - 2s)$ m/s², giving $+6$ and -6 at the endpoints.

The velocity can join rest continuously. The stationary segment has zero acceleration, so this particular cubic cannot also give a continuous acceleration there. Enforcing both endpoint accelerations adds two constraints, motivating a higher-degree segment.

Explore the related calculation with adjustable inputs. Vary a slope or endpoint value and examine how an interpolant changes between observations. Map this exercise's interval to the unit interval before comparing the basis functions.

Identify what a natural endpoint buys

At $x = 0, 1, 2$ a rod has displacements $0, 2, 0$. Construct the first segment of the natural cubic spline and find its slope at 0 and its value at 0.5. Compare with the quadratic through all three data.

▼ **Worked solution**

The natural curvatures are

$$M_0 = M_2 = 0.$$

The interior equation gives $4M_1 = 6(0 - 4 + 0) = -24$, so $M_1 = -6$. Integrate $S'' = -6x$ and impose the first two values to get

$$S = 3x - x^3.$$

The initial slope is 3 and midpoint value 1.375.

The global quadratic is $4x - 2x^2$, with initial slope 4 and midpoint value 1.5. Both match every measured value. The difference is detectable in unmeasured derivatives and in the interior prediction because the quadratic has nonzero end curvature.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Recover a missing quadratic load

On a unit-width panel a power signal is quadratic. Its endpoint values are 2 W and 6 W; its measured energy over one second is 3 W·s. Infer its midpoint value, compare the midpoint and trapezoidal estimates, and determine the sign of curvature.

▼ **Worked solution**

Simpson integrates the quadratic exactly, so $3 = (2 + 4P_m + 6)/6$, giving $P_m = 2.5$ W. The midpoint estimate is 2.5 J and the trapezoidal estimate is 4 J. Their weighted combination $(2 \cdot 2.5 + 4)/3 = 3$ J reproduces the measured energy.

The signal midpoint lies below the endpoint chord value 4 W, so the curvature is positive. These conclusions use the stipulated quadratic law; the same three observations cannot uniquely determine an arbitrary power history.

Explore the related calculation with adjustable inputs. Compare errors from common-node composite rules; the explorer uses an exponential reference, while the polynomial identities in this exercise can be checked exactly.

Does three-level extrapolation integrate every smooth input exactly?

For $f(x) = x^6$ on $[0, 1]$, the trapezoids with one, two and four panels are $1/2,\ 33/128,\ 1421/8192$. Compute the two first-column corrections and the final diagonal value. Compare it with the exact integral and identify the first uncancelled error power.

▼ **Worked solution**

Removing the quadratic terms gives

$$R_{1,1} = (4(33/128) - 1/2)/3 = 17/96,$$
$$R_{2,1} = (4(1421/8192) - 33/128)/3 = 893/6144.$$

Removing the quartic term gives

$$(16(893/6144) - 17/96)/15 = 55/384.$$

The exact value is $1/7$, so the excess is $1/2688$. Both cancellations were exact algebraically, but the trapezoid expansion for this polynomial also contains an h^6 term. Three grids remove two independent error coefficients. Exact integration of every smooth function cannot follow from that finite information.

Explore the related calculation with adjustable inputs. Compare trapezoidal and extrapolated estimates as the power of the integrand changes; track when the smooth-error expansion needed for cancellation is valid.

Revisit chapter 4 · Build a Romberg table from nested panels.

Recover an integral using two biased readings

The four- and eight-interval integrals of the quadratic profile are 306.875 and 306.71875 K·m. Assume their leading error is Ch^2. Eliminate C to estimate the integral and predict the sixteen-interval output. Explain why this elimination is exact for this model on uniform grids but needs a refinement check for arbitrary data.

▼ **Work through the solution**

Write $I_h = I + Ch^2$ and $I_{h/2} = I + Ch^2/4$. Subtract the first relation from four times the second to get $I = (4I_{h/2} - I_h)/3 = 306.66667$ K·m. The eight-interval bias is 0.0520833 K·m. Dividing by four gives a sixteen-interval prediction of 306.67969 K·m. Constant curvature makes the trapezoidal error exactly proportional to h^2 for this quadratic. Curvature variation, lack of smoothness or measurement noise can spoil the assumed ratio.

Use the OneSimulation laboratory to check your answer

Chapter 5 · Gaussian quadrature and exactness

Choose quadrature nodes from moment conditions, prove the polynomial exactness of Gauss rules and map a reference rule to a finite element.

FROM FIRST PRINCIPLES

Choose quadrature nodes by matching moments

A quadrature formula can be designed by asking it to reproduce integrals of successive monomials. On $[-1, 1]$, take two nodes at $\pm a$ with equal weight w, exploiting reflection symmetry of the interval and uniform measure. This model integrates every odd polynomial exactly because its two sampled values cancel. Exactness for the constant and quadratic terms requires

$$2w = \int_{-1}^{1} 1 \, dx = 2,$$

$$2wa^2 = \int_{-1}^{1} x^2 \, dx = \frac{2}{3}.$$

The equations give $w = 1$ and $a = 1/\sqrt{3}$. By linearity the resulting two-point Gauss rule is exact for every cubic polynomial,

$$\int_{-1}^{1} p(x) \, dx = p(-1/\sqrt{3}) + p(1/\sqrt{3}), \qquad \deg p \le 3.$$

Its nodes lie inside the interval, making it useful when endpoints are expensive or shared information is unavailable. Higher-degree accuracy for an arbitrary function depends on the size of the unresolved part of its shape. Mapping another interval affinely requires scaling both the nodes and the weights by the appropriate coordinate transformation.

Background · Quadrature degree

$$Q[x^j] = \int_a^b x^j \, dx \quad (0 \le j \le d)$$

Exactness through degree d is an algebraic property. For a nonsmooth or poorly sampled integrand the observed rate can be much lower.

DERIVATION AND APPLICATIONS

Derive a two-point rule when one endpoint must be sampled

Suppose a measurement at one endpoint is already required for another purpose. How much polynomial exactness remains if the second sample can be placed freely? On $[0,1]$, fix the first node at zero and place the second at a. Denote their weights w_0, w_1. Exact integration of the first three monomials requires

$$w_0 + w_1 = 1,$$
$$w_1 a = \tfrac{1}{2},$$
$$w_1 a^2 = \tfrac{1}{3}.$$

Dividing the quadratic condition by the linear one determines $a = 2/3$; back-substitution fixes both weights. This is the two-node left Gauss–Radau rule. Its first missed monomial can be computed directly,

$$Q(f) = \tfrac{1}{4} f(0) + \tfrac{3}{4} f(2/3),$$
$$\int_0^1 x^3\,dx - Q(x^3) = \tfrac{1}{4} - \tfrac{2}{9} = \tfrac{1}{36}.$$

To see what determines the sign of the error for a more general function, interpolate with a quadratic that matches f at zero and matches both f and its derivative at the interior node. The quadrature sees the same values as that quadratic, which it integrates exactly. For a function with continuous third derivative the interpolation remainder has a factor $x(x-2/3)^2$, nonnegative over the interval. Integrating and applying the weighted mean value theorem gives

$$I - Q = \frac{f'''(\xi)}{6} \int_0^1 x(x-2/3)^2 dx = \frac{f'''(\xi)}{216}, \qquad 0 < \xi < 1.$$

A convexity assumption alone does not determine the sign here, since the error depends on the third derivative. Scaling the panel length by h multiplies the third-derivative error coefficient by h^4.

WORKED EXAMPLE

Integrate a profile on a physical interval

On $0 \le t \le 3$ let a flow rate be $q(t) = 2 + t^2$ in consistent units. The Radau nodes are 0 and 2, and the integral is

$$3[(1/4)2 + (3/4)6] = 15.$$

Direct integration gives $6 + 9 = 15$. The freely chosen interior station resolves quadratic curvature using only one new measurement.

SECOND APPLICATION

Bound the error for a rising exponential

On $[0,1]$ the rule gives $Q(e^x) = 0.25 + 0.75e^{2/3} = 1.71080$, compared with $e - 1 = 1.71828$. Its error 0.00748 is positive. Since $1 \le f''' \le e$, the derived bound brackets the error between $1/216 = 0.00463$ and $e/216 = 0.0126$, consistent with the direct result.

Three nodes sampling a Gaussian ensemble

An ensemble has normalised density $p(x) = e^{-x^2}/\sqrt{\pi}$ over the real line. An instrument records the mean of $\cos(\omega x)$. Its effective response changes from a slowly varying function to a rapidly oscillating one as ω increases. Build a quadrature rule that samples the ensemble only three times. The aim is to learn what exactness for polynomials implies for an observable with a variable frequency.

Exploit symmetry and put nodes at $-c, 0, c$, with equal outer weights w and central weight $1 - 2w$. Odd moments vanish automatically. Integration by parts using $p'(x) = -2xp(x)$ gives ensemble second and fourth moments $1/2$ and $3/4$. Matching them imposes $2wc^2 = 1/2$ and $2wc^4 = 3/4$. Divide the equations to find $c^2 = 3/2$. Substitution gives $w = 1/6$, hence central weight $2/3$. This proves exactness through degree five without needing a table of Gauss–Hermite nodes.

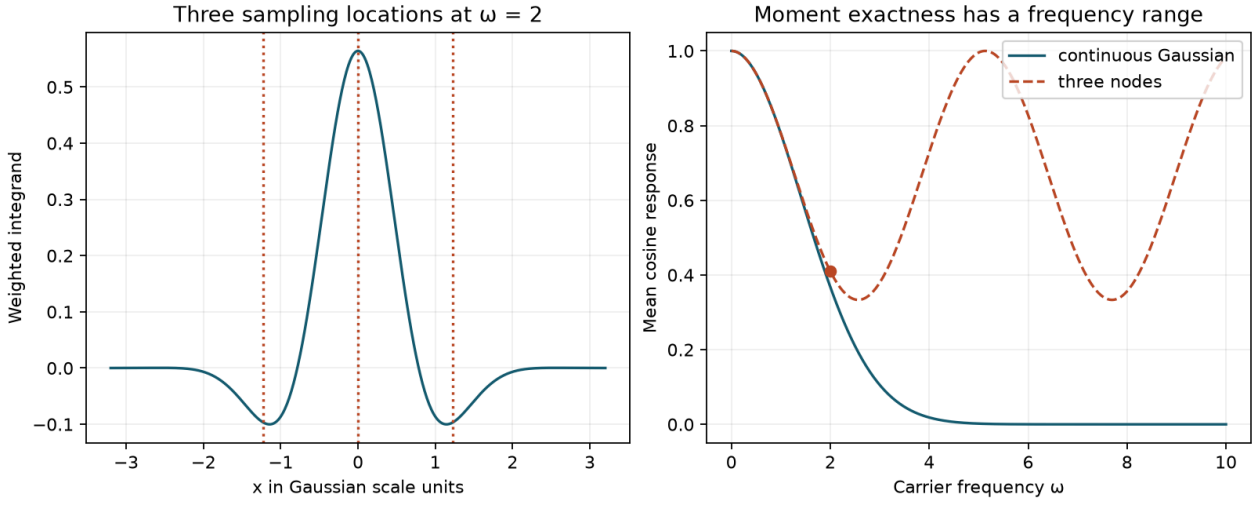
For the instrument response the three readings give

$$Q(\omega) = 2/3 + \cos(c\omega)/3.$$

For comparison derive the exact transform $I(\omega) = \int p(x)\cos(\omega x)\,dx$. Differentiating under the integral and once integrating by parts using the derivative of the Gaussian gives

$$I' = -\omega I/2,$$

with $I(0) = 1$. Hence $I = \exp(-\omega^2/4)$. This independent answer tests the numerical method in the transition from well-resolved response to nearly complete cancellation.



The rule reproduces the first terms of the response at small frequency. Three fixed point measurements cannot capture cancellation for all frequencies of the continuous ensemble.

Expand the cosine in even powers near zero. Its constant, quadratic and quartic terms are integrated exactly. The first possible error has order ω^6; the sixth ensemble moment is $15/8$, while the rule produces $9/8$. The difference begins as $Q - I = \omega^6/960 + O(\omega^8)$. This is a local frequency statement. At high frequency the exact response tends to zero, but the discrete sum repeatedly returns to one. Increasing the node count extends the useful frequency range, as the notebook's independent 64-node comparison illustrates.

A sensor that records only these three phase samples cannot distinguish a large carrier with phases differing by integer multiples of a full cycle from zero carrier at the sampling nodes. This is a property of the information supplied to the quadrature, not a failure of the root-finding procedure that produced the nodes. A useful extra observation would be at a coordinate whose phase differs for the two suspected frequencies; merely repeating a measurement at the central node cannot resolve them.

The Gaussian weight in this example has variance one half. If a laboratory reports a Gaussian standard deviation σ in physical units, first set the physical coordinate to $\sqrt{2}\sigma x$. The carrier in the dimensionless quadrature then becomes $\sqrt{2}\sigma\omega_{\mathrm{physical}}$. Scaling the nodes without scaling the frequency in the reference calculation would compare two different instruments.

Find a nonzero frequency at which all three individual measurements return exactly one and calculate the true ensemble mean there.

▼ **Work through the solution**

Choose $\omega = 2\pi/c$. Then $\omega c = 2\pi$, so both outer cosines and the central cosine equal one. The true mean is

$$\exp[-\pi^2/c^2] = \exp(-2\pi^2/3) \simeq 0.00139.$$

The error here is large despite exactness for five polynomial degrees; the integrand's scale has changed relative to the chosen sampling nodes.

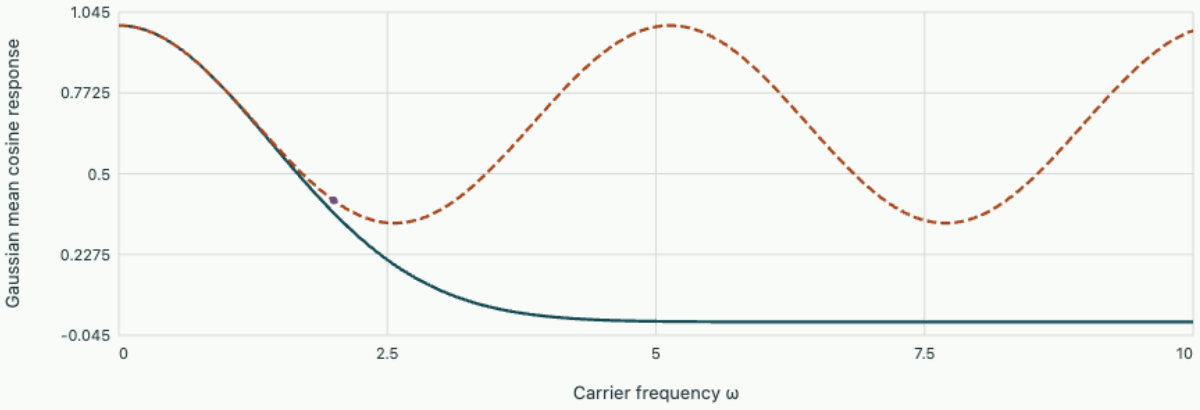
Change parameters · Explore this notebook example

Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Selected carrier frequency ω 2



At ω = 2, three nodes give 0.41003 versus exact 0.36788. Polynomial moment exactness through degree five does not control every oscillatory signal.

Open editable notebook Download .ipynb

Spend samples where they are most useful

When the integrand is a callable function, quadrature nodes need not coincide with pre-existing sensors. A rule with n distinct nodes and n weights has $2n$ scalar parameters. Can these integrate every polynomial through degree $2n-1$? Merely counting parameters does not prove existence. The Gauss construction supplies nodes for a positive weight function on an interval and establishes the exactness by orthogonality.

Start on $[-1,1]$ with a symmetric two-node rule of the form $w[f(-c)+f(c)]$. Testing the constant function gives $2w=2$, hence $w=1$. Symmetry automatically integrates the odd monomials x, x^3. Testing x^2 gives $2c^2=2/3$, or $c=1/\sqrt{3}$. These four independent moment conditions establish exactness on every cubic by linearity. Testing x^4 produces $2/9$, whereas its integral is $2/5$. The error appears at degree four exactly as the parameter count suggested.

Prove the general degree claim

Let P_n be the degree-n orthogonal polynomial for the prescribed positive integration weight $w(x)$. Use its n distinct interior zeros as nodes. Define the weights by integrating the Lagrange basis on these nodes, so every polynomial of degree at most $n-1$ is already integrated exactly. To treat any polynomial f of degree at most $2n-1$, divide it by P_n, giving $f = P_n q + r$ with $\deg q, \deg r \le n-1$.

The integral of $P_n q$ vanishes by orthogonality, and every quadrature evaluation of that product vanishes because the nodes are zeros of P_n. Both the integral and the rule therefore reduce to the same value for r. This proves exactness. There is a sharp obstruction to one extra degree: P_n^2 is nonnegative, positive except at finitely many points, and all its sampled values vanish. Its positive integral cannot be reproduced by this rule. Positivity of the individual Gauss weights follows by applying exactness to the square of each Lagrange basis polynomial.

Transfer the rule into a physical element

An element $[a,b]$ is obtained by the affine change $x = m + \ell\xi$, with midpoint $m = (a+b)/2$, half-length $\ell = (b-a)/2$, and $dx = \ell\, d\xi$. Both the node locations and their weights change. For $[2,4]$, the two physical nodes are $3 \pm 1/\sqrt{3}$, with unit physical weights. The integral of x^3 is

$$(4^4 - 2^4)/4 = 60.$$

Summing the nodal cubes gives

$$2 \cdot 3^3 + 6 \cdot 3(1/3) = 60.$$

A missing Jacobian would escape detection on this particular element where $\ell = 1$; test a second element of a different length.

In a linear finite element with constant density, each consistent mass entry integrates a product of two linear basis functions, hence a quadratic. Two Gauss points integrate it exactly. On a unit element the true mass matrix has diagonal $1/3$ and off-diagonal $1/6$. Using just the midpoint makes all four entries $1/4$. The alternating nodal vector $(1,-1)^T$ then has zero discrete mass although its interpolated field has a positive squared integral. This loss of a mode is a practical consequence of insufficient quadrature. With curved mappings or varying coefficients, determine the degree and smoothness of the entire transformed integrand before reusing a quadrature order.

Distinguish a load integral from a stiffness integral

On the same unit element, a constant axial stiffness integrand made from two derivatives of linear shape functions is itself constant. A single midpoint therefore reproduces the stiffness exactly, despite failing for the consistent mass. If $EA(x) = 1 + x^2$, the same midpoint now approximates its integral by 1.25, whereas the exact result is $4/3$. The difference is $1/12$ in the chosen units. A rule sufficient for one element contribution may thus be insufficient for another in the same analysis.

To select a quadrature order systematically, multiply out the shape functions or their derivatives, coefficient field and mapping Jacobian in the reference coordinate. For discontinuous coefficients, a single smooth polynomial degree may not exist on the entire element. Splitting at the interface, or integrating separate material subcells, then addresses the modelling geometry more directly than increasing the polynomial exactness of one unbroken-panel rule.

Retain both endpoints and derive the best three-point rule

Some instruments already provide readings at the ends of every panel, or adjacent elements need to communicate values at a shared boundary. If endpoints must be quadrature nodes, the unrestricted Gaussian degree $2n-1$ is generally unavailable. To see the obstruction on $[-1,1]$ with n nodes including both ends, form the polynomial $(1-x^2)\prod_{\text{interior}}(x-x_j)^2$. It vanishes at every evaluation point and is nonnegative throughout the interval. Its true integral is positive but every quadrature rule using just these values returns zero. Its degree is $2n-2$, so exactness can extend at most through $2n-3$.

For three symmetric nodes the interior point is zero. Symmetry forces equal endpoint weights a, with central weight b. Reproducing the constants and the square gives $2a+b=2$ and $2a=2/3$, so

$$a=1/3,\ b=4/3.$$

Odd powers integrate to zero by symmetry, hence the rule is exact through degree three, attaining its constrained upper bound. Under affine mapping this is the three-node Gauss–Lobatto rule, also known as the single-panel Simpson rule. Its weights and node placement now follow directly from a measurement constraint and a moment calculation.

Contrast endpoint inclusion with free node placement

Consider the local mass entry on one dimensionless interval $[-1,1]$ for a quadratic middle-node shape function $\phi(x)=1-x^2$ and uniform density. Its exact consistent mass contribution is

$$\int \phi^2 dx=\int (1-2x^2+x^4)dx=2-4/3+2/5=16/15.$$

The endpoint-constrained three-node rule sees values $0,1,0$, giving $4/3$, an overestimate of $4/15$, or 25 percent. The discrepancy arises solely from the quartic term. For a quadratic field, reproducing the field at all interpolation nodes does not imply correct integration of its square.

A three-point Gauss rule has free nodes at zero and $\pm\sqrt{3/5}$, with weights $8/9, 5/9, 5/9$. The outer samples have $\phi^2=(1-3/5)^2=4/25$. Their total contribution is $2(5/9)(4/25)=8/45$; adding $8/9$ recovers $16/15$. In a nodal element the endpoint-constrained rule makes off-diagonal mass entries vanish when basis functions are evaluated only at their own interpolation nodes. This diagonal mass approximation facilitates explicit updates even while changing the exact form.

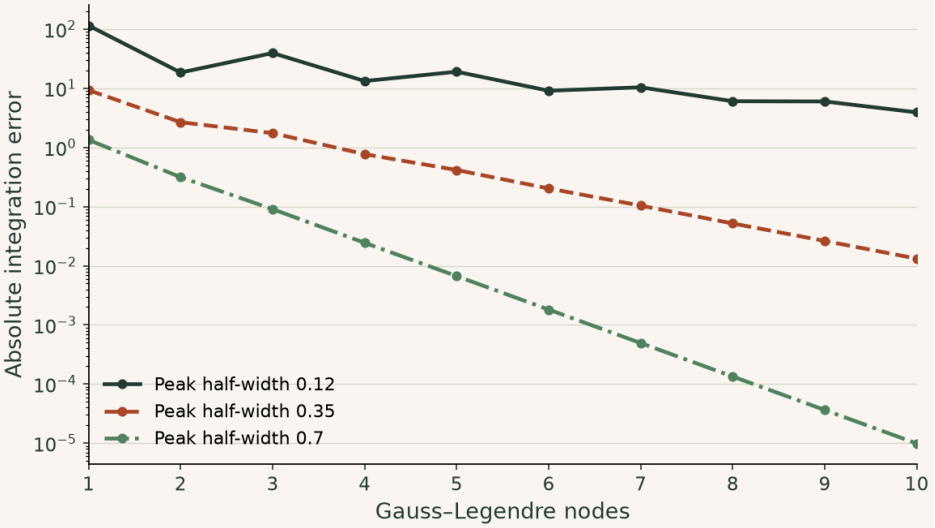
Selecting either rule is therefore a physical and numerical modelling choice. A modal frequency depends on the ratio of stiffness energy to mass energy, so changing one mass entry may shift computed frequencies even if a static patch test passes. If the density varies across the element, include its degree

or smoothness in the integrand before claiming exactness from polynomial degree counting. A discontinuity inside a panel defeats this counting argument until the integration interval is split.

As a check on assembly, sum all rows of the mass matrix. The shape functions form a partition of unity, so the total mass of a constant field is the integral of the density. The nodal diagonal rule reproduces this total for uniform density despite the incorrect middle diagonal entry just calculated. Testing the mass of a constant alone would therefore fail to expose the approximation affecting internal motion.

Does a high polynomial degree resolve a narrow peak?

Integrate $1/(x^2 + a^2)$ on $[-1, 1]$. The exact answer is $2\arctan(1/a)/a$. Use the ordinary Legendre nodes without centring a separate subinterval on the peak; vary only its half-width a.



Odd orders sample the centre directly and even orders straddle it. When the peak is narrow both can be substantially wrong despite high polynomial exactness. Choosing subintervals that resolve the physical width can help more than an extra node.

▶ **Read values from the plot**

Change parameters · Does a high polynomial degree resolve a narrow peak?

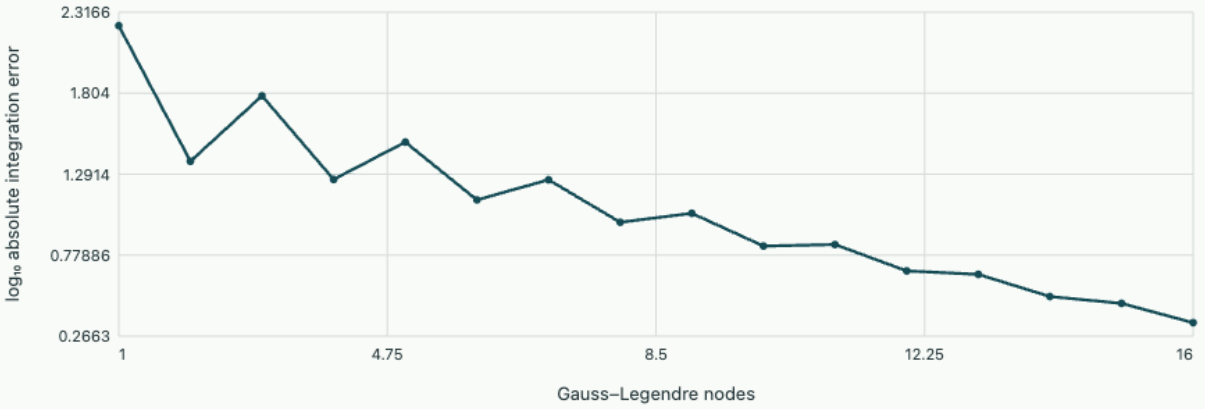
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Peak half-width a **0.1**

Highest Gauss order **16**



Unsplit interval [−1, 1]

Exact integral 29.423. Odd orders hit the peak centre; even orders straddle it. Decrease a before increasing the node budget.

Chapter 6 · Adaptive integration and error estimates

Construct a local error estimate from nested Simpson panels, allocate an absolute error budget and recognise signals that defeat sampling-based adaptivity.

FROM FIRST PRINCIPLES

Estimate a missing integral from two mesh sizes

Suppose a quadrature approximation on a fixed interval has the asymptotic form $Q_h = I + Ch^p + o(h^p)$, where $p > 0$ is known from the rule and smoothness assumptions. Evaluate the same rule after halving all panel widths. Its leading error is divided by 2^p, so eliminating C yields

$$Q_{h/2} - Q_h \simeq Ch^p(2^{-p} - 1),$$

$$I - Q_{h/2} \simeq \frac{Q_{h/2} - Q_h}{2^p - 1}.$$

The magnitude estimates the remaining fine-grid error. For composite Simpson quadrature $p = 4$ and the denominator is 15. An adaptive algorithm repeats this comparison on individual panels. Giving children absolute budgets that sum to their parent's budget makes local bounds additive,

$$|I - Q| \le \sum_j |I_j - Q_j| \le \sum_j \varepsilon_j.$$

The first inequality is exact, while the estimated local errors require the leading-power model to hold. Two rules can agree accidentally on unresolved structure, which makes independent refinement or smoothness information valuable for difficult integrands.

DERIVATION AND APPLICATIONS

Set a meaningful integration tolerance when contributions cancel

An adaptive integrator can refine its mesh until truncation error is small, yet the requested relative accuracy may be limited by uncertainty in the integrand. Consider an integrable real f on $[a, b]$ with pointwise error bounded by $|\delta f(x)| \le \eta |f(x)|$. This can represent a numerical evaluation error or a relative tolerance on measured input values. The resulting data-induced error in the exact integral is bounded directly by

$$|\delta I| = \left| \int_a^b \delta f(x)\, dx \right| \le \eta \int_a^b |f(x)|\, dx.$$

If arbitrary signs of perturbation are allowed, the bound is attained by taking $\delta f = \eta |f|$. Let I_+ and I_- be the respective positive magnitudes of the positive and negative contributions. For $I \neq 0$ the relative sensitivity to the data is therefore

$$\kappa_I = \frac{I_+ + I_-}{|I_+ - I_-|}, \qquad \frac{|\delta I|}{|I|} \le \eta \kappa_I.$$

It equals one for a one-sign integrand and can diverge as the net area tends to zero. If the algorithm has an independent absolute discretization error budget E_Q, add the separate errors before deciding whether the requested tolerance is meaningful,

$$|\widehat{I} - I| \le E_Q + \eta(I_+ + I_-).$$

Splitting at known sign changes makes estimates of the separate contribution magnitudes easier to audit. It does not undo cancellation in the underlying quantity. Correlated data errors can yield a less severe budget when their correlation is independently known; a uniform multiplicative calibration error, for example, scales the net integral itself.

WORKED EXAMPLE

Budget a nearly balanced energy flux

Heat-flow contributions over an interval are $I_+ = 1001\,\mathrm{J}$, $I_- = 999\,\mathrm{J}$. The net is $2\,\mathrm{J}$, and $\kappa_I = 1000$. With relative data uncertainty 10^{-4}, the data alone allow a $0.20\,\mathrm{J}$ discrepancy, or ten percent of the net value. Reducing the quadrature budget from 0.010 to $0.001\,\mathrm{J}$ cannot deliver one percent total accuracy.

SECOND APPLICATION

Select an absolute budget from a reporting target

A second experiment has $I_+ = 8.0$, $I_- = 3.0$ and $\eta = 0.002$. The data contribution is at most 0.022. Reporting net area 5.0 to one percent allows absolute error 0.050. A quadrature budget of 0.028 or smaller fits the remaining allowance. Two subinterval budgets of 0.014 each would satisfy the same total constraint.

Use two resolutions to estimate a missing integral

Suppose a composite rule on a smooth integrand has $Q_h = I + Ch^p + O(h^{p+2})$. Halving the uniform spacing gives

$$Q_{h/2} = I + Ch^p/2^p + \cdots.$$

Subtract the two approximations to eliminate I. The fine-grid error satisfies

$$I - Q_{h/2} \simeq (Q_{h/2} - Q_h)/(2^p - 1).$$

The sign of this estimate matters if correcting the numerical answer. Its magnitude supplies an error indicator. The relation assumes the same leading power operates on both grids and that the error exceeds the noise in the data.

For Simpson's rule $p = 4$. On a parent interval compare one Simpson panel S with two half-panels whose sum is S_2. The fine estimate of the local integral error is $|S_2 - S|/15$, and an extrapolated value is $S_2 + (S_2 - S)/15$. The comparison needs only two new function evaluations because the endpoints and original midpoint are reused. If the indicator is too large, split the interval and apply the same test to each child separately.

Build a global tolerance from local decisions

Let the total absolute tolerance be ϵ. Begin with the whole interval assigned that budget. When a panel splits, a simple allocation gives half its budget to each child. By induction the sum of all final panel budgets equals the original tolerance. If each local *actual* error lies within its budget, the triangle inequality bounds the global error by ϵ. In practice the indicators estimate those actual errors, so this last statement is conditional. Around a discontinuity the fourth-order expansion can fail, and tiny intervals can run into evaluation noise or floating-point spacing.

For signed integrals with strong cancellation, relative error alone is hard to interpret. A large positive area and nearly equal negative area may give a tiny sum that changes sign under rounding. Use an absolute tolerance meaningful for the physical output, possibly combined with a relative term based on a stable scale. Reusing shared endpoint values helps both reproducibility and cost, especially when each evaluation invokes a constitutive model or a simulation.

Inspect one accepted panel and one failure

On $[0, 1]$ use $f(x) = x^4$. One Simpson panel gives $S = 0.20833333$, and two panels give $S_2 = 0.20052083$. Their difference divided by 15 is -0.00052083. The corrected result is 0.2, exactly the known integral up to the printed rounding. Here the fourth derivative is constant and there are no higher powers in the integrand, so cancellation removes the entire leading error. An estimated error of 5.21×10^{-4} would pass a local tolerance of 10^{-3} while 10^{-4} would trigger another split.

Now consider a narrow triangular pulse entirely between 0.10 and 0.11 of the interval, with height one. Its exact area is 0.005. The five points initially used by adaptive Simpson are $0, 0.25, 0.5, 0.75, 1$. Every sampled ordinate is zero. The algorithm would accept zero with a zero indicator, even with a very small tolerance. This failure follows from lack of information, and it applies to any finite deterministic sample set when arbitrary narrow features are permitted. If event times or material interfaces are known, partition at them before sampling. If smoothness bounds are available, they supply an independent scale for an initial mesh. A small estimated local error should be interpreted together with those physical and mathematical assumptions.

Choose an independent comparison

A practical confirmation is to integrate the same signal using a different initial split and a different set of quadrature nodes. If a narrow pulse missed on one mesh is intersected on the other, their discrepancy reveals that the earlier indicator was misleading. Agreement between both estimates improves confidence but still cannot rule out an arbitrarily narrow feature absent from both sample sets.

When a physical bound $|f'| \le B$ is known, it supplies complementary information. Across an unsampled gap of width h whose endpoint values are both zero, a possible positive pulse is bounded by two straight ramps of slopes $+B$ and $-B$. Integrating the triangular envelope gives a maximum hidden area $Bh^2/4$. A credible slope bound therefore turns a statement about unsampled behaviour into a quantitative reason to reduce the largest gap.

Build an independent certificate for an accepted panel

Adaptive extrapolation estimates error by comparing sampled values. A rigorous alternative is available when the model supplies a valid bound M on the fourth derivative throughout an interval. For a single Simpson panel of full width H, the absolute error is bounded by $MH^5/2880$. One derivation uses the cubic Hermite interpolant through both endpoint values and the midpoint value with its derivative at the midpoint. Its remainder is $f^{(4)}(\xi)(x-a)(x-m)^2(x-b)/24$ for $m=(a+b)/2$, assuming sufficient smoothness. The factor multiplying the derivative has one sign throughout the interval.

Integrating that polynomial factor in absolute value gives $H^5/120$, and dividing by 24 gives the stated denominator 2880. Simpson quadrature integrates the cubic interpolant exactly and sees precisely the same three values as the original function. The midpoint derivative enters this argument as a mathematical device but never has to be measured. With independent derivative bounds on separate panels, summing their absolute bounds gives a global certificate without assuming that leading error terms cancel. Such bounds can be conservative, especially when worst allowed curvature occurs only over a tiny region.

A peak that agrees at every initial sample

On $[0,1]$ consider $f(x)=x^2(1-x)^2$, with zero endpoints and midpoint value $1/16$. Its single Simpson estimate is $1/24$. The exact integral is $1/3-1/2+1/5=1/30$, hence the overestimate is $1/120$. The fourth derivative is exactly 24. Substituting $M=24, H=1$ gives the same bound $1/120$, attained in this case because the fourth derivative has constant sign and magnitude. Dividing into two equal Simpson panels lowers the total bound to $2(1/2)^5/120=1/1920$.

A different candidate integrand makes the danger of relying on the sampled discrepancy explicit. Let $g(x)=1-\cos(8\pi x)$. The original three samples and the two additional quarter-point samples are all zero. Both the coarse and refined Simpson estimates are therefore zero, so the usual difference indicator vanishes. Yet its exact integral on the unit interval is one. An independently known derivative bound $M=(8\pi)^4$ gives a positive a priori certificate; the coarse computation cannot legitimately claim a small guaranteed error from those samples.

For equal panels the total fourth-derivative bound here is $M/(2880n^4)$ for n Simpson panels. To ensure an absolute error below 0.01 using only this information requires $n>(M/28.8)^{1/4}$, about 10.85, so eleven panels suffice. The guarantee costs more evaluations than an indicator that happens to stop immediately, but it addresses a different question. Before using it for noisy measured data, check whether a fourth derivative is physically bounded at all and account for sampling error in addition to quadrature.

On the oscillatory example an offset sample at $x=1/8$ gives value two, exposing the failure immediately. That single observation is powerful diagnostically, but its inclusion alone does not provide an accuracy

bound for every other admissible integrand. The derivative certificate applies to the whole class with the stated derivative limit and therefore answers a stronger question.

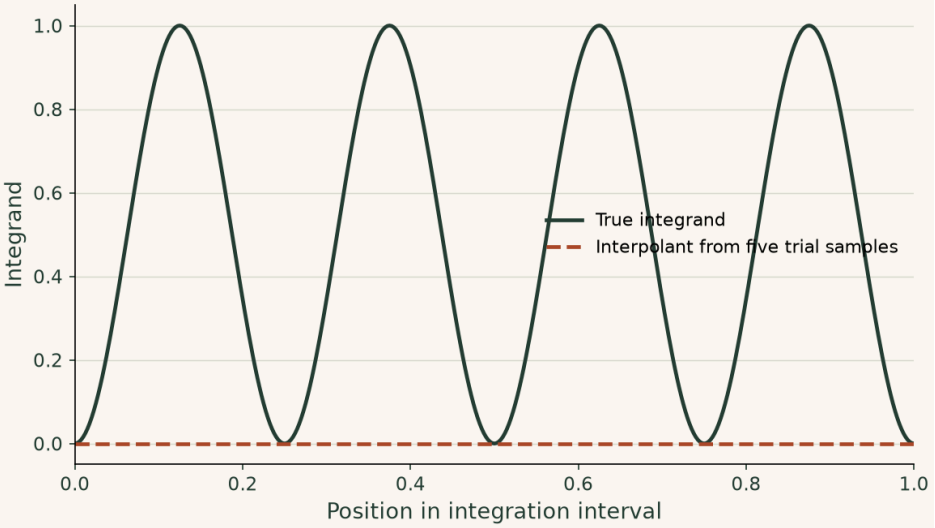
Allocate a certified global Simpson budget

Suppose two smooth regions have lengths L_i and known bounds M_i on the fourth derivative. Give region i exactly n_i equal Simpson panels. Its contribution to the absolute error bound is $M_i L_i^5/(2880 n_i^4)$. For a fixed total panel count the best continuous allocation minimises the sum of these two expressions. Differentiating with a multiplier gives $M_i L_i^5/n_i^5$ constant across regions, hence n_i proportional to $L_i M_i^{1/5}$. The exponent is one fifth because each extra panel reduces the fourth-power global error at a rate proportional to the inverse fifth power of the local panel count.

For two half intervals with $M_1 = 1, M_2 = 32$ and twelve panels in total, this gives four panels first and eight second. Cancelling their common factor $L^5/2880$, the uniform six–six design has bound $33/6^4$. The proposed four–eight design gives $1/4^4 + 32/8^4 = 3/256$, about 46 percent of the uniform bound. Both designs use the same number of panels, integrate through the common interface, and rely on the same two derivative certificates. This does not guarantee that the actual integral error is reduced by the same factor, because the worst allowed derivatives need not occur everywhere. It quantifies the improved guarantee before sampling begins, entirely independently of a potentially aliased adaptive indicator.

Construct a false convergence test for adaptive quadrature

Integrate $\sin^2(4\pi x)$ from zero to one using a trial Simpson panel and its two half-panels. All five evaluation points are multiples of 1/4. Predict the two estimates and the error inferred from their difference.



Both estimates and their difference vanish in exact arithmetic. The actual integral is one half, visible between the nodes. Changing the initial partition or adding independent samples tests the unresolved oscillations that a nested difference could not detect.

▶ **Read values from the plot**

Change parameters · Construct a false convergence test for adaptive quadrature

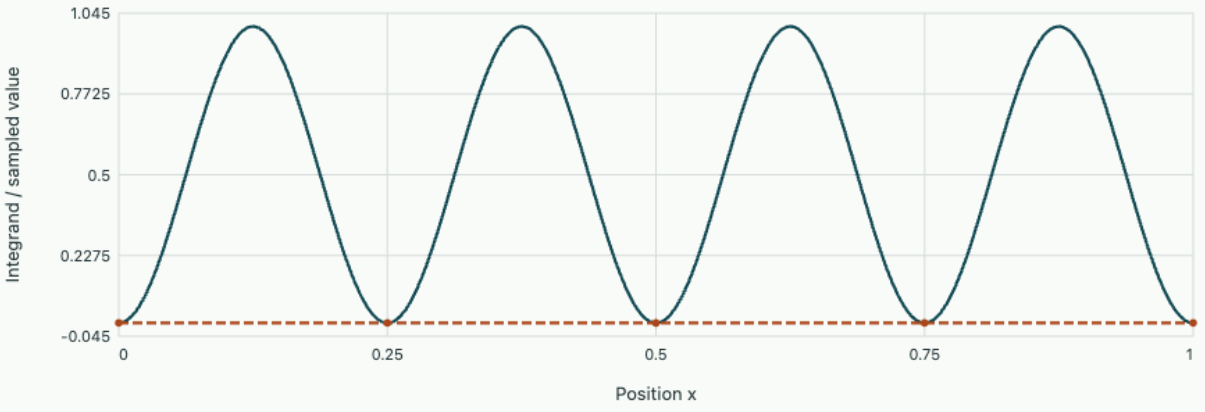
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Oscillations of sin² across interval **4**

Phase offset (radians) **0**



Continuous integrand Nested Simpson samples

Exact integral 1/2. Coarse Simpson 7.9987e-32, refined Simpson 7.9987e-32, difference-based estimate 0. Agreement can be false convergence when every node misses the oscillation.

Study the full derivation

Integrate the oscillations analytically and sample the envelope · Function approximation and numerical integration

A sensor model asks for many Fourier integrals of a slowly changing envelope. Direct trapezoids can return a full-scale signal where the true integral nearly cancels. Derive a Filon rule from exact oscillatory moments and find where it needs a different local model.

Problem sheet 3 · Accuracy with few evaluations

Derive mapped Gaussian weights, expose an underintegrated element and design a tolerance check for a sharply localised signal.

Map a rule and check its units

Use the two-node Gauss rule to integrate x^2 over $[1, 4]$. Give physical nodes and weights. Compare with the exact antiderivative, then explain what happens if the reference weights are used without the mapping factor.

▼ **Worked solution**

The midpoint is 2.5 and half-length is 1.5, so nodes are $2.5 \pm 1.5/\sqrt{3}$ and weights are both 1.5. The sum of squared nodal values is $2[2.5^2 + (1.5/\sqrt{3})^2] = 14$, giving integral 21. The antiderivative gives

$$(64 - 1)/3 = 21.$$

Using weights equal to one would report 14. When x is measured in metres, the weights carry metres and the integral carries m³. Ignoring the Jacobian is a scale error as well as a dimensional inconsistency.

Try this calculation. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

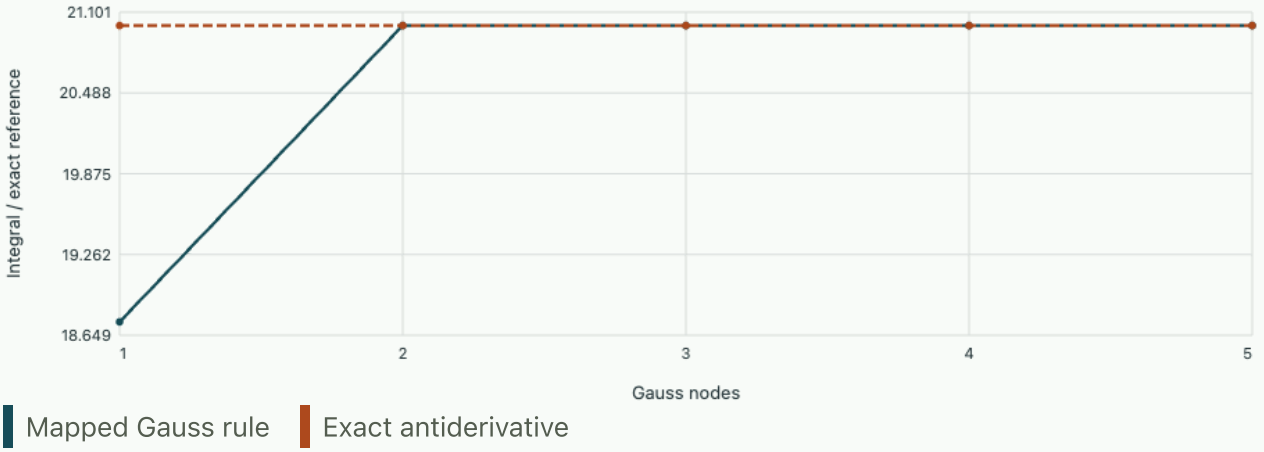
Open this adjustable calculation

Try the calculation

Interval left endpoint 1

Interval length 3

Monomial degree 2



Physical quadrature weights include interval half-length 1.5. For this degree-2 monomial, n nodes are exact once 2n−1 ≥ 2.

Measure mass of an alternating displacement

On $0 \le x \le 1$, linearly interpolate nodal values $(1, -1)$. Compute the exact squared norm, the one-point midpoint approximation and the two-point Gauss result. Explain which field escapes the one-point sampling.

▼ **Worked solution**

Linear interpolation gives $u = 1 - 2x$, so

$$\int_0^1 u^2 dx = 1 - 2 + 4/3 = 1/3.$$

The midpoint value is zero, giving zero numerical norm with one point. The two Gauss nodes are $1/2 \pm 1/(2\sqrt{3})$, where $u = \mp 1/\sqrt{3}$. Weights $1/2$ give $1/3$.

The field changes sign inside the element and its node lies exactly at the one-point sample. Treating zero discrete norm as proof of zero field would introduce a spurious unconstrained mode.

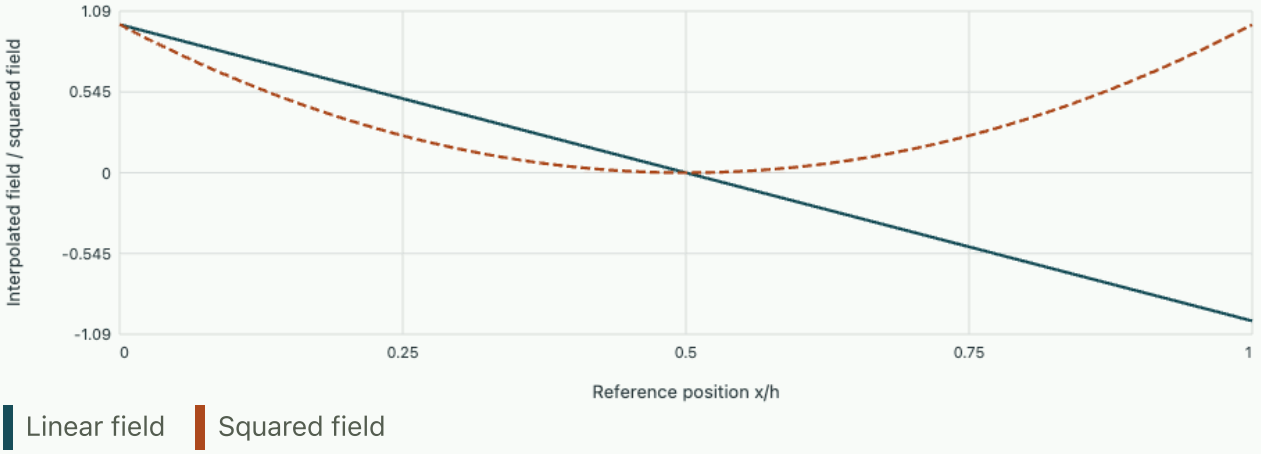
Try this calculation. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Open this adjustable calculation

Try the calculation

Right nodal value -1

Element length 1



Exact squared norm 0.33333, midpoint quadrature 0. Two-point Gauss recovers the exact quadratic integral. An antisymmetric nonzero field can be invisible at the midpoint.

Interpret an adaptive estimate

For a smooth interval, one Simpson panel gives 1.2000 and two half-panels give 1.1850. Estimate the signed correction, the corrected integral and whether an assigned tolerance of 0.0005 is met. Separately, a narrow pulse is known to lie between two sample positions. How would you modify the initial partition?

▼ **Worked solution**

The correction to the fine value is

$$(1.1850 - 1.2000)/15 = -0.0010.$$

The extrapolated estimate is 1.1840. The local error indicator 0.0010 exceeds the budget, so refine both halves and reuse the existing nodal evaluations.

For a pulse with known edges, introduce breakpoints at the edges and sample its interior. An indicator assembled from two rules that both miss the pulse can be zero even when the physical integral is finite.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Notebook practice · Three nodes sampling a Gaussian ensemble

Chapter 7 · Divided differences and derivative data

Build an interpolating polynomial one measurement at a time and incorporate a measured slope at a repeated node.

FROM FIRST PRINCIPLES

Append a node without changing earlier interpolation values

Suppose p_{k-1} already interpolates f at k distinct nodes $x_0, \dots, x_{k-1}$. A new polynomial can preserve all these values by adding a multiple of a product which vanishes at every old node. The new data at x_k then determine the only free coefficient,

$$p_k(x) = p_{k-1}(x) + a_k \prod_{j=0}^{k-1} (x - x_j),$$

$$a_k = \frac{f(x_k) - p_{k-1}(x_k)}{\prod_{j=0}^{k-1} (x_k - x_j)}.$$

This coefficient is the divided difference $f[x_0, \dots, x_k]$. For two nodes it is the ordinary secant slope. The higher coefficients can also be obtained by subtracting divided differences for overlapping node lists and dividing by the total interval width,

$$f[x_i, \dots, x_j] = \frac{f[x_{i+1}, \dots, x_j] - f[x_i, \dots, x_{j-1}]}{x_j - x_i}.$$

Each level has one more inverse power of the input unit than the preceding one. When nodes coincide, these formulas require limits; for example the first divided difference tends to $f'(x_i)$ when the function is differentiable.

DERIVATION AND APPLICATIONS

Convert a second divided difference into a curvature estimate

Three unequally spaced readings carry information about how fast a signal's slope changes. Let the times be $x_- = x_0 - a,\ x_0,\ x_+ = x_0 + b$, with $a, b > 0$. Divide each consecutive value difference by its own interval before subtracting the two slopes. The second divided difference is that slope change divided by the full interval $a + b$. Twice this value is the exact second derivative of the interpolating quadratic,

$$D_2 = \frac{2}{a + b}\left[\frac{f(x_+) - f(x_0)}{b} - \frac{f(x_0) - f(x_-)}{a}\right].$$

To infer its error as an estimate of the true curvature at x_0, expand the right and left secants about the central reading. Assuming four continuous derivatives locally, their leading terms are

$$s_+ = f' + \frac{b}{2} f'' + \frac{b^2}{6} f''' + O(b^3),$$

$$s_- = f' - \frac{a}{2} f'' + \frac{a^2}{6} f''' + O(a^3).$$

Subtract and divide by half the full span. For $h = \max(a, b)$ the result is

$$D_2 = f''(x_0) + \frac{b - a}{3} f'''(x_0) + O(h^2).$$

The leading error changes sign if the sampling geometry is reflected through x_0. A symmetric panel cancels it and produces a second-order estimate; an arbitrary asymmetric panel with both spacings scaled together generally gives first-order accuracy at the middle node. This geometric error can matter even if every reading is exact.

WORKED EXAMPLE

Detect curvature bias from asymmetric time stamps

Take $f(t) = t^3$ and times $-1, 0, 2$. The left secant is 1 and the right secant is 4. The formula gives $D_2 = 2(4 - 1)/3 = 2$, whereas $f''(0) = 0$. The predicted error is $(2 - 1)6/3 = 2$, exactly since the function is cubic. For the same data the estimator measures the exact curvature at $t = 1/3$.

SECOND APPLICATION

Propagate bounded noise into an acceleration measurement

In a different experiment the times are $-0.10, 0, +0.10\,\mathrm{s}$, with independent unknown position errors bounded in magnitude by $0.001\,\mathrm{m}$. The second-derivative weights are $100, -200, 100\,\mathrm{s}^{-2}$. Thus the worst noise error is

$$(100 + 200 + 100)(0.001) = 0.400\,\mathrm{m/s^2}.$$

Doubling the spacing quarters this noise bound while increasing smooth-signal truncation error, a practical trade-off even on a symmetric grid.

Start from this chapter. A quadratic has constant second divided difference. Use that coefficient to explain why the four- and eight-cell runs show a factor of four at midpoints.

OneSimulation laboratory · Measure what interpolation does to a heat integral

Keep the previous fit

A new measurement arrives after a polynomial has been fitted to $x_0, \dots, x_{k-1}$. To preserve those measurements, add a multiple of $\omega_k(x) = \prod_{j=0}^{k-1}(x - x_j)$. At each old node the correction vanishes. At a

new distinct node x_k its coefficient is fixed by $c_k = (f(x_k) - p_{k-1}(x_k))/\omega_k(x_k)$. This gives a triangular construction. Once a coefficient has been determined it remains in every subsequent polynomial provided the original order is retained.

The coefficients can also be obtained using a table. Put $f[x_i] = f(x_i)$ and subtract adjacent entries of the previous order before dividing by the distance between the two outer nodes. The subtraction cancels shared lower-order information. Algebraically the leading coefficient of the polynomial through $x_i, \dots, x_{i+k}$ is

$$f[x_i, \dots, x_{i+k}] = \frac{f[x_{i+1}, \dots, x_{i+k}] - f[x_i, \dots, x_{i+k-1}]}{x_{i+k} - x_i}.$$

This follows by combining the two lower-degree interpolants with factors $x - x_i$ and $x - x_{i+k}$. Their difference divided by the outer spacing agrees with every tabulated value and has the required leading coefficient.

Three temperatures and a test point

Suppose a temperature sensor reports 2, 5 and 17 degrees at 0, 1 and 3 hours. First divided differences are 3 and 6 degrees per hour. The second difference is $(6 - 3)/3 = 1$ degree per hour squared. Hence $p(x) = 2 + 3x + x(x - 1)$. At hour 2 this gives 10 degrees; the original straight line would give 8. The correction shows how strongly the additional observation changes that prediction. Neither interpolant by itself certifies that the temperature really did reach its predicted value between readings.

Nested evaluation uses $2 + x[3 + (x - 1)]$ here. Each new layer adds one multiplication and one addition. Reordering the nodes changes the intermediate table, but uniqueness ensures the same final polynomial in exact arithmetic. Strongly unequal spacings can magnify noise in higher differences and make numerical results sensitive to the order.

Let two nodes coalesce

If a second node approaches the first, the corresponding difference quotient tends to $f'(x_0)$ for differentiable f. An independently available derivative therefore supplies the diagonal entry directly. For a value 1 and slope 2 at zero and a value 4 at one, the required quadratic is $1 + 2x + cx^2$; matching the final value gives $c = 1$. Its prediction at one half is 2.25. Perturb the initial slope by 0.1 while keeping endpoints fixed. The correction becomes $0.1x(1 - x)$, attaining 0.025 at the midpoint. Endpoint agreement alone would miss that uncertainty.

Rescale before comparing coefficients

The coefficients have different units at successive orders. In the temperature example, expressing time in minutes divides the first divided differences by 60 and the second difference by 60^2. Their smaller printed magnitudes do not mean the curvature has become less significant. Evaluating the second term at the same physical instant multiplies by two correspondingly larger node differences and recovers exactly the

original contribution. Scaling the node interval near unit size can make intermediate arithmetic more convenient; interpret a coefficient together with the product of distances it multiplies.

Differentiate a local interpolant on an uneven grid

For samples at $x_0 - a, x_0, x_0 + b$, where $a, b > 0$, the derivative of their quadratic interpolant gives a three-point estimate of $f'(x_0)$. Write it as $w_- f(x_0 - a) + w_0 f(x_0) + w_+ f(x_0 + b)$. Reproducing the derivatives of a constant, the centred coordinate and its square gives three equations: the weights must sum to zero, $-aw_- + bw_+ = 1$, and $a^2 w_- + b^2 w_+ = 0$. Solving yields

$$w_- = -b/[a(a+b)],\ w_0 = (b-a)/(ab),\ w_+ = a/[b(a+b)].$$

These are the same weights one obtains by differentiating Newton's divided-difference polynomial at the central node.

A Taylor expansion exposes the leading cubic contribution. On $f(x) = (x - x_0)^3$, the estimate is $-a^3 w_- + b^3 w_+ = ab$, although the exact derivative is zero. Thus the leading smooth-data bias is $abf'''(x_0)/6$, with higher-order terms controlled by higher derivatives and panel geometry. Moving the nearest observation closer can reduce this deterministic bias while greatly increasing the magnitude of the weights. The calculation reveals the same conflict between localisation and noise that occurs when derivatives are inferred from independently measured values.

Estimate a heating rate from unsymmetrical sampling times

Choose $x_0 = 0$, $a = 1$ minute and $b = 2$ minutes. The weights are $-2/3,\ 1/2,\ 1/6$ per minute. Let the exact temperature over this short interval be $T(t) = 10 + 3t + t^3/10$ degrees. The sampled readings are

$$T(-1) = 6.9,\ T(0) = 10,\ T(2) = 16.8.$$

Applying the formula gives $-4.6 + 5 + 2.8 = 3.2$ degrees per minute. The analytical derivative at zero is three. Its 0.2 bias equals $abT'''/6 = (1)(2)(0.6)/6$, exactly because the temperature has no higher powers.

If each sample has independent standard deviation 0.12 degrees, the rate standard deviation is $0.12\sqrt{4/9 + 1/4 + 1/36} = 0.12\sqrt{26}/6 \approx 0.102$ degrees per minute. Halving both spacings reduces the cubic bias by four to 0.05, but doubles every derivative weight, increasing standard deviation to approximately 0.204. For the original grid, combining bias squared and variance gives mean squared error about 0.0504; for the refined grid it is about 0.0441. A further halving would make random error dominate still more strongly.

The weights sum to zero, so a truly common temperature offset cancels, while time-dependent drift need not. Also, using a symmetric central-difference formula on these asymmetric sample times would estimate a slope centred elsewhere and introduce a contribution from any quadratic term. Keeping the acquisition times in the coefficient equations is essential when observations are irregular because of missed triggers or variable sampling intervals.

A further check compares sampling the same temperatures with a fit that assumes the instrument clock is correct only to a tenth of a minute. At the central point the true local slope is three degrees per minute, so a timing error of 0.1 minute can by itself move a value by approximately 0.3 degrees. This is larger than the assumed independent temperature standard deviation. Errors in the sample locations would enter the divided differences through both numerator and denominator, and the simple fixed-time variance calculation would no longer describe the experiment.

Change parameters · Divided differences and derivative data

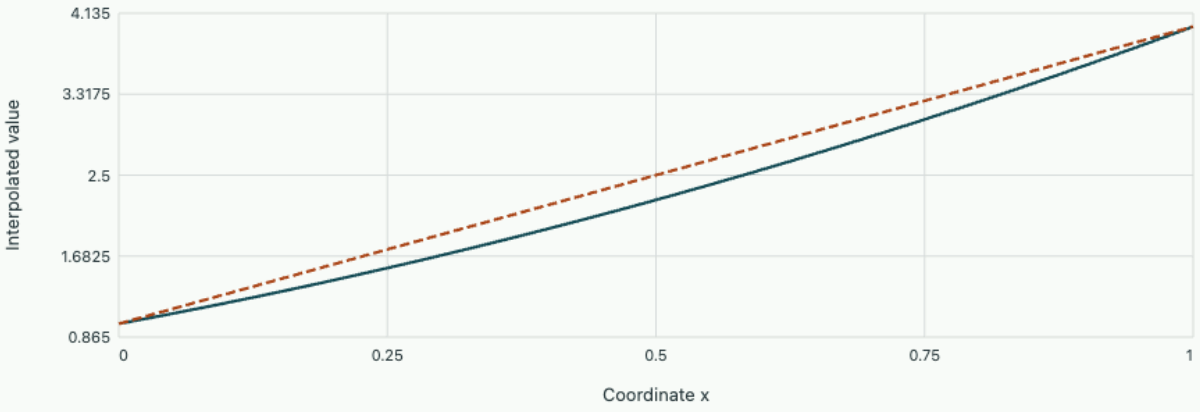
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Initial slope **2**

Final observed value **4**



Midpoint value 2.25. Changing the initial slope by Δ changes the interior by Δx(1 – x), with no endpoint change.

Study the full derivation

Approximating a resonance with a rational function · Function approximation and numerical integration

A smooth response on a measured interval can be governed by a nearby pole. Derive a low-order Padé fit, construct a barycentric rational approximant from samples and check whether a predicted pole survives independent data.

Chapter 8 · Uniform approximation and alternation

Certify a best uniform polynomial by finding alternating errors and compare its guarantee with interpolation at fixed nodes.

FROM FIRST PRINCIPLES

Why alternating errors can certify a uniform approximation

To approximate f on a closed interval, measure the largest error in absolute value. A degree-n minimax polynomial is one for which no other polynomial of the same degree has a smaller largest error. Suppose a candidate p has absolute residual bounded by $E > 0$, and at $n + 2$ ordered points the residual attains alternating extrema,

$$f(x_i) - p(x_i) = (-1)^i E, \qquad x_0 < x_1 < \cdots < x_{n+1}.$$

If another polynomial q had error strictly below E, the difference $q - p$ would have to be positive at each positive residual extremum and negative at each negative one, or vice versa depending on the first sign. Continuity then forces

$$\#\{\text{distinct zeros of } q - p\} \geq n + 1.$$

A nonzero polynomial of degree at most n cannot have that many roots. Thus the candidate is optimal once its global error bound and the alternating points are verified. Values sampled only at trial extrema do not certify that a still larger error is absent between them.

DERIVATION AND APPLICATIONS

Remove the leading monomial with the least possible uniform error

Suppose a polynomial fit can represent every degree up to $n - 1$, but the signal contains one additional monomial x^n. On $[-1, 1]$ the residual $x^n - p_{n-1}(x)$ is a monic polynomial of degree n, regardless of the coefficients chosen for the fit. The Chebyshev polynomials provide a monic residual with particularly small uniform magnitude. Define $T_n(\cos\theta) = \cos(n\theta)$. The cosine addition formula gives

$$\begin{aligned} T_0 &= 1, \\ T_1 &= x, \\ T_{n+1} &= 2xT_n - T_{n-1}. \end{aligned}$$

By induction the coefficient of x^n is 2^{n-1} for $n \geq 1$. Multiplying T_n by the reciprocal of that coefficient gives a monic polynomial bounded by 2^{1-n} on the entire interval. Subtract it from the original target so that the highest powers cancel,

$$p_{n-1}^{*}(x) = x^n - 2^{1-n}T_n(x),$$
$$\|x^n - p_{n-1}^{*}\|_\infty = 2^{1-n}.$$

At the $n+1$ points $x_j = \cos(j\pi/n)$ the chosen residual equals $2^{1-n}(-1)^j$. These alternating extrema certify optimality: a strictly better approximant would force its difference from p_{n-1}^{*} to change sign at least n times, exceeding the degree $n-1$. For a leading term ax^n on a physical interval with midpoint m and half-width h, the affine variable $t = (x-m)/h$ shows that the optimal error scales as

$$E_{\min} = |a|h^n 2^{1-n}.$$

Lower-degree terms can be carried into the approximant exactly. This result applies to uniform absolute error; changing to a mean-square criterion or weighting locations differently changes the optimization.

WORKED EXAMPLE

Compress a cubic response to a quadratic model

For $n = 3$, $T_3 = 4x^3 - 3x$, hence the best degree-at-most-two approximation of x^3 on $[-1, 1]$ is $3x/4$. Its residual has magnitude $1/4$ at $-1, -1/2, 1/2, 1$, with alternating signs. The unused quadratic coefficient follows from the odd symmetry of the target and interval.

SECOND APPLICATION

Scale the guarantee to a physical interval

Approximate $2x^4 + 3x$ by a cubic on $[0, 4]$. The half-width is 2, so the least possible uniform error is

$$2(2^4)/8 = 4.$$

An explicit optimum is $2x^4 + 3x - 4T_4((x-2)/2)$; its fourth-degree term cancels. Halving the interval width reduces the optimal error to $4/16 = 0.25$ for the same polynomial target.

An error budget that applies everywhere

A calibration curve may have a specification requiring its largest absolute deviation anywhere in an interval to stay below a threshold. Assign a polynomial p of degree at most m and consider $E(p) = \max_{a \le x \le b} |f(x) - p(x)|$ for a continuous target. Interpolating selected measurement locations fixes the polynomial there. It need not allocate the remaining error efficiently. Shifting an interpolant by a constant sometimes halves its largest deviation at once.

Suppose $m+2$ ordered locations attain the same error magnitude with alternating signs. Any polynomial with strictly smaller maximum error must move up at each positive extreme and down at each negative extreme. Its difference from the candidate then has at least one zero between each consecutive pair. This requires $m+1$ zeros in a polynomial of degree at most m, a contradiction. Those extreme points thus certify optimality globally once the candidate's error is bounded over the entire interval.

Find the best line by hand

Approximate x^2 on $[-1, 1]$ by $a + bx$. Averaging that line with its reflection gives the constant a. For each x, the averaged error is the mean of the errors at x and $-x$, so the uniform norm cannot increase. For constants the target covers the whole range from zero to one. The midpoint $a = 1/2$ balances its two extreme distances. At $-1, 0, 1$, the errors are $+1/2, -1/2, +1/2$. These three points certify the optimum among all lines, including nonsymmetric ones.

The endpoint interpolant is the constant 1. Its error is zero at both ends and reaches magnitude 1 at zero. Lowering the line sacrifices exact agreement with the endpoints in exchange for a factor of two in the worst case. If the endpoint values encode compulsory boundary conditions, such a shift is inadmissible and the optimisation problem must include those constraints.

Exchange trial extrema

For an unknown target, guess $m + 2$ extreme locations and solve the linear system $f(x_i) - p(x_i) = (-1)^i E$ for the coefficients and E. Search the full interval, including its boundaries, for the largest deviation. When it exceeds the trial level, exchange a support point while retaining sign alternation and repeat. For smooth targets, the zeros of $f' - p'$ supply candidate interior extrema. A coarse plotting grid alone cannot certify the resulting maximum.

For a constant approximation to e^x on $[0, 1]$, monotonicity identifies the endpoints immediately. Equalising $a - 1 = e - a$ gives $a = (1 + e)/2$ and error $(e - 1)/2$, approximately 0.85914. Replacing the upper endpoint by a larger value raises both quantities. The constant optimum depends on the range, providing a simple independent check on an implementation of the exchange procedure.

A lower bound from three measurements

Suppose measurements at $-1, 0, 1$ give the exact values $1, 0, 1$ but the interior shape is otherwise unknown. Any line still has maximum error at least one half on those three measurements: averaging its two endpoint predictions recovers its value at zero, so it cannot be within less than one half of both endpoint targets and the central target. The lower bound survives with sparse data. The matching upper bound requires information between the measurements, supplied in the earlier example by the analytic formula x^2.

Exploit symmetry to approximate a kink by a quadratic

The absolute-value function on $[-1, 1]$ has a derivative jump at zero. To fit it uniformly with a quadratic p, average $p(x)$ with $p(-x)$. The maximum error cannot increase because the target is even and the error of the average is bounded by the average of the two original absolute errors. The optimisation can therefore start with an even polynomial $a + bx^2$ and examine only $0 \le x \le 1$. Its signed error is

$$e(x) = x - a - bx^2.$$

A plausible minimax candidate has equal negative errors at both ends of this half interval and a positive interior maximum.

Equal endpoint errors require $-a = 1 - a - b$, hence $b = 1$. Differentiation then puts the positive maximum at $x = 1/2$, where the error is $1/4 - a$. Equalising its magnitude with the endpoint magnitude a gives $a = 1/8$. Directly completing the square gives $e(x) = 1/8 - (x - 1/2)^2$ on the half interval. Its entire range is $[-1/8, 1/8]$, so this construction has established the upper error bound at every point, including locations between the supposed extrema.

Certify the optimum and compare a smoother error measure

On the original interval the errors at $-1, -1/2, 0, 1/2, 1$ are $-1/8, +1/8, -1/8, +1/8, -1/8$. If any quadratic had a strictly smaller uniform error, its difference from the constructed polynomial would have to alternate sign at those five locations. Continuity would require at least four distinct real zeros. A nonzero quadratic cannot have four real zeros, so the proposed approximation is globally optimal even among polynomials with a linear term. The unusually long alternation is a consequence of symmetry and the central kink; four alternating points would already suffice for degree two.

For comparison, the quadratic interpolating the two ends and the central kink is x^2. Its error on the positive half is $x - x^2$, ranging from zero to $1/4$. Shifting it upward by one eighth cuts the maximum error in half. Now suppose the instrument is used predominantly near the centre, where its true ideal value is zero. The minimax polynomial introduces a nonzero prediction of $1/8$ there even though the original interpolant was exact. The preference for uniform accuracy amounts to a decision about the acceptable errors at those specific operating points.

A differentiability-based search that checks only stationary points can also fail in this problem. The target has a corner at zero, where its left and right derivatives differ. The error maximum in absolute value there is determined by the one-sided behaviour. For piecewise-smooth targets, include breakpoints and endpoints along with the roots of the error derivative on every smooth segment. Omitting a kink can produce an apparent equioscillation pattern with an incorrect reported bound.

For uniform approximation at a tolerance of one tenth, the optimum error of one eighth proves that no quadratic can meet the requirement on this interval. A failure by one fitted polynomial alone would not establish this. The alternation certificate tells an engineer when changing the polynomial class or splitting at the kink is mathematically necessary.

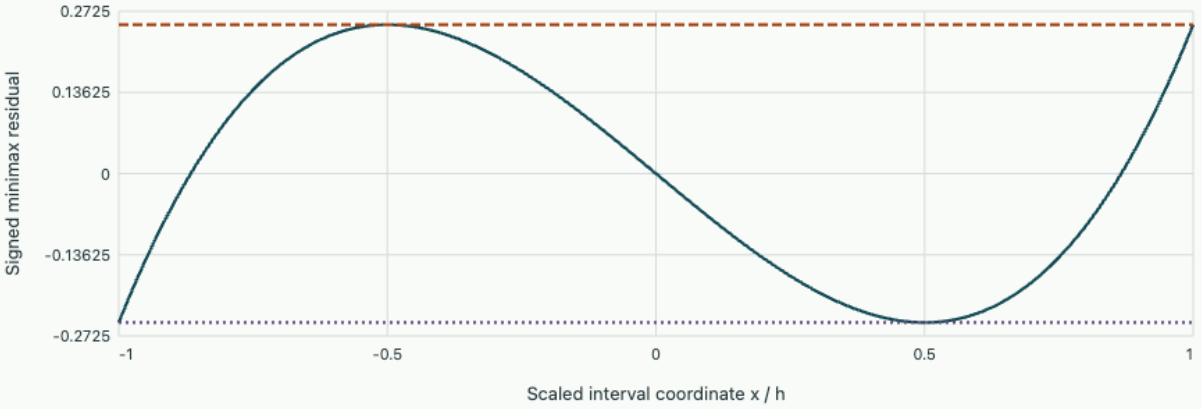
Change parameters · Explore the worked calculation

Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Target monomial degree	3
Interval half-width	1
Leading monomial coefficient	1



Scaled Chebyshev residual | Positive uniform bound | Negative uniform bound

For leading coefficient 1 and degree 3 on an interval of half-width h, the best degree-2 error is 0.25. Lower-degree target terms and an interval translation do not change this bound. There are 4 alternating extreme points.

Problem sheet 4 · Data and uniform error

Update measurements in Newton form and certify a uniform approximation by following the signs of its error.

Add a fourth observation

At times 0, 1 and 2 a displacement takes values 1, 2 and 5. Construct its Newton polynomial. Append a value of 13 at time 3 and evaluate the revised prediction at 1.5.

▼ **Worked solution**

The first differences are 1 and 3; the second is 1. The quadratic is

$$1 + x + x(x-1) = 1 + x^2.$$

At 3 its value is 10.

The new coefficient is

$$(13 - 10)/(3 \cdot 2 \cdot 1) = 1/2.$$

The correction at 1.5 is $0.5(1.5)(0.5)(-0.5) = -0.1875$. The revised value is

$$3.25 - 0.1875 = 3.0625.$$

All three original nodes annul the added term.

Explore the related calculation with adjustable inputs. Vary a slope or endpoint value and examine how an interpolant changes between observations. Map this exercise's interval to the unit interval before comparing the basis functions.

Combine a slope with two values

Find a quadratic with $p(1) = 4$, $p'(1) = -1$ and $p(3) = 6$. If the last measurement changes by 0.2, find the change of the midpoint prediction while preserving the first two conditions.

▼ **Worked solution**

Write $p = 4 - (x-1) + c(x-1)^2$. At 3 this gives $2 + 4c = 6$ and hence $c = 1$. The midpoint prediction at 2 is 4.

The revised curvature coefficient rises by $0.2/4 = 0.05$, so $p(2)$ rises by 0.05. The effect at the repeated node and its derivative both vanish, as the original conditions demand.

Explore the related calculation with adjustable inputs. Vary a slope or endpoint value and examine how an interpolant changes between observations. Map this exercise's interval to the unit interval before comparing the basis functions.

Certify a line

Approximate x^2 on $[0, 2]$ by a line parallel to its endpoint interpolant. Choose the vertical shift minimising the uniform error. Show that allowing a different slope cannot improve the result.

▼ **Worked solution**

The interpolant is $2x$. Its error $x^2 - 2x = (x-1)^2 - 1$ lies between -1 and 0. Dropping the line by 0.5 centres the error range, giving $p = 2x - 0.5$.

At 0, 1 and 2 the errors are $+0.5, -0.5, +0.5$. Any strictly better line would differ from this one by a polynomial changing sign twice. A nonzero linear polynomial cannot have those two distinct roots. Thus 0.5 is globally optimal.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Chapter 9 · Fourier projection and sampled frequencies

Recover periodic modes by projection and identify signals which become indistinguishable on a uniform sample grid.

FROM FIRST PRINCIPLES

Recover a Fourier coefficient by orthogonality

For a square-integrable signal of period 2π, represent its main oscillations by harmonics e^{ikx}. To choose the best coefficients in the integrated squared-error sense, set each residual inner product with an included harmonic to zero. Integration of a complex exponential gives the orthogonality relation

$$\frac{1}{2\pi}\int_0^{2\pi} e^{ikx}e^{-imx}\,dx = \begin{cases}1, & k=m,\\ 0, & k\neq m.\end{cases}$$

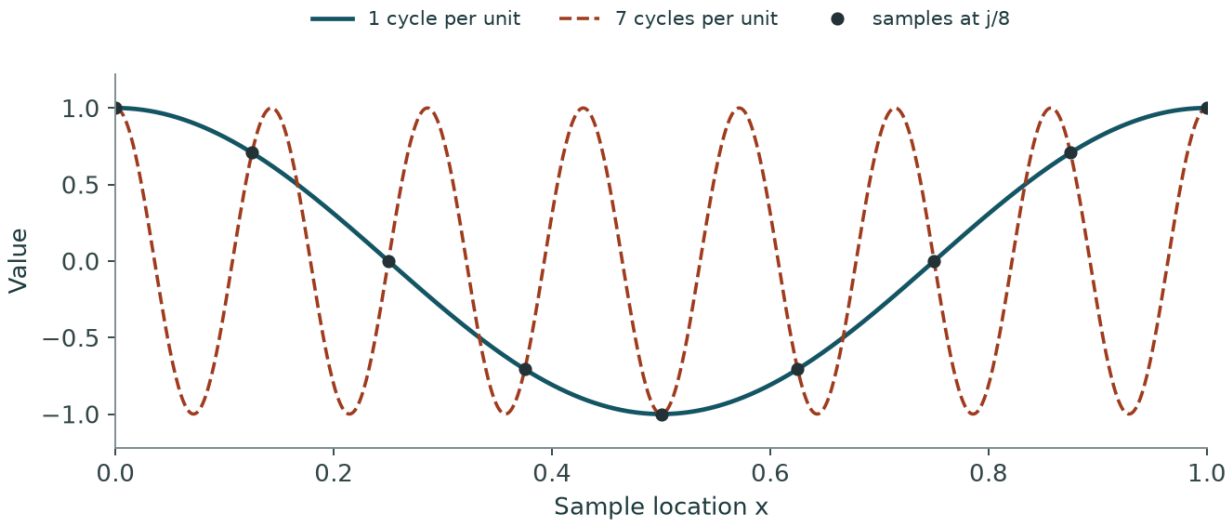
Consequently projection onto mode m gives

$$c_m = (2\pi)^{-1}\int f(x)e^{-imx}\,dx.$$

On a uniform grid $x_j = 2\pi j/N$, replacing this integral by the sample mean defines the normalized discrete coefficient,

$$\hat{c}_m = \frac{1}{N}\sum_{j=0}^{N-1} f(x_j)e^{-imx_j}.$$

The finite geometric series now vanishes unless two frequency labels differ by a multiple of N. Such labels have identical sampled values, so the sampled coefficients can combine several underlying continuous frequencies. A physical sampling rate and restrictions on the input bandwidth are needed to associate a discrete coefficient with a unique physical harmonic.



Two different functions can supply exactly the same input data

At x = j/8 the two sampled phases are 2πj/8 and 2πj − 2πj/8. Their cosines coincide at every grid point. A reconstruction using only these samples cannot choose the underlying frequency. Take one more sample at x = 1/16 and predict which two values the competing models give.

DERIVATION AND APPLICATIONS

Predict spectral leakage for a tone between Fourier bins

A frequency may lie well below the sampling limit and still distribute energy across many discrete Fourier bins. Consider N equally spaced samples of a single complex tone of unit amplitude with ν cycles across the record. Using a rectangular measurement window, $x_j = e^{2\pi i \nu j/N}$ for $j = 0, \dots, N-1$. Under the normalization giving unit amplitude for a tone exactly on its bin, the coefficient in bin k is

$$X_k = \frac{1}{N} \sum_{j=0}^{N-1} e^{2\pi i (\nu - k) j/N}.$$

Set $\delta = \nu - k$. If δ is not a multiple of N, sum the geometric progression and factor a phase out of both numerator and denominator. This gives

$$X_k = e^{i\pi\delta(N-1)/N} \frac{\sin(\pi\delta)}{N \sin(\pi\delta/N)}.$$

At a bin-centred frequency the corresponding limiting coefficient is one and all other bins vanish. At a fractional bin the tails are a consequence of observing a finite record. The total squared spectral weight remains fixed by orthogonality of the discrete basis,

$$\sum_{k=0}^{N-1} |X_k|^2 = \frac{1}{N} \sum_{j=0}^{N-1} |x_j|^2 = 1.$$

Amplitudes in selected bins can therefore be smaller without any loss of input energy. The formula assumes a complex tone. A real cosine contains two complex tones at opposite frequencies, whose coefficients must be added with phase before comparing amplitudes near overlapping lobes.

WORKED EXAMPLE

Quantify a half-bin mismatch

Take $N = 8$ and $\nu = 2.5$ cycles. Bins 2 and 3 each have magnitude $1/[8\sin(\pi/16)] = 0.641$. Together they contain $2(0.641)^2 \simeq 0.821$ of the tone's spectral energy. The remaining fraction is distributed among the other six bins even though the input contains only a single frequency.

SECOND APPLICATION

Distinguish grid spacing from intrinsic frequency content

A recorder sampling at $800\,\mathrm{Hz}$ takes 80 samples of a $125\,\mathrm{Hz}$ complex tone. Its bin spacing is $10\,\mathrm{Hz}$, placing the tone halfway between bins 12 and 13; each magnitude is

$$1/[80\sin(\pi/160)] = 0.637.$$

Extending the actual record to 160 samples gives $5\,\mathrm{Hz}$ spacing and exactly 25 cycles. The normalized magnitude in bin 25 becomes one when the tone remains coherent across the longer measurement.

Orthogonality on a period

For a square-integrable 2π-periodic signal define $\langle f, g\rangle = (2\pi)^{-1}\int_0^{2\pi} f(x)\overline{g(x)}\,dx$. The exponentials e^{ikx} are orthonormal for integer k. Integrating their product gives one when the frequencies coincide and zero otherwise because a nonconstant exponential completes an integer number of cycles. Minimising the squared mean discrepancy over modes $|k| \le m$ therefore gives

$$c_k = \langle f, e^{ikx}\rangle.$$

The residual is orthogonal to every retained mode.

Orthogonality also splits energy. The squared norm of the original signal equals the sum of the retained squared coefficient magnitudes plus the remaining squared error. This identity gives the truncation error exactly for a finite trigonometric series. For a general square-integrable periodic signal the projections converge in the mean-square norm. That statement concerns an integral; a narrow region of large pointwise error can contribute little to it.

Background · Why does a noninteger cycle spread power over several bins?

Project $e^{i\nu x}$ on the integer harmonic e^{ikx} over a single observed interval of length 2π. Integrating the exponential exactly gives

$$c_k = e^{i\pi(\nu-k)}\,\frac{\sin[\pi(\nu-k)]}{\pi(\nu-k)}.$$

For an integer ν, the limit at $k=\nu$ is one and every other bin receives zero. A tone with $\nu = 3.5$ cycles has equal coefficient magnitudes $2/\pi$ at bins 3 and 4. A longer record at the same sampling rate narrows the physical frequency interval between bins. Applying a taper changes the leakage shape because it changes the projection window. Adding zeros after the existing record samples that same windowed spectrum more densely, without extending the observed duration.

A resolvable frequency and an invisible one

Take $f(x) = 2 + \cos(2x) + 0.4\sin(5x)$. Projecting onto frequencies through three retains the constant and cosine. The remaining mean-square error is $0.4^2/2 = 0.08$, so its root mean square is about 0.28284. Now sample at $x_j = 2\pi j/8$. At those points $5x_j$ differs from $-3x_j$ by $2\pi j$. Thus the data are equally compatible with $2 + \cos(2x) - 0.4\sin(3x)$. A discrete fit through frequency three can have zero sample residual yet disagree between nodes.

At $x = \pi/8$, the two sine contributions are $0.4\sin(5\pi/8)$ and $-0.4\sin(3\pi/8)$. Their difference is about 0.7391. This unsampled point exposes the false low-frequency reconstruction. More generally the finite sum of $e^{2\pi i(k-l)j/N}$ vanishes unless $k-l$ is divisible by N. This geometric-series argument explains frequency equivalence on the grid and also supplies the discrete orthogonality behind a Fourier transform.

Factor the discrete Fourier transform into short transforms

The discrete Fourier transform (DFT) maps N equally spaced samples x_n to $X_k = \sum_{n=0}^{N-1} x_n W_N^{kn}$, with $W_N = e^{-2\pi i/N}$. In this convention the forward transform has no normalization. The discrete Fourier coefficients are X_k/N. Computing every sum directly uses order N^2 operations. There is redundant work because the sample weights are powers of the same root of unity.

For even N, separate $n = 2j$ and $n = 2j+1$. Since $W_N^2 = W_{N/2}$, the two partial sums are the length-$N/2$ transforms E_k, O_k of the even and odd samples. Increasing k by $N/2$ leaves both short transforms unchanged and reverses the sign of W_N^k. A single product therefore supplies two entries:

$$X_k = E_k + W_N^k O_k,$$

$$X_{k+N/2} = E_k - W_N^k O_k, \qquad 0 \le k < N/2.$$

For the four samples $(1,2,3,4)$, the even subsequence $(1,3)$ gives $(4,-2)$, while the odd subsequence $(2,4)$ gives $(6,-2)$. The multipliers are 1 and $-i$. Combining the two pairs produces

$(10, -2+2i, -2, -2-2i)$ in frequency order. The zero-frequency term recovers the sum of the samples. Conjugate symmetry holds because the samples are real. The Nyquist entry is real as it must be.

For $N = 2^m$, repeat the split down to length one. Each of the m stages contains $N/2$ paired combinations, often called butterflies. The work scales as $N \log_2 N$. The FFT computes exactly the same mathematical transform as the direct formula, up to rounding. An aliased high frequency still occupies the same discrete bin. Padding an already sampled record can interpolate its displayed spectrum but cannot supply independent measurements at the missing times.

Practise on sheet 5 · Find the zeros in an eight-sample spectrum without eight long sums. Further reading · MIT · 2.161 Signal Processing, Continuous and Discrete · The Fast Fourier Transform · The radix-2 FFT with decimation in time.

Smoothness and endpoint matching

Integrating a Fourier coefficient by parts divides by its frequency. The boundary term vanishes for a smooth periodic continuation. Repeating the operation shows faster coefficient decay when successive derivatives also match at the endpoints and remain sufficiently regular. A mismatch when a finite record is periodically repeated creates a jump. Near a jump partial sums develop oscillations, and adding modes confines them to a narrower region while a peak overshoot persists.

Start a numerical study by refining the sample grid with a fixed retained bandwidth. Then hold that refined grid and enlarge the bandwidth. The first experiment probes whether the coefficient estimates are polluted by unresolved modes. The second probes how much physical detail a deliberately truncated representation omits. A single refinement that changes both quantities makes these explanations difficult to distinguish.

The special frequency at half the sample rate

With eight equally spaced samples the mode $\cos(4x)$ gives $1, -1, 1, -1, \ldots$, whereas $\sin(4x)$ gives zero at every point. A phase shift of a sinusoid at precisely this frequency therefore changes its observed amplitude and can hide the entire oscillation. This explains the special treatment of the highest coefficient in an even-length real Fourier transform. Ten samples over the same period distinguish both sine and cosine components at frequency four because their sample values are now linearly independent.

Average partial Fourier sums to obtain a positive kernel

Ordinary Fourier projection cuts sharply from a retained coefficient to a discarded one. Instead average the first $m+1$ symmetric partial sums, counting the constant-only projection as the first. Mode k then appears in $m+1-|k|$ of those sums for $|k| \le m$. Its coefficient is multiplied by $1 - |k|/(m+1)$. This triangular weighting is called a Fejér mean. Its convolution kernel is the squared modulus of the finite geometric sum $\sum_{j=0}^{m} e^{ijx}$, divided by $m+1$. Expanding that squared modulus counts exactly the same number of pairs with each index difference k.

The resulting kernel is nonnegative and has unit integral when measured with the normalised period measure $dx/(2\pi)$. The reconstructed value is thus a weighted average of the true values. If a real continuous signal lies between zero and one everywhere, this averaged approximation remains in the same interval. For continuous periodic functions the averages converge uniformly: continuity makes nearby signal values uniformly close, while away from the origin the geometric-sum formula bounds the kernel by a constant divided by $m + 1$. Splitting the convolution into these two regions makes the approximation argument explicit.

Compare the reconstructed amplitudes at fixed bandwidth

Take $f(x) = \cos x + 2\cos(3x)$, retaining bandwidth through three. Ordinary projection reproduces this finite Fourier series exactly. Averaging the four partial sums of orders 0 through 3 retains the first cosine with weight $3/4$ and the third with weight $1/4$. The reconstruction is $0.75\cos x + 0.5\cos(3x)$. At zero it predicts 1.25 instead of three. Since the omitted amplitudes are 0.25 and 1.5, orthogonality gives squared mean error $(0.25^2 + 1.5^2)/2 = 1.15625$. The loss is substantial precisely because the real signal contains strong energy near the edge of the chosen band.

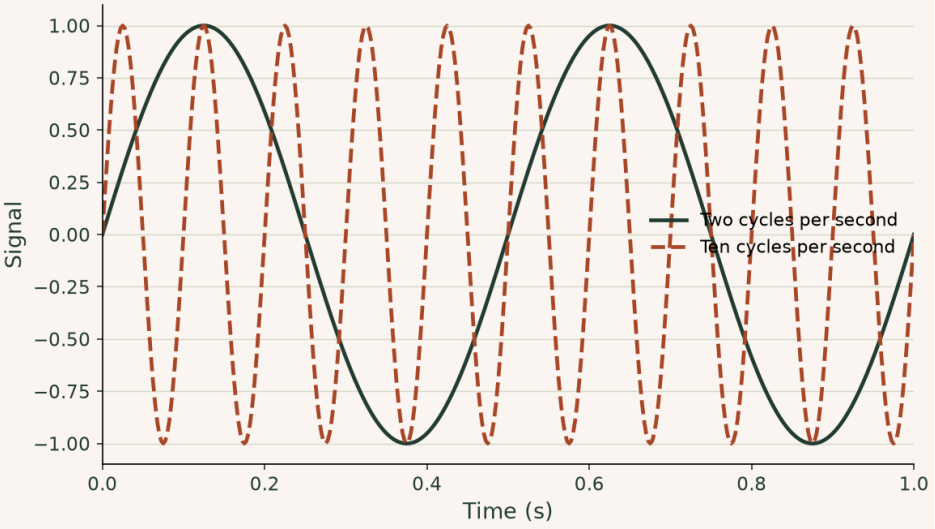
For a periodic piecewise-constant indicator, on the other hand, preservation of the interval between zero and one can be more important than full transmission of high modes. At an isolated jump separating two flat values, symmetry of the kernel assigns equal weight to sufficiently close points on either side of the jump. As the bandwidth grows, the averaged limit at a jump is the midpoint of its one-sided limits under these conditions. It cannot reproduce both sides at the mathematical point of discontinuity, regardless of the convention used to label that point in a data file.

These guarantees concern the continuous kernel and accurate coefficients. Aliased coefficients computed from an insufficient grid represent a different periodic function; applying triangular weights after such sampling cannot recover the missing distinctions between frequencies. Separate the choice of smoothing weights from the acquisition resolution when comparing a nonnegative reconstruction with actual measurements.

The smoothing window also supplies a direct frequency-response test. Feeding the averaging routine a unit cosine at frequency two with bandwidth three should return half its original amplitude. Feeding a constant should return the constant unchanged. Together these probes distinguish a correctly normalised triangular response from an accidental averaging operation that also attenuates the mean.

Make two different waves agree at every sample

Sample each continuous signal eight times per second at $t = j/8$. At these instants the phases of the two components differ by $2\pi j$. A discrete Fourier transform of those measurements cannot determine which continuous waveform produced them.



At t = 1/8 s both waves equal one, but between recorded times the ten-hertz signal executes an extra cycle. Increasing the acquisition rate with an appropriate analogue bandwidth is different from interpolating more display points.

▶ **Read values from the plot**

Change parameters · Make two different waves agree at every sample

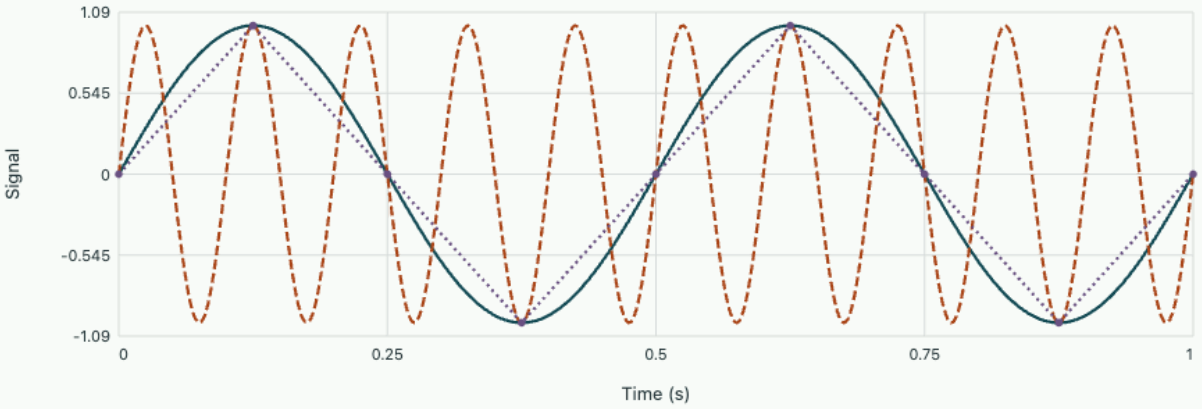
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Samples per second **8**

Base frequency (Hz) **2**



Frequencies 2 and 10 Hz agree at every sample j/8. Adding display points between these readings does not recover the lost band information.

Study the full derivation

Integrating cosine interpolants on nested grids · Function approximation and numerical integration

Chebyshev points turn polynomial interpolation into a cosine series. Derive the exact moments of each basis polynomial, explain aliasing and find a case where two consecutive quadrature grids confidently agree on the wrong answer.

Chapter 10 · Weighted projections and Legendre polynomials

Derive a best mean-square polynomial and determine how a change of weight changes what that polynomial predicts.

FROM FIRST PRINCIPLES

Project a signal under a specified measurement weight

Choose a nonnegative weight $w(x)$ and define the squared error of a model p by integration of $w|f-p|^2$ over the interval of interest. Assume the candidate basis functions have finite, positive weighted norms and are mutually orthogonal. To find the best coefficient for a basis function ϕ_j, vary that coefficient while holding the others fixed. The derivative of the squared error is

$$\frac{\partial}{\partial a_j}\int w\left(f-\sum_k a_k\phi_k\right)^2 dx = -2\left\langle f-\sum_k a_k\phi_k,\phi_j\right\rangle_w .$$

At the minimum this derivative vanishes. All cross terms involving other basis functions disappear by orthogonality, leaving

$$a_j=\frac{\langle f,\phi_j\rangle_w}{\langle \phi_j,\phi_j\rangle_w}.$$

Consequently each coefficient can be computed independently. For unit weight on $[-1,1]$, the Legendre polynomials are one natural choice. If an experiment samples unevenly or certain regions matter more, the implied weight and hence the best fit can change even though the basis spans the same set of functions.

DERIVATION AND APPLICATIONS

Quantify the return from adding one orthogonal basis function

Given a limited coefficient budget, one may want to know exactly how much an extra term reduces integrated squared error. Consider a real signal f, a positive weight defining an inner product, and an existing least-squares projection p onto a finite-dimensional space. Its residual $r=f-p$ is orthogonal to every included basis function. Suppose a proposed function ψ overlaps the original space. Remove its projection first,

$$v=\psi-P\psi,$$
$$\langle v,p_{\text{old}}\rangle=0 \quad \text{for every } p_{\text{old}} \text{ in the original space.}$$

If $v=0$, the added function contains no new direction. Otherwise the expanded space splits into the original space and the direction v. The new projection is $p+av$; expansion and minimization give

$$\|r - av\|^2 = \|r\|^2 - 2a\langle r, v\rangle + a^2\|v\|^2,$$

$$a = \frac{\langle r, v\rangle}{\|v\|^2}.$$

The benefit of adding the term follows immediately,

$$\|r\|^2 - \|r - av\|^2 = \frac{|\langle r, v\rangle|^2}{\|v\|^2}.$$

This expression is unchanged if the candidate is multiplied by a nonzero constant. It also explains why a large raw correlation with a candidate function need not imply new explanatory power: part of that candidate can be represented already. The chosen weight must be used consistently in all projections. A continuous integral and a discrete sum over experimental samples may define different gains.

WORKED EXAMPLE

Calculate the gain from adding curvature

With unit weight on $[-1, 1]$, fit $f = x^2$ first by a constant. The constant fit is $1/3$. Adding $\psi = x^2$ introduces $v = x^2 - 1/3$, equal to the residual. The integrated reduction is

$$\int_{-1}^{1} (x^2 - 1/3)^2 dx = 2/5 - 2/9 = 8/45.$$

The expanded fit reproduces the signal exactly.

SECOND APPLICATION

Compare two candidate corrections to a measured constant

Take three equally weighted readings $f = (2, 3, 7)$. The constant projection gives 4 and residual $(-2, -1, 3)$. Two proposed centered shapes are $u = (-1, 0, 1)$ and $v = (1, -2, 1)$. Their gains are $5^2/2 = 12.5$ and $3^2/6 = 1.5$ respectively. The first leaves squared error $14 - 12.5 = 1.5$. The other centered shape is orthogonal to the first, so adding both eliminates the residual.

Distance depends on the observation

A large error over a small part of an interval may be acceptable when an instrument averages over the interval. With a positive integrable weight w, minimise $J(p) = \int_a^b w(x)|f(x) - p(x)|^2\,dx$ over polynomials of degree at most m. The first variation in an arbitrary allowed direction q gives $\int w(f - p)q = 0$. Conversely this orthogonality implies

$$J(p + q) = J(p) + \int wq^2.$$

This Pythagorean identity proves uniqueness and explains why a residual with large local extrema can still be optimal in the chosen metric.

In a monomial basis these conditions couple every coefficient to every other. Choose polynomials ϕ_j orthogonal for the same weight instead. Then $p = \sum c_j \phi_j$ with $c_j = \int w f \phi_j / \int w \phi_j^2$. Increasing the degree preserves every coefficient already computed. Using quadrature to estimate these integrals introduces another approximation. A rule that samples too sparsely can make two distinct basis vectors look identical under the discrete inner product.

Construct a basis on a symmetric interval

For unit weight on $[-1, 1]$, start with $1, x, x^2$. Odd integrands integrate to zero, so the first two are orthogonal already. Subtract the projection of x^2 onto the constant, namely $1/3$, and retain $x^2 - 1/3$. Scaling this polynomial to take value one at the right endpoint gives $P_2 = (3x^2 - 1)/2$, the second Legendre polynomial. Its squared integral is $2/5$, obtained by integrating its expanded fourth-degree square.

Approximating x^2 by a line retains just $1/3$. Squaring the residual and integrating gives

$$2/5 - 4/9 + 2/9 = 8/45.$$

The mean-square error is half that number and its square root is about 0.29814. The worst local error occurs at the endpoints and equals $2/3$. The constant $1/2$ has smaller uniform error but a larger squared integral of $7/30$. The objective function explains the difference without appealing to how a fitted curve looks.

Move the weight toward one endpoint

On $[0, 1]$ choose density $w(x) = 2x$, normalised to integrate to one. Fitting a constant to $f(x) = x$ now gives its weighted mean $2 \int_0^1 x^2 \, dx = 2/3$. The weighted mean-square error is

$$2 \int x^3 \, dx - (2/3)^2 = 1/18.$$

Equal weighting would instead place the constant at $1/2$. Selecting a weight thus amounts to specifying the region over which a loss should be economical. When interpreting coefficients state the interval, basis normalisation and weight together; otherwise two reported coefficients can be different descriptions of the same fitted polynomial.

A discrete inner product can give a different optimum

If only the endpoints of $[-1, 1]$ are sampled with equal weight, then x^2 takes value one at every observation and the best discrete constant is one. Adding the midpoint with the same weight changes it to $2/3$. Simpson weights on those three samples instead give $(1 + 4 \cdot 0 + 1)/6 = 1/3$, the exact

continuous coefficient, because that particular integral is quadratic. A mismatch between a discrete fit and a continuous projection can thus arise entirely from the observation design, even with error-free measurements.

Represent multiplication by the independent variable

Orthogonal polynomials supply a bridge from projection to quadrature. Choose an orthonormal family $\phi_0, \phi_1, \ldots$ on an interval with positive weight w. Multiplication of ϕ_j by x raises its degree by one. Its inner product with ϕ_i for $i < j - 1$ vanishes, because $\langle x\phi_j, \phi_i\rangle = \langle \phi_j, x\phi_i\rangle$ and $x\phi_i$ has degree below j. Consequently $x\phi_j$ is a linear combination of just $\phi_{j-1}, \phi_j, \phi_{j+1}$. Choosing positive leading coefficients makes the outer recurrence coefficients positive.

Truncating this multiplication map to the first n basis functions gives a symmetric tridiagonal matrix. At a root ξ of ϕ_n the omitted last recurrence term vanishes. The column $(\phi_0(\xi), \ldots, \phi_{n-1}(\xi))^T$ is then an eigenvector of the truncated matrix with eigenvalue ξ. Thus a polynomial root-finding task for Gaussian quadrature nodes becomes a symmetric eigenvalue problem. This correspondence uses the same positive weight throughout. If projection coefficients were assembled with a different effective inner product from the quadrature moments, the nodes would describe a different measure.

Derive two-point Gaussian quadrature from a small matrix

On $[-1, 1]$ with unit weight, use $\phi_0 = 1/\sqrt{2}$ and $\phi_1 = \sqrt{3/2}\, x$. Both diagonal entries $\int x\phi_i^2 dx$ vanish by oddness. The off-diagonal entry is

$$(\sqrt{3}/2)\int_{-1}^{1} x^2 dx = 1/\sqrt{3}.$$

The truncated multiplication matrix has zeros on the diagonal and $1/\sqrt{3}$ off it, so its eigenvalues are $\pm 1/\sqrt{3}$. The two normalised eigenvectors have first-coordinate magnitude $1/\sqrt{2}$. Symmetry makes the quadrature weights equal and their sum must be the mass of the interval, two; hence the two weights are both one.

A further spectral check reconstructs low moments without an integration table. Write the matrix as J. Since $J^2 = I/3$, the quadrature estimate of the second moment is $2e_1^T J^2 e_1 = 2/3$, whereas odd moments through three vanish. The rule therefore integrates every cubic exactly by linearity. At degree four the same formula predicts $2/9$, while the true fourth moment is $2/5$. Multiplication by x repeatedly can produce a component outside the retained basis that returns later to affect the observed moment; truncation eventually loses that information.

For nonsymmetric weights the recurrence diagonal is usually nonzero and the quadrature nodes need no longer occur in opposite pairs. A practical diagnostic remains available: all weights must be positive and their sum must equal the integral of the constant. Passing these checks is necessary but does not certify the node positions. Testing further independently known moments can expose a normalisation or indexing error that the constant check alone would miss.

For instance adding the same spurious offset to both computed quadrature nodes leaves the weights positive and unchanged. The integral of the constant still passes exactly, while the first moment becomes twice that offset instead of zero. On an interval invariant under reflection, testing one odd and one even moment quickly distinguishes centring errors from errors in the spacing or overall weight normalisation.

Change parameters · Explore the worked calculation

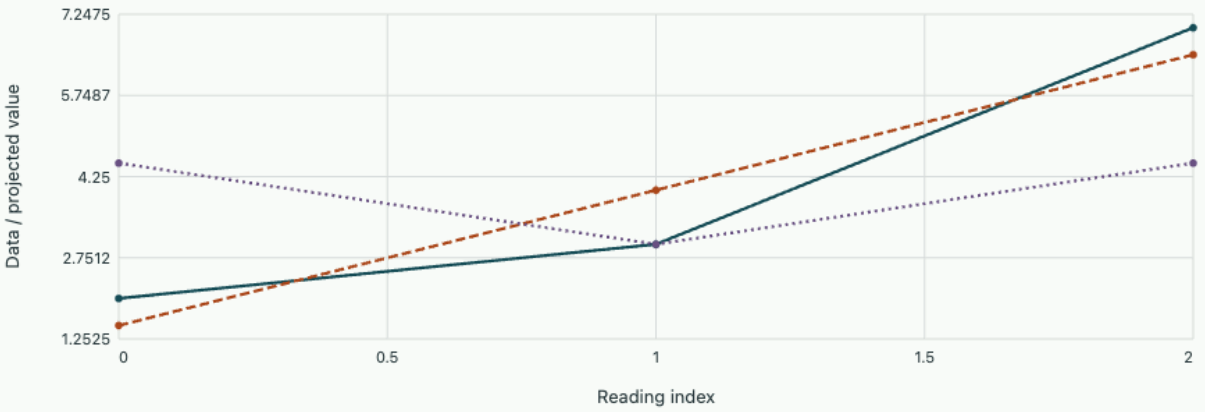
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Middle reading 3

Last reading 7



The two candidate squared-error reductions are 12.5 and 1.5. Choose the larger gain only after using the same observation metric for both shapes.

Problem sheet 5 · Spectral measures of error

Distinguish mean-square projections on continuous and sampled intervals and interpret the residual energy.

Recover the invisible frequency

A recorder samples $\sin(7x)$ at $x_j = 2\pi j/10$. Identify its low-frequency alias. Give a point where the two functions disagree, and find the mean-square energy of their difference over a period.

▼ **Worked solution**

Since 7 is congruent to −3 modulo 10, the sampled signal agrees with $-\sin(3x)$. At $x = \pi/10$ their difference is $2\sin(3\pi/10) \approx 1.618$.

On the continuous period the two sine modes are orthogonal. Each has mean square $1/2$, so their difference has mean square 1. A vanishing discrepancy at every sample therefore coexists with an order-one continuous error.

Explore the related calculation with adjustable inputs. Inspect continuous waves that agree at all recorded samples; the frequency-counting argument remains necessary to establish a whole-band guarantee.

Which modes deserve space?

Let $f(x) = 1 + 0.6\sin x + 0.8\cos(4x)$. Compute its mean square, the root-mean-square error after retaining frequencies at most one, and the best additional single real basis function among all higher-frequency sine and cosine modes.

▼ **Worked solution**

Pairwise orthogonality gives

$$1 + 0.36/2 + 0.64/2 = 1.5.$$

Omitting the fourth cosine leaves root-mean-square error $\sqrt{0.32} \approx 0.566$.

The only omitted nonzero coefficient multiplies $\cos(4x)$. Adding this mode removes the entire remaining mean-square error. Choosing an intermediate frequency for its lower index would make no change for this specific signal.

Explore the related calculation with adjustable inputs. Inspect continuous waves that agree at all recorded samples; the frequency-counting argument remains necessary to establish a whole-band guarantee.

Project an odd function

Find the best linear mean-square approximation to x^3 on $[-1, 1]$ with unit weight. Compute its squared integrated error and compare the optimum slope with endpoint interpolation.

▼ **Worked solution**

Symmetry forces zero constant. Orthogonality to x requires

$$a = \int x^4 / \int x^2 = (2/5)/(2/3) = 3/5.$$

Expanding the square gives

$$2/7 - 2(3/5)(2/5) + (9/25)(2/3) = 8/175.$$

Endpoint interpolation gives slope 1. The least-squares line trades endpoint accuracy for lower integrated loss.

Explore the related calculation with adjustable inputs. Compare the reduction produced by two orthogonal correction directions. The continuous integral and the finite observation sum use the same projection identity but different inner products.

Find the zeros in an eight-sample spectrum without eight long sums

A sensor returns $(1, 2, 0, 0, 1, 2, 0, 0)$ at eight equally spaced times. Before computing any transforms, predict which four Fourier bins vanish. Use a four-point transform to find the other bins and verify the discrete energy identity for the unnormalised DFT.

▼ **Worked solution**

Pair sample n with $n + 4$. The second weight differs by $e^{-i\pi k} = (-1)^k$, so all odd bins cancel. For even $k = 2j$ the two halves reinforce and give twice the four-point DFT of $(1, 2, 0, 0)$. Its entries are $3, 1 - 2i, -1, 1 + 2i$. Thus

$$X_0 = 6,\ X_2 = 2 - 4i,\ X_4 = -2,\ X_6 = 2 + 4i.$$

The sample energy is

$$2(1^2 + 2^2) = 10.$$

The frequency-domain squared magnitudes sum to $36 + 20 + 4 + 20 = 80$. Dividing by eight gives 10. Both the zeros and the energy check depend on the adopted transform convention. Reversing the exponent sign conjugates these coefficients and leaves their squared magnitudes unchanged.

Explore the related calculation with adjustable inputs. Inspect continuous waves that agree at all recorded samples; the frequency-counting argument remains necessary to establish a whole-band guarantee.

Revisit chapter 9 · Factor the discrete Fourier transform into short transforms.

Chapter 11 · Quadrature near singularities and oscillations

Transform integrable endpoint singularities and use phase information to diagnose a misleading quadrature result.

FROM FIRST PRINCIPLES

Remove an integrable endpoint singularity by substitution

An integrand can be unbounded and still have finite area. Consider $f(x) = x^{-\alpha} g(x)$ on $[0, 1]$, where g is continuous and bounded and $0 < \alpha < 1$. The power law has a finite integral near zero since

$$\int_0^{\varepsilon} x^{-\alpha}\, dx = \frac{\varepsilon^{1-\alpha}}{1-\alpha}.$$

Ordinary endpoint rules may still try to sample $f(0)$, which is undefined. Substitute $x = t^q$ and include the Jacobian $dx = q t^{q-1} dt$. The transformed integral becomes

$$\int_0^1 x^{-\alpha} g(x)\, dx = \int_0^1 q\, t^{q(1-\alpha)-1} g(t^q)\, dt.$$

Taking $q = 1/(1-\alpha)$ removes the explicit power. For a square-root singularity the familiar substitution $x = t^2$ gives a regular factor $2g(t^2)$. Continuity of the transformed integrand makes endpoint evaluation meaningful, while the convergence order of a high-order rule still depends on how many transformed derivatives are bounded.

DERIVATION AND APPLICATIONS

Subtract a known endpoint contribution before integrating the remainder

A substitution is one way to handle an integrable endpoint singularity. If an expansion of its smooth coefficient is known, a second route is to integrate a few singular terms analytically. Consider $I = \int_0^1 g(x)/\sqrt{x}\, dx$ with g twice continuously differentiable and $|g''| \le M$. Taylor's theorem about $x = 0$ separates the constant and slope from a remainder bounded by $Mx^2/2$,

$$g(x) = g(0) + x g'(0) + R(x), \qquad |R(x)| \le M x^2/2.$$

The first two contributions have finite known moments. The residual now vanishes at the problematic endpoint and can be sampled continuously there,

$$I = 2g(0) + \frac{2}{3} g'(0) + \int_0^1 \frac{R(x)}{\sqrt{x}}\, dx, \qquad \left|\frac{R(x)}{\sqrt{x}}\right| \le \tfrac{M}{2} x^{3/2}.$$

If only the short initial interval $[0, h]$ is targeted, the two analytic terms scale as $\sqrt{h}$ and $h^{3/2}$. Bounding the remaining contribution yields a direct local error certificate,

$$\left| \int_0^h \frac{g(x)}{\sqrt{x}}\,dx - 2g(0)\sqrt{h} - \frac{2}{3}g'(0)h^{3/2} \right| \le \frac{M}{5}h^{5/2}.$$

The two known coefficients have to use the same units and coordinate origin as h. Numerically it can be advantageous to evaluate the residual through a stable expression or a local series, because subtracting its first two terms from a nearly equal value of $g(x)$ can lose relative precision.

WORKED EXAMPLE

Integrate a polynomial endpoint layer

For $g(x) = 1 + 3x + 2x^2$ and $h = 0.04$, the two retained terms give

$$0.4 + 2(0.008) = 0.416.$$

The remaining integral is $2(2/5)h^{5/2} = 0.000256$, giving total 0.416256. The error bound with $M = 4$ also equals 0.000256, attained because the curvature has constant sign and magnitude.

SECOND APPLICATION

Choose an endpoint panel from a tolerance

A different coefficient has verified $|g''| \le 10$. To limit the omitted endpoint area to 10^{-5}, require $2h^{5/2} \le 10^{-5}$, or $h \le (5 \times 10^{-6})^{2/5} = 0.00758$. Choosing $h = 0.005$ gives a bound 3.54×10^{-6}. The remaining regular part of the integral still needs its own quadrature budget.

When derivative bounds fail

The usual trapezoid error bound needs a bounded second derivative. The integral of $x^{-1/2}$ from zero to one is finite, yet the integrand itself is undefined at the first endpoint. Calling a closed quadrature formula on its original samples is already impossible. Simply omitting that endpoint also omits part of a finite area, and its size scales as the square root of the omitted interval. A substitution can encode the known local behaviour exactly.

Set $x = t^q$. A factor $x^{\alpha}dx$ becomes $q t^{q(\alpha+1)-1}\,dt$. For $\alpha > -1$, choosing $q = 1/(\alpha+1)$ removes the leading power. Any additional factor $g(t^q)$ still determines the transformed smoothness. For $\alpha = -1/2$, $q = 2$ and $g(x) = 1 + x$, the integral becomes

$$\int_0^1 2(1 + t^2)\,dt = 8/3.$$

Composite Simpson quadrature on just two equal panels is exact on this quadratic. Assign the endpoint of the transformed expression by its limit to avoid evaluating zero times infinity.

A deceptively constant sample

Now consider $\int_0^1 \cos(16\pi x)\,dx$. The exact value is zero, because eight periods fit the interval. At the nine nodes of an eight-panel uniform trapezoid rule each value is one. The computed area is one, a full unit of error for an analytic function. A finer grid of sixteen panels samples alternating signs and returns zero. Repeating the first calculation on four panels also returns one. Agreement between the four- and eight-panel estimates therefore gives a false impression of convergence.

The aliasing follows from summing a geometric progression of sampled complex exponentials. When the phase step is an integer multiple of 2π, neighbouring samples coincide. Changing the offset slightly can expose the problem even before changing the number of nodes. A bound involving derivatives recognises the difficulty because the second derivative grows quadratically with the phase frequency.

Use cancellation as part of the method

When the oscillatory factor is known analytically, integrate it rather than forcing the grid to resolve every period. For a differentiable amplitude g, integration by parts gives

$$\int_a^b g(x)\cos(\omega x)\,dx = \left[\frac{g(x)\sin(\omega x)}{\omega}\right]_a^b - \frac{1}{\omega}\int_a^b g'(x)\sin(\omega x)\,dx.$$

This identity shows how smooth amplitudes can produce small integrals even when pointwise values are large. For $g = x$ and $\omega = \pi$ on $[0, 1]$, the boundary term vanishes. The remaining sine integrates to $2/\pi$, giving $-2/\pi^2$. A quadrature scheme that approximates the slowly varying amplitude by lines and integrates each line against the phase reproduces this result exactly for this example. If the phase itself is uncertain, that modelling error must be assessed separately from quadrature error.

Test integrability before applying the change

For $f(x) = 1/x$ on $(0, 1]$ the substitution $x = t^q$ leaves q/t, still nonintegrable for every positive exponent. A numerical rule excluding zero may nevertheless return an ordinary finite number. The integral from a cutoff δ to one is $-\log\delta$, so reducing the cutoff by a factor of ten adds $\log 10 \approx 2.3026$ every time. That persistent change signals a divergent problem definition. It calls for a physically justified finite cutoff or a different mathematical quantity before refining a quadrature rule.

Integrate the singular weight exactly on each panel

Sometimes a substitution changes an integrable singularity into a regular function but also complicates the geometry or an expensive supplied amplitude. An alternative is product integration. Suppose $I = \int_0^1 x^{-1/2} g(x)\,dx$ where only the smooth amplitude g is sampled. Replace g by its piecewise linear interpolant while keeping the weight exactly inside each local integral. The contribution of an endpoint sample equals the integral of the corresponding linear basis function against the singular weight. None of these weights requires evaluating $x^{-1/2}$ at zero.

On a single panel covering the unit interval the basis functions are $1-x$ and x. Their weights are $2-2/3=4/3$ and $2/3$, respectively. They are positive and sum to two, the exact integral of the singular factor. The rule reproduces constant and linear amplitudes exactly. If $|g''|\le M$, the pointwise error satisfies

$$|g-I_1 g|\le Mx(1-x)/2.$$

Multiplying by the positive weight and integrating gives absolute error at most $(M/2)(2/3-2/5)=2M/15$. This derivation tracks smoothness of the amplitude that is actually interpolated.

Compare a weighted panel with an ordinary endpoint sample

Let $g(x)=1+x^2$. The analytically integrated value is $2+2/5=2.4$. The weighted panel predicts

$$(4/3)1+(2/3)2=8/3.$$

The excess is $4/15$, exactly the bound for $M=2$, because the interpolating line lies above this convex amplitude throughout the open interval. A direct unweighted trapezoid evaluation would have tried to sample an infinite value at the left endpoint. The finite weights here contain all information about the integrable divergence.

For a more refined calculation divide at $x=1/4$. On the first panel, rescaling $x=t/4$ multiplies the singular measure by $1/2$, so the local weights become $2/3$ and $1/3$. On the second panel use the basis $(1-x)/(3/4)$, $(x-1/4)/(3/4)$. The weight of its left endpoint is

$$(4/3)[2\sqrt{x}-(2/3)x^{3/2}]_{1/4}^{1}=5/9.$$

The remaining weight is $4/9$, since that panel's total singular measure is one. Combining adjacent contributions gives global weights $2/3, 8/9, 4/9$ at the three sampled amplitudes.

The estimate is now $2/3+(8/9)(17/16)+(4/9)2=2.5$. Its positive error has fallen from $4/15$ to 0.1. Subdivision concentrated near the singular endpoint is one design choice, while error here also depends on where the smooth amplitude bends. If the amplitude itself has a discontinuity, splitting at that physical interface is necessary before invoking the second-derivative bound on either side.

The positive weights also control bounded amplitude uncertainty. If each of the three amplitude evaluations is uncertain by at most 0.01, the worst possible additional integral error is

$$0.01(2/3+8/9+4/9)=0.02.$$

The bound is achieved by a shared positive offset of 0.01. Unlike the deterministic interpolation discrepancy, this worst-case observation error will remain 0.02 under any further positive-weight refinement that still integrates constant amplitudes exactly. A complaint about slow convergence near the endpoint could therefore reflect either amplitude resolution or the data error floor.

Repeated amplitude measurements can reduce independent sampling variance but cannot remove the effect of a shared calibration offset.

Change parameters · Explore the worked calculation

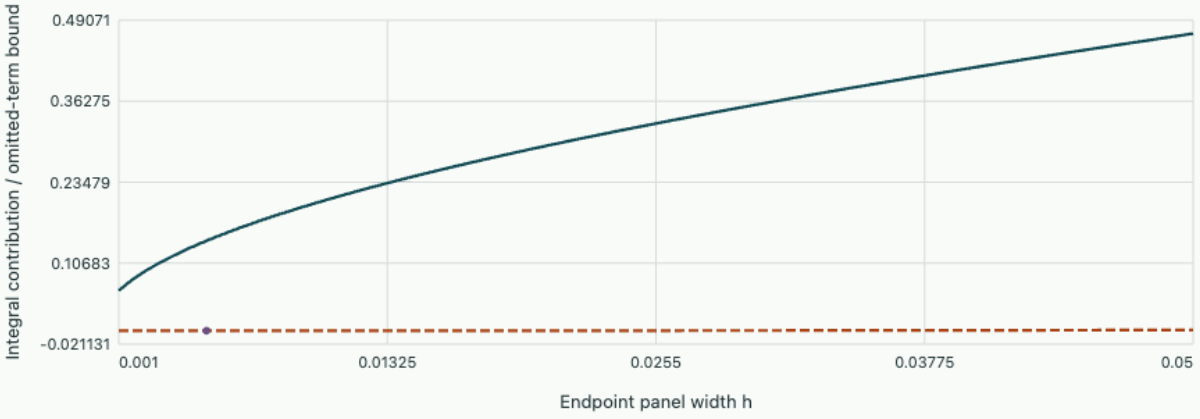
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Second-derivative bound M **10**

Selected endpoint panel width **0.005**



For g(x)/√x with g(0) = 1 and g′(0) = 3, the selected remainder bound is 0.0000035355. The bound assumes |g″| ≤ M over that panel; the rest of the integral has a separate budget.

Chapter 12 · Monte Carlo integration and stratification

Estimate an integral from random samples, attach a sampling uncertainty and derive the gain from stratification in a simple case.

FROM FIRST PRINCIPLES

Derive a Monte Carlo estimate and its sampling error

Draw independent, uniformly distributed points X_i on $[a, b]$, and suppose $f(X_i)$ has finite variance σ_f^2. The expected value of one function sample is its integral divided by the interval length. This immediately gives an unbiased integral estimator,

$$I = (b - a)\,\mathrm{E}[f(X)],$$

$$\widehat{I}_N = \frac{b - a}{N} \sum_{i=1}^{N} f(X_i).$$

For independent draws, the variance of a sum is the sum of the variances. Dividing by the square of the sample count gives

$$\operatorname{var}(\widehat{I}_N) = \frac{(b - a)^2 \sigma_f^2}{N}.$$

Replacing σ_f by the empirical sample standard deviation estimates the standard error. The $1/\sqrt{N}$ decrease concerns repeated sampling; it is not a deterministic bound on one estimate. Quadrupling the number of evaluations halves that standard error. Dependence between samples, an infinite variance, or a biased point generator changes the reasoning and the interpretation of the reported uncertainty.

DERIVATION AND APPLICATIONS

Predict when a reflected random sample reduces integration noise

Uniform draws on $[0, 1]$ have the same distribution after reflection about $1/2$. This permits antithetic sampling, in which each independent draw U is paired with $1 - U$. For a square-integrable integrand f, define the per-pair estimate Z as the average of its two evaluations. Both terms have the desired integral as their expectation, so

$$Z = \tfrac{1}{2}[f(U) + f(1 - U)],$$

$$\mathrm{E}\,Z = \int_0^1 f(x)\,dx.$$

Assume m independently generated pairs and finite single-sample variance $\sigma^2 > 0$. The two members within a pair can be correlated. Let ρ be their correlation. Expanding the variance of their sum gives

$$\operatorname{var}(Z) = \frac{\sigma^2}{2}(1+\rho),$$

$$\operatorname{var}(\overline{Z}_m) = \frac{\sigma^2}{2m}(1+\rho).$$

For the same $2m$ evaluations, independent sampling has variance $\sigma^2/(2m)$; thus negative correlation helps. The source of any reduction can be seen by decomposing f about the midpoint into symmetric and antisymmetric pieces,

$$f_{\text{even}}(x) = \tfrac{1}{2}[f(x) + f(1-x)],$$

$$f_{\text{odd}}(x) = \tfrac{1}{2}[f(x) - f(1-x)].$$

Pair averaging keeps the first and cancels the second exactly. The two pieces are orthogonal under uniform integration, so cancellation removes the antisymmetric contribution to variability. It also uses two evaluations for each independent pair. Standard errors should therefore be calculated across the pair averages.

WORKED EXAMPLE

Calculate the gain for a quadratic integrand

For $f(x) = x^2$,

$$\sigma^2 = 1/5 - 1/9 = 4/45.$$

The covariance with its reflected partner is $\int_0^1 x^2(1-x)^2 dx - 1/9 = 1/30 - 1/9 = -7/90$, giving $\rho = -7/8$. At a fixed evaluation budget the variance falls by a factor of eight and the standard error by $\sqrt{8}$. Each pair integrates the linear part about the midpoint exactly.

SECOND APPLICATION

Recognize when the second evaluation is redundant

For $f(x) = (x - 1/2)^2$, reflection gives the same function value. Here $\rho = 1$ and the paired method has twice the variance of independent sampling at equal cost. With 200 evaluations and single-value variance $1/180$, the independent standard error is $1/\sqrt{36000} = 0.00527$; using 100 independent pairs gives

$$1/\sqrt{18000} = 0.00745.$$

Turn area into an expectation

For an integrable f on $[a, b]$, let U be uniform on the interval. Then $I = (b-a)\mathbb{E}[f(U)]$. Independent draws give an unbiased estimator by averaging their function values and multiplying by the length. With finite variance $\sigma^2 = \operatorname{Var}(f(U))$, independence eliminates cross-covariances and gives

$$\operatorname{Var}(\hat{I}) = (b-a)^2\sigma^2/n.$$

The typical size of the sampling error decreases with the square root of the sample count. Multiplying the budget by four halves the standard error.

In a numerical run replace the unknown variance by the sample variance with denominator $n-1$. For a sufficiently large sample with a suitable distribution, a normal approximation provides an approximate confidence interval. A particular interval need not contain the exact answer. Heavy tails, correlated draws or a nearly unsampled small peak can make the estimated uncertainty unreliable. Recording the random seed reproduces the same sequence of sample values; independent seeds are needed for independent replications.

Four draws and the uncertainty they reveal

For $f(x) = x$ on $[0,1]$, four illustrative draws at 0.1, 0.2, 0.7 and 0.8 have mean 0.45. Their squared deviations sum to 0.37, yielding sample variance $0.37/3$ and estimated standard error $\sqrt{0.37/12} \approx 0.176$. The exact integral is 0.5, so the observed error is 0.05. Because the uniform variance is exactly $1/12$, the true standard deviation for repeated sets of four draws is $\sqrt{1/48} \approx 0.144$. These three numbers answer different questions about this one run and its possible replications.

Guarantee some coverage before sampling

Divide the interval into n equal strata and draw one independent point uniformly inside each. Sum each sample value times its stratum width. The estimator remains unbiased since each local expectation equals the mean over its stratum. For $f(x) = x$ every local variance is $1/(12n^2)$. Multiplying by the squared width and summing n independent contributions gives a total variance of $1/(12n^3)$. For four samples the standard deviation drops to about 0.0361. The gain comes from fixing the amount of sampling effort in each region.

A single sample per stratum cannot estimate its own local variance. Independent repetitions of the complete stratified procedure or more samples per stratum permit empirical uncertainty estimates. In high dimension a full Cartesian partition requires rapidly growing numbers of cells. Simple Monte Carlo retains the same square-root rate under finite variance, but the variance constant and the cost of sampling can still depend strongly on dimension. Identify both the estimator and the method used to measure its uncertainty when reporting a numerical integral.

Use a known integral to reduce variability

For $f(x) = x + x^2$, the integral of the first term is known to be one half. Sampling only x^2 and adding one half leaves the expectation unchanged. The original sample variance per uniform draw is

$$1/3 + 1/2 + 1/5 - (5/6)^2 = 61/180.$$

The residual sample variance per draw is $1/5 - (1/3)^2 = 4/45$, less than one third as large. This control-variate construction uses a known mean and a correlation with the original integrand. An arbitrary replacement by a superficially similar function without restoring its known mean would bias the estimate.

Sample more often where the integrand is large

For a nonuniform positive probability density q on the integration region, the identity $\int f(x)\,dx = \mathbb{E}_q[f(X)/q(X)]$ holds whenever the sampling distribution covers the integrand's support. Independent samples produce an unbiased estimate. Its variance per draw is $\int f(x)^2/q(x)\,dx - I^2$, provided this integral is finite. A density concentrated near large contributions can reduce this variance even if generating a nonuniform draw costs little more than generating a uniform one. A density that is too small where the integrand is large can have the opposite effect.

Cauchy–Schwarz applied to $\int |f| = \int (|f|/\sqrt{q})\sqrt{q}$ gives

$$(\int |f|)^2 \le \int f^2/q.$$

Equality requires q proportional to $|f|$. For a nonnegative integrand that ideal density makes the sample weight constant and the variance zero. Constructing its normalising constant would generally require knowing the integral itself. The result is thus a design guide: choose a tractable density approximating the shape of the large contributions, with known normalisation and sufficiently broad coverage.

Restore finite variance for an integrable endpoint peak

Take $f(x) = x^{-1/2}$ on $(0, 1]$. The integral is two. Under uniform sampling each weight is finite almost surely and has that finite mean, but its second moment is

$$\int_0^1 x^{-1} dx = \infty.$$

The familiar standard-error formula based on a finite variance and a square-root sampling rate is therefore unjustified, even if a modest run happens to avoid tiny x and reports a reassuringly small sample variance.

Choose $q(x) = (3/4)x^{-1/4}$. Its cumulative distribution is $x^{3/4}$, so one draws $X = U^{4/3}$ for a uniform U. The importance weight is

$$W = (4/3)X^{-1/4}.$$

The second moment is $(4/3)\int_0^1 x^{-3/4} dx = 16/3$, giving variance $16/3 - 4 = 4/3$. For 300 independent draws the exact standard deviation of the mean is

$$\sqrt{(4/3)/300} = 1/15.$$

This result quantifies an uncertainty that uniform sampling could not express through a finite variance.

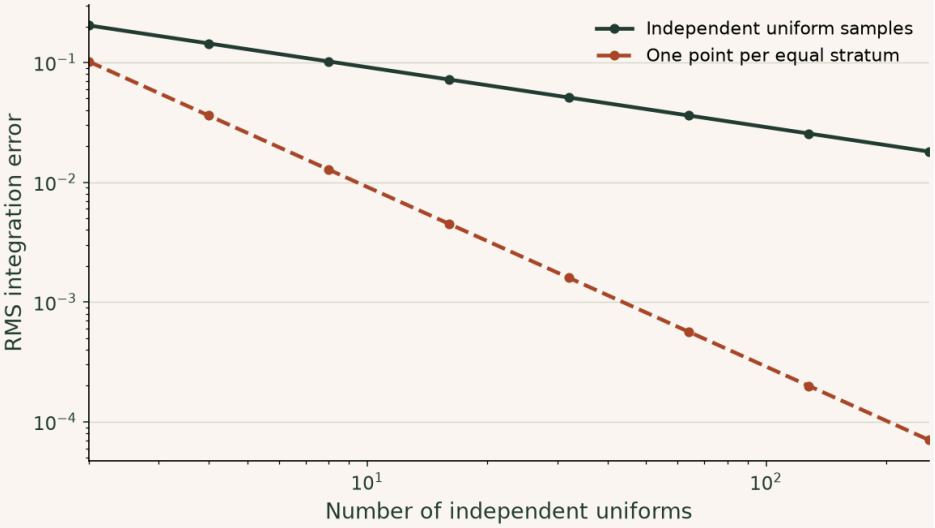
For the ideal comparison $q_* = x^{-1/2}/2$, the cumulative distribution is $\sqrt{x}$; generating $X = U^2$ gives $W = 2$ on every draw. Change the integrand to $x^{-1/2}(1 + x)$ while retaining this density and the weights become $2(1 + X)$, bounded between two and four. The method still controls the singular behaviour without needing exact knowledge of the new integral. For a sign-changing integrand, even sampling from $|f|$ retains sign fluctuation and cannot eliminate cancellation between large positive and negative contributions.

To check the improved density numerically, record both the sample mean and the largest weight across independent replications. A computation that clips extremely small sampled values of X to avoid a large weight changes the target expectation. Under the proposed q, the omitted integrand area from zero to a cutoff δ is $2\sqrt{\delta}$. Even a cutoff of 10^{-4} removes 0.02 integral units, which is already thirty percent of the quoted standard error for 300 samples. Tail handling is therefore part of defining the estimator.

A stable implementation can generate the weight directly from the sampled uniform variable and check this omitted-area estimate.

Compute the expected benefit of stratifying a smooth integrand

Estimate $\int_0^1 x\,dx$ by sampling function values. A uniform draw has variance 1/12; a draw restricted to a stratum of width $1/N$ has variance $1/(12N^2)$. Combine N independent equal-weight draws in each design.



At N = 16 the stratified estimate has sixteen times smaller RMS error for this linear integrand. These are exact ensemble RMS values; any one random realisation can have a larger or smaller error.

▶ **Read values from the plot**

Change parameters · Compute the expected benefit of stratifying a smooth integrand

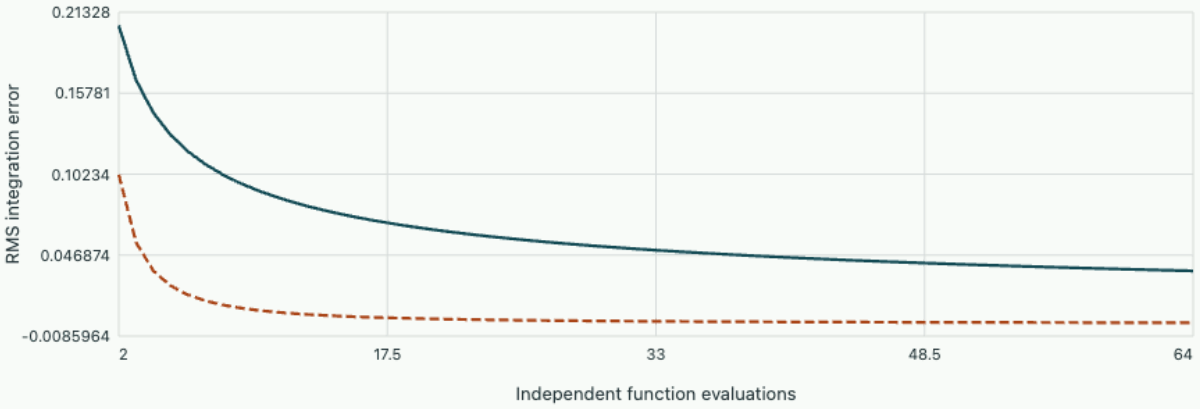
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Integrand exponent p **1**

Maximum sample count **64**



Independent uniform samples · One independent sample per stratum

For x^1, exact integral 0.5. These are ensemble RMS errors computed from exact within-cell moments, not one random run.

Study the full derivation

Spend random samples where they reduce uncertainty · Function approximation and numerical integration

A Monte Carlo estimate can be unbiased and still too noisy to be useful. Derive control variates, allocate a fixed budget across strata, and distinguish variance reduction from an underestimated standard error.

Problem sheet 6 · Choosing where to sample

Exploit endpoint structure, test aliasing and design a stratified random estimate with a known variance.

Remove a singular factor

Evaluate $\int_0^1 x^{-2/3}(1+x)\,dx$ using $x = t^q$. Select q, derive the transformed integrand, and decide whether one Simpson panel pair gives the exact integral.

▼ **Worked solution**

To neutralise the power require $q(1-2/3)-1=0$, giving $q=3$. Then $x^{-2/3}dx = t^{-2}3t^2dt = 3dt$ and the finite endpoint value is 3.

Integrating $3(1+t^3)$ gives

$$3 + 3/4 = 3.75.$$

Simpson is exact on cubics; its three samples are $3, 3.375, 6$, and their weighted sum divided by six is 3.75.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Expose a false agreement

For $f(x) = \cos(12\pi x)$ on $[0,1]$, find the trapezoid results with 3 and 6 panels. Test 12 panels and explain why the initial equal answers cannot justify a zero error estimate.

▼ **Worked solution**

With three panels the phase advances 4π at each sample; with six it advances 2π. Both return the constant sample average 1.

With twelve panels samples alternate between 1 and –1. The half-weight endpoints together count as one positive value, so all contributions cancel. The exact integral is zero. Correlated aliasing on the first two grids invalidates a smooth-error extrapolation.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Two strata with different variability

A response has variance 1 on the first half of $[0,1]$ and 4 on the other half. Estimate the integral from twelve independent evaluations. Compare the variances for allocations (6,6) and (4,8). Derive the continuous optimum under the same fixed budget.

▼ **Worked solution**

Each stratum mean has variance σ_i^2/n_i and weight $1/2$, so total variance is $(1/n_1 + 4/n_2)/4$. Equal allocation gives $5/24$; (4,8) gives $3/16$, slightly smaller.

Differentiating $1/n_1 + 4/(12 - n_1)$ gives $-1/n_1^2 + 4/(12 - n_1)^2$. Its zero requires

$$12 - n_1 = 2n_1.$$

Positive second derivative yields the optimum (4,8), allocating in proportion to each stratum's standard deviation.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Choose a quadrature variable before choosing a rule

Estimate the integral of exp(−x)/√x from 0 to 1. A composite trapezoidal rule in x fails at its first endpoint. Remove the singularity by substitution, compute Simpson estimates using two and four panels in the new variable, and derive a practical error estimate. Explain why fitting a polynomial to the original endpoint would change the question.

▼ **Worked solution**

Set x = t². The Jacobian 2t cancels 1/√x for t positive, and the integrand extends smoothly to g(0) = 2. The integral is now the integral of 2 exp(−t²) over [0,1]. The endpoint value is a limit implied by the model and the substitution.

With two equal panels Simpson gives [g(0)+4g(0.5)+g(1)]/6 ≈ 1.49436. With four panels the weights are 1,4,2,4,1 times 1/12, giving about 1.49371. If the leading error is proportional to h⁴, the estimated fine-grid error magnitude is |S₄−S₂|/15 ≈ 0.000043.

The exact benchmark √π erf(1) ≈ 1.4936483 shows that the fine-grid estimate is already close; the asymptotic estimate is approximate on these coarse grids. Replacing the infinite original endpoint by a fitted finite number amounts to imposing an arbitrary sampling model. The substitution preserves the exact integral and also explains why a higher-order rule now has the smoothness it needs.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Chapter 13 · Evaluating an interpolant without rebuilding its polynomial

Derive a quotient from the Lagrange basis, evaluate the removable singularity at a data point and separate rounding error from sensitivity to the data.

One product shared by all the basis functions

Let distinct nodes $x_0, \dots, x_n$ carry data y_j. The cardinal function $\ell_j(x)$ is one at node j and zero at the other nodes. Its product formula is expensive if every evaluation repeats the same differences. Set $w_j = 1/\prod_{k\neq j}(x_j - x_k)$ and $P(x) = \prod_k (x - x_k)$. Away from all nodes,

$$\ell_j(x) = P(x) w_j/(x - x_j).$$

The polynomial interpolating constant data equal to one is one everywhere. Dividing the interpolating polynomial by this partition of unity cancels $P(x)$ to give

$$p(x) = \frac{\sum_j w_j y_j/(x - x_j)}{\sum_j w_j/(x - x_j)}.$$

At a node the written quotient divides infinities. Multiply numerator and denominator by $x - x_j$ and take the limit to recover y_j exactly. An implementation should return the stored value when the evaluation coordinate equals a node. Multiplying every weight by the same nonzero factor leaves the quotient invariant, a useful freedom when storing weights.

Background · Barycentric weight

$$w_j = \prod_{i\neq j} (x_j - x_i)^{-1}$$

Only relative weights enter the rational expression for the interpolating polynomial. At a node return the recorded value directly.

Worked interpolation and data sensitivity

Take the nodes $-1, 0, 1$. Their weights are proportional to $1, -2, 1$. With data $1, 0, 1$ at $x = 0.5$, the numerator is $2/3 - 2 = -4/3$ and the denominator is

$$2/3 - 4 - 2 = -16/3.$$

Hence $p(0.5) = 1/4$, as the polynomial x^2 requires.

Change the middle datum from 0 to 0.01. Its cardinal function is $1 - x^2$, so the prediction at 0.5 rises by 0.0075. For arbitrary data perturbations bounded by δ, the triangle inequality gives

$$|\Delta p(x)| \le \delta \sum_j |\ell_j(x)|.$$

Large values of this sum signal a sensitivity of the interpolation problem. Evaluating the same polynomial by a smarter formula cannot remove this source of error.

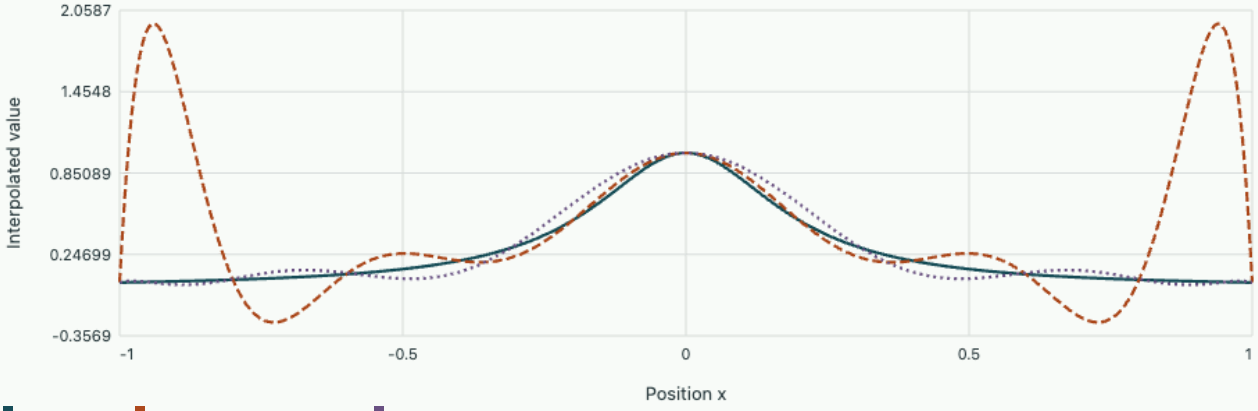
Put more sensors near the endpoints

Approximate $f(x) = 1/(1+25x^2)$ on $[-1, 1]$. Both curves use the same number of exact samples and the same barycentric algorithm. One uses equally spaced nodes and the other uses projections of equally spaced angles on a semicircle. Check the interior before inspecting the ends.

Open this experiment online

Try the calculation

Polynomial degree **10**



Maximum observed error on 501 test points: uniform 1.9156, cosine 0.13217.

Tutorial on what the formula guarantees

Why do three weights with alternating signs produce an answer between zero and one in the first calculation? Must it always lie between the smallest and largest data?

▼ **Worked solution**

The example reconstructs x^2, which stays within that range on this interval. General cardinal functions may be negative. Unlike a convex combination, their weighted sum can overshoot the sampled range.

For the three nodes above, perturb only the datum at -1 by 0.04. What is the change at $x = 0.5$?

▼ **Worked solution**

The left cardinal polynomial is $x(x-1)/2$, equal to $-1/8$ at 0.5. The change is -0.005, with the opposite sign to the perturbation.

Read the university treatment

Compare the construction of weights with the theorem on the barycentric form. Reuse the same nodes to reproduce the data-perturbation calculation. Driscoll and Braun · Barycentric interpolation

Chapter 14 · When a trapezoidal sum counts the wrong frequency perfectly

Use a finite geometric series to explain why a very accurate periodic quadrature rule can completely miss an oscillation.

Summing one Fourier mode

A smooth closed curve parameterised by angle provides a natural periodic integral. On N equally spaced points $\theta_j = 2\pi j/N$, write the trapezoidal estimate of a mean as $T_N = N^{-1}\sum_{j=0}^{N-1} f(\theta_j)$. The endpoint 2π equals zero angle and is counted only once. For $f = e^{ik\theta}$, put $q = e^{2\pi ik/N}$. If $q \neq 1$ then

$$(1-q)(1+q+\cdots+q^{N-1}) = 1 - q^N = 0.$$

If $q = 1$ all N samples equal one.

For a Fourier series whose absolute convergence permits termwise summation, this gives

$$T_N = \sum_{\ell\in\mathbb{Z}} \widehat{f}_{\ell N},$$

$$T_N - \widehat{f}_0 = \sum_{\ell\neq 0} \widehat{f}_{\ell N}.$$

The continuous mean sees only the constant mode. The grid silently adds every integer multiple of its sample count to that mode. Fast decay of smooth Fourier coefficients usually makes this error small. A single deliberately aligned harmonic can defeat the rule.

> **Background · Fourier alias**
>
> $$e^{2\pi i(k+n)j/n} = e^{2\pi ikj/n}$$
>
> On n equally spaced periodic samples two distinct continuum frequencies can yield precisely the same observations. Interpolation cannot disambiguate them without more information.

Two rings of sensors

A ring carries the temperature $20 + 3\cos(8\theta)$ °C. Eight sensors at zero angular offset all read 23 °C. The true mean is 20 °C. Nine equally spaced sensors give the exact mean for this harmonic. Moving the eight sensors through an offset ϕ changes their mean to $20 + 3\cos(8\phi)$ while the ring is unchanged.

A different signal $2 + \cos(3\theta) + 0.2\cos(10\theta)$ on five sensors gives 2.2 because the tenth harmonic aliases to zero frequency. Increasing the number to ten gives the same wrong answer. Independent node offsets or a noncommensurate sample count can challenge an apparently converged result.

Move the sampling ring

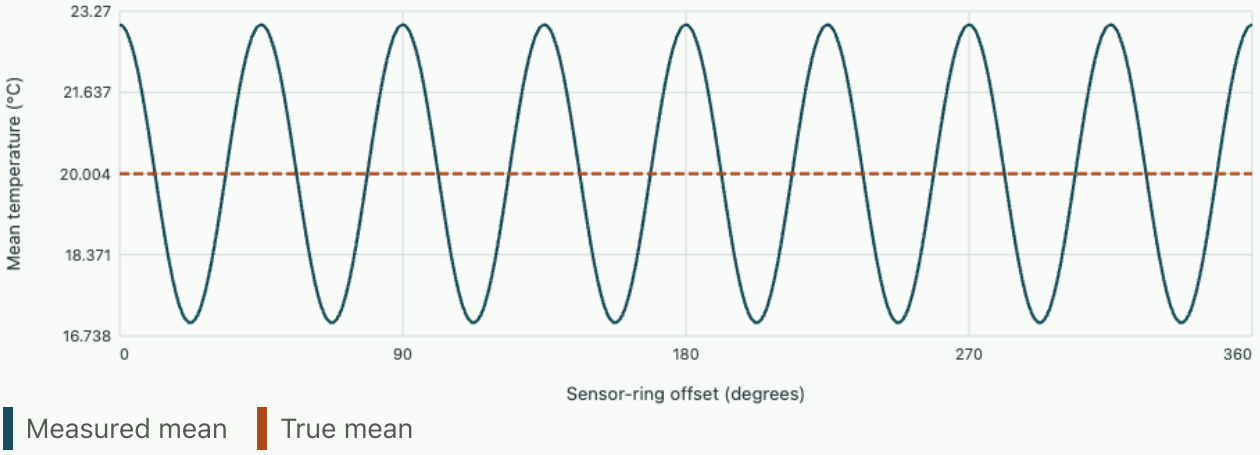
The curve gives the mean reconstructed from rotated sensors over one revolution. Change the number of sensors and the physical harmonic separately. Predict the counts that will make the measurement depend on the angular offset.

Open this experiment online

Try the calculation

Sensors N 8

Physical harmonic k 8



At zero offset the sensors read 23 °C. The largest possible mean error for this grid is 3 °C in exact arithmetic.

Tutorial on resolution

A discrete test gives the exact mean of $\cos(7\theta)$ on eight equally spaced nodes. Can eight samples reconstruct that frequency uniquely?

▼ **Worked solution**

Its mean vanishes, but samples of $\cos(7\theta)$ coincide with those of $\cos(\theta)$ on this grid. Accurate integration of the mean does not imply unique reconstruction.

Average two N-point rules offset by π/N. Which aliases survive?

▼ **Worked solution**

Together the rings form a $2N$-point uniform grid. Terms at odd multiples of N change sign between grids and cancel. Terms at even multiples survive.

Study the full derivation

Integrating cosine interpolants on nested grids · Function approximation and numerical integration

Chebyshev points turn polynomial interpolation into a cosine series. Derive the exact moments of each basis polynomial, explain aliasing and find a case where two consecutive quadrature grids confidently agree on the wrong answer.

Chapter 15 · Deriving Romberg integration and recognising a broken error model

Build an extrapolation table from nested trapezoidal sums and explain when cancelling a presumed leading error makes a poor estimate look precise.

Cancelling a nuisance coefficient

Suppose a smooth integrand gives

$$T(h) = I + ah^2 + bh^4 + O(h^6).$$

Halving the mesh gives

$$T(h/2) = I + ah^2/4 + bh^4/16 + O(h^6).$$

Multiply the second formula by four, subtract the first and divide by three. The unknown a disappears without being estimated explicitly. The new leading error is $-bh^4/4$.

At level j a smooth even-power expansion starts with a power h^{2j} to be removed from the preceding column. The same two-row elimination yields

$$R_{k,0} = T(2^{-k}h_0),$$

$$R_{k,j} = R_{k,j-1} + \frac{R_{k,j-1} - R_{k-1,j-1}}{4^j - 1}.$$

Each refinement reuses all previous function values. Add only the midpoints of old panels, because every old endpoint becomes an even-indexed fine node. This is especially attractive when obtaining a sample means running an expensive smooth simulation.

Background · Richardson extrapolation

$$I \approx I_{h/2} + (I_{h/2} - I_h)/(2^p - 1)$$

This cancels an assumed leading error proportional to h to power p. Several refinements can test whether that asymptotic regime is present.

Check the arithmetic twice

For $\int_0^1 x^4\,dx = 1/5$, the first three trapezoidal estimates are 0.5, 0.28125 and 0.220703125. The first extrapolation column contains 0.2083333 and 0.20052083. Combining them with coefficient 15 gives 0.2. This is an exact polynomial check on indexing and column orientation.

For $\int_0^1 \sqrt{x}\,dx = 2/3$, the first two trapezoidal values are 0.5 and approximately 0.603553. Their extrapolate is about 0.63807. The derivative is singular at zero. The endpoint panel alone has an error

proportional to $h^{3/2}$, the same scale that dominates the global error. A factor of four cannot cancel it.

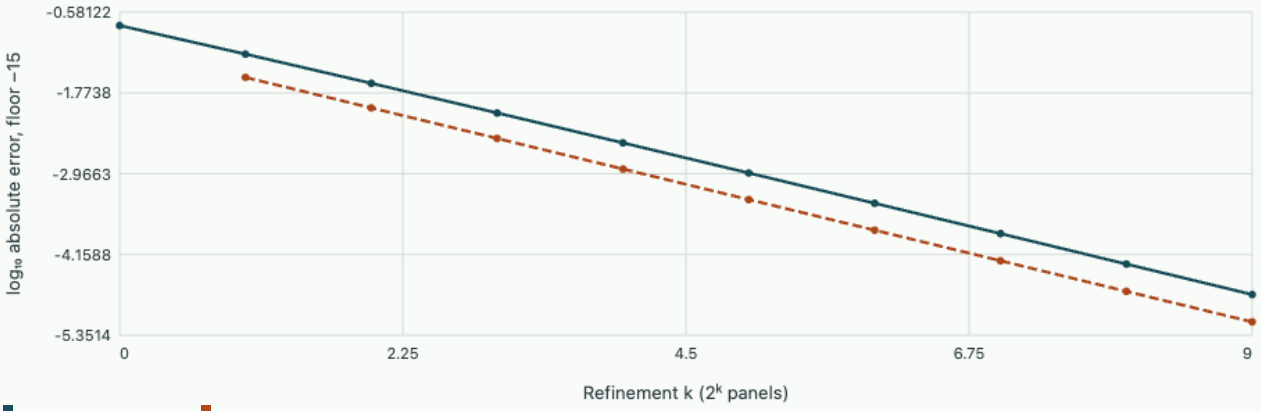
Change the regularity at the endpoint

Integrate x^p on the unit interval. Read absolute error for the raw trapezoidal estimates and their first extrapolation. The ordinate is the base-ten logarithm, with roundoff-level results displayed at the stated floor. Compare a smooth integer exponent and a fractional one.

Open this experiment online

Try the calculation

Power p **0.5**



Exact integral 0.66667. At 512 panels, trapezoidal error 0.000017785 and extrapolated error 0.0000070076.

Tutorial on interpreting a convergence table

Given three successive estimates with differences 0.008 and 0.002, what observed order do they suggest on uniform bisection?

▼ **Worked solution**

A factor of four between signed differences is consistent with $2^p = 4$, hence order two. A further refinement tests whether this is an asymptotic regime.

Remove the square-root difficulty by a change of variable. What estimate becomes exact after one extrapolation?

▼ **Worked solution**

Set $x = t^2$, $dx = 2t\,dt$. The integral becomes $\int_0^1 2t^2\,dt$. The raw trapezoidal error for a quadratic has only an h^2 term, so first-column Richardson extrapolation is exact in exact arithmetic.

Read the university treatment

Work through their extrapolation example, then determine the observed order before extrapolating the square-root data in this chapter. Driscoll and Braun · Numerical integration

Chapter 16 · Summing a Chebyshev series by working backwards

Derive a stable evaluation scheme directly from a three-term polynomial identity and diagnose the failure that occurs when a fitted law is used outside its calibrated interval.

From cosines to a polynomial recurrence

Set $x = \cos\theta$ and $T_k(x) = \cos(k\theta)$. The angle-addition formula gives

$$T_{k+1} = 2xT_k - T_{k-1},$$

with $T_0 = 1, T_1 = x$. Consider $p = \sum_{k=0}^{n} c_k T_k$ with the constant coefficient included in full. We could generate all the polynomials forward. Clenshaw evaluation instead gathers their weighted coefficients from the high end, without converting to monomials.

Set $b_{n+1} = b_{n+2} = 0$ and $b_k = c_k + 2xb_{k+1} - b_{k+2}$ for $k = n, \dots, 1$. Substitute $c_k = b_k - 2xb_{k+1} + b_{k+2}$ into the original sum. Every interior b_j multiplies $T_j - 2xT_{j-1} + T_{j-2} = 0$. Only two boundary terms survive:

$$p(x) = c_0 + xb_1 - b_2.$$

The derivation also identifies a common convention error. If the source defines its zeroth coefficient with a half-weight then it must be halved once before using this formula.

Two useful arithmetic checks

For coefficients $(1, 2, 3)$ at $x = 0.5$, $b_2 = 3$, $b_1 = 2 + 2(0.5)3 = 5$. The answer is

$$1 + 0.5(5) - 3 = 0.5.$$

Independently $T_2(0.5) = -0.5$, giving $1 + 1 - 1.5 = 0.5$.

On the calibration interval each $|T_k| \le 1$, so an isolated coefficient error of 0.001 changes the value by at most 0.001. At $x = 1.2$, $T_{10}(1.2) = \cosh(10\,\operatorname{arcosh}(1.2))$ is around 252. The same tiny coefficient error now changes a prediction by roughly a quarter of a unit. Accurate summation does not make extrapolation reliable.

Follow one uncertain coefficient

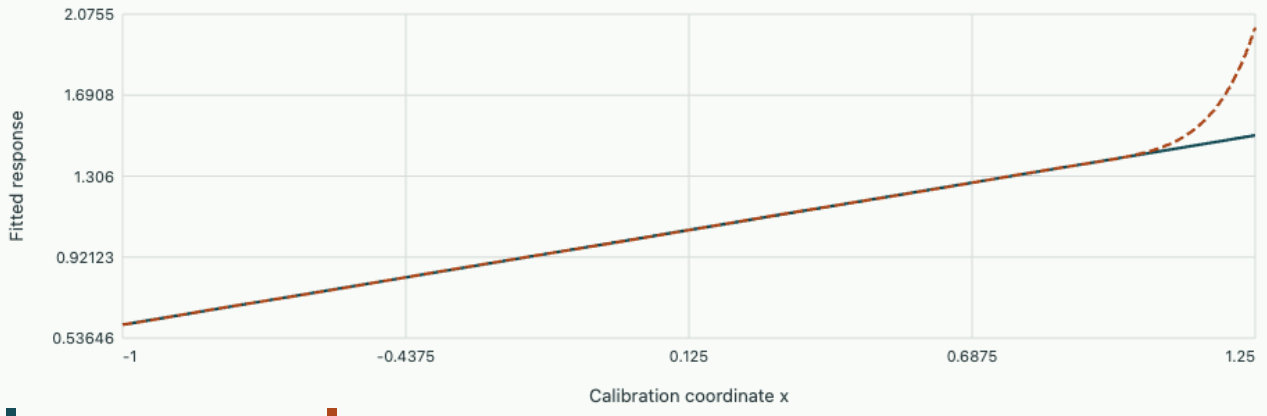
A simple calibration $1 + 0.4x$ has an uncertain extra coefficient at a selectable order. Find the extreme differences inside the interval, then inspect the growth just beyond the right endpoint. The calculation uses the backward recurrence at every position.

Open this experiment online

Try the calculation

Order of uncertain coefficient 10

Coefficient error 0.001



Difference at x = 0.8: 0.0009885. At x = 1.2: 0.25227.

Tutorial on mapping a laboratory interval

A calibration covers $10 \le T \le 50$ °C. Which coordinate should enter the recurrence and what does $T = 54$ correspond to?

▼ **Worked solution**

Centre at 30 and divide by the half-width 20, so $x = (T - 30)/20$. At 54 °C this gives 1.2, already outside the bounded Chebyshev interval.

Give a computable uniform error bound if three stored coefficients have errors no larger than 0.002 each.

▼ **Worked solution**

For $|x| \le 1$, the triangle inequality and $|T_k| \le 1$ give 0.006. Exact values may cancel at some positions; this bound does not rely on such cancellation.

Read the university treatment

Compare coefficient decay inside the approximation interval with sensitivity to extrapolation beyond it. Lloyd N. Trefethen · Approximation Theory and Approximation Practice

Study the full derivation

Integrating cosine interpolants on nested grids · Function approximation and numerical integration

Chebyshev points turn polynomial interpolation into a cosine series. Derive the exact moments of each basis polynomial, explain aliasing and find a case where two consecutive quadrature grids confidently agree on the wrong answer.

Chapter 17 · Where should a random quadrature spend its samples?

Derive the weight that corrects a nonuniform sampling density and see why an unbiased quadrature can nevertheless have infinite variance.

Change the experiment while preserving its mean

For $I = \int_0^1 f(x)\,dx$, choose a probability density $p(x) > 0$ wherever $f \neq 0$. If X is drawn from p, then averaging $W = f(X)/p(X)$ gives

$$E[W] = \int (f/p)p = I.$$

With independent draws and finite second moment, the estimator variance is $[\int f^2/p - I^2]/N$. Merely changing the point distribution without changing the weight would compute a different integral.

For nonnegative f, Cauchy–Schwarz gives

$$(\int f)^2 \leq (\int f^2/p)(\int p).$$

Equality holds if p is proportional to f. The ideal distribution would give identical weights, although constructing its exact normaliser frequently requires already knowing the answer. It remains a guide to choosing an approximate density.

Background · Monte Carlo standard error

$$\mathrm{SE}(\bar{f}) = \sigma_f/\sqrt{N}$$

Independent samples with finite variance give this ideal error scale. Importance weights and correlations can substantially alter the variance of the actual estimator.

A finite expectation with an infinite error bar

For $f(x) = x^{-1/2}$ the integral is 2. Uniform sampling produces weights $X^{-1/2}$ whose second moment is

$$\int_0^1 x^{-1}\,dx = \infty.$$

The estimate has the right expected value, but its usual finite-variance standard error has no foundation.

Take $p_a(x) = (a+1)x^a$ with $a > -1$. An inverse-CDF draw is $X = U^{1/(a+1)}$ for a uniform U. For $f = x^b$, the second moment is $1/[(a+1)(2b-a+1)]$ when $2b - a > -1$. At $b = a = -1/2$, $X = U^2$ and every weight equals exactly two. At $b = -1/2, a = -1/4$, the second moment is

$$1/(0.75 \times 0.25) = 5.3333.$$

The variance is 1.3333 and 100 independent draws have standard deviation about 0.1155.

Compare sampling plans on the same draws

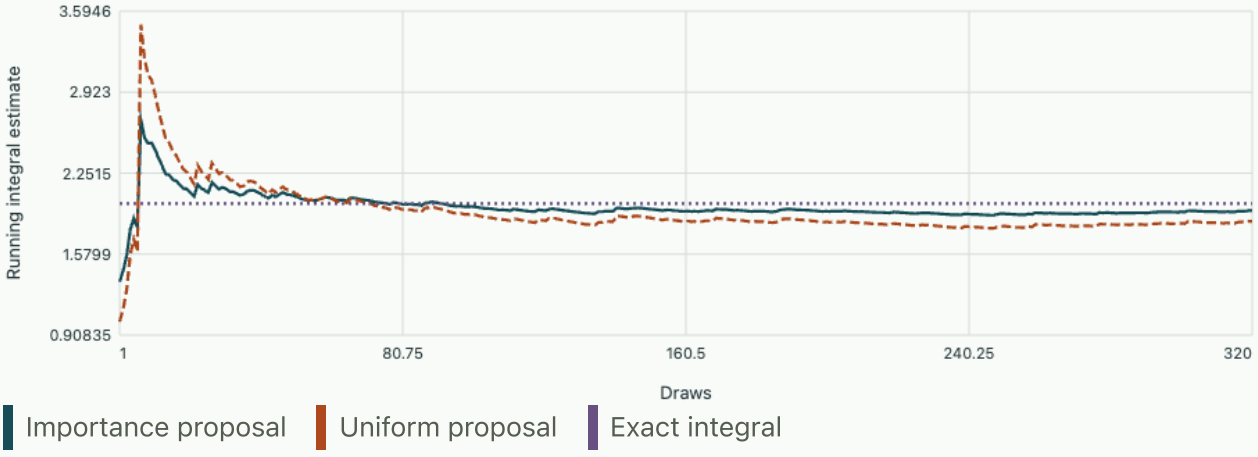
Use the same fixed pseudo-random uniform sequence for the two plans so changes reflect the transformation. Sweep the proposal exponent through the finite-variance boundary and inspect the running mean. One finite trajectory can look benign even when the theoretical second moment diverges.

Open this experiment online

Try the calculation

Integrand exponent b **-0.5**

Proposal exponent a **-0.25**



Exact integral 2. Theoretical single-weight variance of the selected proposal: 1.3333.

Tutorial on signs and diagnostics

If f changes sign, can sampling proportional to $|f|$ remove every fluctuation?

▼ Worked solution

Then the weight is a constant magnitude multiplied by the sign of f. Its variance vanishes only if that sign is constant almost everywhere. Cancellation between positive and negative contributions can still make relative accuracy poor.

A proposal has zero density on a short interval where f is nonzero. Does increasing the number of samples repair the estimate?

▼ **Worked solution**

No sample can reach that interval, so the corresponding integral contribution is absent. This is a bias from missing support; the derivation of $E[W] = I$ fails there.

Study the full derivation

Spend random samples where they reduce uncertainty · Function approximation and numerical integration

A Monte Carlo estimate can be unbiased and still too noisy to be useful. Derive control variates, allocate a fixed budget across strata, and distinguish variance reduction from an underestimated standard error.

Chapter 18 · Integrating an infinite endpoint with a finite table of samples

A density rises without bound near the edge of a specimen, yet the total inferred amount can be finite. Choose nodes that measure the edge contribution accurately. Derive the tanh–sinh transformation, split two different quadrature errors and detect when refining a grid can no longer answer the physical question.

Determine whether a finite measurement exists

Let x be the distance from the edge divided by the specimen width. Take the idealised density $f(x) = x^{-\alpha}$ on the open interval between zero and one. The exponent is a known input in this experiment. Start with $0 < \alpha < 1$. If the instrument samples exactly at zero, its readout would be undefined in this model. That does not make the integrated signal infinite. The contribution of the first ε of the interval is found by integrating from a positive lower bound and taking that bound to zero.

$$I = \frac{1}{1-\alpha},$$

$$I_{[0,\varepsilon]} = \frac{\varepsilon^{1-\alpha}}{1-\alpha}.$$

This sets the physical scale of the numerical task. With $\alpha = 1/2$, the first one ten-thousandth of the spatial window contributes one per cent of the complete signal. For $\alpha = 0.9$, the same strip contributes almost forty per cent. An instrument with finite edge resolution may require a separate sub-resolution material model. In the numerical exercise the stated power law is exact down to arbitrarily small distance. That assumption permits an independent reference answer and lets us attribute a mismatch to the quadrature rather than to unseen edge physics.

A uniform midpoint rule avoids the forbidden endpoint, using $h \sum_{j=0}^{N-1} f((j+1/2)h)$ with $h = 1/N$. The exact contribution of the first cell is $h^{1-\alpha}/(1-\alpha)$. Its midpoint approximation is $2^{\alpha} h^{1-\alpha}$. The leading error already visible in this cell decays slowly when the exponent is close to one. Since the density is convex, every midpoint panel underestimates its integral. The errors do not cancel the missing first-cell contribution. The usual smooth-function bound in terms of a uniform second derivative cannot be applied here.

Build a transformation that resolves both ends

Map a real parameter t onto the physical interval. A convenient first choice is a logistic map, which approaches either endpoint exponentially in the parameter. To concentrate the nodes more aggressively near the endpoints, first replace that parameter by $\pi \sinh t$. Written using a hyperbolic tangent, the resulting map is the usual tanh–sinh transformation. Its derivative becomes small at precisely those points where the original density becomes large.

$$x(t) = \frac{1}{1+\exp[-\pi \sinh(t)]} = \frac{1+\tanh[(\pi/2)\sinh(t)]}{2},$$

$$x'(t) = \pi \cosh(t)\, x(t)[1-x(t)].$$

Apply the change of variables before choosing any quadrature rule. The new integrand is $g(t) = \pi \cosh(t) x(t)^{1-\alpha}[1 - x(t)]$, integrated along the entire real line. To inspect its left tail take $t = -u$ with u large. Then $\sinh u$ and $\cosh u$ both grow approximately as $e^u/2$, while $x(-u)$ behaves as $\exp[-(\pi/2)e^u]$. Their product decays like a factor proportional to e^u multiplied by $\exp[-(1-\alpha)(\pi/2)e^u]$. The second exponential dominates eventually. As α approaches one its coefficient shrinks and the relevant left tail extends farther out.

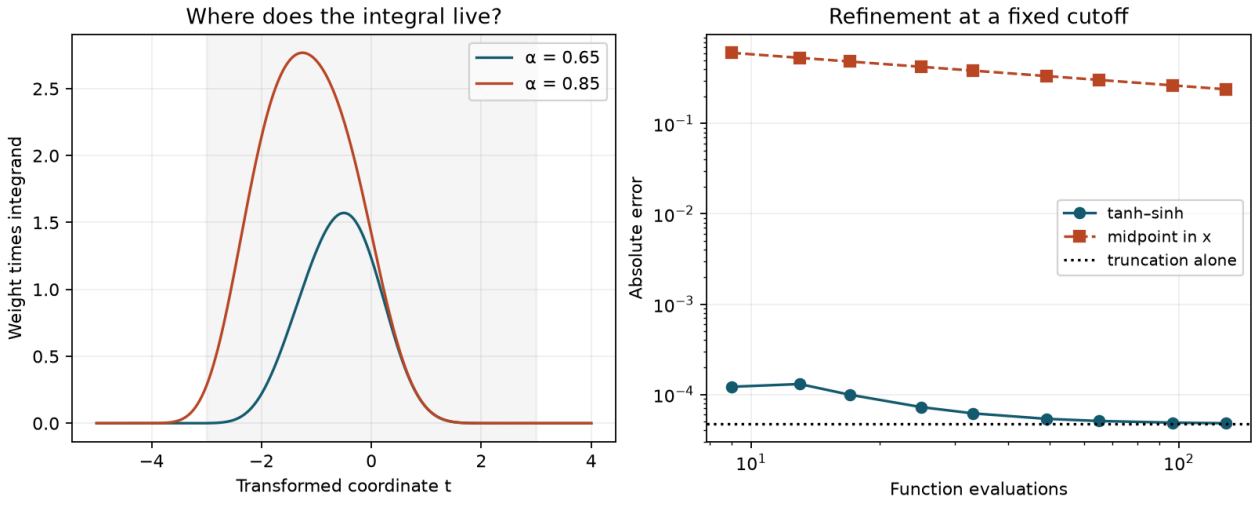
At the right end $x(t)$ approaches one and the factor $1 - x(t)$ supplies the rapid decay. A symmetric truncation in t therefore covers two generally unequal tail contributions. On a real density that is singular at both ends, repeat the asymptotic argument separately for each endpoint. For a strong interior peak the present transformation has no reason to concentrate near the peak. Splitting the interval at a known interior singularity would give each transformed subproblem a meaningful endpoint.

Distinguish tail loss from the panel approximation

Choose a finite half-window T, then a step h_t such that a whole number of panels fits on either side of zero. The notebook applies the composite trapezoid with half weights on both boundary nodes of $[-T, T]$. This calculation leaves out the two infinite tails even if the finite interval is integrated perfectly. There is a second error from replacing the finite integral by sampled trapezoids. At fixed window the panel errors can become tiny while an appreciable area is still missing.

$$I - Q_{T,h_t} = \int_{-\infty}^{-T} g(t)\,dt + \int_{T}^{\infty} g(t)\,dt + \left[\int_{-T}^{T} g(t)\,dt - Q_{T,h_t}\right].$$

In the power-law test the omitted tails can themselves be integrated explicitly. Write $a = x(-T)$ and $b = x(T)$. Their sum is $[a^{1-\alpha} + 1 - b^{1-\alpha}]/(1-\alpha)$. With $\alpha = 0.65, T = 3, h_t = 0.25$, there are 25 evaluations and the notebook gives about 2.85707045 against the reference 2.85714286. The omitted tail is approximately $4.70\,10^{-5}$, while the total difference is about $7.24\,10^{-5}$. The gap between these values belongs to resolution inside the chosen window. Recalculating at smaller h_t without increasing T tests only the latter.



The tail moves to the left as the singularity strengthens. At fixed transformed window a horizontal error floor remains after the middle of the curve has been resolved.

Compute the weighted integrand before rounding the endpoints

If x is stored as exactly one, the expression $1 - x$ loses a small positive number. If x underflows to zero, multiplying $f(x) = \infty$ by a derivative rounded to zero causes another failure. Both effects are artefacts of evaluation order. Here the product $f(x(t))x'(t)$ has a simple algebraic form that is finite everywhere at finite t. Working in logarithms keeps the two endpoint distances separately.

Put $s = \pi \sinh(t)$. A stable evaluation of $\log x$ is $-\log(1 + e^{-s})$, and that of $\log(1 - x)$ is $-\log(1 + e^{s})$. The notebook calls NumPy's log-add-exp operation on zero and the respective exponent to avoid directly forming a very large exponential. The log of the complete integrand is now the sum of $\log \pi + \log \cosh t$, $(1 - \alpha) \log x$ and $\log(1 - x)$. Exponentiating this final quantity may underflow in sufficiently distant tails. There that represents a negligible weighted contribution, rather than the disappearance of the physical interval itself.

In an integral supplied by a black-box material law the algebraic combination may be less accessible. If the law can accept a distance and the logarithm of a distance, the same strategy works. If measurements stop before either endpoint, no quadrature transformation recovers information about the missing function automatically. A defensible result should record where the model was extrapolated as well as how the integration was performed.

Design two distinct convergence experiments

First vary the number of transformed panels at fixed half-window and plot the absolute error. Use absolute error near a zero integral because a relative percentage becomes ill-conditioned as the answer tends to zero. Next increase T while keeping the transformed step approximately fixed. More points are then added to the previously discarded tails. If changing the half-window matters and halving the step does not, the initial allocation spent too much effort in the centre. Repeat this reasoning before making claims about asymptotic convergence from only two almost identical values.

For this particular reference family there is an even simpler independent calculation: map $x = u^{1/(1-\alpha)}$ on the unit interval. The diverging density and the vanishing derivative combine into the constant $1/(1 - \alpha)$. A single properly weighted node can integrate a constant exactly. This yields a stringent check of the implementation while illustrating a general design principle. When the singularity exponent is known and the remaining factor is smooth, a tailored substitution may be the easiest route to the required integral. Tanh–sinh is especially attractive when the two endpoints need protection and their detailed local structure is not fully known.

Compare two uses of a fixed sample budget

Take 33 transformed nodes and put 16 panels on either side of the origin. A half-window of two now gives a step of one eighth. Doubling the half-window with the same number of samples doubles the spacing while extending the tails. There is no general reason this second calculation should be more accurate. It trades centre resolution for coverage. Plot the estimates against window length at fixed sample count and identify the interval where both errors are small. Move the exponent closer to one and repeat; the useful window typically shifts because the physical edge carries a larger fraction of the signal.

For a more expensive black-box integrand, retain samples when halving the transformed step at fixed cutoff. All old abscissae lie on the finer grid. The trapezoid at the finer step is the sum of half the old approximation and the new midpoint samples times the finer spacing. This reuse affects computational cost, not the mathematics of the tail correction. It supports two separate reports: how much the estimated integral changed and how many genuinely new evaluations were required to obtain that change.

A local OneSimulation comparison in the notebook uses twenty-four line elements on the retained window. Eight mapped Gaussian points per element sample the transformed density at different locations from the trapezoid. Compare that sum with the analytical finite-window integral. Agreement here isolates the numerical integration on the interval. The previously calculated tails must still be included when reporting the complete physical signal.

Open editable notebook Download .ipynb

Tutorial on endpoint mass and stopping criteria

1. An instrument cannot resolve the first millionth of the width. For $\alpha = 2/3$, what fraction of the total signal would be omitted if this strip were simply discarded?

▼ **Work through the solution**

The fraction is $\varepsilon^{1-\alpha}$. Substituting 10^{-6} and $1/3$ gives 10^{-2}, or one per cent. The full integral is three, so the omitted integral is 0.03 in the dimensionless units of the model. This error would persist even for an exact integration over the resolved remainder.

---

2. With $\alpha = 1/2$, compare the first-panel midpoint contribution to its exact value. What follows for the number of uniform panels needed merely to halve this contribution to the error?

▼ **Work through the solution**

The exact value is $2\sqrt{h}$, and the midpoint gives $\sqrt{2h}$. Their difference is $(2 - \sqrt{2})\sqrt{h}$. To halve it one must quarter h, hence quadruple the number of panels on the fixed overall interval. A second-order rule for a smooth function would suggest a much smaller increase.

---

3. Two calculations with $T = 2$ agree to ten printed digits. Does this suffice to report ten digits in the full integral? Propose an additional diagnostic for a density without a known antiderivative.

▼ **Work through the solution**

Agreement supports resolution inside the selected window. Extend both tails at comparable step and record the changes from the new bands separately. Inspect their decay and, if the model supplies an asymptotic bound, integrate that bound beyond the last sampled band. Almost equal numbers at one truncation do not constrain a contribution that neither run sampled.

Further reading · Masatake Mori, discovery and development of the double-exponential transformation

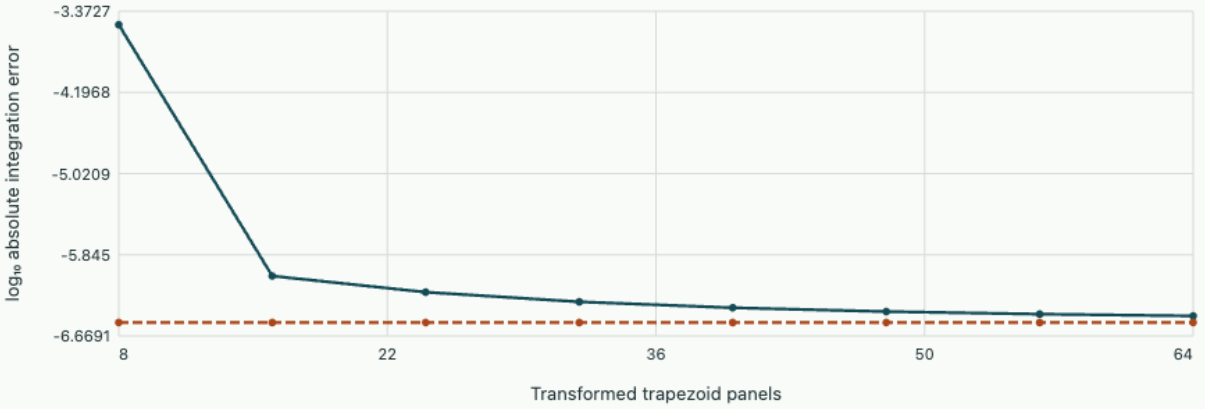
Change parameters · Integrating an infinite endpoint with a finite table of samples

Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Endpoint exponent α	0.5
Truncation |t| ≤ T	3
Finest transformed panels	64



Tanh–sinh truncation plus sampling error | Exact omitted-tail contribution

Exact integral 2; omitted tails contribute 2.9306e-7. Increasing panel count cannot remove a fixed truncation error. The transformed integrand uses logarithms to retain tiny endpoint weights.

Problem sheet 7 · Project · How many sensors measure the mean around a ring?

Design, challenge and repair a quadrature for periodic temperature measurements. Distinguish a sampled mean, reconstruction of a continuous profile and error in a reference calculation.

Predict three instrument readings

A manufactured test field is $T(\theta) = 20 + 2\cos(5\theta) + 0.8\cos(12\theta)$ °C. Determine its exact spatial mean. Predict measurements from rings with 5, 6 and 7 equally spaced sensors at zero offset. Calculate their mean using a finite geometric sum.

▼ **Work through the solution**

The continuous mean is 20 °C because both cosines integrate to zero over a full turn. For mode k set $z = e^{2\pi i k/N}$. When $z \neq 1$, the discrete average is

$$(1 - z^N)/[N(1 - z)] = 0.$$

When $z = 1$, every term is one. Taking real parts leaves only harmonics divisible by the sensor count. Five sensors see a mean of 22, six give 20.8 and seven give 20. These results describe three different instruments observing exactly the same field. A 12-sensor instrument also measures 20.8, so agreement with a coarser grid can be misleading.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Build a diagnostic from a second run

Turn the six-sensor ring by 15°. Determine which part of the signal changes in its mean. Can the average of the two runs recover the continuous mean? Test the 12th harmonic with the rotatable sensor ring. Its test harmonic has amplitude 3 °C. Compare the sign reversal and rescale its mean offset by 0.8/3 to match this instrument.

▼ **Work through the solution**

The twelfth harmonic changes sign since $12(15^\circ) = 180^\circ$. The fifth harmonic still averages to zero. The second mean is 19.2 °C; the two-run average is 20. The offset is a quarter of the old sample interval. A half-interval rotation of 30° would bring the twelfth harmonic through a full cycle and keep the erroneous mean. Choose offsets using the alias you intend to diagnose.

Use an interval integral for an independent calibration

A second sensor samples a reference profile $u(x) = x^4$ along a straight rail with scaled coordinate $0 \le x \le 1$. On one, two and four equal panels its trapezoidal values are 0.5, 0.28125 and 0.220703125. Build the first two extrapolation columns. Then change the exponent to one half in the extrapolation experiment. Explain where using the wrong error power would distort the calibration.

▼ **Work through the solution**

Two first extrapolates are

$$(4 \times 0.28125 - 0.5)/3 = 0.2083333,$$

$$(4 \times 0.220703125 - 0.28125)/3 = 0.20052083.$$

Combining them with denominator 15 gives 0.2. The square-root profile lacks a bounded derivative at the endpoint, so a full even-power error expansion cannot be assumed. The observed convergence ratio is the diagnostic in this run.

State what the specification guarantees

The design must integrate any trigonometric polynomial with modes through 12 and also reconstruct it uniquely if asked. Find sufficient uniform sensor counts for each task and explain their difference.

▼ **Work through the solution**

A 13-point ring integrates all these modes exactly because none of $1, \dots, 12$ is divisible by 13. To recover all sine and cosine coefficients through mode 12 one needs $2(12) + 1 = 25$ real degrees of freedom. A 25-point uniform ring suffices. The mean is a single linear functional and can be identifiable when the full field is not.

Explore the related calculation with adjustable inputs. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Problem sheet 8 · Integrate a rapidly oscillating signal without resolving every crest

A lock-in detector measures the in-phase integral $I(k) = \int_0^1 f(x)\cos(kx)\,dx$. The carrier frequency is known exactly from the instrument. The envelope is smooth and expensive to evaluate. Design a quadrature rule that spends those evaluations on the envelope, and find a case where even the improved method gives a deceptively small answer.

Predict a complete alias on a coarse scan

Let $f(x) = 1 + x^2$ and $k = 16\pi$. Use eight uniform panels and include both endpoints with half weight in the trapezoidal rule. What does the instrument predict, and what is the true integral?

▼ **Work through the solution**

The carrier completes exactly one full cycle per panel. Its values at every sampled point are therefore one, so the naive sum integrates only the envelope. It gives

$$1 + 1/3 + 1/(6 \cdot 8^2) = 1.3359375.$$

Integrating the oscillatory expression by parts twice gives

$$I(k) = 2\sin k/k + 2\cos k/k^2 - 2\sin k/k^3.$$

At 16π the true result is

$$2/(16\pi)^2 = 0.00079157.$$

The small true signal comes from cancellation of lobes with slightly unequal envelope weights. Nothing in the sampled carrier values reveals those lobes. A small residual between the trapezoidal estimates of a few adjacent sampled segments would be uninformative here.

Explore the related calculation with adjustable inputs. Change the parameters in the experiment on this same project page and compare its output with this part of the derivation.

Record which quantities are known before choosing a rule. If the underlying carrier were unknown and only sampled instrument voltages were available, this example would be an information problem: the samples are also consistent with an unmodulated envelope. Here the oscillator frequency and phase are controlled and available between sample positions. A numerical method can exploit that extra knowledge without evaluating the expensive part of the integrand any more frequently.

Integrate a linear envelope against the exact carrier

Approximate only the envelope on a panel of width h and midpoint m by $a + bt$, where $t = x - m$. Find a, b from endpoint samples and derive the exact panel integral of this approximation.

▼ **Work through the solution**

Use $a = (f_L + f_R)/2$, $b = (f_R - f_L)/h$, and set $d = kh/2$. Expand $\cos[k(m + t)]$ into cosine and sine products. Odd integrands integrate to zero on the symmetric interval. The constant part gives $2a\cos(km)\sin d/k$. The slope term gives $-2b\sin(km)(\sin d - d\cos d)/k^2$, after an integration by parts of $t\sin(kt)$. Add the two terms and sum over panels. This is a piecewise-linear member of the Filon family: interpolate a slowly varying amplitude and compute its oscillatory moments. It integrates every globally linear envelope exactly for every carrier frequency, up to roundoff.

Explore the related calculation with adjustable inputs. Change the parameters in the experiment on this same project page and compare its output with this part of the derivation.

The individual moment expressions require care when d is very small. In particular $\sin d - d\cos d = d^3/3 - d^5/30 + \cdots$; evaluating it by direct subtraction near zero loses significant digits. In the limit $k \to 0$ the constant moment becomes h, the odd contribution vanishes, and the rule reduces to an ordinary trapezoid for the envelope. For high frequency the same formula behaves quite differently from sampling the product. The analytic moments retain the signs and magnitudes of oscillatory contributions between the nodes.

Explain an apparently perfect zero and recover curvature

Return to eight complete periods on eight panels. The fitted linear rule now gives zero. Is this answer exact? What new measurement can recover the missing signal?

▼ **Work through the solution**

Here $d = \pi$ and $\sin(km) = 0$ on every panel. Both moment contributions vanish separately, regardless of the endpoint amplitudes. The constant and linear parts of each local envelope make no net contribution over a centred complete cosine period aligned this way. The exact quadratic curvature does contribute. Introduce one midpoint envelope measurement and fit a quadratic through the endpoints and midpoint; its exact oscillatory second moment recovers the polynomial test identically. On these aligned panels a simpler check is to compute the residual $f(x) - p_1(x) = (x - x_L)(x - x_R)$. The integral of its quadratic term supplies exactly the signal discarded by the linear rule. The linear answer is far closer in absolute terms than the aliased trapezoid, yet its relative error for the small target is still one hundred percent.

Explore the related calculation with adjustable inputs. Change the parameters in the experiment on this same project page and compare its output with this part of the derivation.

Dimensional reasoning makes the restoration plausible. For a twice differentiable amplitude the pointwise error of linear interpolation on one panel is bounded by $\max|f''|h^2/8$. Multiplying by $|\cos| \le 1$ and integrating across all panels yields a uniform absolute error bound of the same order on this unit interval. It says little about relative error when the exact integral is a small difference of much larger signed contributions. An acceptance criterion for a detector with a known noise floor should therefore include an absolute tolerance expressed in the actual output units.

Test resolution and report a useful tolerance

Vary the number of panels and sweep the carrier through adjacent frequencies in the live graph. Which measurements distinguish genuinely small coherent output from values produced by aliasing?

▼ **Work through the solution**

Compute an independent reference by integrating the polynomial test analytically. Compare absolute errors on both sides of an integer number of cycles, including the exact alignment at eight cycles. The aliased trapezoid produces peaks when sampled phases line up. The Filon result follows the reference more closely once the envelope is resolved, with the special zero at the exactly aligned coarse case explained above. Repeat after doubling panel count and changing the start phase in an actual experiment. The true in-phase projection also changes with phase, so compare against the corresponding analytic reference for that phase instead of demanding agreement between physically different projections.

Explore the related calculation with adjustable inputs. Change the parameters in the experiment on this same project page and compare its output with this part of the derivation.

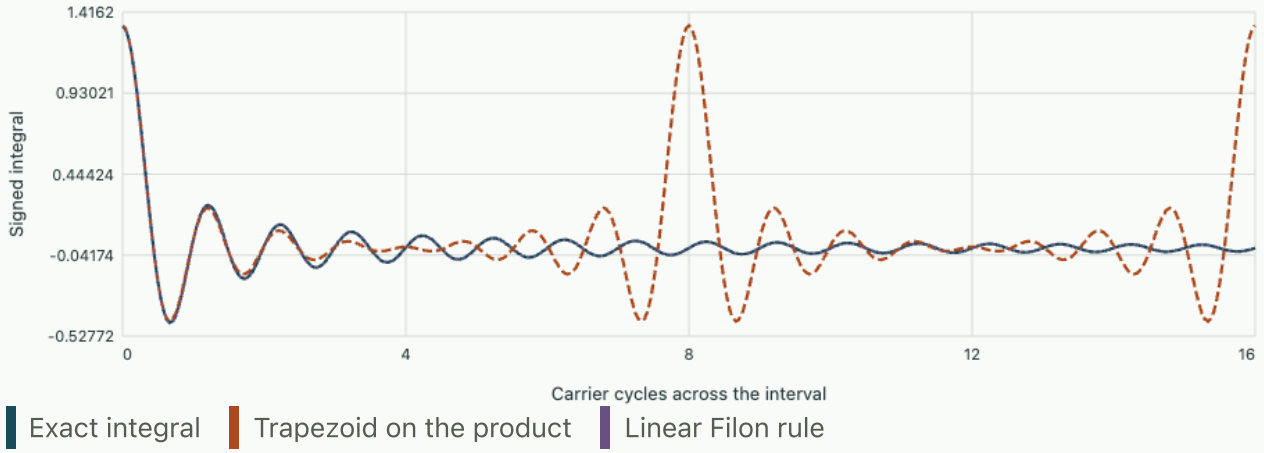
The graph displays integrals versus the number of carrier cycles, holding the number of envelope panels fixed over the sweep. Start with eight panels and include the neighbourhood of eight cycles. Then increase the panel count to sixteen. Read the signed reference and the two calculated estimates at the selected frequency. A rapidly oscillating but known carrier is a structural clue: a rule using its exact moments and an independently tested envelope approximation can avoid catastrophic sampling aliases.

Open this experiment online

Try the calculation

Envelope panels **8**

Selected carrier cycles **8**



At 8.000 cycles: exact 0.0007916, sampled-product trapezoid 1.3359375, linear Filon -0.0000000.

L. N. G. Filon's original paper on quadrature for trigonometric integrals.

Notebook practice · Integrating an infinite endpoint with a finite table of samples

Function approximation and numerical integration

Construct approximations from the information actually available. Derive their error and observe what changes when the function, data or sampling rule violates an assumption.

1. ## Approximating a resonance with a rational function

   A smooth response on a measured interval can be governed by a nearby pole. Derive a low-order Padé fit, construct a barycentric rational approximant from samples and check whether a predicted pole survives independent data.

2. ## Integrating cosine interpolants on nested grids

   Chebyshev points turn polynomial interpolation into a cosine series. Derive the exact moments of each basis polynomial, explain aliasing and find a case where two consecutive quadrature grids confidently agree on the wrong answer.

3. ## Integrate the oscillations analytically and sample the envelope

   A sensor model asks for many Fourier integrals of a slowly changing envelope. Direct trapezoids can return a full-scale signal where the true integral nearly cancels. Derive a Filon rule from exact oscillatory moments and find where it needs a different local model.

4. ## Spend random samples where they reduce uncertainty

   A Monte Carlo estimate can be unbiased and still too noisy to be useful. Derive control variates, allocate a fixed budget across strata, and distinguish variance reduction from an underestimated standard error.

5. ## How many neighbouring samples belong in a slope estimate?

   A wider local fit suppresses measurement noise, yet it can smooth away real changes in slope. Derive the competing errors, then use noisy synthetic data to find which differentiator wins at a chosen spacing.

6. ## What can smoothness buy at a fixed polynomial degree?

   Prove density for every continuous signal, derive a faster rate for smooth signals and identify the persistent overshoot when a jump violates the assumptions.

7. ## When should the denominator be fitted too?

   Develop a minimax certificate for rational approximants, find the pole of a measured response and identify the obstruction that appears when a trial denominator vanishes inside the interval.

8. ## Measuring an area integral with finitely many samples

   Build deterministic multidimensional rules from moments and geometry. Compare tensor products, symmetric triangle rules and sparse combinations at a fixed number of evaluations.

Function approximation and numerical integration · 1. Approximating a resonance with a rational function

A smooth response on a measured interval can be governed by a nearby pole. Derive a low-order Padé fit, construct a barycentric rational approximant from samples and check whether a predicted pole survives independent data.

Match series coefficients without requiring a polynomial model

Let $f(x) = c_0 + c_1 x + c_2 x^2 + \cdots$ near zero. A polynomial Taylor approximation fixes coefficients but cannot represent a finite pole. A type $[m/n]$ Padé approximant $P_m(x)/Q_n(x)$ uses polynomials of degree at most m, n, and fixes the arbitrary common scaling by $Q_n(0) = 1$ where possible. Require

$$Q_n f - P_m = O(x^{m+n+1}).$$

For powers above m the unknown numerator cannot contribute; they give linear equations $\sum_{j=0}^{n} q_j c_{k-j} = 0,\ k = m+1, \ldots, m+n$ for the denominator, with missing negative-index coefficients taken zero. Lower coefficients then give

$$p_k = \sum_{j=0}^{\min(k,n)} q_j c_{k-j}.$$

Although the desired quotient is nonlinear in its coefficients, series matching has reduced the construction to linear algebra.

For $f = e^{-x}$, matching $1 - x + x^2/2$ gives

$$[1/1] = (1 - x/2)/(1 + x/2).$$

Its expansion agrees through second degree; at $x = 1$ it gives $1/3$, compared with $e^{-1} = 0.3679$ and Taylor degree two 0.5. But the rational has a pole at -2 even though the exponential is entire, and becomes negative for $x > 2$. A good local approximation does not establish that every pole corresponds to physical resonance or every extrapolated value respects a known constraint such as positivity. The model's domain and validation data determine the legitimate interpretation.

Interpolate at selected support points while fitting other samples

When samples, rather than derivatives at one origin, are available, a barycentric form is convenient:

$$r(z) = \frac{\sum_{j=1}^{m} w_j f_j/(z - z_j)}{\sum_{j=1}^{m} w_j/(z - z_j)}.$$

At a support node use the exact stored f_j, interpreting the apparent singularity by its limit when $w_j \neq 0$. Common rescaling of all weights leaves r unchanged. For another data pair (z_i, f_i), a small cross-multiplied discrepancy requires $\sum_j w_j (f_i - f_j)/(z_i - z_j) \approx 0$. With several unsampled points this is a homogeneous least-squares problem for the weights; impose unit norm and take the right singular vector

belonging to the smallest singular value. The adaptive AAA algorithm selects an additional support point where current error is large and rebuilds this fit until a stopping criterion is met.

The cross-multiplied error and true value error differ by the rational denominator. A very small denominator can amplify a small algebraic residual, so validate the fitted quotient directly on independent points and check for poles in the intended domain. Closely paired numerator and denominator roots can almost cancel. Such a weak-residue pole–zero pair can encode numerical noise rather than a well-resolved physical mode. Testing another grid, another noise realisation and the residue is more informative than printing every computed root with equal confidence.

Compare local and global descriptions on a known response

Take $f(x) = 1/(1 + \beta x^2)$ for $\beta > 0$. Its poles are at $x = \pm i/\sqrt{\beta}$, outside the real line but at a finite distance from the origin. The even Taylor series is geometric in $-\beta x^2$. Through power x^{2m} the relative error magnitude is $(\beta x^2)^{m+1}$. It converges for $|x| < 1/\sqrt{\beta}$, fails to converge outside that radius and can show oscillatory partial values across a perfectly smooth real measurement interval. Its type $[0/2]$ rational reconstruction from $c_0 = 1, c_1 = 0, c_2 = -\beta$ is exact in noiseless arithmetic. That success rewards a model family matching the analytic structure; it does not imply unknown response functions must have just this pair of poles.

Compare a Taylor and a rational continuation

The plot holds the rational response above as known truth. Vary the pole distance through β, then change the polynomial truncation. Watch where increasing the expansion order improves agreement and where it worsens it. A noisy curvature estimate changes the fitted rational pole locations; use the readout to compare the pole error with the predicted local change. The experiment intentionally exposes the truth, which is unavailable for a real instrument and must be replaced by out-of-sample checks.

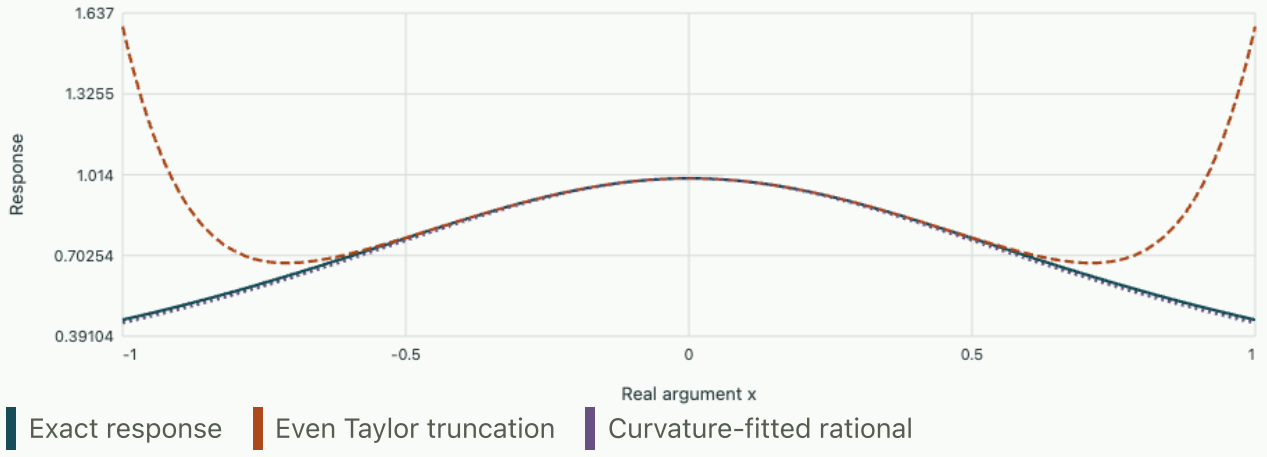
Open this experiment online

Try the calculation

Curvature β 1.2

Highest even power / 2 4

Curvature bias (%) 5



True pole distance 0.913, fitted pole distance 0.891. The Taylor convergence boundary is |x| = 0.913.

Test the assumed response law on an unused reading

A value away from the centre tests curvature without differentiating noisy samples. Under the assumed normalisation $f(0) = 1$, solving $y = (1 + \beta x^2)^{-1}$ gives

$$\beta = (1/y - 1)/x^2.$$

At $x = 1/2$, a reading $y = 10/13$ yields $\beta = 1.2$. Before including any other reading, predict $f(1) = 5/11$. A measured value of 0.50 differs by about 0.0455. More Taylor coefficients of the fitted rational function all inherit the same discrepancy. They cannot establish an independent agreement.

The derivative of the inferred parameter with respect to a measured value is $-1/(x^2 y^2)$. At the calibration point its magnitude is 6.76. An absolute measurement uncertainty of 0.005 propagates to about 0.034 in β, and then about 0.0070 in the predicted reading at one. A disagreement of 0.0455 would warrant examining calibration of the unused reading, the assumed normalisation and the model itself. Comparing two computed continuations built from the same reading would miss this test.

Tutorial · separate a useful fit from a physical pole

Infer a relaxation constant from two Taylor coefficients

A response starts $f(s) = 4 - 6s + 9s^2 + \cdots$. Construct a type $[0/1]$ fit and test the next coefficient rather than using it to determine the parameter.

▼ **Work through the solution**

The constant numerator is 4 and $4/(1 + qs) = 4 - 4qs + 4q^2s^2 + \cdots$. Matching slope gives $q = 1.5$ and predicted quadratic coefficient 9, consistent with the extra observation. The pole is $-2/3$ in the units of s. A further measured coefficient or a separate frequency interval can test the single-pole model. Agreement of three coefficients alone does not preclude additional weak modes.

Calculate a pole from two barycentric nodes

Use nodes $z_1 = 0, z_2 = 1$, samples $1, 1/2$ and weights $1, -2$. Simplify the quotient and locate its pole. What happens at the support nodes?

▼ **Work through the solution**

Multiplying numerator and denominator by $z(z - 1)$ gives numerator $(z - 1) - z = -1$ and denominator $(z - 1) - 2z = -z - 1$. Hence $r = 1/(1 + z)$, with pole at -1. The original sums are singular separately at 0 and 1, but their limits give exactly 1 and $1/2$. Direct numerical evaluation should return stored values at support nodes rather than form infinity divided by infinity.

Estimate the sensitivity of an inferred pole distance

The sampled response yields curvature coefficient $-\beta = -4.0 \pm 0.4$. Infer the distance of the conjugate poles and its linearised relative uncertainty. Would this prove the real system has that exact physical singularity?

▼ **Work through the solution**

The fitted distance is $\beta^{-1/2} = 0.50$, with relative uncertainty approximately $\frac{1}{2}(0.4/4.0) = 5\%$, or absolute 0.025. A systematic error in the assumed response shape can dominate this local parameter uncertainty. Test independently sampled values farther from zero, compare alternative rational orders, and examine whether pole positions and residues persist. Noise estimates conditional on one chosen model are not a proof of analytic structure.

Further reading

- Nakatsukasa, Sète and Trefethen · the original AAA rational approximation algorithm

- Course investigation · approximation error across an interval
- Investigate interpolation error in a heat integral

Function approximation and numerical integration · 2. Integrating cosine interpolants on nested grids

Chebyshev points turn polynomial interpolation into a cosine series. Derive the exact moments of each basis polynomial, explain aliasing and find a case where two consecutive quadrature grids confidently agree on the wrong answer.

Use a change of coordinate to simplify both the grid and the basis

The Chebyshev polynomial $T_k(x)$ is defined by $T_k(\cos\theta) = \cos(k\theta)$. For $N+1$ nodes $x_j = \cos(j\pi/N)$, sampling a degree-N polynomial in Chebyshev coordinates becomes sampling a finite cosine series at equally spaced angles. A discrete cosine transform computes the unique interpolant's coefficients without a Vandermonde solve. Clenshaw–Curtis quadrature integrates that interpolant over $[-1, 1]$. The integral is unweighted dx, so in angle coordinates there is a $\sin\theta$ Jacobian; it is not merely the trapezoidal average of the sampled values over angle.

Let $I_k = \int_{-1}^{1} T_k(x)dx = \int_0^\pi \cos(k\theta)\sin\theta\, d\theta$. Odd k gives zero by parity. Using a product-to-sum identity for even k gives $I_k = 2/(1-k^2)$, including $I_0 = 2$. With coefficients in the convention $p_N = \sum_{k=0}^{N} a_k T_k$, the quadrature is simply

$$Q_N = \sum a_k I_k.$$

Some DCT implementations store doubled endpoint coefficients. Convert to the polynomial convention before using these moments; otherwise even the constant function can integrate to four instead of two. A quick audit is exact integration of $1, x, x^2$ where the grid supports those degrees.

Explain why unsampled higher modes can imitate low modes

At sampling angles $\theta_j = j\pi/N$, $\cos[(2N-k)\theta_j] = \cos(k\theta_j)$ for integer j. Frequencies folded about multiples of $2N$ are indistinguishable on this grid. A single basis function T_{2N-k} consequently yields the same interpolant and quadrature as T_k, even though the exact integrals differ. At $N = 8$, T_{14} aliases to T_2, so $Q_8 = -2/3$ while the exact integral is

$$-2/195 \simeq -0.01026.$$

At $N = 16$ the polynomial is fully resolved and integrated exactly. Many practical smooth integrands have rapidly decaying high-mode coefficients, limiting this effect; smoothness without a quantitative resolution check is insufficient.

The error for a convergent Chebyshev expansion can be analysed as a sum of high-mode coefficients times differences between true and aliased moments. The even moments decay roughly as k^{-2}, so exact integration of many high modes contributes little, but a high mode folding to T_0 can cause a large erroneous contribution. This helps explain how a rule exact for degree only N can often perform competitively with Gaussian quadrature while still suffering carefully designed aliasing failures. Number of exact polynomial moments and actual performance on an analytic function are different comparisons.

Reuse data under refinement while looking for unresolved structure

Doubling the number of intervals gives $x_{2j}^{(2N)} = x_j^{(N)}$, so all previous samples can be reused. This is valuable when evaluating one function call involves a time-dependent model or laboratory observation. For a general interval $[a, b]$, map x affinely and multiply every integral by $(b - a)/2$. Compare successive Q_N, decay of the latter Chebyshev coefficients and an independent nonnested or local check. If the integrand has an unresolved local spike or frequency intentionally locked to the grid, two nested results can agree spuriously.

For a diagnostic let $f = T_{4N}$. Both the N- and $2N$-interval grids sample the constant value one, so both return 2 despite the exact integral $2/(1 - 16N^2)$ being small and negative. Once the $4N$-interval grid resolves the highest polynomial the answer changes dramatically. This does not imply nested algorithms should always quadruple a mesh unquestioningly. When physical bounds on derivatives, a known bandwidth or local event locations are available, use them to select an initial resolution and targeted check. Without such information no deterministic finite sampling can certify arbitrary unknown integrands.

Background · Can two consecutive nested rules agree on the wrong integral?

Use the Chebyshev polynomial $T_{4N}(x)$ on $[-1, 1]$. On the fine grid $x_j = \cos(j\pi/(2N))$, its sampled values are $\cos(2j\pi) = 1$. The coarse grid is a subset and sees exactly the same constant. Both rules consequently return the integral of one, namely 2. But the unweighted integral follows by setting $x = \cos\theta$ and integrating $\sin\theta\cos(4N\theta)$:

$$\int_{-1}^{1} T_{4N}(x)\,dx = \frac{2}{1 - (4N)^2}.$$

At $N = 4$ both sampled estimates are 2 while the actual integral is $-2/255$. An independent off-grid evaluation, or a refinement sequence chosen using prior frequency information, can expose the oscillation. This example pinpoints which assumption a difference-based stopping test needs: the current samples must already represent the significant scales.

Make two nested rules agree on a false integral

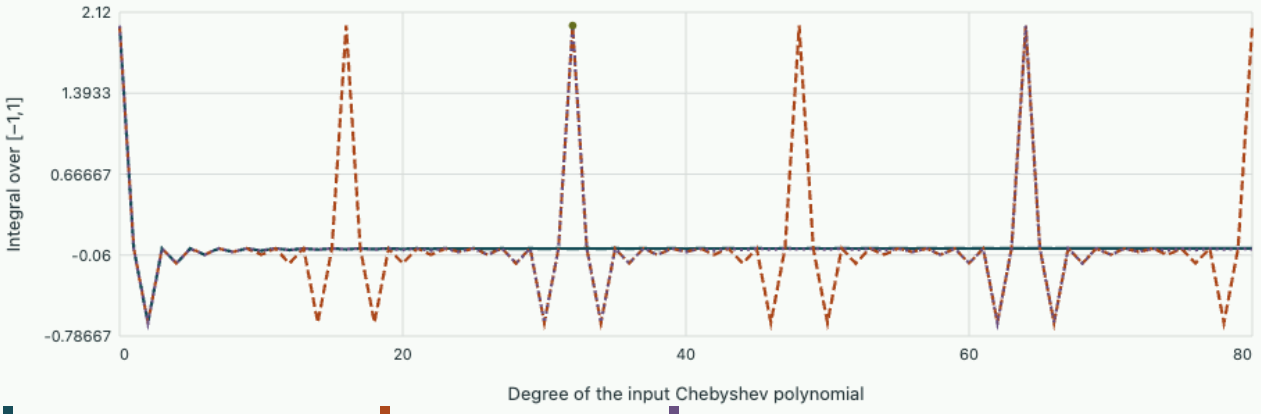
Take eight coarse intervals and choose degree 32. Read the two interpolated moments and derive the correct integral analytically. At degree 31 both rules also agree. Explain why they are now exact. At degree 16 determine which grid returns 2 and which resolves the true moment. What evidence distinguishes symmetry from a spurious agreement due to aliasing?

Open this experiment online

Try the calculation

Intervals on the coarse cosine grid **8**

Chebyshev polynomial degree to inspect **32**



Exact polynomial moment · Coarse cosine grid · Twice as many intervals · Chosen degree on coarse grid

Degree 32 integrates to -0.001955. The two nested estimates are 2 and 2.

Tutorial · inspect coefficients and stopping signals

Build a three-point rule from moments

With $N = 2$ the nodes are $1, 0, -1$. Use symmetry and moments of $1, x^2$ to obtain the weights. Integrate x^4 and explain the error.

▼ Work through the solution

Symmetry gives endpoint weights w and centre weight v. Exactness of the constant requires $2w + v = 2$; for x^2 it gives $2w = 2/3$, hence

$$w = 1/3, v = 4/3.$$

The estimate for x^4 is $2/3$ whereas the exact integral is $2/5$, an error of $4/15$. Although the rule integrates odd cubics by symmetry, polynomial degree four is not guaranteed. Expanded in Chebyshev modes, its unresolved T_4 aliases to the constant.

Derive the error caused by one small high-frequency contaminant

At $N = 8$ a signal is $1 + 0.01T_{16}$. Compare the sampled integral with the exact one and inspect a second grid with 16 intervals.

▼ **Work through the solution**

Every coarse node samples 1.01, yielding 2.02. The exact integral is

$$2 + 0.01[2/(1 - 256)] = 1.9999216.$$

The error around 0.0201 is large relative to the true high-mode integral because its samples look constant. At 16 intervals the true mode is represented, giving the exact value up to roundoff. A small-amplitude frequency component can contribute disproportionately to a coarse low-mode estimate when it aliases.

Determine whether an adaptive error diagnostic has a rigorous guarantee

Two nested rules agree to 10^{-10} and all computed high coefficients are small. Does that certify the integral for an otherwise unconstrained analytic function? Construct an explicit counterexample.

▼ **Work through the solution**

At any fixed finite N, the polynomial T_{4N} above is analytic and all coarse coefficients except the constant vanish on both grids. Adding a multiple of $T_{4N} - 1$ to an arbitrary candidate leaves both sampled grids unchanged but shifts its exact integral by nearly minus twice that multiple. Thus agreement is evidence conditional on a smoothness or scale model, not an unconditional bound over every analytic function. A genuine analytic error theorem also needs quantitative information such as a bound on an analytic continuation region and its magnitude.

Further reading

- Trefethen · Is Gauss Quadrature Better than Clenshaw–Curtis?
- Course investigation · Fourier coefficients, sampling and aliasing
- Investigate interpolation error in a heat integral

Function approximation and numerical integration · 3. Integrate the oscillations analytically and sample the envelope

A sensor model asks for many Fourier integrals of a slowly changing envelope. Direct trapezoids can return a full-scale signal where the true integral nearly cancels. Derive a Filon rule from exact oscillatory moments and find where it needs a different local model.

Estimate an oscillatory integral before choosing a grid

Consider $I(\omega) = \int_a^b f(x)e^{i\omega x}dx$ with smooth envelope f and known constant phase slope. The complex result contains cosine and sine transforms as real and imaginary parts. Direct quadrature can work if the grid resolves both envelope and oscillation. But a uniform grid with spacing $h = 2\pi/\omega$ samples the same complex phase at every point and can report a large entirely spurious response. A calculation accurate for the slow envelope alone need not be accurate for its product with the carrier. The explicit oscillating factor contains information we can integrate analytically rather than sampling it repeatedly.

Integration by parts gives

$$I = [f(x)e^{i\omega x}/(i\omega)]_a^b - (1/(i\omega))\int_a^b f'(x)e^{i\omega x}dx.$$

For bounded derivatives the integral is typically $O(1/\omega)$ at high frequency, and endpoint cancellations or endpoint vanishing can improve the decay. A result as large as $\int |f|dx$ at a very high frequency is therefore suspicious unless the phase has a stationary point, rapid structure exists in the envelope or a special physical resonance changes the setup. Absolute error is a better validation measure near an exact transform zero than relative error, which divides by a possibly vanishing number.

Derive a two-node Filon panel about its midpoint

On a panel $x = c + s,\ \ -h \le s \le h$, approximate the envelope by $A + Bs$ through its endpoints, with $A = [f(c+h) + f(c-h)]/2$ and $B = [f(c+h) - f(c-h)]/(2h)$. Retain the carrier exactly. Define $z = \omega h$. Its zeroth moment is

$$M_0 = \int_{-h}^{h} e^{i\omega s}ds = 2h\sin z/z.$$

Differentiating with respect to ω gives $\partial_\omega M_0 = iM_1$, so

$$M_1 = \int_{-h}^{h} se^{i\omega s}ds = 2ih^2\frac{\sin z - z\cos z}{z^2},$$

$$Q_{\text{panel}} = e^{i\omega c}(AM_0 + BM_1).$$

M_0 is real and even in frequency and M_1 imaginary and odd, as symmetry requires. The rule is exact for every linear envelope at every frequency, regardless of the number of oscillations in the panel. At $\omega = 0$ it returns the trapezoidal integral of the linear interpolant. Directly subtracting $\sin z - z\cos z$ for tiny z loses

precision. Instead use its series $z^3/3 - z^5/30 + \cdots$, and evaluate $\sin z/z$ by a stable sinc routine or a short series near zero.

For $f(x) = 1 + x,\ x \in [0, 1]$ and $\omega = 20\pi$, the constant contribution vanishes over ten cycles. Integration by parts on x gives

$$I = 1/(i\omega) = -i/(20\pi).$$

One linear Filon panel gives the exact $-0.0159155i$. A composite trapezoid with ten subintervals samples $e^{i\omega x} = 1$ at every node, returning 1.5 real. Both methods evaluated the same envelope but represented the carrier quite differently. Refining blindly by factors that preserve a frequency coincidence can produce misleading error estimates on the first few grids.

Separate envelope approximation from phase modelling

Since the carrier has magnitude one for real phase, $|I - Q| \le \int |f - p| dx$. A piecewise-linear interpolation error bound on a panel of width H is $\max |f''| H^2/8$, giving a global absolute bound $(b - a) \max |f''| H^2/8$ independent of ω. Oscillatory cancellation can make the actual error much smaller, but this bound already identifies a sensible grid criterion based on envelope curvature. Higher-order Filon methods integrate more polynomial moments or use derivatives at endpoints. Stable moment recurrences and consistent treatment near $\omega = 0$ matter as much as theoretical order.

For a nonlinear phase $e^{i\omega g(x)}$, changing variable to $u = g(x)$ can reduce to a linear phase on a monotone interval, with transformed envelope $f(x(u))/g'(x(u))$. At a stationary point $g' = 0$ this substitution becomes singular. In that case oscillations slow locally and the contribution can decay as $\omega^{-1/2}$ for a simple quadratic stationary point, often larger than endpoint terms. Splitting the domain and using a local quadratic-phase or stationary-phase treatment is more informative than declaring the original linear-phase rule universally efficient. Identify the physical phase law before attributing a numerical discrepancy to envelope smoothness.

Make a fast carrier invisible to the sampling grid

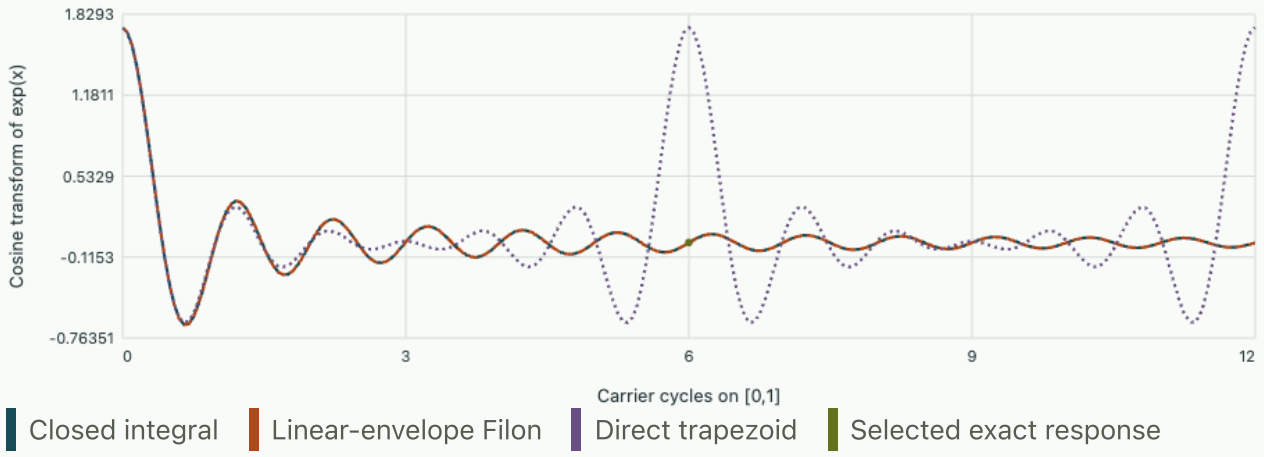
The envelope is exp(x) on the unit interval. Keep six integration panels and select six carrier cycles. Every direct trapezoidal sample sees a cosine of one. Over each full cycle both a constant and a linear envelope have zero cosine integral here. Why does the curved exponential still give 0.0012082? Change to seven panels at the same frequency and compare absolute errors. Finally set the frequency to zero: both methods integrate the same piecewise-linear envelope, so their common value can still differ from the exact integral.

Open this experiment online

Try the calculation

Uniform integration panels 6

Selected carrier (cycles in the interval) 6



At the selected frequency the exact response is 0.0012082. Filon gives -1.0674e-16 and trapezoids give 1.7223.

Tutorial · check the carrier convention and asymptotic scale

Integrate a symmetric odd envelope

For $f(x) = x$ on $[-1, 1]$ and $\omega = \pi$, compute I from the first moment. Which part vanishes?

▼ **Work through the solution**

Use $h = 1, c = 0, A = 0, B = 1$. Since $\sin \pi = 0, \cos \pi = -1$, $I = 2i/\pi$. The cosine part vanishes by parity and $x \sin(\pi x)$ is nonnegative on both sides, confirming the positive imaginary sign. Using a Fourier convention $e^{-i\omega x}$ instead would conjugate the result. A test on this odd envelope detects a sign error that any even real envelope would miss.

Give a sufficient interpolation-based error budget

On a unit interval a calibrated envelope obeys $|f''| \le 6$. How many equal panels suffice for absolute error 10^{-4} using the elementary bound? Does doubling frequency change the stated count?

▼ **Work through the solution**

The bound is $6/(8N^2)$. Requiring it at most 10^{-4} gives $N \ge \sqrt{7500}$, so 87 panels suffice. The bound makes no frequency assumption beyond a known real linear phase and does not require more panels

merely because frequency doubles. Floating-point phase reduction for extremely large arguments and uncertainty in the physical frequency are separate sources of error not included in the envelope bound.

Estimate sensitivity to a miscalibrated frequency

Show that a frequency error $\delta\omega$ can perturb the integral by at most $|\delta\omega| \int |x f(x)| dx$. Evaluate for the previous $f = 1 + x$ on $[0, 1]$ and $|\delta\omega| = 0.01$.

▼ **Work through the solution**

Differentiate under the integral to obtain $I'(\omega) = i \int x f(x) e^{i\omega x} dx$, then bound its norm and integrate over the frequency error. Here $\int_0^1 x(1 + x) dx = 5/6$, giving an absolute perturbation at most 0.00833. Relative to the small exact response at 20π, this can be substantial. An exactly integrated carrier on the wrong frequency is therefore not a high-accuracy physical prediction merely because the quadrature test is exact.

Further reading

- Iserles · numerical quadrature of highly oscillating Fourier integrals, full paper
- Course investigation · refinement and quadrature verification
- Investigate interpolation error in a heat integral

Function approximation and numerical integration · 4. Spend random samples where they reduce uncertainty

A Monte Carlo estimate can be unbiased and still too noisy to be useful. Derive control variates, allocate a fixed budget across strata, and distinguish variance reduction from an underestimated standard error.

Subtract a fluctuation with an independently known mean

To estimate $I = \mathbb{E}[f(X)]$, suppose evaluating $f(X)$ also supplies a cheap correlated quantity $g(X)$ with known exact mean μ_g. For any fixed coefficient β, $Y = f(X) - \beta[g(X) - \mu_g]$ has the same mean I. Its variance is $V_f - 2\beta C_{fg} + \beta^2 V_g$. Differentiating gives optimum $\beta_* = C_{fg}/V_g$ if $V_g > 0$ and minimum $V_f(1 - \rho_{fg}^2)$. For a crude physical surrogate it is accuracy of its fluctuations, not its absolute scale, that matters. A negatively correlated surrogate can also reduce variance by a negative coefficient. A constant control carries no information and should not be divided by zero variance.

As an exact example take $X \sim U[0,1]$, $f = X^2$, $g = X$. We know $\mu_g = 1/2$ and seek $I = 1/3$. Direct moments give

$$V_f = 1/5 - 1/9 = 4/45,\ V_g = 1/12,\ C = 1/4 - 1/6 = 1/12.$$

Hence $\beta_* = 1$ and residual variance $4/45 - 1/12 = 1/180$, sixteen times smaller than the original. At the same sample count the standard deviation falls by four. Each controlled sample $X^2 - X + 1/2$ is not equal to the true answer, but it fluctuates much less. An independent pilot run can estimate β for unknown covariances. Reusing the same small sample to optimise β and estimate the mean can introduce finite-sample bias and overstate achieved variance reduction, especially when many controls are fitted.

Partition the input space and allocate observations

Divide the domain into disjoint strata with known probabilities w_j, then draw n_j independent observations within each. The stratified estimator is $\hat{I} = \sum_j w_j \bar{f}_j$ with variance $\sum_j w_j^2 \sigma_j^2 / n_j$, where σ_j^2 is the conditional variance. Correct weighting is crucial: the allocation probabilities need not equal the target distribution's stratum probabilities. At a fixed total sample count $N = \sum n_j$, differentiating with a Lagrange multiplier gives $n_j \propto w_j \sigma_j$, known as Neyman allocation. For unequal sample costs c_j under fixed $\sum c_j n_j$, the corresponding optimum is $n_j \propto w_j \sigma_j / \sqrt{c_j}$. Pilot uncertainty and minimum counts per stratum matter before rounding to feasible integers.

For two equal-probability strata with conditional standard deviations 1 and 4, $N = 100$ calls yields optimal counts 20 and 80. The variance becomes $(1/4)(1/20 + 16/80) = 0.0625$, compared with 0.085 from 50 each. The difference is material but smaller than one might guess from the 4:1 standard deviation ratio. Overallocating all observations to the difficult stratum would leave the other conditional mean unidentified. A deterministic stratum whose mean is known separately can be removed; one whose variance merely looked zero in a tiny pilot may still need independent sampling. For rare heavy-tail excursions, a naive pilot can miss the very fluctuations it was intended to characterise.

Choose an antithetic pair with a justified covariance

For symmetric $U \sim U[0,1]$, both U and $1-U$ have the target distribution. The pair average $Z = [f(U) + f(1-U)]/2$ is unbiased. Its variance is $[V_f + \mathrm{cov}(f(U), f(1-U))]/2$. With the same two function evaluations, two independent samples would have variance $V_f/2$. Antithetic pairing therefore helps exactly when the paired function values are negatively correlated, as with a monotone function over this scalar interval. If $f(U) = f(1-U)$ the pair is redundant; the method then doubles cost per independent value. Compute standard errors across independent pairs, not by treating both values inside each pair as independent calls.

For $f = X^2$, $Z = X^2 - X + 1/2$, algebraically the same controlled random variable encountered above. Its variance is $1/180$ per pair and it costs two evaluations. The linear control cost only one expensive f evaluation when $g = X$ was nearly free. The two methods yield identical random fluctuations per independent draw of X here, but different error per unit computation. Comparing only the algorithm's reported sample count without accounting for pairs, pilot runs and surrogate evaluations can reverse a practical recommendation.

When does the control increase the noise?

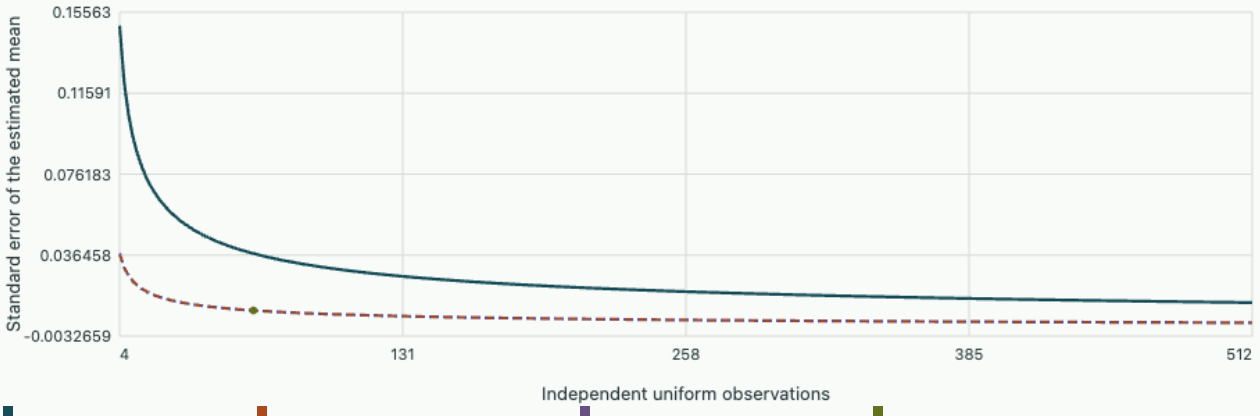
Draw X uniformly on the unit interval and estimate the mean of X². The known mean of X is one half. Compare the standard errors at coefficients zero, one and three. Use the graph to find how many direct samples match the uncertainty of 32 optimally controlled samples, then confirm the answer from the exact variance ratio.

Open this experiment online

Try the calculation

Coefficient β in X² − β(X − ½) **1**

Sample count to inspect **64**



Variance ratio controlled/direct = 0.0625. With 64 independent samples the controlled standard error is 0.0093169.

Tutorial · verify variance and bias separately

Budget calls with a measured correlation

A model response has standard deviation 6 and a validated fixed-coefficient control has correlation 0.95. Estimate calls for standard error 0.10 with and without control, neglecting control cost.

▼ **Work through the solution**

Plain sampling requires $6^2/0.1^2 = 3600$ independent calls. The controlled variance is $36(1 - 0.95^2) = 3.51$, requiring 351 independent calls. These counts assume the reported correlation and optimal coefficient apply to the production distribution. Validate with held-out paired observations. A mistaken known control mean introduces bias $\beta(\mu_{\text{used}} - \mu_{\text{true}})$, which averaging more calls cannot remove.

Find the cost-aware allocation

Two equal-probability strata have standard deviations 1 and 4. Calls in the second cost four times those in the first. Allocate 180 cost units, measuring cheap calls in units of one.

▼ **Work through the solution**

The optimum ratio $n_2/n_1 = (4/\sqrt{4})/(1/\sqrt{1}) = 2$. With $n_1 + 4n_2 = 180$ this gives

$$n_1 = 20, n_2 = 40.$$

Their variance is

$$(1/4)(1/20 + 16/40) = 0.1125.$$

Assigning half the calls to each would instead permit 36 and 36 with variance about 0.118. Assigning half the budget to each gives a different allocation again. Specify the constrained physical resource before quoting an optimum.

Identify an antithetic failure and repair the error estimate

For $f(x) = (x - 1/2)^2$, an analyst generates 100 independent antithetic pairs but computes an ordinary sample error as though all 200 values were independent. What is wrong?

▼ **Work through the solution**

Each pair consists of identical function values. There are only 100 independent observations of this response; the variance of the paired estimator is $V_f/100$, whereas the false independent count suggests $V_f/200$. Estimate uncertainty from the variance of the 100 pair averages or use 200 genuinely

independent calls if possible. Here the pairing has no cancellation because the function is symmetric under precisely the transformation used to generate the second member.

Further reading

- ETH Zürich · numerical methods for CSE, quadrature and statistical estimation
- Course investigation · statistical error in random quadrature
- Investigate interpolation error in a heat integral

Function approximation and numerical integration · 5. How many neighbouring samples belong in a slope estimate?

A wider local fit suppresses measurement noise, yet it can smooth away real changes in slope. Derive the competing errors, then use noisy synthetic data to find which differentiator wins at a chosen spacing.

Fit a polynomial in local coordinates

Suppose we observe $y_j = f(jh) + \epsilon_j$ at equally spaced times around zero. The errors are independent with mean zero and common standard deviation σ. Fit $a + bx + cx^2$ by minimizing the sum of squared differences on five points with $j = -2, -1, 0, 1, 2$. Differentiate that sum with respect to b. Because $\sum j = \sum j^3 = 0$, the odd column of the design matrix is orthogonal to both even columns. The fitted slope is therefore

$$\hat{b} = \frac{\sum j y_j}{h \sum j^2} = \frac{-2y_{-2} - y_{-1} + y_1 + 2y_2}{10h}.$$

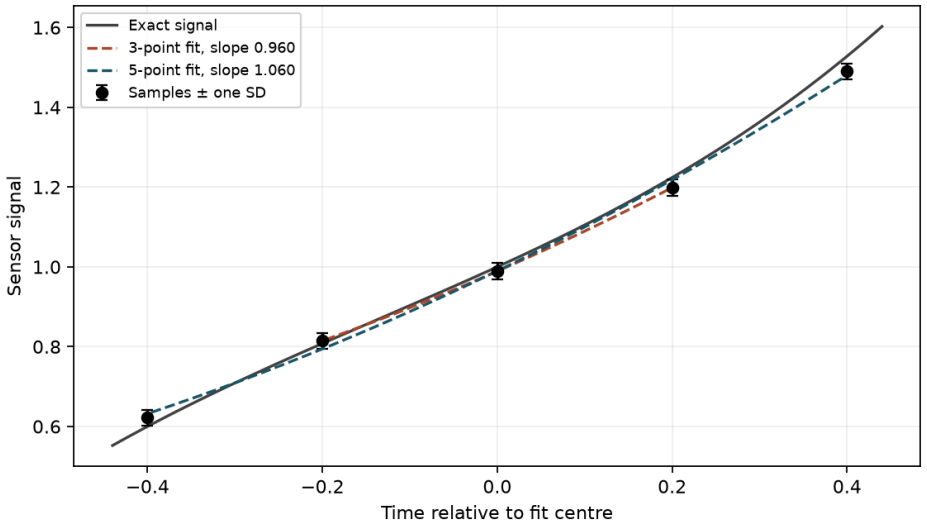
Its weights have zero sum and give exactly one on a unit linear ramp. A quadratic term cancels by symmetry. A cubic term βx^3 contributes $\beta h^2 \sum j^4 / \sum j^2 = 3.4 \beta h^2$ to the estimate. Meanwhile the independent noise variances add after weighting. Their sum is $\sigma^2 / (10h^2)$. For a three-point fit the same calculation gives a cubic bias of βh^2 and a variance of $\sigma^2 / (2h^2)$.

Write the leading bias as Bh^2 and the noise standard deviation as S/h. The mean square error around the true derivative is $B^2 h^4 + S^2 / h^2$. Its minimum is at $h = (S^2 / (2B^2))^{1/6}$ when both contributions are present. Repeated noisy measurements could reduce the noise term too. Correlated measurement errors require the covariance matrix between samples in the variance calculation.

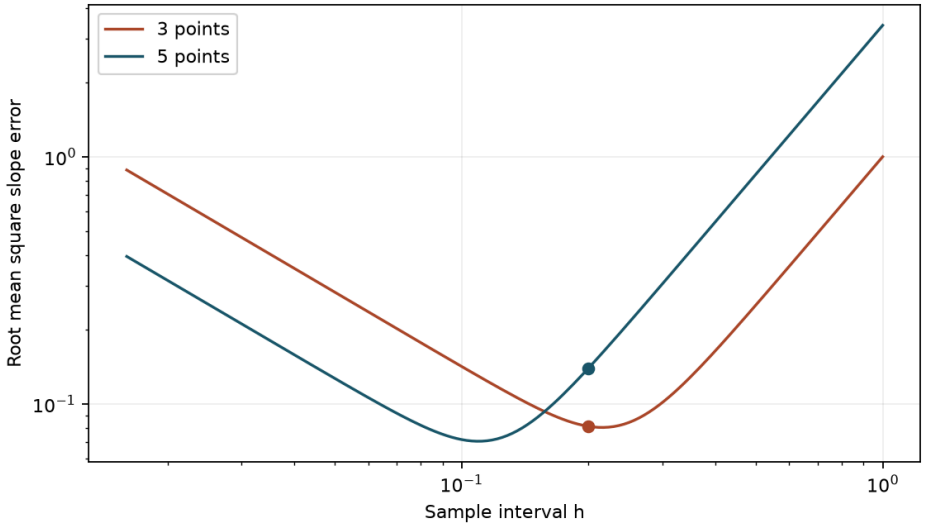
Diagnose a surprising result at two spacings

Use a signal $f(x) = 1 + x + 0.4x^2 + x^3$ and noise standard deviation 0.02. The target derivative at zero is one. At $h = 0.20$ the five-point bias is 0.136 and its noise standard deviation is about 0.0316, giving a root mean square error of 0.1396. The three-point bias is 0.040, its standard deviation is 0.0707 and its total error is 0.0812. The shorter fit wins despite its greater sensitivity to random noise.

At $h = 0.08$ the five-point values change to 0.02176 and 0.07906, giving a total of 0.0820. The three-point result is about 0.1769. Changing the spacing has exchanged the dominant source of error. A single realization can put its estimate on either side of the derivative. Change the random seed while keeping the spacing fixed to distinguish this variability from the average error in the second plot.



The three-point and five-point local fits use nested samples. Inspect the tangent at the centre, where the derivative of the underlying cubic signal equals one.



For a cubic signal the bias-plus-variance formula is exact. The location of the minimum is an experimental design choice controlled by signal curvature change and noise.

Exercise · predict a single perturbation

A five-point experiment at spacing 0.2 measures the rightmost sample 0.03 above its true value. By how much does that change the slope? What happens if the same offset appears on all five samples?

▼ **Work through the solution**

The rightmost derivative weight is $2/(10h) = 1$ in these units. The slope rises by 0.03. A common offset contributes its magnitude times the sum of all weights, which is zero. This cancellation also explains why an uncertain but constant instrument zero does not affect this derivative.

Exercise · choose a spacing when noise increases

If an otherwise identical experiment has eight times the noise standard deviation, how should the optimum spacing change? Will this scaling hold for a perfectly quadratic signal?

▼ Work through the solution

With fixed cubic coefficient B, the optimum is proportional to $\sigma^{1/3}$. Eight times the standard deviation therefore doubles the best spacing. For a quadratic the cubic bias vanishes identically. The stated model then predicts decreasing noise error as the spacing grows. In a real record the usable window can be limited by unmodelled higher derivatives, nearby boundaries or changing noise statistics.

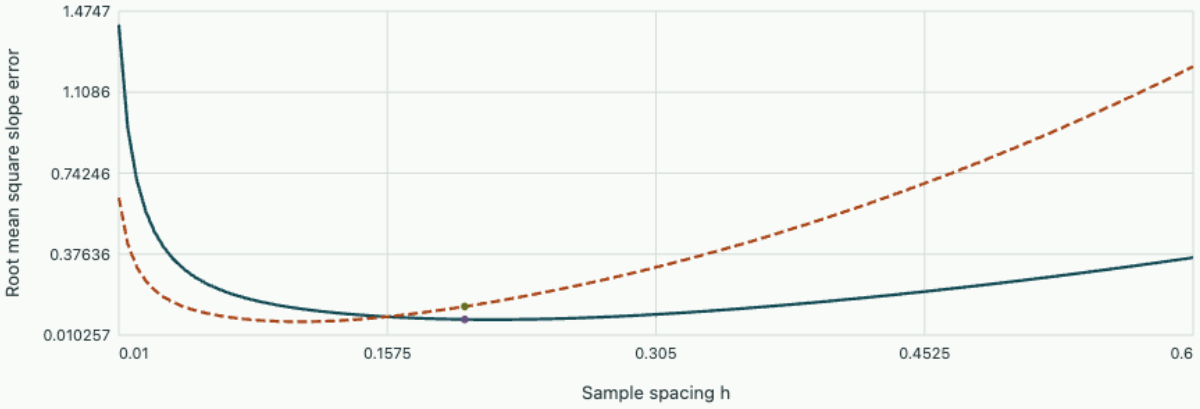
Change parameters · Explore this notebook model

Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Noise standard deviation	**0.02**
Cubic coefficient	**1**
Selected sample spacing	**0.2**



Three-point quadratic fit · Five-point quadratic fit · Selected three-point error · Selected five-point error

At h = 0.2, slope RMSE is 0.08124 for three samples and 0.13963 for five. The wider fit reduces noise but admits more cubic bias.

Open editable notebook Download .ipynb

Function approximation and numerical integration · 6. What can smoothness buy at a fixed polynomial degree?

Prove density for every continuous signal, derive a faster rate for smooth signals and identify the persistent overshoot when a jump violates the assumptions.

Construct a convergent polynomial without differentiating the data

For continuous f on $[0,1]$, let K count successes in n independent Bernoulli trials of success probability x. The expected sample $B_n f(x) = E[f(K/n)]$ is a polynomial of degree at most n. The weights sum to one and $E[K/n] = x,\ \mathrm{Var}(K/n) = x(1-x)/n$. Let $\omega_f(\delta)$ be the largest change in f between points at distance at most δ.

$$|B_n f(x) - f(x)| \le \omega_f(\delta) + 2\|f\|_\infty P(|K/n - x| > \delta) \le \omega_f(\delta) + \frac{\|f\|_\infty}{2n\delta^2}.$$

Given a target error, choose δ using uniform continuity first, then n to control the second term. The bound holds simultaneously for every x. For an L-Lipschitz signal, Cauchy–Schwarz improves the argument directly to $\|B_n f - f\|_\infty \le L/(2\sqrt{n})$. For $f(x) = x^2$, taking the second moment gives the exact excess $x(1-x)/n$. Existence of a uniformly good polynomial gives no guarantee that interpolating arbitrary samples will find it.

Relate interpolation error to best achievable error

For distinct nodes let I_n interpolate and ℓ_j be the cardinal polynomials. Its operator norm in the maximum norm is $\Lambda_n = \max_x \sum_j |\ell_j(x)|$. At a maximising point, data with the signs of the cardinal functions attain the bound and can be extended continuously between nodes. For any polynomial p of degree at most n, $I_n p = p$, so subtraction gives

$$\|f - I_n f\|_\infty \le (1 + \Lambda_n) \inf_{\deg p \le n} \|f - p\|_\infty.$$

For Chebyshev extrema the angular representation $x = \cos\theta$ makes the cardinal functions shifted trigonometric kernels. Their absolute contributions away from the nearest angular node are bounded by a constant times $1/j$, using $\sin u \ge 2u/\pi$ on $[0, \pi/2]$. The nearest term is bounded. Summing yields $O(\log n)$ growth. At equally spaced nodes the barycentric weights are proportional to $(-1)^j \binom{n}{j}$, giving very large alternating interior weights near an endpoint as degree increases. To obtain an explicit lower bound, measure distance in grid spacings and evaluate at $1/2$. Since $\prod_{i=0}^{n} |1/2 - i|/n! = \binom{2n}{n}/(2\,4^n)$, the absolute value of the jth cardinal polynomial there is this factor times $\binom{n}{j}/|j - 1/2|$. For $j = \lfloor n/2 \rfloor$, the central-binomial estimate gives growth at least proportional to $2^n/n^2$. A single cardinal function is already enough to prove exponential growth of the norm. This explains why the same best error can coexist with very different interpolation errors.

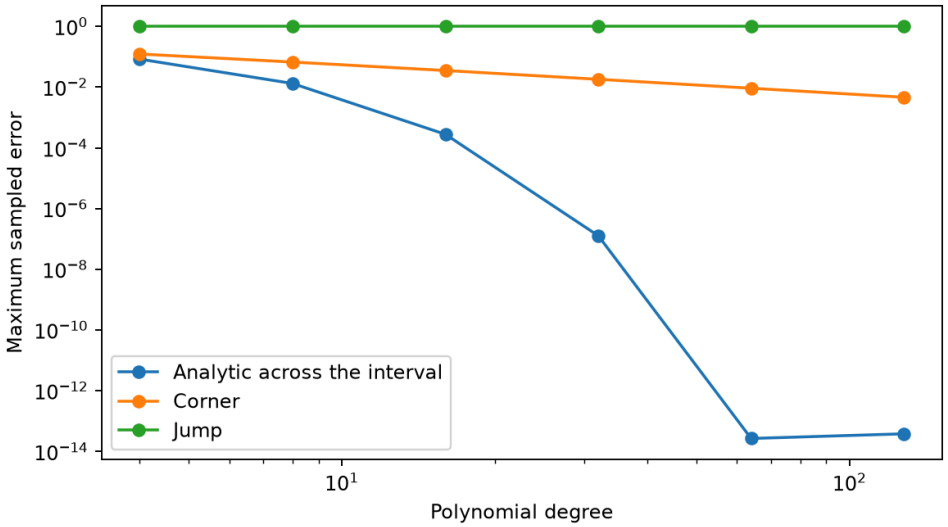
Compare an analytic continuation, a corner and a jump

Put $g(\theta) = f(\cos\theta)$. Chebyshev coefficients are cosine coefficients of an even periodic function. If the periodic g has m absolutely continuous derivatives and the next derivative is integrable, integration by parts $m+1$ times bounds its coefficients by $Ck^{-(m+1)}$. For $m \geq 1$, summing the tail gives $O(n^{-m})$. Piecewise smooth functions can be treated by integrating each smooth interval and retaining internal jump terms. For $|\cos\theta|$, continuity removes the first boundary term while the derivative jumps, giving coefficients of order k^{-2} and a tail of order $1/n$. Analytic extension to an ellipse instead gives the geometric tail developed in the course's interpolation chapter.

At a jump of height J, continuous approximants cannot have a uniform error below $|J|/2$ on a domain containing both sides. For a periodic step of height 2, its sine series involves $(4/\pi)\sin((2j+1)x)/(2j+1)$. At $x = z/(2N)$ the first N terms form a midpoint Riemann sum tending to $(2/\pi)\int_0^z \sin(t)/t\,dt$. Its first maximum, at $z = \pi$, is about 1.179. The overshoot persists while its width shrinks. A mean-square error can still decrease.

How should a convergence plot treat the jump itself?

Report the maximum over the whole interval, a second maximum excluding a fixed physical neighbourhood of the jump and an integrated squared error. Sampling at a fixed small set of points can miss the narrowing peak entirely. Increase the sampling resolution with the polynomial degree.



The same Chebyshev sampling rule resolves an analytic profile quickly, converges more slowly at a corner and retains a large maximum error near a jump. The last analytic points reveal finite-precision effects.

Tutorial on changing the order of limits

Obtain an explicit degree requirement for a bounded slope

▼ **Work through the solution**

For a target maximum error 0.01 and $L = 2$, the Lipschitz estimate guarantees $n \geq 10^4$. This is a sufficient degree. For the specific quadratic the exact worst error is $1/(4n)$, requiring only $n \geq 25$.

Predict the effect of measurement error on both meshes

▼ **Work through the solution**

Bounded nodal errors of size ϵ perturb the interpolant by at most $\Lambda_n \epsilon$. Even if the noiseless tail decreases geometrically, at fixed instrument accuracy the noise term eventually dominates. On an equally spaced mesh increasing the degree can make the prediction worse.

Trefethen · approximation, regularity and interpolation

Change parameters · What can smoothness buy at a fixed polynomial degree?

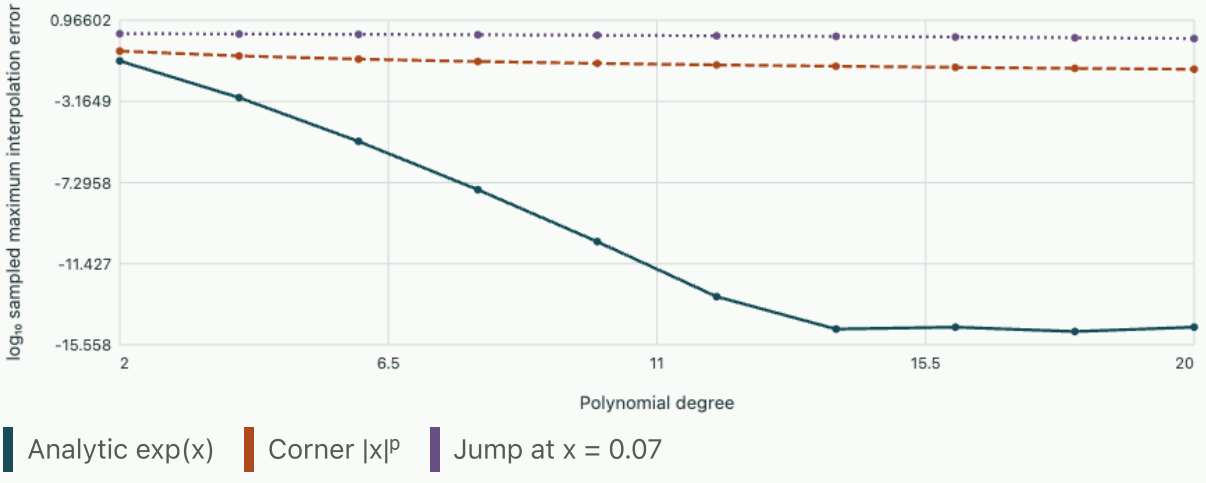
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Largest even Chebyshev degree **20**

Corner exponent p **1**



Errors are sampled on 402 fixed test locations, not certified supremum bounds. Smoothness changes the attainable rate; the jump remains incompatible with uniform convergence by continuous polynomials.

Function approximation and numerical integration · 7. When should the denominator be fitted too?

Develop a minimax certificate for rational approximants, find the pole of a measured response and identify the obstruction that appears when a trial denominator vanishes inside the interval.

Count independently adjustable coefficients

Write $r(x) = p_m(x)/q_n(x)$ on a compact real interval. Scaling numerator and denominator together leaves the response unchanged. Require a pole-free denominator of fixed sign and a normalisation, such as $q(0) = 1$ when admissible. For r, s of the same type,

$$r - s = \frac{p_m \tilde{q}_n - \tilde{p}_m q_n}{q_n \tilde{q}_n}.$$

The numerator has degree at most $m + n$. If the error of a candidate takes alternating signs with magnitude E at $m + n + 2$ ordered points and never exceeds E, any strictly better pole-free approximant would force $m + n + 1$ sign changes of the numerator. That is impossible unless the difference vanishes identically. This proves sufficiency. Common factors or unused degrees cause defects and can reduce the necessary alternation count. The certificate requires the global maximum between tested points to be checked.

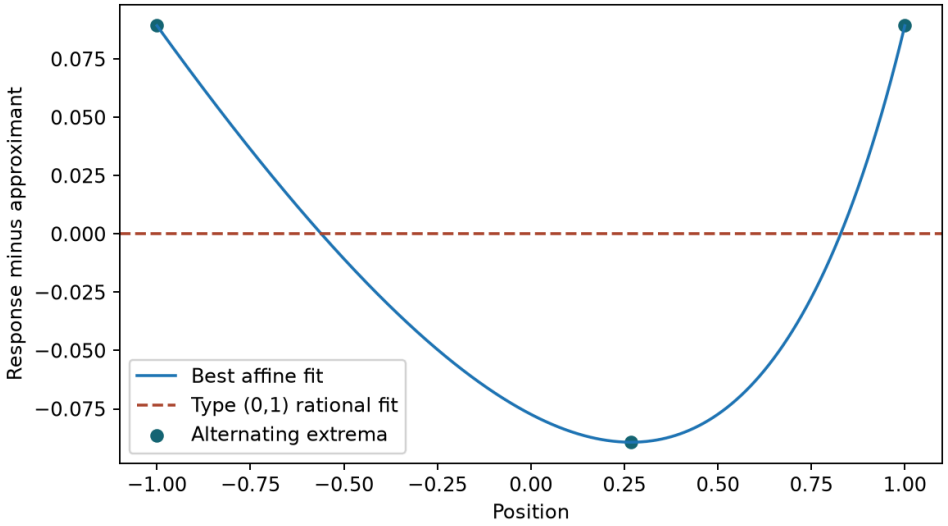
Recompute the extreme points after changing the pole

For $m = n = 1$, $q = 1 + cx$, impose $p(x_i) = [f(x_i) - (-1)^i E](1 + cx_i)$ at four proposed points. Products Ec make this nonlinear. One Remez-type iteration solves those four equations for the coefficients and E, searches between the old points for extrema of $f - r$, then exchanges points until magnitudes agree and the denominator retains its sign. Locating roots of $(f - r)'$, plus testing endpoints, is essential. A dense sample alone gives only a lower bound on the supremum.

For an explicit example approximate $f(x) = 1/(2 - x)$ on $[-1, 1]$. A linear best approximant must have the endpoint-error line as a starting shape. The endpoint secant is $s(x) = 2/3 + x/3$. Convexity makes $s - f \geq 0$, maximised where $1/(2 - x)^2 = 1/3$, hence $x = 2 - \sqrt{3}$. Set $d = s(x) - f(x) = (4 - 2\sqrt{3})/3$ there. Moving the line down by $d/2$ produces three equal alternating errors and proves optimality for degree one. The type $(0, 1)$ rational function $p = 1/2,\ q = 1 - x/2$ gives zero error. A comparison at equal parameter count has a physical interpretation when a material response genuinely has a nearby pole.

Why does matching a series at the origin give a different optimisation problem?

Padé chooses derivatives at one point. Uniform best approximation chooses the largest error over an interval. For a response with a small additional smooth background the Padé coefficients still reproduce the local series, but its largest positive and negative errors generally have unequal magnitudes. Use held-out values near the interval boundaries.



The three signed extrema certify the best affine fit. Allowing one denominator parameter puts a pole at the physical location and reproduces this response exactly.

Tutorial on poles and certificates

Give an interval-wide error bound when the fitted pole is uncertain

▼ **Work through the solution**

For $f_a = 1/(a - x)$ with $a > 1 + \epsilon$ and $|\Delta a| \le \epsilon$, subtraction gives $|f_{a+\Delta a} - f_a| \le \epsilon/[(a-1)(a-1-\epsilon)]$. A tightly fitted pole near the endpoint can still create a large uncertainty in the response.

Construct a counterexample to using a few residuals as the certificate

▼ **Work through the solution**

Samples at $-1, 0, 1$ all avoid $x = 0.3$. The function $\eta/(x - 0.3)$ has arbitrarily small values at those samples when η is small, yet unbounded interior magnitude. Verify the denominator is nonzero over the entire interval before interpreting alternation at measured points.

Trefethen · rational approximation and minimax theory

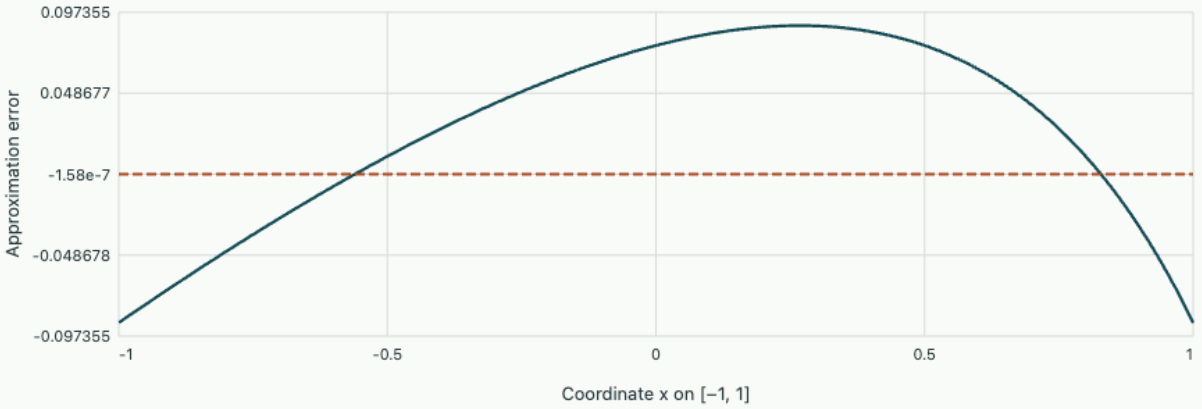
Change parameters · When should the denominator be fitted too?

Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Pole location a outside interval	2
Fitted pole error	0



Best affine approximation Rational with chosen pole

Best affine error magnitude 0.089316 with interior extremum x = 0.26795. A correctly placed single pole reproduces this target exactly; pole error introduces a structured bias.

Function approximation and numerical integration · 8. Measuring an area integral with finitely many samples

Build deterministic multidimensional rules from moments and geometry. Compare tensor products, symmetric triangle rules and sparse combinations at a fixed number of evaluations.

Which polynomials does a product rule integrate?

On a rectangle Fubini's theorem factors integrals of $x^i y^j$. Applying a one-dimensional rule in each direction forms weights $w_a w_b$ at pairs (x_a, y_b). If each factor integrates through degree $2n-1$, the tensor rule integrates every monomial with each exponent at most $2n-1$, including some of total degree above that number. For d axes the price is n^d values. Unequal smoothness can justify different orders on different axes.

On the reference triangle $x, y \geq 0,\ x + y \leq 1$, integration of the inner monomial followed by the beta integral gives

$$\int_T x^i y^j \, dA = \frac{i!\,j!}{(i+j+2)!}.$$

Seek a symmetric three-node rule at permutations of barycentric coordinates (a, b, b). Symmetry integrates affine functions once the weights each equal $1/6$. Matching x^2 gives $(a^2 + 2b^2)/6 = 1/12$ with $a + 2b = 1$. One solution uses $a = 2/3, b = 1/6$. The quadratic cross moment follows from barycentric symmetry and $(x + y + z)^2 = 1$. For a physical triangle multiply the reference weights by the absolute Jacobian determinant. For a curved mapping the determinant may itself vary across nodes.

Move the singularity into the Jacobian

Suppose a surface flux density on the same triangle is $1/\sqrt{x+y}$. With $s = x + y,\ v = y/(x+y)$, the map $x = s(1-v), y = sv$ has determinant s. Hence the total is $\int_0^1 \int_0^1 s^{1/2}\, dv\, ds = 2/3$. Substituting $s = u^2$ gives integrand $2u^2$. Two Gauss nodes along u and one along v now suffice in exact arithmetic. A Cartesian mapping $x = u, y = (1-u)v$ leaves an endpoint singularity even after multiplication by its Jacobian.

OneSimulation experiment · Integrate a singular field over a triangle

Vary the tensor order. Predict which transformed integrand is exactly integrated first.

Download the executable notebook · Python source

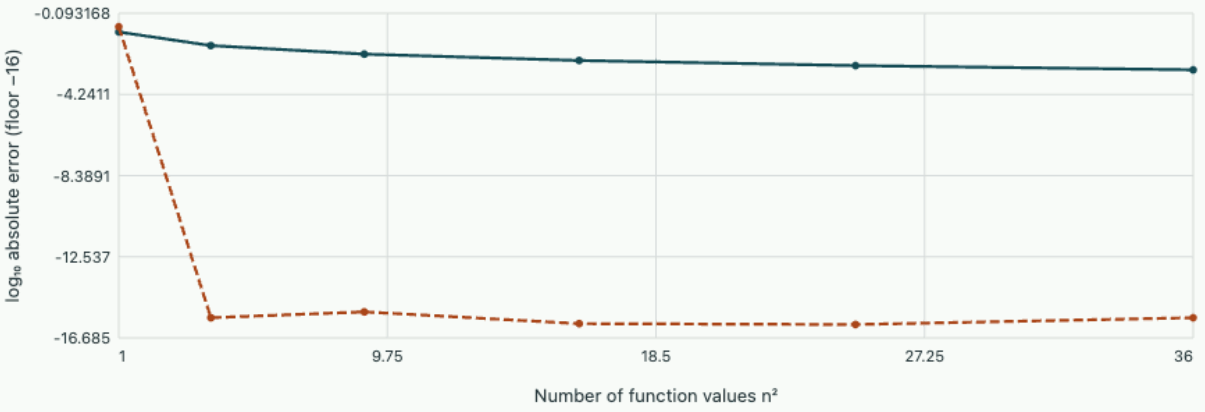
Adjust the experiment in the browser. Generalise the flux to $(x+y)^{-p}$, with $0 \leq p < 2$. Writing $s = x + y$ gives $\int_0^1 s^{1-p}\, ds = 1/(2-p)$. The squared radial map has transformed integrand $2u^{3-2p}$. Change the exponent and quadrature order, then predict when a polynomial integrand permits exact Gauss integration. In particular, increasing p past $3/2$ leaves an integrable singularity even in the radial coordinate. The browser constructs the Gauss nodes by Legendre root iteration; the notebook obtains the same rules from OneSimulation for the $p = 1/2$ reference case.

Open the live experiment

Try the calculation

Corner exponent p **0.5**

Maximum Gauss order **6**



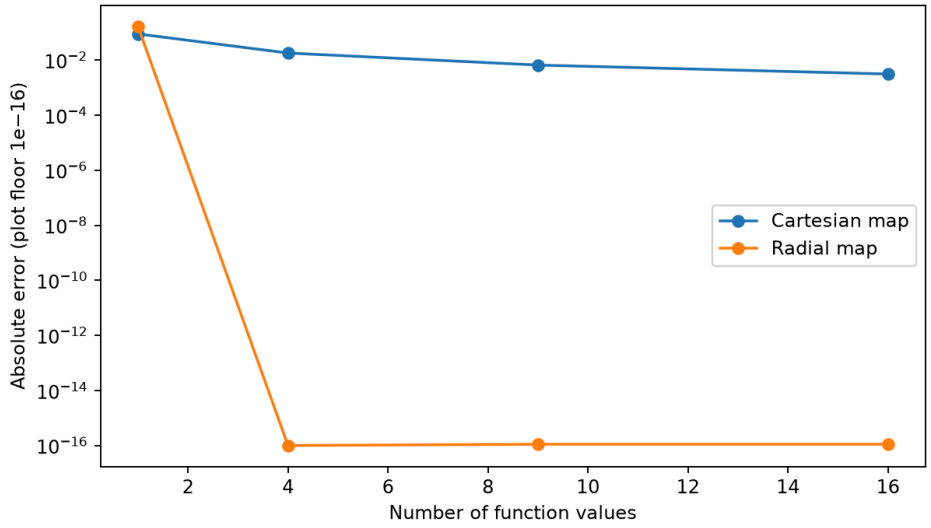
Exact integral 1/(2 – p) = 0.66667. With 36 evaluations, errors are 0.0010391 (Cartesian) and 2.2204e-16 (radial). For p = 1/2 the radial integrand is 2u², so order 2 already integrates it to rounding accuracy.

```
"""Integrate a corner singularity over the unit right triangle by two maps."""
import matplotlib.pyplot as plt
import numpy as np
from onesim.numerics.discretization.geometry.carriers.kernels.quadrature.rules.reference_quadrature
import gauss_legendre

orders = [1, 2, 3, 4]
errors = [[], []]
for order in orders:
    points, weights = np.asarray(gauss_legendre(order)).T
    nodes, weights = (points+1)/2, weights/2
    u, v = np.meshgrid(nodes, nodes, indexing="ij")
    # First map x=u, y=(1-u)v. Second map x=u²(1-v), y=u²v.
    values = [(1-u) / np.sqrt(u+(1-u)*v), 2*u**2 * np.ones_like(v)]
    for sequence, integrand in zip(errors, values):
        sequence.append(abs(weights @ integrand @ weights - 2/3))
print(f"Order {orders[-1]} errors: Cartesian map {errors[0][-1]:.6g}, radial map {errors[1][-1]:.3g}")
assert max(errors[1][1:]) < 1e-13
fig, ax = plt.subplots(figsize=(7, 4), layout="constrained")
for err, label in zip(errors, ["Cartesian map", "Radial map"]):
    ax.semilogy(np.square(orders), np.maximum(err, 1e-16), "o-", label=label)
ax.set(xlabel="Number of function values", ylabel="Absolute error (plot floor 1e-16)")
ax.legend()
fig.savefig("numerics-triangle-cubature.png", dpi=170)
```

Output from the reference run

```
Order 4 errors: Cartesian map 0.00311717, radial map 1.11e-16
```



Two transformations of the same triangle lead to different quadrature errors.

Buy mixed resolution only where it matters

Let Q_l be nested one-dimensional rules of growing order, $Q_0 = 0, \Delta_l = Q_l - Q_{l-1}$. Telescoping the full tensor rule gives a sum of products of increments. In two dimensions retain terms with $l_1 + l_2 \le q, l_i \ge 1$. This construction spends few evaluations at points fine in both directions. If mixed derivatives control the product increments, many omitted terms are small. A sharp oblique ridge can violate that premise. Refining only along individual axes then misses the ridge.

For $q = 3$, the combination is $Q_2 \otimes Q_1 + Q_1 \otimes Q_2 - Q_1 \otimes Q_1$. Take midpoint and Simpson as nested levels on $[0, 1]$. Five distinct points replace the nine in the full tensor Simpson grid. A test on $x^2 y^2$ gives $1/12 + 1/12 - 1/16 = 5/48$, compared with exact $1/9$. The difference is exactly the mixed increment omitted from this level.

How can two deterministic estimates agree for the wrong reason?

A compact nonnegative bump placed between all sampled nodes returns zero on both grids. To certify an error, use an available derivative or interval bound on each region, or change the sampling geometry independently. An observed difference is an estimator whose reliability depends on resolving the integrand.

Tutorial on domains and evaluation budgets

Transport the triangle rule to vertices (0,0), (4,0), (1,2)

▼ **Work through the solution**

The matrix mapping coordinate differences has columns $(4, 0), (1, 2)$ and determinant 8. The three weights are $8/6$. Their sum is the area 4. Compute each physical node as the same barycentric combination of the physical vertices.

Decide which quadrature to buy for three dimensions

▼ **Work through the solution**

With five levels of points per coordinate, the tensor rule costs $5^3 = 125$. If the integrand is a sum $g(x) + h(y) + k(z)$, product increments involving refinement in two directions vanish because differences annihilate constants. A sparse rule can recover the separate one-dimensional accuracy. A dependence on $x + y + z$ generally reintroduces mixed terms.

Davis and Rabinowitz · Methods of Numerical Integration

Approximating a measured field

Reconstruct a potential from weighted modes. Compare orthogonality in a rectangle, cylinder and sphere, and determine which coefficients can be recovered from noisy measurements.

1. Rectangular geometries, orthogonality and sources that drive individual modes
2. Cylindrical geometry, coaxial boundaries and Bessel spectra
3. Normalised spherical harmonics and the field of an arbitrary three-dimensional source
4. Reconstructing a source from noisy voltage measurements
5. Weighted least squares and residual diagnostics
6. How model error masquerades as a fitted physical parameter

Rectangular geometries, orthogonality and sources that drive individual modes

Electrostatic field theory and measurement

Fourier modes are solutions selected by their electrode walls. On a rectangle they turn a boundary-voltage problem into independent vertical equations. When a charge fills the rectangle the same eigenfunctions give a different denominator and a different physical response.

Determine which sign of a separation constant can fit the walls

Take a vacuum rectangle $0 < x < a,\ 0 < y < b$, invariant along z. Ground the left, right and bottom edges and impose $V(x, b) = f(x)$ at the top. If the boundary data disagree at a corner, the continuum problem can have large edge fields; we first choose smooth f vanishing at both endpoints. In the charge-free interior, a product solution $X(x)Y(y)$ gives $X''/X = -Y''/Y$, equal to a constant wherever the product is nonzero. If $X'' = k^2 X$, the only combination of hyperbolic functions zero at both $x = 0, a$ is zero. The same is true for the linear case $k = 0$. Thus $X'' = -k^2 X$. The left boundary picks a sine and the right boundary enforces $k = n\pi/a,\ n = 1, 2, \ldots$.

Then $Y'' = k^2 Y$, with the bottom condition choosing $\sinh ky$. Choosing its top value as unity gives the transfer function $Y_n = \sinh(k_n y)/\sinh(k_n b)$. Summing products,

$$V(x, y) = \sum_{n \geq 1} f_n \sin(k_n x) \frac{\sinh(k_n y)}{\sinh(k_n b)},$$

$$f_n = \frac{2}{a} \int_0^a f(x) \sin(k_n x)\, dx.$$

These coefficients follow by multiplying the top boundary by $\sin(k_m x)$ and integrating. For $n \neq m$ the product reduces to cosines with integer half-wave multiples and integral zero; for equal indices its integral is $a/2$. This orthogonality incorporates the edge boundary values. Periodically repeating a voltage function without regard to its endpoint conditions can give cosine or constant terms that violate the grounded side walls.

Track how geometry filters electrode detail

A mode's vertical derivative is $k_n \cosh(k_n y)/\sinh(k_n b)$, so the field normal to the electrode weights high modes differently from potential. At the top the magnitude multiplier is $k_n \coth(k_n b)$. For a half-space, high-frequency field content can be large very near a patterned electrode even if it disappears rapidly a short distance inside. For large $k_n b$, direct numerical evaluation of hyperbolic sines may overflow. The algebraically equivalent ratio $e^{-k_n(b-y)}[1 - e^{-2k_n y}]/[1 - e^{-2k_n b}]$ stays bounded; for very small arguments evaluate $1 - e^{-t}$ with a numerically stable exponential-difference routine.

Suppose the rectangle is square and the electrode has $f = 4\sin(\pi x/a) - 2\sin(3\pi x/a)$ V. At mid-height $y = a/2$, the fundamental transfer factor is $1/[2\cosh(\pi/2)] = 0.1993$, while the third is

$$1/[2\cosh(3\pi/2)] = 0.00898.$$

At $x = a/2$ the first sine is 1 and the third is -1, so the potential is 0.815 V approximately, compared with the 6 V drive at the same x on the electrode. Taking the simple arithmetic average of the top and bottom voltages at that x would predict 3 V and miss leakage to the grounded side walls.

If a second boundary carries a pattern, superpose the source-free solution produced by it with the other boundaries grounded. For example a nonzero bottom voltage pattern uses $\sinh[k_n(b-y)]/\sinh(k_n b)$. This works because the field equation and boundary conditions in fixed-voltage form are linear. A floating conductor imposes a different constraint coupling many modal coefficients, not the prescription of one extra arbitrary voltage profile.

Integrate a discontinuous electrode and distinguish two kinds of convergence

Let a short strip centred on the top boundary span $a/2 - w/2 < x < a/2 + w/2$ with voltage V_d, the remaining top grounded. Integrating the sine over that interval gives

$$f_n = 4V_d \sin(n\pi/2)\sin(n\pi w/(2a))/(n\pi).$$

Even modes vanish by mirror symmetry. Individual allowed coefficients decrease as $1/n$ at most because the boundary has jumps. On the boundary partial sums overshoot close to a jump. Adding terms narrows that region rather than removing its limiting fractional overshoot. At fixed interior distance the exponential transfer renders the series much better behaved.

For $w = a/3$, the first and third coefficients are $2V_d/\pi$ and $-4V_d/(3\pi)$. Along the centreline both contribute positively because their sine factors have opposite signs. A finite-width electrode cannot be captured exactly with a single spatial mode even when its interior signal far away is mainly fundamental. Recovering the electrode width from a distant scan by dividing attenuated high modes by their transfer factors will amplify measurement errors. Forward propagation and inverse reconstruction need different treatments of frequency content.

Let the volume rather than the electrode excite the modes

Ground all four edges and embed a known charge density $\rho(x, y)$. Choose double sines $\sin(m\pi x/a)\sin(n\pi y/b)$. They already vanish on all edges and are eigenfunctions of the negative Laplacian with eigenvalues $\lambda_{mn} = (m\pi/a)^2 + (n\pi/b)^2$. Project the charge using $\rho_{mn} = 4(ab)^{-1}\int_0^a \int_0^b \rho \sin(m\pi x/a)\sin(n\pi y/b)\,dy\,dx$. Applying Poisson's equation coefficient by coefficient yields

$$V_{mn} = \frac{\rho_{mn}}{\epsilon_0 \lambda_{mn}}.$$

High spatial frequencies of a volume charge produce weaker potentials by an inverse-square spectral factor. The corresponding normal field on a boundary involves an additional derivative. This polynomial filtering differs from the exponential decay of a boundary pattern measured away from its source. Both follow from the same PDE but answer different experiments.

For a manufactured test choose a square of side 10 mm with $\rho = \rho_0 \sin(\pi x/a)\sin(2\pi y/a)$. The voltage magnitude peaks at $\rho_0 a^2/(5\pi^2\epsilon_0)$. With $\rho_0 = 1.0\ \mu\mathrm{C/m^3}$ it is 0.229 V. The total charge integrates to zero, since the second sine is odd about the mid-height. The field and stored energy are nonzero. At $x = a/2, y = a/4$ the voltage is a positive maximum, consistent with a positive local source. At $x = a/2, y = 3a/4$ the same amplitude is negative.

Use an exact discrete eigenvalue for a rigorous code check

On a uniform square grid of spacing $h = a/N$, replace $-\partial_x^2$ by its centred three-point formula. Acting on the sampled sine gives eigenvalue $4h^{-2}\sin^2(m\pi h/(2a))$. The two-dimensional sum of the corresponding factors is therefore the exact discrete eigenvalue for this interior mode with zero edges. A computed solution can be compared first to $\rho_{mn}/(\epsilon_0\lambda_{mn,h})$, which tests the solver algebra, and then to the continuous $\rho_{mn}/(\epsilon_0\lambda_{mn})$, which tests spatial discretisation. Reducing the iterative residual alone cannot remove the difference between these two physical amplitudes.

At $N = 10$, modes $(1,2)$ have $a^2\lambda_h = 4N^2[\sin^2(\pi/20) + \sin^2(\pi/10)] = 47.99$, versus $5\pi^2 = 49.35$. The discrete amplitude is about 2.8% high despite solving its linear equations exactly. For fixed mode under grid halving, expansion of the sine gives an eigenvalue error proportional to h^2. At wavelengths only a few cells wide this expansion is unreliable. Try a sequence of grids with the same physical mode and compare actual amplitude ratios. Holding instead a fixed number of cells per wavelength while changing the physical wavelength tests a different convergence question.

Two transverse dimensions select a different attenuation length

Extend the rectangle into a rectangular conducting shaft $0 < x < a,\ 0 < z < c$, running to $y > 0$. Ground all four side walls, require bounded potential far down the shaft and prescribe a voltage on its mouth at $y = 0$. Separation in the cross section now gives products $\sin(m\pi x/a)\sin(n\pi z/c)$ with both $m, n \geq 1$. The axial equation is

$$Y'' = [(m\pi/a)^2 + (n\pi/c)^2]Y.$$

Each boundary coefficient therefore propagates as $e^{-\gamma_{mn} y}$,

$$\gamma_{mn} = \pi\sqrt{m^2/a^2 + n^2/c^2}.$$

A narrow second dimension suppresses penetration even when the first dimension is very broad. Approximating the shaft by a two-dimensional strip removes an entire contribution to the decay constant.

If the aperture is driven uniformly at V_0, projecting gives $f_{mn} = 16V_0/(\pi^2 mn)$ for odd m, n, zero otherwise. The ideal mouth voltage jumps at the grounded rim. Deep inside, only the lowest odd pair survives and its amplitude exceeds V_0 before the exponential is applied; the higher terms restore the actual boundary value at $y = 0$. With $a = 20$ mm and $c = 10$ mm the slowest decay length is

$$1/\gamma_{11} = 2.85\,\mathrm{mm}.$$

A two-dimensional strip ignoring the 10 mm side would instead give

$$a/\pi = 6.37\,\mathrm{mm}.$$

After 10 mm penetration the first-mode attenuation factors are 0.0298 and 0.208, almost a factor of seven apart.

A simple numerical test samples the centreline voltage at several heights and calculates adjacent log slopes. Close to the mouth many odd modes matter, while deep inside the ratio between two positions Δy apart approaches $e^{-\gamma_{11}\Delta y}$. Compare with a second geometry that swaps a, c. The two lowest-order decay lengths must agree because the eigenvalue is symmetric in the two dimensions; an implementation that uses only the nominal 'width' may fail this check. A second observation path away from the centre can sense different mode weights, even though every mode's exponential is unchanged.

Suppose the mouth instead has opposite voltages on the left and right halves, antisymmetric about $x = a/2$. Its coefficients vanish for odd m, so the $(1, 1)$ mode is absent by symmetry. The leading allowed mode can be $(2, 1)$, shortening the decay length. A centreline probe would then read zero everywhere by antisymmetry and could not fit any length at all. Move the probe to $x = a/4$ while keeping it at mid-height of the cross section. This illustrates how an excited spectrum, a geometry's complete set of eigenvalues, and an instrument's observed spectrum may differ.

Why the energy of an ideal electrode voltage step depends on edge resolution

Return to a two-dimensional strip or half-plane. To focus on a single boundary Fourier mode take $V = A\sin(kx)e^{-ky}$ in $0 < x < a,\ y > 0$, with $ka = n\pi$. Its gradients in x and y have equal prefactor kAe^{-ky}. The sum of their squared sine and cosine factors is one, so the energy per perpendicular length is

$$U_n' = \epsilon a k A^2/4.$$

For many orthogonal sine modes cross terms vanish after integration. Thus the stored field energy involves the weighted sum $\sum_n n|f_n|^2$, while the mean-square boundary voltage itself involves simply $\sum_n |f_n|^2$. The extra factor of n comes from the steep near-electrode field of short-wavelength structure.

For a finite electrode patch abruptly stepping from 0 to V_d, the allowed Fourier amplitudes typically decay as $1/n$ along nonzero subsequences. The voltage series converges in mean square, but its energy sum can contain a logarithmically diverging $\sum 1/n$. This is the local singularity of the zero-width voltage discontinuity. It does not prevent a well-defined voltage at any interior distance, but it prevents assigning a finite total energy or a cut-off-independent peak electric field to that exact sharp-edge idealisation. Mesh refinement can produce progressively higher energy estimates even as interior voltage curves appear perfectly stable.

There is a local way to see the same conclusion without a Fourier series. Consider a small half-disk of vacuum around a junction where an otherwise flat boundary jumps between voltages zero and V_d. In polar coordinates about the junction, the leading harmonic solution depends only on angle: $V = V_d\phi/\pi$ with

angle origin chosen on the grounded side. Its field magnitude is $V_d/(\pi r)$. Integrating $\epsilon E^2/2$ over a half-annulus of inner radius δ and outer radius R gives

$$U' = \epsilon V_d^2 \ln(R/\delta)/(2\pi).$$

Halving the microscopic length changes this single-edge contribution by $\epsilon V_d^2 \ln 2/(2\pi)$. The independence of R in that change is a practical test of the logarithmic asymptotic regime.

A finite insulation seam between real electrodes supplies a different boundary problem. For a deliberately smoothed transition of width δ, the spectrum falls faster above $k \sim 1/\delta$; the local singularity is cut off. One can compare a family of specified physical seam widths while separately refining each numerical grid until it resolves its own seam. Shrinking seam width along with mesh spacing instead changes the physical geometry at each refinement. Report the interior voltage, near-edge energy within a named observation region and the physical seam model separately. Each can converge at a different rate even in a technically correct Laplace solver.

A discrete residual can fall quickly while the wrong spatial scale remains

The separated sine basis makes it possible to analyse the convergence of a concrete numerical Laplace solver as well as the continuum solution. On a square grid with equal spacing, the five-point discretisation in a charge-free cell says its voltage equals the mean of its four neighbours. A Jacobi sweep sets all new values to this mean of the old grid, preserving fixed boundary values meanwhile. Subtract the exact discrete solution from the current approximation: the error evolves by the same averaging step with zero boundary values, independently of the inhomogeneous source that generated the original solution.

Insert an error pattern $e_{ij} = \sin(m\pi i/N)\sin(n\pi j/N)$. Elementary sine addition on the two x-neighbours produces $2\cos(m\pi/N)e_{ij}$; the y-neighbours produce the analogous factor. Hence one sweep multiplies this error mode by $\mu_{mn} = [\cos(m\pi/N) + \cos(n\pi/N)]/2$. For the longest $(1,1)$ mode, $\mu \simeq 1 - \pi^2/(2N^2)$. Its amplitude takes roughly $2N^2/\pi^2$ sweeps to drop by $1/e$. Doubling resolution therefore makes this relaxation about four times slower per physical longest mode even before counting the four times as many cells per sweep.

Choose $N = 64$. Then $\mu_{11} = 0.99880$. After 100 sweeps about 88.6% of that initial smooth error still remains, although the update per cell has become small. At the opposite corner of the spectrum $(N-1, N-1)$ gives nearly -1, so a checkerboard mode also decays slowly while changing sign on successive sweeps. Weighted Jacobi replaces an old value by $(1-\omega)$ times itself plus ω times the neighbour mean. At $\omega = 2/3$ that checkerboard amplification becomes approximately $-1/3$, removing short-wavelength oscillations quickly. The long smooth mode still has amplification near one. An attractive smooth colour map after a few sweeps can thus be misleading.

A multigrid correction exploits these distinct errors. First relax high-frequency errors on the fine grid. Compute the residual of the original discrete equation and restrict it to a coarser grid. A long fine-grid mode becomes easier to change there. Solve the coarse error equation, prolong its correction back to the fine grid, and smooth again. The correction must solve for error with the appropriate homogeneous boundary constraints, not repeat the original boundary excitation a second time. The residual and

prolongation should respect the same physical units and grid volumes as the original source. Otherwise a visually plausible coarse correction can inject net charge at each level.

For a manufactured check initialise the fine-grid error to a known sine mode on top of an exact discrete solution. After one weighted sweep measure the inner product of the new error with the original mode and divide by its squared norm. The observed ratio must equal $1 - \omega + \omega\mu_{mn}$ before roundoff or boundary effects. Repeat with $(1, 1)$, a mixed $(1, N - 1)$ mode, and the checkerboard-like $(N - 1, N - 1)$ mode. This directly tests the implementation's indexing, old-versus-new updates and boundary handling; it does not require the continuum equation to be discretisation exact.

Finally compare iterative and truncation errors in measurable volts. For a discrete mode, if the algebraic residual coefficient is r_{mn} in the form $-\nabla_h^2 e = r$, its error coefficient is $r_{mn}/\lambda_{mn,h}$. The same residual amplitude in a long-wavelength mode produces a larger voltage error than in a high-frequency mode. An absolute solver tolerance with no physical scaling is therefore insufficient to certify a voltage reading. Stop iterating when the solver error in the actual observable is small compared with discretisation and experimental uncertainty, then refine the spatial representation separately.

Resolve a short-wavelength electrode pattern

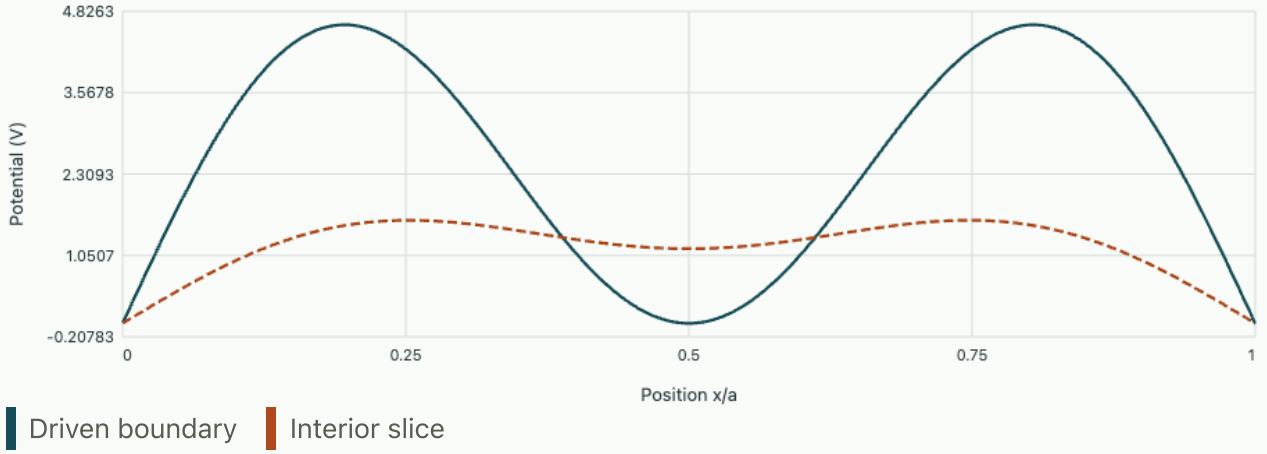
Increase the transverse mode number. Predict how far the pattern reaches into the strip before inspecting the voltage profile.

Open this experiment online

Try the calculation

Height y/a **0.18**

Third harmonic at the boundary (V) **3**



At the middle of the strip the potential is 1.15 V. The third harmonic retains 18.3% of its boundary amplitude.

Build a resistor-network analogue of the rectangular boundary problem

Use Kirchhoff measurements to check a field solver

A square lattice of nodes is connected by identical resistors R between nearest neighbours. Outer nodes are held at prescribed voltages. Derive the interior equation from charge conservation at a node. Which measured electrical quantity can test a normal field at the boundary? What changes on an anisotropic grid?

▼ **Derive and check the result**

For a steady interior node with no supplied current, the sum of currents $(V_{ij} - V_{\text{neighbour}})/R$ is zero. Rearrangement gives the same mean-of-four equation as the square-grid discrete Laplacian. Thus the resistor array reproduces the *discrete* Dirichlet solution for the same pinned boundary voltages. Its Joule dissipation $\sum_{\text{edges}} (V_i - V_j)^2/R$ is stationary against variations of the unfixed voltages, giving the same equations a second way. The numerical voltage maps can be compared node by node; the physical power and the electrostatic field energy are different dimensional quantities even though their quadratic forms are analogous.

To estimate the normal field near a flat boundary, measure the signed current from a boundary node along the edge connecting it to its nearest interior neighbour. This equals the observed voltage difference divided by R; multiplying by R/h converts it to the corresponding one-sided field magnitude in volts per metre. At sharp corners the one-sided difference has large discretisation error and the continuum physical field may be singular. Summing all currents supplied to a terminal edge gives a more robust integrated flux analogue. Account for any links between nodes at different prescribed boundary voltages if those links contribute separate current to the laboratory power supplies.

For horizontal spacing h_x and vertical spacing h_y, replacing every resistor by the same value no longer matches the isotropic Cartesian Laplacian. A finite-volume cell requires conductance proportional to face length divided by edge length: horizontal links have conductance proportional to h_y/h_x, vertical links to h_x/h_y. Their ratio is h_y^2/h_x^2. This restores a discrete conservation law with the correct physical area factors. A measured profile agreeing exactly with an equal-resistor lattice on a stretched geometry could therefore validate the wrong continuum material or geometry unless those weights were checked first.

Tutorial · design a boundary measurement and a solver benchmark

Identify a forbidden mode

A symmetric positive top electrode has its centre at $x = a/2$. A numerical sine transform reports a large $n = 2$ coefficient. Is it evidence of physical high-frequency content? Which controlled change could verify the sign of the defect?

▼ **Work through the solution**

Mirror substitution $x \mapsto a - x$ preserves the boundary voltage but changes $\sin(2\pi x/a)$ to its negative. Its exact projection is zero. Sharp electrode edges may require many odd modes; they cannot explain an even coefficient under exact symmetry. Check coordinate origin, edge alignment on the grid and measurement offsets. Shift the actual patterned patch by a small amount to one side, then the same amount to the opposite side. A genuine even coefficient changes sign under these mirrored displacements, while odd-mode amplitudes agree. The response at interior height has the same parity because the transfer factor depends on mode number alone.

Measure where electrode charge resides

For a single sinusoidal top voltage $V_d \sin(\pi x/a)$, derive the total free charge per perpendicular length on the upper electrode face. Why does it depend on the bottom distance even though the top voltage profile is the same?

▼ **Work through the solution**

At the upper metal the normal from metal into the dielectric is $-\hat{\mathbf{y}}$. With $E_y = -\partial_y V$, the metal charge density is

$$+\epsilon_0 \partial_y V = \epsilon_0 k V_d \coth(kb) \sin(kx),$$

$$k = \pi/a.$$

Integrating along $0 < x < a$ yields $2\epsilon_0 V_d \coth(\pi b/a)$ per length. Bringing the bottom closer increases $\coth(kb)$. For $kb \ll 1$ the expression tends to $\epsilon_0 b^{-1} \int_0^a f(x)dx$, the local parallel-plate limit. A segmented electrode supplies a nonuniform fixed profile; the result counts its integrated face charge and assumes the ideal voltage boundary remains prescribed in each segment.

Further reading

- Move the measurement plane above a patterned electrode
- Evaluate a grounded-enclosure boundary problem
- Griffiths, 5th ed., §3.3.1 · Cartesian separation

Cylindrical geometry, coaxial boundaries and Bessel spectra

Electrostatic field theory and measurement

The radial coordinate of a cylinder leads to logarithms in one experiment and to Bessel functions in another. The difference is set by the variation along the axis and around the circumference. Recovering the radial weight from flux clarifies both.

A conducting annulus selects a logarithmic radial potential

Put a long cylindrical metal electrode of radius a inside a coaxial grounded tube of radius b, with homogeneous dielectric of permittivity ϵ in the gap. Long means that probes are far enough from both ends for axial variation to be negligible. With no free gap charge and no angular variation, Laplace's equation reduces to $\partial_r(r\partial_r V) = 0$. One integration gives $rV' = C$, a statement of conserved radial flux through cylinders; the second gives a logarithm. With inner electrode at V_0,

$$V(r) = V_0 \frac{\ln(b/r)}{\ln(b/a)},$$

$$E_r = \frac{V_0}{r \ln(b/a)},$$

$$C' = \frac{2\pi\epsilon}{\ln(b/a)}.$$

Here C' is capacitance per length. To obtain it, integrate D_r over the inner metal face per unit axial length and divide by V_0. Another route integrates the energy density $\epsilon E_r^2/2$ times $2\pi r\,dr$, giving $C'V_0^2/2$. For $a = 2$ mm, $b = 8$ mm and $\epsilon = 2\epsilon_0$, $C' = 80.3$ pF/m. At $r = \sqrt{ab} = 4$ mm the voltage is $V_0/2$; the half-voltage radius is the geometric mean. An equally spaced radial ruler does not sample equal voltage steps.

The logarithm is not allowed at the axis of a solid charge-free cylinder because it diverges and carries line-source flux. In the annular domain the singular axis is outside the field region. The same radial function can therefore be admissible or inadmissible in two problems with very similar geometry. Before discarding a mathematical solution, check which radii actually lie in the domain and whether singular sources were explicitly prescribed.

Nonuniform wall voltages in a two-dimensional disk

Next take a full disk of radius a with prescribed side voltage $f(\phi)$, invariant in z. Seek $R(r)e^{im\phi}$. Single-valuedness after $\phi \to \phi + 2\pi$ forces integer m. The radial equation $r^2R'' + rR' - m^2R = 0$ is solved by $r^{|m|}$ and $r^{-|m|}$ for $m \neq 0$; regularity at the centre chooses the growing power. The zero mode is constant, discarding its logarithmic partner. Thus an angular Fourier component on the wall penetrates as $(r/a)^{|m|}$. At $r = a/2$, a third harmonic retains one eighth while a sixth harmonic retains only one sixty-fourth. Both can coexist on the wall without being equally visible in the interior.

For $f(\phi) = 2 + 3\cos(2\phi)$ V, the interior is

$$2 + 3(r/a)^2\cos(2\phi) = 2 + 3(x^2 - y^2)/a^2.$$

The potential has a saddle at the centre, with zero field there and opposite Cartesian curvatures. A scan along a single diameter could show a conspicuous maximum without detecting any charge. The two-dimensional Laplacian is zero because the curvatures cancel. This is also a useful finite-difference benchmark: a centred Cartesian Laplacian differentiates this quadratic exactly in the interior when the circular boundary has been represented correctly.

End effects produce a Sturm–Liouville problem

Now ground the side of a circular tube, excite its end at $z = 0$ and look for fields decaying into $z > 0$. For an axisymmetric source-free potential $V = R(r)Z(z)$, choose $Z = e^{-kz}$ for decay. Laplace's equation becomes

$$(rR')' + k^2 rR = 0.$$

The solutions are $J_0(kr)$ and $Y_0(kr)$. The former is regular with value one at zero; the latter diverges logarithmically. Grounding at $r = a$ requires $J_0(ka) = 0$, so $k_n = j_n/a$ with positive roots $j_1 = 2.40483,\ j_2 = 5.52008, \ldots$. Unlike sine wavenumbers they are not evenly spaced at low index.

Multiplying two such eigen-equations by the opposite radial function, subtracting and integrating gives $(k_m^2 - k_n^2) \int r R_m R_n \, dr$ equal to the endpoint term $r(R_n' R_m - R_m' R_n)$. It vanishes at a grounded wall and at a smooth axis. Hence modes at distinct eigenvalues are orthogonal with weight r. This weight arises physically from the circumference of a radial ring. Normalize dimensionless radius $x = r/a$. For a zero j_n, the squared modal norm is $J_1(j_n)^2/2$. An end profile $f(x)$ therefore gives

$$V(r, z) = \sum_n c_n J_0(j_n r/a) e^{-j_n z/a},$$

$$c_n = \frac{2}{J_1(j_n)^2} \int_0^1 x f(x) J_0(j_n x) \, dx.$$

For $f = V_0$ across the cap, $\int_0^1 x J_0(jx) dx = J_1(j)/j$ by differentiating $x J_1(jx)$. Then $c_n = 2V_0/[j_n J_1(j_n)]$, with alternating signs. The cap voltage jumps at the wall corner, making its near-rim reconstruction slow. A smooth end pattern that already vanishes at a yields faster coefficient decay. In either case a narrow radial strip at the axis should not be given as much integration weight as an equal width near the circumference.

Two ways to see which axial scale survives

Far enough from the cap the first excited nonzero coefficient dominates. Its length is

$$a/j_1 = 0.416a.$$

Ratios of distant on-axis voltages can then identify tube radius without absolute calibration of the cap voltage. For $a = 15.0$ mm, an added 5.0 mm of distance should give factor $e^{-2.40483/3} = 0.449$. Repeat this estimate at several starting positions. The ratio only approaches that value once competing modes

have decayed. A datum deep enough to suppress all higher modes may also be close to the instrument noise level, which biases a fitted decay length.

Instead impose an axial sinusoid e^{ikz} on a very long cylinder's side wall. The radial equation now has the opposite sign, $(rR')' - k^2 rR = 0$, whose regular solution is the modified Bessel function $I_0(kr)$. The potential is $V_0 I_0(kr)\cos(kz)/I_0(ka)$. The experiment selects a real oscillation along z, not an exponential attenuation away from an end. Near the axis $I_0(\xi) = 1 + \xi^2/4 + \cdots$; the field generated by a long-wavelength side modulation is nearly uniform across a narrow cross section. At large ka, $I_0(ka)$ grows exponentially and the axis scarcely senses a short-wavelength wall pattern. Calling every radial solution simply a Bessel mode without tracking the separation sign can exchange these two physical predictions.

Construct a grid that treats the axis as a regular point

A conservative radial discretisation starts from $\int_{r_-}^{r_+} r\,\nabla_r^2 V\,dr = [rV']_{r_-}^{r_+}$. Use annular cells and fluxes at shared ring faces. Near $r = 0$ the left face has zero circumference, requiring no unknown point beyond the axis. For a nodal grid with Taylor expansion $V(r) = V(0) + Ar^2 + O(r^4)$, the radial operator at zero is $4A$, so $\nabla_r^2 V(0) \simeq 4[V(h) - V(0)]/h^2$. Simply setting $1/r = 0$ and using the Cartesian one-dimensional second derivative misses a factor of two.

Compare a numerical response to the analytic J_0 mode at a fixed z/a. Test the wall value, axis derivative, radial energy norm and distant decay rate. These are independent observables. A discretisation can reproduce the slowest length while misrepresenting an edge profile composed of many faster modes. Conversely, matching a discontinuous cap perfectly at mesh nodes does not demonstrate that high-order interior amplitudes are accurate.

Can a coaxial dielectric be graded to produce a uniform field?

The homogeneous coaxial field grows as $1/r$ and reaches its highest magnitude at the inner wire. A designer proposes grading the dielectric radially so the full annular gap experiences one common radial field. Assume both metal faces are equipotentials, the scalar permittivity $\epsilon(r)$ may vary continuously, and there is no free volume charge. Cylindrical symmetry still gives $D_r = \lambda/(2\pi r)$, independent of how the material is graded. A constant electric-field magnitude E_g requires $\epsilon(r) = \lambda/(2\pi E_g r)$, that is, permittivity proportional to $1/r$. This construction replaces a uniform material by a known spatial distribution; applying $\epsilon(r)\nabla^2 V = 0$ would miss the material gradient and give the wrong field.

Write $\epsilon(r) = \epsilon_a a/r$. At applied voltage V_0, the constant field is $E_g = V_0/(b-a)$, and charge per length is

$$\lambda = 2\pi\epsilon_a a V_0/(b-a).$$

Thus $C' = 2\pi\epsilon_a a/(b-a)$. For $a = 2.0$ mm, $b = 6.0$ mm, $\epsilon_a = 6\epsilon_0$, the outer relative permittivity is 2, and $C' = 167$ pF/m. At 12 V every radius experiences 3000 V/m. For a homogeneous dielectric at the same voltage the inner surface field would be $12/[0.002\ln 3] = 5461$ V/m regardless of its constant permittivity. The gradient reduces peak field at fixed voltage. Whether a real material benefits also depends on its local field limit, losses and interfaces.

Where is the polarisation charge that makes this possible? In the dielectric the true total charge satisfies

$$\rho_{\mathrm{tot}} = \epsilon_0 \nabla \cdot \mathbf{E} = \epsilon_0 E_g / r.$$

All of it is bound since $\rho_f = 0$. Direct evaluation from $\mathbf{P} = (\epsilon(r) - \epsilon_0) E_g \hat{\mathbf{r}}$ gives $-r^{-1}\partial_r(rP_r) = +\epsilon_0 E_g / r$, the same result. Material surfaces add bound charges $P_r(b)$ on the outer radial face and $-P_r(a)$ on the inner face. Integrating their surface contributions and the volume charge over a full annulus returns zero net bound charge, as required for a polarised neutral body. A constant radial electric magnitude can therefore coexist with zero free charge and nonzero total charge.

A manufactured graded material is often approximated by concentric homogeneous shells. For shell j from r_j to r_{j+1}, its voltage drop at fixed λ is $\lambda \ln(r_{j+1}/r_j)/(2\pi\epsilon_j)$. To reproduce the exact voltage drop of the continuously graded design over each shell, choose $\epsilon_j = \epsilon_a a \ln(r_{j+1}/r_j)/(r_{j+1} - r_j)$. This matches electrode capacitance even with relatively few shells. Yet the field inside each discrete shell still falls as $1/r$ and jumps at its interfaces. Matching terminal capacitance alone does not verify the target uniform-field profile. Compute the field just inside each shell edge and refine the radial layering against a local peak-field tolerance.

What a small displacement of the end electrode excites

An axisymmetric tube calculation retains only angular order $m = 0$. Now impose at the tube entrance a voltage profile proportional to $r\cos\phi$ near the axis and taper it to zero at the wall. This is a dipolar lateral pattern with positive voltage on one side and negative on the other. Separating $R(r)\cos(m\phi)e^{-kz}$ into Laplace's equation yields

$$(rR')' + \left(k^2 r - \frac{m^2}{r}\right) R = 0.$$

The regular solutions are $J_m(kr)$. Near the axis their leading powers are $r^{|m|}$, so an $m = 1$ potential vanishes on the centreline while its transverse gradient generally does not. The grounded wall selects zeros of J_m. The first zero of J_1 is 3.83171, larger than the first axisymmetric zero, so the slowest lateral disturbance decays faster than the slowest axisymmetric voltage. Probe design therefore changes which decay laws are experimentally accessible.

For a cap shaped exactly as $f = V_1 J_1(jr/a)\cos\phi$, $j = 3.83171$, the whole solution is a single mode. Near the axis $J_1(jr/a) \simeq jr/(2a)$; using $r\cos\phi = x$,

$$V \simeq V_1 j x e^{-jz/a}/(2a).$$

Hence $E_x(0,z) = -V_1 j e^{-jz/a}/(2a)$. With $V_1 = 2.0$ V and $a = 10$ mm, the lateral field starts at 383 V/m and falls to 8.30 V/m after an axial distance equal to one radius. A centreline voltmeter sees zero at both positions and gives no hint of the finite transverse force. A pair of voltmeters at small opposite x or a field probe oriented along x measures the relevant derivative.

For a realistic shifted electrode the input can contain both $m = 0$ and $m = 1$. Place detectors at $(r, 0, z)$ and (r, π, z). Their average selects even azimuthal orders; their difference selects odd ones. The ratio of

the difference at two heights tests $m = 1$ decay after higher odd modes subside. The average also includes $m = 2, 4, \ldots$, so attributing it to $m = 0$ requires symmetry information or more angular sensors. Reversing the applied electrode displacement changes the sign of an odd response but leaves even corrections unchanged to leading order. This experiment links angular Fourier selection and radial eigenvalues to a measurable alignment tolerance without requiring an entire three-dimensional voltage map.

Project a ring source without calling an axis cell a point detector

Deposit a thin circular ring of total free charge q at radius s in the plane $z = z_0$ inside a long grounded cylindrical tube. Its volume density in an axisymmetric model is $q\,\delta(r - s)\delta(z - z_0)/(2\pi s)$, for $s > 0$. The $2\pi s$ removes the circumference introduced by the volume element. A radial code that deposits q directly as a density in one mesh bin, without dividing by annular volume, changes the total charge each time the radius or mesh spacing changes. In this calculation one can track that normalisation all the way to a voltage coefficient.

Take the tube infinite in both axial directions for simplicity. Expand $V = \sum_n v_n(z) J_0(j_n r/a)$ in the grounded radial eigenfunctions. Project Poisson's equation using weight $r\,dr$ and the norm $a^2 J_1(j_n)^2/2$. The ring produces

$$[-d_z^2 + (j_n/a)^2] v_n = \frac{q J_0(j_n s/a)}{\pi \epsilon a^2 J_1(j_n)^2}\,\delta(z - z_0).$$

The infinite-line Green response for $-d_z^2 + k^2$ is $e^{-k|z-z_0|}/(2k)$, as can be checked by integrating its derivative across z_0. Thus each modal voltage has a finite amplitude and two decaying tails, and the complete field is obtained by summing radial modes. At a detector separated axially from the ring the exponential factors ensure rapid convergence. Exactly on the mathematical ring the local potential diverges logarithmically like that of a thin filament; comparing finite-element peak voltages there requires a finite wire cross section or a specified detector average.

An unexpected and testable consequence concerns the source radius. The coupling to mode n is proportional to $J_0(j_n s/a)$. It becomes exactly zero when the source sits on a radial node of that mode. The second mode has an interior node at $s/a = j_1/j_2 \simeq 0.4357$. A ring there still carries charge and excites many other modes but contributes no second modal coefficient. As an experimental analogy, a weighting potential that changes sign across part of a sensor can receive cancelling contributions from an extended excitation. Here the cancellation is encoded exactly in orthogonality and the pointwise coupling value.

A check independent of the ring-charge normalisation uses reciprocity. Imagine two narrow coaxial source/detector rings of radii s, r, each with the same definition of total charge and voltage measured around its circumference. The modal response contains the symmetric product $J_0(j_n s/a) J_0(j_n r/a)$ multiplied by a factor that depends only on axial separation. Swapping radius and axial position therefore leaves voltage per total charge unchanged. If a numerical table instead looks symmetric per charge density but not per total ring charge, check whether the radius-dependent integration weights were applied on both deposition and sampling.

Another useful test fixes total deposited charge but reduces a resolved annular thickness around s. First compare with an integral of the analytic ring solution across the same finite distribution. Then take the thin-ring limit at a detector away from the deposition region. Taking grid thickness and source thickness to zero together while comparing a self-voltage at the source conflates numerical refinement with a mathematically singular observable. Stating detector location, normalisation and order of limits turns a superficially straightforward ring source into a robust benchmark for cylindrical Poisson solvers.

Measure a radial cylindrical mode

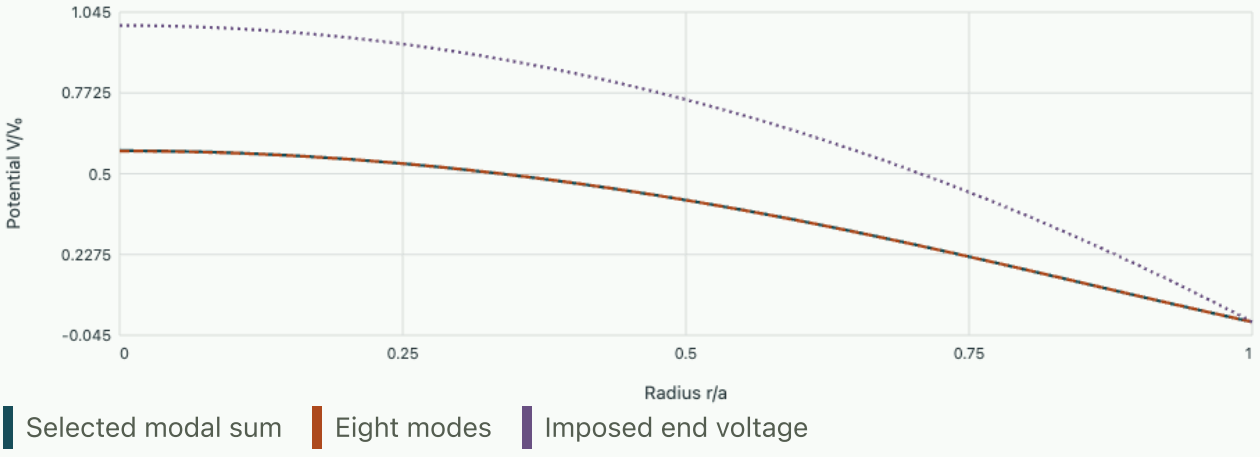
Change the electrode mode and compare radial nodes with axial decay. A cylindrical boundary chooses a different eigenbasis from a rectangular one.

Open this experiment online

Try the calculation

Height z/a **0.25**

Retained modes **3**



The on-axis potential is 0.5774 V_0 with the selected modes, versus 0.5765 V_0 using eight. The slowest mode has e-folding length 0.4158 a.

A cylindrical interface can carry induced bound charge with an entirely radial field

Separate the two charges detected at a coaxial material boundary

A long coaxial gap contains two homogeneous radial layers with interface at s, no free sheet there and fixed inner line charge $+\lambda$. Find the bound interface charge per length, verify it from the field jump, and determine the interface radius at which the layers take equal voltages. Evaluate the result for $a = 2$ mm, $b = 18$ mm and relative permittivities 2 and 4.

▼ **Derive and check the result**

For both regions $D_r = \lambda/(2\pi r)$. In material i the polarisation is $P_i = D_r(1 - \epsilon_0/\epsilon_i)$ along the radial direction. At the interface the outer face of material 1 has bound density $+P_1$, the inner face of material 2 has $-P_2$. Their sum times the circumference gives bound line charge $\lambda_b = \lambda(\epsilon_0/\epsilon_2 - \epsilon_0/\epsilon_1)$. It equals $-\lambda/4$ for the given ordering; the value per length is independent of s because the $1/s$ field cancels the interface circumference. A local density meter would see a different value at a different radius even though the integrated result stays the same.

Direct use of Gauss's law for the true electric field gives $\epsilon_0\, 2\pi s(E_{2r} - E_{1r}) = \lambda[\epsilon_0/\epsilon_2 - \epsilon_0/\epsilon_1]$, confirming the sign and magnitude. A check with displacement alone gives zero jump and thus correctly reports no *free* charge at that same surface. The two field laws answer different questions. Extending the shell calculation to include both exterior material faces returns zero total bound charge per material, so the internal negative interface charge never represents a net creation of charge.

The voltage drops are proportional to $\ln(s/a)/\epsilon_1$ and $\ln(b/s)/\epsilon_2$. Equality gives $s^{\epsilon_1+\epsilon_2} = a^{\epsilon_2} b^{\epsilon_1}$, or $s = a(b/a)^{\epsilon_1/(\epsilon_1+\epsilon_2)}$, a dimensionally transparent form. Numerically $s = 2(9)^{1/3} = 4.16$ mm. A half-voltage location at 4.16 mm is far from the arithmetic centre of this gap. It is also distinct from the 6.0 mm geometric mean for a homogeneous annulus, because the inner lower-permittivity layer needs less geometric thickness to take the same voltage.

Tutorial · radial modes with different physical boundaries

Replace the grounded wall by a reflecting boundary

A long tube has $\partial_r V = 0$ on its side instead of $V = 0$, and its end voltage profile is nonuniform. What happens to the radial eigenvalues and to the assertion that every end disturbance dies away?

▼ **Work through the solution**

The wall now requires

$$J_0'(ka) = -J_1(ka) = 0.$$

In addition a constant radial mode with $k = 0$ satisfies both side and axis conditions. Its axial equation is $Z'' = 0$, so boundedness allows a constant component. The end profile's weighted cross-sectional mean supplies that persistent offset. Only the components orthogonal to the constant decay. Requiring the potential to tend to zero at infinity while prescribing a nonzero mean cap voltage would be incompatible for this insulated side model unless other boundaries or sources are introduced.

Infer a radial charge density from a coaxial reading

Suppose the measured potential in a coaxial gap is $V = A\ln(b/r) + B(b^2 - r^2)$ with constant permittivity. Determine the volume charge and the signs of its contribution to the two face fluxes. Which

part can be changed by externally altering the applied voltage at fixed charge density?

▼ **Work through the solution**

The logarithmic term is harmonic away from the excluded axis. For the quadratic term $\nabla_r^2[-Br^2] = -4B$, giving uniform $\rho = 4\epsilon B$. The radial field is $A/r + 2Br$. The net outward displacement flux through the gap per length is $2\pi\epsilon[bE(b) - aE(a)] = 4\pi\epsilon B(b^2 - a^2)$, exactly $\rho\pi(b^2 - a^2)$. At positive B the quadratic term contributes positive outer flux and negative inward-face flux, with the larger outer contribution from both area and field. Changing only the source-free boundary excitation adds a logarithmic term and, if the reference permits it, a constant. It alters both field readings without altering the inferred uniform volume density.

Further reading

- Reconstruct a radial electrode profile with one to eight Bessel modes
- Fließbach, chapter 10 · cylindrical problems

Normalised spherical harmonics and the field of an arbitrary three-dimensional source

Electrostatic field theory and measurement

An axial voltage scan can determine Legendre coefficients only if the source is rotationally symmetric. A general spherical electrode or molecule needs angular modes with two indices. Their normalisation fixes the conversion from a measured multipole coefficient to a physical voltage.

Separate three coordinates without imposing axial symmetry

In spherical coordinates the source-free Laplace equation is $r^{-2}\partial_r(r^2\partial_r V) + r^{-2}\Delta_\Omega V = 0$, where $\Delta_\Omega = (\sin\theta)^{-1}\partial_\theta(\sin\theta\,\partial_\theta) + (\sin^2\theta)^{-1}\partial_\phi^2$. Assume $V = R(r)Y(\theta,\phi)$. Separation requires

$$-\Delta_\Omega Y = \lambda Y.$$

Multiply by Y^* and integrate over the unit sphere. Integration by parts has no actual surface boundary for globally smooth fields and gives

$$\lambda \int |Y|^2 d\Omega = \int |\nabla_\Omega Y|^2 d\Omega \geq 0.$$

The constant mode has eigenvalue zero.

Separate azimuth as $e^{im\phi}$. Continuity after a full rotation forces integer m. With $u = \cos\theta$, the remaining equation becomes

$$[(1-u^2)P']' + [\lambda - m^2/(1-u^2)]P = 0.$$

Requiring regular behaviour at both poles yields $\lambda = \ell(\ell+1)$ with integers $\ell \geq |m|$. The corresponding angular functions are the associated Legendre functions. For each ℓ there are $2\ell+1$ independent m's, all with identical radial behaviour. A rotation mixes coefficients within this subspace but cannot convert a dipole into a quadrupole.

Fix one convention and verify its first modes

Choose $P_\ell(1) = 1$ and for $m \geq 0$ include the Condon–Shortley phase in $P_\ell^m(u) = (-1)^m(1-u^2)^{m/2}d^m P_\ell/du^m$. Define

$$Y_{\ell m} = \sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}}\,P_\ell^m(\cos\theta)e^{im\phi},$$

$$Y_{\ell,-m} = (-1)^m Y_{\ell m}^*.$$

With this convention $\int Y_{\ell m}^* Y_{\ell' m'}\,d\Omega = \delta_{\ell\ell'}\delta_{mm'}$, using $d\Omega = \sin\theta\,d\theta\,d\phi$. In particular $Y_{00} = 1/\sqrt{4\pi}$,

$$Y_{10} = \sqrt{3/(4\pi)}\cos\theta,$$

$$Y_{11} = -\sqrt{3/(8\pi)}\sin\theta e^{i\phi}.$$

To check the dipole normalisation independently,

$$\int \cos^2\theta \, d\Omega = 4\pi/3.$$

For Y_{11},

$$\int \sin^2\theta \, d\Omega = 8\pi/3.$$

The complex basis is an accounting device. A real potential obeys $v_{\ell,-m} = (-1)^m v^*_{\ell m}$, leaving $2\ell+1$ real angular degrees of freedom at each order.

On a sphere of radius a suppose segmented electrodes prescribe $f(\theta,\phi)$. Its coefficients are $v_{\ell m} = \int Y^*_{\ell m} f \, d\Omega$. These have units of volts because the angular functions and solid angle are dimensionless. The radial equation $(r^2R')' - \ell(\ell+1)R = 0$ has solutions $r^\ell, r^{-\ell-1}$. For an empty interior, regularity selects $V_{\text{in}} = \sum v_{\ell m}(r/a)^\ell Y_{\ell m}$. For empty exterior decaying at infinity,

$$V_{\text{out}} = \sum v_{\ell m}(a/r)^{\ell+1} Y_{\ell m}.$$

Nonuniform prescribed surface voltage requires electrically distinct electrode patches or a specified ideal boundary; one continuous metal conductor could not sustain that static variation.

Represent a tilted dipole in the laboratory frame

Impose $f = V_0 \hat{\mathbf{p}}\cdot\hat{\mathbf{r}}$ with $\hat{\mathbf{p}} = (\sin\alpha, 0, \cos\alpha)$. In the rotated frame with polar axis along $\hat{\mathbf{p}}$ this is a pure $m=0, \ell=1$ pattern. In the laboratory it is $V_0[\cos\alpha\cos\theta + \sin\alpha\sin\theta\cos\phi]$. Using the explicit harmonics,

$$v_{10} = V_0\sqrt{4\pi/3}\cos\alpha,$$

$$v_{11} = -V_0\sqrt{2\pi/3}\sin\alpha,$$

$$v_{1,-1} = +V_0\sqrt{2\pi/3}\sin\alpha.$$

The sum of their squared magnitudes is $4\pi V_0^2/3$, independent of tilt. The interior solution simplifies to $V_0\hat{\mathbf{p}}\cdot\mathbf{r}/a$, giving uniform $\mathbf{E} = -V_0\hat{\mathbf{p}}/a$. At $a = 40$ mm, $V_0 = 2.0$ V and $\alpha = 30^\circ$, $E_x = -25.0$ V/m, $E_z = -43.3$ V/m. An axis-only fit would infer the correct axial component while losing the transverse one. A full vector field measurement at one interior point distinguishes the rotated dipole from a weaker aligned one.

Tie angular modes to an actual charge distribution

The addition theorem expresses $P_\ell(\hat{\mathbf{r}}\cdot\hat{\mathbf{s}}) = [4\pi/(2\ell+1)]\sum_m Y_{\ell m}(\hat{\mathbf{r}})Y^*_{\ell m}(\hat{\mathbf{s}})$. Rotational invariance explains its structure: the sum is unchanged by simultaneous rotation of both directions, so it depends only on their angle. Put $\hat{\mathbf{s}}$ at the north pole, where only $m=0$ survives, and the proportionality constant follows from $Y_{\ell 0}$ and $P_\ell(1)=1$. Checking $\ell=1$ recovers the dot product of the two unit vectors.

Use the generating series $1/|\mathbf{r}-\mathbf{s}| = \sum_{\ell\ge 0}(r_<^\ell/r_>^{\ell+1})P_\ell(\hat{\mathbf{r}}\cdot\hat{\mathbf{s}})$. For an observation radius outside every source point, define $q_{\ell m} = \int \rho(\mathbf{s}) s^\ell Y^*_{\ell m}(\hat{\mathbf{s}})\, d^3s$. Inserting the addition theorem into the Coulomb integral gives

$$V(\mathbf{r}) = \frac{1}{\epsilon_0}\sum_{\ell,m}\frac{q_{\ell m}}{(2\ell+1)r^{\ell+1}}Y_{\ell m}(\hat{\mathbf{r}}).$$

Unlike boundary voltage coefficients, $q_{\ell m}$ carries units C m$^\ell$. At $\ell=0$, $q_{00} = Q/\sqrt{4\pi}$ and multiplication by Y_{00} returns $Q/(4\pi\epsilon_0 r)$, fixing the prefactor. At $\ell=1, m=0$,

$$q_{10} = \sqrt{3/(4\pi)}\,p_z.$$

The two transverse coefficients reconstruct p_x, p_y. Measuring the voltage coefficient at two radii can distinguish a decaying exterior multipole from a regular potential imposed by distant equipment, whose local expansion instead grows as r^ℓ.

Design an angular quadrature that measures the claimed coefficient

Uniformly spaced latitude rings do not have equal area. In angular integrals substitute $u=\cos\theta$, giving $d\Omega = du\,d\phi$. Equal spacing in u and ϕ already avoids overweighting the polar caps. More accurate quadrature uses Gauss–Legendre nodes in u with their weights, combined with equispaced azimuthal samples for the Fourier factors. If a known angular pattern contains $|m|\le M$, at least $2M+1$ azimuthal samples are needed to distinguish these signed orders on a fixed ring. A higher order present in an unknown pattern can alias into the fitted low orders, so adding more sample angles and reserving independent test directions is informative.

A robust reconstruction checks the constant mode on $f=1$, the total power $\sum|v_{\ell m}|^2 = \int |f|^2 d\Omega$ for an exactly represented pattern, and the predicted angular field at withheld directions. With noisy overdetermined data use a weighted linear least-squares fit to the modes, including weights for sampling geometry or measurement uncertainty as appropriate. The two choices are physically different: one approximates an angular integral, the other quantifies independent noisy observations. State which objective your fit minimises. An angular reconstruction can predict a voltage while still having very poor sensitivity to high-order exterior source moments, whose signals decay rapidly with radius.

A displaced charge supplies an exact all-orders convergence experiment

Place a point charge q at $s\hat{\mathbf{z}}$ within an otherwise empty sphere used as a measurement surface. At radius $R > s$ the potential is $q/[4\pi\epsilon_0\sqrt{R^2 + s^2 - 2Rs\cos\theta}]$. The axial symmetry makes only $m = 0$ coefficients nonzero, giving the exact Legendre series $[q/(4\pi\epsilon_0 R)]\sum_{\ell\geq 0}(s/R)^\ell P_\ell(\cos\theta)$. This example exhibits the convergence parameter directly. It is the radius of the enclosing source relative to the measurement radius, not the magnitude of the charge or the number of sampled angles.

On the ray nearest the point charge, $\theta = 0$, every Legendre polynomial is unity. Truncation at order L there gives a geometric partial sum and a relative error exactly $(s/R)^{L+1}$ when compared with the exact voltage at that point. At $s/R = 0.40$, six retained orders $\ell = 0, \ldots, 5$ leave 0.410% relative error. At $s/R = 0.85$ the same approximation leaves 37.7%. Achieving 1% in the latter case requires $L + 1 > \ln(0.01)/\ln(0.85) = 28.3$, so at least 29 orders. A claim that retaining through quadrupole always suffices has no physical meaning without an observation radius and a desired error.

On the opposite ray $P_\ell(-1) = (-1)^\ell$, making the terms alternate. The absolute voltage error there has a smaller prefactor $1/(1 + s/R)$, in contrast to $1/(1 - s/R)$ on the near side. For this particular geometric series its relative error magnitude is still $(s/R)^{L+1}$, despite the alternating signs. At other angles the angular polynomial oscillations can produce much smaller error but not a uniformly improved bound. The simple inequality $|P_\ell(u)| \leq 1$ gives an absolute bound $|q|(s/R)^{L+1}/[4\pi\epsilon_0 R(1 - s/R)]$ for the entire angular surface. For an arbitrary signed source replace $|q|$ by integrated absolute charge and s by a radius enclosing all source points. A nearly neutral distribution can have a small exact voltage at some angles, making any comparable relative bound much less informative.

Now rotate the point charge through polar and azimuthal angles θ_s, ϕ_s without changing its radius. The coefficient is exactly $q_{\ell m} = q s^\ell Y^*_{\ell m}(\theta_s, \phi_s)$. Summing $|q_{\ell m}|^2$ over m gives $q^2 s^{2\ell}(2\ell + 1)/(4\pi)$, by setting the two angular arguments equal in the addition theorem. This rotational invariant is a strong test of a numerical angular projection: a rotation changes individual complex phases, but the total power at each angular order remains fixed. Rotating a sensor array rather than the source should produce the equivalent opposite rotation of the fitted description.

A spherical split electrode drives an interior field without net charge

Construct a spherical boundary of radius a whose northern half is held at $+V_0$ and southern half at $-V_0$, separated by a thin insulating equatorial seam. In the mathematical limit take $f(u) = V_0 \operatorname{sgn} u$, $u = \cos\theta$. The problem is axisymmetric and changes sign under reflection through the equator, so only odd ℓ occur. Using the Legendre (rather than unit-normalised harmonic) coefficients, the series inside is $V = \sum b_\ell (r/a)^\ell P_\ell(u)$, where $b_\ell = (2\ell + 1)V_0 \int_0^1 P_\ell(u)\, du$ for odd ℓ. There is no monopole, hence no change to the net charge when the two voltages are applied symmetrically in a symmetric environment.

The first two nonzero coefficients can be found without a general table. $P_1 = u$ gives

$$b_1 = 3V_0/2.$$

$P_3 = (5u^3 - 3u)/2$ integrates from zero to one to $-1/8$, so $b_3 = -7V_0/8$. At the centre all terms with $\ell \geq 2$ have vanishing gradient, leaving an exact central field $E_z(0) = -3V_0/(2a)$. Even though the electrode voltages jump abruptly at the equator, a single low-order coefficient determines this point measurement exactly. A finely resolved electrode charge density near the seam requires a very different number of modes and a model of the seam width.

Set $a = 30$ mm and $V_0 = 2.0$ V. The central field is -100 V/m. At $z = 0.20a$ on the northern axis, the $\ell = 1$ voltage contributes 0.600 V and $\ell = 3$ adds -0.0140 V. The first omitted odd term scales as $(0.2)^5$ times its angular coefficient. At $z = 0.9a$ the same rapidly convergent two-term approximation is no longer trustworthy. A student can compare the actual surface voltages estimated from a truncated expansion to predictions at selected interior radii. A converged central field and a conspicuously rounded numerical step can coexist in the same exact analytic model.

For a finite insulating angular seam suppose instead the boundary interpolates continuously and antisymmetrically near $u = 0$. Only a thin band changes relative to the sign function. The dipole coefficient depends on $\int f(u) u \, du$; the extra weight u suppresses the influence of perturbations concentrated around the equator. If the band spans $|u| < \delta$ and the change in potential is bounded by V_0, the dipole coefficient error is at most $3V_0\delta^2/2$. High-order edge charge can nevertheless change markedly. This estimate shows how a real imperfection can have a small impact on one observable and a large impact on another. It also provides a quantitative tolerance when using the ideal split boundary to predict a centre-field calibration.

Separate material inside a measurement sphere from fields imposed outside it

A laboratory sometimes measures voltages on two concentric source-free spherical surfaces, without knowing whether the disturbance comes from an object inside both surfaces or from distant electrodes outside both. A single angular map determines the coefficient of each measured harmonic at that radius but cannot in general distinguish these two origins. In an annulus lacking sources, each angular component takes the radial form $v_{\ell m}(r) = A_{\ell m} r^{\ell} + B_{\ell m} r^{-\ell-1}$. The regular term is compatible with external forcing, and the decaying term with sources inside the inner radius. Both are harmonic in the measurement shell.

For a single $\ell = 1$ component, denote its coefficient of an ordinary $\cos\theta$ pattern by $v(r) = a_1 (r/R_1) + b_1 (R_1/r)^2$, so a_1, b_1 are voltages at the first radius. If two scans give $v(R_1) = 2.00$ V and $v(2R_1) = 1.30$ V, the equations are

$$a_1 + b_1 = 2.00, \; 2a_1 + b_1/4 = 1.30.$$

Multiply the first by one quarter and subtract, giving $a_1 = 0.4571$ V and $b_1 = 1.5429$ V. At a third radius $1.5R_1$ the predicted value is 1.371 V. A pure internal dipole would have returned 0.50 V on the second sphere, while a pure uniform external field would have returned 4.00 V. Fitting a single dipole to the first reading alone would grossly misdescribe the second.

For general order the determinant of the two normalised equations is $t^{-\ell-1} - t^{\ell}$ with $t = R_2/R_1$. As the two radii approach each other, this determinant tends to zero and the coefficients cannot be separated

reliably. Increasing separation improves discrimination of the two radial powers up to the point where one component falls below the measurement noise or additional sources enter the annulus. The monopole case deserves separate attention: its growing branch is just a constant voltage offset, while its decaying branch reveals total charge. A voltage reference common to both scans is therefore indispensable when using their difference to infer monopole charge.

Suppose the field was measured rather than voltage. Radial field coefficients are $-\ell A_{\ell m} r^{\ell-1} + (\ell+1) B_{\ell m} r^{-\ell-2}$. Measuring both radial field and voltage on one sphere can, in principle, supply the two independent equations needed to separate regular and decaying parts. For $\ell = 0$ the constant reference disappears from the radial field. An independently calibrated force probe then constrains total internal charge even when the voltmeter has a common offset. Numerically estimated radial derivatives from two very close voltage scans, however, share measurement errors with those voltages; treating them as two independent precise observations overstates the information gained.

A final issue is the coordinate origin. Moving it can create higher moments from lower ones for a nonneutral distribution. If the actual source is a displaced point charge and the coordinate origin shifts toward it, the series converges faster at a given observation sphere. If the charge distribution is unknown, choosing the origin solely to minimise apparent high-order coefficients can hide genuine structure or amplify noise. Record the physical origin of the measuring geometry, report withheld predictions, and compare the integrated monopole flux, which is independent of the choice of origin as long as the source remains enclosed. This separates a useful coordinate convention from an inferred physical symmetry.

Propagate measured boundary harmonics

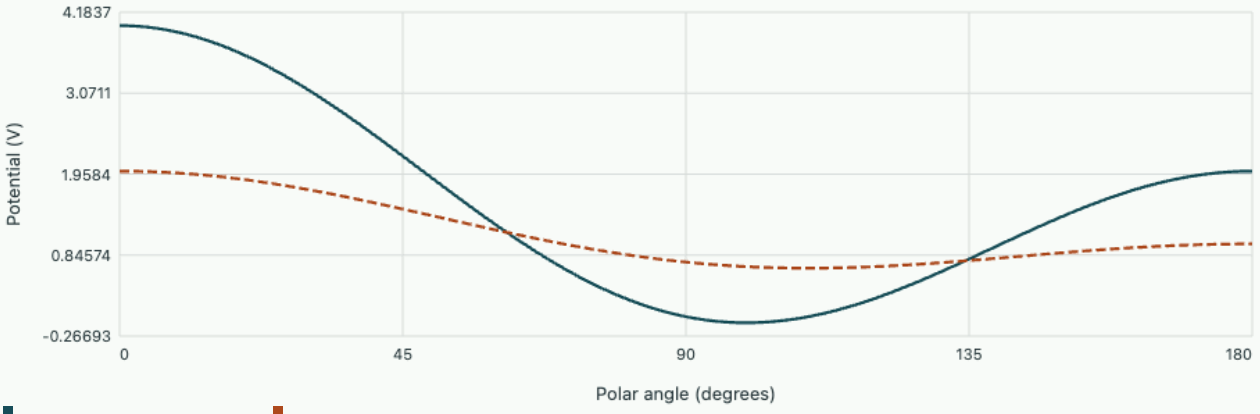
Choose angular boundary data and examine its attenuation away from the spherical surface. Relate the relative loss of detail to harmonic degree.

Open this experiment online

Try the calculation

Observation radius r/a **0.5**

Quadrupole on the sphere (V) **2**



Boundary at r = a Selected radius

The north pole and equator read 2 V and 0.75 V at this radius. Both branches coincide on the boundary.

Predict an angular sampling requirement before buying more probes

How many angular measurements are needed for a declared band limit?

A real boundary pattern is known, independently of the current measurement, to contain only harmonics with $\ell \le 3$. Count its real degrees of freedom. Design a separable quadrature that exactly projects such a pattern onto the same harmonics in exact arithmetic. Explain what can still go wrong when the true band limit was assumed incorrectly.

▼ **Derive and check the result**

There are $\sum_{\ell=0}^{3}(2\ell+1) = 16$ independent real degrees of freedom. Sixteen arbitrary readings need not determine them: their locations must give a full-rank observation matrix. Equatorial-only data, for example, miss every axisymmetric odd pattern there. A straightforward overdetermined grid uses uniform azimuths and Gauss–Legendre nodes in $u = \cos\theta$. For $L = 3$ use at least $2L+1 = 7$ equally spaced azimuths so that differences $m - m'$ up to 6 cannot alias a nonzero azimuthal difference into the constant component of a projection.

After that exact azimuthal sum, only equal m terms survive. Their products in u are proportional to $P_\ell^m(u)P_{\ell'}^m(u)$. Each contains $(1-u^2)^{|m|}$ times a polynomial and is itself a polynomial of degree at most $\ell + \ell' \le 6$. A four-point Gauss–Legendre rule integrates polynomials through degree 7 exactly. Thus four latitudes times seven azimuths, 28 weighted samples, suffice to reconstruct all 16 coefficients for the

declared band limit. Both polar caps are represented by quadrature weights; no point must lie precisely on either pole.

If the true field contains order $m = 7$, then $e^{i7\phi} = 1$ at the seven uniformly spaced azimuths starting at zero. Its azimuthal variation can masquerade as a constant on each sampled circle. A second scan shifted by half an azimuthal spacing flips the sign of this harmonic but preserves a true constant, exposing the alias. Unmodelled high latitudinal orders can similarly corrupt low coefficients through the finite radial quadrature. With noisy measurements evaluate withheld off-grid directions and report both the sensitivity to additional modes and the precision assumed in the protected low orders. An exactly integrating grid under a band-limit hypothesis is not a proof that the physical source obeyed that hypothesis.

Tutorial · derive a surface charge and test identifiability

Two spheres agree on their voltage patterns

The same prescribed $v_{\ell m}$ is realised on an infinitesimally thin spherical electrode array between identical dielectric media of permittivity ϵ. Find the coefficient of its free charge density. Which integrated measurement isolates the monopole?

▼ **Work through the solution**

The interior and exterior radial derivatives on $r = a$ are $\ell v_{\ell m} Y_{\ell m}/a$ and $-(\ell+1) v_{\ell m} Y_{\ell m}/a$. With outward normal pointing from interior to exterior, $\sigma_f = \epsilon(E_{\mathrm{out},r} - E_{\mathrm{in},r})$, giving coefficient $\epsilon(2\ell+1)v_{\ell m}/a$. A whole-surface integral of $\sigma_f a^2 d\Omega$ retains only the monopole since all nonconstant harmonics are orthogonal to Y_{00}. For uniform boundary voltage V_0, $v_{00} = \sqrt{4\pi} V_0$ and total charge $4\pi\epsilon a V_0$, the capacitance of an isolated sphere times voltage. High-order voltage components can instead demand strong local charge densities despite zero net added charge.

A ring of detectors misses an entire pattern

Twelve probes placed uniformly on the equator read zero voltage. Can this exclude a substantial $\ell = 1$ component? Give an explicit counterexample and a second configuration still invisible after adding a north-pole probe.

▼ **Work through the solution**

A pure Y_{10} vanishes all along the equator, whatever its amplitude. A north-pole reading would detect it. But the real quadrupole proportional to $\sin\theta\cos\theta\cos\phi$, corresponding to xz/r^2, vanishes on both the equator and at either pole. An intermediate ring or a set of oblique directions is needed. Increasing the number of equatorial samples alone cannot resolve a lack of checks at other latitudes. Under a declared band limit compute the rank and singular values of the probe-by-harmonic matrix before interpreting a null reading as an absent multipole.

Further reading

- Compare interior and exterior harmonics with a polar scan
- Infer a tilted quadrupole from six independent probe readings
- Fließbach, chapter 11 · spherical harmonics
- Griffiths, 5th ed., §§3.3–3.4 · spherical separation and multipoles

Reconstructing a source from noisy voltage measurements

Electrostatic field theory and measurement

A forward electrostatic calculation can be unique and stable while the inverse problem constructed from it is unstable or nonunique. A scanned electrode, a finite collecting area and an explicitly chosen noise level make these statements experimentally testable.

Write down what the instrument actually measures

Pattern a long electrode at $z = 0$ with an unknown voltage $f(x)$, periodic over length L for this model. The region above is homogeneous, charge free, and contains a sufficiently weak probe at height h. For a Fourier component $f_m e^{ik_m x}$, $k_m = 2\pi m/L$, the bounded harmonic solution at height h is $f_m e^{-|k_m|h} e^{ik_m x}$. Treat the constant component according to the actual reference. Suppose the probe reports an average over a rectangular window of width w centred at x_j, plus noise e_j. For uniform probe sensitivity the forward relation is

$$y_j = \sum_m H_m f_m e^{ik_m x_j} + e_j,$$

$$H_m = e^{-|k_m|h}\,\mathrm{sinc}(k_m w/2),$$

$$\mathrm{sinc}\, t = \frac{\sin t}{t}.$$

The exponential is physical propagation from boundary to probe; the sinc factor is averaging by the detector. The latter has exact zeros in this idealised model. For modes at those zeros no noise-free reading at the same width and height can recover the lost amplitude. More closely spaced scan positions increase the sampling bandwidth but cannot undo either such instrumental nulls or the exponential decrease of the true high-frequency signal.

A nonuniform probe shape replaces the sinc by its Fourier transform. Unknown height adds an additional calibration problem: increasing h and increasing high-frequency boundary amplitudes can yield the same readings if the amplitudes are unrestricted. Before inferring boundary detail, calibrate the probe width and decide whether height is known, measured with uncertainty or jointly estimated using an independent constraint.

A direct inverse amplifies uncertainty in predictable bands

In an ideal complete periodic scan the discrete Fourier transform separates resolvable modes. Let $\hat{y}_m = H_m f_m + \eta_m$. Inverting one coefficient gives $f_m^{\mathrm{dir}} = \hat{y}_m / H_m$ whenever $H_m \neq 0$. A perturbation $\delta\hat{y}_m$ becomes

$$\delta f_m = \delta\hat{y}_m / H_m.$$

With $H_m = e^{-|k|h}$, the amplification grows exponentially with spatial frequency. The mode values can be uniquely determined in ideal noiseless data of infinite precision, yet the inverse is unbounded as increasingly fine modes are admitted. Finite data can additionally produce exact nonuniqueness.

For $h = 2.0$ mm and negligible aperture, boundary wavelengths 20, 10 and 4 mm have transfers $0.533, 0.285, 0.0432$, respectively. With 1 mV noise in the measured amplitude of each Fourier component, inverse noise levels are $1.87, 3.51, 23.1$ mV. A genuine 10 mV boundary ripple at wavelength 4 mm contributes only 0.432 mV at the probe, smaller than that noise. Plotting its raw inverted coefficient with a smooth line does not improve its information content. Reducing height to 1 mm raises that ripple's signal to 2.08 mV and changes the experiment's resolving power before any reconstruction choice is made.

Derive a regularised inverse and expose its bias

Assume coefficients are inferred separately and use a quadratic penalty. For each m, minimise $J_m = |H_m f - \hat{y}_m|^2 + \alpha^2 w_m^2 |f|^2$ with positive weights w_m and positive regularisation parameter α. Splitting real and imaginary parts, or differentiating with respect to f^*, gives

$$\widehat{f}_m = \frac{H_m^* \hat{y}_m}{|H_m|^2 + \alpha^2 w_m^2},$$

$$W_m = \frac{|H_m|^2}{|H_m|^2 + \alpha^2 w_m^2},$$

$$\mathbb{E}[\widehat{f}_m] = W_m f_m$$

for zero-mean measurement noise. The reconstruction suppresses noise near a small transfer by introducing deterministic bias $(W_m - 1) f_m$. Its variance for known mode-noise variance s_m^2 is $|H_m|^2 s_m^2 / (|H_m|^2 + \alpha^2 w_m^2)^2$. Neither low variance alone nor a small data residual is evidence that a spatial ripple was recovered. Setting α very large produces a smooth, nearly zero estimate regardless of the data. Setting it very small can reproduce noise almost exactly.

For $w_m = 1, \alpha = 0.10$ and wavelength 4 mm in the example above, $W_m \simeq 0.157$. A true 10 mV ripple is estimated at only 1.57 mV on average. The inversion gain is about 3.64 rather than 23.1. At 20 mm, the same penalty retains 96.6% of the true amplitude. If the purpose is to quantify a short ripple, improving the experiment can be preferable to changing the penalty. A derivative penalty $w_m \propto |k_m|$ encodes a stronger preference against fine structure; a constant mode must be handled separately if the voltage reference or the derivative penalty leaves it unconstrained.

What survives away from a uniform grid or simple geometry?

For arbitrary electrode shapes and detector positions, represent source or boundary data in a chosen basis $f = \sum_j a_j b_j$. Solve the forward field problem for each unit coefficient and sample it exactly as the real probe would, giving matrix entries A_{ij}. A noise covariance Σ weights the misfit through $\|\Sigma^{-1/2}(A\mathbf{a} - \mathbf{y})\|^2$. For an identity penalty apply the singular-value decomposition to the whitened matrix,

$$\Sigma^{-1/2} A = U S V^*.$$

Each right singular vector is a source pattern; its singular value tells how strongly it appears in the whitened measurement space. Multiplication by U^* diagonalises the data misfit and reduces the problem to the scalar filters above.

A zero singular value means the chosen sensors are insensitive to an entire source combination. No choice of numerical precision or regularisation changes that observational fact. A small nonzero value means one needs sufficiently small noise and modelling errors to identify the corresponding component. A practical resolution matrix $R = (A^*\Sigma^{-1}A + \alpha^2 I)^{-1}A^*\Sigma^{-1}A$ maps a noiseless true coefficient vector to the regularised estimate. A column of R shows which neighbouring coefficients respond when one ideal source coefficient is present. Testing such impulse responses describes the reconstruction's spatial smoothing more honestly than quoting the spacing of its output grid.

Use withheld measurements and identifiable physical quantities

Reserve some independently measured positions, heights or probe widths. Fit the rest and use the recovered candidates to predict the withheld measurements by applying the forward operator. The test is more informative when the withheld observations have different sensitivity patterns. A second distant scan can verify low modes while remaining insensitive to every high-frequency ambiguity. A second close scan or a different probe width can fill modes attenuated or nulled by the first. Correlated calibration errors, such as one common voltage offset, should be modelled as such, rather than being counted as many independent observations.

A distinct nonuniqueness occurs when trying to infer an arbitrary three-dimensional charge density solely from exterior voltages. Take any smooth compactly supported potential $g(\mathbf{x})$ inside the source region whose value and derivatives vanish outside. Add $\delta\rho = -\epsilon_0\nabla^2 g$ to a candidate charge distribution. Its induced potential is g and therefore vanishes at every external detector. Unless additional sign, support or constitutive assumptions constrain the source, infinitely many internal distributions remain consistent with exact exterior data. This is stronger than amplification of measurement noise. Total charge from a closed-surface flux or low multipole moments can still be identifiable when the internal density is not.

See an ill-conditioned inverse in a two-source instrument

Two possible source patches sit symmetrically beneath two voltage probes. A calibrated forward response, in volts of probe signal per chosen source unit, is $A = \begin{pmatrix} 1 & 0.98 \\ 0.98 & 1 \end{pmatrix}$. Each patch affects both probes almost equally. The sum pattern $(1, 1)/\sqrt{2}$ and difference pattern $(1, -1)/\sqrt{2}$ are orthonormal eigenvectors as well as singular vectors, with gains 1.98 and 0.02. The condition number is 99. The instrument localises total source strength much more reliably than it locates its division between the two patches.

Suppose the actual source vector is $(1, 0)$. The predictions are $(1.00, 0.98)$ V. A measurement perturbation $(+0.01, -0.01)$ V, though small compared with the signal, lies in the weak difference mode. The exact inverse interprets it as source error $(+0.50, -0.50)$, giving a reconstructed distribution $(1.50, -0.50)$. A common perturbation $(+0.01, +0.01)$ instead produces almost negligible source

changes $+0.00505$ each. A single rms error statistic for the voltage readings misses how their spatial correlation changes uncertainty in the reconstruction.

Use an identity Tikhonov penalty $\alpha = 0.10$. The response factors for the two true source directions become

$$W_{+} = 1.98^2/(1.98^2 + 0.10^2) = 0.9975,$$
$$W_{-} = 0.02^2/(0.02^2 + 0.10^2) = 0.0385.$$

In the absence of noise the estimated source from $(1, 0)$ is approximately $(0.518, 0.480)$. Both patches light up almost equally. This is the resolution matrix in action: the model predicts the observed voltages well yet cannot confidently identify which of the two patches generated them. Under the same opposite noise perturbation the additional reconstructed difference is only 0.0192 on each patch in opposite directions. A visually smooth result is reproducible but biased towards nearly equal strengths.

Add a third probe designed to respond as $(0.6, -0.6)$. This linear combination may be implemented physically by a differential electrode, so negative coefficients are meaningful. The squared singular gain of the old difference direction increases from 0.0004 to $0.0004 + 2(0.6)^2 = 0.7204$, while the common gain is unchanged. The new difference singular value is 0.849, reducing the ratio of the largest to smallest singular values to about 2.33 if the three channels have comparable independent noise in the specified units. Adding another copy of the first common-sensitive probe would have been much less effective. Designing measurements against the weak singular vector can improve source resolution far more than collecting many redundant readings.

Separate numerical error from height uncertainty and unknown backgrounds

Return to a sinusoidal source with negligible aperture. If its true height is h but one divides by an assumed transfer computed at $h + \delta h$, then even exact noiseless data return $\widehat{f}/f = e^{|k|\delta h}$. For a 4 mm wavelength and positive height error 0.10 mm, the fractional amplitude bias is

$$e^{0.157} - 1 = 17.0\%.$$

Unlike random noise this bias does not shrink by averaging repeated scans at the same miscalibrated height. A 20 mm calibration pattern at that same error suffers only about 3.2% bias, so testing one long-wavelength standard need not guarantee short-wavelength accuracy.

For an unknown but stable boundary pattern recorded at two known relative probe heights, $\hat{y}_m(h_2)/\hat{y}_m(h_1) = e^{-|k_m|(h_2 - h_1)}$ for modes above noise when the aperture is unchanged. The source amplitude and common aperture factor cancel. This measures a relative vertical displacement but not the absolute height if the true boundary amplitudes are unknown. Add a calibration electrode of known voltage and wavelength, or an independently measured mechanical height, to determine the absolute scale. Near a sensor's aperture zero the ratio is exceptionally noise sensitive and should not be used for the height fit.

Consider also a slowly varying stray background produced by distant conductors. In a source-free layer of finite thickness a Fourier mode can contain both $Ae^{-|k|z}$ and $Be^{+|k|z}$. The growing term is excluded in a bounded half-space continued to infinity, but can represent a field launched from the upper boundary of a finite domain. For a single spatial mode measured at two heights one can solve a 2×2 system for A, B. Its determinant is $2\sinh[|k|(h_2 - h_1)]$, up to a sign fixed by the column order. As the two heights coalesce their rows become nearly dependent. A measurement showing unexpected height scaling may thus indicate the wrong outer boundary in the forward model rather than a faulty inverse algorithm.

A reproducible inverse analysis can be audited along three independent axes. First manufacture noise-free readings from a source whose exact forward voltage is known, then refine only the forward discretisation. Second add noise with a specified amplitude and covariance and repeat enough draws to estimate recovered variation in identified modes. Third perturb physical nuisance parameters such as probe width, height and background independently of the noise draws. For each stage state what quantity is being recovered: total charge, a few spatial coefficients, or a real-space image biased by a chosen penalty. These stages expose different error mechanisms even when they yield similar changes in a plotted voltage trace.

Build an exactly invisible internal charge perturbation

The general construction of a silent source can be made explicit for a spherical support of radius a. Pick a smooth interior potential perturbation $g(r) = V_g[1 - (r/a)^2]^3$ for $0 \le r < a$, and set $g = 0$ outside. At $r = a$ both the potential and its first radial derivative vanish; its second radial derivative also approaches zero, avoiding unmodelled surface sheets. It is regular at the origin. Therefore g itself is the potential relative to infinity of the compact charge distribution $\delta\rho = -\epsilon_0 r^{-2} d(r^2 g')/dr$. With $x = r/a$, expansion gives

$$\delta\rho = \frac{\epsilon_0 V_g}{a^2}(18 - 60x^2 + 42x^4), \qquad 0 \le x < 1.$$

The perturbation is positive near the centre and negative in an outer band if $V_g > 0$. Its enclosed charge is

$$-4\pi\epsilon_0 r^2 g' = 24\pi\epsilon_0 V_g a\, x^3(1 - x^2)^2.$$

It vanishes at both ends of the radius interval and is positive at intermediate radii. Gauss's law outside then gives exactly zero electric field and the chosen reference gives zero exterior voltage. A numerical inverse recovering only zero perturbation from exterior data is observationally consistent with both the absence of charge and this particular structured distribution. Integrating the polynomial with its correct $4\pi r^2 dr$ weight confirms the exact zero total charge independently.

For $a = 20$ mm, $V_g = 0.10$ V, the internal centre voltage differs by 100 mV while every ideal exterior detector, no matter how accurately calibrated, records no difference whatsoever. At $r = a/2$, $g = 0.0422$ V and the enclosed perturbation charge is 9.39×10^{-14} C. An internal measurement can therefore break this particular ambiguity. Moving external probes closer while keeping them outside the sphere cannot. Positivity restrictions on the total density can reduce ambiguity, but adding a sufficiently small signed perturbation to a sufficiently positive background can preserve nonnegativity of both candidate densities and still leave their exterior data identical.

This example suggests a practical way to report results when full source reconstruction is impossible. Ask whether the quantity of interest changes along a silent perturbation. Total charge and all external moments stay unchanged here, whereas the central voltage and internal charge profile change. A claim about total charge can be appropriate even when detailed shell-by-shell densities are not identifiable. Conversely a fitted compact central peak may depend on the chosen smoothness penalty or positivity constraints even when every reported exterior measurement is reproduced.

A multiscale numerical exercise can demonstrate the effect reproducibly. Discretise both g and its associated $\delta\rho$, solve the independent forward Poisson problem, and evaluate internal and external probes. Coarse grids may show a small exterior signal if source deposition fails to preserve its neutral moments. Refine the source representation, plot the exterior residual separately from the interior voltage difference and compare with the analytic polynomial. A residual that tends to zero while the interior difference remains finite is the expected physical outcome. Using exactly the same discrete operator to manufacture data and to recover them can mask representation errors, so a second mesh, an analytic source integral or a differently implemented forward solver is especially valuable.

See which part of a specimen electrodes can resolve

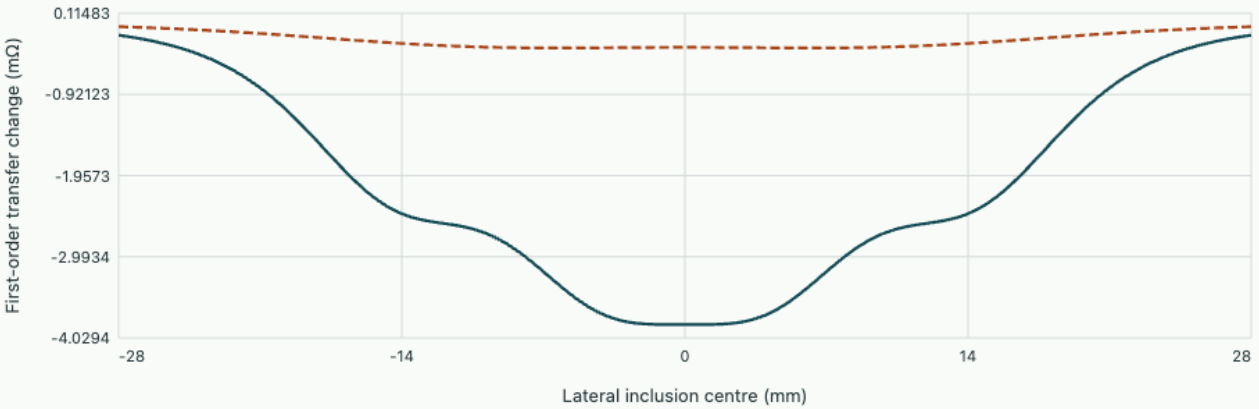
Compare two contact configurations and their sensitivity to a local material change. An accurate potential at the electrodes does not necessarily resolve the interior.

Open this experiment online

Try the calculation

Inclusion depth (mm) 8

Sensing half-spacing (mm) 5



Centre perturbation -3.858 mΩ; unperturbed transfer 79.58 Ω. Sphere radius 1 mm, conductivity increase 0.020 S/m.

Can two independent estimates agree for the wrong reason?

Separate common uncertainty from independent noise in two height scans

A probe measures a 10 mm wavelength mode at heights 1.0 and 2.0 mm with negligible aperture. The two Fourier amplitudes are 5.34 and 2.85 mV, each with independent 0.10 mV random uncertainty, but they share an unknown common voltage gain. Estimate the boundary amplitude and its random error under unit gain. Decide what the agreement can establish about calibration.

▼ **Derive and check the result**

The two transfers are

$$H_1 = e^{-0.2\pi} = 0.5335,$$
$$H_2 = e^{-0.4\pi} = 0.2846.$$

Each individually predicts about 10.0 mV at the boundary. For independent equal-variance amplitude noise, minimise $\sum (y_i - H_i f)^2$. Its solution is $f = (H_1 y_1 + H_2 y_2)/(H_1^2 + H_2^2)$, and random standard deviation $0.10/\sqrt{H_1^2 + H_2^2} = 0.165$ mV. Simply averaging the two directly inverted amplitudes gives the noisier second inversion equal weight and hence a larger variance.

The measured amplitude ratio is about 0.534, as predicted by the additional 1 mm of propagation. A common multiplicative calibration factor cancels from this ratio. If the true voltage gain were 1.10, both unit-gain inversions would agree at about 10 mV even though the physical amplitude would be 9.1 mV. Repeated scans reduce the random uncertainty but cannot remove this systematic scale. A known reference electrode or independent calibrated voltage source is needed.

If instead both probes share a small additive voltage offset in every spatial sample, a full ideal Fourier analysis would place that constant entirely in the zero mode. Incomplete sampling, finite windows or background leakage can mix the offset into the fitted sinusoid. Including a constant nuisance term in a weighted least-squares fit is one way to test that effect directly. Identifying the form of a shared error is as important as quoting its magnitude: common gain, common offset and relative height uncertainty enter the inverse problem in different directions and respond to different calibration experiments.

Tutorial · compare an experimental redesign with a stronger prior

Resolve a window null without changing sample spacing

A voltage camera averages over $w = 4$ mm. A suspected boundary pattern has wavelength 4 mm. Find its measured transfer and propose a second window width that recovers sensitivity. Include the effect of a 1 mm measurement height.

▼ **Work through the solution**

At equal aperture and wavelength the sinc argument is π, giving zero. With a 2 mm window the aperture transfer becomes

$$\sin(\pi/2)/(\pi/2) = 2/\pi.$$

Height attenuation is $e^{-2\pi(1/4)} = 0.2079$, so the combined transfer is 0.132. A 20 mV boundary amplitude therefore gives 2.65 mV in the second scan and exactly zero in the first idealised scan. Collecting more centres with the original width would still give zero. Calibrating the second aperture and its orientation is part of the forward model, especially near a sinc null where small geometry errors strongly change inference.

Choose a regularisation level by its permitted bias

For a mode with transfer 0.20 and unit penalty weight, require the mean reconstructed amplitude to retain at least 80% of the true value. Find the allowed maximum α. For 2 mV noise in the measured coefficient, compare inverse standard deviations with and without this penalty.

▼ **Work through the solution**

Set $W = 0.04/(0.04 + \alpha^2) \geq 0.80$. This gives $\alpha \leq 0.10$. At the limiting value the noise gain is $0.20/(0.04 + 0.01) = 4$, so the reconstructed noise standard deviation is 8 mV. Without regularisation the inverse noise is 10 mV. This modest improvement costs a 20% amplitude bias. The calculated tradeoff may or may not be acceptable for the physical question. Stating both numbers allows a reader to decide whether this experiment can support a claimed amplitude difference.

Further reading

- Measure the transfer of two spatial frequencies
- Reconstruct a tensor from finite electrostatic observations
- MIT, Smith · electric-field inversion and models of interacting electrodes
- Christian Clason · Regularization of Inverse Problems, spectral filters and Tikhonov regularisation

Weighted least squares and residual diagnostics

Inverse problems and least squares

Derive the fitting criterion from measurement error and separate the uncertainty of an estimated mean from the scatter of a new observation.

FROM FIRST PRINCIPLES

Derive the least-squares distance from the instrument noise

Suppose $y = Ax + \epsilon$, with Gaussian measurement noise of zero mean and known positive definite covariance Σ. Independent errors of unequal sizes should not count equally in raw units, and correlated errors should not be counted twice. Factor $\Sigma = LL^T$ and multiply the model by L^{-1}. The transformed noise has unit covariance,

$$b = L^{-1}y,$$
$$B = L^{-1}A,$$
$$\mathrm{Cov}(L^{-1}\epsilon) = I.$$

For a candidate state, the negative log likelihood differs by a constant from half the squared norm of the transformed residual. This follows from the Gaussian density,

$$\tfrac{1}{2}(y - Ax)^T\Sigma^{-1}(y - Ax) = \tfrac{1}{2}\|b - Bx\|_2^2.$$

The likeliest state therefore minimizes ordinary squared residuals in these whitened coordinates. Solves with the triangular factor L preserve this interpretation without forming its inverse. If the covariance is estimated or state dependent, the stated probability model and resulting uncertainty estimates need to include that extra information.

Background · Whitened residual

$$r = \Sigma^{-1/2}(y - G(m))$$

For a known positive covariance this expresses mismatch in noise units and removes the modelled correlations between measurements.

DERIVATION AND APPLICATIONS

Construct a model check that requires no fitted parameters

A residual can be compared with noise most cleanly when the particular combination of readings was forced to vanish by the model itself. Consider $y = Ax + \epsilon$ with zero-mean noise of known positive definite

covariance Σ. Choose a nonzero contrast c satisfying $A^T c = 0$. No value of the unknown x can change the predicted contrast because its mean is

$$\mathrm{E}(c^T y) = c^T A x = 0,$$
$$\mathrm{var}(c^T y) = c^T \Sigma c.$$

Under a Gaussian noise model its standardized value is a unit normal random variable. Squaring it gives a one-degree-of-freedom discrepancy,

$$Z = \frac{c^T y}{\sqrt{c^T \Sigma c}}, \qquad Z^2 \sim \chi^2_1.$$

The corresponding fit interpretation follows by whitening $b = L^{-1} y,\ B = L^{-1} A$. The vector $u = L^T c / \sqrt{c^T \Sigma c}$ has unit norm and is perpendicular to every column of B. Thus every fitted residual has projection Z along the same direction,

$$u^T (b - B\widehat{x}) = u^T b = Z, \qquad \|b - B\widehat{x}\|^2 \geq Z^2.$$

When there is exactly one independent residual direction, equality holds for the least-squares fit. The contrast is a model check only to the extent that the covariance and zero-mean error assumption are justified. It identifies a combination of inconsistent readings, without uniquely locating a faulty sensor.

WORKED EXAMPLE

Check an affine trend from three measurements

At times $0, 1, 2$ a linear trend predicts zero second difference. Observations $2, 3, 6$, each with independent standard deviation 0.20, give contrast $2 - 2(3) + 6 = 2$. Its standard deviation is $0.20\sqrt{6} = 0.490$, hence $Z = 4.08$ and $Z^2 = 16.7$. Fitting the intercept and slope cannot remove that discrepancy.

SECOND APPLICATION

Include a shared calibration mode in a conservation test

Three independent channels should satisfy a balance with contrast $c = (1, 1, -1)^T$. Suppose their independent errors each have deviation 0.10 and a common additive offset has independent deviation 0.20. Since $c^T (1, 1, 1)^T = 1$, the offset contributes 0.20^2 to contrast variance. For readings $1.0, 2.0, 2.6$ the discrepancy is $0.4 / \sqrt{0.03 + 0.04} = 1.51$ standard deviations. Dropping the shared offset would inflate it to 2.31.

A distance in data space depends on the instrument

Let $y = Ax + \epsilon$ with zero-mean error of known positive-definite covariance Σ. If the errors are jointly Gaussian, the negative log likelihood differs from $(y - Ax)^\mathsf{T} \Sigma^{-1} (y - Ax)/2$ only by constants independent of x. This expression still defines a weighted least-squares criterion without normality.

Correlated readings can vary together substantially while their difference is very precise; individually dividing them by standard deviations fails to capture that information.

Factor $\Sigma = LL^{\mathsf{T}}$ and solve $Lb = y,\ LB = A$ by triangular substitution. Then $b = Bx + \eta$ with $\operatorname{Cov}\eta = I$; $\|b - Bx\|^2$ is the original weighted distance. This is whitening. It also cures a units problem: residuals measured in different units are compared after division by their uncertainty scales. For truly independent readings of standard deviations s_i, the transformation simply divides each row and response by s_i.

Derive the projection and compute it without losing precision

Differentiating the squared norm gives

$$B^{\mathsf{T}}(b - B\hat{x}) = 0.$$

Thus $B\hat{x}$ is the orthogonal projection of b onto the column space. For full column rank write a thin QR factorisation $B = QR$ with $Q^{\mathsf{T}}Q = I$. The minimiser solves $R\hat{x} = Q^{\mathsf{T}}b$. Forming the normal matrix $B^{\mathsf{T}}B$ instead squares its two-norm condition number. QR avoids this extra source of precision loss, though it cannot create information absent from the measurements.

Suppose two instruments measure one constant as 10 and 14 with independent standard deviations 1 and 2. The weighted mean is $(10 + 14/4)/(1 + 1/4) = 10.8$, with standard deviation $(1 + 1/4)^{-1/2} = 0.894$. The less precise instrument still contributes, but with one quarter of the weight. Giving it half the weight would confuse standard deviation with variance. If both instruments share an unknown calibration bias, the stated independence model overstates what averaging can achieve.

Background · Why can two whitening transformations give the same fitted curve?

A symmetric inverse square root and a triangular Cholesky solve generally give different coordinates for the whitened residual. Let both transformations W_1, W_2 whiten the same positive definite covariance Σ. Then the change of coordinates $Q = W_2 W_1^{-1}$ obeys

$$QQ^T = W_2 \Sigma W_2^T = I.$$

Whitened residual vectors are therefore related by an orthogonal transformation. Their squared lengths and hence their least-squares minimisers agree. Two individual whitened components can nevertheless have different magnitudes. If a diagnostic needs to identify which physical sensor was faulty, interpret its residual in the original measurement coordinates alongside the correlation model. A large component obtained after rotation cannot by itself identify a physical channel.

Work with residuals after fitting

The fitted whitened response is Hb with $H = QQ^{\mathsf{T}}$, and residual $r = (I - H)b$ has covariance $I - H$. Consequently raw whitened residuals after fitting do not have equal independent unit variance.

Observation i has leverage H_{ii}; a high-leverage point can pull the fit toward itself and show a deceptively small residual. Under the stated model, $\mathbb{E}\|r\|^2 = m - p$ for m observations and p independent fitted coefficients. With Gaussian errors the sum has a chi-square distribution with those degrees of freedom. Plot signed residuals against input, time and sensor identity to look for omitted curvature, drift or serial correlation.

Uncertainty in a fitted trend and a future measurement

In the full-rank linear model $\operatorname{Cov}\hat{x} = (B^{\mathsf{T}}B)^{-1}$. For the model mean $c^{\mathsf{T}}\hat{x}$ the variance is $c^{\mathsf{T}}\operatorname{Cov}(\hat{x})c$; solve $R^{\mathsf{T}}v = c$ to obtain it as $\|v\|^2$. Predicting a future *individual reading* requires adding its independent measurement variance. This distinction matters when validating fitted response curves. A single point can legitimately lie far outside a narrow interval for the mean. When only relative weights are known, estimate a common variance scale from the residual degrees of freedom and state the additional assumption.

Separate common and independent errors before estimating a mean

Measurements made after one calibration often share a random offset. Write $y_i = \mu + b + e_i$, where the common offset has mean zero and variance τ^2, and the individual errors are independent of one another and of b, with variances s_i^2. The covariance matrix is the diagonal matrix of individual variances plus the rank-one matrix $\tau^2 \mathbf{1}\mathbf{1}^T$. Consider any linear estimate $w^T y$ that reproduces an arbitrary constant, so that $\sum_i w_i = 1$. Its variance separates into $\tau^2 + \sum_i w_i^2 s_i^2$. The first term survives every admissible choice of weights because the shared error is indistinguishable from the unknown mean within this set of readings.

Minimising the second term with a multiplier for the sum constraint gives

$$2 s_i^2 w_i = \lambda.$$

Normalisation yields weights proportional to the individual precisions, with variance $\tau^2 + 1/\sum_i s_i^{-2}$. In particular, estimating the weights from the marginal variances $s_i^2 + \tau^2$ as if every reading were independent gives both an incorrect covariance and generally incorrect relative weights. This derivation uses only second moments. Gaussian assumptions enter if the standard deviation is subsequently turned into a normal-theory probability interval.

Decide whether to repeat a reading or recalibrate

A balance has a random calibration uncertainty of 0.30 g. The first reading is 12.40 g with individual standard deviation 0.10 g. The second, acquired during a shorter averaging window, is 11.90 g with individual standard deviation 0.20 g. The individual precisions are 100 and 25 per g², so the relative weights are 0.8 and 0.2. The estimate is 12.30 g. Its variance is $0.09 + 1/125 = 0.098$ g², giving standard deviation 0.313 g. A long sequence of additional readings using this same calibration can only push that standard deviation toward 0.30 g.

The difference between these two readings removes both μ and the shared offset. Its standard deviation is

$$\sqrt{0.10^2 + 0.20^2} = 0.224\,\mathrm{g}.$$

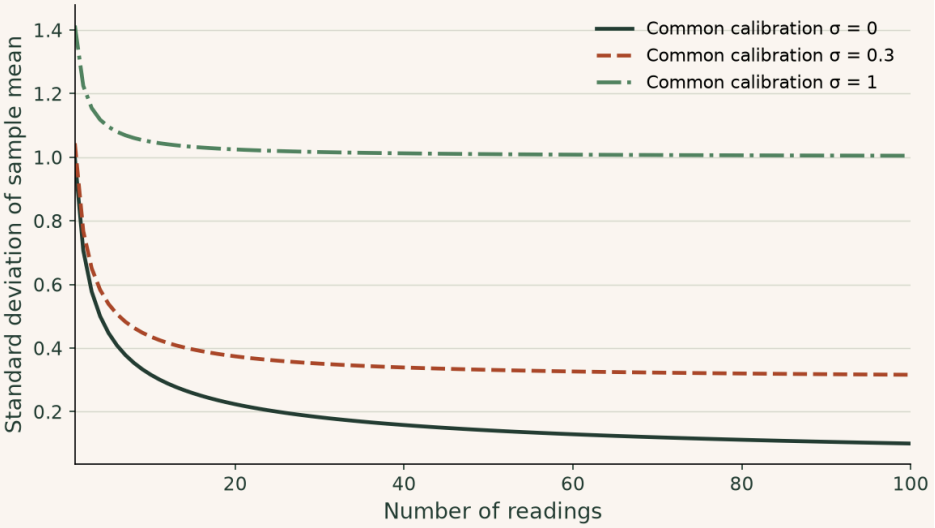
The observed difference of 0.50 g is about 2.24 such standard deviations. That discrepancy is relevant to the model for individual readings and averaging windows. It supplies no direct evidence for or against a common calibration offset of a particular size, because that offset has cancelled algebraically.

What if the entire two-reading procedure is repeated after an independent calibration? Averaging the two independently calibrated group estimates gives variance $0.098/2 = 0.049$ g² and standard deviation 0.221 g. Taking the same number of readings with one calibration instead gives variance $0.09 + 1/250 = 0.094$ g². The choice of sampling protocol changes the uncertainty even when the balance, total reading count and individual variances are identical. If calibrations share an uncertain reference standard, independence between groups must itself be revised.

A separate reference mass weighed during each calibration provides another way to estimate a shared offset. Subtracting that reference reading removes the offset but also introduces the reference measurement error into every corrected value. The corrected readings are therefore correlated again, now through the shared reference. Their covariance should be recomputed from this modified measurement protocol.

Find a measurement floor that averaging cannot remove

Each reading has independent noise of variance one and the same additional zero-mean calibration offset. Repeating the measurement divides the independent variance by the number of readings. The calibration error is shared by every term in their average.



With common standard deviation 0.3, averaging one hundred readings with independent sensor noise would still leave total standard deviation about 0.316. Fitting a diagonal covariance and extrapolating a 1/√N law misses this limiting uncertainty.

▶ **Read values from the plot**

Change parameters · Find a measurement floor that averaging cannot remove

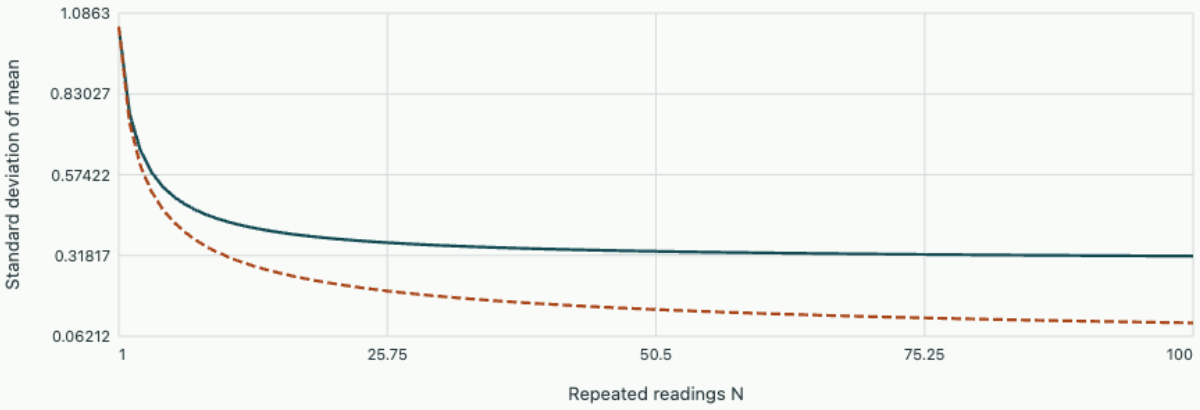
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Shared calibration SD **0.3**

Independent reading SD **1**



The limiting standard deviation is the shared calibration SD 0.3. Repeating the same instrument does not remove its common unknown offset.

How model error masquerades as a fitted physical parameter

Inverse problems and least squares

Project an omitted spatial effect into the fitted parameter space. Work out when residual checks reveal the omission and when they are silent.

A misspecified instrument

Suppose $d = Ax_* + b + \eta$, with zero-mean measurement noise and a deterministic but omitted contribution b. Assume A has independent columns. Least squares gives

$$\widehat{x} = (A^T A)^{-1} A^T d.$$

Taking expectations leaves the parameter bias $A^+ b$. The fitted residual is $(I - AA^+)(b + \eta)$. Thus the same model error separates into a component that biases the parameters and an orthogonal component that remains visible to residual checks.

For a single sensitivity column a, decompose $b = b_{\parallel} + b_{\perp}$, where $b_{\parallel} = a(a^T b)/(a^T a)$. Repeating the measurement reduces independent random error; the deterministic projection remains. Error bars computed solely from measurement variance can become more misleading as more readings are acquired.

Calibration with an omitted offset

Fit $d_i = mt_i$ at $t = (1, 2, 3)$ to observations actually generated by $d_i = 2t_i + 0.3$. The unmodelled constant has nonzero projection onto t. The bias is $0.3(1 + 2 + 3)/(1 + 4 + 9) = 0.12857$, giving fitted slope 2.12857. The residuals at the three points are approximately 0.17143, 0.04286 and −0.08571. Their signed trend suggests a missing intercept. At $t = 6$ the model predicts 12.7714 while the actual law gives 12.3; an apparently modest training residual has grown into a substantial extrapolation error.

Now use a discrepancy exactly proportional to time, $b_i = 0.1t_i$. The fitted slope becomes 2.1 with zero residual at every training point. Additional times under the same model cannot separate the true slope from this contribution. A separate calibration that suppresses the extra physical effect can.

Rotate the missing physics

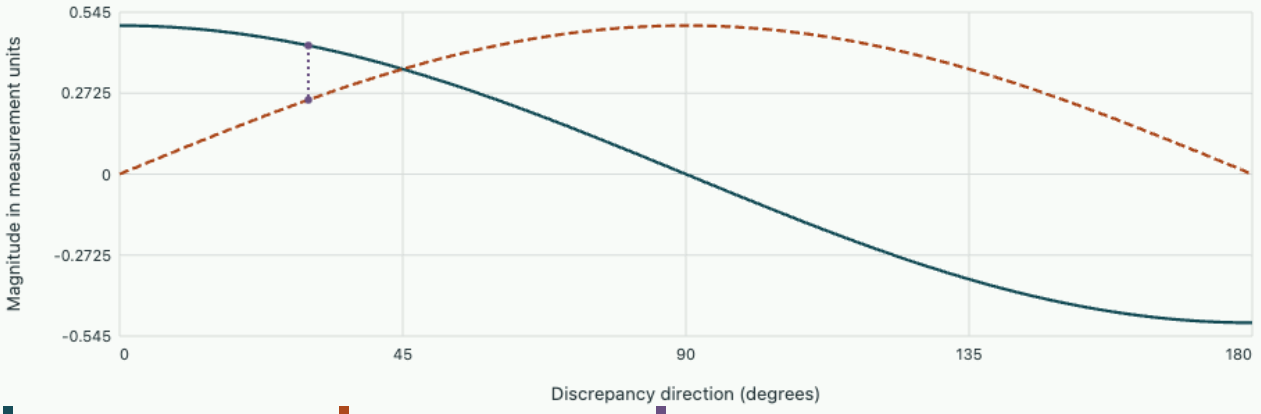
In a two-reading instrument, the model sensitivity is the unit horizontal vector. Vary the magnitude and direction of the omitted contribution. Observe parameter bias and visible residual separately, and locate cases with the same discrepancy norm but completely different fit diagnoses.

Open this experiment online

Try the calculation

Omitted-effect norm **0.5**

Direction (degrees) **30**



Signed parameter bias | Visible residual norm | Selected direction

At 30° the parameter bias is 0.43301 and residual norm 0.25.

Tutorial on experimental design

Repeat the offset example at symmetric times $(-1, 0, 1)$. Which part of the fitted result changes?

▼ **Worked solution**

The constant is now orthogonal to t, giving zero slope bias. The residual is a constant 0.3 in every reading. The experiment has made an offset visible without corrupting the slope.

Can adding an unrestricted discrepancy parameter for every datum solve the identifiability problem?

▼ **Worked solution**

It permits exact fits for any physical parameter by assigning the remaining signal to the discrepancy. Constrain the discrepancy from theory, independent data or a justified prior to extract the physical parameter.

Read the university treatment

Compare errors that project into a fitted singular direction with errors seen in the orthogonal data residual. Per Christian Hansen · Discrete Inverse Problems

Sampling, propagation and unresolved scales

Apply quadrature and spectral approximation to a real aperture, a decaying field and a radiated-power measurement. Look for an observed convergence rate and explain any plateau.

1. Diffraction from the near field to the focal plane
2. What can a mesh measure about a point charge?
3. Separate diffraction from a finite observation window
4. When differentiating a measured dipole invents radiated power
5. Verification, refinement and local error
6. What a periodic grid does to a decaying Fourier mode
7. Diffraction as a Fourier transform
8. How a grating separates wavelengths
9. Scattering from a central potential and what angular data can recover
10. Action, angle and the slowing clock at a separatrix
11. What a momentum eigenstate means in a finite experiment
12. How a turning point supplies the missing semiclassical phase
13. Measure gravity with a pendulum of finite amplitude

Diffraction from the near field to the focal plane

Waves, radiation and relativity

An illuminated aperture is a boundary condition for a wave. A screen a millimetre away, an intensity camera at a lens focus and a microscope measuring an evanescent field sample different consequences of the same condition. We will derive operators for propagating the field, estimate the approximations each instrument makes and use energy and sampling checks to tell numerical structure from a measurable fringe.

How each spatial frequency propagates

Take a monochromatic scalar field with physical value given by the real part of $U(x,y,z)e^{-i\omega t}$. This approximates one transverse component when the polarisation remains nearly uniform and angles are small. In a homogeneous transparent half-space it obeys the Helmholtz equation. Specify the complex amplitude in the plane $z=0$ and require outgoing or decaying waves for positive z. Fourier transform the two transverse coordinates, leaving a differential equation in the third.

$$\partial_z^2 \widehat{U} + (k^2 - k_x^2 - k_y^2)\widehat{U} = 0,$$

$$\widehat{U}(k_x,k_y,z) = \widehat{U}(k_x,k_y,0)e^{ik_z z},$$

$$k_z = \sqrt{k^2 - k_x^2 - k_y^2}.$$

Choose the root with positive real part for propagation and positive imaginary part for decay. Each propagating component advances in phase. Hence power integrated over a plane is unchanged for a paraxial beam with no absorbing aperture downstream. For angles approaching grazing incidence the weight for normal flux contains the direction cosine. Components with transverse wavenumber exceeding k decay exponentially. An aperture map thus contains structures invisible to a distant measurement even if the map itself is known exactly.

In vacuum a sinusoidal field structure of period 400 nm illuminated at 600 nm has $k_x/k = 1.50$. Its amplitude has fallen to $1/e$ after $1/(k\sqrt{1.5^2-1}) = 85.4$ nm. At 500 nm the remaining amplitude is 0.00287 and the squared amplitude is roughly eight parts in a million of its plane value. A numerical back-propagation that reinstates the original component also multiplies measurement errors by about 350. Improving detector gain does not restore the unknown phase.

A propagation kernel with a controlled phase error

For small transverse wavenumber, expand $k_z = k - k_\perp^2/(2k) - k_\perp^4/(8k^3) + \cdots$. The first correction becomes a chirp in Fourier space. Completing the square in its inverse transform gives the Fresnel kernel. Write $\mathbf{r}$ for a transverse observation point and $\boldsymbol{\xi}$ for a point in the aperture.

$$U(\mathbf{r},z) \simeq \frac{e^{ikz}}{i\lambda z}\int U(\boldsymbol{\xi},0)\exp\left[\frac{ik}{2z}|\mathbf{r}-\boldsymbol{\xi}|^2\right]d^2\xi.$$

This expression retains phase curvature across the aperture. The coefficient ensures that a spatially uniform boundary amplitude reproduces a uniform plane wave. An uncovered infinite plane is a useful test for a numerical convolution, although a finite computation needs padding or a truly periodic uniform test. A measured slit has sharp edges with broad spatial spectra, so peak error near an edge can exceed the error suggested by a small angle evaluated for the central lobe.

At an observation angle θ a propagation distance z accumulates neglected phase of order $kz\sin^4\theta/8$. For $z = 0.10$ m, $\lambda = 500$ nm and $\theta = 0.01$ rad this is 0.00157 rad. At 0.05 rad it is close to one radian. A faint sidelobe at a moderately large angle can require exact angular-spectrum propagation even while a paraxial estimate describes the central spot beautifully.

Which quadratic term can the far field lose?

Expand the kernel's square into an observation phase, a source phase and a bilinear term. Only the source phase can be dropped without changing the predicted intensity directly through that factor. For source radius a its maximum value is $\pi a^2/(\lambda z)$. The remaining integral is the Fourier transform of the aperture at spatial frequencies $x/(\lambda z), y/(\lambda z)$. Define the Fresnel number using the half-width a. A convention based on full slit width gives a value four times larger.

$$U(\mathbf{r}, z) \simeq \frac{e^{ikz + ikr^2/(2z)}}{i\lambda z}\, \widehat{U}\!\left(\frac{kx}{z}, \frac{ky}{z}, 0\right),$$

$$N_F = \frac{a^2}{\lambda z}.$$

For a uniformly illuminated slit of full width $w = 0.20$ mm at 532 nm and $z = 1.0$ m, $N_F = 0.0188$. Direct integration of the linear aperture phase gives

$$I/I_0 = \operatorname{sinc}^2(\pi w x/\lambda z),$$

where $\operatorname{sinc}(u) = \sin(u)/u$. The first zeros should be 2.66 mm either side of centre. The edge phase is 0.059 rad. Moving the screen to 30 mm gives $N_F = 0.627$ and edge phase 1.97 rad. Substituting the smaller distance in the far-field formula yields a number but no longer a controlled approximation.

Compare near- and far-field predictions in the slit integral

In that experiment reduce the mesh interval before moving the screen. If the exposed variable counts panels, doubling panels at fixed aperture halves the interval. A converged difference between the full and far-field integrals is a physical failure of the approximation. A feature that moves each time the number of source samples doubles is a numerical failure. Keep the same physical horizontal axis and the same normalisation when comparing adjacent screens.

From a circular pupil to a resolution measurement

When the pupil is a disk, polar coordinates considerably simplify the Fourier integral. Align the transverse transform vector along the horizontal axis. The angular average of $e^{-iqr\cos\phi}$ defines $J_0(qr)$. One radial

integration, using $d[uJ_1(u)]/du = uJ_0(u)$, then gives the result. Divide by the on-axis amplitude, proportional to aperture area.

$$\frac{U(q)}{U(0)} = \frac{2}{a^2}\int_0^a rJ_0(qr)\,dr = \frac{2J_1(qa)}{qa},$$

$$q = k\sin\theta.$$

At $qa = 0$ use $J_1(s) \sim s/2$. The first zero is at $qa = 3.8317$. Hence a disk of diameter D has its first dark ring at $\sin\theta = 1.22\lambda/D$. Two mutually incoherent identical points contribute the sum of their intensities. Rayleigh's conventional separation places each peak on the first zero of its neighbour. A coherent pair instead has an interference term that depends on the unknown relative phase. The same separation criterion cannot predict its measured trough without that phase.

A 5.0 mm pupil on a 100 mm focal-length lens at 550 nm gives a first-zero radius of 13.4 micrometres on the focal plane. A camera with 6.5 micrometre pixels spans the full bright core with about four pixels. Fitting the central pixel alone to find the image position would depend strongly on whether the spot lay on a pixel centre or a corner. A joint fit to integrated pixel intensities can estimate a position more precisely than the image width when enough photons are detected. That precision describes one isolated spot and does not directly give a two-source resolution.

Why a lens puts the spectrum close at hand

A thin positive lens adds the quadratic phase $-kr^2/(2f)$ to a field immediately in front of it. At the plane $z = f$, this cancels the source term of the Fresnel kernel exactly within paraxial theory. The focal intensity therefore samples a Fourier spectrum even for a tabletop lens standing much nearer than the free-space far-field distance. The physical exit pupil remains part of the amplitude. Closing it removes high transverse spatial frequencies in an imaging arrangement and broadens the spot of a point emitter.

For illumination with Gaussian amplitude $U \propto e^{-r^2/w^2}$, completing the transverse Fourier integrals gives another Gaussian, with focal-plane amplitude radius $w_0 = \lambda f/(\pi w)$ for a broad lens and an incident flat wavefront. If $w = 1.0$ mm, $f = 100$ mm and $\lambda = 633$ nm, $w_0 = 20.1$ micrometres. This is the $1/e$ field radius, also the $1/e^2$ intensity radius. The Rayleigh length $\pi w_0^2/\lambda$ is 2.02 mm. Moving a detector one Rayleigh length off the waist increases the radius by $\sqrt{2}$ and halves its on-axis intensity for fixed total beam power.

Use the ray-bundle experiment to locate the geometric focus before estimating its diffractive width

From a point image to transmitted object contrast

A microscope may resolve individual test objects yet deliver very little contrast for a periodic specimen at a similar length scale. The reason is that incoherent imaging acts linearly on intensity, while coherent imaging acts linearly on field amplitude. Start with mutually incoherent object points. Their expected image intensities add, so the image is the convolution of object intensity with a nonnegative intensity point-spread function $h = |h_A|^2$, where h_A is the coherent amplitude impulse response. The Fourier

transform of h, normalised to one at zero spatial frequency, is the optical transfer function. Its magnitude describes attenuation of a sinusoidal intensity modulation.

For a focused unaberrated system the coherent amplitude response in Fourier space is the pupil function. The convolution theorem applied to $|h_A|^2$ shows that the incoherent transfer function is the pupil autocorrelation. For a clear circular pupil this amounts to asking how much two equal disks overlap as their centres separate. The overlap consists of two circular sectors minus their triangular parts. With normalised frequency $s = f/f_c$ and incoherent cutoff $f_c = 2\,\mathrm{NA}/\lambda_0$ in object space, the result is

$$M(s) = \frac{2}{\pi}[\arccos s - s\sqrt{1-s^2}], \quad 0 \le s \le 1,$$

$$M(s) = 0 \quad (s > 1).$$

Here λ_0 is vacuum wavelength and $\mathrm{NA} = n \sin\theta_{\max}$. At $s = 0$ the disks coincide and the full mean intensity passes. At the cutoff their edges just touch and sinusoidal contrast vanishes. The coherent pupil transmits field spatial frequencies only through NA/λ_0. These two cutoffs answer different questions about different object and illumination statistics. A coherent phase grating viewed in exact focus can yield an almost uniform intensity image despite transmitting several field components.

For $\mathrm{NA} = 0.25$ at 550 nm, the incoherent cutoff is 909 cycles/mm. A sinusoidal target of 450 cycles/mm has $s = 0.495$ and $M \simeq 0.396$. An object with 0.60 sinusoidal intensity visibility is predicted to show about 0.238 before detector averaging. If the object-referred pixel width is 0.75 micrometres, rectangular pixel integration multiplies this contrast by $\mathrm{sinc}(\pi f p) = 0.823$; the recorded visibility is about 0.196. Nyquist frequency for this pixel spacing is 667 cycles/mm. Hence even some frequencies that the optics can transmit lie above the detector Nyquist limit and can alias back to lower apparent frequencies.

This prediction can be tested with an adjustable magnification while keeping entrance pupil and object fixed. Object-referred pixel width falls as magnification rises; a pixel-limited measurement should approach the optical transfer curve. Raising magnification also spreads light over more pixels, changing per-pixel count noise, so compare fitted contrasts with uncertainty. At frequencies where the ideal transfer is small, a high-contrast target and low read noise matter. Dividing by an arbitrarily small transfer function to sharpen the final image can turn noise into a convincing periodic pattern.

A pupil-phase error redistributes energy without absorbing it

Introduce a known defocus into a clear circular pupil. Let the phase relative to its centre vary as $\phi(r) = \Phi r^2/a^2$. The on-axis focal amplitude is proportional to the pupil average of $e^{i\phi}$, since all ideal contributions arrive in phase there. Integrate in polar coordinates and substitute $u = r^2/a^2$. Relative to the perfect on-axis intensity the result is

$$S = \left| \int_0^1 e^{i\Phi u} du \right|^2 = \left[\frac{\sin(\Phi/2)}{\Phi/2} \right]^2.$$

This special-case Strehl ratio assumes a uniform pupil and pure quadratic defocus. Removing a phase piston makes no change to intensity. Expanding for small Φ yields

$$S = 1 - \Phi^2/12 + \cdots.$$

The pupil-phase variance after removing the mean piston is $\Phi^2/12$, which explains the leading connection between reduced peak intensity and wavefront variance without assuming that every type of aberration has the same pattern.

A quarter of a wave of centre-to-edge optical path error gives $\Phi = \pi/2$ and $S = 0.811$. One half wave gives 0.405. In an ideal phase-only pupil the missing on-axis intensity appears elsewhere in the image; total transmitted power over a sufficiently wide observation plane is unchanged. Contrast this with an absorbing pupil that removes a uniform fraction of intensity everywhere. Measuring only the brightest pixel confuses the two. Measure integrated power over increasing radii and fit a spatial profile as well as a peak.

To turn this into an axial calibration, derive the edge phase from the distance mismatch. For a paraxial marginal ray in air making angle approximately NA, an axial shift Δz introduces a relative phase magnitude near $k\,|\Delta z|\,\mathrm{NA}^2/2$. At $\mathrm{NA} = 0.10$, $\lambda = 500$ nm and $|\Delta z| = 25$ micrometres this is 1.57 rad, the quarter-wave case. With a larger numerical aperture use actual longitudinal wavevectors instead of the quadratic approximation and model polarisation where required. The formulas predict the same on-axis drop for positive and negative defocus. A single peak reading cannot give the sign of an axial misplacement; scanning through focus supplies that missing sign and can reveal an asymmetric aberration.

Distinguish aperture physics from quadrature error

Refine the aperture samples and vary their physical extent independently. A moving numerical null and a true diffraction minimum have different refinement signatures.

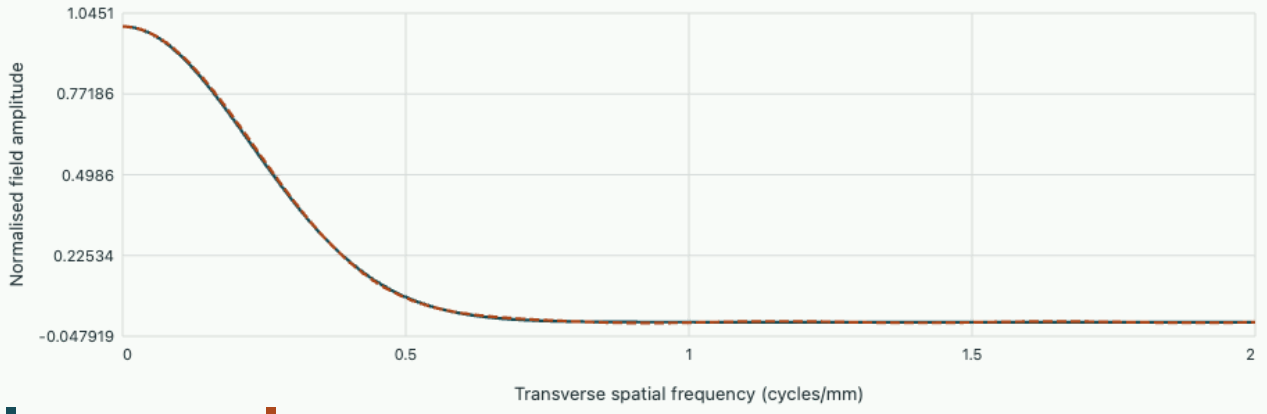
Open this experiment online

Try the calculation

Gaussian aperture radius w (mm) 1

Window half-width / w 2

Samples across window 30



Sample spacing 0.13333 mm, Nyquist frequency 3.75 cycles/mm. Measured amplitude at 1 cycle/mm: -0.0014811.

Worked optical problem · determine how much light fits into one Airy disk

Two groups compare image quality using respectively the peak intensity and the energy contained inside the first dark ring. Derive the encircled-energy fraction for a uniform circular pupil and obtain its value at the first ring. Would closing the aperture increase the useful energy fraction on a detector of fixed physical radius?

▼ **Integrate the radial intensity with its area measure**

Write $v = ka \sin\theta$, and use the small-angle focal-plane variable proportional to image radius. Focal intensity varies as $[2J_1(v)/v]^2$. The annular area element supplies a factor $v\,dv$. Use $J_0' = -J_1$ and $J_1' = J_0 - J_1/v$ to differentiate $J_0^2 + J_1^2$. The derivative is $-2J_1^2/v$, giving the fraction inside radius v as $1 - J_0(v)^2 - J_1(v)^2$. It tends to zero at the origin and to one at very large radius, which fixes the integration constants and normalisation without needing an absolute incident amplitude.

At $v = 3.8317$, $J_1 = 0$ and $J_0 \simeq -0.4028$. The encircled fraction is therefore 0.838. The remaining 16.2 percent is outside the central disk even in a perfect ideal image. If the physical detector radius corresponded to this value before closing the pupil, reducing the pupil radius decreases v at the detector edge and therefore decreases the collected fraction. It also ordinarily reduces the total incident power transmitted by the pupil.

Peak-normalised plots rescale each pupil's image to one at its centre, which hides both effects. For a fair power comparison record the total incident flux, the pupil transmission and the energy on the fixed

detector. A phase aberration or obscuration can leave the nominal first-zero position a poor boundary for comparing different pupils. The same physical encircled radius offers a reproducible metric in those cases, provided the beam centre and background are estimated independently.

A phase ramp moves a diffraction pattern without changing its shape

Illuminate a thin aperture with a tilted plane wave in the paraxial regime. At the aperture plane this multiplies its complex amplitude by a linear phase ramp. The Fourier shift theorem then moves the entire focal-plane amplitude by a transverse amount $f \sin \theta_i$ along the tilt direction. Its intensity pattern retains its shape as long as the illumination envelope fills the pupil uniformly and the lens has enough acceptance. Tilting a narrow 532 nm beam by one milliradian in front of a 100 mm lens moves the pattern by about one tenth of a millimetre.

This provides a practical separation of phase tilt from aperture narrowing. A tilt moves both first zeros in the same direction, whereas changing the width of a centred slit moves the zeros apart or together. A position fit using one zero alone cannot separate these perturbations. Jointly fitting both zeros yields a mean position sensitive to tilt and a separation sensitive to width. A scanned detector with unequal position calibration on the two sides can mix the two observables.

For a spatially incoherent angular source, integrate intensities of the shifted patterns over the angular source distribution. An extended source then blurs zeros that would be sharp under one coherent angle. A coherent superposition of the same angular components instead requires adding the shifted amplitudes and can produce additional fringes. The illumination statistics determine which calculation describes the recorded image.

Tutorial · Distinguish a pupil defect from pixel integration

In a slit focal pattern the central lobe has one-sided first-zero distance 40 micrometres. An analyst samples precisely at that first zero using a centred pixel 8 micrometres wide and finds positive intensity. Estimate the finite-pixel fraction and propose a measurement that would expose an additional scattered background.

▼ Show the pixel calculation and a discriminating test

Let $x = x_0 + s$, where $|s|$ is much smaller than x_0. The sinc amplitude near its zero is approximately $-s/x_0$, giving intensity $I/I_{\mathrm{peak}} \simeq s^2/x_0^2$. Averaging across a pixel of width p gives $p^2/(12x_0^2)$, about 3.33×10^{-3}. This estimate uses a flat response within each pixel and a sample centred on the true minimum. A two-micrometre centring error contributes about another 0.0025 at leading order.

Use magnification that doubles the first-zero distance in pixel units while keeping pupil and wavelength fixed. Pixel leakage at a centred zero should then fall by four relative to the peak. A background proportional to local scattered light may respond differently and a dark offset can be estimated from

opaque-frame counts. Calibrate differing exposure and throughput, subtract dark counts prior to normalisation and fit both flanks to locate the true minimum.

Tutorial · Design a propagation-grid check

A 512-point transverse grid with 2 micrometre spacing represents an aperture at 600 nm. What angle can the grid represent? Could its periodically repeated boundary contaminate a nominal 0.08 rad bundle after 10 mm propagation, if the beam spreads from the centre? Explain two refinements that test distinct errors.

▼ **Show the angular and spatial budgets**

Nyquist gives $|k_x| \leq \pi/\Delta x$, hence

$$|\sin\theta| \leq \lambda/(2\Delta x) = 0.15.$$

The 0.08 rad direction is resolved in principle, although modes near Nyquist would need a separate resolution study. Its lateral displacement over 10 mm is about 0.802 mm, larger than the 0.512 mm half-window even for a beam initially confined to its centre. A periodic FFT can therefore wrap light around into the opposite side.

Double the number of points with spacing fixed. This doubles the window while keeping the represented angular limit fixed. Then halve the spacing while doubling the number again to leave this enlarged window fixed and test transverse discretisation independently. Compare conserved flux, central intensity and the location of the first minimum, retaining absolute normalisation. Flux conservation in a periodic window can hold perfectly even when light has wrapped across its boundary. A plot of that single diagnostic would miss the contamination.

MIT 2.71 Optics · Fourier optics, apertures and propagation

What can a mesh measure about a point charge?

Numerical field laboratory

Distinguish conserved charge and measurable exterior fields from a numerical peak that diverges as the source becomes more localised.

Ask three detectors different questions

Imagine three instruments looking at a small charged body in vacuum. One measures the voltage on a spherical probe outside it, one measures the integrated electric flux through a closed surface, and one estimates the largest electric field somewhere near the source. If the body is idealised as a point, which readings have finite limits? The distinction matters when a calculation displays a dramatic change in maximum field after a perfectly reasonable mesh refinement.

For a positive point charge q, the exterior potential $V = kq/r$ tends to infinity as the observation point approaches the source. Its radial field magnitude kq/r^2 also diverges. Gauss's law gives a different answer for the flux through any sphere of positive radius:

$$4\pi r^2 E = q/\epsilon_0.$$

A finite integral and an unbounded maximum are compatible. Knowing what the instrument measures determines which numerical output should be expected to converge.

Use a smooth source with a known answer

Replace the potential by $V_a = kq/\sqrt{r^2 + a^2}$, with $a > 0$ a distance. This is a particular physical regularisation. Differentiating before sampling gives

$$E_a = kqr/(r^2 + a^2)^{3/2}.$$

Divide the flux by q/ϵ_0. The putative enclosed-charge fraction becomes

$$F_a(r) = \frac{r^3}{(r^2 + a^2)^{3/2}}.$$

To recover the corresponding density, differentiate $qF_a(r)$ with respect to radius and divide by shell area. The terms involving r^4 cancel, leaving $\rho_a(r) = 3qa^2/[4\pi(r^2 + a^2)^{5/2}]$. This density is smooth, nonnegative and integrates over all space to q, since $F_a(0) = 0$ and $F_a(\infty) = 1$. It has an extended tail. The parameter a is a smoothing scale rather than an outer radius where charge abruptly stops.

At $r = a$ only $1/(2\sqrt{2})$, about 35.36%, of the charge is enclosed. At $r = 2a$ the fraction is about 71.55%. A common intuition that two smoothing distances already contain essentially everything fails for this distribution. Measure the probe radius in units of the source size before blaming the flux calculation.

Hold the observable fixed while refining

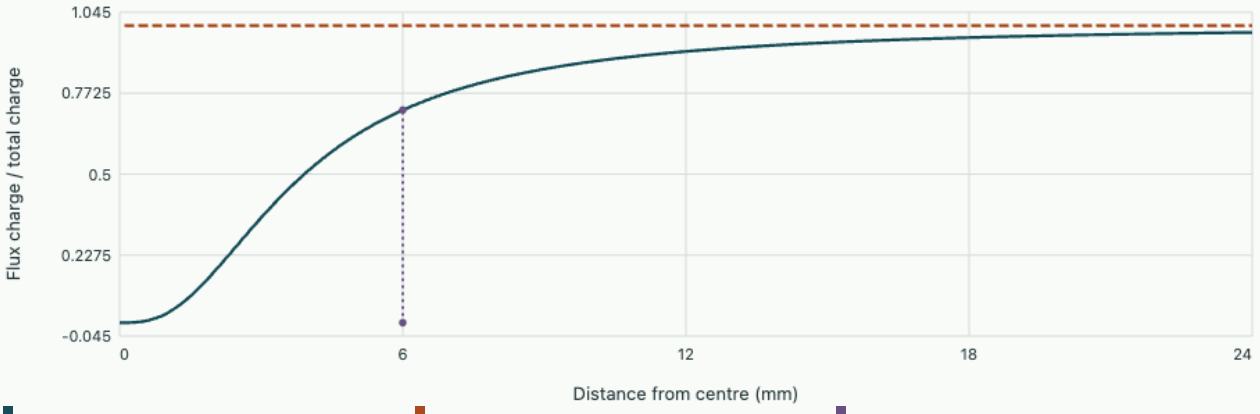
The exterior-field ratio relative to the true point charge is $[1 + (a/r)^2]^{-3/2}$. At fixed physical r, expanding the bracket gives $1 - \frac{3}{2}(a/r)^2 + O((a/r)^4)$. Thus a probe far enough away sees a second-order regularisation error. Locating the maximum of E_a, however, requires setting $a^2 - 2r^2 = 0$, giving $r = a/\sqrt{2}$ and a peak proportional to $1/a^2$. Refine a and move the observer with the maximum: the reported peak grows even as a fixed exterior probe reading converges.

Open the live experiment

Try the calculation

Charge-cloud radius a (mm) 3

Probe radius r (mm) 6



The probe encloses 71.554% of the charge. Reducing a by half gives 91.308% at the same physical probe.

Keep the probe at 6 mm and reduce a from 3 to 1.5 mm. The flux rises from about 71.55% to 91.31% of the point-charge value. Now instead halve both the probe radius and a. The measured fraction stays the same. The two procedures have distinct limiting observations; writing “halve the mesh” without recording the source representation and probe definition would make the experiment irreproducible.

Use continuity as a numerical diagnostic

Suppose a code deposits a prescribed charge into nearby cells and then solves for a nodal potential. Summing deposited cell charges tests normalisation of the source. Summing outward face fluxes over a larger box tests a separate constraint on the computed field. A correct finite-volume implementation cancels interior face pairs. Interpolating a nodal field to a smooth plotting surface with an unrelated quadrature need not conserve the same discrete quantity. A box-flux test using the solver's own numerical face fluxes and a smooth-surface test using the reconstructed field answer complementary questions.

A second check concerns energy. For a true point charge the field-energy integral near the origin scales as $\int_0 r^2 (r^{-2})^2 dr$, which diverges. For a body of radius a this short-distance contribution typically scales like $1/a$. Comparing two source regularisations through their total self-energy without specifying a physical charge radius is therefore an ill-defined validation exercise. Interaction energy with a smooth distant field, in contrast, can have a well-defined point-particle limit.

Tutorial

How far from this smooth charge should a flux probe be placed to enclose 99%?

▼ **Work through the solution**

Solve $[1 + (a/r)^2]^{-3/2} = 0.99$. This gives

$$r/a = [0.99^{-2/3} - 1]^{-1/2} \simeq 12.2.$$

The answer depends on the chosen source shape. A uniform finite ball would have a different exterior behaviour.

A code uses $a = 2h$ and reports the field one cell away, $r = h$, on every refined mesh. What do you expect for its flux fraction and pointwise field?

▼ **Work through the solution**

The fixed ratio $a/r = 2$ gives $F = 5^{-3/2} \simeq 0.0894$ at every resolution. The field still scales as r^{-2} times that factor and grows like h^{-2}. The observer is moving towards the singular source.

If total deposited charge differs from q by 2% independently of mesh size, can adjusting a cure the exterior result?

▼ **Work through the solution**

At large radius the flux follows the total charge that was actually deposited. Changing width cannot repair an incorrect integral. Fix the source normalisation and cell-volume weights before interpreting a convergence slope.

Separate diffraction from a finite observation window

Numerical field laboratory

Recover the far-field angular spectrum of a smooth aperture, then isolate the errors from truncating and sampling the aperture field.

Begin with a transform that can be checked analytically

Consider a long slit system whose transmitted scalar field varies along one transverse coordinate x. At a distant observation point in direction θ, a small aperture element accumulates an approximate relative phase $-kx\sin\theta$. The Fraunhofer amplitude is proportional to $\int A(x)e^{-2\pi i f x}\,dx$ with spatial frequency $f=\sin\theta/\lambda$. The approximation neglects the quadratic phase across the aperture; either a suitable lens or sufficient propagation distance is needed. A numerical Fourier transform cannot cure violation of this optical assumption.

Take the ideal aperture field $A(x)=e^{-x^2/w^2}$ and normalise far-field amplitude by its on-axis value. Integrating by parts or completing the square gives

$$F(f)/F(0)=e^{-(\pi w f)^2}.$$

One derivation avoids any complex contour: differentiate F with respect to f and replace $xA=-(w^2/2)A'$. Integrating by parts gives

$$F'=-2\pi^2 w^2 f F.$$

Solve this first-order equation with the measured $F(0)$. Thus a wider beam on the aperture narrows the far-field distribution. Squaring the amplitude gives the intensity measured by an ordinary detector.

At $w=1$ mm, the $1/e$ amplitude spatial frequency is $1/(\pi w)\simeq 0.3183$ cycles/mm. For 500 nm light this is a small angle of 1.59×10^{-4} radians. An instrument calibrated in radians per millimetre rather than cycles per millimetre would need a factor 2π; stating the Fourier convention prevents that frequent scale discrepancy.

What the instrument has really sampled

A scanner covers $-W\le x\le W$ and gives one amplitude at each of N equally spaced cell midpoints. The numerical transform is the weighted sum $\Delta x\sum_j A(x_j)e^{-2\pi i f x_j}$, where $\Delta x=2W/N$. By symmetry its imaginary part vanishes for a perfectly centred real Gaussian; the program sums the cosine projection. It normalises by the numerical on-axis amplitude, allowing us to compare shapes on a common vertical scale.

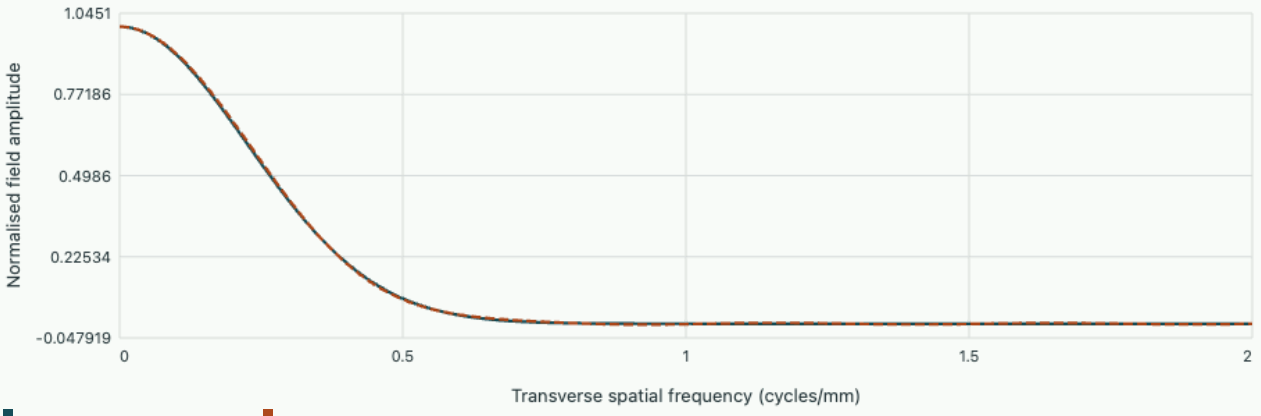
Two separate approximations have entered. Truncating the aperture multiplies the field by a rectangle. In frequency space this convolves the Gaussian spectrum with a sinc and can produce signed sidelobes. Sampling at regular intervals makes the spectrum repeat up to a phase convention, with alias copies separated by $1/\Delta x$. Increasing W at fixed N reduces truncation but worsens sampling. Widening the field of view and adding more detector pixels are not interchangeable changes.

Design a convergence test with two axes

Open the live experiment

Try the calculation

Gaussian aperture radius w (mm)	1
Window half-width / w	2
Samples across window	30



Sample spacing 0.13333 mm, Nyquist frequency 3.75 cycles/mm. Measured amplitude at 1 cycle/mm: -0.0014811.

Keep $w = 1$ mm and select 40 samples. Compare $W = w$ and $W = 3w$. The aperture field at the boundary falls from e^{-1} to e^{-9}, so an apparent sidelobe associated with the abrupt window should decrease strongly. Next fix $W = 3w$ and lower the number of samples. The displayed Nyquist frequency falls as Δx grows. When a plotted point lies above it, a smooth-looking value is an aliased sample of the periodic discrete spectrum.

At $W = 2w$, a centred Gaussian retains the on-axis amplitude fraction $\operatorname{erf}(2) \simeq 0.9953$ if integrated exactly. The normalised plot hides this 0.47% loss at zero frequency. If absolute power calibration matters, repeat the experiment without normalisation and independently integrate $|A|^2$ across the aperture. A shape comparison and a throughput comparison are different measurements, both meaningful.

Recover a size from a far-field measurement

Suppose a detector measures the intensity falling to one half at $f_{1/2}$. Since $I/I_0 = e^{-2(\pi w f)^2}$, the inferred w is $\sqrt{\ln 2/2}/(\pi f_{1/2})$. For $f_{1/2} = 0.1874$ cycles/mm the radius is about 1 mm. If the instrument actually reported amplitude half-width the corresponding formula would have $\sqrt{\ln 2}$ in the numerator; the radius would differ by a factor $\sqrt{2}$. The raw measured quantity and its threshold should accompany any quoted fitted radius.

A positive intensity cannot distinguish a negative sidelobe amplitude from a positive amplitude of the same magnitude. To observe the sign change shown by the finite-window amplitude curve, interfere the field with a phase reference. Conversely, when comparing with a camera image, square the signed numerical amplitude once and apply the detector pixel average. Comparing a real-amplitude plot directly with a recorded intensity array is an avoidable model mismatch.

Tutorial

How could one distinguish an intrinsic physical sidelobe from ringing produced by the scanner window without a known reference aperture?

▼ **Work through the solution**

Increase the scanned width while keeping Δx fixed by adding samples. Window artefacts should change as the omitted tails shrink. Repeat with a different smooth weighting and report its effect on resolution. An intrinsic feature that survives these tests can then be compared against an independent physical model.

At $W = 3$ mm and $N = 24$, what spatial frequencies can be represented without aliasing, and what does adding zero padding do?

▼ **Work through the solution**

$\Delta x = 0.25$ mm, so the Nyquist limit is 2 cycles/mm. Zero padding supplies more displayed frequency samples of the same discrete transform. It adds no measured aperture detail and cannot resolve aliasing already present in the original samples.

A beam shifted by x_0 within a sufficiently wide window has the same far-field intensity. What information changes?

▼ **Work through the solution**

Substituting $x' = x - x_0$ in the Fourier integral multiplies its amplitude by $e^{-2\pi i f x_0}$. The phase ramp records the displacement; the intensity is invariant. A truncated window may break this invariance by clipping different amounts of field.

When differentiating a measured dipole invents radiated power

Numerical field laboratory

Infer dipole radiation from a sampled displacement and show quantitatively why position noise can dominate a seemingly well-resolved power estimate.

From trajectory to a physically meaningful quantity

An electrically small source carries an oscillating electric dipole moment $p(t) = p_0 \cos \omega t$ in a fixed direction. In nonrelativistic electric dipole radiation, the total instantaneous far-zone power associated with retarded time is

$$P = |\ddot{p}|^2/(6\pi\epsilon_0 c^3).$$

The average over a single sinusoidal cycle is $p_0^2\omega^4/(12\pi\epsilon_0 c^3)$. The fourth power of frequency expresses a physical sensitivity, but a numerical second derivative can introduce a separate sensitivity to high-frequency measurement errors.

A sensor supplies moment or displacement samples at times $t_j = jh$. If the source is a known rigid displacement $x(t)$ of charge q, then $p = qx$. The constant conversion from moment to displacement cancels when comparing a numerical power with a reference power. We therefore scale amplitude by p_0 and time by $1/\omega$. The true acceleration trace is $-\cos t$. Comparing it with the estimated acceleration before squaring reveals signs and phases that a power curve conceals.

Derive the frequency response of the differentiation formula

Use $D_2 p_j = (p_{j-1} - 2p_j + p_{j+1})/h^2$. On a complex exponential $e^{i\nu jh}$ the symbol is $-4\sin^2(\nu h/2)/h^2$. Dividing its magnitude by the exact ν^2 gives a squared sinc factor. At a fixed physical frequency the error is quadratic in h. For a signal oscillating in sign from one sample to the next, $\nu h = \pi$, and the numerical second derivative multiplies its amplitude by $4/h^2$. A numerical derivative smooths the very highest frequency relative to an exact high-frequency derivative yet still strongly amplifies the raw high-frequency component relative to a low-frequency physical acceleration.

Add a deliberately reproducible alternating error $\epsilon(-1)^j$ to the position readings. Its acceleration error is $-4\epsilon(-1)^j/h^2$. Squared and averaged over a complete even grid, the alternating error and the fundamental are orthogonal. Therefore the normalised inferred mean power becomes

$$\frac{\langle P_{\mathrm{est}} \rangle}{\langle P_{\mathrm{true}} \rangle}$$

=\left[\frac{\sin(h/2)}{h/2}\right]^4+2\left(\frac{4\epsilon}{h^2}\right)^2 .

The first term tends to one from below as sampling is refined. At a fixed per-sample alternating displacement amplitude the second grows as h^{-4}. The experiment holds a particular *sampling-related* error fixed; it does not claim that all real noise changes frequency when the sampling rate changes. A fixed-frequency physical disturbance obeys the symbol at its own fixed ν.

A numerical example in which refinement hurts

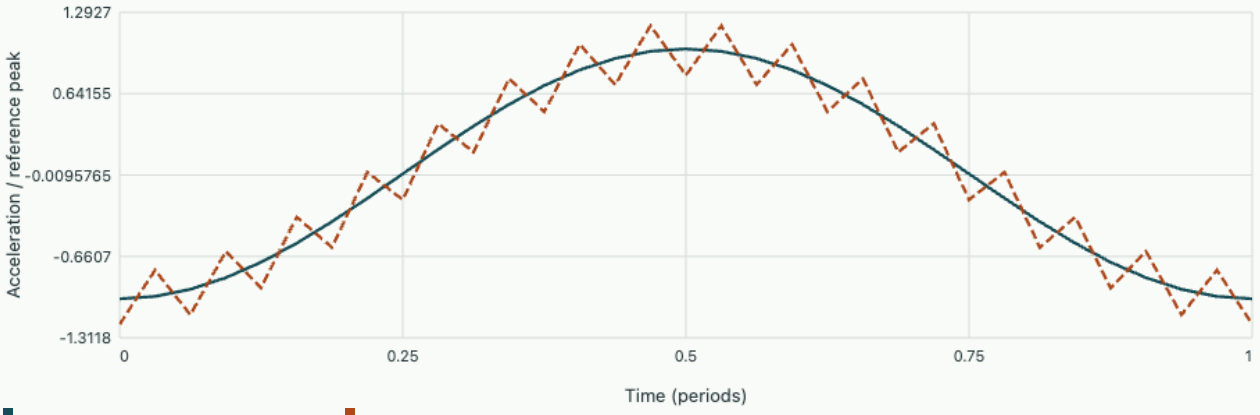
For 32 samples per period, $h = 2\pi/32 \simeq 0.19635$, and a positional disturbance two parts per thousand of full amplitude gives a spurious acceleration peak about 0.2075 of the true peak. Position data plotted at ordinary scale can look nearly perfect. The normalised mean-square acceleration is already about 1.080. At 64 samples per period with the same alternating error the spurious acceleration amplitude grows fourfold, and its contribution to mean power grows sixteenfold.

Open the live experiment

Try the calculation

Samples per physical period **32**

Position disturbance / amplitude **0.002**



Mean squared acceleration / exact value 1.0797. The alternating error contributes peak acceleration 0.20751 in the same units.

Set the perturbation to zero first and verify the approach to one as the physical sinusoid becomes increasingly well resolved. Restore the perturbation and determine which combination of resolution and disturbance first incurs a ten-per-cent error. The error can initially decrease and later increase with sample rate. Choosing the estimate that happens to lie closest to one at a single resolution can hide a cancellation between physical-mode discretisation error and positive noise power.

Recover a physical model without claiming information you did not measure

If an independent measurement shows the source contains only one sinusoidal frequency, fit its sine and cosine coefficients to the positions rather than differentiating every neighbour. On a full uniformly sampled cycle these coefficients are orthogonal to the alternating offset. Multiplying the fitted phasor by $-\omega^2$ gives the model-based acceleration. This can be a much more precise estimate, but it assumes other harmonics are absent or irrelevant. A weak physical second harmonic has an acceleration amplification of

four and power weighting of sixteen relative to equal moment amplitude. Validate the assumed spectrum before suppressing it.

For broad-band signals one may fit a smooth basis with an explicit high-frequency cutoff or use a model of detector bandwidth. Record the chosen cutoff, its physical justification and the uncertainty propagated into power. Repeating the computation with nearby cutoffs is a test of robustness, not a derivation of the correct filter. Where possible, compare the inferred radiated energy with a distant calibrated flux measurement over the same retarded-time interval. These two measurements have different systematic errors.

Tutorial

Independent, zero-mean random sample errors have variance σ_p^2. What is the variance in one centred second derivative and why need adjacent derivative errors not be independent?

▼ **Work through the solution**

The weights are $1, -2, 1$ divided by h^2. Independence of the original samples gives variance $6\sigma_p^2/h^4$. Adjacent derivative estimates share two original samples with nonzero weights, so their covariance is generally nonzero.

A second physical harmonic has amplitude 5% of the fundamental. How much does it contribute to the total time-averaged electric-dipole radiation, assuming orthogonal harmonics over the record?

▼ **Work through the solution**

Its acceleration amplitude relative to the fundamental is

$$0.05 \times 2^2 = 0.2.$$

Its mean power is $0.2^2 = 0.04$ of the fundamental contribution, about 3.85% of their sum. A source spectrum small in displacement need not be equally small in radiated power.

Why must a measured power spectrum specify whether it represents one-sided or two-sided frequencies?

▼ **Work through the solution**

A real signal contains conjugate spectral amplitudes at opposite frequencies. A one-sided positive-frequency spectrum usually combines their powers. Counting both after applying such a convention doubles the inferred variance and radiation power.

Verification, refinement and local error

Numerical partial differential equations

Use a manufactured solution to verify an implementation, estimate observed convergence and identify when adaptivity is needed for the physical quantity of interest.

FROM FIRST PRINCIPLES

Derive the observed order and an extrapolated continuum output

Verification compares the discretized problem against a known solution or controlled refinement. Choose one scalar output Q, the same geometry and the same physical observation point on each mesh. Assume a smooth leading error $Q_h = Q + Ch^p + o(h^p)$ on geometrically similar meshes, with refinement ratio $r > 1$. Subtract successive estimates to remove the unknown continuum output:

$$Q_h - Q_{h/r} \simeq Ch^p(1 - r^{-p}),$$
$$Q_{h/r} - Q_{h/r^2} \simeq Ch^p r^{-p}(1 - r^{-p}).$$

Their ratio approaches r^p, provided the leading coefficient is nonzero. Solve for the order, then use the finest pair to remove the leading error term:

$$p_{\text{obs}} = \frac{\ln\left|(Q_h - Q_{h/r})/(Q_{h/r} - Q_{h/r^2})\right|}{\ln r},$$
$$Q_{\text{ext}} = Q_{h/r^2} + \frac{Q_{h/r^2} - Q_{h/r}}{r^{p_{\text{obs}}} - 1}.$$

The ratio is meaningful only when a consistent leading error dominates. Differential changes of solver tolerance, geometry or timestep across the meshes can contaminate it. A manufactured solution independently tests discretization and boundary terms by inserting a chosen exact field into the differential operator and supplying the resulting source.

Background · Observed convergence order

$$p_{\text{obs}} = \log(e_h/e_{h/2})/\log 2$$

Errors must refer to the same observable and norm. Cancellation, singular boundaries and a reference solution of insufficient accuracy can distort the result.

DERIVATION AND APPLICATIONS

How the measured norm changes an observed convergence rate

A verification study may report very different rates for the same discrete field. Choose a fixed one-dimensional domain of length one, divided into cells of width h. Suppose the solution is smooth away from an interface, where a nearby discontinuity, boundary approximation or reconstruction produces a localized error. Let the error magnitude be Ah^{α} on a fixed number m of cells near the interface, and at most Bh^{p} elsewhere. The count m stays fixed under refinement. For a finite $q \geq 1$ use the physically weighted norm instead of an unweighted vector norm:

$$\|e_h\|_q^q = \sum_i h|e_i|^q \lesssim mA^q h^{q\alpha+1} + B^q h^{qp}.$$

As h decreases the term with the smaller exponent decays more slowly. If the indicated leading magnitudes are actually attained, the corresponding rate of the norm is

$$\|e_h\|_q = O(h^{\min(\alpha+1/q,p)}),$$
$$\|e_h\|_\infty = O(h^{\min(\alpha,p)}).$$

The extra $1/q$ comes from the shrinking physical measure of the affected region. It disappears in a maximum norm. A single signed integral of error can give still another result because opposite local errors cancel:

$$\left|\sum_i he_i\right| \leq \sum_i h|e_i| = \|e_h\|_1.$$

A convergence statement therefore requires the reported norm, physical location, and mesh sequence. If an interface stays at a changing subcell position as the grid is refined, the leading coefficient can vary; observed slopes from just two grids can then be misleading.

WORKED EXAMPLE

Calculate three norms for a displaced step

Let one cell of a piecewise-constant reconstruction differ from the exact field by 1.00, with zero error elsewhere. At $h = 1/64$, the L^1 error is $1/64 = 0.015625$, the L^2 error is $\sqrt{1/64} = 0.125$ and the maximum error is 1.00. Refining to $h = 1/256$ gives 0.003906, 0.0625 and 1.00. The measured rates are one, one half and zero for precisely the same displacement.

SECOND APPLICATION

Compare cancellation with local accuracy

Suppose two adjacent cells at $h = 1/100$ have errors $+0.500$ and -0.500. The signed integral discrepancy is zero; the L^1 error is 0.0100 and the L^2 error is

$$\sqrt{0.005} = 0.0707.$$

Checking only total inventory would certify the output but miss the incorrect distribution. A local measurement or absolute-error norm verifies a different and often more demanding quantity.

Build a test with a known answer

Consider $-\nabla \cdot (D\nabla u) + \kappa u = f$ on the unit square. Select the exact field $u_\star = \sin(\pi x)\sin(\pi y)$ and prescribe its value at the boundary. Now choose $D = 1 + x$, $\kappa = 2$, using coordinates normalised to a metre. Differentiation gives

$$\Delta u_\star = -2\pi^2 u_\star,$$
$$u_{\star,x} = \pi \cos(\pi x)\sin(\pi y).$$

The required source is therefore

$$f = [2(1+x)\pi^2 + 2]u_\star - \pi\cos(\pi x)\sin(\pi y).$$

The derivative of D is essential: inserting only the local value of D times the Laplacian specifies a different differential equation. Manufacturing a source tests whether the code solves the stated model and boundary conditions. Experimental validation asks a separate question about fidelity to physical observations. A polished plot resembling the exact field does not substitute for a numerical error measure with defined sample positions and units.

Read a three-grid study

On successively halved mesh spacings suppose the error in one norm follows $E_h = Ch^p + o(h^p)$. Taking ratios gives

$$p \simeq \log_2(E_h / E_{h/2}).$$

Three grids provide two such estimates, indicating whether a consistent asymptotic trend is present. A single ratio may result from cancellation between terms of different order. State the norm, grid location, geometry representation, time step if applicable, and the residual tolerance. Comparing a cell average on one grid with a nodal peak on another can produce a meaningless rate.

For the variable-diffusion square, maximum nodal errors at 24, 48 and 96 intervals per axis are approximately 1.33345×10^{-3}, 3.33763×10^{-4}, 8.34283×10^{-5}. The observed orders are approximately 1.998 and 2.000. The finest-grid algebraic residual is already far below the field error in its respective normalisation. Tightening the iteration tolerance alone will not remove the leading spatial defect. Raising the sine mode while holding the mesh fixed increases derivatives of the exact solution and tests a different resolution regime. Inspect the grid and mode experiment.

Decide where further resolution matters

A manufactured sine is smooth everywhere. Actual domains may contain a re-entrant corner, a material jump or a narrow source. Derivative bounds used in theoretical convergence estimates can fail locally there. Uniformly halving a two-dimensional mesh roughly quadruples the cell count while spending most new cells away from a localised feature. Local refinement attempts to concentrate new degrees of freedom where they matter to a chosen output. Which cells merit refinement depends on the equation, boundary data and the output being measured.

For a conforming finite-element approximation, one class of error indicators combines an element residual with jumps in normal flux across interior faces. The residual tests whether each element satisfies the differential balance, while the jumps measure disagreements between neighbouring reconstructed fluxes. In a constant-coefficient Poisson example, a piecewise linear solution can have zero elementwise Laplacian even when it bends globally; the slope changes then appear at interelement faces. An indicator using only the interior derivative would miss this information. Mesh-length factors and coefficient weights are necessary to combine the terms in a specific norm.

Finish a refinement study by tracking the physical observable itself. If the objective is integrated heat flow, a changing maximum pointwise gradient near one corner may have little influence on that integral. Conversely, apparently stable domain averages can hide an unresolved peak. Report observed change with resolution and an independent balance or analytic check, specifying exactly which quantity has been verified.

Choose cells to refine from residuals and flux jumps

A mesh-convergence table needs several complete solutions. To spend the next refinement locally, examine the equation that the present field still fails to satisfy. For $-u'' = f$ with zero Dirichlet endpoints, let u_h be its conforming piecewise-linear Galerkin solution. Integrating the error equation by parts on each cell K gives

$$\int_0^1 (u - u_h)' v' \, dx = \sum_K \int_K (f + u_h'') v \, dx - \sum_z J_z v(z),$$

$$J_z = u_h'(z^-) - u_h'(z^+).$$

The sum over z includes interior nodes only. Even when the cell derivative is constant and $u_h'' = 0$ in each open cell, its jumps measure failures of flux continuity. Dropping these terms would miss part of the weak residual. Galerkin orthogonality lets us substitute $v - I_h v$ for v, where $I_h v$ is a local conforming approximation. Local interpolation and trace estimates supply the mesh-length factors in an indicator. One conventional allocation on a uniform mesh is

$$\eta_K^2 = h_K^2 \int_K (f + u_h'')^2 \, dx + \frac{h_K}{2} \sum_{z \in \partial K \text{ interior}} J_z^2.$$

The half shares a node between its two neighbors. An upper bound on the energy error involves the square root of the summed indicators and a mesh-independent constant. Efficiency estimates can also involve unresolved variations in the load. The indicator is a refinement criterion, not a measured cell error.

Spending one refinement near a source

On the two half-intervals let $f = 0$ on the left and $f = 4$ on the right. The central hat function integrates to $1/4$ on the right, so the interior load is 1. Its stiffness is 4. The central temperature is $1/4$. The cell slopes are $1/2, -1/2$, so $J = 1$. The jump contributes $1/4$ to each indicator squared. The source contributes zero on the left and $(1/2)^2(16/2) = 2$ on the right. Thus the right cell accounts for $2.25/2.5 = 90\%$ of the total. A bulk-marking rule targeting at least 75% marks this one cell.

An exact check explains the distinction between indicator and error. Integrating the differential equation gives the exact slope $1/2$ throughout the left cell and $1/2 - 4(x - 1/2)$ on the right. The true squared energy error is zero on the left and $\int_0^{1/2}(1 - 4s)^2 ds = 1/6$ on the right. The indicator assigns some weight to the neighbor because both cells share a flux jump. Its largest contribution nevertheless identifies the curved part of the field.

Check refinement at an identical location

Suppose a displacement is reported at the centre of a square. On an even grid that point may coincide with a node, whereas on an odd grid it can lie at the centre of a cell. Taking the maximum over mesh nodes produces a different sample on each grid. A refinement curve can then combine field discretisation error with the movement of the evaluation point. Choose a fixed physical point and evaluate each solution there with its own consistent interpolation, or compare a well-defined integral over the same physical region.

Similarly, compare gradients on compatible supports. Nodal averages of element stresses can look smoother than the individual element values, yet smoothing can alter extrema and convergence rates. Record the native support and any recovery operator. This makes an apparent improvement reproducible and gives students enough information to explain a discrepancy between two otherwise credible simulation reports.

Separate space and time with a crossed refinement study

Suppose a reported output has asymptotic error $Q(h, k) - Q_\star = Ah^p + Bk^q$ plus smaller terms, with spatial spacing h and time step k. Refining both coordinates in a fixed relation examines only one line across this two-dimensional error surface. Opposite signed contributions can cancel on that line. A genuine spatial order check uses several h at fixed sufficiently small k, and a temporal check uses several k at controlled h. When exact values are available, the individual error signs provide particularly clear evidence of cancellation.

A small crossed design with coarse and fine settings in both directions already offers a structural test. Denote its four outputs by $Q_{cc}, Q_{cf}, Q_{fc}, Q_{ff}$, with the first subscript indicating the spatial mesh. For an additive leading error the spatial change $Q_{cc} - Q_{fc}$ equals $Q_{cf} - Q_{ff}$. Their difference, $Q_{cc} - Q_{cf} - Q_{fc} + Q_{ff}$, measures a mixed contribution or departure from the leading separated model. A small difference supports that model over the tested settings but cannot prove that the assumed orders are correct; additional refinement levels are needed to estimate orders from the data.

Recover a hidden temporal bias from four runs

Take a manufactured output $Q = 10 + 2h^2 - k$. The benchmark value is ten. Choose $h_c = 0.2,\ h_f = 0.1,\ k_c = 0.1,\ k_f = 0.05$. The four outputs are respectively $9.98,\ 10.03,\ 9.92,\ 9.97$. At either time step, halving the spatial mesh reduces the output by 0.06. At either spatial mesh, halving the time step increases the output by 0.05. The mixed difference is zero. Refining both together moves the value by only −0.01, concealing spatial and temporal changes five or six times larger.

If the orders are independently known as $p = 2,\ q = 1$, the fine-grid spatial bias is

$$(Q_{cf} - Q_{ff})/(2^2 - 1) = 0.02.$$

The fine-time bias is

$$(Q_{fc} - Q_{ff})/(2^1 - 1) = -0.05.$$

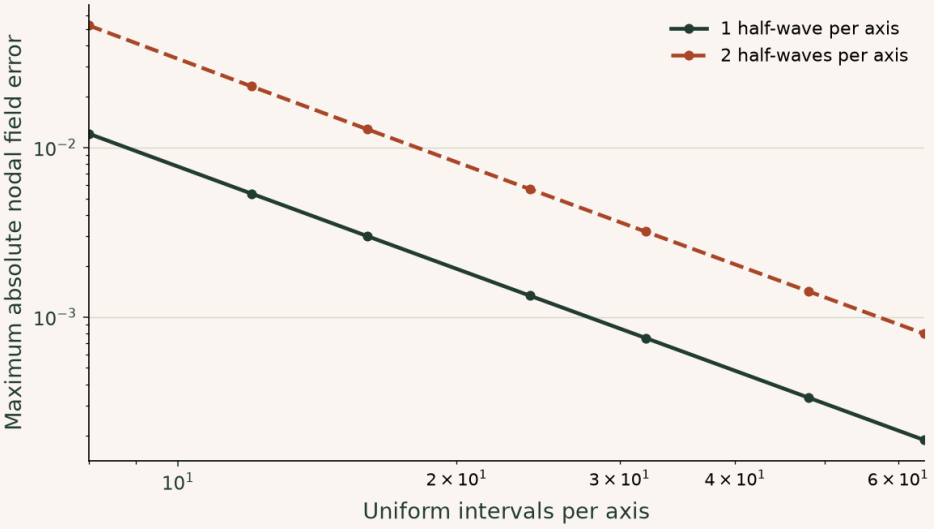
Subtracting both from the fine result gives $9.97 - 0.02 + 0.05 = 10$, exactly recovering the limit in this idealised expansion. Choosing instead $k = 2h^2$ on every paired run would return exactly ten for every mesh, including a very coarse one. A convergence plot along that line would demonstrate cancellation rather than independent verification.

Explicit stability restrictions can make unrestricted crossing impractical. If a coarse time step is unstable on the finest spatial grid, choose a stable common range or use an implicit reference to isolate spatial error. Keep output locations, boundary representation and integration tolerances consistent across the crossed study, since changes in those other approximations can also enter the mixed difference.

A fifth run can interrogate the assumption that the fine-grid spatial change is asymptotic. At $h = 0.05,\ k = 0.05$ the manufactured output is 9.955. The change from 9.97 is 0.015, one quarter of the earlier spatial change 0.06 at the same time step. This independent ratio supports $p = 2$ for the tested family; it does not distinguish a true zero cross term from a smaller mixed term below the numerical resolution.

Resolve a second spatial frequency on the same grid

Set the exact field on the square to $\sin(m\pi x)\sin(m\pi y)$ and manufacture the corresponding source for diffusion coefficient $1 + x$ and reaction coefficient 2. Run the downloadable OneSimulation model with m = 1 and 2 on each grid. The boundary data agree, while the source and the exact interior curvatures change.



Both series approach a slope of minus two on the log–log axes. At fixed grid count the higher-frequency solution is less accurate. Doubling its grid count restores the points per wavelength; compare those shifted entries to see how closely that simple resolution argument predicts the actual errors.

▶ **Read values from the plot**

Download the OneSimulation model (Python)

Change parameters · Resolve a second spatial frequency on the same grid

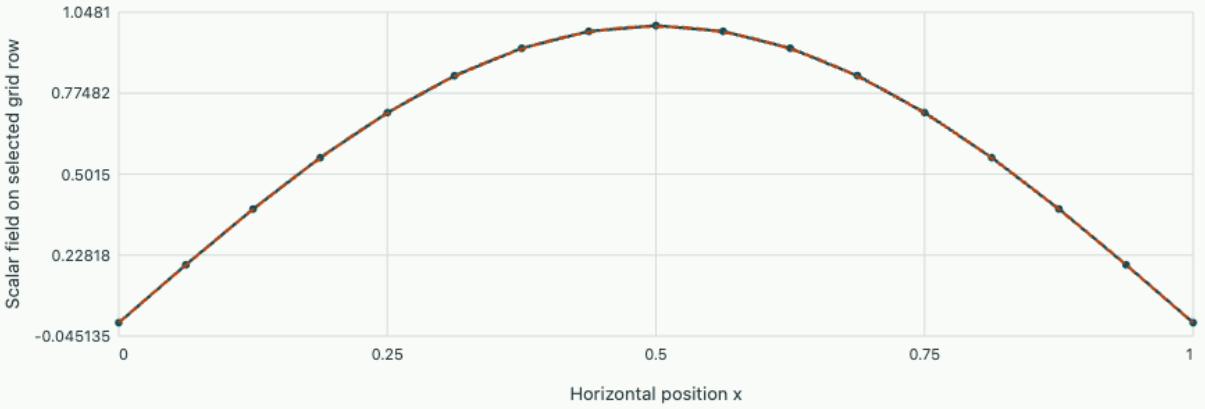
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Mesh intervals per axis **16**

Sine half-waves per direction **1**



Variable-coefficient five-point solve Manufactured continuum field

For −∇·[(1+x)∇u] + 2u = f, maximum nodal error 0.0030008, relative algebraic residual 4.2371e-11. Slice y = 0.5; 17 Jacobi-preconditioned CG iterations. Refinement measures discretisation after the algebraic solve has converged.

What a periodic grid does to a decaying Fourier mode

Numerical partial differential equations

A periodic diffusion solver predicts the shape of a sampled wave while its amplitude decays too slowly. Fourier analysis separates spatial differentiation error, an unresolved physical mode and errors caused by the time integrator.

Diagonalise the differential operator on the periodic interval

Use a dimensionless coordinate $0 \le x < 2\pi$ and solve $u_t = \kappa u_{xx}$ with periodic values and periodic derivatives. The complex exponentials e^{imx}, with integer m, satisfy both conditions and are eigenfunctions of the second derivative. Put $u(x,t) = A_m(t)e^{imx}$ and cancel the common exponential. The scalar amplitude equation is

$$A_m' = -\kappa m^2 A_m.$$

Thus every cosine or sine mode decays by $\exp(-\kappa m^2 t)$, with no phase shift, while the m=0 mean is conserved.

On a uniform grid $x_j = 2\pi j/N$, the samples of m and $m + \ell N$ agree for every integer ℓ. A discrete Fourier transform returns coefficients in one representative band. For the real cosine test with even N we can fold a positive physical mode into $k = \min(r, N-r)$, $r = m \bmod N$. A Fourier spectral Laplacian multiplies that coefficient by $-k^2$. For a resolved mode it differentiates the trigonometric interpolant exactly, up to arithmetic. For unresolved data it differentiates the aliased representative; no operation using only those samples can identify the original m.

Apply the centred difference stencil to a single wave

The standard second difference is $(u_{j+1} - 2u_j + u_{j-1})/h^2$, $h = 2\pi/N$. Put the grid wave $u_j = e^{ikjh}$ into the stencil. Factor out the common phase and simplify with $e^{i\xi} + e^{-i\xi} = 2\cos\xi$. This gives the symbol and its small-h expansion:

$$\frac{e^{ikh} - 2 + e^{-ikh}}{h^2} = -\frac{4}{h^2}\sin^2(kh/2) = -k^2 + \frac{k^4 h^2}{12} + O(k^6 h^4).$$

The effective squared wavenumber is smaller than k^2 for every nonconstant mode in the resolved band, so this spatial approximation decays too slowly if the time equation is integrated exactly. The error is modest when $kh \ll 1$, but at the Nyquist mode $kh = \pi$ the ratio is

$$4/\pi^2 = 0.4053.$$

A stable time integrator cannot repair this spatial attenuation error without additional modeling. The experiment exponentiates both discretizations' symbols exactly, isolating space and sampling here.

Measure separate errors with the same sampled initial data

Choose $\kappa = 0.5,\ m = 3,\ N = 16$. The true squared wavenumber is 9 and the difference stencil gives 8.006. Starting at unit amplitude, at $t = 0.20$ exact and spectral amplitudes are 0.40657; finite differences give 0.44906. Doubling N to 32 changes the difference result to 0.41716. The absolute amplitude error falls from 0.04249 to 0.01059, a ratio of 4.01 consistent with the second-order expansion. This is a spatial-convergence study because time evolution introduces no truncation error.

Now take $m = 12,\ N = 16$. Every sampled cosine equals the low mode $k = 4$ at $t = 0$. At $t = 0.02$, the continuum amplitude for the true physical mode is $e^{-1.44} = 0.237$, whereas the spectral grid predicts

$$e^{-0.16} = 0.852.$$

A spectral computation can therefore agree exactly with its trigonometric interpolant and disagree strongly with the evolution of the intended signal. Sampling on a finer grid or obtaining prior bandwidth information resolves that ambiguity. The mean mode is particularly revealing when a nonconstant physical cosine aliases to a constant.

Start at mode 3 and successively double N from 8 to 32. Compute two amplitude-error ratios at the same time. Then use mode 12 on 16 points to observe an alias, and increase N until the Fourier grid represents the intended mode.

Open this experiment online

Try the calculation

Periodic grid points	16
Physical cosine mode	3
Diffusivity	0.5



The grid represents mode 3 as 3. At t = 0.20, amplitudes are 0.40657 for the physical mode, 0.40657 for the Fourier grid and 0.44906 for the difference stencil.

Tutorial · resolve a damping rate to a target tolerance

For a mode well inside the grid bandwidth, estimate the number of samples per wavelength needed for a relative error below 1% in the decay rate with centred differences. Test the estimate for $m = 2$. How could the time integrator contaminate a measurement of this spatial error?

▼ **Show the resolution estimate**

Writing p for samples per wavelength gives

$$kh = 2\pi/p.$$

The relative rate error is approximately $(kh)^2/12$. Requiring at most 0.01 gives $p \gtrsim 18.14$; try 19 samples per wavelength, or $N = 38$ for $m = 2$. The exact ratio $[\sin(\pi/19)/(\pi/19)]^2$ is about 0.991. A forward-Euler step would replace exponential propagation by $1 - \kappa k_{\mathrm{eff}}^2 \Delta t$. Its repeated-step error could partly mask the underdamping of the spatial stencil. Compare with the exponential of the discrete symbol before attributing a measured rate to space alone.

Further reading · Lloyd N. Trefethen, Spectral Methods in MATLAB · periodic grids, Fourier differentiation and downloadable programs

Check the discrete eigenvalue with OneSimulation

A spatial sine also diagonalises the five-point Dirichlet Laplacian on a square. Use the same modified wave-number argument in both coordinate directions. Run `python poisson-mode.py --mode 2 --cells 8 16 32` and compare the observed nodal error with the discrete eigenvalue prediction in the CSV. This is a steady forced problem whose spatial operator can be checked independently of any time integrator. Download the OneSimulation verification model.

Diffraction as a Fourier transform

Waves and optics

Derive the far-field pattern from a finite aperture and identify which features control its scale and its sidelobes.

FROM FIRST PRINCIPLES

Sum the field across a uniformly illuminated slit

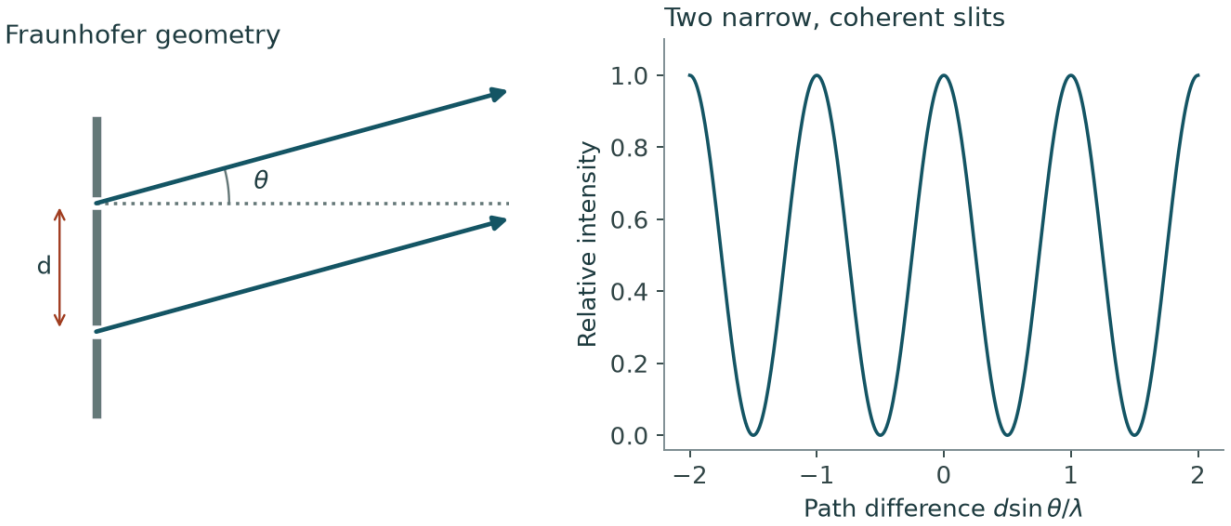
A distant screen receives contributions from across the whole aperture. Consider monochromatic scalar illumination normal to a long slit of width a. In the far field, paths from two points separated by distance dx differ by approximately $dx \sin\theta$ in direction θ. Distance-dependent amplitudes are nearly constant across the slit, leaving a linear phase variation. Put $k = 2\pi/\lambda$ and choose the origin at the slit centre.

$$U(\theta) = C \int_{-a/2}^{a/2} e^{-ikx\sin\theta}\, dx.$$

Integrating the exponentials from the two edges and normalising to forward amplitude yields a sinc function. The central limit is one by continuity even though the displayed ratio takes the form zero divided by zero there.

$$\frac{U(\theta)}{U(0)} = \frac{\sin z}{z},$$
$$z = \frac{\pi a \sin\theta}{\lambda}.$$

Intensity is the squared magnitude of the coherent field sum. The first noncentral zero therefore occurs at $a|\sin\theta| = \lambda$. Widening the slit narrows its central angular peak. For a general transmission profile the same weighted integral is its Fourier transform sampled at transverse wavenumber $k\sin\theta$.



Read fringe order from the geometry

Illuminate two narrow slits normally so they emit equal, in-phase amplitudes with the same polarisation. Projecting their separation onto the far-field observation direction gives a path difference d sin θ. A path difference of half a wavelength adds the second amplitude with the opposite sign. A finite slit width introduces a separate envelope over these fringes.

Explore the related parameter study beside the computed figure.

DERIVATION AND APPLICATIONS

Locate missing interference orders in a double slit

A pair of finite slits can erase an interference maximum even when the detector is perfectly aligned. Consider two identical uniformly illuminated slits of width a, centre separation $d > a$, and long dimension perpendicular to the observation plane. Assume coherent monochromatic light of wavelength λ, scalar diffraction and far-field detection. At observation angle θ define transverse spatial frequency $q = 2\pi \sin\theta/\lambda$. Each point at coordinate x contributes phase e^{-iqx}. A centred slit therefore contributes

$$U_1(q) = C \int_{-a/2}^{a/2} e^{-iqx}\, dx = Ca\, \frac{\sin(qa/2)}{qa/2}.$$

Translating a slit multiplies its transform by a phase factor. The two translated copies have centres at $\pm d/2$. Add their fields, square the magnitude, and divide by the central intensity. The aperture width and slit separation now enter different factors.

$$\frac{I(\theta)}{I(0)} = \left[\frac{\sin(qa/2)}{qa/2}\right]^2 \cos^2(qd/2).$$

The centres of the ideal two-point interference orders satisfy

$$d \sin\theta = m\lambda.$$

Evaluate the single-slit envelope at those positions. The order disappears exactly when its location coincides with a noncentral zero of the envelope.

$$a \sin\theta = p\lambda,$$

$$m = p\frac{d}{a},$$

$$p = \pm 1, \pm 2, \ldots.$$

If d/a is a rational number, only those choices of p producing integer m erase interference orders. Finite width can also shift the actual maxima of the product slightly away from the nominal interference centres. The coincidence of a zero with an order centre remains exact in this ideal model.

WORKED EXAMPLE

Explain an absent green fringe

Change the inputs and investigate this derivation.

Two slits 50.0 µm wide have centre spacing 200 µm. At $\lambda = 500$ nm the fourth order lies at $\sin\theta = 4\lambda/d = 0.0100$. This is the first zero of the single-slit amplitude. A lens of focal length 200 mm would place that missing fringe about 2.00 mm from the centre. The first order samples an envelope factor $[\sin(\pi/4)/(\pi/4)]^2 = 0.811$.

SECOND APPLICATION

Reconstruct slit dimensions from a pattern

Change the inputs and investigate this derivation.

A different apparatus uses 600 nm light and a 300 mm focal length. Near the axis, the measured nominal interference spacing is 0.900 mm, giving $d = f\lambda/\Delta x = 200$ µm. The first envelope zero lies 2.25 mm from the centre, giving $a = f\lambda/x_1 = 80.0$ µm. Their ratio is 2.50. The first envelope zero lies halfway between two nominal orders, while the second erases order five at 4.50 mm.

Where a geometrical shadow breaks down

Illuminate a thin screen at normal incidence with a monochromatic plane wave of vacuum wavelength λ in air. Light transmitted through different parts of an opening reaches a distant observation point along slightly different paths. Their phases matter even where geometrical rays would draw a hard shadow. Use scalar diffraction, suited to features much larger than the wavelength and modest angles. Represent the complex field just beyond the screen by an aperture function $A(x, y)$, incorporating amplitude and phase.

For a far-away observation direction $\mathbf{n}$, the propagation distance from aperture point $\mathbf{r}'$ is $r - \mathbf{n}\cdot\mathbf{r}'$ to first order. Replacing the inverse distance in the wave amplitude with $1/r$ while retaining this linear phase yields

$$U(\theta_x, \theta_y) \propto \iint A(x, y) e^{-ik(x\sin\theta_x + y\sin\theta_y)}\, dx\, dy.$$

The angular field samples a spatial Fourier transform. The neglected quadratic path term produces a phase of order $\pi a^2/(\lambda r)$ for aperture half-width a. The free-space far-field approximation requires this quantity to be small. A lens supplies a quadratic phase that creates the Fourier pattern in its back focal plane even at a practical laboratory distance.

Add the contributions of a uniform slit

For a very tall slit of width w, study the horizontal cross-section. Uniform amplitude across the width gives an integral of the exponential between $-w/2$ and $w/2$. The result is $w \sin u/u$ with $u = \pi w \sin\theta/\lambda$. Taking squared magnitude and normalizing at the origin yields

$$I(\theta)/I(0) = \left(\frac{\sin u}{u}\right)^2.$$

The first zeros satisfy

$$\sin\theta = \pm\lambda/w.$$

A broader opening makes the central spot narrower because a smaller angular change suffices to accumulate a full cycle across it. The oscillating tails arise from the abrupt edges. A gradual transmission taper reduces those sidelobes at the cost of a wider central maximum. Peak brightness additionally depends on the transmitted amplitude and collection geometry.

Measure width in a focal plane

Illuminate a 120 µm slit with 600 nm light and use a focal-length 200 mm lens. In the paraxial focal plane $X \simeq f \sin\theta$. The first minima occur at $X = \pm 1.00$ mm, giving a central width of 2.00 mm. At half that distance $u = \pi/2$, so the intensity is $4/\pi^2 = 0.405$ of its central value. These two measurements separately test angular scale and profile shape.

If the slit is moved sideways, the Fourier shift theorem supplies only a phase factor in the far field, and the intensity pattern remains centred in this ideal arrangement. If it is illuminated at a small angle instead, the incident field carries a linear phase across the slit and the whole pattern moves. This distinction becomes useful when diagnosing whether a displaced focal-plane spot comes from a decentered opening or a tilted incoming beam.

A narrow receiving slit or detector pixel also integrates across part of the pattern. If it straddles a predicted zero, it collects light from either side and reports a positive count even for perfect diffraction. Reducing detector width tests this averaging interpretation.

Observe an aperture before its far field forms

A finite opening can produce bright and dark features even on its axis as the detector moves away. Consider a uniformly illuminated circular aperture of radius a in an opaque thin screen. Use a

monochromatic scalar plane wave and a paraxial on-axis detector at distance z large compared with a, yet not necessarily large compared with a^2/λ. At aperture radius r the extra path is $\sqrt{z^2+r^2}-z \simeq r^2/(2z)$. The Fraunhofer approximation drops this quadratic phase; Fresnel propagation keeps it.

Relative to the unobstructed advancing plane wave, the on-axis field is proportional to $\int_0^a (2\pi r/i\lambda z) e^{i\pi r^2/(\lambda z)}\,dr$. The factor $r\,dr$ makes the substitution $u = \pi r^2/(\lambda z)$ immediate. Using the free propagation normalization gives

$$\frac{U(0,z)}{U_{\text{free}}(0,z)} = 1 - e^{i\pi N_F},$$

$$\frac{I(0,z)}{I_{\text{free}}} = 4\sin^2\left(\frac{\pi N_F}{2}\right),$$

$$N_F = \frac{a^2}{\lambda z}.$$

Successive annuli adding half a wavelength of path alternately reinforce and cancel on axis. An even integer Fresnel number gives perfect cancellation in the ideal model. An odd integer makes the on-axis intensity four times the unobstructed value. There is no excess total power: interference redistributes light in the transverse plane. The result concerns local on-axis intensity at a specified distance rather than integrated transmission.

Predict successive axial extrema

Choose $a = 0.500$ mm and $\lambda = 500$ nm. Then

$$a^2/\lambda = 0.500\,\text{m}.$$

A detector at $z = 0.250$ m sees $N_F = 2.00$, predicting an axial null. Moving it to $z = 0.500$ m gives $N_F = 1.00$ and a fourfold maximum. The aperture-to-distance ratio is at most 0.002 in these two measurements, supporting the small-angle path expansion even though the far-field criterion is not satisfied.

At $z = 5.00$ m the Fresnel number is 0.100 and the exact Fresnel expression gives

$$I/I_{\text{free}} = 0.0979.$$

Its far-field expansion is $(\pi N_F)^2 = 0.0987$, now within about one percent. This limit recovers the inverse-square on-axis intensity at fixed opening and incident irradiance. A real photodiode averages several transverse points, which can fill an axial null, and a broadband source averages slightly different Fresnel numbers. Repeating the axial scan with a smaller active area or a narrower spectral band tests which averaging effect dominates the observed minimum.

A complementary screen can help verify that the observed oscillations arise from aperture interference. Within the same scalar boundary approximation, replacing the circular opening by an opaque disk makes the two complex transmitted fields add to the unobstructed field. On axis the disk's normalized field is therefore $e^{i\pi N_F}$ in this model, with unit intensity and a distance-dependent phase. A bright point behind

an opaque disk may seem surprising from geometric rays, but neighbouring angles still redistribute the blocked power. Finite coherence, source size and a detector averaging over the narrow central feature can strongly reduce its measured contrast.

From Fresnel zones to a diffractive lens

The axial null from a clear circular opening results from destructive addition of neighbouring half-period zones. A zone plate reverses that outcome by deliberately blocking every other annulus. At desired focal distance f, a ray from the edge of zone number m has path excess $m\lambda/2$ relative to the centre. Solving $\sqrt{f^2 + r_m^2} - f = m\lambda/2$ gives

$$r_m^2 = m\lambda f + \frac{m^2\lambda^2}{4} \simeq m\lambda f.$$

The approximation requires the outer zone radius to remain small compared with focal distance. Consecutive radii therefore grow approximately as the square root of zone number, and outermost zones are narrowest. Leaving odd zones transparent and even zones opaque removes contributions with the opposite phase at the design focus. Power is also distributed into other diffraction orders and an undiffracted background, so this simple plate does not concentrate all incident light into one focus.

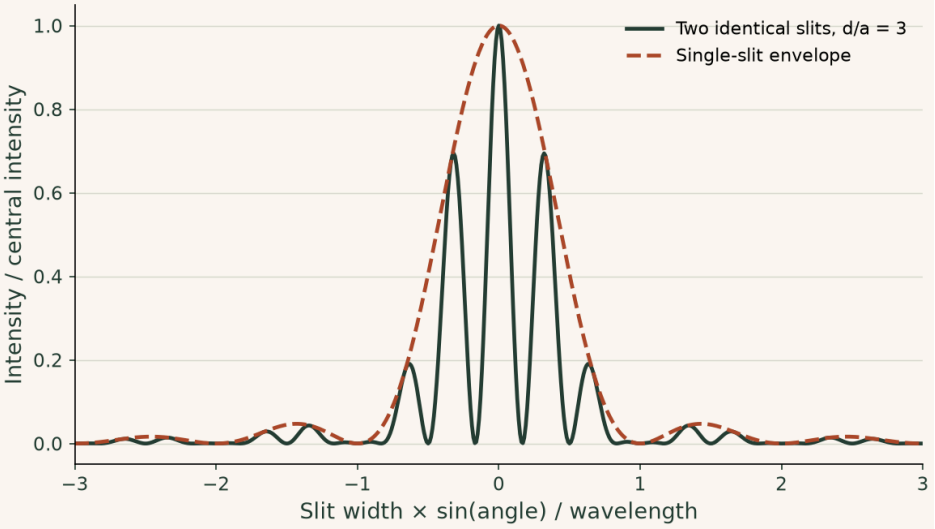
Set $\lambda = 500$ nm and design for $f = 0.200$ m with twenty zones. The radius of the first zone is $\sqrt{\lambda f} = 0.316$ mm, and the twentieth radius is 1.414 mm. The width of that outer zone is $\sqrt{\lambda f}(\sqrt{20} - \sqrt{19}) = 35.8$ µm, setting a fabrication scale substantially finer than the overall aperture. If the plate is used at 600 nm without changing its radii, its first-order focal length becomes

$$r_1^2/\lambda = 0.167\,\mathrm{m}.$$

This shortening with increasing wavelength follows from the required relative phase and is an observable distinction from simply changing the clear aperture of a refracting lens. Scanning on-axis intensity at two colours tests both the zone layout and its predicted chromatic shift.

Which interference orders disappear under diffraction?

Two uniformly illuminated slits have width a and centre spacing $3a$. The Fraunhofer field is the product of the single-aperture transform and a two-source phase sum. Normalise intensity at zero angle and use $u = a\sin\theta/\lambda$.



The nominal third interference maximum occurs at $u = 1$ and is extinguished by the first zero of the slit envelope. A missing order diagnoses the width-to-spacing ratio; it need not indicate missing or blocked illumination.

Try it · Double slit

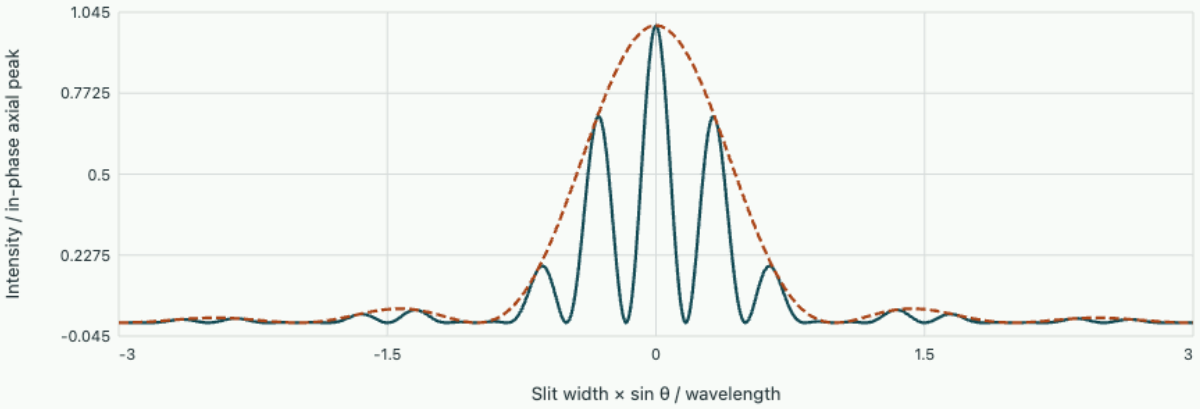
Change the inputs and compare the recalculated curves with the preceding derivation and baseline figure. The axes give the units; the numerical readout states the model assumptions and any limiting cases. Reset the inputs to return to the reference setting.

Open this live calculation

Try the calculation

Slit spacing / slit width **3**

Relative incident phase (degrees) **0**



Two finite slits Single-slit envelope

For zero phase and integer d/a, the corresponding interference order is removed by the first envelope zero. Fraunhofer scalar model with equal slit illumination.

▶ **Read values from the plot**

How a grating separates wavelengths

Waves and optics

Derive the angular dispersion and resolving power of a finite grating and diagnose an order overlap.

FROM FIRST PRINCIPLES

Derive a grating peak from equal phase steps

Take N equally spaced, equally illuminated grooves of pitch d under normal monochromatic illumination. Observe in the far field at angle θ to the grating normal. Neighbouring positions produce a path difference $d\sin\theta$. Ignoring the common single-groove envelope for the moment, each groove contributes an equal field with phase increment δ. Their sum is

$$\delta = \frac{2\pi d}{\lambda}\sin\theta,$$

$$U = U_1 \sum_{j=0}^{N-1} e^{ij\delta} = U_1 e^{i(N-1)\delta/2}\,\frac{\sin(N\delta/2)}{\sin(\delta/2)}.$$

The fields line up at $\delta = 2\pi m$, giving $d\sin\theta = m\lambda$. The nearest zero requires an additional phase step of $2\pi/N$. Convert that change to a small angle and compare with the angular separation of two neighbouring wavelengths at fixed order.

$$\Delta\theta_{\text{zero}} \simeq \frac{\lambda}{Nd\cos\theta},$$

$$\Delta\theta_{\lambda} \simeq \frac{|m|\Delta\lambda}{d\cos\theta}.$$

Equating the separations gives the ideal Rayleigh resolving power $\lambda/\Delta\lambda = |m|N$ for nonzero order. It is the finite illuminated groove count that fixes the intrinsic width; optical aberrations, slit width and pixel sampling enter a measured line profile separately.

DERIVATION AND APPLICATIONS

Convert an entrance-slit width into spectral broadening

An ideal grating can distinguish wavelengths separated by about $\lambda/(mN)$ in order m with N illuminated grooves. A finite entrance slit adds another width because light from different points on the slit arrives at different incidence angles. Consider a slit of full width w in the focal plane of a collimator of focal length f_c. Use small angular deviations around incidence α and diffraction angle β, both defined in the convention $m\lambda = d(\sin\alpha + \sin\beta)$ for groove spacing d. Differentiating at fixed wavelength gives

$$0 = d(\cos\alpha\, d\alpha + \cos\beta\, d\beta),$$

$$|d\beta| = \frac{\cos\alpha}{\cos\beta}|d\alpha|.$$

The angular interval incident on the grating is w/f_c. A camera lens of focal length f_{cam} converts the outgoing angular spread into a full image width W at its focal plane. Measure local displacements perpendicular to the central diffracted ray.

$$W = f_{\mathrm{cam}}\frac{\cos\alpha}{\cos\beta}\frac{w}{f_c},$$

$$\left|\frac{dx}{d\lambda}\right| = \frac{f_{\mathrm{cam}}m}{d\cos\beta}.$$

The second relation follows by varying wavelength at fixed incidence. Dividing the slit image width by this local dispersion gives the wavelength interval corresponding to the full geometric slit image.

$$\Delta\lambda_{\mathrm{slit}} = \frac{d\cos\alpha}{m}\frac{w}{f_c}.$$

The camera focal length cancels because it magnifies both the slit image and the spectrum equally. Actual recorded lines are convolutions of the projected slit, diffraction response and detector response. Their distinct width definitions must be retained when fitting line profiles.

WORKED EXAMPLE

Identify the dominant width in a green spectrometer

Change the inputs and investigate this derivation.

For $d = 1.00$ µm, $\alpha = 0$, $m = 1$, $f_c = 100$ mm and $w = 50.0$ µm, the geometric width is 0.500 nm. At 500 nm, $N = 10{,}000$ illuminated grooves give a diffraction resolution scale of 0.0500 nm. The slit width corresponds to ten such intervals. Doubling the camera focal length would double the image width in pixels while leaving the geometric 0.500 nm wavelength interval unchanged.

SECOND APPLICATION

Size a slit for an angled input

Change the inputs and investigate this derivation.

A separate instrument uses $d = 2.00$ µm, $\alpha = 60.0^\circ$, $m = 2$ and $f_c = 200$ mm. To obtain a full geometric spectral width of 0.100 nm, invert the result: $w = mf_c\Delta\lambda/(d\cos\alpha) = 40.0$ µm. At $\lambda = 600$ nm the grating equation gives $\sin\beta = 0.600 - 0.866 = -0.266$, a physically accessible direction. This angle check matters when changing high orders or incidence settings.

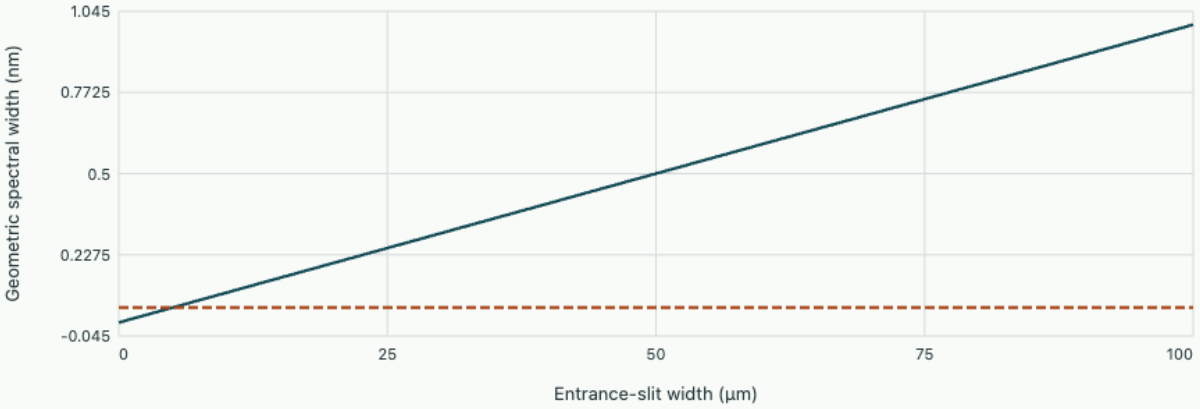
Try it · Slit resolution

Vary the physical parameters in the derived relation. Use the units and assumptions in the readout to connect the curves to the two worked examples. A parameter study can test the forward prediction; the same relation can then be inverted to infer a parameter from a measurement.

Open this live calculation

Try the calculation

Collimator focal length (mm)	100
Incidence angle (degrees)	0
Diffraction order	1



1 µm grating spacing Diffraction scale: 500 nm, 10000 grooves

Geometric width d cosα w/(m fc); diffraction gives λ/(mN). Reducing the slit width eventually reaches the diffraction floor rather than an unlimited resolving power.

Interference across equally spaced openings

Illuminate N identical narrow slits separated by distance d with normally incident monochromatic light. In the far field every slit produces the same single-slit envelope. At observation angle θ adjacent contributions differ in optical path by $d\sin\theta$, so their phases differ by $\phi = 2\pi d\sin\theta/\lambda$. The array field equals the single-slit field multiplied by a sum of N unit phasors. Taking its geometric sum and squared magnitude gives the array intensity factor

$$G(\phi) = \left[\frac{\sin(N\phi/2)}{\sin(\phi/2)}\right]^2.$$

At $\phi = 2\pi m$ the expression has a removable limit N^2. Those directions satisfy the grating equation $d\sin\theta_m = m\lambda$. Each peak collects contributions from all illuminated slits in phase. Some orders can

nevertheless be faint or missing if the single-slit envelope is small or zero there. Transmission efficiency into a selected order thus depends on the groove or opening profile in addition to the grating period.

Dispersion and finite width

At fixed order different wavelengths peak at different angles. Differentiate the grating equation to obtain

$$d\theta_m/d\lambda = m/(d\cos\theta_m).$$

A large angular dispersion aids spatial separation at the detector. To assess resolution, also compute the width of each peak. The adjacent zeros of the numerator lie one phase increment $2\pi/N$ either side of the maximum. Converting the increment back to angle gives a one-sided width $\lambda/(Nd\cos\theta_m)$.

Choose the conventional criterion that one wavelength's peak falls on the neighbour's first zero. Equate the angular separation due to $\Delta\lambda$ to this one-sided width and obtain $\lambda/\Delta\lambda = |m|N$. Broad illumination of a grating therefore improves the ideal resolving power even with unchanged dispersion. The derivation assumes uniform coherent illumination, collimation, adequate detector sampling, and an entrance slit narrow enough to preserve this diffraction-limited width.

A small spectrometer

A 600-lines/mm grating has $d = 1.667$ µm. A collimated 10.0 mm beam illuminates 6000 lines. At $\lambda = 600$ nm first order occurs at $\arcsin(0.360) = 21.1^\circ$. The ideal smallest separation is

$$600/6000 = 0.100\,\mathrm{nm}.$$

The angular dispersion is 6.43×10^{-4} rad/nm. A camera of focal length 200 mm centred on that order maps a 0.100 nm change to about 12.9 µm near its axis. Detector pixels that are large compared with this separation reduce the measurable resolution.

At that same angle, 300 nm light in second order has the same value of $m\lambda$. Increasing the beam width narrows both peaks without separating their centres. An order-selecting filter can restrict the allowed band before measurement. A second grating or prism dispersing perpendicular to the first can instead separate overlapping orders into different traces. This illustrates why calibration with only one visible line cannot certify the identity of every line in a wide-band spectrum.

Test the finite-array argument with the two-slit acoustic field. For two equally excited slits the array factor has its second positive-angle zero at $d\sin\theta = 3\lambda/2$. With $d/\lambda = 3$, predict 30.0° and compare with the sampled 30.50° at receiver radius 0.90 m. At $d/\lambda = 2$ predict 48.59°. Compare the measured angle at both receiver radii to see how the finite observation distance affects the far-field limit.

Map overlapping orders in a high-order spectrometer

Using high diffraction order increases the ideal resolving power for a fixed number of illuminated grooves. It also reduces how much spectrum belongs unambiguously to a single order. Consider a reflection grating

in a geometry where incidence and diffraction are on the same side of the normal in the usual grating-angle convention. The path difference between neighbouring grooves is $d(\sin\alpha+\sin\beta)$, so peaks obey $m\lambda = d(\sin\alpha+\sin\beta)$. In a Littrow setting the selected diffracted beam retraces the incident direction and $\alpha=\beta=\theta$.

At fixed angle two wavelengths overlap exactly if $m\lambda_m = (m+1)\lambda_{m+1}$. Solving for their separation gives $\lambda_m - \lambda_{m+1} = \lambda_m/(m+1)$, approximately λ_m/m at high order. This is the scale of a free spectral range in wavelength. Combined with resolving power mN for N uniformly illuminated grooves, the number of independently resolved spectral elements per order is approximately N. Higher order makes each element narrower while proportionately reducing the unambiguous range at a fixed spatial coordinate.

Real echelle gratings use inclined groove facets to redirect much of the reflected power into the chosen high orders. The groove spacing sets peak positions and order overlap; the groove profile sets an efficiency envelope. A second dispersive element ordinarily separates the superposed orders along another detector coordinate. Its angular dispersion can be lower than the primary grating's because it only needs to separate different orders, not resolve the closest spectral lines within one.

Design the spectral window at one detector position

Take groove spacing $d = 31.6$ µm, order $m = 100$ and central wavelength 550 nm. The Littrow relation requires $\sin\theta = m\lambda/(2d) = 55.0/63.2 = 0.870$, hence $\theta = 60.5^\circ$. Illuminating 1000 grooves gives ideal wavelength resolution $550/(100\times1000) = 0.00550$ nm. The wavelength in order 101 landing at the same position is

$$550(100/101) = 544.55\,\text{nm}.$$

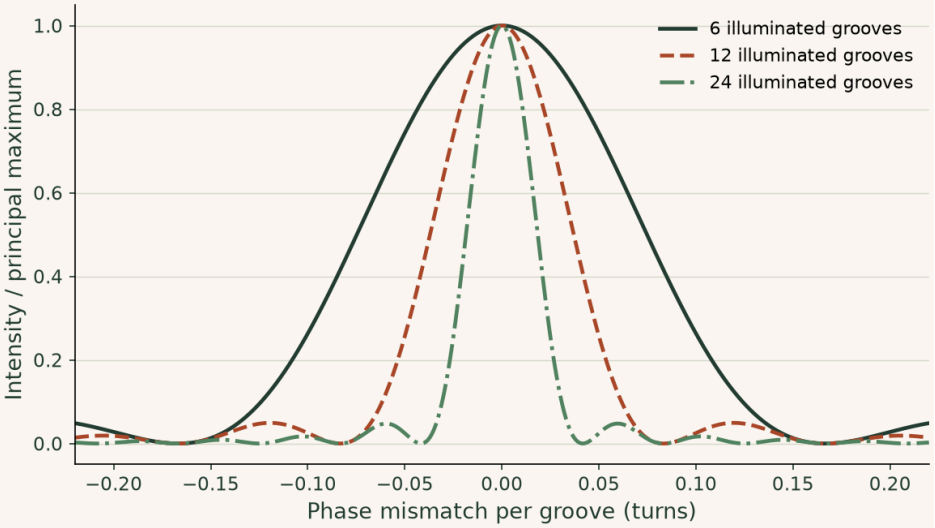
Thus the separation between these coincident-order wavelengths is 5.45 nm, about 990 resolution intervals.

A narrow-band filter passing 548 to 552 nm would reject that particular neighbouring order at the central position while retaining many resolvable lines near 550 nm. An instrument meant to record hundreds of nanometres at once needs cross-dispersion or several separate exposures. As incidence approaches grazing, simple angular-dispersion estimates can become very large, but finite beam footprint, usable groove count, diffraction efficiency and detector size limit the physical gain. Accounting for order separation, illuminated footprint and sampling is necessary before describing two nearby spectral features as resolvable in the laboratory.

A rough detector-format estimate follows directly from this example. If each independent spectral resolution interval requires at least two samples, one full free spectral range near the chosen order needs about two thousand useful pixels along the high-dispersion direction. That count ignores additional margins, optical distortion and the need to sample the entrance-slit profile adequately. The perpendicular axis must also have room for the traces of neighbouring orders and for a background estimate between them. Resolution, spectral coverage and detector format therefore impose independent constraints on what a single exposure can realistically measure.

Separate groove count from grating period

Keep the grating period and incident wavelength fixed. Adding N equal amplitudes as a geometric series gives a central peak whose first zeros occur at phase mismatch $\pm 1/N$ in turns. Each plotted peak is normalised to its own height.



Doubling illuminated groove count halves the first-zero width without moving the principal order. Normalising hides the N^2 increase in peak intensity for fixed incident field per groove; spectral width and absolute throughput answer different questions.

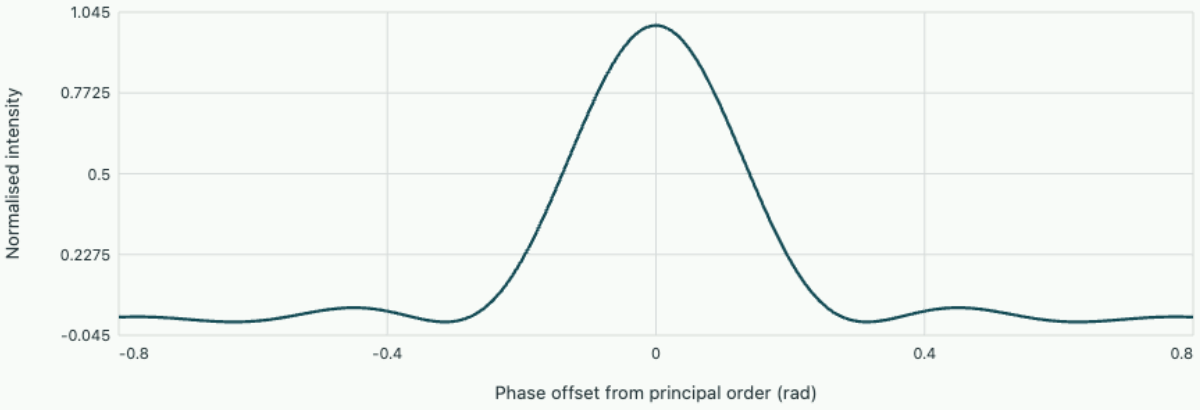
Try it · Grating

Change the inputs and compare the recalculated curves with the preceding derivation and baseline figure. The axes give the units; the numerical readout states the model assumptions and any limiting cases. Reset the inputs to return to the reference setting.

Open this live calculation

Try the calculation

Illuminated grooves **20**



Finite uniform grating

First zeros at ±2π/N = ±0.31416 radians. Absolute peak intensity scales as N² at fixed field per groove.

▶ **Read values from the plot**

Scattering from a central potential and what angular data can recover

Analytical mechanics from constraints to invariants

A detector records angles, whereas the force law refers to separations that were never directly observed. Derive the forward map from incoming impact parameter to angle, distinguish the several trajectories a detector may count in one angular bin, and invert sufficiently complete signed deflection data under stated hypotheses.

Build the deflection integral from a radial turning point

Use relative mass μ, asymptotic relative speed v_∞, incoming energy $E = \mu v_\infty^2/2 > 0$ and impact parameter $b \geq 0$. With a central interaction that tends to zero at infinity, angular momentum is $\ell = \mu v_\infty b$. For a single repulsive encounter let r_0 be the closest approach. Energy gives

$$\dot r^2 = v_\infty^2[1 - U(r)/E - b^2/r^2];$$

r_0 is the outer turning root reached from infinity. Dividing $\dot\varphi = \ell/(\mu r^2)$ by the radial speed and integrating over the outgoing half trajectory gives

$$\Phi(b) = \int_{r_0}^{\infty} \frac{b\,dr}{r^2\sqrt{1 - U(r)/E - b^2/r^2}},$$

$$\chi(b) = \pi - 2\Phi(b).$$

The straight, free trajectory sweeps $\Phi = \pi/2$ between closest approach and infinity and gives zero deflection. For repulsive bend away from the centre this sign convention gives positive χ. A weak attraction can give negative signed deflection even though a detector reporting the angle between incoming and outgoing velocities may supply only its magnitude. If the path winds by more than a complete turn, angular data modulo 2π also lose winding information. The square-root singularity in the integral for an ordinary turning point is integrable. A substitution $r = r_0 + s^2$ makes the near-endpoint quadrature regular. A double turning point requires separate analysis because it can make the angular advance diverge.

Incoming trajectories with parameters between b and $b + db$ represent annular area $2\pi b\,db$ for an axially symmetric beam. At a detected polar angle θ the solid angle of a ring is $2\pi \sin\theta\,d\theta$. Conservation of incoming counts gives

$$\frac{d\sigma}{d\Omega} = \sum_j \frac{b_j}{\sin\theta}\left|\frac{db_j}{d\theta}\right|.$$

The sum includes all incoming impact parameters reaching that angular bin. Correcting measured counts requires incident flux, target number, angular acceptance and efficiency. The formula is a density in solid angle. Its apparent divergence at an isolated angle need not imply an infinite number in a finite acceptance bin. Integrate the density against the actual bin solid angle or use the preimage areas of the bin directly.

Solve a hard sphere and an inverse-square repulsion

A hard sphere of effective exclusion radius a specularly reverses the incoming normal velocity at the impact surface. The angle between incoming and outgoing directions is $\theta = 2\arccos(b/a)$ for $b < a$. Therefore $b = a\cos(\theta/2)$,

$$|db/d\theta| = (a/2)\sin(\theta/2).$$

The factors cancel to give $d\sigma/d\Omega = a^2/4$, independent of polar angle, and total cross section πa^2. Equal polar angular bins do not give equal counts: their solid angles differ by $\sin\theta$. This geometric distinction often accounts for a measured histogram that does not look uniform even for a constant differential cross section.

For $U(r) = K/r$ with $K > 0$, the orbit is a repulsive hyperbola. The analytic relation is

$$\tan(\theta/2) = K/(2Eb).$$

It can be derived from the energy and the angle through the deflection integral: put $u = b/r$ so that $\Phi = \int_0^{u_0} du/\sqrt{1 - (K/Eb)u - u^2}$. Complete the square and evaluate the arcsine at the root to get

$$\Phi = \pi/2 - \arctan[K/(2Eb)].$$

Define $d = K/(2E)$; then $b = d\cot(\theta/2)$,

$$d\sigma/d\Omega = d^2/[4\sin^4(\theta/2)].$$

High precision near the forward direction is physically limited by finite beam size, detector acceptance and any screening of the long-range interaction.

For $d = 2.0\,\mathrm{mm}$ in an abstract mechanical analogue, a deflection of 60° corresponds to $b = 3.46\,\mathrm{mm}$, and the differential cross section is $16\,\mathrm{mm}^2/\mathrm{sr}$. The cross section for deflections greater than 60° is the area $\pi b^2 = 37.7\,\mathrm{mm}^2$. The differential value and integrated area answer different laboratory questions. Doubling energy at fixed interaction halves d, quarters the cross section into a fixed angular interval and reduces the deflection at a fixed incoming b. Measuring only a total count without specifying angular threshold is inadequate for a long-range interaction.

Recover impact parameter from angular counting under monotonicity

Suppose the deflection decreases continuously from π at $b = 0$ to zero at large b. Integrating counts for all angles at least θ then gives precisely the incident disk $\pi b(\theta)^2$. The recovered map satisfies

$$b(\theta)^2 = 2\int_\theta^\pi (d\sigma/d\Omega)(\theta')\sin\theta'\,d\theta'.$$

Applying this formula to constant hard-sphere intensity gives $b^2 = a^2(1+\cos\theta)/2$, the geometrical result. It also exposes the information needed from a detector: a finite upper cutoff can leave the reconstruction with an unknown disk-area offset.

Failing monotonicity, the same cumulative counts give the total area of all incoming regions producing deflections above threshold, which can be a collection of annuli. No unique $b(\theta)$ follows from that one number. Near a smooth extremum $\theta(b) = \theta_r + c(b-b_r)^2/2 + \cdots$, two branches contribute inverse square-root factors proportional to $1/\sqrt{|\theta-\theta_r|}$. A finite bin contains finite counts because integrating that singularity gives a square root. Tagging the incoming impact parameter in a controlled beam or measuring trajectory time delay can separate branches invisible in a final-angle histogram.

Invert a signed deflection function at one energy

Assume smooth $U(r)$ with $\rho(r) = r\sqrt{1-U(r)/E}$ strictly increasing on the radial interval sampled by the trajectories. The turning condition is

$$\rho(r_0) = b.$$

Write $W(\rho) = \ln[r(\rho)/\rho]$ and transform the deflection integral to $\Phi = b\int_b^\infty [d\ln r/d\rho]\, d\rho/\sqrt{\rho^2-b^2}$. The term from $d\ln\rho/d\rho$ gives the free angle $\pi/2$; the remainder has $\chi = -2b\int_b^\infty W'(\rho)\, d\rho/\sqrt{\rho^2-b^2}$. This is an Abel-type map. Multiply by $1/\sqrt{b^2-s^2}$ and integrate from $b=s$ to infinity. Interchanging the integrals produces the inner integral

$$\int_s^\rho \frac{b\,db}{\sqrt{b^2-s^2}\sqrt{\rho^2-b^2}} = \frac{\pi}{2}.$$

Substitute $b^2 = s^2 + (\rho^2-s^2)\sin^2\alpha$ to check the identity. With $W(\infty)=0$ the inversion is $W(s) = \pi^{-1}\int_s^\infty \chi(b)\, db/\sqrt{b^2-s^2}$. Recover $r = se^{W(s)}$ and finally $U(r) = E[1-(s/r)^2]$ parametrically. This derivation explains the restrictions. A trajectory only samples radii outside its turning point, the signed angle must retain branch information, and the map to ρ must be single valued. Unknown large-b tails influence the integral even when only comparatively small radii are of interest.

Derive the weak-deflection formula from a straight incoming trajectory

When the deflection is small one can calculate the transverse impulse along the undeflected line before evaluating a nonlinear radial integral. Put the incoming velocity along x and the impact point at $y = b > 0$, so the unperturbed trajectory is $x = v_\infty t$, $r = \sqrt{x^2+b^2}$. The transverse force is $-U'(r)b/r$. Dividing its time integral by the incoming momentum μv_∞ gives the signed leading deflection

$$\chi(b) \simeq -\frac{b}{E}\int_b^\infty \frac{U'(r)\,dr}{\sqrt{r^2-b^2}}.$$

The factor follows from replacing dt by dx/v_∞, using symmetry between positive and negative x, and the identity $dx = r\,dr/\sqrt{r^2 - b^2}$. A decreasing positive potential gives a positive angle, consistent with repulsive scattering in the chapter's convention. The approximation neglects longitudinal momentum change at leading order; a finite speed change far from the interaction would also be inconsistent with elastic energy conservation. Near a turning barrier or for a large angle the original deflection integral must replace this straight path.

An analytically tractable finite-range model is $U(r) = U_0 \exp[-r^2/(2a^2)]$. Along the incoming straight line the transverse force is $(U_0 b/a^2)\exp[-(x^2 + b^2)/(2a^2)]$. The Gaussian x integral gives

$$\chi(b) \simeq \sqrt{\pi/2}(U_0/E)(b/a)\exp[-b^2/(2a^2)].$$

A central hit has zero transverse force by symmetry if the particle crosses through the weak smooth potential rather than being reflected. Distant particles see little interaction, so the deflection must peak at an intermediate impact parameter. Differentiation places the maximum at $b = a$, with height $\sqrt{\pi/(2e)}U_0/E$.

For $U_0/E = 0.10$ the peak is about $0.0760\,\mathrm{rad} = 4.36^\circ$, comfortably small for this leading calculation. Two impact-parameter families can contribute to one smaller nonzero angle. For example $b = 0.50a$ gives $0.0553\,\mathrm{rad}$; a trajectory with b between $1.5a$ and $1.7a$ produces a similar deflection on the declining branch. A detector counting more events near four degrees may be recording the merging of these branches, rather than a particularly large local force at the turning separation. Reconstructing $b(\theta)$ from monotonically accumulated disk areas would be invalid for this smooth repulsive bump even though every encounter is elastic.

Measure the energy dependence to distinguish the weak limit. For fixed b/a the leading angle scales as $1/E$. Doubling energy halves the predicted peak angle. A finite spread of incident energies therefore broadens the observed rainbow before detector angular resolution is considered. A position-tagged incident beam can furnish $\chi(b)$ point by point; a detector measuring final angles alone mixes weakly and more strongly penetrating paths. Adding incoming-position information changes what inverse problem is identifiable, without improving the detector's angular precision.

Limit a scattering inference using finite range and measurement bandwidth

The Abel reconstruction requires data to arbitrarily large b. A finite experiment measures up to $b_{\max}$, and the unmeasured tail contributes an integral to $W(s)$. For the example of a known asymptotic $\chi = C/b$, the tail from $b_{\max}$ to infinity evaluates to $C\,\arcsin(s/b_{\max})/(\pi s)$ for $s < b_{\max}$. Substitute $b = s/\cos\alpha$ to obtain this result. For $s \ll b_{\max}$, it approaches $C/(\pi b_{\max})$. Setting the tail to zero therefore introduces an offset across the reconstructed log-radius even at small s. Bounded angular noise and uncertainty in the tail model should be propagated separately.

As a numerical illustration take $C = 5.0\,\mathrm{mm}$,

$$b_{\max} = 100\,\mathrm{mm},$$
$$s = 20\,\mathrm{mm}.$$

The missing contribution is about 0.0160, corresponding to roughly a 1.6% multiplicative error in reconstructed radius if other terms were exact. At that same s, the leading whole weak-tail value $C/(2s) = 0.125$ would be biased appreciably by this omission. A model with a short exponential tail would give a quite different correction. Adding a physically motivated tail with an uncertainty range is more informative than extending a noisy measured curve by an unrestricted polynomial.

The lower integration endpoint has a square-root weight. Uniform samples in b can be wasteful near a selected reconstruction radius and coarse interpolation there can dominate the result. Substitute $b = s\cosh u$ to obtain

$$db/\sqrt{b^2 - s^2} = du.$$

The Abel integral becomes a straightforward integral of $\chi(s\cosh u)$ over u, with a finite upper limit if the data stop at $b_{\max}$. This substitution removes a numerical singularity but adds no information about missing impact parameters. Validate a reconstruction by inserting its inferred potential back into the original forward integral and predicting angular measurements withheld from the fit. Agreement with the very data used in the inversion alone can conceal an incorrect tail or a misassigned branch.

Convert centre-of-momentum angles into laboratory angles

An ordinary collision experiment often launches a particle of mass m_1 into a stationary particle of mass m_2. In Newtonian elastic scattering the centre-of-mass speed is $V = m_1 v_{\mathrm{in}}/(m_1 + m_2)$, while the projectile's speed in that frame is $u_1 = m_2 v_{\mathrm{in}}/(m_1 + m_2)$. Rotate that velocity by the centre-of-mass angle χ in the scattering plane, then add the centre velocity back. The outgoing laboratory velocity components are $V + u_1\cos\chi$ longitudinally and $u_1\sin\chi$ transversely. Consequently $\theta_{\mathrm{lab}} = \operatorname{atan2}(\sin\chi, m_1/m_2 + \cos\chi)$. The centre-of-mass solid-angle density derived from the impact parameter does not transfer unchanged to a lab angular histogram.

If $m_1 = m_2$, the identity $1 + \cos\chi = 2\cos^2(\chi/2)$ gives $\theta_{\mathrm{lab}} = \chi/2$ except when the projectile stops and loses a well-defined direction. A heavy projectile can be restricted to a narrow forward cone. Let $r = m_1/m_2 > 1$. Differentiating the laboratory angle gives

$$d\theta_{\mathrm{lab}}/d\chi = (1 + r\cos\chi)/(1 + 2r\cos\chi + r^2).$$

It vanishes at $\cos\chi = -1/r$. At that angle $\sin\theta_{\max} = 1/r$; hence $r = 2$ limits the projectile to $\theta_{\max} = 30^\circ$. Two centre-of-mass angles can generate the same nonzero laboratory angle below this maximum. This laboratory-frame fold may appear even for a monotone centre-of-mass repulsive potential.

Energy helps separate the branches. The outgoing laboratory projectile energy divided by its incident energy is $[m_1^2 + m_2^2 + 2m_1m_2\cos\chi]/(m_1 + m_2)^2$. For $m_1 = 2m_2$, head-on reversal in the momentum frame gives energy fraction $1/9$, while forward scattering gives one. Both can land near the laboratory beam direction. A detector reporting angle only cannot distinguish a barely deflected fast projectile from a strongly deflected slow one emerging along almost the same line. Recording time of flight to a downstream detector or measuring the recoil supplies the missing information. Instrumental and kinematic multiplicity should therefore be addressed before attributing a two-branch angular signal to an exotic force law.

Ask what a flight-time measurement adds to angle data

At a given impact parameter choose a large comparison sphere of radius R enclosing a short-range potential. The transit time from its incoming boundary to the outgoing boundary is $t_U(R,b) = 2v_\infty^{-1}\int_{r_0}^{R} dr/\sqrt{1-U/E-b^2/r^2}$. The free reference at the same b is $t_0 = 2\sqrt{R^2-b^2}/v_\infty$. Their difference is a classical time delay under a precisely specified pair of surfaces. It combines a different path length with the altered speed along the path. A repulsive barrier can cause either a shorter geometrical penetration or longer dwell near a turning point, so the sign cannot be inferred merely from the word repulsive.

For an impenetrable sphere of radius $a > b$, the particle moves at free speed on each straight segment. The radial turning point is replaced by specular contact at a, so $t_{\mathrm{hard}} = 2[\sqrt{R^2-b^2} - \sqrt{a^2-b^2}]/v_\infty$. The delay is $-2\sqrt{a^2-b^2}/v_\infty$. Using $a = 3.0\,\mathrm{mm}$, $b = 1.8\,\mathrm{mm}$, and $v_\infty = 2.0\,\mathrm{m/s}$, one finds $-2.4\,\mathrm{ms}$. The hard sphere takes less transit time between the two points where the respective trajectories cross the large sphere. Those points differ from the endpoints of the free line, so this negative delay implies no acceleration above the fixed free speed.

For a potential with a true long-range $1/r$ tail the simple free-reference difference can acquire a logarithmic dependence on the comparison radius. A reported time delay then needs its reference geometry and tail model, just as a measured forward angle needs its angular acceptance. In the inverse problem, combining deflection and delay can help separate long dwell near a nearly unstable circular orbit from a large geometrical bend at an ordinary turning point. A large angle alone contains insufficient information to assign the time spent close to the scattering centre.

Reconstruct a finite-depth potential well from two deflections

A circularly symmetric square well provides a forward problem in which the trajectory is straight in each region and changes direction at the potential boundary. Let $U(r) = -U_0 < 0$ for $r < a$ and zero outside. A particle with $b < a$ enters with speed v_∞ and travels inside at $v_{\mathrm{in}} = v_\infty\sqrt{1+U_0/E} \equiv n v_\infty$. Angular momentum conservation across the radial impulse gives

$$v_\infty \sin i = v_{\mathrm{in}} \sin r_i,$$

where i and r_i are angles of the velocity to the normal outside and inside. Thus $\sin i = b/a,\ \sin r_i = b/(na)$. At each crossing the direction bends toward the centre. The magnitude of net deflection is $2(i - r_i)$, with negative signed deflection in the convention used for attraction here.

Take $n = 1.25,\ b = 0.60a$. Then $i = 36.87^\circ$, $r_i = 28.69^\circ$, and the total deflection has magnitude 16.35°. As b approaches a from below, the magnitude approaches $2[\pi/2 - \arcsin(1/n)]$, about 73.7°. For $b > a$ the trajectory does not touch the ideal sharp boundary and is undeflected. A smooth finite-width transition removes the ideal discontinuity at grazing incidence and can generate trajectories that linger near the outer part of the well. Examining a well with measured edge width is therefore different from trying to fit all grazing events with a perfectly sharp step.

If a is known, an incoming position tag and a measured deflection at one b determine $n = \sin i/\sin(i - |\chi|/2)$, hence $U_0/E = n^2 - 1$. The preceding numbers give

$$U_0/E = 0.5625.$$

For an independently known E this determines the well depth. If a is unknown too, one angle supplies only one equation for two parameters. Multiple impact parameters, an independent threshold for entering the well, or a measured time of flight can supply the additional scale. Near $b = 0$ the deflection is approximately $2(b/a)(1 - 1/n)$, so a series of very small impact parameters measures mainly that one combination; broad coverage is needed to resolve a and n individually.

The time inside the well is $2a \cos r_i/(nv_\infty)$, because the interior chord has offset b/n and speed nv_∞. At the same b, the time between intersections of the free path with the circle is $2a \cos i/v_\infty$. For $n = 1.25, b = 0.6a$, their ratio is $\cos 28.69^\circ/[1.25 \cos 36.87^\circ] \approx 0.877$. The faster interior motion crosses a longer chord in less time. For grazing free paths the reference chord becomes very short and the comparison can reverse. Reporting which part of the trajectory the timing equipment resolves is therefore as important as measuring the dwell accurately.

Recover incoming annuli from scattering angles

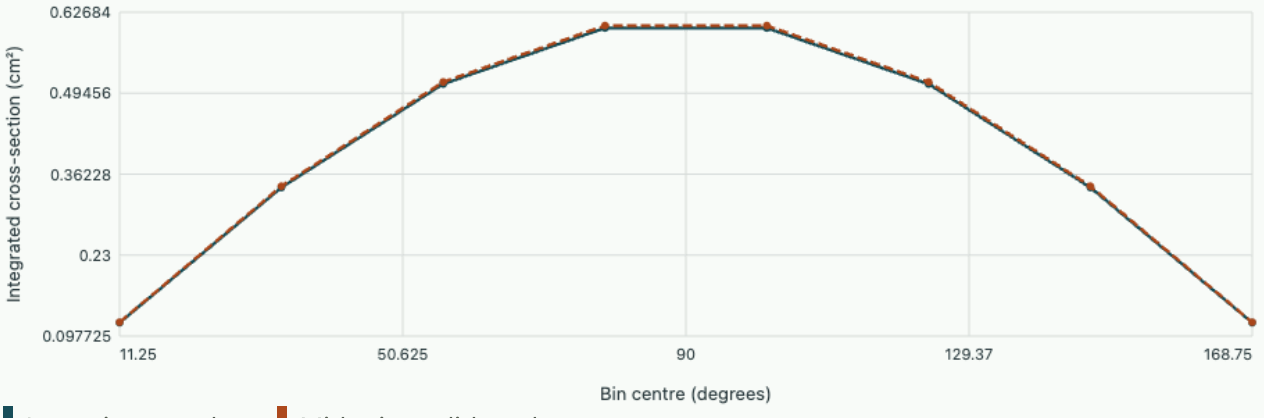
Bin the exactly solvable hard-sphere reflection in polar angle and compare the integrated cross-section with the incoming geometric area.

Open this experiment online

Try the calculation

Polar detector bins **8**

Effective radius (cm) **1**



Exact sum 3.1416 cm². Midpoint sum 3.1619 cm². Differential cross-section 0.25 cm²/sr.

Propagate uncertainties from a detector ring to an inferred force

Suppose N counts are recorded in an angular ring and the product of incident fluence, target count and calibrated efficiency is C. The estimate of bin-averaged cross section is $N/(C\Delta\Omega)$, with $\Delta\Omega = 2\pi[\cos\theta_{\mathrm{low}} - \cos\theta_{\mathrm{high}}]$ for full azimuth. Approximately independent counting events have fractional standard error $1/\sqrt{N}$, but the fluence and efficiency create correlated scale errors across angular bins. When integrating bins to reconstruct $b^2(\theta)$, a shared five-percent normalisation uncertainty does not shrink just because the number of angular bins grows. It propagates into an approximately 2.5% scale uncertainty in the recovered impact parameter.

For an ideal hard sphere, let an experiment integrate the area feeding angles greater than 90° and find $14.1 \pm 0.7\,\mathrm{mm}^2$. Since this is $\pi b(90^\circ)^2 = \pi a^2/2$, the inferred radius is $\sqrt{2(14.1)/\pi} = 3.00\,\mathrm{mm}$ with uncertainty approximately $0.075\,\mathrm{mm}$ for small errors. If the unobserved backward final ten degrees are simply omitted, an extra solid angle $\Delta\Omega = 2\pi[1 - \cos 10^\circ]$ is missed. Its small size for a smooth hard-sphere law can be estimated, but a general inverse analysis cannot assume that backward intensity is smooth without testing it.

Is multiplying each measured bin centre by the angular density there sufficient for a steep forward-scattering law?

▼ **Work through the solution**

A finite bin measures an integral against solid angle, not the value at its centre. For a rapidly varying $\sin^{-4}(\theta/2)$ density near zero, the lower-angle part of the bin can dominate its count. Use the exact preimage annulus $\pi[b(\theta_{\mathrm{low}})^2 - b(\theta_{\mathrm{high}})^2]$ or integrate the model numerically over the detector's angle-dependent efficiency. Repeat with narrower bins as a check. If the lower acceptance overlaps the direct beam, separate the unscattered population and account for its instrumental angular spread before assigning a cross section to small deflections.

Work directly in counts for a finite annular detector

For the repulsive inverse-square example take $d = 2.0\,\mathrm{mm}$ and a full-azimuth ring that accepts polar angles from 55° to 65°. The expected incident area is $\pi d^2[\cot^2(27.5^\circ) - \cot^2(32.5^\circ)]$, about $15.4\,\mathrm{mm}^2$. Multiplying by an incident fluence of 100 particles per square millimetre and efficiency 0.80 predicts about 1230 counts. A Poisson fluctuation scale is then roughly 35 counts. The exact integral supplies a meaningful reference even when the plotting software reports a differential value at the centre of the bin.

Moving both angular boundaries one degree outward decreases the count because the repulsion sends fewer particles through larger angles. The change can be calculated by reevaluating the two cotangents, keeping the fluence and efficiency fixed. This sensitivity allows a calibration uncertainty in the angular zero to be propagated into a count prediction without treating it as a force uncertainty. If fluence is independently measured the dependence on d^2 can then estimate the interaction scale. If fluence is itself unknown, a single count measures only their product; independent calibration or measurements across energies are needed to separate them.

Tutorial on inverse scattering

An incoming uniform beam scatters from a hard disk in three-dimensional impact geometry with effective radius $3.0\,\mathrm{mm}$. What area feeds the polar bin from 60° to 90°?

▼ **Work through the solution**

The accepted impact parameters run from $a\cos 45^\circ$ to $a\cos 30^\circ$. The difference of their squared radii is $9(0.75-0.50)=2.25\,\mathrm{mm}^2$, giving area $2.25\pi = 7.07\,\mathrm{mm}^2$. Alternatively the bin has solid angle $2\pi(\cos 60^\circ-\cos 90^\circ)=\pi$ steradians and differential cross section $a^2/4 = 2.25\,\mathrm{mm}^2/\mathrm{sr}$. The agreement checks both the Jacobian and the angular convention.

At large impact parameter a weak repulsion gives $\chi(b)\simeq C/b$ with C a length. Use the inversion to identify the leading tail of the potential.

▼ **Work through the solution**

Substitute $b = s/\cos\alpha$ into the Abel integral. One obtains

$$W(s)\simeq C/(2s).$$

Thus $r\simeq s + C/2$ and $U/E = 1-e^{-2W}\simeq C/s\simeq C/r$. The potential tail is EC/r. For the exact repulsive Coulomb result $C = K/E$, matching the original K/r. This check uses the weak-tail expansion at large radii only; continuing it to small radii would discard higher deflection terms.

A measured angular bin near a rainbow doubles in width towards its allowed side. By what approximate factor do its counts change if background is negligible?

▼ **Work through the solution**

At a simple fold the angular density grows as the inverse square root of distance from the extremal angle. For a bin of width $\Delta\theta$ starting at the fold, the integral scales as $\sqrt{\Delta\theta}$ provided $\sin\theta_r$ is nonzero and other weights are locally smooth. Doubling the width gives a factor $\sqrt{2}$, unlike the factor two from a smooth background. Finite angular resolution rounds off the divergence and can obscure this scaling in very narrow bins.

Compare angular bins and conserved incoming area.

Fließbach's mechanics text treats trajectories in central fields.

Action, angle and the slowing clock at a separatrix

Analytical mechanics from constraints to invariants

A phase portrait contains enough information to extract a period by differentiating an area. Use that fact to build a clock for finite-amplitude pendulum motion, determine precisely where slow-parameter invariance fails, and distinguish a librating trajectory from one that makes complete rotations.

Turn a closed phase-space loop into an action variable

For a one-degree-of-freedom Hamiltonian $H(q,p;\lambda)$ at fixed parameter λ, take a connected periodic orbit of energy E. Define its action by $J(E,\lambda)=(2\pi)^{-1}\oint p\,dq$, oriented along the actual motion. The units are energy times time. A smooth family of nested librating loops has $2\pi J$ equal to the enclosed canonical phase area. For circulation on a coordinate cylinder, such as an angle increasing through 2π, the integral still defines an action even though the loop winds around the cylinder. Its area interpretation and sign need a reference branch.

At fixed q,λ, differentiating $H(q,p(E,q))=E$ gives

$$\partial_E p = 1/H_p = 1/\dot{q}.$$

Differentiating the action integral yields $\partial_E J = T/(2\pi)$, because $dq/\dot{q}=dt$ around the orbit. At ordinary turning points $\dot{q}=0$, the derivative integrand has an integrable singularity while the two endpoints of the closed loop agree; one can justify the derivative using a local square-root coordinate. Inverting gives

$$\omega(J)=\partial_J E.$$

If θ is the angle conjugate to J, the transformed Hamiltonian is $E(J)$ and $\dot{\theta}=\omega(J)$ is constant. This canonical angle generally differs from the geometric angle reported by a camera.

For a unit-mass harmonic oscillator the energy curve $p^2+\omega_0^2q^2=2E$ is an ellipse. Its area is $2\pi E/\omega_0$, giving $J=E/\omega_0$ and a frequency independent of amplitude. One may take $q=\sqrt{2J/\omega_0}\sin\theta,\ p=\sqrt{2J\omega_0}\cos\theta$, whose bracket is one. The position turns nonuniformly in coordinate space even though the phase angle advances uniformly. Under a slow change in ω_0, constant J predicts $E\propto\omega_0$ and amplitude proportional to $1/\sqrt{\omega_0}$. Constant action does not mean constant mechanical energy while work is performed by a changing spring.

Integrate the finite-amplitude pendulum

Let $I=ml^2$, $\omega_0^2=g/l$, and measure q from downward vertical. The Hamiltonian $H=p^2/(2I)+I\omega_0^2(1-\cos q)$ has separatrix energy $E_s=2I\omega_0^2$. For a libration of turning angle $q_m<\pi$, define $k=\sin(q_m/2)=\sqrt{E/E_s}$. The positive momentum branch is $p=2I\omega_0\sqrt{k^2-\sin^2(q/2)}$. Set $\sin(q/2)=k\sin u$. Evaluating four quarter-arcs of the loop gives

$$J_{\mathrm{lib}} = \frac{8 I \omega_0}{\pi}\big[\mathsf{E}(k) - (1-k^2)\mathsf{K}(k)\big],$$

$$T_{\mathrm{lib}} = \frac{4}{\omega_0}\mathsf{K}(k).$$

Here $\mathsf{K}(k) = \int_0^{\pi/2} (1 - k^2 \sin^2 u)^{-1/2} du$ and $\mathsf{E}(k) = \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2 u}\, du$. To verify the first formula after substitution, note that $k^2 \cos^2 u = (1 - k^2 \sin^2 u) - (1 - k^2)$. The transformed integral of $k^2 \cos^2 u / \sqrt{1 - k^2 \sin^2 u}$ splits exactly into the two quoted complete elliptic integrals. The period follows either by $dq/\dot q$ or by differentiating the action. That independent check is useful when numerical differentiation of noisy action estimates amplifies sampling errors.

At small k,

$$\mathsf{K} = (\pi/2)(1 + k^2/4 + \cdots),$$

$$T \approx (2\pi/\omega_0)(1 + q_m^2/16 + \cdots).$$

At $q_m = 60^\circ$, $k = 1/2$ and $\mathsf{K} \simeq 1.68575$, giving $T/T_{\mathrm{small}} = 1.0732$. A pendulum of length $1.00\,\mathrm{m}$ then has period about $2.153\,\mathrm{s}$, where the small-angle answer is $2.006\,\mathrm{s}$. Timing ten periods rather than one helps resolve the accumulated difference. Measuring the arc amplitude from a projected image requires a geometric correction if the swing plane is tilted relative to the camera.

Approach the saddle from either side

At the upright position $q = \pi$, the potential is maximal and the equilibrium is a hyperbolic saddle in phase space. Put $q = \pi + s$ with small s; the energy near the saddle is $H \simeq E_s + p^2/(2I) - I\omega_0^2 s^2/2$. For an orbit at $E_s - \Delta E$, the time spent in a fixed small neighbourhood includes an integral proportional to $\int ds / \sqrt{s^2 - 2\Delta E/(I\omega_0^2)}$. As ΔE decreases, this develops a logarithm. Equivalently $\mathsf{K}(k) \sim \ln[4/\sqrt{1-k^2}]$ as $k \to 1^-$. The separatrix encloses finite canonical area but its period is infinite. Hence $\partial_E J$ diverges although J itself remains finite.

On the positive rotation branch $E > E_s$, set $k_r = \sqrt{E_s/E} < 1$. A period is the time for q to increase through 2π, so there is only one positive branch of momentum in its action integral. Direct integration gives

$$J_{\mathrm{rot}} = (2/\pi)\sqrt{2IE}\,\mathsf{E}(k_r),$$

$$T_{\mathrm{rot}} = 4\sqrt{I/(2E)}\,\mathsf{K}(k_r).$$

At very high energy, the potential is a small modulation and these formulas approach free-rotor values $\sqrt{2IE}$ and $2\pi\sqrt{I/(2E)}$. At the separatrix the positive rotation action approaches $4I\omega_0/\pi$, half the limiting libration action $8I\omega_0/\pi$. The two actions count different closed paths on a cylinder. The factor two is a topological change in the cycle, not a discontinuous jump in the mechanical momentum at a point.

Test slow conservation against the actual orbit period

To treat slowly varying parameters, define instantaneous action–angle coordinates for each fixed λ. The generating function $W(q, J, \lambda)$ acquires explicit time dependence when $\lambda = \lambda(t)$, adding $\dot{\lambda}\,\partial_\lambda W$ to the transformed Hamiltonian. Its angle dependence produces an oscillating $\dot{J} = -\dot{\lambda}\,\partial_\theta \partial_\lambda W$. On a regular periodic orbit, the integral of an angle derivative over a cycle vanishes. If λ changes little during that cycle and there is no resonant complication, this cancellation explains why the action changes little on average. The requirement uses the actual orbit period at the chosen energy. Near a separatrix it cannot be justified merely by comparing a drive rate to the small-amplitude period.

Consider a parameter ramp lasting $20\,\mathrm{s}$ for the one-metre pendulum. Near the bottom this covers about ten periods. For $1 - E/E_s = 10^{-6}$, the asymptotic $\mathsf{K} \simeq \ln 4000 = 8.29$ gives a period of approximately $10.6\,\mathrm{s}$, so the same ramp spans less than two cycles. Even longer ramps fail the scale separation sufficiently close to the saddle. If a varying parameter changes the separatrix area until it meets the trajectory, continued use of a single analytic action branch is especially unreliable. A trajectory can pass into circulation or return to libration depending on its phase and the ramp history.

Compute an action without measuring momentum continuously

Suppose a camera tracks one-dimensional positions over one full cycle, and velocity estimates are noisy. For a known position-dependent potential and calibrated mass, an independent measurement of the turning energy allows phase-area integration without differentiating the camera trace. On a bound orbit $p_\pm(q) = \pm\sqrt{2m[E - V(q)]}$, so $J = \pi^{-1}\int_{q_-}^{q_+} \sqrt{2m[E - V(q)]}\,dq$. One needs a potential calibration, both turning positions and a common energy datum. An additive constant in the potential cancels when E is shifted by the same amount. An unrecognised dissipative force makes the inward and outward energy profiles different and invalidates the assumed single closed energy curve.

Take a particle in the potential $V(q) = a|q|$, $a > 0$, which can approximate opposed constant restoring forces away from a narrow central smoothing zone. Its turning displacements are $\pm E/a$. Integrate on the positive half and use symmetry to find

$$J = 4\sqrt{2m}E^{3/2}/(3\pi a).$$

Differentiating gives

$$T = 2\pi\,dJ/dE = 4\sqrt{2mE}/a.$$

The same period follows by constant acceleration: starting at rest at a positive turning point, the fall to the origin takes $\sqrt{2mE}/a$, one quarter period. Agreement across a nondifferentiable potential cusp is a useful check that the action calculation handles the turning boundaries rather than requiring sinusoidal motion.

For $m = 0.10\,\mathrm{kg}$, $a = 2.0\,\mathrm{N}$, and $E = 0.20\,\mathrm{J}$, the amplitude is $0.10\,\mathrm{m}$, action is $0.00849\,\mathrm{J\,s}$, and period is $0.400\,\mathrm{s}$. If the restoring strength a doubles slowly at fixed action, $E \propto a^{2/3}$ increases by $2^{2/3} = 1.59$, while the turning amplitude shrinks by $2^{-1/3} = 0.794$. A harmonic-oscillator action rule $J = E/\omega$ using a single measured frequency would be incorrect here: the actual $E/\omega = ET/(2\pi)$ exceeds the action by a

factor $3/2$. The equality found for a harmonic oscillator follows from its linear energy–action relation and need not generalise.

Separate the angle accumulated during a cycle from the action it encloses

In a slowly modulated oscillator an almost constant action can coexist with a measurable phase error. Consider $\ddot{q} + \omega(t)^2 q = 0$ with smooth positive ω. An approximate oscillation is $q = A(t)\cos\Phi(t)$. At leading order $\dot{\Phi} = \omega$. Cancelling the first-order sine term gives $2\dot{A}\omega + A\dot{\omega} = 0$, so $A\sqrt{\omega}$ is approximately constant. The amplitude law reproduces conservation of E/ω after averaging. The accumulated phase is $\int \omega(t)\,dt$, which can change by many radians during a slow parameter cycle even if the initial and final frequency and action agree.

For $\omega(t) = 10[1 + 0.10\sin(2\pi t/T_r)]\,\mathrm{s}^{-1}$ and $T_r = 20\,\mathrm{s}$, integration over one full ramp gives $200\,\mathrm{rad}$ of leading phase. The largest value of $|\dot{\omega}|/\omega^2$ is only a few 10^{-3}. To first order the final amplitude returns to its initial value, but the position at the end is set by a phase that is generally not an integer multiple of 2π. A calculation that expects to return to its initial point after the parameter cycle confuses action invariance with angle closure. A small constant frequency calibration error, integrated over a long ramp, can dominate the final phase error while producing little amplitude error.

A sudden frequency switch gives an instructive opposite limit with exact phase dependence. At fixed q, p, changing ω_1 instantly to ω_2 changes the action to $J_2 = p^2/(2m\omega_2) + m\omega_2 q^2/2$. Write the initial state at action J_1 using the harmonic angle. Then

$$J_2/J_1 = (\omega_1/\omega_2)\cos^2\theta + (\omega_2/\omega_1)\sin^2\theta.$$

For a frequency doubling, the ratio ranges from one half at a passage through the centre to two at a turning point. For a uniformly sampled ensemble the average ratio is 1.25, although the pre-switch and post-switch Hamiltonians are both perfectly integrable.

An experiment can therefore test slowness by preparing the same initial action at different phases and comparing the distribution of final actions. A broad phase-dependent distribution for a rapid switch narrows as the ramp becomes slow on a regular family of tori. Near a separatrix a slow external ramp can still yield a broad distribution because the orbital clock itself stalls. The size of the instantaneous energy, measured without phase and the final Hamiltonian, does not distinguish these mechanisms.

Resolve a separatrix in time without integrating arbitrarily close to the saddle

At exactly E_s, energy conservation gives $p = \pm 2I\omega_0\cos(q/2)$ on $-\pi < q < \pi$. Choose the increasing branch passing through $q = 0$ at $t = 0$. Separation of variables in $\dot{q} = 2\omega_0\cos(q/2)$ gives

$$\tan(\pi/4 + q/4) = e^{\omega_0 t}.$$

Therefore $q(t) = 4\arctan(e^{\omega_0 t}) - \pi$,

$$p(t) = 2I\omega_0 \operatorname{sech}(\omega_0 t).$$

This exact orbit approaches the upright saddle exponentially at both infinite endpoints. The two endpoints represent the same physical angle on the cylinder even though they differ by 2π on an unwrapped plot.

The separatrix momentum falls to one tenth of its maximum after $\omega_0 t = \operatorname{arcosh} 10 \approx 2.993$. At $\omega_0 = 3.13\,\mathrm{s}^{-1}$ that takes only $0.956\,\mathrm{s}$. Reaching one hundredth takes $\operatorname{arcosh} 100/\omega_0 \approx 1.69\,\mathrm{s}$. The logarithmic increase in time per retained significant digit explains why an apparently captured upright state can simply be a trajectory very close to the saddle. Measuring angular velocity as well as angular displacement makes the exponential approach easier to distinguish from a truly stationary upright solution.

The phase area enclosed by both separatrix branches can also be recovered directly. Integrating their difference gives

$$\int_{-\pi}^{\pi} 4I\omega_0 \cos(q/2)\,dq = 16I\omega_0.$$

Dividing by 2π yields the libration limit $8I\omega_0/\pi$ obtained using elliptic integrals. By contrast a trajectory winding once on the upper branch integrates only the upper momentum, giving half that value. Testing these two integrals on an unwrapped numerical phase trace can reveal a missing lower turning branch or a routine that silently wraps the angle before accumulating $p\,dq$.

Find a frequency directly from noisy orbital areas

Suppose a calibrated experiment measures the phase areas $A(E) = \oint p\,dq$ at several energies. With $J = A/(2\pi)$, the identity $dA/dE = T$ says that an area-versus-energy slope directly has units of time. Take three close energies $0.09, 0.10, 0.11\,\mathrm{J}$, with measured areas $0.180, 0.202, 0.225\,\mathrm{J\,s}$. A central difference gives $T \simeq (0.225 - 0.180)/0.020 = 2.25\,\mathrm{s}$ at the centre. If each area has an independent uncertainty $0.001\,\mathrm{J\,s}$, the propagated standard uncertainty of that difference is about $0.071\,\mathrm{s}$, before accounting for energy uncertainty and curvature of $A(E)$.

Decreasing the energy spacing would reduce smooth finite-difference bias but amplify the effect of those same area errors. A local fit over several energies can trade variance against the assumption of a smooth slope. Near a separatrix the actual slope varies rapidly, so a wide polynomial fit may smooth away precisely the period increase being sought. Compare with direct return times at selected energies, or fit the logarithmic form justified by the saddle expansion when the orbit is demonstrably in its asymptotic regime. A measured phase area can be highly accurate while an inferred frequency remains poorly constrained because it is a derivative.

A practical area estimate from ordered camera samples uses the polygon formula $\tfrac{1}{2}\sum_i (p_i + p_{i+1})(q_{i+1} - q_i)$, including the return to the initial point with a consistent sign. Very long time sampling can overrepresent the slow turn neighbourhood but need not corrupt a properly oriented geometric line integral. It can make a uniform-clock finite difference of velocity near a turnaround noisy, however. If a dataset includes several cycles with dissipation, do not simply close the final state back to

the beginning of the first cycle. Find individual return events, estimate the action per cycle and test whether its change is consistent with measured losses.

Construct the canonical angle on two sides of a turning point

For a bound one-dimensional potential take a reference crossing at the lower turning point and choose the increasing-q branch. A Hamilton–Jacobi generating integral is

$$W(q, J) = \int_{q_-}^{q} p(q', E(J))\, dq'.$$

Its angle is $\theta = \partial_J W = (dE/dJ) \int_{q_-}^{q} dq'/\dot{q}' = \omega t_{\mathrm{flight}}$ on that branch. At the upper turning point $\theta = \pi$ for a time-reversal-invariant kinetic-plus-potential Hamiltonian. On the return branch the same coordinate q must be assigned the complementary angle $2\pi - \theta_{\mathrm{outgoing}}(q)$, because its momentum has opposite sign. The coordinate alone is insufficient to reconstruct a unique phase.

For a harmonic oscillator beginning at its leftmost position a quarter-amplitude point is crossed at two different times. In coordinates $q = -A\cos\theta$ take $q = -A/2$. The increasing branch gives

$$\theta = \pi/3;$$

the returning branch gives $5\pi/3$. Their velocities have opposite signs. Using the principal arccosine for every time sample forces both branches onto the same interval and makes an apparently discontinuous phase that never completes a cycle. Measuring the sign of velocity or interpolating successive turning events restores the appropriate branch. Near a saddle, tiny velocities make this branch identification particularly sensitive to camera noise.

Once a reliable angle is available, the phase deviation from the uniform prediction $\theta(t_0) + \omega(t - t_0)$ becomes a useful diagnostic. A static potential with a wrongly calibrated shape can cause a phase error periodic in position even when its return time is fitted correctly. Dissipation usually changes the action as well, while a constant clock error causes a steadily accumulating phase drift. Compare the residual over several cycles at different amplitudes and plot it against reconstructed angle as well as laboratory time. These alternate plots test different physical explanations of the same timing mismatch.

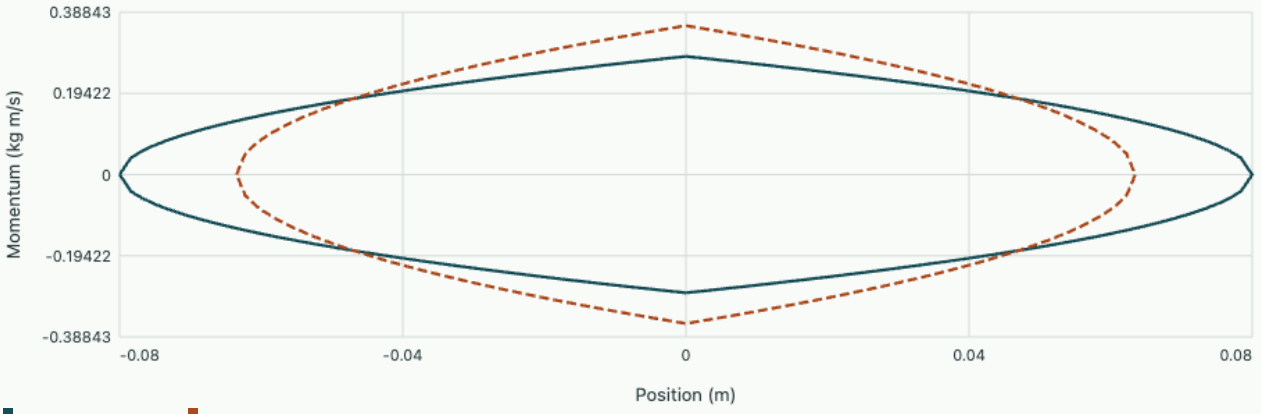
Measure the area of a closed orbit

Inspect how deformation of a potential changes an energy contour. Compute its action before using it to infer an adiabatic frequency.

Open this experiment online

Try the calculation

Final to initial force ratio 2



Initial orbit | Slowly adjusted orbit

Final amplitude 0.0635 m, energy 0.254 J and period 0.3564 s.

Separate a true separatrix crossing from a change of angle convention

Imagine slowly reducing the pendulum's gravitational strength while a bob initially librates. The separatrix action $J_s = 8I\omega_0/\pi$ decreases with ω_0. Away from the crossing the orbit approximately maintains its own action, so its energy as a fraction of the instantaneous barrier rises. When the shrinking separatrix action reaches that of the orbit, no closed librating torus of that action remains. The system can enter one of the two circulation branches. Its action convention changes because the relevant closed curve on the cylinder now winds rather than traversing two momentum branches. An apparent numerical factor-of-two jump caused by this convention is distinct from the actual dynamics through the slow saddle neighbourhood.

For an initial libration action equal to 0.40 of the original separatrix action at fixed I, a parameter protocol reducing ω_0 to half its original value leaves J/J_s about 0.80, assuming adiabatic tracking remains valid. Reducing it to 0.30 of the initial value would formally give $J/J_s = 1.33$, impossible for a libration. The naive crossing estimate is $\omega_0 = 0.40\omega_{0,\mathrm{initial}}$, but the timing and selected rotation direction require integrating the actual trajectory near the saddle; the adiabatic argument loses uniform accuracy there. A local measure of p and unwrapped q supplies the physical distinction.

A real parameter protocol must specify how gravity strength changes. A pendulum length ramp changes both $I = ml^2$ and $\omega_0 = \sqrt{g/l}$, so its separatrix area scales as $I\omega_0 = m\sqrt{g}\,l^{3/2}$. Reducing length to 0.64 of its starting value reduces the separatrix action by $0.64^{3/2} = 0.512$, whereas replacing only ω_0 in a fixed-I model would predict the opposite trend. A literal change of length also brings radial kinetic energy and actuator work, as derived for the reelled pendulum. Specify the physical control path in parameter space before applying a numerical action law.

An unwrapped angle increases by 2π over a recorded interval but the phase portrait seems to contain a bounded loop. Can both observations be trusted simultaneously?

▼ **Work through the solution**

A true circulation winds around the cylindrical phase space, whereas a libration has zero net unwrapped displacement after a full return. A plot that wraps angles into a fixed interval and draws a line across its boundary can display an artificial closing segment. Conversely, measurement noise near $\pm\pi$ can make a naive unwrapping code insert a spurious 2π jump. Inspect signed momentum and the time-resolved crossing of the angle boundary, then reconstruct the loop in adjacent copies of the coordinate interval. Explicitly identifying the cycle on which the action was integrated resolves the apparent contradiction.

Infer work during an adiabatic change from a phase average

Differentiate the action integral at fixed energy with respect to the external parameter λ. The energy relation gives $\partial_\lambda p|_{E,q} = -H_\lambda/H_p$, so $J_\lambda|_E = -(2\pi)^{-1}\oint H_\lambda\,dt$. Combine this with $J_E = T/(2\pi)$ and differentiate at fixed J. The result is

$$\partial_\lambda E|_J = \langle \partial_\lambda H\rangle_{\text{time}}.$$

A slowly changed parameter therefore does net reversible work at the orbit-averaged rate dictated by its conjugate generalised force. The average must weight time, not distance along the plotted phase curve.

For a harmonic spring with stiffness k, $H_k = q^2/2$ and $\langle q^2\rangle = A^2/2 = E/k$, giving $\partial_k E|_J = E/(2k)$. Integrating yields $E \propto \sqrt{k}$, consistent with $J = E/\omega$. A 4% slow stiffness increase raises energy by about 1.98% exactly under this law. Changing stiffness at a single turning point would instead add 4% to the energy for that sudden change. The difference quantifies why observing only identical start and end stiffnesses does not identify the work protocol.

Tutorial on interpreting an action experiment

A harmonic oscillator has energy $0.030\,\mathrm{J}$ and frequency $5.0\,\mathrm{Hz}$. Its stiffness is increased slowly until the frequency doubles. Determine action, final energy and relative amplitude.

▼ **Work through the solution**

The action is

$$0.030/(2\pi\cdot 5.0) = 9.55\times10^{-4}\,\mathrm{J\,s}.$$

Approximately constant action gives a final energy of $0.060\,\mathrm{J}$. With fixed mass and $E = m\omega^2A^2/2$, the final amplitude is $1/\sqrt{2}$ times its initial value. A sudden change performed at a particular phase would in general give a different final action.

A simulation estimates pendulum periods by differentiating sampled action values on a coarse uniform energy grid. Why is the last estimated period below the separatrix likely to be inaccurate even if the

individual areas are well converged?

▼ **Work through the solution**

A finite difference averages $dJ/dE = T/(2\pi)$ over a finite energy interval. The derivative rises logarithmically near the saddle, so the average can underestimate its endpoint value markedly. Refine adaptively in the distance to E_s, use the elliptic quadrature or measure return times directly. The numerical integrator must also locate the long dwell near the unstable equilibrium without stepping across it artificially.

Two trajectories have energy $E_s + \delta$, one rotating clockwise and one anticlockwise. Can energy and the positive magnitude of rotation action distinguish them?

▼ **Work through the solution**

Both have the same speed at each angle and the same positive action magnitude and period. The sign of momentum or a signed orientation of the fundamental rotational cycle distinguishes the branches. Near the separatrix a numerical classifier looking only at energy and unsigned area can lose track of this information. Count the unwrapped angular change and inspect the sign of p to identify circulation direction.

Measure phase area and compare its change under a parameter ramp.

David Tong develops action–angle variables and adiabatic invariance.

What a momentum eigenstate means in a finite experiment

Quantum states, dynamics and measurement

A detector never counts a perfectly sharp momentum eigenket. Derive continuum normalisation from a finite box, change the spectral measure from momentum to energy and predict the full evolution of a normalisable Gaussian packet.

Follow a normalised state as its allowed box grows

Consider a particle moving in one spatial dimension. A periodic box of length L is a mathematical model for an observation made far from any reflecting wall. The inner product is the integral of $\psi^*(x)\chi(x)$ over the box. The momentum operator is $-i\hbar\, d/dx$ with periodic boundary conditions. Integration by parts cancels its boundary form for periodic differentiable functions. Its normalised modes $L^{-1/2}e^{ik_n x}$ have $k_n = 2\pi n/L$ and momenta $p_n = \hbar k_n$. Their overlap is one if the integers coincide and zero otherwise. Increasing L reduces the spacing $\Delta p = 2\pi\hbar/L$ but also reduces the amplitude of each individual box mode at a fixed position. If one simply replaces the Kronecker delta by a Dirac delta while carrying over $L^{-1/2}$, the units and probabilities will fail.

Let c_n be the unit-normalised amplitudes with $\sum_n |c_n|^2 = 1$. For a smooth normalisable packet the probability in an interval dp has a limit $|\phi(p)|^2 dp$. Thus $c_n \simeq \sqrt{\Delta p}\,\phi(p_n)$. Substituting in the synthesis sum and using $L\Delta p = 2\pi\hbar$ produces

$$\psi(x) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} e^{ipx/\hbar}\phi(p)\, dp,$$

$$\phi(p) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} e^{-ipx/\hbar}\psi(x)\, dx.$$

The distributions $\langle x|p\rangle = (2\pi\hbar)^{-1/2}e^{ipx/\hbar}$ have $\langle p|p'\rangle = \delta(p-p')$. The delta acts under integration; $\delta(0)$ has no finite interpretation as the probability for a detector. A physical probability is $\int_{p_1}^{p_2} |\phi(p)|^2 dp$. Changing to wave number gives

$$|\phi_k(k)|^2 = \hbar|\phi(\hbar k)|^2.$$

A graph labelled density in k with a vertical axis copied directly from the momentum plot loses the Jacobian and in general ceases to normalise to one.

Change from momentum bins to energy bins

For a free particle $E = p^2/(2m)$. Each positive energy has two distinct momenta $p_\pm = \pm\sqrt{2mE}$. A detector measuring energy but not direction sums over these mutually orthogonal alternatives. Apply the ordinary change-of-variables rule on each monotone half-line:

$$w(E) = \frac{m}{\sqrt{2mE}}\left[|\phi(+\sqrt{2mE})|^2 + |\phi(-\sqrt{2mE})|^2\right], \qquad E > 0.$$

The factor $m/|p|$ is $1/|dE/dp|$, with units of momentum per energy. Measuring a finite energy interval around zero can have a density diverging as $E^{-1/2}$ if $\phi(0) \neq 0$, yet the integrated probability is finite because $\int_0^\delta E^{-1/2} dE = 2\sqrt{\delta}$. The same effect appears in a box as relatively dense low-energy modes. Standing-wave box boundary conditions change the counting of travelling directions; include degeneracy once and use the appropriate mode spacing, rather than doubling both.

On an infinite line it is useful to introduce $|E, +\rangle = \sqrt{m/|p|}\,|p\rangle$ at $p > 0$, and a corresponding negative branch. In the overlap,

$$\delta(p(E) - p(E')) = \delta(E - E')/|dp/dE|;$$

the two square-root factors give $\delta(E - E')$. The resolution of identity contains $\sum_{\pm} \int_0^\infty dE\, |E, \pm\rangle\langle E, \pm|$. Measuring energy while coherently manipulating the two degenerate directions requires keeping both amplitudes. Adding their probabilities is sufficient only for an instrument whose outcome identifies energy without mixing the directions before detection.

Integrate the free evolution of a Gaussian

Prepare a minimum-uncertainty wave packet centred at x_0 with average momentum p_0 and initial position standard deviation s:

$$\psi(x, 0) = (2\pi s^2)^{-1/4} \exp[-(x - x_0)^2/(4s^2) + i p_0 (x - x_0)/\hbar].$$

Its squared modulus is a Gaussian with variance s^2, which explains the factor four in the amplitude exponent. Complete the square in the Fourier integral to find a momentum probability Gaussian centred at p_0 of variance $\hbar^2/(4s^2)$. The initial state has zero symmetrised position–momentum covariance. Multiplying by a chirp factor $\exp[i\kappa(x - x_0)^2]$ leaves its initial position density unchanged but changes momentum variance and covariance. A snapshot of one density does not determine its later free spread.

Each momentum component evolves by $e^{-ip^2 t/(2m\hbar)}$. Split $p = p_0 + q$. The constant term in p^2 produces a common phase; the linear term shifts the envelope by $p_0 t/m$; the quadratic term changes the width. Carrying out the remaining Gaussian integral yields

$$|\psi(x, t)|^2 = \frac{\exp[-(x - x_0 - p_0 t/m)^2/(2s_t^2)]}{\sqrt{2\pi}\, s_t},$$

$$s_t^2 = s^2 \left[1 + \left(\frac{\hbar t}{2m s^2}\right)^2\right].$$

An independent route uses the Heisenberg operator $x(t) = x(0) + p(0)t/m$. Its variance is $s^2 + (t/m)\langle\{\Delta x, \Delta p\}\rangle + (t/m)^2(\Delta p)^2$. Inserting zero covariance and the Gaussian momentum width reproduces s_t^2. For a chirped packet the covariance term can be negative; momentary narrowing during free evolution is consistent with uncertainty because momentum width and the phase correlation must be included. The density integral stays one, so spreading lowers the height of the peak without removing particles.

Take a model atom of mass 1.44×10^{-25} kg prepared with $s = 1.00\,\mu\mathrm{m}$. Its spread time $t_s = 2ms^2/\hbar = 2.73$ ms. At $t = t_s$ the standard deviation is 1.41 µm. With mean velocity 2.00 mm/s its mean has advanced 5.46 µm. At long flight times $s_t \simeq t\Delta p/m$. Measuring position after a known long flight therefore maps the momentum distribution for this free packet. At $0.5t_s$ the neglected initial position variance is still dominant, so the same inference would substantially overestimate Δp.

A detector window and a Fourier grid do different jobs

Suppose the position-sensitive detector has a Gaussian point-spread function of width d. The measured density is a convolution of the true density with that response if detections are independent and gain is uniform. Gaussian variances then add:

$$s_{\mathrm{obs}}^2 = s_t^2 + d^2.$$

This broadening belongs to the observation model; it does not increase the particle's momentum spread before the measurement. For $d = 1.20\,\mu\mathrm{m}$ at $t = t_s$, the observed standard deviation becomes $1.85\,\mu\mathrm{m}$. A fit to a single image with unknown d cannot disentangle the two contributions. A sequence of flight times constrains the coefficient of t^2 independently of a fixed instrument width.

A discrete Fourier propagation represents a periodic interval sampled at spacing Δx. Its wave-number spacing is $2\pi/L$, and the highest resolved wave number is order $\pi/\Delta x$. Enlarging L at fixed Δx samples the momentum profile more finely and delays periodic wraparound. Reducing Δx at fixed L raises the bandwidth without moving the periodic boundary. Either operation alone can leave the other error intact. A norm-preserving FFT propagator can therefore return a perfectly normalised wave packet that has wrapped through an unphysical boundary or whose carrier momentum has aliased. Compare the phase, group velocity and width against independent formulas, then refine the two scales separately.

Why boundary conditions are part of a momentum observable

The expression $p = -i\hbar\, d/dx$ does not by itself specify a complete operator. On a finite interval $[0, L]$, integrate two sufficiently smooth square-integrable functions by parts:

$$\langle f, pg\rangle - \langle pf, g\rangle = -i\hbar[f^*(x)g(x)]_0^L.$$

A symmetric domain makes this boundary form vanish for every pair in the domain. A self-adjoint operator additionally has the same domain as its adjoint. This distinction matters for a physical generator of unitary translations and for assigning sharp spectral outcomes. Imposing a real or imaginary matrix representation in a finite basis can conceal boundary conditions that become essential as the basis is enlarged.

For a ring allow quasi-periodicity $g(L) = e^{i\theta}g(0)$ with fixed θ common to all functions in the domain and square-integrable first derivative. Then the boundary form vanishes. Conversely, for an adjoint candidate f the requirement that it vanish for all possible $g(0)$ forces $f(L) = e^{i\theta}f(0)$, so the same condition reappears. Solving $pg = pg$ yields the momenta $p_n = \hbar(2\pi n + \theta)/L$. Thus boxes with precisely the same local differential expression have momentum spectra shifted by an externally fixed phase. In a physically realised ring such a phase can encode a magnetic flux through a gauge-transformed boundary

condition; a gauge consistent with that interpretation must also be used in the Hamiltonian and physical momentum.

Compare with forcing $g(0) = g(L) = 0$ as in a hard-wall well. The boundary form vanishes for all f, g satisfying both zero conditions, so the derivative is symmetric there. But the adjoint boundary form vanishes for arbitrary endpoint values of f when g vanishes at both ends. Consequently this two-zero derivative domain is smaller than its adjoint domain. There is no ordinary rigid translation that carries every state of a hard-wall interval back into that same interval without assigning what happens at its boundary. Meanwhile the kinetic energy $-\hbar^2 d^2/(2m\,dx^2)$ with suitable Dirichlet domain is self-adjoint and has a well-defined discrete spectrum. The symbol $p^2/(2m)$ in a Dirichlet well does not turn its first derivative with the same two endpoint conditions into a self-adjoint sharp momentum observable.

A release-and-detect experiment makes this distinction concrete. Prepare the first Dirichlet state $\sqrt{2/L}\sin(\pi x/L)$ supported in $0 < x < L$, then remove the walls suddenly and measure momentum on the full line. Its momentum wavefunction is the Fourier transform of the spatial function extended by zero outside. Multiplying the exponential and sine and integrating gives a continuous amplitude proportional to $[1 + e^{-ipL/\hbar}]/[(\pi/L)^2 - (p/\hbar)^2]$. The apparent singularities at $p/\hbar = \pm\pi/L$ are removable because the numerator also vanishes. It is a mistake to predict two delta peaks by replacing the standing wave with the sum of two infinite plane waves. The finite support multiplies that sum by a spatial window and broadens both momentum components.

The mean momentum for this real function is zero. Its second moment after release is $\hbar^2 \int |\psi'|^2 dx = \pi^2\hbar^2/L^2$, consistent with the original kinetic energy. Thus a continuous released momentum distribution can have an rms momentum whose square exactly recovers a discrete energy of the original bounded system. This is a useful quantitative example where the equality of moments across two measurement procedures does not imply equality of their full spectral probability distributions. Adding a short laboratory impulse to every particle after release should translate the whole distribution by the corresponding momentum kick, preserving its width. That response provides an independent check on a momentum detector's sign and scale.

Derive the free propagator with a convergent limiting prescription

Partitioning the free unitary into momentum eigenstates produces its position kernel

$$K(x, t; x', 0) = \frac{1}{2\pi\hbar} \int_{-\infty}^{\infty} \exp\left[\frac{i}{\hbar}\left(p(x - x') - \frac{p^2 t}{2m}\right)\right] dp.$$

It is oscillatory rather than absolutely integrable. Define it by first damping $e^{-\varepsilon p^2}$, evaluating a convergent Gaussian and then taking $\varepsilon \downarrow 0$ on smooth normalisable test packets. Completing the square in the exponent gives, for $t > 0$,

$$K = \sqrt{\frac{m}{2\pi i\hbar t}} \exp\left[\frac{im(x - x')^2}{2\hbar t}\right].$$

The phase convention of the square root for positive time follows continuously from the regularised Gaussian. The units are inverse length so $\psi(x, t) = \int K(x, t; x', 0)\psi(x', 0)\,dx'$ has the same units as the

initial wavefunction. As $t \to 0^{+}$, the kernel does not approach a finite spike pointwise. Its increasingly fast oscillations across x' cancel except near $x' = x$, recovering the initial state in the appropriate strong or distributional sense. For a constant initial amplitude one may recover unity formally, but such an amplitude on the full line is itself not a normalised test state.

Two checks involve no wave packet fitting. Inserting two kernels for t_1, t_2 and integrating over an intermediate position gives the kernel for $t_1 + t_2$, either by another Gaussian integration or, more directly, by composing the momentum phase factors. Complex conjugation and reversing the time arguments yield the inverse evolution. Neither statement holds for an arbitrary choice of magnitude prefactor assigned merely to obtain the correct phase. In Fourier numerics the corresponding checks are unit norm after repeated evolution and recovery of the original sampled complex state after a forward followed by backward propagation. These are necessary algorithm checks, even though a periodic or aliased computational model can still satisfy them exactly.

This kernel also clarifies the long-time momentum mapping. Factor $e^{imx^{2}/(2\hbar t)}$ from K and identify the Fourier transform of $e^{imx'^{2}/(2\hbar t)}\psi(x', 0)$ evaluated at $p = mx/t$. If the occupied initial region is narrow enough that $mx'^{2}/(2\hbar t)$ varies little across it about the selected origin, the extra phase can be neglected and

$$|\psi(x, t)|^{2} \simeq \frac{m}{t}\left|\phi\Big(\frac{mx}{t}\Big)\right|^{2}.$$

The factor m/t is the Jacobian dp/dx. When the initial packet is displaced, use a suitable shifted origin or account for the induced linear phase explicitly. This derivation applies well beyond Gaussians and gives a scale for testing the approximation before interpreting a flight image as a momentum histogram. The quadratic residual phase provides the matter-wave analogue of the near-field correction in diffraction.

Distinguish a detector crossing from a snapshot in space

A detector counts particles passing through an aperture over time. Its signal is related to flux, whereas a camera image of identical preparations at one time probes a position distribution. For a one-dimensional free wavefunction the continuity equation is $\partial_t |\psi|^2 + \partial_x j = 0$ with $j = (\hbar/m)\,\mathrm{Im}(\psi^{*}\partial_x \psi)$. Integrating over the region $x < X$ gives the change in probability to the left of a fixed plane. If there is no flux at negative infinity, the net integrated current through X equals the reduction in that probability. A positive-reading counter with absorption is a more detailed instrument and cannot in general be identified with a signed current in the presence of reflected waves or interference.

For a minimum-uncertainty Gaussian initially at x_0, the exact phase in the freely expanding envelope has a quadratic term. Differentiating the phase gives the local velocity entering current,

$$\frac{j(x, t)}{|\psi(x, t)|^{2}} = \frac{p_0}{m} + \frac{t}{t_s^{2} + t^{2}}\left(x - x_0 - \frac{p_0 t}{m}\right).$$

At the centre this equals the group velocity p_0/m, but ahead of the centre it is larger and behind it smaller. A single constant velocity multiplying density is inadequate for a significantly spreading packet. For the earlier atom at $t = t_s$, one current-standard-deviation distance ahead corresponds to $1.41\ \mu\mathrm{m}$

and the additional speed is about 0.259 mm/s. The trailing side gives the opposite correction. Direct differentiation of its Gaussian density verifies that this current satisfies continuity. A code reproducing s_t but evolving the wrong complex phase may fail this current test; judging an entire propagation by the changing width alone would miss the error.

A useful inverse question is whether two time-of-flight images suffice to recover an unknown complex wavefunction. In the long-time limit the image yields $|\phi(p)|^2$, which contains no direct momentum-dependent phase. Translation of the initial wave packet changes $\phi(p)$ by a linear phase, leaving the momentum density unchanged. Complex conjugate or chirped preparations can also share particular intensity records. Additional calibrated evolution, known reference interference or structural assumptions can provide phase information. Even an exact free propagator only defines the forward map; its invertibility on complex amplitudes does not imply intensity-only data identify every possible input. The experiment must specify which class of preparations and which measurement settings render the desired parameters estimable.

Track a continuous superposition through focus

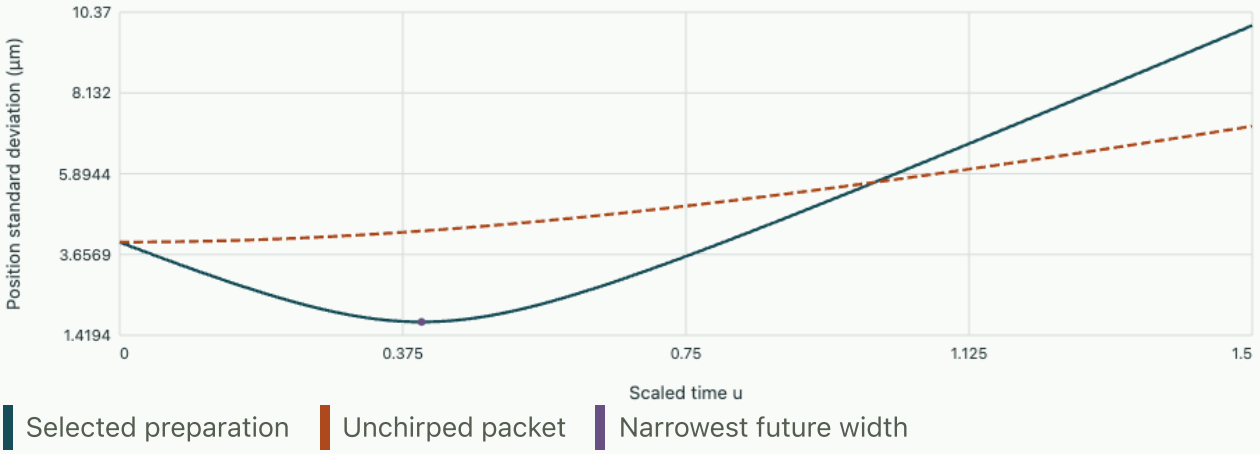
Change a position-momentum correlation and compare the full packet width with an initially unchirped packet. A narrow width at one time does not determine its subsequent spread.

Open this experiment online

Try the calculation

Initial phase curvature b

-2



For u ≥ 0 the minimum width is 1.789 µm at u = 0.4.

Plan an FFT comparison using physical position and momentum resolution

A grid propagator can track the same packet analytically solved earlier, then carry over its verified conventions to less elementary potentials. A common trap is to regard the number of points N as a single accuracy parameter. Define a periodic interval of physical length $L = N\Delta x$. Discrete Fourier modes have $k_j = 2\pi j/L$, with negative wave numbers represented by negative indices after reordering. Their spacing is $\Delta k = 2\pi/L$; the largest distinct magnitude is close to $\pi/\Delta x$. A narrow structure in momentum requires large L. A high carrier momentum requires small Δx. These conditions can fail independently even when the norm stays exactly one.

If $\psi_l = \psi(x_l)$ is a sampled wavefunction in inverse square-root metres and the forward transform sums without $1/N$, the approximated continuum k-amplitude is

$$\phi_k(k_j) \simeq \Delta x \sum_l \psi_l e^{-ik_j x_l}/\sqrt{2\pi}.$$

The phase includes the physical coordinate origin adopted for the grid. Parseval then gives

$$\Delta k \sum_j |\phi_k(k_j)|^2 = \Delta x \sum_l |\psi_l|^2.$$

A library returning a unitary discrete transform instead assigns a factor $1/\sqrt{N}$ to both directions; multiplying its output by $\sqrt{\Delta x/\Delta k}$ converts sample units. Plotting a library output directly against physical momentum without these scale factors is especially misleading when comparing two meshes. Total discrete ℓ^2 norm and continuum probability density are different normalisations of the same samples.

For the atom of mass 1.44×10^{-25} kg, $s = 1.00$ µm and $v_0 = 2.00$ mm/s,

$$k_0 = mv_0/\hbar = 2.73\,\mu\mathrm{m}^{-1},$$

$$\Delta k_{\mathrm{rms}} = 1/(2s) = 0.50\,\mu\mathrm{m}^{-1}.$$

A box 80 µm wide gives spacing $0.0785\,\mu\mathrm{m}^{-1}$, about 6.4 modes per rms momentum width. With $N = 256$, grid spacing is 0.3125 µm and Nyquist wavenumber is $10.05\,\mu\mathrm{m}^{-1}$, exceeding $k_0 + 5\Delta k_{\mathrm{rms}} = 5.23\,\mu\mathrm{m}^{-1}$. This comfortably samples the initial carrier distribution. It is still necessary to check that the envelope remains inside the periodic box at the desired final time.

At $t = 10$ ms the packet centre has moved 20 µm and $s_t \simeq 3.80$ µm. Starting at the centre of an interval spanning -40 to 40 µm, its five-standard-deviation front approaches 39 µm and begins testing the boundary margin. At 20 ms the centre alone reaches the periodic boundary. A perfectly unitary calculation will show it reappearing on the opposite side. This is correct for a ring and incorrect for a free-space flight image. Enlarging N while keeping the same L improves the Nyquist limit but does nothing to postpone this crossing. An open-boundary absorber can delay reflected artefacts but attenuates norm; a lost norm in that model should be compared with counted escape rather than naively enforced to zero.

Infer displacement and speed from a wrapped packet without mistaking periodicity for reflection

A code on a periodic box of length 80 µm reports a sharp packet peak at $+35$ µm in one frame and -36 µm in the next frame 4.50 ms later. Find a consistent uniform drift and identify the ambiguity if no speed prior is available.

▼ **Derive and check the result**

The smallest forward displacement is $+9$ µm after adding one full box length to the second position. It corresponds to 2.00 mm/s and agrees with the earlier preparation. Direct subtraction without unwrapping would give -71 µm, or -15.8 mm/s, an apparent reversal. All displacements $9 + 80n$ µm for integer n yield the same wrapped pair, so the frames alone cannot identify the number of turns. A momentum measurement or a bound on speed relative to frame spacing resolves the branch. For an extended packet crossing a seam, use a periodic circular moment such as $\langle e^{2\pi i x/L} \rangle$ so an ordinary spatial mean does not spuriously jump toward the box centre. If that circular moment is nearly zero, even its angular centre is ill-defined and a full density comparison is preferable.

Tutorial · design a momentum measurement

Distinguish an unresolved line from a finite-width packet

A periodic calculation of length $L = 100\,\mu\mathrm{m}$ contains mode n. A detector bins momentum in intervals $1.0 \times 10^{-29}\,\mathrm{kg\,m/s}$. How many box-mode spacings fit into a bin? What changes if the box length doubles at fixed physical packet?

▼ **Work through the solution**

Here $\Delta p = h/L = 6.626 \times 10^{-30}\,\mathrm{kg\,m/s}$, so the detector bin spans about 1.51 spacings. The precise number of grid points falling within a bin depends on its endpoints. Doubling L gives roughly 3.02 spacings per bin. Each corresponding coefficient squared decreases approximately in proportion to Δp, while the number included increases. The total probability should approach a stable integral on a progressively finer grid. If a graph compares $|c_n|^2$ across the boxes, divide by the corresponding Δp before calling the ordinate a continuum density.

Calculate a threshold energy interval without evaluating a delta at zero

A packet's momentum density is nearly constant w_0 over $-p_c < p < p_c$. Obtain the probability for $0 < E < E_c = p_c^2/(2m)$ by two different integrations. What changes when only positive-momentum particles reach the counter?

▼ **Work through the solution**

Direct integration in momentum gives $2p_c w_0$. In energy the two branches contribute $2mw_0/\sqrt{2mE}$. Its integral gives $4mw_0\sqrt{E_c}/\sqrt{2m} = 2p_c w_0$ as required. A positive-direction selector halves this particular probability when applied to the original ensemble. If the entire subsequent experiment is renormalised to the selected particles, divide again by the total positive-momentum probability. Those two reported probabilities correspond to different conditioning and should be labelled explicitly.

Infer the packet width from two images

The model atom has an initial measured width 1.56 µm and a width 1.85 µm after 2.73 ms. Assume a time-independent Gaussian instrument response and a freely evolving unchirped minimum-uncertainty state. Infer s and d, and identify an additional check.

▼ **Work through the solution**

Subtract squared widths to remove the detector:

$$s_{\mathrm{obs}}(t)^2 - s_{\mathrm{obs}}(0)^2 = \hbar^2 t^2/(4m^2 s^2).$$

The observed difference is about $0.99\,\mu\mathrm{m}^2$, giving $s \simeq 1.0\,\mu\mathrm{m}$ and $d = \sqrt{s_{\mathrm{obs}}(0)^2 - s^2} \simeq 1.2\,\mu\mathrm{m}$, within rounding. A third flight time tests the predicted quadratic in t. A term linear in time would point to a nonzero initial correlation or to measurement drift. Independent calibration of the point-spread function distinguishes those interpretations from a nonminimum momentum width.

Further reading

- Fließbach · Eigenwerte und Eigenfunktionen
- TUM · Theoretische Physik 3, representations and momentum states
- Course investigation · probability density and current

How a turning point supplies the missing semiclassical phase

Quantum states, dynamics and measurement

Far from a turning point the wavelength follows the classical momentum. At the turning point that approximation becomes singular. Derive the Airy connection, quantise a smooth well, distinguish a hard wall and estimate a tunnelling exponent with a defensible error check.

Expand the action before interpreting a local wavelength

For a stationary one-dimensional state with smooth potential $V(x)$, write $\psi = e^{iS(x)/\hbar}$. Substitution into Schrödinger's equation gives

$$S'^2 - i\hbar S'' = 2m(E - V) = p^2(x).$$

Expand $S = S_0 + (\hbar/i)S_1 + \cdots$. The leading terms give $S_0' = \pm p$ and $S_1' = -p'/(2p)$ in an allowed region where $p > 0$. Thus travelling waves have form $Cp^{-1/2}\exp[\pm i\int^x p\,dx/\hbar]$. The amplitude factor is physically essential: for a single such branch the current to leading order is $\pm|C|^2/m$, independent of x, as required for steady flux in a real potential. A constant amplitude with variable wavenumber would predict current piling up in a stationary region.

A local condition for this expansion is $\hbar|p'|/p^2 \ll 1$ together with control over higher derivatives on the wavelength scale. Away from a barrier the phase can remain accurate even if absolute amplitude normalisation must be fixed globally. Where $E < V$, define $\kappa = \sqrt{2m(V - E)}$ with units of momentum. The local solutions become $\kappa^{-1/2}\exp[\pm\int \kappa\,dx/\hbar]$. A single real decaying branch carries zero probability current by itself. Stationary tunnelling flux through a finite barrier involves a complex combination satisfying both exterior matching conditions, even though the leading size of transmission can often be inferred from the decaying exponential.

Resolve the apparent divergence near a simple turning point

Let x_t satisfy

$$V(x_t) = E,$$

with slope $F = V'(x_t) > 0$. Linearise $V - E \simeq F(x - x_t)$. The local Schrödinger equation becomes

$$\psi'' = (2mF/\hbar^2)(x - x_t)\psi.$$

Introduce length $\ell = (\hbar^2/(2mF))^{1/3}$ and $\xi = (x - x_t)/\ell$, leaving $d^2\psi/d\xi^2 = \xi\psi$. Its two independent solutions are Ai and Bi. On the positive side $\mathrm{Ai}(\xi)$ decays, whereas $\mathrm{Bi}(\xi)$ grows. A bound state extending into a semi-infinite forbidden region selects the former. At $\xi = 0$ the Airy function is finite; the apparent divergence $p^{-1/2}$ came from extending the outer approximation beyond its range.

The asymptotic forms for $s > 0$ large are

$$\mathrm{Ai}(s) \sim \frac{e^{-2s^{3/2}/3}}{2\sqrt{\pi}\, s^{1/4}}, \qquad \mathrm{Ai}(-s) \sim \frac{\sin(2s^{3/2}/3 + \pi/4)}{\sqrt{\pi}\, s^{1/4}}.$$

For a linear slope, $2s^{3/2}/3$ equals $|\int_{x_t}^{x} p\, dx|/\hbar$ on the allowed side or $\int_{x_t}^{x} \kappa\, dx/\hbar$ on the forbidden side. Matching in an interval where $|x - x_t| \gg \ell$ but the potential remains effectively linear gives the connection from a decaying tail of coefficient C to $2C\, p^{-1/2} \sin[\hbar^{-1} \int_{x}^{x_t} p\, dx + \pi/4]$. The coefficient two and phase $\pi/4$ both come from the same local solution. A turning point with zero slope, a sharp step or two turning points closer than their Airy lengths requires a different local treatment.

Join two decaying tails to a standing wave

For a smooth well with two well-separated simple turning points x_1, x_2, the left tail matches to $\sin[\hbar^{-1} \int_{x_1}^{x} p\, dx + \pi/4]$. The right tail matches to $\sin[\hbar^{-1} \int_{x}^{x_2} p\, dx + \pi/4]$. Their zeros and logarithmic derivatives must coincide throughout the overlap, up to an overall sign. The sum of the two phase arguments is the full one-way action divided by $\hbar$ plus $\pi/2$, which must be an integer multiple of π. Taking $n \geq 0$ gives

$$\int_{x_1}^{x_2} p(x, E_n)\, dx = \pi\hbar(n + 1/2).$$

For a harmonic oscillator the turning point magnitude is $a = \sqrt{2E/(m\omega^2)}$ and $p = m\omega\sqrt{a^2 - x^2}$. The action is the area of a semicircle times $m\omega$, equal to $\pi E/\omega$. The rule reproduces $E_n = \hbar\omega(n + 1/2)$ exactly for this special spectrum, even though its outer WKB wavefunction is still inaccurate near each turning point and globally for low quantum number. Equality of a spectral result does not imply equality of spatial densities or agreement of all observables.

Different boundaries change the phase budget. Take a hard wall at $x = 0$ and a linearly rising potential Fx for $x > 0$. The solution decaying to the right is $\mathrm{Ai}[(x - E/F)/\ell]$. Its exact energies follow from $\mathrm{Ai}(-E/F\ell) = 0$. The wall imposes a node without an Airy phase at the left. Setting the right-matched standing wave to zero at the wall gives

$$\int_{0}^{E/F} p\, dx = \pi\hbar(n + 3/4).$$

Thus $E/(F\ell) \simeq [3\pi(n + 3/4)/2]^{2/3}$. For $n = 0$ this is about 2.320 versus the first exact zero 2.338; with $n = 4$ the approximation is about 7.94. Replacing the wall by a second smooth turning point would instead introduce an erroneous half offset.

Separate a credible exponential estimate from its prefactor

For a barrier with two simple turning points x_a, x_b, a large forbidden action $I = \int_{x_a}^{x_b} \kappa\, dx/\hbar$ suppresses the transmitted flux approximately as $T \simeq e^{-2I}$, up to prefactors and corrections that depend on geometry and matching. For a triangular forbidden region with height $V_0 - E = \Delta$ at $x = 0$ falling as

$\Delta - Fx$ to zero at Δ/F, its action is $I = 2\sqrt{2m}\Delta^{3/2}/(3\hbar F)$. A discontinuous incident-side boundary may affect the prefactor substantially. It does not justify assigning a second smooth-turning-point phase to the discontinuity. For a rectangular barrier the exact thick-barrier prefactor is $16E(V_0 - E)/V_0^2$ times the decaying exponential at equal exterior potentials. It can exceed one while total transmission remains small.

Suppose two measurements change only the forbidden width of a flat barrier from d to $d + \Delta d$, keeping both interfaces and asymptotic energies the same. In the thick limit the ratio $T(d + \Delta d)/T(d) \simeq e^{-2\kappa \Delta d/\hbar}$ cancels the common prefactor. If $\kappa/\hbar = 2.0\,\mathrm{nm}^{-1}$ and $\Delta d = 0.50$ nm, this ratio is

$$e^{-2} = 0.1353.$$

An offset error in barrier width that leaves this ratio unchanged can still cause a large error in absolute transmission. Near the top two turning points can merge, I becomes small, and the raw exponential lacks a unitary connection to the over-barrier reflection regime. Use a suitable uniform approximation or a direct scattering solution there.

Predict a high-energy spectrum without a closed-form wavefunction

Consider a pure quartic trap $V(x) = \lambda x^4$ with $\lambda > 0$. Classical turning points lie at $x = \pm a,\ a = (E/\lambda)^{1/4}$. Unlike the harmonic oscillator its period changes with energy, so its levels cannot be equally spaced. The semiclassical action can be evaluated once in dimensionless coordinates and then rescaled to any m, λ, E:

$$A(E) = \int_{-a}^{a} \sqrt{2m(E - \lambda x^4)}\, dx = 2\sqrt{2m}\, \lambda^{-1/4} E^{3/4} I_4,$$

$$I_4 = \int_0^1 \sqrt{1 - y^4}\, dy.$$

With $s = y^4$,

$$I_4 = \frac{1}{4} B(1/4, 3/2) \simeq 0.8740.$$

Quantising $A(E_n) = \pi\hbar(n + 1/2)$ predicts $E_n \simeq [\pi\hbar(n + 1/2)\lambda^{1/4}/(2\sqrt{2m} I_4)]^{4/3}$. The dependence $E \propto n^{4/3}$ and the dimensional factor $\hbar^{4/3}\lambda^{1/3}m^{-2/3}$ can be checked independently from scaling the full Schrödinger equation. That dimensional argument does not supply the coefficient or the half-integer offset. The action and boundary phases supply the additional physics.

Choose units such that $H = -d^2/dx^2 + x^4$. Then $2m = \hbar = 1$ in the scaled equation, and the leading energies are $[\pi(n + 1/2)/(2I_4)]^{4/3}$. For $n = 0$ this gives about 0.868, below the accurately computed ground value around 1.060. An independent variational Gaussian $\psi = (\beta/\pi)^{1/4} e^{-\beta x^2/2}$ has energy $\beta/2 + 3/(4\beta^2)$, minimised at $\beta^3 = 3$ and giving 1.082. Its proximity to the true ground energy has a different status: the Gaussian energy is an upper bound by the variational principle, while leading WKB

provides no universal bound for this potential. Using agreement between two approximate numbers as proof of a rigorous interval would require actual lower and upper estimates.

At larger n the period relation provides an additional prediction. Since $A \propto E^{3/4}$,

$$T_{\mathrm{cl}} = 2\,dA/dE \propto E^{-1/4};$$

the frequency increases at larger amplitude, and the local spacing should grow as $E^{1/4}$. Numerically compute a handful of high levels with a direct finite-difference or spectral Hamiltonian, then plot $E_n^{3/4}$ against $n + 1/2$. A nearly linear trend is a more sensitive diagnostic of the predicted law than comparing a single apparent power from a log plot at low n. Increase domain size and basis cutoff independently. A truncation that confines high-energy tails can falsely steepen the spacing, while a coarse grid can flatten the kinetic-energy dispersion. Both errors become more visible for rapidly oscillating states.

Relate a barrier action to a coherent population oscillation

Take two identical minima separated by a smooth barrier. At energies below the barrier top but above each isolated-well ground, there are two classically allowed intervals and a central forbidden region. If the barrier is sufficiently opaque, construct approximate localised states $|L\rangle, |R\rangle$ with equal one-well energy E_0. In their span the simplest symmetric Hamiltonian is $H = E_0 I - J(|L\rangle\langle R| + |R\rangle\langle L|)$ with $J > 0$ in a suitable phase convention. The even and odd combinations have energies $E_0 - J, E_0 + J$. The tunnel splitting is $2J$, set to leading exponential order by the overlap of amplitudes through the central forbidden interval.

A single under-barrier wavefunction amplitude carries e^{-I},

$$I = \int \kappa dx/\hbar.$$

For an open barrier the transmission probability is proportional to e^{-2I}. For a symmetric double-well splitting the relevant quantity is a coherent off-diagonal energy amplitude, so its leading exponential is $J \propto e^{-I}$, times a prefactor involving the oscillation scales and matching in the two wells. Confusing these exponents can lead to a severe error when using a calculated spatial barrier to predict a coherent tunnel frequency. The exact prefactor depends on the potential and turning-point normalisations. A direct symmetric numerical eigenproblem giving the difference of two nearly equal energies requires enough precision to resolve an exponentially small splitting.

Prepare $|L\rangle$ at time zero. Its evolution under the two-state Hamiltonian gives

$$P_R(t) = \sin^2(Jt/\hbar);$$

the first full transfer time is $\pi\hbar/(2J)$. If the observed energy splitting expressed as ordinary frequency is $2J/h = 20$ Hz, the first transfer occurs at 25 ms. Changing only the barrier action by $+0.70$ at fixed prefactor multiplies the splitting by $e^{-0.70} = 0.497$ and approximately doubles the transfer time. An open-barrier transmission experiment over the same change instead sees factor $e^{-1.4} = 0.247$. Both tests depend exponentially on barrier shape but interrogate different quantities.

Now add a calibrated tilt $\epsilon = E_L - E_R$. The 2×2 Hamiltonian after trace subtraction is $(\epsilon/2)\sigma_z - J\sigma_x$, and the maximum transfer falls to $4J^2/(\epsilon^2 + 4J^2)$. At $\epsilon = 2J$ it is one half while the oscillation angular frequency increases from $2J/\hbar$ to $2\sqrt{2}J/\hbar$. A loss of transfer contrast need not be environmental decoherence; static tilt alone yields this combination of increased frequency and reduced amplitude. Measuring both quantities and reversing the tilt tests the two-level model. If a higher intrawell level lies within the tilt or excitation range, a single tunnelling doublet no longer spans the relevant motion and a larger basis is required.

Understand the angular correction at a radial turning point

In three dimensions the reduced radial equation on $r > 0$ has a centrifugal term $\hbar^2 l(l+1)/(2mr^2)$. Applying the flat-coordinate WKB rule directly to this equation can give incorrect phases at its singular endpoint, especially for low l. A way to see the appropriate angular correction without guessing is to set $r = r_0 e^x$, mapping the half-line to all real x, and write $u(r) = e^{x/2} w(x)$. The arbitrary reference r_0 keeps the logarithm dimensionless. Since $d/dr = r^{-1} d/dx$, differentiating twice yields

$$\frac{d^2u}{dr^2} = \frac{e^{x/2}}{r^2}\left(w'' - \frac{1}{4}w\right).$$

The resulting equation for w has squared local action momentum in x proportional to $2mr^2(E - V) - \hbar^2(l + 1/2)^2$. Thus returning its action $\int p_x dx$ to r uses $l(l+1) \to (l+1/2)^2$ in the radial semiclassical momentum. This is the Langer correction. It does not change the exact angular eigenvalue $L^2 = \hbar^2 l(l+1)$; it accounts for the coordinate transformation and endpoint behaviour of the semiclassical representation. Confusing those statements would alter the exact radial equation to repair an approximate integration rule.

An instructive check is the isotropic oscillator. For $V = m\omega^2 r^2/2$ and effective classical angular momentum $L = (l + 1/2)\hbar$, the radial action over its two roots is

$$A_r = (\pi/2)(E/\omega - L).$$

One can obtain this without an explicit contour integral by noting that radial motion repeats at twice the Cartesian frequency, so

$$dA_r/dE = \pi/(2\omega).$$

Fix the integration constant at the circular-orbit threshold $E = \omega L$, where the two radial turning points meet and $A_r = 0$. Quantising $A_r = \pi\hbar(n_r + 1/2)$ gives $E = \hbar\omega(2n_r + l + 3/2)$, matching the exact three-dimensional oscillator spectrum. For $l = 0$ the naive uncorrected centrifugal term has no inner turning point; transferring a generic two-smooth-turning-point formula into that half-line problem without reconsidering its origin is especially unsafe.

Calculate an exponentially small transmission without subtracting

nearly equal numbers

In a classically forbidden interval the local solution contains a growing and a decaying branch. In a long interval numerical transfer matrices can contain elements as large as e^{I}, while the quantity sought is as small as e^{-2I}. Straight propagation of a decaying boundary condition in the unstable direction can contaminate it with a tiny growing roundoff component which eventually dominates. A plot matching the expected slope over the first few decay lengths gives no guarantee that a final tiny flux is accurate. Numerically stable alternatives include propagating a logarithmic derivative with controlled handling of its poles, matching independently normalised left and right solutions, or composing scattering matrices for short slices while retaining decaying propagation factors.

For a square barrier the exact transmission at $E < V_0$ in equal exterior regions is $T = [1 + V_0^2 \sinh^2(\kappa d/\hbar)/(4E(V_0 - E))]^{-1}$. With $E = V_0/2$ this simplifies to $\operatorname{sech}^2 I$. At $I = 20$ its value is approximately $4e^{-40} = 1.70 \times 10^{-17}$. Computing $T = 1 - R$ in ordinary double precision can give exactly zero since the reflection probability rounds to one. Compute transmitted amplitude or its logarithm directly and compare $\log T \simeq \log 4 - 2I$. At a moderate I, where both currents are accurately representable, independently verify $R + T = 1$. Moving the benchmarking point across the barrier-opacity range tests both the physical approximation and the floating-point representation of the desired observable.

Resolve a wave through a classical turning point

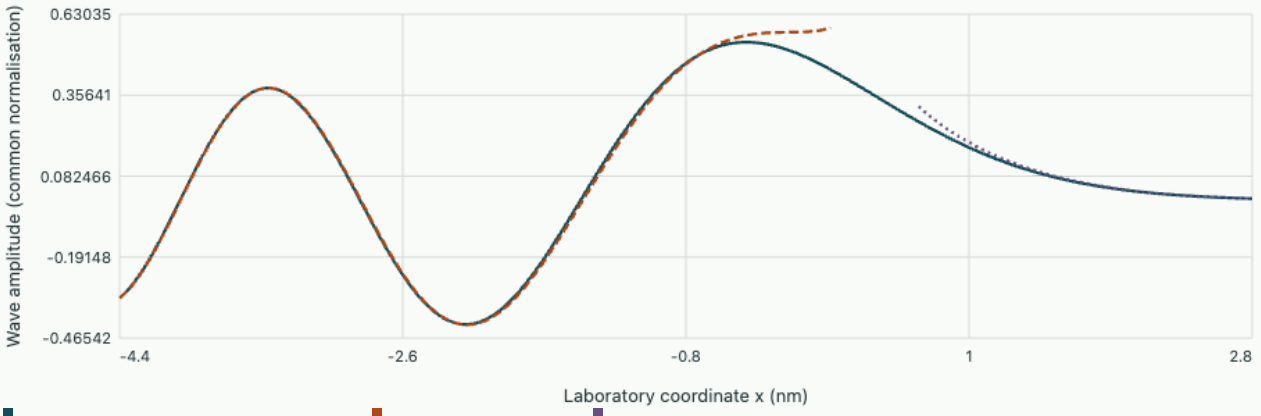
Vary the physical length scale in a locally linear potential. Compare the numerical Airy solution and the two semiclassical branches near and away from the turn.

Open this experiment online

Try the calculation

Turning-point length (nm) **0.8**

Turning-point position (nm) **0.4**



Numerically integrated Ai | Allowed WKB | Forbidden WKB

The turn at 0.4 nm has finite Ai value 0.35503. The characteristic transition length is 0.8 nm; the asymptotic curves omit the nearest 0.35 lengths.

Connect semiclassical action to counting states in phase space

The action quantisation rule can also be read as a statement about the number of energy levels below a chosen energy. In a one-dimensional smooth confining well the classically accessible region in (x, p) has area $\Omega(E) = \int dx \int_{-p(x)}^{p(x)} dp = 2A(E)$. Roughly one stationary state fits per phase-space area $2\pi\hbar$, with an endpoint correction from the turning-point phases. The smoothed density of states follows by differentiating $\bar{N}(E) \simeq \Omega(E)/(2\pi\hbar)$, giving $\bar{\rho}(E) = T_{\mathrm{cl}}/(2\pi\hbar)$. This counts levels after smoothing over individual spacing; the exact count is a staircase and its derivative a sum of delta functions. A formula for a smooth density cannot predict each state in a near-degenerate doublet without additional phase and tunnelling information.

For a particle in a two-dimensional rigid rectangular region, apply the same counting idea in four-dimensional phase space. With area $\mathcal{A} = L_x L_y$, accessible momentum disk $p_x^2 + p_y^2 < 2mE$, the scalar smoothed count is $\bar{N}(E) \simeq \mathcal{A}\pi(2mE)/(2\pi\hbar)^2 = \mathcal{A}mE/(2\pi\hbar^2)$. Its derivative is a constant, unlike the $E^{-1/2}$ free one-dimensional density and the $\sqrt{E}$ three-dimensional density. This leading area law can be checked by directly counting positive integer lattice points satisfying $(n_x\pi/L_x)^2 + (n_y\pi/L_y)^2 < k^2$ in a Dirichlet rectangle. It works after smoothing when many modes fit in the domain and makes no assumption that the rectangular shape itself models all possible billiards.

A more sensitive check accounts for the excluded axes in that first-quadrant lattice count. For a Dirichlet rectangle subtract half of the positive-axis contributions at each boundary of the quadrant. In terms of perimeter $\mathcal{P} = 2(L_x + L_y)$, the resulting first two terms are $\bar{N}(k) \simeq \mathcal{A}k^2/(4\pi) - \mathcal{P}k/(4\pi) + O(1)$ after appropriate smoothing, with fluctuations from discrete modes beyond this local expansion. Neumann boundaries reverse the sign of the perimeter term because modes constant along one coordinate are

admitted. Area alone determines the bulk growth, whereas boundary condition affects the next correction. The finite corner term and spectral fluctuations require more care than assigning one universal integer offset to every billiard.

The distinction between smoothing and exact prediction also explains why counting every computed eigenvalue only up to an arbitrarily sharp threshold can produce apparent convergence jumps as the mesh improves. If a true eigenvalue lies near the threshold, a small numerical shift changes an integer count by one. Compare a range of cumulative counts, track nearby eigenvalue residuals and study the smoothed density with a bandwidth large enough to include several levels but small compared with the energy scale of interest. Repeating for two geometries of equal area and different perimeter tests actual geometric information beyond an overall energy scale.

Measure the boundary correction in a square cavity

A scalar mode calculation uses a square of side 10.0 length units and reports all Dirichlet modes with $k < 5.00$ inverse length units. Find the leading area estimate, add the perimeter correction, and compare with a direct lattice-point count.

▼ **Derive and check the result**

The area contribution is $100(25)/(4\pi) = 198.94$; the perimeter subtraction is $40(5)/(4\pi) = 15.92$, giving a smoothed estimate 183.03 before corner and oscillatory terms. The exact count uses positive integers n_x, n_y satisfying $n_x^2 + n_y^2 < (50/\pi)^2 \simeq 253.30$. For each $n_x = 1, \dots, 15$, count $n_y \le \lfloor \sqrt{253.30 - n_x^2} \rfloor$ and sum. The values are $15, 15, 15, 15, 15, 14, 14, 13, 13, 12, 11, 10, 9, 7, 5$, giving 183. Agreement at this threshold is close, while the leading area term alone is about sixteen high. Nearby thresholds can have different residual fluctuations, so the close integer agreement at this one wave number should not be treated as an exact two-term counting rule. The same calculation on a different aspect ratio at fixed area would change both the perimeter and the lattice pattern.

Measure a reflection phase separately from the travel action

In a one-dimensional cavity a quantisation condition can be written as a round-trip phase budget. Let an allowed region accumulate one-way phase $\theta(E)$ and let reflections at its left and right bounding regions have phase shifts $\phi_L(E), \phi_R(E)$ for unit-modulus reflection. A self-reproducing wave requires

$$2\theta(E) + \phi_L(E) + \phi_R(E) = 2\pi n,$$

with an integer depending on the chosen phase origin. A rigid Dirichlet wall reflects with amplitude -1; a smooth decaying turning point contributes a different reflection phase that the Airy matching fixed. Adding 2π to one reflection phase simply relabels the integer, but an energy-dependent boundary phase alters the spacing and cannot be removed by a constant relabelling.

A useful boundary that can be solved without approximation is a finite step $V = 0$ for $x < 0$, $V = V_0 > E$ for $x > 0$. With incident wave e^{ikx}, reflected amplitude re^{-ikx} and exterior evanescent $te^{-\kappa x}$, where k, κ here are inverse lengths, continuity of the value and derivative gives

$$1 + r = t,\ ik(1 - r) = -\kappa t.$$

Solving yields

$$r = (k - i\kappa)/(k + i\kappa).$$

Its magnitude is one and its phase can be written $-2\arctan(\kappa/k)$. As V_0 tends to infinity with E fixed, the phase tends to $-\pi$, giving the rigid-wall limit. Just below the step threshold it approaches zero. Two mirrors with equally complete reflection can therefore produce different cavity eigenfrequencies.

Near the rigid-wall limit expand $r \simeq -e^{2ik/\kappa}$. To leading order the exterior behaves like a rigid reflecting plane displaced a penetration length $1/\kappa$ beyond the interface. For a deep symmetric finite well of width L, the phase condition approximately replaces L by $L + 2/\kappa(E)$. This shift lowers levels relative to a rigid well of the same geometric width. It is only an approximation because penetration depends on energy. Matching logarithmic derivatives independently on the two sides is preferable near threshold or when the barrier shape varies over the penetration length.

Infer a boundary penetration from a measured free spectral interval

A long one-dimensional cavity with identical deep finite barriers has a measured low-mode wavenumber interval 2% smaller than π/L. Under the constant penetration approximation infer $1/\kappa$ relative to L and identify a check on the approximation.

▼ **Derive and check the result**

Set $\pi/(L + 2/\kappa) = 0.98\pi/L$. Then $2/\kappa L = 1/0.98 - 1 = 0.0204$, or a penetration distance $1/\kappa = 0.0102L$ at each wall. For $L = 100$ nm this is about 1.02 nm. Measure several successive intervals to see whether one penetration length describes them. An increasing penetration with energy changes the reflection-phase derivative and therefore the spacing; the physical group delay then contains travel time in the allowed region plus boundary delay. A geometrically longer rigid cavity would produce a similar single interval but a different energy dependence, so an isolated observation does not uniquely identify the cause.

Tutorial · determine what a semiclassical estimate can certify

Recover the period and level spacing from an action curve

For $A(E) = \int_{x_1}^{x_2} p\,dx$ in a smooth bound well, derive dA/dE and infer the local energy spacing. Apply it to a 100 kHz oscillator and express spacing in both joules and electronvolts.

▼ **Work through the solution**

The endpoint derivative terms vanish because $p = 0$ at both ends. Inside, $\partial p/\partial E = m/p = 1/v$, so dA/dE equals the classical one-way travel time $T_{\mathrm{cl}}/2$. From the quantisation rule increasing n by one

gives local spacing $\Delta E \simeq 2\pi\hbar/T_{\mathrm{cl}} = h f_{\mathrm{cl}}$. At 100 kHz this is 6.626×10^{-29} J or 4.136×10^{-10} eV. In an anharmonic potential the classical period generally depends on energy; use the local slope or integrate between adjacent solutions rather than assuming a universal ladder spacing.

Check whether the Airy overlap interval exists

For a model atom with mass 1.44×10^{-25} kg let $F = 1.0 \times 10^{-21}$ N at the turning point. Find the Airy length. If $|V''| = 1.0 \times 10^{-16}\,\mathrm{N/m}$, estimate the relative quadratic correction at 5ℓ.

▼ **Work through the solution**

The length is $(\hbar^2/2mF)^{1/3} \simeq 3.38 \times 10^{-8}$ m. The ratio of quadratic potential term to linear term at 5ℓ is $|V''|5\ell/(2F) \simeq 0.00845$. Thus several Airy lengths can fit in a region with small curvature correction for these numbers. The parameter $|x - x_t|/\ell = 5$ should still be checked against the desired asymptotic precision; the result supports an overlap interval but does not guarantee a prescribed decimal accuracy. Bringing a second turning point within a few such lengths changes the matching problem qualitatively.

Compare a triangular action with a rectangular shortcut

A linearly falling barrier has initial forbidden wavenumber $\kappa_0/\hbar = 3.0\,\mathrm{nm}^{-1}$ and width $d = 2.0$ nm. Compare its exponential estimate with that of a rectangle using initial height throughout. What would a rectangle at the average potential height predict?

▼ **Work through the solution**

For the triangle, $\kappa(x) = \kappa_0\sqrt{1 - x/d}$, so $I = (2/3)(\kappa_0/\hbar)d = 4.0$, giving $e^{-8} = 3.35 \times 10^{-4}$. The initial-height rectangle gives $I = 6$ or $e^{-12} = 6.14 \times 10^{-6}$. Averaging the potential height first gives wavenumber $\kappa_0/\sqrt{2}$, $I = 4.24$ and $e^{-8.49} \simeq 2.06 \times 10^{-4}$. The action averages the square root of the local energy gap, not the gap itself. Agreement in mean barrier height is insufficient for predicting an exponential observable. Exact matching or a direct numerical probability current would be needed to determine the prefactor.

Further reading

- Fließbach · Näherungsmethoden
- Cambridge · WKB, Airy matching and semiclassical quantisation
- Course investigation · semiclassical quantisation and tunnel splitting

Measure gravity with a pendulum of finite amplitude

Mechanics in the laboratory

A metre-long pendulum is an appealing clock because its small-angle period depends only on length and gravity. A precise clock can nevertheless give a precisely wrong value of gravity. Derive the full period, trace the systematic error to an identifiable approximation and design a measurement that distinguishes a length error from a release-angle error.

Define the measured length and time

Use a point bob on a massless, inextensible string attached to a fixed pivot. Let L be the distance from pivot to the bob's centre of mass, and θ the signed departure from the downward vertical. For a rigid extended bob, rotational inertia about its own centre adds to mL^2; a length measured to the bottom of the bob is neither of these dynamical lengths. Measure the elapsed time between successive crossings of a photogate in the same direction. Crossings in opposite directions give half of the period in a symmetric oscillation. A stationary turning point is a poor timing marker because the position error there produces a large time error.

At rest at the release angle θ_0, the energy is $mgL(1-\cos\theta_0)$. At a later angle the speed is $L|\dot\theta|$. Equating kinetic plus potential energy to the release value gives

$$\dot\theta^2 = (2g/L)(\cos\theta - \cos\theta_0).$$

Only a quarter cycle is needed. Inverting the angular speed and integrating from zero to θ_0 produces a square-root singularity at release. The period is finite since integrating the inverse square root over a small interval gives a square root.

Resolve the turning point before approximating

Write $k = \sin(\theta_0/2)$ and substitute $\sin(\theta/2) = k\sin u$. The vanishing square root cancels a factor $\cos u$ from the differential. One obtains the complete elliptic integral of the first kind with modulus k.

$$T = 4\sqrt{\frac{L}{g}}\,K(k),$$

$$K(k) = \int_0^{\pi/2} \frac{du}{\sqrt{1-k^2\sin^2 u}}.$$

Expanding its integrand yields

$$K(k) = (\pi/2)(1 + k^2/4 + 9k^4/64 + \cdots).$$

Since $k \simeq \theta_0/2$ at small amplitude, the first fractional period correction is $\theta_0^2/16$ with angle in radians. It is positive. At larger amplitude the tangential component of gravity rises more slowly with angle than the linear restoring law predicts. The pendulum spends longer near the turns. Passing degrees directly into the quadratic correction would give nonsense even if the timer and recorded angle were perfect.

Infer the bias in gravity

Suppose the true acceleration is $9.81\,\mathrm{m\,s^{-2}}$, the length is 1.000 m and the release angle is 30 degrees. The small-angle period is about 2.0061 s and the finite-amplitude period about 2.0410 s. Interpreting the longer measured period by the linear formula would give about $9.48\,\mathrm{m\,s^{-2}}$. The fractional gravity error is about twice the fractional period error with opposite sign because gravity is inferred from T^{-2}. Timing a hundred periods reduces random timing error but leaves this systematic shift.

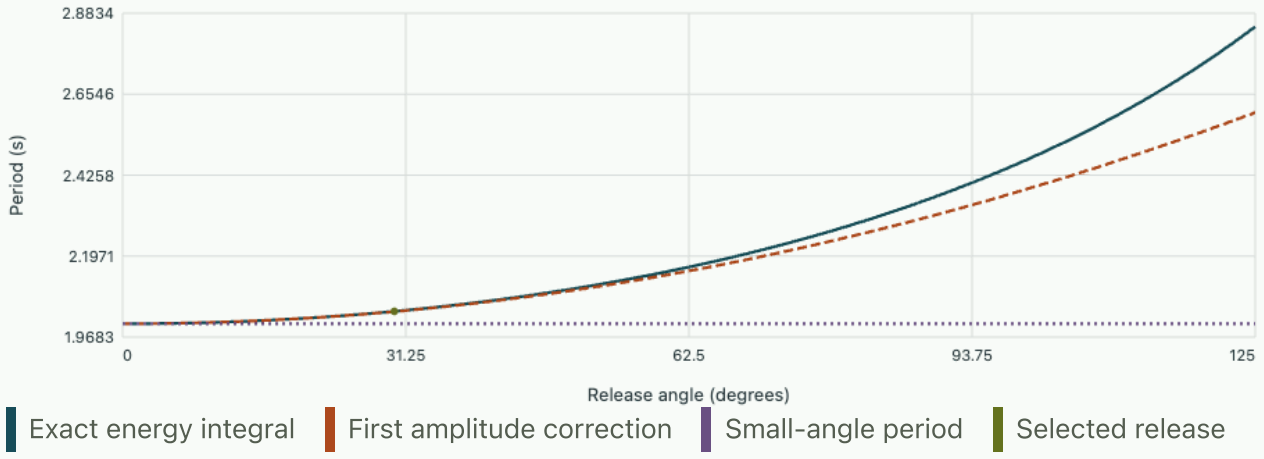
The numerical curve evaluates the complete elliptic integral. It uses the arithmetic–geometric mean identity $K(k) = \pi/[2\,\mathrm{AGM}(1, \sqrt{1-k^2})]$, so no sampled point sits on the turning point. The single-correction curve truncates the amplitude series after its first nonconstant term. Vary angle and length, comparing both the predicted period and the result a linear analysis would infer. Which change rescales both period curves together? Which changes their relative discrepancy?

Open this experiment online

Try the calculation

Release angle (degrees) 30

Pendulum length (m) 1



Period 2.041 s. Small-angle analysis would infer g = 9.4772 m/s².

Design a two-release check

For the same pendulum the ratio $T(\theta_a)/T(\theta_b) = K(k_a)/K(k_b)$ cancels the unknown length and local gravity. Measure at two amplitudes before deciding whether to recalibrate the length. If the ratio agrees but both absolute times differ from prediction, check the effective length, clock rate and model of the bob. If the ratio changes during a decaying swing, record initial and final amplitudes for the timed sequence. The period associated with a long sequence is an average over its slowly changing amplitude. Substituting the initial angle throughout would overcorrect the measurement.

Tutorial

At small amplitude, what release angle gives a fractional period bias of approximately one part in a thousand?

▼ Work through the solution

Set $\theta_0^2/16 = 0.001$. This gives 0.1265 rad, about 7.25 degrees. The gravity bias from an uncorrected measurement would be approximately two parts in a thousand. A tolerance on one quantity is not automatically the same tolerance on the other.

A photogate records successive interruptions every 1.02 seconds for the metre pendulum. A calculation gives nearly four times local gravity. Identify one possible error in the observable.

▼ Work through the solution

The gate may have counted both directions and returned a half period. Doubling that interval gives about 2.04 s. Because the inferred acceleration is inversely proportional to the square of the period, the missing factor of two gives a factor of four.

Why does reducing the grid spacing in an integration of the linear equation fail to recover the finite-amplitude curve?

▼ Work through the solution

Integration error vanishes as the step is refined, but the restoring torque in that equation is $mgL\theta$. The true torque is $mgL\sin\theta$. A converged numerical solution can therefore provide independent evidence of an inadequate physical approximation.

For the equation of motion, phase portrait and dissipation, see the MIT treatment of the simple pendulum.

Applied study

Reconstruct a field from samples and estimate how interpolation changes its integral.

What lies between two measurements?

A probe records temperatures at equally spaced positions along a rod. To estimate the mean temperature, first decide what profile those samples represent. For neighbouring points x_j, x_{j+1} at distance h, write $x = x_j + sh$ with $0 \le s \le 1$. The straight line matching both readings is

$$(I_h f)(x) = (1 - s) f(x_j) + s f(x_{j+1}).$$

The weights are nonnegative and sum to one. This reproduces every affine function, keeps interpolated values between their neighbours, and amplifies bounded sample errors by at most a factor of one. A finer display grid can draw this line more densely. Recovering new curvature requires new source information.

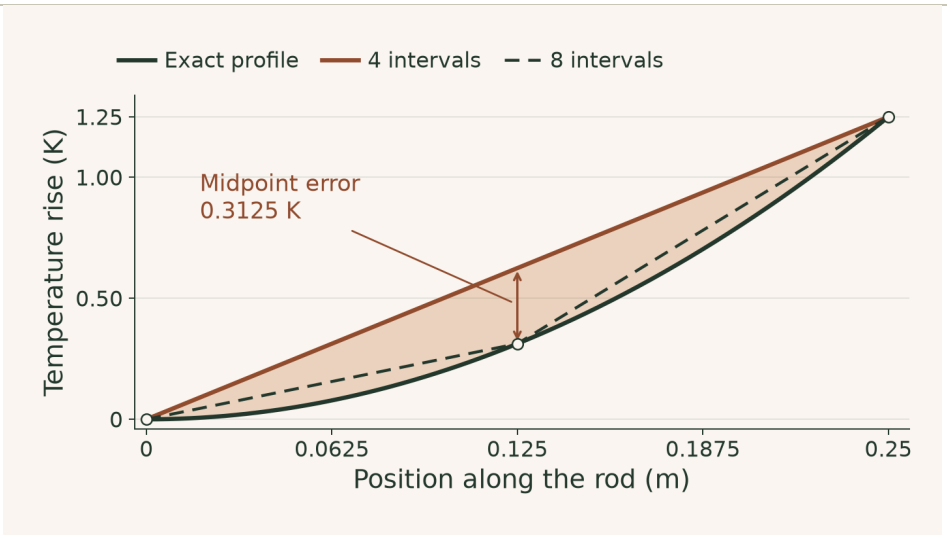
Use curvature to predict the sign and size of error

For twice continuously differentiable f, the interpolation remainder on one panel is $f(x) - I_h f(x) = \tfrac{1}{2} f''(\xi)(x - x_j)(x - x_{j+1})$ for some ξ on that panel. The product is nonpositive and reaches magnitude $h^2/4$ at the midpoint. Thus the maximum error is bounded by $h^2 \max |f''|/8$. For positive curvature the chord lies above the curve.

Integrating the chord gives $h[f(x_j) + f(x_{j+1})]/2$, the trapezoidal rule. Integrating the bound with the quadratic factor retained gives a panel integral error at most $h^3 \max |f''|/12$. On a fixed interval of length L there are L/h panels, so the composite error is bounded by $L h^2 \max |f''|/12$. The pointwise and total error both decrease quadratically, for different reasons.

Integrate a curved temperature profile

Use the exact profile $T(x) = 300 + 20(x/L)^2$ K with $L = 1$ m. Four equal panels have $h = 0.25$ m. On the first panel the endpoints are 300 and 301.25 K. Their interpolated midpoint is 300.625 K, while the exact midpoint is 300.3125 K. The overestimate of 0.3125 K is the same on every panel because the curvature is constant.



The shaded area is the extra integral from the first of four trapezoids on the 1 m rod. Sampling its midpoint splits that trapezoid into two. Subtracting 300 K from both curves makes their separation easier to read.

The exact integral is $300 + 20/3 = 306.6667$ K·m on this unit-length rod. Here $T'' = 40$ K/m² and the trapezoidal overestimate equals $Lh^2T''/12 = 0.20833$ K·m. Dividing the integral by length gives mean temperature. To obtain thermal energy above a reference temperature, additional material properties and cross-sectional area are needed.

Problem class

Choose the sensor spacing

For the quadratic rod, select the minimum number of equal intervals so the midpoint error never exceeds 0.05 K. How many measurements are needed including both endpoints?

▼ Show the calculation

With N intervals on a metre, the bound is

$$40/(8N^2) = 5/N^2 \,\mathrm{K}.$$

Requiring $5/N^2 \le 0.05$ gives $N \ge 10$, hence 11 measurements. At constant curvature this bound is attained at the midpoints.

Infer an integral from two estimates

Four and eight panels give 306.875 and 306.71875 K·m. Assume $Q_h = Q + ch^2$. Eliminate c and estimate Q. Would adding sensors only to the display grid make this correction possible?

▼ Show the calculation

Multiply $Q_{h/2} = Q + ch^2/4$ by four and subtract Q_h. This gives $Q = (4Q_{h/2} - Q_h)/3 = 306.66667$ K·m. For the quadratic profile the error law is exact. Subdividing the same source chords on screen preserves their total area and supplies no second source approximation.

ONESIMULATION LABORATORY

Measure what interpolation does to a heat integral

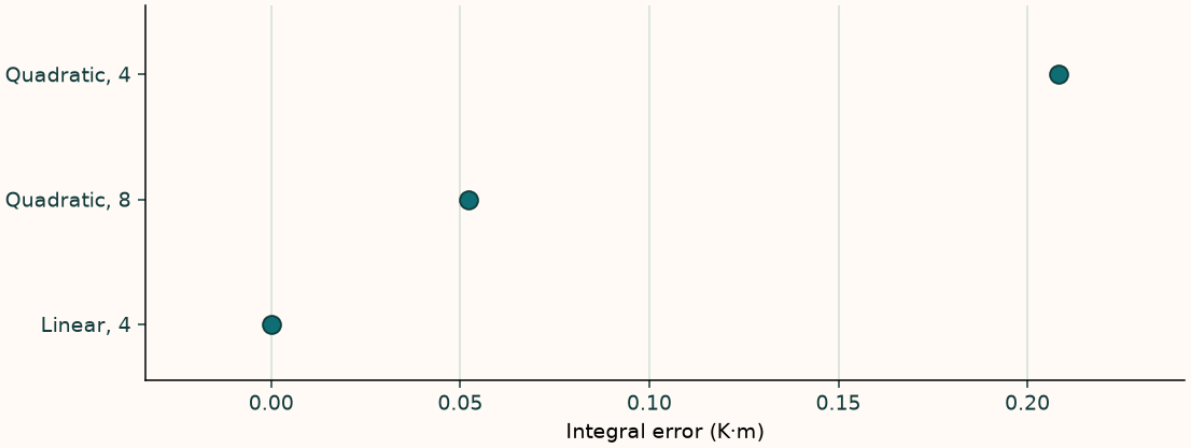
Compare recorded OneSimulation runs on this page. To test your own inputs, download the Python model and run the model in an activated OneSimulation environment.

OneSimulation maps source samples to the original nodes and their midpoints. Integrating those segments gives the area under the interpolant. Keep the quadratic temperature law and double the number of source intervals. Then restore four intervals and replace the quadratic law with $300 + 20x/L$. Download the Python experiment · OneSimulation point interpolation.

```
python approximation-and-quadrature.py
python approximation-and-quadrature.py --cells 8
python approximation-and-quadrature.py --law linear
```

Errors against the exact source function and its integral.

Profile, intervals	Max. midpoint error (K)	Integral (K·m)	Integral error (K·m)
Quadratic, 4	0.3125	306.875	+0.208333
Quadratic, 8	0.078125	306.71875	+0.0520833
Linear, 4	0	310	0



For a quadratic field, doubling the sampling density cuts the integral bias by four. The linear field has zero interpolation bias at the coarser spacing.

Halving the spacing divides both quadratic-profile errors by four. The linear-profile result tests exactness on the function space represented by the interpolant.

Change parameters · Change the laboratory inputs

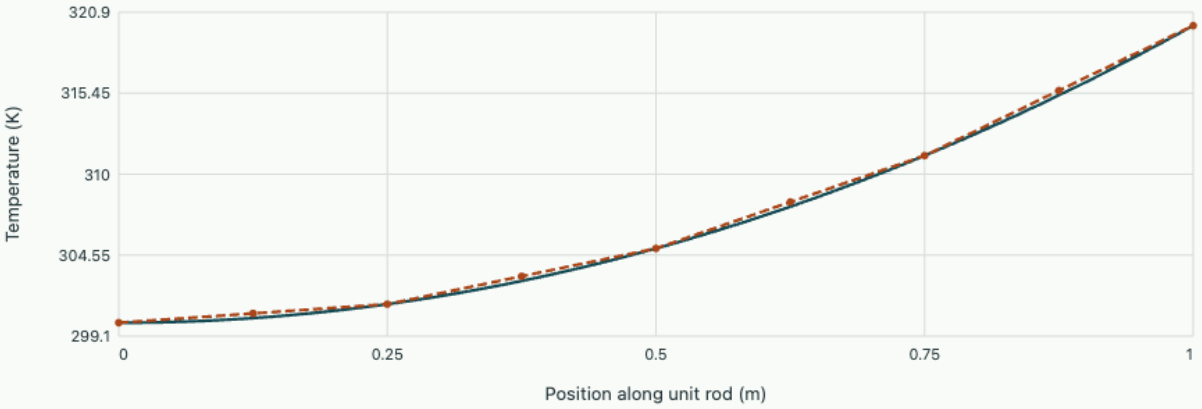
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Source intervals 4

Polynomial profile power 2



Source temperature law | Transferred piecewise-linear field

Integrated transferred field 306.88 K·m; exact integral 306.67 K·m; bias 0.20833 K·m. Refining the target samples alone does not change the source interpolant.

Problem sheet 2 · Recover an integral using two biased readings

Notation and assumptions

Read each defining equation with its assumptions and units, then follow the link to the argument that gives the equation its meaning.

Barycentric weight

$$w_j = \prod_{i \neq j} (x_j - x_i)^{-1}$$

Only relative weights enter the rational expression for the interpolating polynomial. At a node return the recorded value directly.

Derivation · Evaluating an interpolant without rebuilding its polynomial

Fourier alias

$$e^{2\pi i (k+n) j/n} = e^{2\pi i k j/n}$$

On n equally spaced periodic samples two distinct continuum frequencies can yield precisely the same observations. Interpolation cannot disambiguate them without more information.

Derivation · When a trapezoidal sum counts the wrong frequency perfectly

Interpolation condition

$$p(x_i) = f_i$$

The polynomial matches the tabulated values at distinct nodes. Small nodal discrepancies do not guarantee a small error between nodes.

Derivation · Polynomial interpolation and node placement

Monte Carlo standard error

$$\mathrm{SE}(\bar{f}) = \sigma_f / \sqrt{N}$$

Independent samples with finite variance give this ideal error scale. Importance weights and correlations can substantially alter the variance of the actual estimator.

Derivation · Where should a random quadrature spend its samples?

Quadrature degree

$$Q[x^j] = \int_a^b x^j \, dx \quad (0 \le j \le d)$$

Exactness through degree d is an algebraic property. For a nonsmooth or poorly sampled integrand the observed rate can be much lower.

Derivation · Gaussian quadrature and exactness

Richardson extrapolation

$$I \approx I_{h/2} + (I_{h/2} - I_h)/(2^p - 1)$$

This cancels an assumed leading error proportional to h to power p. Several refinements can test whether that asymptotic regime is present.

Derivation · Deriving Romberg integration and recognising a broken error model

Further reading

- Use the Chebyshev examples to connect coefficient decay and smoothness. Compare a function with an interior corner to one with a nearby complex singularity. Lloyd N. Trefethen · Approximation Theory and Approximation Practice
- Chapters 5 and 9 give complementary local and global approximations. Compare those constructions on identical samples. Driscoll and Braun · Fundamentals of Numerical Computation

Compare specific topics with the university sources