Numerical linear algebra

Separate conditioning from algorithmic error, solve dense and sparse systems, and use residuals and spectral information to assess the computed result.

Read the subject in depth

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

- **Matrix computations and spectral accuracy**
  Trace how algebraic structure, representation and numerical precision affect the solution of linear problems. Verify each conclusion with an independent diagnostic.

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. **Operators derived from a physical balance**
   Construct the matrix of a constrained network, a mechanical mode or a spatial field. Determine its nullspace, symmetry and natural inner product before choosing a factorisation.
2. **Ill conditioning and information**
   Inspect the small singular directions in actual observation operators. Distinguish a good residual from a well-determined physical parameter.

How the course fits together

1. **Chapters 1–4**
   Separate input sensitivity from arithmetic error. Elimination solves square systems and orthogonal projection solves overdetermined ones. A stationary iteration replaces one direct solve with a sequence of residual corrections.
2. **Chapters 5–8**
   Use the information from successive matrix products to choose Krylov corrections. Eigenvalues, orthogonal reflections and singular values then explain both algorithm behaviour and sensitivity to perturbations.
3. **Chapters 9–12**
   Exploit sparsity before factorising, build reduced eigenproblems with Arnoldi, improve a computed answer by solving for its residual, and use block elimination when only selected variables are required.

Contents

1. Chapter 1 · Conditioning and backward error
2. Chapter 2 · Elimination, pivoting and factorisation
3. Problem sheet 1 · Measurements and elimination
4. Chapter 3 · Least squares and orthogonal factorisation
5. Chapter 4 · Stationary iterations and smoothing
6. Problem sheet 2 · Fitting and relaxation
7. Chapter 5 · Krylov subspaces and preconditioning
8. Chapter 6 · Eigenvalues, modes and spectral gaps
9. Problem sheet 3 · Search directions and modes
10. Chapter 7 · Householder reflections and practical QR
11. Chapter 8 · Singular values and minimum-norm solutions
12. Problem sheet 4 · Orthogonal factorizations
13. Chapter 9 · Graph structure and fill-in
14. Chapter 10 · Arnoldi iteration and GMRES
15. Problem sheet 5 · Graphs and search spaces
16. Chapter 11 · Iterative refinement and backward error
17. Chapter 12 · Eliminating constraints with Schur complements
18. Problem sheet 6 · Accurate constrained solves
19. Chapter 13 · Why a positive energy has a Cholesky factor
20. Chapter 14 · Removing unequal scales before an iteration
21. Chapter 15 · How a moderately accurate mode can give a precise frequency
22. Chapter 16 · How to rotate two point clouds into agreement
23. Chapter 17 · Reading an effective resistance from a singular graph matrix
24. Chapter 18 · Counting vibration modes without calculating their shapes
25. Problem sheet 7 · Project · Can two displacement sensors distinguish soft and stiff modes?
26. Problem sheet 8 · Can an eigensolver certify the pitch of a continuous string?
27. Matrix computations and spectral accuracy
28. Operators derived from a physical balance
29. Ill conditioning and information
30. Applied study
31. Notation and assumptions
32. Further reading

Chapter 1 · Conditioning and backward error

Distinguish an inaccurate measurement from an inaccurate solve by tracing how each changes the answer to a small equilibrium problem.

FROM FIRST PRINCIPLES

Derive the relative error in a linear solve

Suppose measured data b determine unknowns through $Ax = b$, with A square and nonsingular. Keep the coefficients fixed and change only the measured data by δb. Subtracting the two equations removes the unperturbed solution exactly,

$$A(x + \delta x) - Ax = \delta b,$$
$$\delta x = A^{-1}\delta b.$$

Choose a vector norm and its induced matrix norm, defined as the largest stretch of a unit vector. Applying the induced norm first to the perturbation and then to the original data gives

$$\|\delta x\| = \|A^{-1}\delta b\| \le \|A^{-1}\|\,\|\delta b\|,$$
$$\|b\| = \|Ax\| \le \|A\|\,\|x\|.$$

For nonzero b and x, combine these two inequalities to express both errors as fractions,

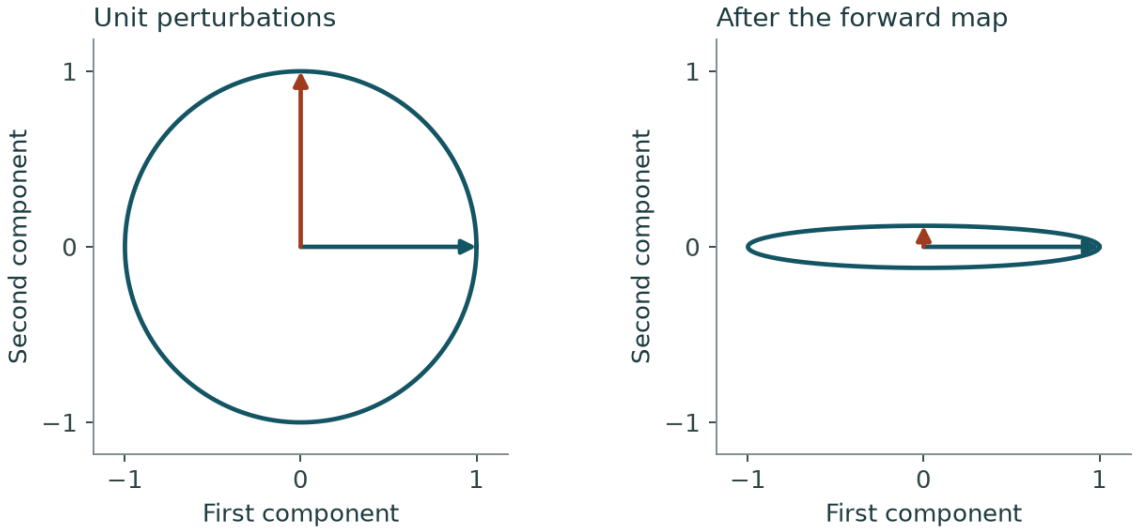
$$\frac{\|\delta x\|}{\|x\|} \le \|A\|\,\|A^{-1}\|\,\frac{\|\delta b\|}{\|b\|}.$$

The product is the normwise condition number. It concerns the problem even when the solver is exact. For a computed solution one can instead insert its measured residual as the data perturbation. The norm bound allows the most sensitive possible orientation of that perturbation, so a particular residual can give a considerably smaller error.

Background · Condition number

$$\kappa(A) = \|A\|\|A^{-1}\|$$

For an invertible square matrix choose a compatible norm. Conditioning describes data sensitivity, while stability refers to the algorithm processing the data.



Which measurement error gets amplified in the inverse?

The image of the unit circle under diag(1, 0.12) supplies both singular values geometrically. An error of 0.012 in the second measured component becomes 0.1 in the reconstructed input. The identical measurement error along the horizontal axis stays 0.012. The direction of the error matters alongside its magnitude.

DERIVATION AND APPLICATIONS

Bound each recovered coordinate using instrument tolerances

Sensors often have different tolerances expressed in different physical units. A perturbation ball based on an overall norm discards this information. Assume A is known and invertible, and let the calibrated uncertainty of reading j be an independent interval $|\delta b_j| \le t_j$. Here independent means all sign combinations are admissible, without a probabilistic assumption. Writing the solution perturbation row by row gives

$$\delta x_i = \sum_j (A^{-1})_{ij} \delta b_j.$$

The triangle inequality provides an upper bound. It is also attainable for a specified coordinate: choose the perturbation signs to agree with the signs of that row. Thus the smallest symmetric enclosing interval for that coordinate is

$$\max_{|\delta b_j| \le t_j} |\delta x_i| = \sum_j |(A^{-1})_{ij}| t_j = (|A^{-1}|t)_i.$$

One can evaluate the same quantity for a reported linear output $z = c^T x$ without forming the whole inverse. Solve one transposed system $A^T w = c$; then the output is $w^T b$ and its extreme uncertainty follows from the same sign argument,

$$|\delta z|_{\max} = |w|^T t,$$
$$A^T w = c.$$

The maximizing perturbations for different coordinates may conflict. A vector of separate coordinate bounds therefore describes an enclosing box; its corners need not be reachable simultaneously. Calibration information tying two errors to the same bias would change the admissible set and can substantially alter the bound.

WORKED EXAMPLE

Bound a difference from two calibrated sensors

Take readings $b_1 = x_1 + x_2$ and $b_2 = x_1 - x_2$, with tolerances 0.020 and 0.080 in common units. Reconstructing $x_2 = (b_1 - b_2)/2$ gives worst error 0.050, attained when the two readings move in opposite directions. For $x_1 + x_2 = b_1$ the error bound is only 0.020. Adding the two individual coordinate bounds would yield 0.100 and miss their shared dependence on the data.

SECOND APPLICATION

Choose the more effective recalibration

A second reconstruction reports $z = 0.1b_1 + 4b_2$, with tolerances 0.50 and 0.050. Its worst error is 0.250. Halving the tolerance of the first sensor reduces it to 0.225, while halving the second reduces it to 0.150. The second recalibration yields a fourfold larger reduction despite its smaller raw tolerance.

Start from this chapter. A small residual controls solution error only through the inverse operator. Compare iterative and direct solutions on identical nodes before changing mesh size.

OneSimulation laboratory · Separate algebraic error from mesh error

Determine which perturbations a machine can represent

A computed residual is assembled from rounded operations. To interpret a small residual we need a scale for each rounding. In a normalized binary system with p significant bits, including the leading one, positive numbers in $[2^e, 2^{e+1})$ have the form $2^e(1 + k\,2^{1-p})$. Increasing the integer k by one advances by 2^{e+1-p}. Thus absolute spacing grows with the size of the number. At a power of two, the neighbours on the two sides even have different spacings.

Rounding to nearest incurs at most half a spacing. Dividing by the smallest magnitude 2^e in that interval gives the relative bound $u = 2^{-p}$, called unit roundoff. For a normal result of suitable magnitude one can write $\mathrm{fl}(z) = z(1 + \delta)$, $|\delta| \le u$. In binary64, $p = 53$. The distance from one to its larger neighbour is 2^{-52}, twice the unit roundoff. Subnormal numbers maintain a constant absolute spacing towards zero. The uniform relative estimate above does not extend through zero or to overflows.

Suppose an instrument adds the increments 2^{-53} twice to a unit baseline. The exact result is $1 + 2^{-52}$. With round-to-nearest, ties-to-even, $1 + 2^{-53}$ is halfway between one and its successor. The lower

endpoint has even trailing significand, so it rounds to one. Adding the second increment separately hits the same tie again. Grouping the increments first produces the representable 2^{-52}. Adding this to one then advances one spacing. Identical mathematical data can therefore yield different rounded sums.

Now take $1 + 2^{-52}$ and subtract one. The difference 2^{-52} is represented exactly, even though it is tiny relative to the original numbers. Cancellation becomes damaging when nearby operands already contain error relative to their difference. Recording the precision, order of accumulation and scale of the data lets us distinguish this effect from sensitivity of the underlying linear system.

Practise on sheet 1 · Predict a dot product in a tiny floating-point system. Further reading · Heidelberg University · R. Herzog, Einführung in die Numerik · Chapter 2, §§5–6 · Zahldarstellung und Rechnerarithmetik · KIT · P. E. Bradley, Numerische Mathematik · §§1.1–1.5 · Gleitkomma-Arithmetik · RWTH Aachen · A. Reusken, Numerische Analysis I · §§2.2.1 and 2.3 · Zahlendarstellungen, Stabilität eines Algorithmus.

Two questions about a reported solution

Suppose you determine two unknown forces by balancing their components in two directions. If the directions are almost parallel, the second measurement repeats much of the first. A machine can solve both written equations very accurately while the inferred forces vary considerably from trial to trial. To understand this, separate the map from data to physical answer from the algorithm used to evaluate that map. Conditioning describes the map locally near the supplied data. Stability describes the calculation that produced an answer from those data.

Let A be a nonsingular real matrix, let $Ax = b$ and suppose that only the right-hand side changes. Subtract $Ax = b$ from $A(x + \delta x) = b + \delta b$. The result $A\delta x = \delta b$ says that A^{-1} maps the measurement disturbance into the answer. In a consistent norm,

$$\|\delta x\| \le \|A^{-1}\| \|\delta b\|.$$

To compare percentages rather than units, use $\|b\| \le \|A\| \|x\|$ and divide through by $\|x\|$. For nonzero b this gives

$$\frac{\|\delta x\|}{\|x\|} \le \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}.$$

The product is the condition number in the specified norm. Rotating the disturbance relative to the singular vectors changes the actual amplification, so a worst-case bound alone cannot tell which measurement caused trouble.

Background · How can changing units change a condition number?

Suppose two independent instruments measure a length and a force. In chosen reference units the measurement matrix is the identity. If the first instrument reports the same length in millimetres rather than metres, its numerical output is multiplied by a thousand. Measuring input and output perturbations with unweighted Euclidean norms then produces a condition number of 1000. The physical readings have exactly the same precision as before.

Represent the instrument uncertainties by standard deviations s_i, and scale each output by its own s_i. Under a change of units, both the first measurement and its uncertainty acquire the factor 1000. Their ratio stays fixed. A physically meaningful sensitivity analysis also scales the unknowns when their units differ. If D_y, D_x contain the respective reference scales, the operator to examine is

$$A_{\text{scaled}} = D_y^{-1} A D_x.$$

Report those scales next to a condition number. An apparently large change in conditioning can then be traced to sensor precision, physical coupling, or simply the coordinates used to report them.

Resolve the sensitive direction

Consider $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1.001 \end{smallmatrix}\right)$ and $b = (2, 2.001)^T$. The exact answer is $(1, 1)^T$. Subtract the first measurement from the second. The difference, 0.001, is exactly 0.001 times the second unknown. If the second measurement increases by 0.0001, its inferred value increases to 1.1. The first unknown must fall to 0.9 to preserve the first measurement. Neither individual equilibrium equation looks unstable by itself. It is their near redundancy that makes the difference sensitive.

The infinity norm makes the scale explicit. The largest absolute row sum of A is 2.001. Its inverse is $1000 \left(\begin{smallmatrix} 1.001 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, whose largest row sum is 2001. The relative data change is $0.0001/2.001$ while the relative solution change is 0.1. Multiplying the data change by the condition number gives about 0.200, a valid upper bound. Had the two measurements increased equally, their difference would have stayed fixed and the same amplification would not occur.

Ask which equations the machine solved

For a computed vector $\hat{x}$, form $r = b - A\hat{x}$. In exact arithmetic $\hat{x}$ solves $A\hat{x} = b - r$ exactly. Consequently $\|r\|/\|b\|$ is a relative backward error when only b may change. In our example the vector $(0.9, 1.1)^T$, tested against the original right-hand side, has a small residual in the second row. It is both an accurate answer to a nearby set of measurements and a poor estimate of the original forces.

If the rows carry different physical units, an unscaled residual can let one equation dominate the judgement. Scale each balance by a stated characteristic magnitude or report the residuals separately. Recompute them from the original data when possible. Feeding the same rounded intermediate quantities back into a consistency check may merely confirm the internal arithmetic. For a sensitive

inverse map, improving the accuracy of the input or measuring an independent direction can accomplish more than tightening the solver tolerance.

Allow uncertainty in the coefficients

If directions as well as forces are measured, the perturbed equations are

$$(A + \delta A)(x + \delta x) = b + \delta b.$$

Keeping first-order terms gives

$$A\delta x \simeq \delta b - \delta A\, x.$$

Uncertainty in a coefficient therefore contributes an effective disturbance proportional to the magnitude of the state it multiplies. Reporting only uncertainty in the measured right-hand side can underestimate the final error.

For a scalar calibration $ax = b$, logarithmic differentiation gives

$$\delta x/x \simeq \delta b/b - \delta a/a.$$

If both measurements are 1% too large with the same sign, their errors cancel exactly in the ratio. Opposite 1% changes give $1.01/0.99 - 1 \simeq 2.02\%$. Componentwise measurement correlations affect typical errors even though a norm bound has to allow the worst admissible combination. In force reconstruction this is a reason to keep calibration information with the measured balances.

Compute a condition number for a particular observable

Suppose a linear solve is used only to predict $q = c^T x$, rather than every component of the state. Let A be fixed and nonsingular. Perturb the load by δb; the output changes by $c^T A^{-1}\delta b$. Solve the dual system $A^T z = c$. Then the exact change is $z^T \delta b$, and Cauchy–Schwarz gives

$$|\delta q| \le \|z\|_2 \|\delta b\|_2.$$

Equality is possible by choosing a load error parallel to z. The absolute condition number of this observable with respect to this load norm is consequently $\|z\|_2$. It can be much smaller than the worst amplification over all outputs.

The error model matters to the result. If each source instead has its own bounded relative uncertainty $|\delta b_i| \le \epsilon |b_i|$, the worst change is $\epsilon \sum_i |z_i b_i|$. The bound is attained by choosing the signs of the perturbations to make every term reinforce the same output change. For independent mean-zero source errors with covariances collected in Σ, the variance is $z^T \Sigma z$, generally smaller than a worst-case bound. A worst-case condition number and a statistical standard deviation answer different questions even when evaluated using exactly the same dual solution.

Which of two coupled readings can be predicted?

Take $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1.02 \end{smallmatrix}\right)$ and load $b = (2, 2.02)^T$. Subtracting the equations gives $0.02 x_2 = 0.02$, and back-substitution gives

$$x = (1, 1)^T.$$

For a prediction of the sum $q_s = x_1 + x_2$, the dual vector is $z_s = (1, 0)^T$: the first equation already reports that sum. For the difference $q_d = x_1 - x_2$, subtract twice the inferred x_2 from the first equation. This gives $q_d = 101 b_1 - 100 b_2$, so $z_d = (101, -100)^T$. These two physically simple outputs depend on the same pair of source readings in strikingly different ways.

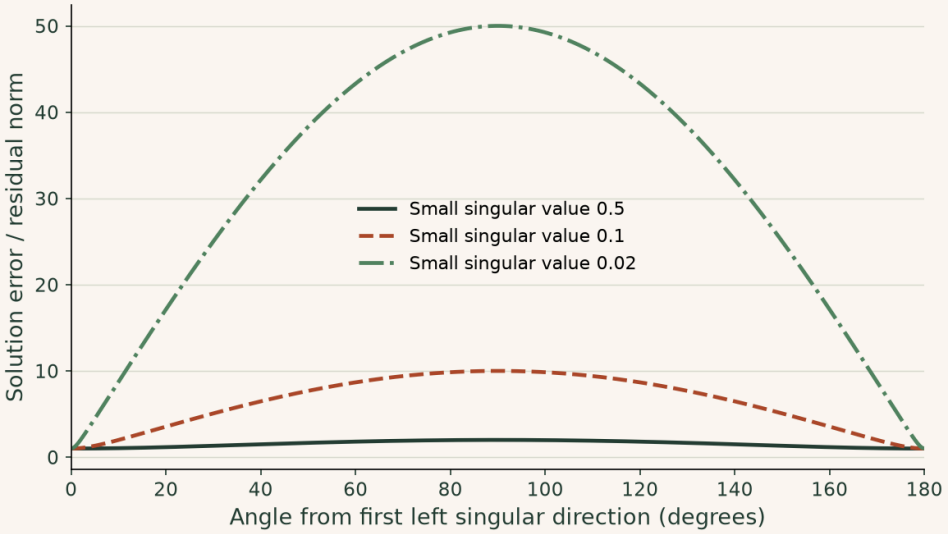
With each reading uncertain by at most 0.001 in absolute units, the sum can change by at most 0.001. The difference can change by as much as 0.201 when the errors have opposite signs. With a single common offset of 0.001 in both readings, however, the difference changes by only 0.001 because the two large terms nearly cancel. The nominal difference is zero, so relative error in that output is undefined. An absolute tolerance or a physically chosen reference amplitude provides a meaningful accuracy requirement.

A designer considering improved load cells should therefore ask whether calibration drift is shared or independent and whether the useful output is a sum or a difference. Reporting just the matrix condition number gives no way to recover either piece of information. If the coefficients of A are uncertain as well, the effective right-hand side also contains the term $-\delta A\, x$ to first order. Its direction can be entirely different from plausible measurement errors in b.

A direct numerical perturbation illustrates the worst bounded case. Increasing b_1 by 0.001 and decreasing b_2 by 0.001 gives $x_2 = 0.90$ and $x_1 = 1.101$. The output difference has changed by 0.201 exactly, demonstrating that the sign choice in the componentwise bound is achievable.

Rotate a residual without changing its norm

Let $A = \operatorname{diag}(1, s)$ and take a unit residual $r = (\cos\theta, \sin\theta)$. Compute the correction solving $A e = r$. Its size depends on alignment even though the residual norm remains exactly one.



For s = 0.02 the correction ranges from one to fifty. The condition number gives the largest possible amplification; a residual norm alone does not identify the direction actually excited.

▶ **Read values from the plot**

Change parameters · Rotate a residual without changing its norm

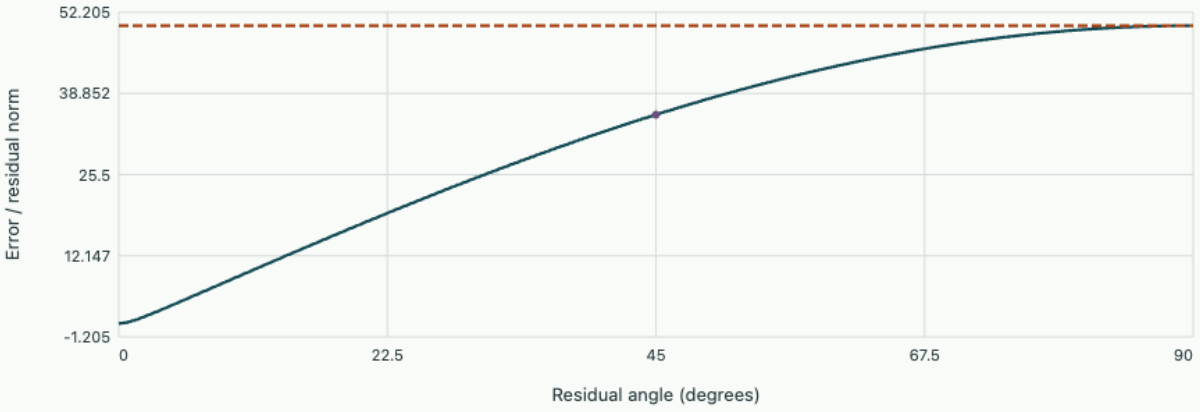
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Weak singular value s **0.02**

Selected residual angle (degrees) **45**



Selected amplification 35.362. All residuals have norm one; $\kappa_2(A) = 50$.

Study the full derivation

When stable eigenvalues allow large transient growth · Matrix computations and spectral accuracy

Two damped states exchange energy in only one direction. Its eigenvalues predict ultimate decay, while a short-time response can grow by a large factor. Compute the matrix exponential, its gain and a resolvent-based sensitivity check.

Chapter 2 · Elimination, pivoting and factorisation

Build an LU factorisation by hand, reuse it for a second load and identify when an innocent-looking elimination step loses useful digits.

FROM FIRST PRINCIPLES

Make one coefficient vanish without changing the solution

Consider a pair of linear balances $ax + by = f$ and $cx + dy = g$. Every solution of both also satisfies their linear combinations. Conversely, adding a multiple of one equation to another is reversible by subtracting the same multiple. For $a \neq 0$, choose the multiplier $m = c/a$ so that subtracting m times the first equation from the second removes x,

$$(d - mb)y = g - mf,$$
$$x = (f - by)/a.$$

The same two operations can be stored as a matrix factorization. The lower triangle records the multiplier and the upper triangle records the remaining coefficients,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ m & 1 \end{pmatrix} \begin{pmatrix} a & b \\ 0 & d - mb \end{pmatrix}.$$

The second pivot $d - mb$ must also be nonzero for a unique solution. Further right-hand sides can reuse both factors because neither depends on f or g. When the first pivot is zero one can try exchanging rows. Exchanging rows to obtain a larger pivot can also reduce rounding errors in the multiplier.

DERIVATION AND APPLICATIONS

Eliminate a variable from a symmetric quadratic energy

For a symmetric stiffness matrix the elimination pivot has a mechanical meaning. Consider the quadratic energy of two displacement coordinates u, v, with stiffnesses a, d, coupling c, and applied loads f, g. Stationarity with respect to each displacement gives the two linear balance equations. Assume the first stiffness $a > 0$,

$$E(u, v) = \tfrac{1}{2}au^2 + cuv + \tfrac{1}{2}dv^2 - fu - gv,$$
$$\partial_u E = au + cv - f.$$

At a prescribed v the energy is minimized at $u = (f - cv)/a$. Complete the square in u before solving the remaining balance. The reduced energy retains both the coupling-induced stiffness change and the force transmitted from the eliminated coordinate,

$$E = \frac{a}{2}\left(u - \frac{f - cv}{a}\right)^2 + \frac{1}{2}\left(d - \frac{c^2}{a}\right)v^2 - \left(g - \frac{cf}{a}\right)v - \frac{f^2}{2a}.$$

The second elimination pivot is the curvature $s = d - c^2/a$. For a stable equilibrium with a unique response one needs $s > 0$. The same result follows by expanding the symmetric factorization

$$A = \begin{pmatrix} 1 & 0 \\ c/a & 1 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & s \end{pmatrix} \begin{pmatrix} 1 & c/a \\ 0 & 1 \end{pmatrix}.$$

Small s means one combined motion changes the energy very little. A large resulting displacement can therefore be a property of the model even if the elimination is performed accurately.

WORKED EXAMPLE

Compute the effective load after eliminating a support

Let $a = 4$, $c = -2$, $d = 3$ in $\mathrm{N/m}$, with loads $f = 2$ and $g = 1\,\mathrm{N}$. The retained stiffness is

$$3 - 4/4 = 2\,\mathrm{N/m}.$$

Its effective load is $1 - (-2)(2)/4 = 2\,\mathrm{N}$, so $v = 1\,\mathrm{m}$ and $u = (2 + 2)/4 = 1\,\mathrm{m}$. Substitution gives original loads $4 - 2 = 2$ and $-2 + 3 = 1$.

SECOND APPLICATION

Diagnose loss of stability without computing eigenvalues

With $a = 9$, $d = 4$, and $c = 6$ the second pivot vanishes. The trial motion $u = -2v/3$ has zero quadratic energy for every v. Changing only c to 6.3 gives

$$s = 4 - 6.3^2/9 = -0.41.$$

Minimizing over u at $v = 2$ and zero load now yields $E = sv^2/2 = -0.82$, demonstrating an unstable direction in the supposed stiffness model.

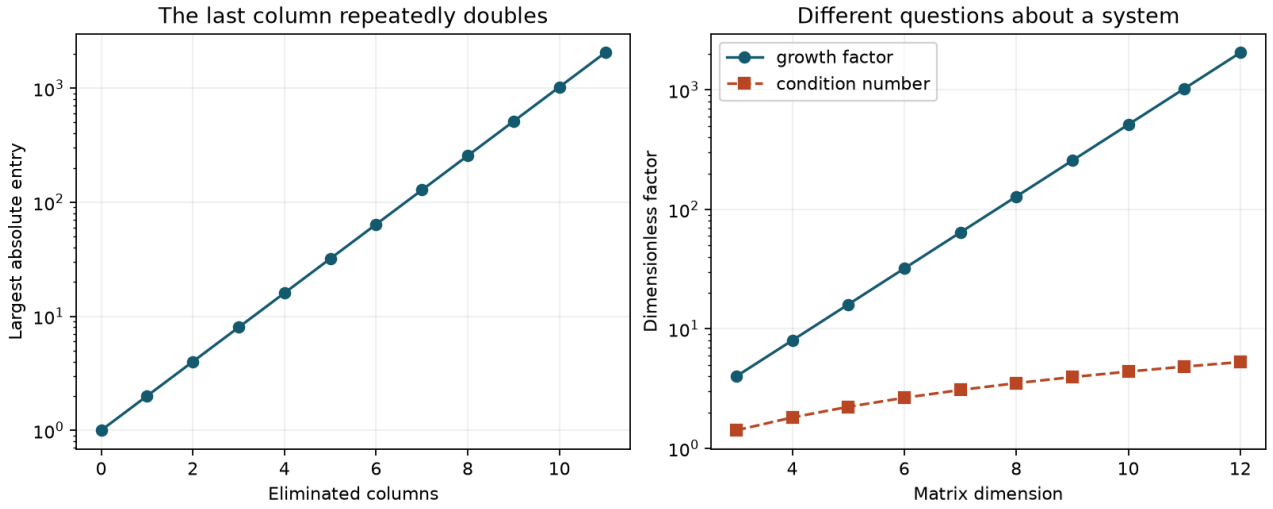
Start from this chapter. Use the direct factorisation as a fixed-mesh reference. Compare it with a converged iterative solve, then check why the peak temperature itself moves on a different mesh.

OneSimulation laboratory · Separate algebraic error from mesh error

How large can the last column grow?

A test matrix can be easy to describe and still produce surprisingly large intermediate values during elimination. Build an n-by-n matrix with ones on the diagonal and in its last column, minus ones strictly below the diagonal outside that column, and zeros everywhere else. Every input coefficient has magnitude at most one. Choose ordinary partial pivoting with the first candidate retained on ties. Every active diagonal pivot in the first $n - 1$ columns equals one and the entries below it equal minus one. There is no tiny pivot and the elimination multipliers all equal minus one.

Follow the last column instead. In the first elimination stage every lower row gains the pivot row. Its last entry changes from one to two. At the next stage every remaining row gains a pivot row whose last entry now equals two. The surviving last entries become four. An induction gives a final pivot of magnitude 2^{n-1}: the growth factor is exponential even though each multiplier is bounded. For six rows the last column of the upper triangular factor reads $1, 2, 4, 8, 16, 32$.



The first curve follows the implementation's arithmetic; the second panel also tests the sensitivity of the original input-output problem. The two quantities describe different possible sources of error.

The plotted condition number is the ratio of the largest and smallest singular values in Euclidean norm. It quantifies the possible relative amplification of perturbations in the original system. The growth factor instead anticipates how large the numbers in intermediate elimination stages can become compared with the input. Standard backward-error estimates contain the absolute sizes of those factors because roundoff is incurred in the intermediate arithmetic. A large growth factor weakens such a guarantee. It does not by itself establish a large solution error for every right-hand side.

In fact all the operations exposed above are small integer additions, sign changes and powers of two. They are exactly representable at the modest dimensions used by this notebook. A demonstration that simply solves this one exact test and prints a small residual would therefore miss the point of the construction. Reverse the row order and repeat the measurement. The singular values of the matrix remain the same, because a row permutation preserves Euclidean norm. Intermediate growth can change dramatically through a different sequence of tied pivot choices.

For a five-row matrix, a solver reports largest intermediate coefficient 16. Another researcher claims this proves that rounding must have lost four binary digits in every solution. Which part of the claim follows?

▼ **Work through the solution**

The factor 16 is correct for the stated tie convention: each of four elimination stages doubles the surviving last entry. Four binary bits describe the ratio of scales. No automatic loss of four actual solution bits follows. Test a perturbed matrix and a specific right-hand side, and report forward error against a reliable reference together with a scaled residual. The exact integer example may incur no rounding in elimination at all.

Open editable notebook Download .ipynb

Eliminate an unknown without changing the solution

Gaussian elimination replaces one equation by itself minus a multiple of another. In exact arithmetic this is reversible: add the same multiple back. In the first column choose a nonzero pivot a_{11} and form multipliers $l_{i1} = a_{i1}/a_{11}$ for the rows below it. Subtract l_{i1} times row one from row i. All lower entries in column one vanish, and the remaining unknowns satisfy a smaller system. Applying the same operations in the same order to the load vector is essential.

Repeating the argument produces an upper triangular matrix U. The inverse row additions can be collected into a unit lower triangular matrix L containing the multipliers. If rows have been permuted, the relation is $PA = LU$, where P records those exchanges. To solve for a new right-hand side, first solve $Ly = Pb$ by forward substitution, then $Ux = y$ by backward substitution. The factorisation depends on the matrix alone. This is valuable in static analysis when the material and supports stay fixed and only the applied loads change.

Factor once and load twice

Consider $A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 2 \end{pmatrix}$. Subtract half the first row from the second to get $(0, 2.5, 1)$. Subtract 0.4 times the new second row from the third to get $(0, 0, 1.6)$. Thus U has diagonal $2, 2.5, 1.6$ with ones immediately above, and L has multipliers 0.5 and 0.4 immediately below its unit diagonal. This matrix needs no pivot exchange. Multiplying the two factors recovers the original entries and gives a quick arithmetic check.

For the load $b = (1, 2, 3)^T$, forward substitution yields

$$y = (1, 1.5, 2.4)^T.$$

Back substitution starts with $x_3 = 2.4/1.6 = 1.5$, then gives

$$x_2 = (1.5 - 1.5)/2.5 = 0,$$
$$x_1 = (1 - 0)/2 = 0.5.$$

A second load $b = (0, 1, 0)^T$ gives

$$y = (0, 1, -0.4)^T,$$
$$x = (-0.25, 0.5, -0.25)^T.$$

Both calculations are short once the triangular factors are available.

Eliminate a chain in two linear-time sweeps

In a chain of nearest-neighbour balances, row i contains only a left coefficient a_i, diagonal d_i, and right coefficient c_i. Subtracting a multiple of the preceding, already reduced row removes the left coefficient and updates just one diagonal and one load. Write the reduced pivot and load as p_i, g_i. The first pair is $p_1 = d_1,\ g_1 = b_1$. Thereafter the multiplier, pivot and load satisfy

$$m_i = a_i / p_{i-1},$$
$$p_i = d_i - m_i c_{i-1},$$
$$g_i = b_i - m_i g_{i-1}.$$

The last equation gives

$$x_n = g_n / p_n.$$

Proceed backwards with $x_i = (g_i - c_i x_{i+1})/p_i$. This is the Thomas algorithm. Each row costs a fixed number of operations and only three coefficient diagonals need storage. On repeated load cases the pivots and multipliers depend only on the original coefficient matrix and can be retained. Symmetric positive definiteness guarantees positive pivots here. An arbitrary nonsingular tridiagonal matrix can still have a zero leading pivot and require pivoting.

A four-node diffusion chain

With conductance one between successive nodes and fixed end values 1 and 2, the four internal balances have diagonal two, neighbours minus one and load $(1, 0, 0, 2)^T$. The reduced pivots are $2, 3/2, 4/3, 5/4$. The corresponding loads are $1, 1/2, 1/3, 9/4$. The last unknown is

$$(9/4)/(5/4) = 9/5.$$

One step back gives $(1/3 + 9/5)/(4/3) = 8/5$, then $7/5$ and $6/5$. Each of the five temperature differences is one fifth, as conservation without an interior source requires. A dense factorisation would return the same field but would obscure why a million-node chain is computationally tractable.

Practise on sheet 1 · Reuse chain factors after a load change. Further reading · University of Limerick · MS6021 Scientific Computation, W. T. Lee · Tridiagonal Matrices: Thomas Algorithm, sections I–II, Stage 1 and Stage 2.

Why exchange rows?

Consider $10^{-5}x + y = 1$ and $x + y = 2$. The exact solution is $x = 1/(1 - 10^{-5})$, $y = 2 - x$. Eliminating with the first coefficient gives multiplier 10^5. The reduced equation contains the difference between numbers of size 10^5. In arithmetic rounded to three significant digits, both the new coefficient and right-hand side round to -1.00×10^5. Back substitution gives $y = 1$, then $x = 0$. Substitution in the second original equation exposes a residual of one.

Swap the rows first. The multiplier is now 10^{-5}, and the reduced equation rounds to $y = 1.00$. Back substitution in the first pivot equation gives $x = 1.00$, close to the exact answer. Partial pivoting chooses the largest absolute available coefficient in the pivot column, bounding the magnitudes of the multipliers by one. It usually limits numerical growth, though it cannot remove sensitivity already present in the original matrix.

For a dense matrix of size n, the triangular factorisation takes order n^3 arithmetic operations and substitutions take order n^2 per load. Sparse discretisations can be much cheaper, but eliminating an unknown may create new nonzero entries between its neighbours. This fill depends on ordering. Ordering for small fill and pivoting for accuracy answer different questions, and production solvers may need both.

Use symmetry when the physics supplies it

For a symmetric positive definite stiffness matrix, Cholesky factorisation represents $A = CC^T$ with a lower triangular C and positive real diagonal. The first diagonal entry is $\sqrt{a_{11}}$. Entries below it are fixed by the first column of the product, and subtracting that column’s outer product leaves a smaller positive definite Schur complement. Positivity follows by minimising the full quadratic energy over the eliminated variable for every fixed nonzero remaining vector.

For $A = \left(\begin{smallmatrix} 4 & 2 \\ 2 & 5 \end{smallmatrix}\right)$, the first column of C is $(2, 1)^T$. The remaining diagonal is

$$\sqrt{5 - 1^2} = 2.$$

A zero or negative candidate under the square root signals that the positive definite assumption or numerical precision needs inspection. Unrestrained mechanical assemblies can have rigid modes of zero energy. Supplying appropriate physical constraints addresses that singularity directly instead of treating it as a pivot-ordering problem.

Trace element growth during pivoted elimination

Partial pivoting chooses the largest available entry in magnitude within the active column. Each multiplier then has magnitude at most one, an important control on an individual elimination step. It does not bound

the size of all updated entries by their original size. The trailing update subtracts a multiplier times the pivot row from another row, and two terms of similar magnitude may add. Repeated reinforcement can build large intermediate entries even when each column choice satisfies the pivot rule.

Measure this effect by the growth factor, the largest magnitude encountered during elimination divided by the largest initial matrix entry. Roundoff in a factorisation is tied to the magnitude of the computed factors, so a large growth factor can weaken a backward-error guarantee. The factor describes a potential loss mechanism. A particular solve can still be exact when every operation happens to be exactly representable, and a large forward error can arise for quite different reasons if the underlying problem is ill-conditioned. These distinctions matter when interpreting numerical experiments designed to stress an algorithm.

Construct a matrix where the final column doubles

Use the four rows $(1, 0, 0, 1)$, $(-1, 1, 0, 1)$, $(-1, -1, 1, 1)$, and $(-1, -1, -1, 1)$. Every initial entry has magnitude at most one. Choose the first available row on tied pivots. Eliminating the first column adds row one to all three lower rows, so their final entries become 2. Eliminating column two adds the second active row to the next two rows, changing their final entries to 4. Eliminating column three adds the third pivot row to the last, producing the final pivot 8. No multiplier exceeded one. The triangular factor has last column $(1, 2, 4, 8)^T$ and the observed growth factor is eight.

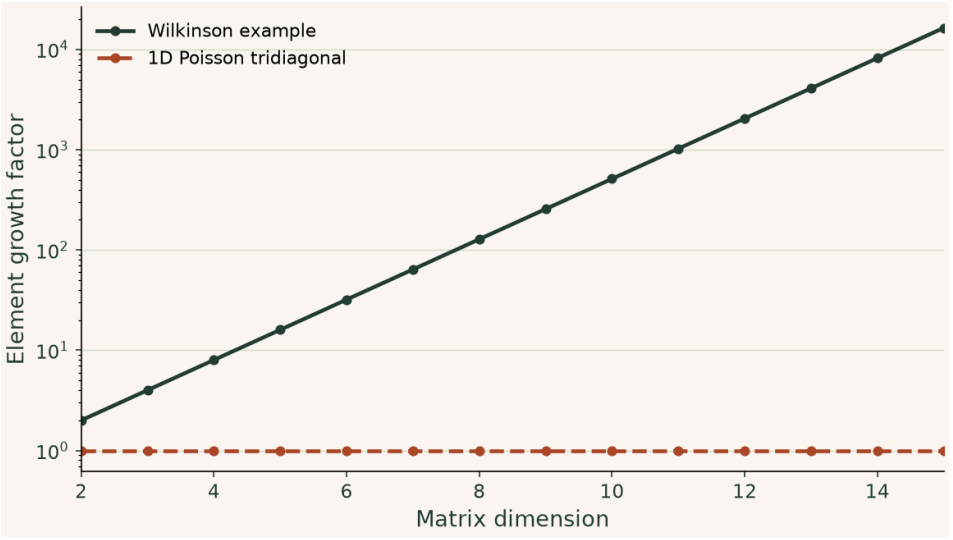
The construction extends directly to size n, with an all-ones last column, unit diagonal before that column, and minus ones below the diagonal. After j stages each remaining row has last entry 2^j, as follows inductively by adding two equal entries during the next elimination. The resulting growth factor is 2^{n-1} for this tie rule. Thus even a rule that prevents division by a small pivot relative to its column admits exponential worst-case growth. Selecting a different row on a tie can change the numerical path without changing the physical system.

For a concrete back-substitution test select exact state $\mathbf{1}$. Multiplication gives load $(2, 1, 0, -2)^T$. Applying the same row operations transforms it to $(2, 3, 5, 8)^T$. Back substitution gives $x_4 = 1$, $x_3 = 5 - 4 = 1$, $x_2 = 3 - 2 = 1$ and $x_1 = 2 - 1 = 1$. Binary arithmetic represents these small integers exactly. A residual check therefore need not reveal anything unusual in this case even while the monitored growth factor exposes an unfavourable factorisation path.

The comparison with no growth is especially clear for a diagonal matrix with entries $(1, 1, 1, 8)$. Its last pivot is also eight, but its largest original entry was already eight and its growth factor is one. A large absolute pivot alone cannot identify element growth; the ratio must refer to the size of the input.

How large can intermediate elimination entries become?

Apply partial pivoting, keeping the first candidate in a tie. The test matrix has ones on the diagonal and final column, minus ones below the diagonal, and zero elsewhere. Record the largest absolute intermediate entry divided by the largest original entry.



Elimination doubles successive last-column entries in the first family. At dimension twelve the growth factor is 2048, versus one for the Poisson example. The graph measures element growth; it does not claim the same factor for the observed solution error.

▶ **Read values from the plot**

Change parameters · How large can intermediate elimination entries become?

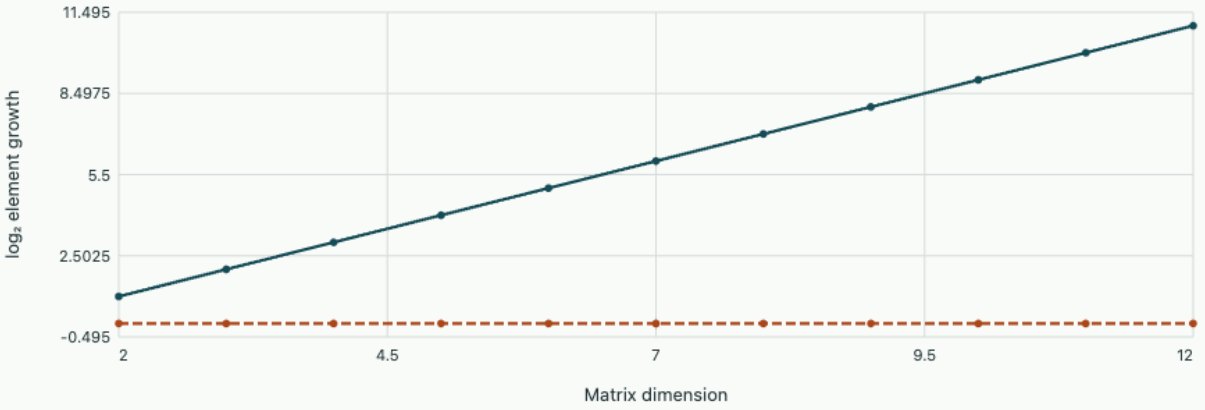
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Largest dimension **12**

Last-column scale **1**



Tie-preserving partial pivoting Poisson tridiagonal family

At dimension 12, growth factor 2048. This is intermediate element growth, not an observed forward-error factor.

Problem sheet 1 · Measurements and elimination

Use explicit perturbations, row exchanges and reusable factors to diagnose errors in small linear balances.

Two scales in a force measurement

A sensor system has $A = \operatorname{diag}(2, 0.02)$ and reports $b = (2, 0.02)^T$. Find the exact unknowns and $\kappa_2(A)$. Compare two additive disturbances of magnitude 0.002, one entirely in the first measurement and one entirely in the second. Finally, explain which sensor needs the tighter absolute calibration.

▼ **Worked solution**

Division by each diagonal entry gives

$$x = (1, 1)^T.$$

The singular values are 2 and 0.02, so $\kappa_2 = 100$. The first disturbance changes the solution by $(0.001, 0)^T$, whereas the second gives $(0, 0.1)^T$. The second direction therefore produces a hundred times more absolute force error for the same measurement error.

For equal admissible absolute errors in the two forces, the second measurement needs a hundred times smaller absolute uncertainty. Rescaling its output by a factor of 100 makes the two matrix rows comparable but also rescales its uncertainty by 100. It cannot turn the original sensor into a more precise one.

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.

Compare a componentwise and a global test

Take $A = \operatorname{diag}(1000, 1)$, $b = (1000, 1)^T$ and $\hat{x} = (1, 0.98)^T$. Compute the infinity-norm relative residual and the componentwise relative residual in each equation. If the instrumentation specification allows 0.5% relative error separately in each reading, does the approximation meet it?

▼ **Worked solution**

The predicted measurements are $(1000, 0.98)^T$, giving residual $(0, 0.02)^T$. The global infinity-norm ratio is

$$0.02/1000 = 2 \times 10^{-5}.$$

In the separate equations the ratios are zero and $0.02/1 = 0.02$.

The second equation misses its 0.5% specification by a factor of four. Global normalisation uses the much larger first equation to set the denominator. One remedy is to divide each equation by its nonzero measurement uncertainty and judge the largest resulting scaled residual.

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.

Pivot and reuse

Factor $A = \left(\begin{smallmatrix} 0 & 2 \\ 3 & 1 \end{smallmatrix}\right)$ with a row permutation. Solve for $b = (4, 5)^T$ and for $c = (2, 1)^T$ using that factorisation. Why does the zero leading entry indicate neither singularity nor lack of a unique solution?

▼ **Worked solution**

Exchange rows with $P = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. The permuted matrix is already upper triangular, so $L = I$ and $U = \left(\begin{smallmatrix} 3 & 1 \\ 0 & 2 \end{smallmatrix}\right)$. For $Pb = (5, 4)^T$, back substitution gives $x_2 = 2$, $x_1 = 1$. For $Pc = (1, 2)^T$, it gives $x_2 = 1$, $x_1 = 0$.

The determinant is -6, so the matrix is nonsingular. A zero at one candidate pivot location merely prevents that ordering of elimination. Applying the row permutation to both the matrix and every new load preserves the original problem.

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

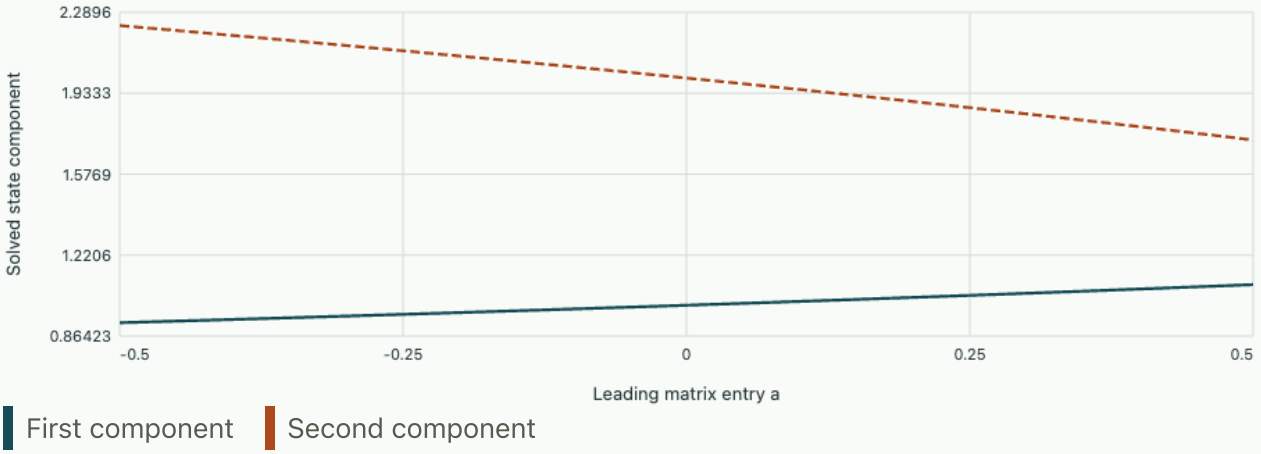
Open this adjustable calculation

Try the calculation

Leading entry a **0**

First load b_1 **4**

Second load b_2 **5**



For A = ((a,2),(3,1)), selected solution (1, 2). Determinant a − 6 = -6 stays nonzero even at a = 0; a row exchange avoids the zero leading pivot.

Reuse chain factors after a load change

A three-node chain has diagonal $(4, 3, 2)$, both off-diagonals equal to -1, and load $(2, 4, 0)^T$. Compute its pivots and solution. Without refactoring, determine the extra response to a unit load on the last node. Explain which step fails for a nonsingular chain with first row $(0, 1)$ and second row $(1, 2)$.

▼ **Worked solution**

The pivots are

$$4,\ 3 - 1/4 = 11/4,\ 2 - 4/11 = 18/11.$$

The reduced loads are

$$2,\ 4 + 2/4 = 9/2,\ (4/11)(9/2) = 18/11.$$

Back substitution gives

$$x_3 = 1,\ x_2 = (9/2 + 1)/(11/4) = 2,\ x_1 = (2 + 2)/4 = 1.$$

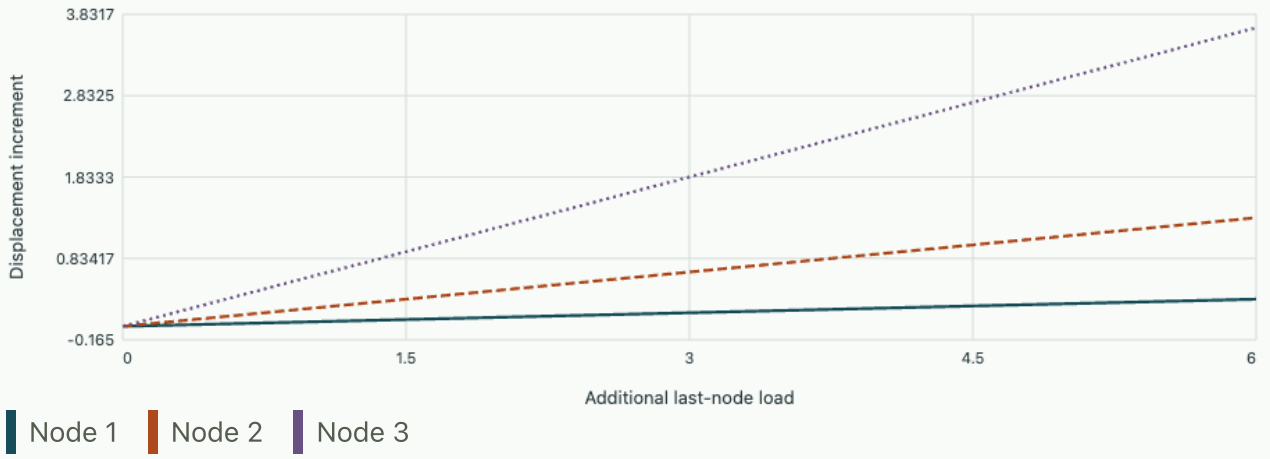
The incremental load transforms to $(0, 0, 1)$. Back substitution through the same factors gives increments $11/18,\ 2/9,\ 1/18$ in reverse node order. They are positive throughout the connected chain. In the final matrix the first pivot is zero, although its determinant is –1. A row exchange can solve it. The unpivoted recurrence cannot divide by its first pivot.

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

Additional load on last node 1



Pivots 4, 11/4 and 18/11 can be reused. At the selected load the solution is (1.0556, 2.2222, 1.6111); the unit response is (1/18, 2/9, 11/18).

Revisit chapter 2 · Eliminate a chain in two linear-time sweeps.

Predict a dot product in a tiny floating-point system

Use four significant binary bits, normalized numbers, round-to-nearest ties-to-even and a sufficiently wide exponent range. Form the dot product $(1, 1, 1) \cdot (8, 0.5, 0.5)$ by adding terms from left to right and then by adding the two smallest terms first. Compute the spacing just above 8. Would accumulating all terms exactly and rounding only the final answer change either result?

▼ **Worked solution**

Above $8 = 2^3$ the spacing is

$$2^{3+1-4} = 1.$$

Both 8 and 9 are representable. The intermediate 8.5 is halfway between them and rounds to the even significand of $8 = (1.000)_2\, 2^3$. Left-to-right accumulation therefore returns 8 after both additions.

The sum of the two half-units is exactly one. Grouping them gives 8+1 = 9, equal to the exact dot product. A single final rounding also gives 9. The products with one are exact here. All of the discrepancy enters through the intermediate additions. A componentwise explanation can identify the two lost half-units even when the input vector is perfectly well conditioned for this sum.

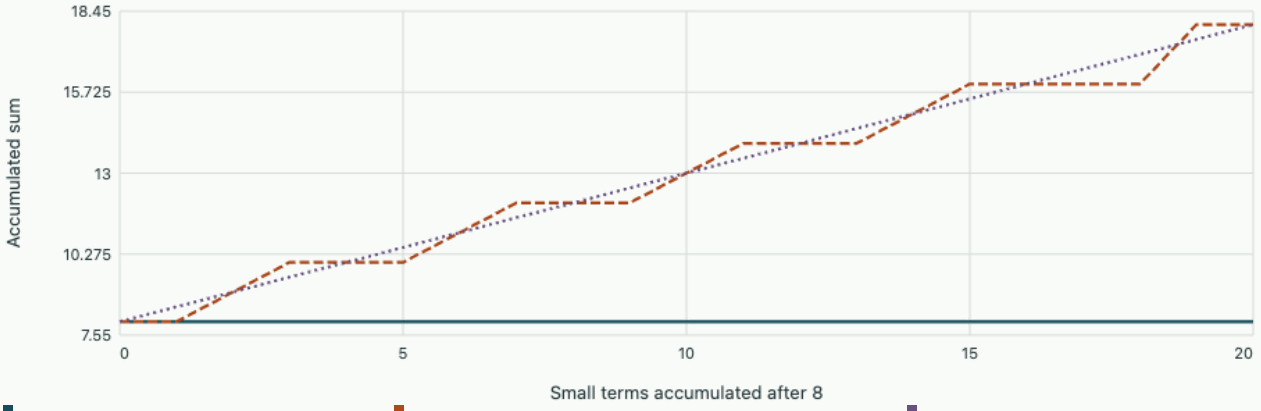
Try this calculation. Change the parameters in the corresponding calculation and compare its independently recomputed values with the equation derived in this solution.

Open this adjustable calculation

Try the calculation

Significant binary bits **4**

Repeated small addend **0.5**



Round after every addition | Accumulate small terms exactly first | Exact real sum

Round-to-nearest, ties-to-even with 4 significant binary bits; spacing above 8 is 1. Only intermediate arithmetic changes between the first two curves.

Revisit chapter 1 · Determine which perturbations a machine can represent.

Notebook practice · How large can the last column grow?

Chapter 3 · Least squares and orthogonal factorisation

Fit a straight line to inconsistent observations, derive its residual conditions and explain why an orthogonal factorisation can preserve accuracy better than the normal equations.

FROM FIRST PRINCIPLES

Derive the orthogonality condition for a fit

With more observations than unknown coefficients, the linear prediction Ax can miss the observed vector b. Suppose squared residuals carry equal weight and choose coefficients to minimize $\|b - Ax\|_2^2$. Try any change $x \mapsto x + td$, where the direction d is fixed but the scalar t can have either sign. Expand with $r = b - Ax$,

$$\|r - tAd\|_2^2 = \|r\|_2^2 - 2t\, d^T A^T r + t^2 \|Ad\|_2^2.$$

At a minimum the linear term must vanish in every direction, so $A^T r = 0$. This says the residual is perpendicular to every vector the model can generate. Rearranging gives the normal equations,

$$A^T A x = A^T b.$$

If A has independent columns then the quadratic term is positive for every nonzero direction and the minimizer is unique. A thin QR factorization with orthonormal columns Q and triangular R yields the equivalent equation $Rx = Q^T b$ and avoids explicitly forming products that can worsen numerical conditioning.

Background · Orthogonal projector

$$P = QQ^{\mathsf{T}}$$

Columns of Q must be orthonormal in the stated real inner product. Least squares removes the component of the data orthogonal to the fitted range.

DERIVATION AND APPLICATIONS

Impose one exact calibration condition on a noisy fit

Some data describe noisy readings while another relationship is an exact conservation or calibration requirement. Model noisy data by $Ax \simeq b$ with independent equal-variance errors, but require $c^T x = d$ exactly. Assume A has full column rank and $c \neq 0$. Write $H = A^T A$, and let x_0 denote the ordinary least-squares solution. Expanding around x_0 removes the cross term because the residual there is orthogonal to every column of A,

$$\|Ax - b\|^2 = \|Ax_0 - b\|^2 + (x - x_0)^T H (x - x_0).$$

Thus the constrained fit is the closest feasible point in the metric determined by H. Introduce a multiplier λ for the exact condition in half the squared objective. Setting its derivatives to zero and solving the first equation gives

$$H(x - x_0) + c\lambda = 0,$$
$$x = x_0 - H^{-1} c \lambda.$$

Applying the constraint determines the multiplier and the increase in squared residual,

$$\lambda = \frac{c^T x_0 - d}{c^T H^{-1} c},$$
$$\Delta \|r\|^2 = \frac{(c^T x_0 - d)^2}{c^T H^{-1} c}.$$

The denominator is positive. In computation the expression $H^{-1}c$ denotes a solve, preferably using existing QR factors. An already satisfied constraint leaves the fit unchanged. Directions weakly constrained by the measured data can accommodate an exact calibration with little increase in residual.

WORKED EXAMPLE

Reconcile two fitted quantities with an exact total

An ordinary fit gives $x_0 = (2, 5)^T$ with $H = \operatorname{diag}(4, 1)$. The correct total is $x_1 + x_2 = 6$. Now $H^{-1}c = (0.25, 1)^T$, the denominator is 1.25, and $\lambda = 0.8$. The adjusted pair is $(1.8, 4.2)$. Its extra squared residual is $4(0.2)^2 + (0.8)^2 = 0.80$, with the less precisely measured second coordinate absorbing most of the correction.

SECOND APPLICATION

Enforce a zero offset in a calibration line

Three equally weighted readings at $t = 0, 1, 2$ are $1, 3, 5$. The free fit is $1 + 2t$. A separate exact zero calibration demands an intercept of zero. Along that feasible line the slope obeys $(0^2 + 1^2 + 2^2)m = 0 + 3 + 10$, giving $m = 2.6$. The three residuals $1, 0.4, -0.2$ have squared norm 1.20, a directly observable cost of the added calibration condition.

Replace exact consistency by an optimisation problem

Three thermometers at different positions need not lie on a common straight temperature profile. Let the rows of A contain the values of candidate basis functions at those positions and let b contain the observations. With more independent measurements than unknown coefficients, $Ax = b$ may have no solution. A least-squares fit chooses x to minimise $\|b - Ax\|_2^2$. Equal unweighted terms assume the

deviations are comparable in units and importance. If one instrument is much noisier than another, that assumption deserves review before interpreting a fitted slope.

Take any coefficient perturbation sd. Expanding the squared mismatch gives $\|b - Ax\|^2 - 2s\, d^T A^T(b - Ax) + s^2\|Ad\|^2$. At a minimum the coefficient of s must vanish for every d. Hence $A^T Ax = A^T b$. This is also sufficient: when the linear term vanishes, every perturbation adds a nonnegative square. Geometrically the residual is perpendicular to the column space of A. Full column rank makes the minimiser unique, since no nonzero coefficient perturbation then leaves the fitted values unchanged.

Background · When is a fitted residual independent of the fitted coefficients?

Assume the design matrix has full column rank and the observation errors have covariance $\sigma^2 I$. Write the parameter error as $A^+\epsilon$ and the residual as $(I - P)\epsilon$, with $P = AA^+$. Their cross-covariance is

$$\operatorname{Cov}(\hat{x} - x, r) = \sigma^2 A^+(I - P)^T = 0.$$

The last equality follows because $I - P$ maps every vector into the orthogonal complement of the columns of A, where A^+ vanishes. For Gaussian errors the two quantities are jointly Gaussian, so zero cross-covariance also gives independence. With non-Gaussian noise their projections may still be statistically dependent. Orthogonality is always an algebraic property of the least-squares fit; independence needs an assumption about how errors were generated.

Fit three imperfect readings

Take positions $t = -1, 0, 1$ and deviations from a baseline temperature of $1, 2, 5$ K. Fit $b(t) \simeq c + mt$, using a constant and a linear column. Their dot product is zero because the positions sum to zero. Their squared lengths are three and two. Multiplying the observation vector by each column gives 8 and 4, so the normal equations are diagonal and yield $c = 8/3$ K and $m = 2$ K per position unit.

The predicted values are $2/3, 8/3, 14/3$ K. The residual $b - Ax = (1/3, -2/3, 1/3)^T$ has zero sum and zero dot product with $(-1, 0, 1)^T$, as required. Changing the intercept cannot lower the sum of squares because the positive and negative residuals already balance; changing the slope has the analogous weighted balance. The squared residual norm is $2/3$ K². This measures fit to these readings. It does not tell whether a line will predict temperatures well outside the sampled interval.

Remove redundant precision loss

If A has full column rank and singular values σ_i, then $A^T A$ has eigenvalues σ_i^2. Consequently

$$\kappa_2(A^T A) = \kappa_2(A)^2.$$

When two model columns are nearly parallel, forming the products explicitly can lose significant information about their small difference. For example, a straight-line fit on positions clustered around a very large offset has nearly proportional constant and position columns. Centering and scaling positions improves the parameterisation and gives an intercept with a local physical interpretation.

Alternatively, factor $A = QR$ with Q having orthonormal columns and R upper triangular. Decompose the residual into its orthogonal parts inside and outside the column space:

$$\|b - Ax\|^2 = \|Q^T b - Rx\|^2 + \|(I - QQ^T)b\|^2.$$

The second term is independent of the fit. Solve $Rx = Q^T b$ by back substitution. Householder reflectors build this factorisation using length-preserving transformations. In the three-temperature example, dividing the constant column by $\sqrt{3}$ and the position column by $\sqrt{2}$ already produces Q. The transformed data are $8/\sqrt{3}$ and $4/\sqrt{2}$. Dividing by the diagonal of R recovers exactly the same fit. When the columns are nearly dependent, inspect the rank explicitly, using column pivoting or singular values, before giving a physical interpretation to individual coefficients.

Include measurement weights explicitly

If an instrument gives independent errors with standard deviations s_i, divide row i of both A and b by s_i. The sum of squares now measures squared differences in units of each sensor’s own standard deviation. For two readings 0 and 10 with standard deviations 1 and 3, a constant fit minimises $c^2 + (10 - c)^2/9$. Differentiation gives $c = 1$, versus the unweighted mean of 5. A reliable low reading exerts more influence on the weighted result.

This weighting still assumes independence and an appropriate error description. If both readings share an uncertain offset, repeated observations can reduce independent fluctuations without removing the offset. Near a rank deficiency, a small weighted residual likewise cannot establish separate physical identifiability of the fitted coefficients. Examine uncertainty or singular directions in the parameter map as well as the final mismatch.

Derive a prediction after withholding one observation

Let A have full column rank and use equal independent data variances. The fitted vector is $\hat{y} = Hy$ with $H = A(A^T A)^{-1} A^T$; implement the projection with a stable factorisation. Suppose row i is withheld and the reduced design stays full rank. Its new prediction at that location, $\hat{y}_i^{(-i)}$, can be obtained without refitting the entire data set. Replace y_i in the full data by this predicted value. The model fitted to the reduced observations now also satisfies the full normal equations because the added row has precisely zero residual at that fit.

Apply the original hat matrix to the modified vector. Its i-th fitted value is $\hat{y}_i + h_{ii}(\hat{y}_i^{(-i)} - y_i)$, which must equal $\hat{y}_i^{(-i)}$. Rearranging gives

$$y_i - \hat{y}_i^{(-i)} = (y_i - \hat{y}_i)/(1 - h_{ii}).$$

This relation shows how leverage can mask an error inside the training data. If $h_{ii} = 1$, the denominator vanishes because that observation alone identifies some parameter combination. Withholding it then requires a different rank convention or additional modelling information.

Predict the end of a calibration line from its other readings

Fit $y = a + bt$ to $(t, y) = (0, 0), (1, 1), (2, 4)$. The mean time is one and the mean response is $5/3$. The slope from centred data is $[(-1)(-5/3) + (1)(7/3)]/[1 + 1] = 2$, giving intercept $-1/3$. The fitted values are $(-1/3, 5/3, 11/3)$ and residuals are $(1/3, -2/3, 1/3)$. Taken individually, either endpoint looks better fitted than the midpoint. But each endpoint also has disproportionate influence on the slope used to fit itself.

For a straight line the intercept and centred-time columns are orthogonal, so $h_{ii} = 1/3 + (t_i - 1)^2/2$. The endpoint leverages are $5/6$, the central leverage is $1/3$, and their sum is two, the number of fitted parameters. The leave-one-out residuals are consequently $2, -1, 2$. Verify the last directly: the first two readings define the line $y = t$; it predicts 2 at the final time, while the held-out observation is 4. For the central reading the two endpoints define $y = 2t$, predicting 2 instead of 1.

The scientific question determines whether this test is relevant. Dropping an endpoint forces extrapolation, whereas dropping the midpoint tests interpolation between two retained settings. Averaging these three squared errors mixes two distinct prediction tasks. If future measurements will be taken only inside the calibrated interval, a separate evaluation there is more informative. Serially correlated measurements also invalidate the interpretation of neighbouring held-out readings as independent validation, although the algebraic leave-one-out identity for a fixed design remains valid.

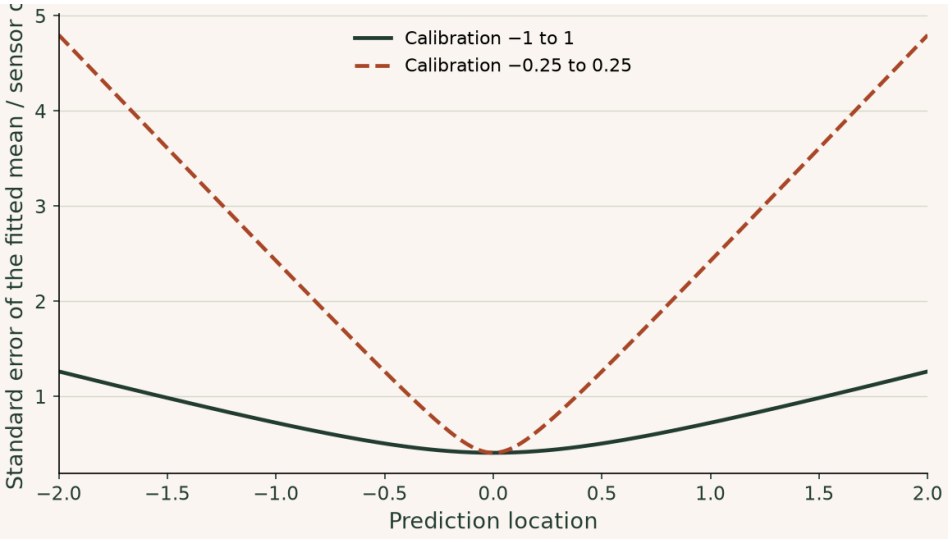
For these three observations the sum of squared training residuals is

$$1/9 + 4/9 + 1/9 = 2/3.$$

The sum after leaving each observation out in turn is $4 + 1 + 4 = 9$. The discrepancy arises even with an exactly implemented fitting routine. Each training prediction has already used the observation against which it is compared. The difference becomes acute near the boundary because that observation has supplied much of the information about its own fitted value.

Estimate a fitted line beyond its calibration interval

Six independent sensors of equal variance calibrate an intercept and slope. For each proposed location use the row $c = (1, x)$. The standard error of the fitted mean is $\sigma \| R^{-\mathsf{T}} c \|$ from the QR factorisation of the design.



Both designs give $1/\sqrt{6}$ times sensor σ at the centre. Narrowing the calibration range by four multiplies the slope standard error by four. To predict an individual future reading add its variance before taking the square root.

▶ **Read values from the plot**

Change parameters · Estimate a fitted line beyond its calibration interval

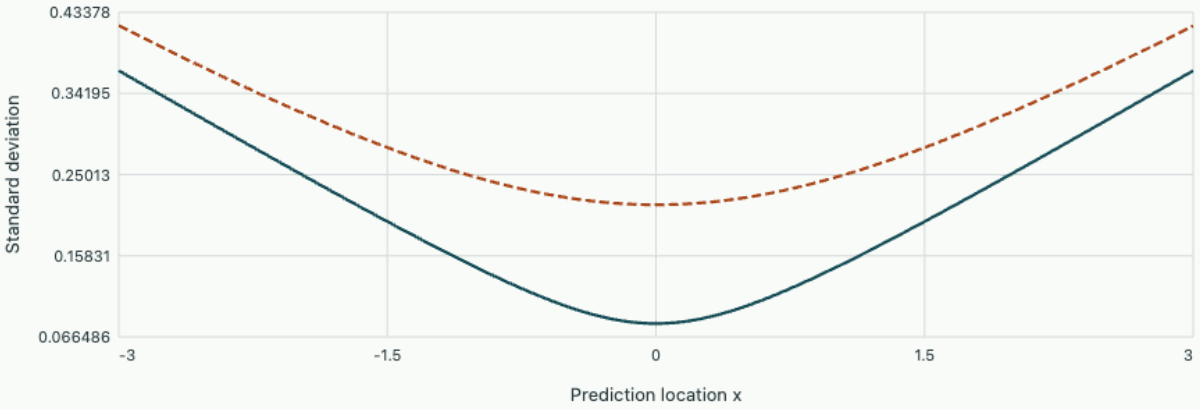
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Calibration half-width **1**

Independent sensor SD **0.2**



Slope SD 0.11952; intercept SD 0.08165. Six equally spaced calibration readings have equal, independent variance.

Study the full derivation

Refitting a sensor as observations arrive or disappear · Matrix computations and spectral accuracy

A calibration line receives one new observation at a time. Update its least-squares factorisation using rotations, quantify the influence of a deleted observation and detect when removing a sample destroys identifiability.

Chapter 4 · Stationary iterations and smoothing

Derive two iterative solvers by splitting a matrix, and decide from their error modes whether repeated updates improve an approximation.

FROM FIRST PRINCIPLES

Turn a balance into a repeatable correction

Suppose a large linear system $Ax = b$ is difficult to solve, but a nearby matrix M is easy to invert. At a trial value x_k, compute the imbalance $r_k = b - Ax_k$. Estimate the correction by solving $Md_k = r_k$, then add it to the trial value,

$$x_{k+1} = x_k + M^{-1}(b - Ax_k).$$

For the true solution x_* the residual vanishes. Subtract this fixed solution from the update and define $e_k = x_k - x_*$. The algorithm propagates its remaining error through one fixed matrix,

$$e_{k+1} = (I - M^{-1}A)e_k,$$
$$e_k = (I - M^{-1}A)^k e_0.$$

A sufficient test for convergence is an induced norm of the iteration matrix below one. More generally, in finite dimension convergence for every starting error requires all eigenvalues to have magnitude below one. Taking M as the diagonal of A gives Jacobi iteration when all diagonal entries are nonzero. Its cheap solves are worthwhile only if the resulting errors decay at a useful rate.

DERIVATION AND APPLICATIONS

Balance the slow and fast modes of a fixed relaxation step

Suppose the linear operator A is real symmetric positive definite and estimates ℓ, u bound its eigenvalues, with $0 < \ell \le u$. Consider a scalar relaxation $x_{k+1} = x_k + \omega(b - Ax_k)$. Because A has an orthonormal eigenbasis, each error component evolves independently. For an eigenvector belonging to λ, its scalar error changes according to

$$e_{k+1,\lambda} = (1 - \omega\lambda)e_{k,\lambda}.$$

The largest possible amplification over the spectral interval occurs at an endpoint. Indeed the absolute value of an affine function is convex. Thus choosing a single step for every component asks for the smallest possible value of

$$q(\omega) = \max\{|1 - \omega\ell|,\ |1 - \omega u|\}.$$

Below the optimum the low-frequency endpoint with smaller eigenvalue decays too slowly; above it the high-eigenvalue endpoint overshoots too far. Equalizing their absolute values with opposite signs gives the best fixed step and its worst-case bound,

$$1-\omega\ell = -(1-\omega u),$$

$$\omega_* = \frac{2}{\ell+u},$$

$$q_* = \frac{u-\ell}{u+\ell}.$$

For coincident endpoints a single step solves the problem. When the ratio u/ℓ is very large, even the optimal scalar choice leaves q_* near unity. The result bounds the Euclidean norm directly because the eigenvectors are orthogonal. For a nonsymmetric matrix, eigenvalue endpoints alone do not establish this norm bound.

WORKED EXAMPLE

Set the step and predict an iteration budget

A model has spectrum in $[2,8]$. Choose $\omega = 0.2$; then the endpoint error factors are $+0.6$ and -0.6. To guarantee reduction by a factor of 1000 requires $k \geq \ln(0.001)/\ln(0.6) = 13.52$, hence 14 steps. An error entirely in the mode $\lambda = 5$ vanishes after the first update, showing why actual runs can beat the guarantee.

SECOND APPLICATION

Diagnose divergence from a missed fast mode

A second solver uses $\omega = 0.25$ on a system initially believed to have largest eigenvalue 6. A newly resolved mode has $\lambda = 9$, giving amplification $1 - 0.25(9) = -1.25$. Its error reverses sign and grows by a factor $1.25^4 = 2.44$ in four steps. With a verified interval $[1,9]$ the corrected fixed step is 0.2, yielding a contraction bound 0.8.

Split a difficult solve into easy solves

A direct factorisation may consume substantial memory on a fine mesh. One alternative repeatedly solves with a matrix M that is easy to invert. Write $A = M - N$. Starting at $x^{(0)}$, solve $Mx^{(k+1)} = Nx^{(k)} + b$. This update can also be written as a correction $x^{(k+1)} = x^{(k)} + M^{-1}(b - Ax^{(k)})$. The exact solution is a fixed point. A convenient choice of M is useful only if these cheap corrections move arbitrary starting guesses toward that point.

Subtract the update from the exact fixed-point equation. The error propagates as $e^{(k+1)} = Ge^{(k)}$, where $G = M^{-1}N = I - M^{-1}A$. If G is diagonalizable, expansion of the initial error in its eigenvectors gives a factor λ_i^k per mode after k steps. The spectral radius below one gives convergence from every initial vector. A mode with magnitude close to one decays slowly. For a nonnormal matrix the eigenvectors may be nearly parallel, so several early error norms can rise even when the ultimate decay is assured.

Update a two-node network

Let $A = \left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$ and $b = (1, 1)^T$, with exact solution $(1, 1)^T$. Jacobi uses $M = \operatorname{diag}(A)$. The updates are

$$x_1^{new} = (1 + x_2^{old})/2,$$
$$x_2^{new} = (1 + x_1^{old})/2.$$

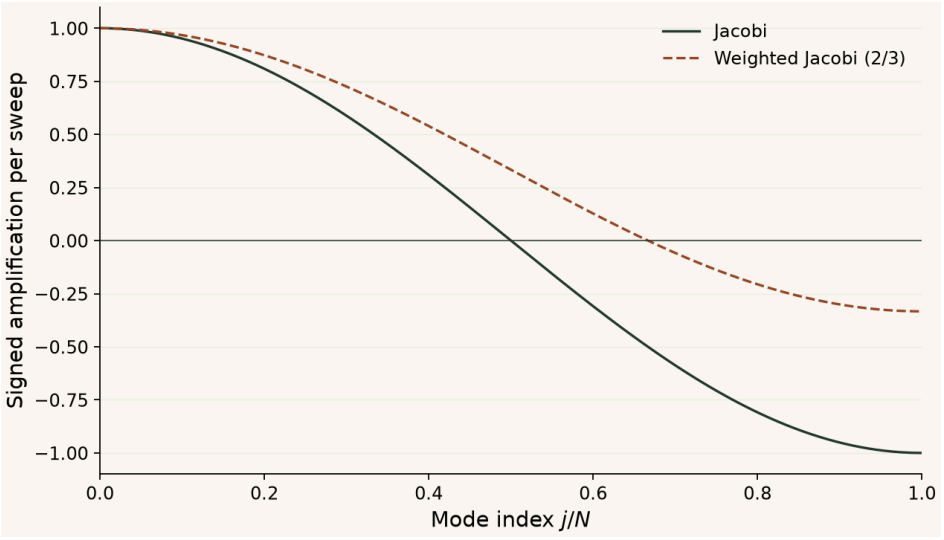
Starting at zero gives $(0.5, 0.5)$, then $(0.75, 0.75)$. Here the initial error lies in the symmetric eigenvector, with factor $+0.5$. An initial antisymmetric error would change sign each step with factor -0.5. Both shrink at the same magnitude despite different histories.

Gauss–Seidel includes the lower triangular part of A in M. It uses the just-computed x_1^{new} in the second equation. From zero, the first sweep gives $(0.5, 0.75)$ and the next gives $(0.875, 0.9375)$. The error propagation matrix for this ordering is $\left(\begin{smallmatrix} 0 & 1/2 \\ 0 & 1/4 \end{smallmatrix}\right)$, whose eigenvalues are zero and one quarter. A sweep uses recent information, but introduces a sequential dependence that plain Jacobi avoids.

Why a good smoother can be a poor solver

On a uniform one-dimensional Dirichlet grid with N intervals, each Jacobi correction for the Poisson equation replaces the error by its neighbours’ average. Substitute the sine-shaped error $e_i = \sin(j\pi i/N)$. The addition identity shows that its amplification is $\cos(j\pi/N)$. Low-frequency modes vary gently between neighbours, so averaging barely changes them. As the grid is refined the slowest factor tends to one. An iteration count measured on a coarse problem therefore gives little indication of fine-grid cost.

Weighted Jacobi mixes one ordinary Jacobi sweep with the old iterate. The factor becomes $1 - \omega + \omega\cos(j\pi/N)$. Choosing $\omega = 2/3$ sends a very oscillatory mode with cosine near minus one to approximately minus one third. The unweighted scheme would retain almost all of its magnitude with alternating sign. For a smooth mode with cosine near one, both updates do little. A multigrid method exploits this distinction: several weighted sweeps remove much of the local oscillation, and a coarser representation corrects the remaining long-scale error. Always evaluate a smoother by its effect on selected frequencies and a standalone solver by the slowest remaining mode.



Modal smoothing on a uniform one-dimensional Poisson grid. A sign change reverses the error pattern; its absolute factor determines how quickly the amplitude decays.

Estimate the cost before extrapolating a timing

For the slowest sine error on a grid of N intervals,

$$\cos(\pi/N) = 1 - \pi^2/(2N^2) + O(N^{-4}).$$

To reduce this component by a factor of ten, solve $\cos(\pi/N)^k \le 0.1$ using logarithms. For large N this gives

$$k \simeq 2N^2 \ln(10)/\pi^2.$$

Doubling the number of intervals therefore roughly quadruples the sweep count for ordinary Jacobi. The work in a one-dimensional sweep also doubles. A timing extrapolated solely from the cost of one sweep misses this additional factor of four.

This estimate starts with an error containing the first sine mode. If a benchmark uses a carefully chosen high-frequency error, a weighted Jacobi run may look much faster. The two measurements describe different aspects of the iteration. Repeating the comparison with both a smooth and a rapidly alternating initial error makes that dependence observable in a short experiment.

Convergence can begin with a large increase

Write an affine stationary iteration as $x_{k+1} = Tx_k + c$. Subtracting its fixed-point equation shows that $e_k = T^k e_0$ for the error in either consistently chosen sign convention. Every initial error decays in finite dimension exactly when the spectral radius of T is smaller than one. Diagonalising a normal T then gives

$$\|T^k\|_2 = \rho(T)^k.$$

For a nonnormal matrix its eigenvectors need not be orthogonal and this norm identity fails. A shrinking asymptotic factor can coexist with substantial early error growth.

Consider $T = rI + N$ with $N^2 = 0$ and $0 < r < 1$. Since I and N commute, the binomial expansion leaves $T^k = r^k I + k r^{k-1} N$. The second term eventually vanishes because exponential decay dominates a polynomial factor. Initially it can be large if the off-diagonal coupling in N is large. This is a precise way to distinguish eventual convergence from contractivity in a specified physical norm. Monitoring only the eigenvalue magnitude hides the size of short-term excursions experienced by the components.

A coupled recurrence that delivers the wrong early verdict

Take $r = 1/2,\ N = \left(\begin{smallmatrix} 0 & 8 \\ 0 & 0 \end{smallmatrix}\right)$, and start from error $e_0 = (0, 1)^T$. The exact formula gives $e_1 = (8, 1/2)^T$, $e_2 = (8, 1/4)^T$, $e_3 = (6, 1/8)^T$, and $e_4 = (4, 1/16)^T$. An investigator who compares the first error norm with its initial value would conclude that the procedure is unstable. Yet both eigenvalues are one half and the formula assures eventual decay. At ten steps the first component is

$$80/512 = 0.15625.$$

This slow recovery after a rapid early amplification is a property of the coupling.

A second experiment starts from $e_0 = (1, 0)^T$. The nilpotent part annihilates that direction and the first component decays as 2^{-k} from the very beginning. Thus observing a successful trajectory from just one initial condition can also conceal the largest early amplification. In a simulation that uses this solve inside another algorithm, those intermediate excursions may violate physical admissibility even if the mathematical linear system would eventually be solved to high accuracy.

Scaling the first coordinate by a factor of sixteen reduces the numerical off-diagonal entry in the scaled iteration from eight to one half. The infinity norm of that scaled matrix is then one; a slightly stronger scaling makes it strictly contractive. This does not change any eigenvalue or physical trajectory. It changes how coordinate errors are measured. An engineer should specify whether the chosen scaled tolerance actually controls both components to their required physical accuracies before using contractivity in that norm as a practical stopping guarantee.

A related diagnostic is to distinguish the one-step map norm from its spectral radius numerically. Here applying T to the unit vector e_2 gives length $\sqrt{64 + 1/4} > 8$, so its two-norm is at least eight. No calculation of eigenvalues can give a comparable lower bound on this one-step gain in a nonnormal problem. Computing a largest singular value would measure exactly that worst initial norm amplification.

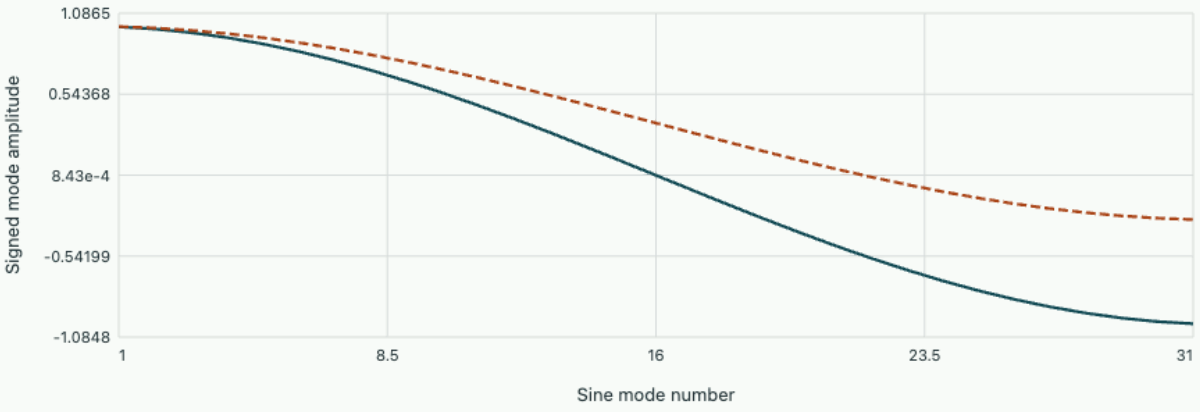
Change parameters · Stationary iterations and smoothing

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	32
Relaxation weight ω	0.65
Sweeps	1



Slowest weighted one-step factor 0.99687. After 1 sweeps the signs distinguish phase reversal from attenuation.

Problem sheet 2 · Fitting and relaxation

Use orthogonality to judge a fit and spectral factors to predict what successive relaxation sweeps remove.

Reweight an unreliable reading

Fit a constant to the readings 1, 2 and 8. First minimise the unweighted sum of squared deviations. Next suppose the third reading has standard deviation twice that of the others, with independent zero-mean errors. Write the weighted objective, find its minimiser, and explain the change.

▼ **Worked solution**

For equal weights, differentiating $\sum (y_i - c)^2$ gives $3c = 11$, so $c = 11/3$. With independent errors, weighting by inverse variance gives relative weights $1, 1, 1/4$. Differentiate $(1-c)^2 + (2-c)^2 + (8-c)^2/4$ to get $2.25c = 5$, or $c = 20/9$.

The high reading has less leverage because its prescribed uncertainty is larger. The weighted residual sum vanishes at the answer: $\sum w_i (y_i - c) = 0$. This argument depends on the uncertainty model, and a common calibration offset cannot be removed by treating measurements as independent.

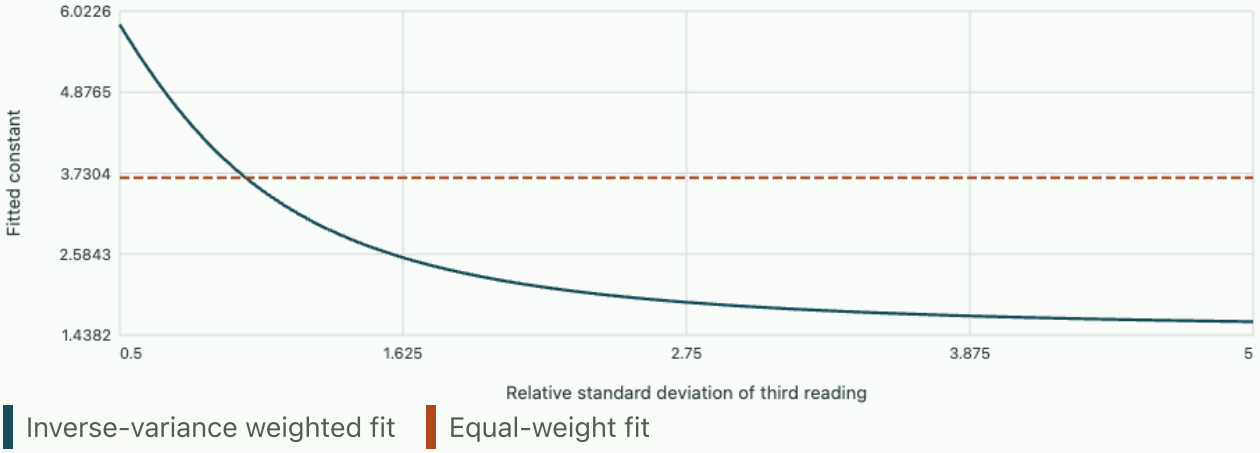
Try this calculation. Change one observation's uncertainty and recompute the weighted projection. More than one fitted coefficient requires the additional basis columns specified in this exercise.

Open this adjustable calculation

Try the calculation

Third reading **8**

Third-reading SD / other SD **2**



Readings are (1,2,8). At the selected uncertainty ratio the fit is 2.2222. The third weight scales as inverse variance, not inverse standard deviation.

Rotate into a better fitting basis

For sample positions $s = 999, 1000, 1001$, consider columns 1 and s. Compute their dot product. Replace the second column by $s - 1000$, compute the new two-by-two Gram matrix, and express the original intercept and slope for a centred fit $5 + 0.2(s - 1000)$. Which intercept most directly represents the experiment?

▼ **Worked solution**

The original column dot product is 3000. The centred positions sum to zero and have squared norm two. The new Gram matrix is diagonal with entries 3 and 2, so the fitted mean and trend separate in the normal equations.

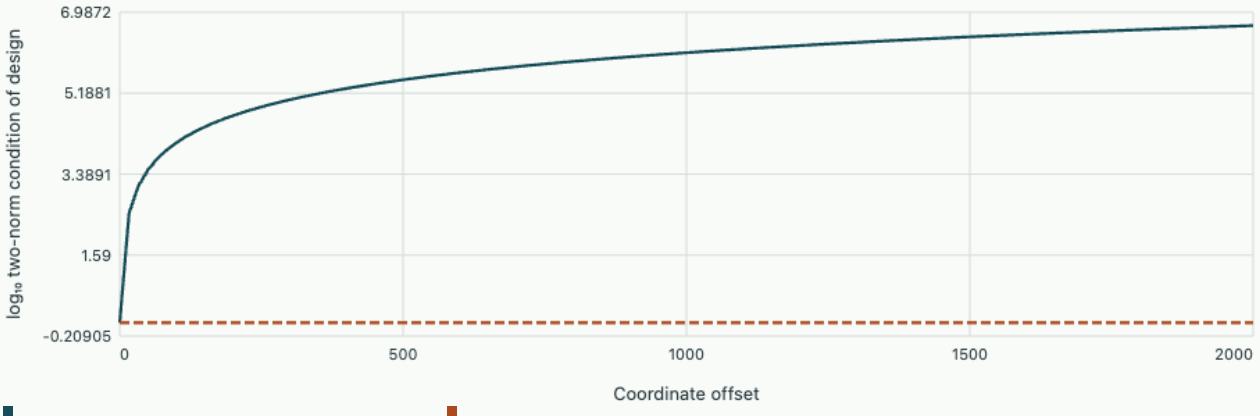
Expanding gives intercept $5 - 200 = -195$ and slope 0.2. The centred intercept is 5 at a measured position. The original intercept extrapolates roughly a thousand position units from the data, so a small slope error can change it considerably.

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

Coordinate offset **1000**



At offset 1000, design condition 1224700. Centring positions (c−1,c,c+1) makes the Gram matrix diag(3,2), preserving the set of fitted straight lines.

Calculate a smoothing factor

On an eight-interval Dirichlet grid take Jacobi error modes $j = 1$ and $j = 7$. Find both amplification factors for ordinary Jacobi and for weighted Jacobi with weight $2/3$. Predict the magnitude of the high-frequency component after three weighted sweeps if its initial amplitude is one.

▼ **Worked solution**

Ordinary Jacobi gives

$$\cos(\pi/8) = 0.92388,$$
$$\cos(7\pi/8) = -0.92388.$$

The weighted factors are

$$1/3 + (2/3)(0.92388) = 0.94925,$$
$$1/3 - (2/3)(0.92388) = -0.28259.$$

Three weighted sweeps leave magnitude $0.28259^3 \simeq 0.0226$ of the high-frequency error. The smooth error retains $0.94925^3 \simeq 0.855$ of its amplitude. A coarse-grid correction should therefore target the smooth component.

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 5 · Krylov subspaces and preconditioning

Derive the first CG steps from a minimum-energy principle and determine what is gained by changing the metric in which a linear system is solved.

FROM FIRST PRINCIPLES

Derive the first Krylov correction from an energy minimum

For symmetric positive definite A, the unique solution of $Ax = b$ minimizes a quadratic energy. For any trial state x, moving along direction p by a scalar α changes that energy according to

$$E(x) = \tfrac{1}{2}x^T Ax - b^T x,$$
$$E(x + \alpha p) - E(x) = -\alpha p^T r + \tfrac{1}{2}\alpha^2 p^T Ap,$$

where $r = b - Ax$ is the residual. Every nonzero direction has $p^T Ap > 0$, so setting the derivative in α to zero gives the unique line minimum. The residual at the updated state is then perpendicular to the search direction,

$$\alpha = \frac{p^T r}{p^T Ap},$$
$$p^T(r - \alpha Ap) = 0.$$

Starting with $p = r$ searches the first Krylov space spanned by the initial residual. Expanding that space to include $Ar, A^2 r, \ldots$ permits a better energy minimum at each step. Conjugate gradients chooses directions mutually orthogonal under the A-weighted inner product so that later line searches preserve the minima already achieved in earlier directions.

DERIVATION AND APPLICATIONS

Determine the decrease delivered by a preconditioned search direction

A preconditioner changes the scale and orientation of the direction used to correct a linear system. To see how that affects the objective, take symmetric positive definite A and preconditioner M, with residual $r = b - Ax$. Solve $Mz = r$ to choose a direction. For nonzero r, the product $r^T z = r^T M^{-1} r$ is positive, so a small positive step along z reduces the quadratic energy $E = \tfrac{1}{2}x^T Ax - b^T x$.

$$E(x + \alpha z) = E(x) - \alpha r^T z + \frac{\alpha^2}{2} z^T Az.$$

The minimum along this direction is at $\alpha_* = (r^T z)/(z^T Az)$. Substitution gives the exact energy decrease. If x_* solves the original system, the full energy difference to the solution is also a quadratic form,

$$E(x) - E(x + \alpha_* z) = \frac{(r^T z)^2}{2 z^T A z},$$
$$E(x) - E(x_*) = \tfrac{1}{2} r^T A^{-1} r.$$

Their ratio is a squared cosine. Let $u = A^{-1/2} r$ and $v = A^{1/2} z$; their ordinary dot product is $r^T z$. The first search removes the fraction

$$\frac{E(x) - E(x + \alpha_* z)}{E(x) - E(x_*)} = \frac{(u^T v)^2}{(u^T u)(v^T v)} \leq 1.$$

A direction proportional to the exact correction removes all the error. A positive scalar multiple of M changes the step length inversely to the direction length, preserving the updated point. These facts concern one search and also apply to the first step of preconditioned conjugate gradients.

WORKED EXAMPLE

Compare two directions on the same system

Take $A = \operatorname{diag}(1, 9)$, $r = (1, 3)^T$ and initial state zero. The available energy decrease is

$$(1 + 9/9)/2 = 1.$$

Without preconditioning $z = r$, so the first step removes $10^2/[2(1 + 81)] = 0.610$. With $M = \operatorname{diag}(1, 3)$, $z = (1, 1)^T$; it removes $4^2/[2(1 + 9)] = 0.800$. Both updates solve the same physical balances.

SECOND APPLICATION

Rescale an exact preconditioner

In a second problem $A = \operatorname{diag}(2, 5)$, $r = (4, 10)^T$ and the available solver uses $M = 2A$. Its direction is $(1, 1)^T$. The minimizing step is $(4 + 10)/(2 + 5) = 2$, giving correction $(2, 2)^T = A^{-1} r$. Judging a preconditioner by whether the step length is close to unity would reject this exact one-step correction.

> **Start from this chapter.** Relax the CG residual tolerance on a fixed grid. Observe the departure from the direct solution, then compare with the shift caused by spatial refinement at tight tolerance.
>
> OneSimulation laboratory · Separate algebraic error from mesh error

A geometrical interpretation of positive definiteness

Assume $A = A^T$ and $v^T A v > 0$ for all nonzero v. The algebraic solution $x_\star$ is the unique minimiser of $J(x) = x^T A x/2 - b^T x$. Indeed, completing the square gives

$$J(x) - J(x_\star) = (x - x_\star)^T A (x - x_\star)/2.$$

For a structural stiffness matrix this has the interpretation of excess elastic potential energy. The negative gradient at an approximation is the residual $r = b - Ax$. Even in a purely algebraic application this energy geometry explains which approximation CG selects.

In a chosen search direction p, minimise $J(x + \alpha p)$ by differentiation. The optimal step is

$$\alpha = p^T r / (p^T A p).$$

Two successive minimisations should not undo each other. If $p_i^T A p_j = 0$ for distinct directions, minimisation along p_j preserves minimisation along p_i. Such directions are called conjugate. CG generates them from successive residuals and matrix-vector products. Starting with $p_0 = r_0$, the exact-arithmetic recurrences are $\alpha_k = (r_k^T r_k)/(p_k^T A p_k)$, $r_{k+1} = r_k - \alpha_k A p_k$,

$$\beta_k = (r_{k+1}^T r_{k+1})/(r_k^T r_k),$$

$$p_{k+1} = r_{k+1} + \beta_k p_k.$$

Work through a two-dimensional example

Take $A = \operatorname{diag}(1, 4)$, $b = (1, 1)^T$, and $x_0 = 0$. Then $p_0 = (1, 1)^T$, $Ap_0 = (1, 4)^T$ and $\alpha_0 = 2/5 = 0.4$. The first approximation is $(0.4, 0.4)^T$; its residual is $(0.6, -0.6)^T$. That residual is perpendicular to the old search direction because the line minimum has already been reached. The update coefficient is $\beta_0 = 0.72/2 = 0.36$, giving $p_1 = (0.96, -0.24)^T$.

Check the conjugacy directly:

$$p_0^T A p_1 = 0.96 - 0.96 = 0.$$

Now $p_1^T A p_1 = 1.152$, so

$$\alpha_1 = 0.72/1.152 = 0.625.$$

The second approximation is $(1, 0.25)^T$, which solves the system. A repeated steepest descent restart would search along the new residual, lose conjugacy, and generally need more than two steps. In floating-point arithmetic, residual orthogonality gradually deteriorates. A fresh residual computed from the final vector is a valuable check.

Improve the geometry of the search

At iteration k, candidates lie in $x_0 + \operatorname{span}\{r_0, Ar_0, \dots, A^{k-1} r_0\}$. The error can be written $p_k(A) e_0$ for a polynomial with $p_k(0) = 1$. If the operator has only a few distinct eigenvalues, a low-degree polynomial

can vanish on them all. More generally, many eigenvalues in a tight cluster can be reduced together. This explains why the condition number gives only a worst-case convergence estimate.

A positive definite preconditioner M corresponds to solving with $M^{-1/2}AM^{-1/2}$. For the diagonal example, $M = \operatorname{diag}(A)$ makes the transformed operator the identity, so one step suffices. For a full matrix, the same choice corrects scales but may leave strong coupling untouched. Applying an approximate inverse also costs time and memory. Compare total solve time, setup cost and physical error over representative loads before choosing a preconditioner. The heated cube experiment reports how changing the tolerance alters the same temperature observable on a fixed mesh. Compare direct and iterative heat balances.

Read an energy-norm convergence estimate

For symmetric positive definite A, a classical CG estimate is $\|e_k\|_A / \|e_0\|_A \le 2[(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)]^k$. To see the square root appear, minimise the largest magnitude of an error polynomial over the entire eigenvalue interval $[\lambda_{\min}, \lambda_{\max}]$, with its value fixed at one at zero. Shifting a Chebyshev polynomial onto that interval produces the stated decay. Treating every point of the interval as a possible eigenvalue makes this a deliberately pessimistic estimate when the actual eigenvalues form tight clusters.

For $\kappa = 100$, the factor is $9/11$. The bound first guarantees a reduction below 10^{-3} after 38 iterations. If effective preconditioning reduces the condition number to 4, the factor becomes $1/3$ and seven iterations suffice for the same guarantee. The comparison gives a budget for the search iterations only. An expensive preconditioner can erase that saving when it is set up for a single small load case. For repeated loads, setup can be spread over many solves, changing the economics without changing the convergence analysis.

Relate the stopping norm to energy

For a symmetric positive definite system the energy error squared equals $r^T A^{-1} r$, since $r = Ae$. Consequently residuals with the same Euclidean length may have different physical significance. With $A = \operatorname{diag}(1, 100)$, residuals $(1,0)^T$ and $(0,1)^T$ both have length one, but their energy errors are 1 and 0.1. A tolerance derived from an energy specification should account for this spectral weighting.

Turn a known spectral interval into a polynomial guarantee

Let A be symmetric positive definite with spectrum in $[a, b]$, $0 < a < b$. After k conjugate-gradient steps from x_0, every available state error has the form $p(A)e_0$, where p has degree at most k and $p(0) = 1$. The constraint at zero follows because the correction is an element of the Krylov space generated by the initial residual Ae_0. Expressing e_0 in orthonormal eigenvectors shows that the squared energy error is $\sum_i \lambda_i |p(\lambda_i)|^2 |e_{0,i}|^2$. Bounding the maximum of the polynomial over the entire spectral interval gives a guarantee independent of the initial error.

Map $[a, b]$ to $[-1, 1]$ through $s = (a + b - 2\lambda)/(b - a)$. The degree-$k$ Chebyshev polynomial $T_k(s)$ has magnitude at most one there, while at $\lambda = 0$ its argument $s_0 = (a + b)/(b - a)$ exceeds one. Thus

$p(\lambda) = T_k(s)/T_k(s_0)$ is an admissible trial polynomial with worst amplitude $1/T_k(s_0)$. Since CG minimises the energy norm over all trial corrections, it performs at least as well as this explicit choice. Positive definiteness is used twice: it defines the energy norm and keeps zero outside the proposed spectral interval.

Compare an interval bound with the actual finite spectrum

Suppose the only eigenvalues are 1, 2 and 4. For two steps take $T_2(s) = 2s^2 - 1$, $s_0 = 5/3$. Then $T_2(s_0) = 41/9$, so the worst relative energy error is at most $9/41 \approx 0.2195$ for the trial polynomial and hence for CG. At the spectral endpoints the polynomial equals $9/41$. At $\lambda = 2$, however, $s = 1/3$ and the factor is $-7/41$. Each mode has acquired a small amplitude even though the signs differ. The actual energy error depends on how much of the initial energy lay in each mode.

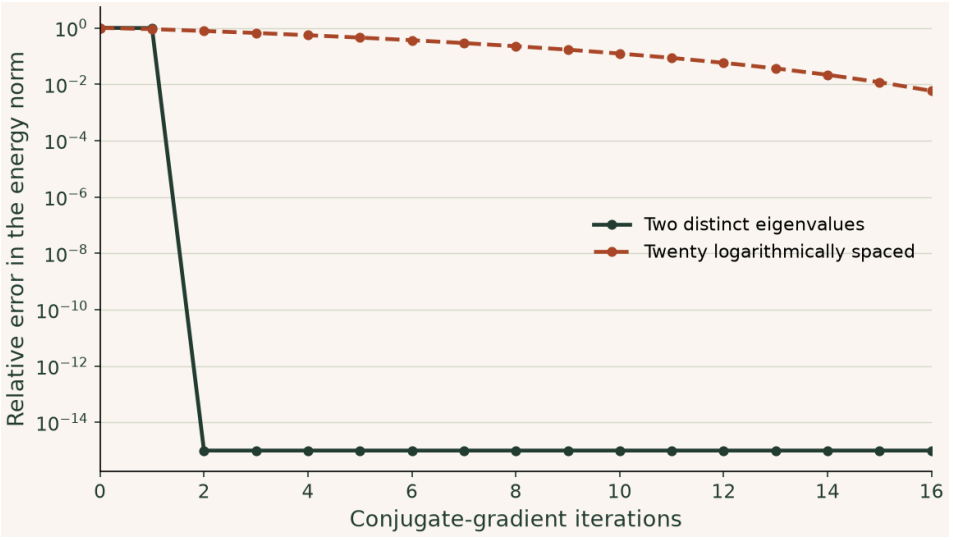
For three steps the spectral interval still has infinitely many possible points. But the actual three eigenvalues permit the polynomial $(1 - \lambda)(1 - \lambda/2)(1 - \lambda/4)$, which equals one at zero and vanishes on the entire spectrum. CG can therefore reach the exact solution on this finite example in at most three steps in exact arithmetic. Repeated eigenvalues do not increase this bound. The interval guarantee loses this information by preparing for every point between the smallest and largest eigenvalues, including values absent from the matrix.

A strong preconditioner can produce a narrow spectral cluster plus a few outliers. A helpful interpretation is to spend a small number of polynomial roots suppressing the outliers and use the rest to approximate zero across the cluster. The number of distinct eigendirections and the distribution of initial energy then complement a simple condition-number estimate. Finite precision can spoil exact finite termination, particularly when supposedly eliminated components reappear through roundoff.

If A is instead indefinite, a polynomial residual argument may still be meaningful for a different Krylov solver, but the quantity $e^T A e$ can vanish for nonzero error or become negative. The preceding minimisation proof then loses the positive measure of approximation on which it depends. Checking symmetry alone is insufficient to apply it.

Same condition number, different Krylov histories

Solve two positive diagonal systems from zero with identical right-hand sides. Both have extreme eigenvalues 1 and 100. Plot the error in $\|e\|_A = \sqrt{e^{\mathsf{T}} A e}$, using the known diagonal inverse as a reference.



For two distinct eigenvalues a degree-two error polynomial can vanish on the whole spectrum. The spread-out spectrum requires many more directions. Values smaller than 10^{-15} are displayed at that floor.

▶ **Read values from the plot**

Change parameters · Same condition number, different Krylov histories

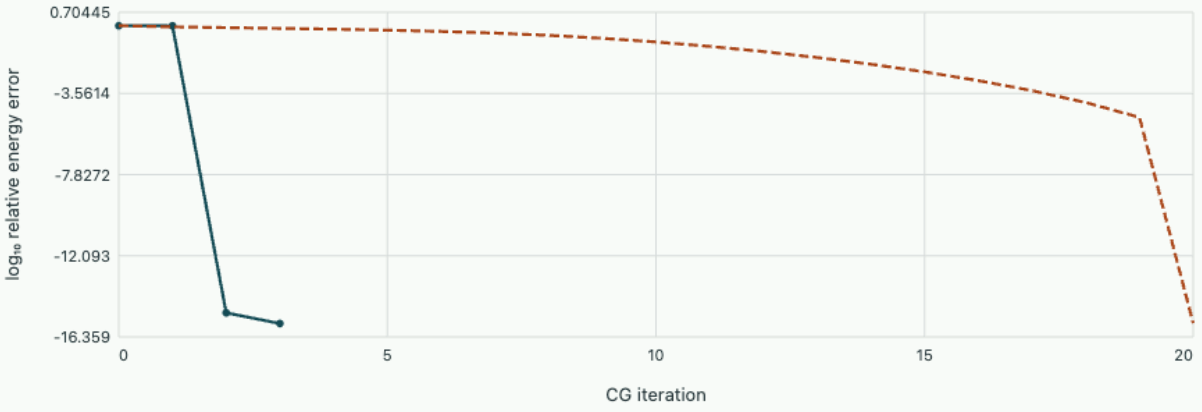
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Extreme eigenvalue ratio **100**

Unknowns **20**



Two distinct eigenvalues Spread-out spectrum

Both matrices have condition number 100. Exact arithmetic needs at most two CG steps for the two-eigenvalue case.

Chapter 6 · Eigenvalues, modes and spectral gaps

Estimate a dominant vibration or diffusion mode, explain the role of eigenvalue separation and use a residual to check a computed eigenpair.

FROM FIRST PRINCIPLES

Watch a power iteration select a mode

Start with a real symmetric matrix A and express a trial vector v_0 in its orthonormal eigenvectors u_j. Order the eigenvalues so $|\lambda_1| > |\lambda_2| \geq \cdots$ and suppose the starting component c_1 is nonzero. Applying A repeatedly multiplies each component by its own eigenvalue,

$$v_0 = \sum_j c_j u_j,$$

$$A^k v_0 = \lambda_1^k \left[c_1 u_1 + \sum_{j>1} c_j (\lambda_j/\lambda_1)^k u_j \right].$$

Normalizing after each application keeps vector lengths manageable without changing the relative size of the components. The largest competing component therefore shrinks relative to the first at a geometric rate. A scalar estimate of the eigenvalue comes from projecting the action back onto the trial vector,

$$\rho(v) = \frac{v^T A v}{v^T v}.$$

For an exact eigenvector the quotient equals its eigenvalue. Equal leading magnitudes can prevent selection of a single direction. A negative leading eigenvalue can make the displayed vector alternate sign even while its one-dimensional direction converges.

DERIVATION AND APPLICATIONS

Derive the eigenvalue shift produced by weak coupling

Two individually measured modes can move apart in frequency or stiffness when they are coupled. Model a real symmetric operator in an orthonormal basis by diagonal values α, β and off-diagonal coupling g. All three numbers must have the same units. A nonzero eigenvector exists only when the determinant of the shifted matrix vanishes,

$$(\alpha - \lambda)(\beta - \lambda) - g^2 = 0.$$

Set $m = (\alpha + \beta)/2$ and $d = (\alpha - \beta)/2$. The quadratic becomes $(\lambda - m)^2 = d^2 + g^2$, so the two eigenvalues and their separation are

$$\lambda_{\pm} = m \pm \sqrt{d^2 + g^2},$$
$$\lambda_{+} - \lambda_{-} = 2\sqrt{d^2 + g^2}.$$

Even when the uncoupled values coincide, a nonzero coupling leaves a gap $2|g|$. For well-separated diagonal entries, assume $\alpha > \beta$ and expand the square root for $|g| \ll \alpha - \beta$. The largest eigenvalue has the shift

$$\lambda_{+} - \alpha = \frac{g^2}{\sqrt{d^2 + g^2} + d} \simeq \frac{g^2}{\alpha - \beta}.$$

The rationalized form gives an accurate small shift without subtracting almost equal eigenvalues. The eigenvalue sum remains exactly $\alpha + \beta$ because the trace is unchanged. A mass-normalized vibration model typically returns squared angular frequencies, so any measured ordinary frequencies must first be converted to the variables of the operator.

WORKED EXAMPLE

Predict a shift away from a crossing

With $\alpha = 10$, $\beta = 6$ and $g = 0.5$, the upper shift is $0.25/(\sqrt{4.25} + 2) = 0.0616$; the leading approximation is 0.0625. The lower eigenvalue is 5.9384 and the upper is 10.0616. The opposite signs of the shifts preserve the trace of 16.

SECOND APPLICATION

Recover coupling and detuning from a measured gap

A second two-mode system has eigenvalues 19 and 25, giving a mean of 22 and half-gap of 3. An independent uncoupled calibration gives $|\alpha - \beta| = 4$, so $d^2 = 4$. The inferred coupling magnitude is

$$\sqrt{3^2 - 2^2} = \sqrt{5} = 2.24.$$

The eigenvalues do not determine the sign of g, which also changes when one basis vector is reversed.

ONESIMULATION LABORATORY

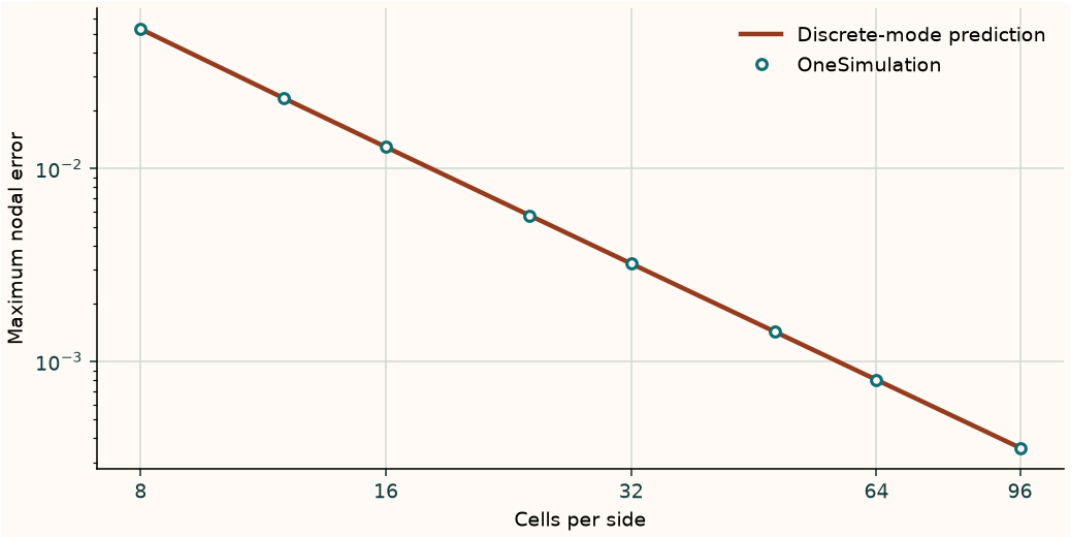
Predict grid error from the wavelength

Can a solver satisfy every discrete balance to roundoff and still overpredict the field? Consider a unit square with zero values on all four sides and a uniform diffusion coefficient. Prescribe the continuous field $u = \sin(ax)\sin(ay)$ with $a = m\pi$ and integer m. Differentiating it twice in each direction gives the source $f = 2a^2u$ for $-\nabla^2 u = f$. This constructs a known solution before any numerical calculation.

Predict the grid's response. For an evenly spaced line with interval $h = 1/N$, add the two neighbouring values of a sampled sine using the angle addition rule. The negative second difference in one direction becomes a multiple of the original sample:

$$\frac{2\sin(ax) - \sin(a(x-h)) - \sin(a(x+h))}{h^2} = \frac{4}{h^2}\sin^2(ah/2)\sin(ax).$$

The square contributes the same factor twice. Its discrete eigenvalue is $\lambda_h = 8\sin^2(ah/2)/h^2$. Since the source was constructed with the continuous eigenvalue, the grid solution is $u_h = (2a^2/\lambda_h)u$ at the nodes. Expanding the sine gives the amplitude ratio $1 + (ah)^2/12 + O((ah)^4)$. At a fixed mode, halving the mesh spacing should reduce the leading amplitude error fourfold.



Maximum nodal error against cells per side on logarithmic axes, with numerical results and a discrete eigenvalue prediction. Half-waves per side = 2.

Download this run (CSV)

Read the comparison. For mode 2, refinement from 24 to 48 cells per side changes the largest nodal error from 0.005731 to 0.001429. The ratio is 4.01. Compare modes at a common mesh size to see why a smooth but shorter-wavelength field can be harder to resolve. The prediction evaluates the maximum on the actual nodes because some grids miss the continuous peaks.

Try a different wavelength. Choose mode 4 on 16, 32 and 64 cells. Compare the output with mode 2 on 8, 16 and 32 cells. These pairs have the same number of cells per half-wave. Predict which error values carry over.

Further reading · Gilbert Strang, Difference matrices and eigenvalues, §5.5 (PDF).

Download the OneSimulation model (Python). Set up the local runtime, then run the downloaded file:

```
python poisson-mode.py --mode 4 --cells 16 32 64
```

What an eigenproblem measures

In an undamped mechanical model $M\ddot{u} + Ku = 0$, insert $u = v\cos(\omega t)$. Nonzero amplitudes require

$$Kv = \omega^2 M v.$$

The normal frequencies are square roots of generalised eigenvalues and the mode shapes are determined only up to scale and sign. For symmetric positive definite M, substituting $w = M^{1/2}v$ gives the ordinary symmetric operator $M^{-1/2}KM^{-1/2}$. This step explains why mass weighting matters when comparing shapes from two different meshes.

For an ordinary symmetric matrix A with orthonormal eigenvectors, a normalised trial vector $v = \sum c_i u_i$ has Rayleigh quotient $\rho = v^T A v = \sum_i \lambda_i c_i^2$. Hence it lies between the smallest and largest eigenvalues. Near an isolated mode its value can be much more accurate than the vector direction: shifting a small amount of amplitude into another orthogonal mode changes the quotient by a term proportional to the square of that amplitude.

Select a mode by repeated application

The power iteration repeatedly forms Av_k and normalises it. Before normalisation,

$$A^k v_0 = \sum c_i \lambda_i^k u_i.$$

If one eigenvalue has strictly largest magnitude and the initial vector overlaps its eigenspace, the ratio of any competing component to the leading one gains a factor $(\lambda_i/\lambda_1)^k$. Normalisation avoids overflow while preserving those ratios. A negative leading eigenvalue can cause sign changes between successive normalised vectors. Comparing their raw difference then misrepresents convergence.

For $A = \operatorname{diag}(4, 2)$ and a starting vector proportional to $(1, 1)^T$, the ratio of the second component to the first is $1/2$ after one multiplication and $1/4$ after two. A vector proportional to $(4, 1)^T$ has Rayleigh quotient $(4 \cdot 16 + 2)/17 = 66/17 = 3.88235$. After four steps the ratio is $1/16$, and the quotient is

$$1026/257 = 3.99222.$$

Check that the quotient error decreases roughly fourfold per additional step, although the directional error decreases only twofold.

Use shifts to find another part of the spectrum

Power iteration selects magnitude. To find a mode near a target σ, apply the inverse of $A - \sigma I$ implicitly by solving a linear system each iteration. The eigenvectors are the same and the eigenvalues are $1/(\lambda_i - \sigma)$. For eigenvalues 2 and 4 with $\sigma = 2.1$, the transformed magnitudes are 10 and about 0.526. The eigenvalue near 2 dominates for a suitable starter. A fixed shift permits factor reuse, but choosing it exactly equal to an eigenvalue makes the shifted solve singular.

Test a candidate pair by forming $r = Av - \rho v$ with $\|v\|_2 = 1$. In the orthonormal basis,

$$\|r\|^2 = \sum_i (\lambda_i - \rho)^2 c_i^2.$$

At least one eigenvalue must lie within $\|r\|$ of ρ. This is a numerical certificate for a symmetric problem. To certify a particular eigenvector, the gap to neighbouring eigenvalues also matters. At a repeated eigenvalue every normalised linear combination of its modes is valid. Comparing subspaces or projecting one span onto the other is then more meaningful than matching individual component signs.

Follow the whole spectrum by QR iteration

To find several eigenvalues of a small dense operator, factor $A_k - \mu I = QR$ and form $A_{k+1} = RQ + \mu I$. Orthogonality gives $A_{k+1} = Q^T A_k Q$, an exact similarity. Hence every step preserves eigenvalues, trace and determinant while changing the coordinates used to represent the operator. This use of QR is different from solving a triangular least-squares system. Repeated unshifted steps amount to an orthogonalised power iteration on multiple directions. A shift near a separated eigenvalue can make the last coordinate nearly invariant and accelerate its separation.

For a symmetric matrix a preliminary sequence of orthogonal similarities reduces it to tridiagonal form. When the coupling between the last two coordinates is negligible relative to the desired accuracy, setting that coupling and its transpose to zero separates a scalar eigenproblem. This removal is called deflation. If the off-diagonal magnitude is e, setting the pair to zero is a symmetric perturbation of spectral norm $|e|$, giving a direct measure for that decision. For a general real nonsymmetric matrix, the converged result can also contain real two-by-two blocks representing complex eigenvalue pairs.

A shifted step with a check on the trailing mode

Take $A = \left(\begin{smallmatrix} 4 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ and $\mu = 2$. Normalising the first shifted column gives $(2, 1)^T/\sqrt{5}$. Complete the orthogonal basis with $(-1, 2)^T/\sqrt{5}$. Multiplying $Q^T A Q$ gives diagonal $(22/5, 8/5)$ and off-diagonal $-1/5$. The trace remains six and the determinant remains seven. The last-coordinate Rayleigh estimate, initially two, is now 1.6 compared with the exact lower eigenvalue $3 - \sqrt{2} \approx 1.586$. The residual of this coordinate has fallen from one to 0.2. A subsequent shift can use the improved estimate. Preserving eigenvalues alone would never certify that a coordinate is ready for deflation.

Practise on sheet 3 · Compare two shifts on the same symmetric operator. Further reading · ETH Zürich · Lecture Notes on Solving Large Scale Eigenvalue Problems · §4.1 The basic QR algorithm and Algorithm 4.4, The Hessenberg QR algorithm with Rayleigh quotient shift.

Which boundary modes are physical?

An elastic assembly suspended in free space admits rigid translations, and in three dimensions also rigid rotations when its geometry allows them. In a correctly assembled model these motions cost zero strain energy, so the stiffness matrix has a nullspace and some generalised eigenvalues are zero. Attempting

unshifted inverse iteration on that stiffness matrix cannot locate a lowest positive vibration by simply inverting the operator. One can instead constrain rigid motions consistently or project into the mass-orthogonal complement of their span.

For a two-mass system with unit masses and a unit spring connecting them,

$$K = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Translation $(1, 1)^T$ has zero squared frequency; opposite displacement $(1, -1)^T$ has squared frequency two. Grounding the first mass eliminates translation but changes the second mode to an oscillator with frequency one. Projection leaves the original free-system vibration at $\sqrt{2}$. These are physically different comparisons; the treatment of a nullspace determines which frequency an eigenvalue calculation is actually reporting.

Convert an eigenpair residual into an angle bound

Given a real symmetric matrix A, a unit-vector approximation v and a scalar θ, write its residual $r = Av - \theta v$. Expand $v = \sum_j c_j q_j$ in an orthonormal eigenbasis. The squared residual is $\sum_j (\lambda_j - \theta)^2 |c_j|^2$. This identity always implies that at least one eigenvalue lies within $\|r\|$ of θ. It does not by itself show whether the reported vector represents one particular physical mode. Two almost coincident eigenvalues can produce a tiny residual for many different mixtures of their eigenvectors.

To bound leakage out of a target eigenspace, let γ be a lower bound on the distance from θ to every eigenvalue outside that space. The squared residual is at least $\gamma^2 \sum_{\text{outside}} |c_j|^2$. Consequently the sine of the angle between v and the target space is at most $\|r\|/\gamma$. This result assumes a symmetric or Hermitian operator under the chosen inner product. For a mass-weighted mechanical problem, transform to mass-normal coordinates or perform the argument in the proper mass inner product. An ordinary Euclidean angle between unscaled coordinates may have little relevance to vibrational energy.

Resolve a cluster more reliably than either member

Take $A = \operatorname{diag}(10, 10.02, 30)$, and $v = (1, 1, 0)^T/\sqrt{2}$. Its Rayleigh quotient is 10.01. Multiplication gives $r = (-0.01, 0.01, 0)^T/\sqrt{2}$, of length 0.01. Relative to the cluster near ten this residual is small, yet the angle from the first coordinate eigenvector is 45 degrees. Using only the other low eigenvalue as a separation gives $\gamma = 0.01$, so the bound on that individual angle has sine at most one and provides no useful orientation guarantee.

Now take the target as the entire first two-dimensional subspace. Its gap from the reported quotient to the outside eigenvalue is

$$30 - 10.01 = 19.99.$$

The same residual implies leakage no greater than $0.01/19.99 \approx 0.000500$. In fact the true leakage of this particular vector is exactly zero. If the measured eigenvalues are squared angular frequencies, the two low modes have nearly the same frequency but can mix strongly under a small asymmetric

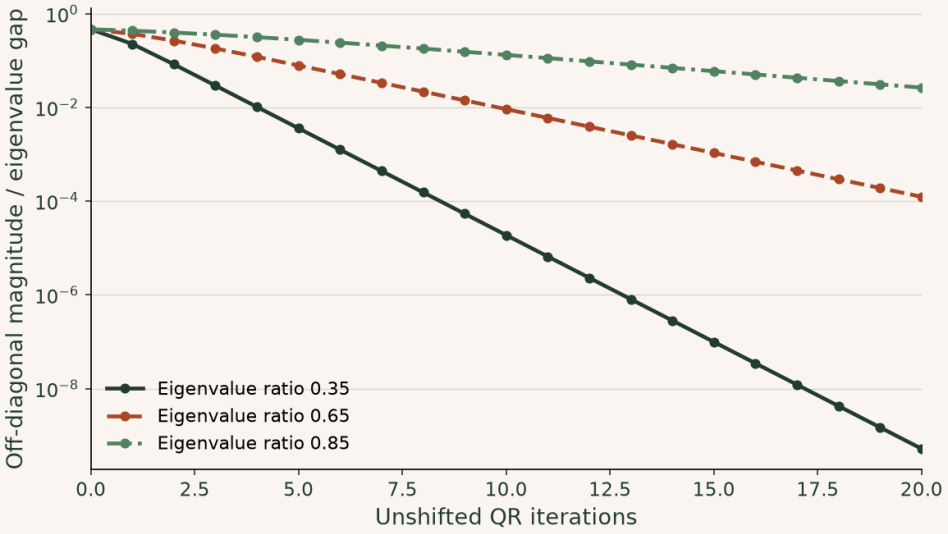
manufacturing perturbation. Their combined subspace may remain very stable and may therefore be the more sensible object to compare with measurements.

A laboratory comparison can project independent excitation patterns onto the computed cluster and compare the total energy in that cluster. Such a result is invariant under rotations of its basis, which a simple table of pairwise eigenvector differences is not. If the gap to the rest of the spectrum closes as well, the subspace estimate becomes weak. Measuring the residual alone cannot repair that missing separation information.

For example the sum of squared projections of the excitation $(1, 0, 0)^T$ onto the first two computed orthonormal modes is one for any orthonormal basis of the exact low cluster. Individual projections can range between zero and one as the basis rotates. A test based on cluster projection assesses the reproducible quantity without requiring a uniquely labelled first mode.

Watch the QR iteration resolve two close eigenvalues

Rotate each diagonal matrix with eigenvalues $1, \rho$ by 0.6 radians, then factor $A_k = Q_k R_k$ and form $A_{k+1} = R_k Q_k$. Each step is an orthogonal similarity. Dividing the off-diagonal by the gap lets all three matrices start with the same normalized mixing.



The tails are straight on a logarithmic vertical axis. Once the dominant direction is nearly aligned, one step multiplies the remaining mixing by approximately the eigenvalue ratio. At a ratio near one, an unchanged spectrum can coexist with slow convergence to a diagonal matrix.

▶ **Read values from the plot**

Change parameters · Watch the QR iteration resolve two close eigenvalues

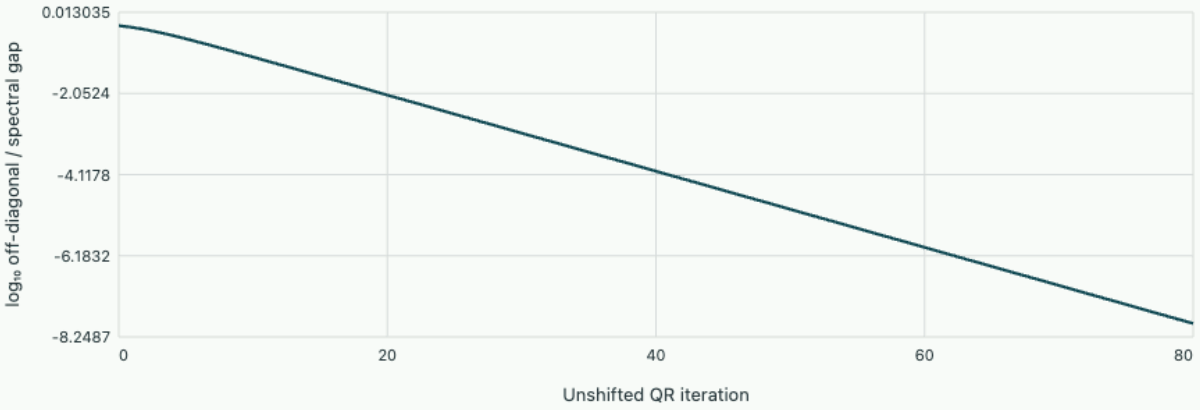
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Smaller eigenvalue ratio ρ **0.8**

Initial rotation (degrees) **35**



Orthogonal-similarity mixing

Asymptotic contraction factor 0.8. The eigenvalues remain 1 and 0.8 throughout the iteration.

Study the full derivation

When stable eigenvalues allow large transient growth · Matrix computations and spectral accuracy

Two damped states exchange energy in only one direction. Its eigenvalues predict ultimate decay, while a short-time response can grow by a large factor. Compute the matrix exponential, its gain and a resolvent-based sensitivity check.

Problem sheet 3 · Search directions and modes

Use trial directions, preconditioners and shifted inverses to control which components of a discrete model survive an iteration.

One exact line minimisation

For $A = \operatorname{diag}(2, 8)$, $b = (2, 2)^T$ and $x_0 = 0$, find the first CG step and residual. Verify residual orthogonality and determine the final exact solution without performing another matrix-vector product.

▼ **Worked solution**

With $p_0 = (2, 2)^T$, $p_0^T r_0 = 8$ and $p_0^T A p_0 = 40$, so $\alpha_0 = 0.2$. Then $x_1 = (0.4, 0.4)^T$ and $r_1 = (1.2, -1.2)^T$. Its dot product with p_0 is zero, as minimisation along that direction requires.

The exact solution follows by division: $x_\star = (1, 0.25)^T$. One remaining conjugate search direction suffices because two distinct eigenvalues occur. From this starting vector both eigendirections are present.

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.

Select a preconditioner

Let $A = \left(\begin{smallmatrix} 100 & 5 \\ 5 & 1 \end{smallmatrix}\right)$. Show that it is positive definite. With the diagonal preconditioner $M = \operatorname{diag}(100, 1)$, form the symmetric transformed matrix and find its condition number. Is diagonal scaling an exact inverse in this case?

▼ **Worked solution**

The leading minors are 100 and 75, both positive. Multiplying left and right by $\operatorname{diag}(0.1, 1)$ gives $\left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$. Its symmetric and antisymmetric modes have eigenvalues 1.5 and 0.5, giving condition number 3.

The scaling removes the difference between diagonal stiffnesses while retaining the off-diagonal coupling. The transformed matrix is distinct from the identity; a general right-hand side can therefore require two CG steps in exact arithmetic.

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.

Locate a vibration mode

A mass-normalised operator has eigenvalues 9, 16 and 25 s⁻². Which natural angular frequencies do they represent? With shift $15\ \mathrm{s}^{-2}$, give the transformed eigenvalues for inverse iteration and identify the

selected mode for a generic initial vector. Finally, specify the units and meaning of a unit-vector Rayleigh residual of 0.01 s⁻².

▼ **Worked solution**

Taking square roots gives angular frequencies 3, 4 and 5 rad/s. The transformed eigenvalues are $1/(9-15) = -1/6$, $1/(16-15) = 1$ and $1/(25-15) = 1/10$ in s². The middle mode has the largest magnitude and dominates if it has nonzero initial overlap.

For a symmetric operator at least one squared frequency lies within 0.01 s⁻² of the Rayleigh quotient. Near $16\ \mathrm{s}^{-2}$, the corresponding frequency uncertainty is about $0.01/(2 \cdot 4) = 0.00125$ rad/s, from differentiating the square root. This last expression is a local linear estimate.

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.

Compare two shifts on the same symmetric operator

For $A = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, compute one unshifted QR step using a positively oriented orthogonal basis. Then use shift one and choose any orthonormal completion if the shifted matrix loses rank. Explain why a singular shifted matrix causes no division by an eigenvalue in this method.

▼ **Worked solution**

The first unshifted column gives

$$q_1 = (2,1)^T/\sqrt{5},\ q_2 = (-1,2)^T/\sqrt{5}.$$

Their Rayleigh values are $14/5$ and $6/5$, and $q_2^T A q_1 = 3/5$. The trailing coordinate still has residual 0.6.

With shift one the first shifted column is $(1,1)^T$. Its normalised direction and the perpendicular direction are already the two eigenvectors. The updated matrix is diagonal with entries 3 and 1, so deflation is exact. A zero last pivot in the QR factor is allowed because the algorithm multiplies factors and never solves with the singular matrix $A - I$.

Explore the related calculation with adjustable inputs. Compare unshifted spectral convergence as the gap narrows; the explicit shifted QR identity in this exercise supplies the separate exact shift check.

Revisit chapter 6 · Follow the whole spectrum by QR iteration.

Bound the effect of stopping a linear solve early

On the common ten-cell mesh the direct and loose-tolerance CG peak temperatures are 0.05708900 and 0.05620782 K. The maximum pointwise difference over all nodes is reported as 8.8×10^{-4} K to two

significant figures. Derive a bound for the difference between maxima even if the two peak nodes differ. Evaluate it here. Does it bound the continuum error?

▼ Work through the solution

For any nodes i,

$$u_i \leq v_i + \|u - v\|_\infty .$$

Maximise both sides, then interchange u, v, to obtain

$$|\max u - \max v| \leq \|u - v\|_\infty .$$

The two listed peaks differ by 0.00088118 K, consistent with the rounded all-node bound. The argument compares solutions sampled on one mesh. It cannot bound the continuous peak error because neither discretisation nor off-node sampling is included.

Use the OneSimulation laboratory to check your answer

Chapter 7 · Householder reflections and practical QR

Build one reflection explicitly, use it to triangularise a rectangular system and locate the residual in transformed coordinates.

FROM FIRST PRINCIPLES

Construct a reflection that zeros a column tail

A reflection reverses the component of a vector perpendicular to a chosen plane and leaves the component in the plane fixed. If a nonzero vector v is normal to the plane, the perpendicular projection of x is $v(v^T x)/(v^T v)$. Subtracting twice that projection defines

$$Hx = x - 2v\frac{v^T x}{v^T v},$$

$$H = I - 2\frac{vv^T}{v^T v}.$$

The reflection preserves lengths because the parallel and perpendicular components remain orthogonal. To reduce a column x, target an axis vector $y = \alpha e_1$ of equal length, so $|\alpha| = \|x\|_2$. Set $v = x - y$. Expanding the dot products gives $2v^T x = v^T v$ because x and y have equal norm. Consequently

$$Hx = x - v = y.$$

Choosing α opposite in sign to x_1 avoids cancellation in v_1. Reflecting only the unprocessed rows of each successive column yields an upper triangular factor while preserving previously generated zeros. An already zero column needs no transformation.

DERIVATION AND APPLICATIONS

Recover the least-squares residual from Householder coordinates

A QR solver can report how much an orthogonal transformation sends into rows below a triangular factor. That part has a precise measurement interpretation. Assume a real matrix A with $m > n$ and independent columns. Apply the same sequence of Householder transformations to the augmented matrix $[A\ b]$. Their product Q^T is orthogonal. Partition its result into an upper block with n rows and a lower block,

$$Q^T[A\ b] = \begin{pmatrix} R & c \\ 0 & d \end{pmatrix}.$$

Because orthogonal transformations preserve the Euclidean norm, the squared residual of any trial vector decomposes into the sum of two nonnegative terms. Only the first can be changed by the fit,

$$\|b - Ax\|_2^2 = \|c - Rx\|_2^2 + \|d\|_2^2.$$

The optimal x solves $Rx = c$, leaving residual norm $\|d\|_2$. In exact arithmetic the original-coordinate residual is obtained by reversing the transforms on a vector whose upper block is zero,

$$r = b - Ax = Q\begin{pmatrix}0\\ d\end{pmatrix},$$

$$A^T r = R^T \begin{bmatrix} I & 0 \end{bmatrix}\begin{pmatrix}0\\ d\end{pmatrix} = 0.$$

The lower coordinates are combinations of observations. Their individual values depend on the particular reflections, whereas their squared norm is invariant. Retaining that norm directly avoids subtracting two large nearly equal squared norms when the fit is excellent. If some triangular pivots are numerically negligible, ranking the columns and checking the model's effective rank become necessary before interpreting all n fitted coefficients.

WORKED EXAMPLE

Inspect residuals after fitting a constant

Fit $A = (1, 1)^T$ to readings $b = (8, 10)^T$. Rotate to the orthonormal directions $(1, 1)/\sqrt{2}$ and $(-1, 1)/\sqrt{2}$. The transformed data are $18/\sqrt{2}$ and $2/\sqrt{2}$, while the transformed model is $(\sqrt{2}, 0)^T x$. The fit gives $x = 9$, residual norm $\sqrt{2}$, and original residual $(-1, 1)^T$. Its sum is zero as the constant column requires.

SECOND APPLICATION

Compare the numerical consequences of subtracting squared norms

Suppose transformed coordinates of another fit have $\|c\| = 10^8$ and $\|d\| = 0.1$. Computing $\|b\|^2 - \|c\|^2$ would subtract two values near 10^{16} to recover 0.01. Binary64 spacing near 10^{16} is 2, so the small addition can disappear entirely. Keeping the transformed tail returns residual norm 0.1 directly, subject to normal transformation errors.

Eliminate a column without stretching it

An orthogonal transformation applied simultaneously to the data matrix and right-hand side leaves every Euclidean residual length unchanged. A Householder reflector is $H = I - 2vv^T/(v^T v)$ for a nonzero vector v. Decompose any vector into components parallel and perpendicular to v. The parallel component changes sign and the perpendicular component stays fixed, proving both length preservation and $H^2 = I$. Implementation requires just a dot product and a scaled vector subtraction rather than a dense matrix.

To make the entries below the first element of a column x vanish, choose $v = x - \alpha e_1$ where $|\alpha| = \|x\|_2$. A short calculation gives $v^T v = 2(\|x\|^2 - \alpha x_1)$ and $v^T x = \|x\|^2 - \alpha x_1$, hence

$$Hx = x - v = \alpha e_1.$$

Choose α with sign opposite to x_1 to avoid subtracting nearly equal positive numbers when computing v_1. A zero column requires a separate decision about rank.

Follow the data into the triangular system

For $x = (3, 4)^T$, use $\alpha = -5$ and $v = (8, 4)^T$. The reflection matrix is $\begin{pmatrix} -.6 & -.8 \\ -.8 & .6 \end{pmatrix}$. Take a one-column least-squares problem with measurements $b = (4, 3)^T$. Multiplication sends x to $(-5, 0)^T$ and b to $(-4.8, -1.4)^T$. The optimum coefficient matches the first transformed measurement, giving $c = 0.96$. The second transformed measurement remains a residual of magnitude 1.4. Back in physical coordinates the predicted vector is $(2.88, 3.84)$, with residual $(1.12, -0.84)$ and the same length.

For several columns, apply the first reflector to the entire remaining block. Construct the next reflector only on rows below the first pivot, protecting the already triangular part. Saving the reflector vectors permits application of Q or Q^T later without explicitly forming either full matrix. Additional right-hand sides can then use the same orthogonal transformations and triangular solve.

Detect when a coefficient is poorly determined

If a new column is nearly in the span of earlier columns, little of it survives below the preceding pivots. A small trailing diagonal of R signals potential coefficient sensitivity, but its scale depends on column units and order. An instrument model with one parameter measured in micro-units can naturally have a small column. Scale the columns according to the intended range of the corresponding parameters before interpreting a near-zero diagonal.

Column-pivoted QR selects a remaining column with a large unrepresented component at each stage. It can help reveal effective rank and identify an independent set of measurements or basis functions. A numerical rank decision still needs a tolerance tied to data uncertainty or floating-point arithmetic. Reporting a fitted coefficient to ten digits after discarding an almost invisible direction would conceal the limitation that the factorisation has revealed.

Check a nearly aligned column

Consider $x = (1, 10^{-8})^T$. Its norm differs from one by only about 5×10^{-17}, below the spacing between one and the next larger double precision number. Computing the exact difference $x_1 - \|x\|$ as that subtraction may therefore give zero. Choosing the target with opposite sign makes the first entry of the reflector vector about two and the remaining small entry is retained. This calculation explains the sign convention geometrically: two opposite points on a sphere provide valid reflection targets, and only one directs the subtraction away from severe cancellation.

Update a triangular fit when observations arrive sequentially

A Householder QR factorisation is efficient when the entire design matrix is available. Suppose instead that a new row arrives after the original observations have been reduced to an upper triangular factor R, a transformed data vector z, and an accumulated residual sum of squares S. The original least-squares objective equals $\|Rx - z\|^2 + S$. Appending measurement row a^T and value b adds $(a^T x - b)^2$. Stacking a^T below R produces a matrix that is triangular except for one row. Orthogonal plane rotations can restore triangular form by eliminating each incoming entry in turn.

To remove a pair $(r, s)^T$, let $\rho = \sqrt{r^2 + s^2}$, $c = r/\rho,\ d = s/\rho$. Multiplication by the rotation with rows (c, d) and $(-d, c)$ gives $(\rho, 0)^T$. Apply precisely the same two-row operation to all remaining columns and to the transformed data. Its final bottom data entry has no corresponding design column, so its square adds to S. If $r = s = 0$ there is nothing to eliminate. A numerical implementation obtains the length with a scaled hypotenuse evaluation to avoid overflow from squaring a large observation.

Insert a point into an existing two-parameter fit

Suppose the existing transformed problem has $R = I,\ z = (1, 2)^T,\ S = 0$. A new sensor reports $x_1 + x_2 = 5$. The best prior estimate was $(1, 2)^T$, which predicts three and therefore disagrees with the incoming reading by two. To reconcile all observations, stack the new row $(1, 1 \mid 5)$ below $(1, 0 \mid 1)$ and $(0, 1 \mid 2)$. Rotate the first and last rows with $c = d = 1/\sqrt{2}$. The first becomes $(\sqrt{2}, 1/\sqrt{2} \mid 6/\sqrt{2})$; the last becomes $(0, 1/\sqrt{2} \mid 4/\sqrt{2})$.

The second rotation combines leading entries 1 and $1/\sqrt{2}$, yielding $\rho = \sqrt{3/2}$, $c = \sqrt{2/3}$, $d = 1/\sqrt{3}$. Its surviving right-hand side is

$$2\sqrt{2/3} + 4/\sqrt{6} = 8/\sqrt{6}.$$

Thus the second-coordinate estimate is

$$(8/\sqrt{6})/\sqrt{3/2} = 8/3.$$

Back substitution in the first row gives

$$x_1 = (6 - 8/3)/2 = 5/3.$$

The final bottom data entry is $-2/\sqrt{3} + 4/\sqrt{3} = 2/\sqrt{3}$, contributing $4/3$ to the residual sum of squares.

Verify directly: the three residual magnitudes are all $2/3$, so their squared sum is $4/3$. The new reading moved both coordinates equally because their original transformed information was identical and its row is symmetric. No historical measurement needed to be reread. Conversely deleting an earlier reading can expose rank deficiency and is a different numerical operation; simply running the same update with a negated data value would add another squared term rather than remove information.

When a stream contains different measurement variances, divide each arriving row and its right-hand side by its independently known standard deviation before applying the rotations. The accumulated S then measures squared standardised discrepancies. If the new reading shares an error with earlier readings, dividing its row alone is insufficient. One must first account for the conditional error and correlated design contribution or refactor a suitable joint whitening transformation. This is a measurement-model limitation, even when the rotations themselves remain perfectly orthogonal.

Repeated readings of a common reference are one practical source of precisely this kind of dependence.

Change parameters · Explore the worked calculation

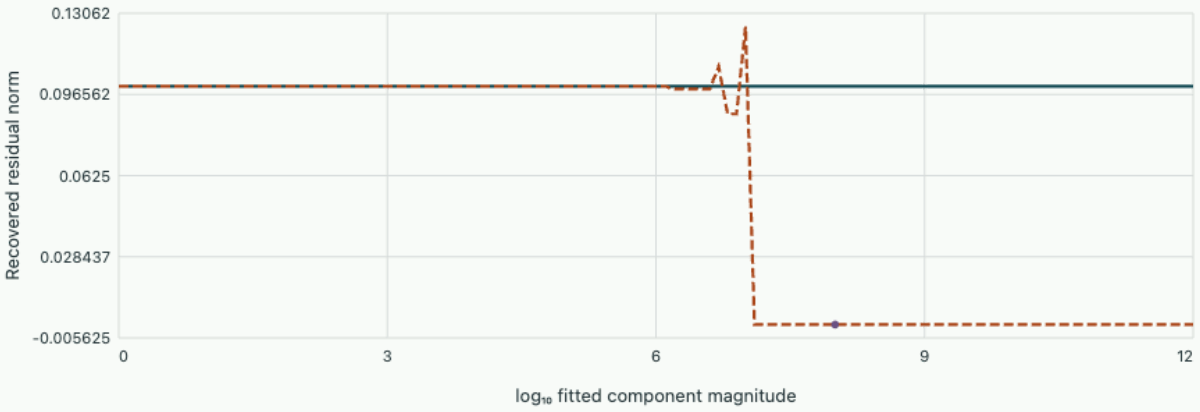
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

$\log_{10}$ fitted component magnitude **8**

Residual-tail magnitude **0.1**



At fitted magnitude 10^8, the subtractive residual is 0 while the preserved tail is 0.1. Binary64 rounding can erase the smaller squared term.

Study the full derivation

Refitting a sensor as observations arrive or disappear · Matrix computations and spectral accuracy

A calibration line receives one new observation at a time. Update its least-squares factorisation using rotations, quantify the influence of a deleted observation and detect when removing a sample destroys identifiability.

Chapter 8 · Singular values and minimum-norm solutions

Resolve a linear measurement into independent directions, compute its pseudoinverse and predict the effect of truncating a weakly observed component.

FROM FIRST PRINCIPLES

Count what a rectangular measurement can recover

For a real matrix A, the squared output length in a unit input direction v is $v^T A^T A v$. Since $A^T A$ is symmetric and nonnegative, it has orthonormal eigenvectors v_j with eigenvalues $\sigma_j^2 \geq 0$. Each positive eigenvalue defines a unit output direction $u_j = Av_j/\sigma_j$. Distinct orthogonal input eigenvectors then have orthogonal output directions because

$$u_i^T u_j = \frac{v_i^T A^T A v_j}{\sigma_i \sigma_j} = \delta_{ij}.$$

Resolve an input into these vectors and the null space. A component along v_j is amplified by σ_j, while every null component disappears. For measured data b, matching the observed output components gives

$$x_{\min} = \sum_{\sigma_j > 0} \frac{u_j^T b}{\sigma_j} v_j.$$

Adding any null vector produces the same prediction but increases the squared norm by its own squared length. This proves the minimum-norm choice. Data perpendicular to all the u_j remain as a least-squares residual, and noise along a small positive singular value can be amplified substantially.

Background · Singular value

$$Av_i = \sigma_i u_i$$

Input and output directions can live in different spaces. A small positive singular value can cause large inverse amplification.

DERIVATION AND APPLICATIONS

Find the smallest reconstruction in physically scaled coordinates

The minimum-norm solution of an underdetermined fit depends on which units and physical scale define the norm. Consider exact independent constraints $Ax = b$ with fewer rows than columns, and let W be a symmetric positive definite matrix of penalties. Minimize $\frac{1}{2} x^T W x$ among all feasible states. A reference

value could first be subtracted if zero represents an inappropriate baseline. Factor $W = L^T L$ and express the optimization in the scaled variable $y = Lx$,

$$Ax = b \quad \Longrightarrow \quad (AL^{-1})y = b,$$
$$x^T W x = \|y\|_2^2.$$

Thus the usual minimum-norm singular-value solution applies to the scaled matrix. An alternative derivation introduces a multiplier λ. Differentiating $\frac{1}{2}x^T W x - \lambda^T (Ax - b)$ gives

$$Wx = A^T \lambda.$$

Substituting in the constraints yields

$$(AW^{-1}A^T)\lambda = b,$$
$$x = W^{-1}A^T(AW^{-1}A^T)^{-1}b.$$

The reduced matrix is positive definite for independent constraint rows. For any feasible perturbation h with $Ah = 0$, the cross term in the objective vanishes:

$$(x+h)^T W(x+h) - x^T W x = 2\lambda^T Ah + h^T W h = h^T W h.$$

This establishes uniqueness and explains the role of the weighting. Directions that data leave unresolved are assigned by a stated metric. Changing units while transforming both A and the metric consistently preserves the physical result.

WORKED EXAMPLE

Distribute a measured total with unequal motion costs

A balance gives $x_1 + x_2 = 10$ and the squared cost is $x_1^2 + 4x_2^2$. Stationarity requires $x_1 = 4x_2$, yielding $(8, 2)$ at cost 80. The unweighted minimum-norm pair $(5, 5)$ satisfies the balance but has weighted cost 125. There is no change in agreement with the measurement.

SECOND APPLICATION

Check invariance under a change of coordinate units

Write a second balance $u + v = 6$ and equal physical cost $u^2 + v^2$. If $w = 100v$, the new row is $(1, 0.01)$ and the cost is $u^2 + 0.0001w^2$. The weighted solution is $u = 3, w = 300$, recovering $v = 3$. Applying an unweighted norm to the rescaled coordinates instead gives $u = 6/(1.0001) \simeq 5.9994$, exposing a unit-dependent extra assumption.

Find directions that a rectangular map measures separately

For real A, the matrix A^TA is symmetric positive semidefinite since $v^TA^TAv = \|Av\|^2$. Choose orthonormal eigenvectors v_i with positive eigenvalues σ_i^2 and define $u_i = Av_i/\sigma_i$. Dotting two such vectors shows that the u_i are orthonormal. Completing both bases gives the singular value decomposition $A = U\Sigma V^T$. Each σ_i is a gain from one parameter direction into a corresponding output direction. Zero singular values indicate directions that no observation in this model can distinguish.

To minimise $\|Ax - b\|_2$, express both vectors in these orthogonal bases. For each positive singular value the residual component vanishes by taking $v_i^Tx = u_i^Tb/\sigma_i$. Output components outside the range remain unavoidable, while input components in the nullspace can take any value without changing the fit. Setting those free components to zero gives the minimum-norm least-squares answer. This derivation identifies the pseudoinverse without assuming either a square matrix or a unique exact solution.

Distinguish an unobserved coordinate from an error

Suppose two instruments read $b_1 = x_1 + x_2$ and $b_2 = 0.02x_3$, giving readings 4 and 0.06. One singular value is $\sqrt{2}$, corresponding to the mean direction of x_1, x_2; another is 0.02 along x_3. The difference $x_1 - x_2$ is unobserved. The minimum-norm exact fit is $(2, 2, 3)^T$, and every $(2 + s, 2 - s, 3)^T$ has identical data. More precision on the mean measurement cannot determine the hidden difference.

A noise change of +0.01 in the second reading raises the reconstructed third parameter by 0.5. If the admissible parameter scale is one, this is a substantial perturbation. Truncating the 0.02 singular direction instead sets $x_3 = 0$. Then the residual grows from zero to the entire second reading, but its noise no longer controls the fitted third coordinate. The decision trades bias against uncertainty; it does not prove that the true third parameter is zero.

Assess a reduced measurement model

Keeping the largest k singular components produces $A_k = \sum_{i \le k} \sigma_i u_i v_i^T$. In the spectral norm its error is σ_{k+1}: expansion in the right singular basis gives this upper bound, and applying the difference to v_{k+1} attains it. Any rank-k alternative must annihilate some unit vector in the span of the first $k + 1$ right singular vectors, on which the original map has norm at least σ_{k+1}. Hence no lower-rank replacement improves the worst-case absolute output error.

That best-approximation result concerns forward predictions in a specified norm. Inverting small singular values magnifies noise, so a good low-rank approximation for forward use does not automatically determine the appropriate inverse for parameter recovery. Work with physically scaled parameters and noise-weighted observations before deciding which singular components are meaningful.

Choose an output before discarding a direction

Suppose an instrument observes x_1 accurately and $0.001x_2$ with uncertainty 0.01. The second variable alone has uncertainty of order ten. The output x_1, however, has no projection onto the poorly observed

direction. Its uncertainty receives no contribution from that amplification. Conversely the output $x_1 + x_2$ inherits the full uncertainty. This distinction can change whether a rank decision is acceptable. The spectrum describes the data acquisition matrix, while the projection of the chosen observable onto right singular vectors determines which losses of information affect the reported result.

Prove that truncation is optimal in two matrix norms

When compressing a dense linear map A, decide which discrepancy is being minimised over all approximations of rank at most k. Its spectral norm measures the largest output error over unit inputs. For any rank-k map B, consider the span of the first $k+1$ right singular vectors of A. Its dimension exceeds the rank of the restriction of B, so there exists a unit vector w in this span satisfying $Bw = 0$. Therefore $\|(A-B)w\| = \|Aw\| \geq \sigma_{k+1}$. Retaining the first k singular pairs attains this lower bound, as its discarded part has largest singular value exactly σ_{k+1}.

For the Frobenius norm, project each column of A into the column space of a candidate B. Orthogonal projection is the closest approximation with columns in that space. If its projector is P, the captured squared norm is $\sum_j \sigma_j^2 \|P u_j\|^2$. The weights $\|P u_j\|^2$ lie between zero and one and sum to at most k. The largest possible weighted sum assigns weight one to the k largest squared singular values. Hence the optimal discarded Frobenius norm is $(\sum_{j>k} \sigma_j^2)^{1/2}$. Both arguments require the norm and coordinate scaling just stated. Entrywise maximum error or application-specific weighted errors generally have different optimisers.

Decide how many responses a reduced map must reproduce

Suppose the four singular values of a unit-scaled instrument model are 8, 3, 2 and 1. A rank-two approximation has best possible worst-case gain error two and Frobenius error $\sqrt{2^2 + 1^2} = \sqrt{5}$. Its relative squared Frobenius loss is $5/(64+9+4+1) = 5/78$, approximately 6.41 percent. Describing this as a loss of just 6.41 percent can be misleading if the experiment deliberately excites the third singular direction. For that unit input, the exact output has length two and the compressed output is zero, giving complete loss of the signal of interest.

Suppose instead inputs have equal independent unit variance in all four right singular coordinates. Then the expected squared error in the output is the sum of the discarded squared gains, exactly five. If the third coordinate is excited with variance one hundred while the others retain variance one, the error from truncating that coordinate alone becomes 400. Under that different usage distribution the meaningful objective is $\|(A-B)C^{1/2}\|_F$, where C is the input covariance.

In these diagonal coordinates the weighted gains become $8, 3, 20, 1$. The best rank-two design for expected output error now retains directions three and one, and the discarded expected squared error becomes

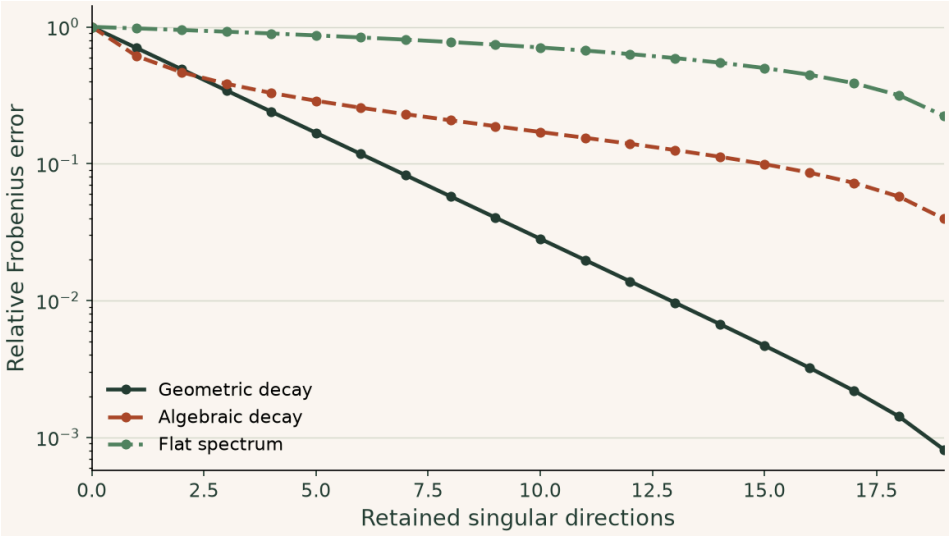
$$3^2 + 1^2 = 10.$$

The ranking changed because the physical question changed. It did not change the optimality theorem for the unweighted matrix; it changed which matrix correctly represents the intended error measure.

When an input covariance is singular, some directions are never excited under the stated design. The weighted objective can be indifferent to large errors in just those directions. A later change of operating regime may activate them, so a reduced model qualified for one input ensemble should retain the conditions under which its ranking was derived.

Decide whether a low-rank approximation is worthwhile

The best rank-k approximation discards the tail of the singular spectrum. Its relative Frobenius error is the square root of discarded squared singular values divided by total squared singular values, with the square root applied to the complete ratio.



For a flat spectrum, retaining ten of twenty directions still leaves an error of $1/\sqrt{2}$. Geometric decay makes the same rank far more effective. The spectral-norm error instead depends only on the largest discarded singular value.

▶ **Read values from the plot**

Change parameters · Decide whether a low-rank approximation is worthwhile

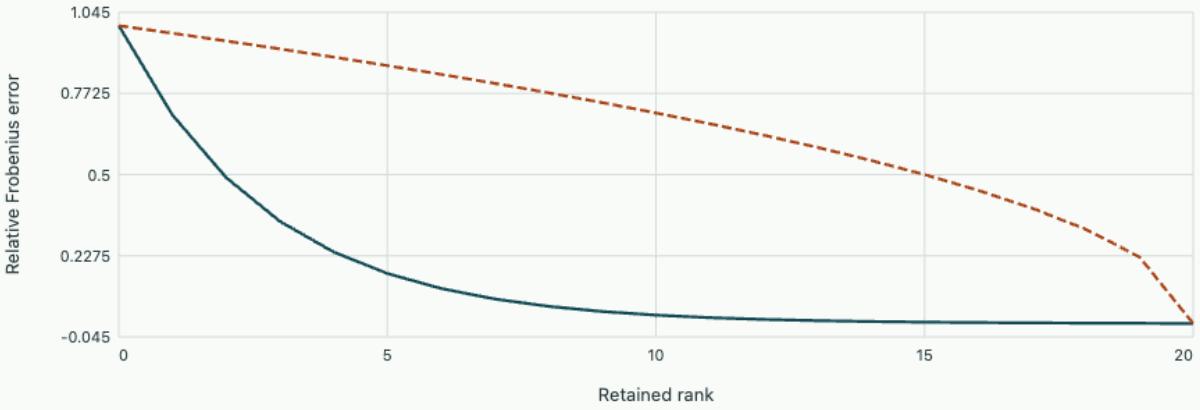
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Singular directions **20**

Geometric decay ratio **0.7**



The best spectral-norm error after rank k is σk+1; this plot instead measures the complete discarded energy.

Study the full derivation

How a random sketch finds a dominant subspace · Matrix computations and spectral accuracy

An instrument returns a matrix too large to factor densely but cheap enough to multiply by a few vectors. Build a low-rank approximation by probing its range, explain what a random draw can miss and design a residual check using fresh probes.

Problem sheet 4 · Orthogonal factorizations

Use geometric reflections and singular directions to separate independent information from redundant measurements.

Build a reflection

A column has leading entries $x = (3, 4)^T$. Using the target $(-5, 0)^T$, construct the reflector, apply it to $y = (0, 5)^T$, and check preservation of the angle between the columns.

▼ **Worked solution**

Take $v = x + 5e_1 = (8, 4)^T$. Then

$$H = I - 2vv^T/80 = \begin{pmatrix} -0.6 & -0.8 \\ -0.8 & 0.6 \end{pmatrix}.$$

It sends x to $(-5, 0)^T$ and y to $(-4, 3)^T$.

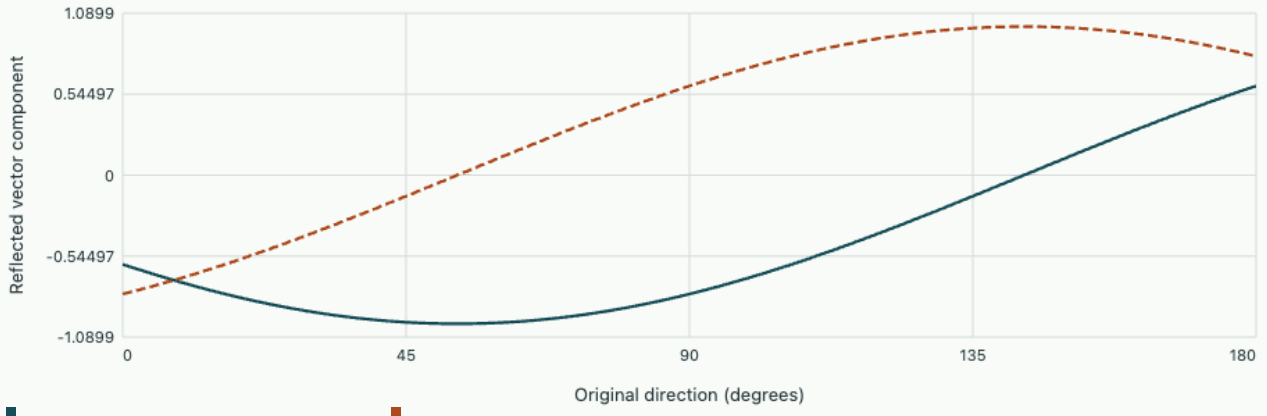
Both lengths stay five. The original dot product is 20 and the transformed dot product is also 20, so the cosine is $20/25 = 0.8$. Retaining the reflector vector suffices to apply the same transformation to another load.

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

Direction of reflected input (degrees) **90**



The reflector sends (3,4) to (−5,0). At the selected unit-vector angle its output is (-0.8, 0.6) with norm one; all angles between pairs are preserved.

Detect a missing direction

A measurement matrix has rows $(1, 1)$ and $(2, 2)$. Compute its singular values. For observations $b = (3, 6)^T$, find every exact solution and the one of smallest Euclidean norm.

▼ **Worked solution**

The Gram matrix has both rows equal to $(5, 5)$, so its eigenvalues are 10 and 0. The singular values are $\sqrt{10}$ and zero. The null direction is $(1, -1)$.

The independent measurement requires $x_1 + x_2 = 3$. Write $x = (1.5 + t, 1.5 - t)^T$; its squared length is $4.5 + 2t^2$, minimised at $t = 0$. Doubling a measurement supplies no independent information about t.

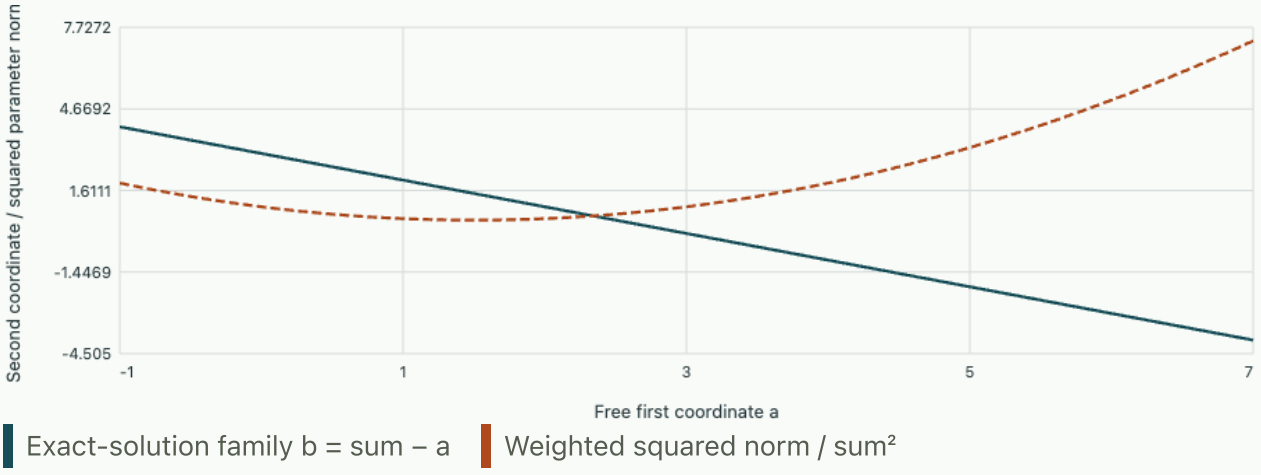
Try this calculation. Explore the continuum of parameter pairs with the same observed sum, and see how a chosen norm selects one pair.

Open this adjustable calculation

Try the calculation

Identified sum a+b **3**

Quadratic weight on second parameter **1**



The minimum weighted norm occurs at (1.5, 1.5). The data identify the sum; the choice within its nullspace comes from the stated parameter metric. The norm curve is normalised by sum².

Decide whether to truncate

Let $A = \mathrm{diag}(4, 0.02)$ and observe $b = (4, 0.03)^T$ with a possible second-component error of 0.02. Find the unregularised estimate, propagate the uncertainty, and compare with a truncated SVD retaining singular values above 0.1.

▼ **Worked solution**

Direct inversion gives $(1, 1.5)^T$; an observational error of 0.02 in the second component propagates to an uncertainty of one unit in x_2. The interval of consistent values is $[0.5, 2.5]$.

Truncating gives $(1, 0)^T$ and residual norm 0.03. Truncation suppresses noise amplification but also discards the second-coordinate signal. The predicted second measurement is now outside the stated uncertainty range, which is evidence against this cutoff for these data.

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 · Graph structure and fill-in

Predict which entries appear when an unknown is eliminated and use graph ordering to compare storage for two equivalent factorisations.

FROM FIRST PRINCIPLES

Derive fill-in from one sparse elimination step

Represent a symmetric matrix as a graph with one vertex per unknown and edges at its nonzero off-diagonal entries. Reorder the first pivot to the top and write its diagonal entry as $a \neq 0$, with couplings c to the remaining vertices and a trailing block B. Solving the pivot equation for its unknown gives

$$ax + c^T y = f,$$
$$x = (f - c^T y)/a.$$

Insert this expression into the remaining equations. The matrix update at a pair of remaining vertices is

$$B_{\text{new}} = B - \frac{cc^T}{a}, \qquad (B_{\text{new}})_{ij} = B_{ij} - \frac{c_i c_j}{a}.$$

If both i and j touch the pivot, a new entry can appear even when $B_{ij} = 0$. This is fill-in. A symbolic elimination graph therefore joins all remaining neighbours before removing the pivot. Exact numerical cancellation can remove a predicted entry, so the graph describes the general pattern. Choosing vertices with fewer or better-connected neighbours can reduce fill and factorization work.

DERIVATION AND APPLICATIONS

Reduce a conducting star and preserve boundary currents

In a resistor network, symbolic fill has an observable magnitude. Consider one interior node of potential u connected to boundary nodes with prescribed potentials v_i through positive conductances g_i. Assume ideal linear resistors and no independent current injected at the interior node. Charge conservation there requires the sum of currents into it to vanish. With $G = \sum_i g_i$, the interior voltage is the conductance-weighted average

$$\sum_i g_i(u - v_i) = 0,$$
$$u = \frac{\sum_i g_i v_i}{G}.$$

The boundary current entering the star at terminal i is $I_i = g_i(v_i - u)$. Substitute the reconstructed u, then collect potential differences relative to the other terminals. The reduced graph contains a physical branch between every pair of formerly separate neighbours,

$$I_i = \sum_{j \neq i} \frac{g_i g_j}{G}(v_i - v_j),$$

$$g_{ij}^{\mathrm{eff}} = \frac{g_i g_j}{G}.$$

Summing the currents gives zero by pairwise cancellation. Multiplying the currents by the corresponding terminal voltages and summing gives the dissipated power, which must agree before and after elimination,

$$\sum_i v_i I_i = \sum_i g_i (v_i - u)^2 = \sum_{i<j} g_{ij}^{\mathrm{eff}} (v_i - v_j)^2.$$

The two graphs can have very different numbers of branches while producing identical boundary measurements. An interior current source would additionally transfer a load vector to the terminals and cannot be dropped in a pure graph reduction.

WORKED EXAMPLE

Reduce a three-terminal network

With conductances $1, 2, 3\,\mathrm{S}$, the new pairwise conductances are $1/3, 1/2, 1\,\mathrm{S}$ for pairs $12, 13, 23$. For terminal voltages $1, 0, 0\,\mathrm{V}$, the original centre sits at $1/6\,\mathrm{V}$. Current from the first terminal is $5/6\,\mathrm{A}$, equal to $(1/3 + 1/2) \times 1\,\mathrm{A}$ in the reduced graph. Both dissipate $5/6\,\mathrm{W}$.

SECOND APPLICATION

Price the fill of a high-degree node

A centre joined to ten isolated boundary nodes by identical $5\,\mathrm{S}$ branches produces $10 \times 9/2 = 45$ reduced branches of $25/50 = 0.5\,\mathrm{S}$. Driving one terminal to $2\,\mathrm{V}$ with the other nine grounded gives current $9(0.5)(2) = 9\,\mathrm{A}$. The unreduced centre is at $0.2\,\mathrm{V}$, giving $5(2 - 0.2) = 9\,\mathrm{A}$ by an independent balance.

Interpret matrix entries as connections

For a symmetric sparse stiffness matrix, put one graph vertex at each unknown and an edge between distinct vertices whose corresponding entry is nonzero. A spring, diffusion face or shared element typically couples nearby unknowns, so the assembled matrix has far fewer entries than a dense array. Gaussian elimination changes this graph. Partition a pivot a, its column below the pivot v and the trailing block B. Eliminating the pivot replaces B by the Schur complement $B - vv^T/a$. For every pair of nonzero entries of v, the outer product can create a coupling between their vertices.

Thus the remaining neighbours of the eliminated vertex can become a clique. These added entries are called fill. Some may cancel for particular coefficients, but symbolic analysis usually retains every position that can be nonzero. It predicts a storage pattern before numerical values are known and can often be reused when material properties or loads change on the same mesh.

Compare two orders on a star

Take four unknowns arranged as a centre joined to three leaves, with no direct leaf-to-leaf connections. Give each diagonal value 4 and each centre-to-leaf coupling –1. Eliminating the centre first subtracts $1/4$ from every leaf diagonal and creates three off-diagonals of $-1/4$ connecting the three pairs of leaves. Although those leaves were initially independent when the centre was specified, eliminating it lets a disturbance at one leaf affect every other.

Instead eliminate one leaf at a time. Each has only one remaining neighbour, so its elimination modifies only the central diagonal and creates no off-diagonal fill. Three eliminations lower the central diagonal from 4 to $4 - 3/4 = 3.25$. The two factorizations give the same physical answer in exact arithmetic, but their intermediate matrices require different storage and work. Relabelling the equations and unknowns together preserves symmetry and changes which order the eliminator encounters.

Use separators on a mesh

A one-dimensional chain eliminated from an end retains its narrow bandwidth. Eliminating interior vertices in an inconvenient order can create distant couplings relative to the original numbering. In a two-dimensional mesh even a careful sweep can leave a long front of mutually coupled vertices. Nested dissection separates the graph into smaller disconnected regions by setting aside a separator, factors the interiors independently and processes that separator later.

There is a practical trade between the cost of constructing an ordering and the savings in repeated factorisations. For a fixed matrix shape with many load cases this overhead can be amortised. For unsymmetric or indefinite systems numerical pivoting may alter the predicted pattern to maintain accuracy. Record both the number of nonzeros in the original operator and the number stored in the factors. Quoting the former alone can severely underestimate memory requirements, particularly for a three-dimensional mesh whose separators contain many more nodes than a one-dimensional chain.

Count the work created by a mesh separator

Ordering influences more than the final number of nonzeros. In a two-dimensional m-by-m grid with nearest-neighbour couplings, remove a cross of order m unknowns and factor the resulting quadrants before the cross. Eliminating each interior communicates among the adjacent separator variables. In a generic factorisation their remaining block is dense: eliminating a separator of size proportional to m then costs order m^3 arithmetic operations and stores order m^2 factor entries. The cost of finding the ordering is separate from this symbolic estimate.

Repeat the argument recursively on the four quadrants of side approximately $m/2$. At level j there are about 4^j independent subdomains with separator side $m/2^j$. Total dense-block storage for this level is proportional to $4^j(m/2^j)^2 = m^2$. There are order $\log m$ levels, yielding storage $O(m^2 \log m)$. Arithmetic at that level is proportional to $4^j(m/2^j)^3 = m^3/2^j$; summing a geometric series gives $O(m^3)$. In terms of the total unknown count $N \simeq m^2$, these are $O(N \log N)$ storage and $O(N^{3/2})$ arithmetic.

Compare memory growth on a sequence of grids

To see why the two resource estimates differ, use an idealised separator count with top separator length 64 and seven levels through length one. Ignore constants and store the square of each separator length. The root contributes $64^2 = 4096$ units; the four children of length 32 contribute $4 \cdot 32^2 = 4096$ as well. Seven levels therefore contribute 28,672 units. For work counted as the cube of each length, the root contributes 262,144 units and its children 131,072. Summing all seven contributions gives

$$262144(1 + 1/2 + \cdots + 1/64) = 520192.$$

Refining each spatial direction by a factor of two produces top length 128 and eight levels. Storage becomes $8 \cdot 128^2 = 131072$, about 4.57 times as much. Work becomes $128^3(2 - 1/128) = 4177920$, about 8.03 times as much. The unknown count grew by four. A budget estimate assuming linear growth in both memory and factorisation time would miss the dominant cost early in this idealised sequence. The constants are structural units, not predictions in bytes or seconds.

These estimates assume balanced geometric partitions, bounded local connectivity and a factorisation that does not introduce unexpected pivot exchanges. Strong global couplings can destroy small separators; in three dimensions the separator of an m^3 grid is a surface of order m^2 unknowns. Substituting surface size into the same argument immediately predicts much heavier top-level work. Before extrapolating timings, inspect actual separator sizes and factor nonzeros for the physical mesh family at hand.

The same symbolic factorisation can often serve many load cases on an unchanged mesh. After numerical factors have been computed, each new pair of triangular substitutions has work proportional to the stored factor entries rather than the cubic separator work. A memory increase can therefore affect every repeated solve even when the one-time factorisation cost is amortised over many sources.

Predict the cost of moving from a sheet to a solid

For a three-dimensional m-by-m-by-m grid, divide a cube into eight smaller cubes separated by coordinate planes. An idealised top separator has order m^2 unknowns rather than order m. At level j the separator storage is proportional to $8^j(m/2^j)^4 = m^4/2^j$, and separator arithmetic to $8^j(m/2^j)^6 = m^6/8^j$. Both level sums are geometrically dominated by the largest separators. In terms of $N = m^3$ unknowns the resulting estimates are $O(N^{4/3})$ storage and $O(N^2)$ factorisation work. The logarithmic storage factor of the two-dimensional sheet has given way to a different power of system size.

Doubling resolution along each axis multiplies N by eight, asymptotic factor storage by roughly sixteen, and factor arithmetic by roughly sixty-four. Suppose the factors of a current three-dimensional problem occupy 3 GB and take 90 seconds to compute. Applying these scalings gives about 48 GB and 96 minutes after such refinement, subject to ordering and machine efficiency. By contrast, applying the existing factors to each extra load scales primarily with factor entries. If memory is limited, the predicted top separator may rule out the planned direct factorisation before the nominal total number of input matrix entries does. An iterative alternative should then be compared at the actual required tolerance and with the cost of its preconditioner included.

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.

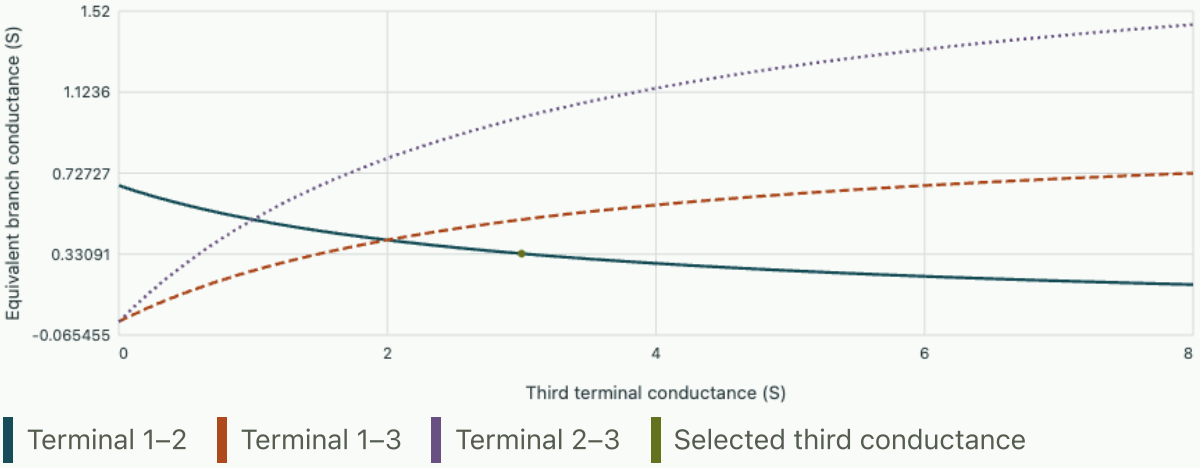
Open this calculation with adjustable inputs

Try the calculation

First terminal conductance (S) 1

Second terminal conductance (S) 2

Selected third conductance (S) 3



At g_3 = 3 S the eliminated 1–2 branch has conductance 0.33333 S. All three couplings come from the same Schur complement.

Study the full derivation

Solve a singular network without inventing a ground current · Matrix computations and spectral accuracy

A floating resistor network has unique voltage drops but no absolute voltage reference. Derive its nullspace, test charge balance, and distinguish adding a computational gauge from changing the physical boundary conditions.

Chapter 10 · Arnoldi iteration and GMRES

Construct an orthonormal basis for a nonsymmetric operator and use a small least-squares problem to choose an approximation.

FROM FIRST PRINCIPLES

Project a large operator onto successive Krylov directions

For $Ax = b$, start with residual $r_0 = b - Ax_0 \neq 0$. Choose the first basis vector as $v_1 = r_0/\beta$, where $\beta = \|r_0\|_2$. Repeated matrix products generate a search space, but their columns can become nearly parallel. At step j, remove previously resolved components of Av_j before normalizing the remainder,

$$h_{ij} = v_i^T A v_j,$$

$$w = Av_j - \sum_{i=1}^{j} h_{ij} v_i,$$

$$h_{j+1,j} = \|w\|_2.$$

If $h_{j+1,j} > 0$, take $v_{j+1} = w/h_{j+1,j}$. The resulting Arnoldi identity $AV_k = V_{k+1}\bar{H}_k$ expresses each large product using a small upper-Hessenberg matrix. A correction $V_k y$ gives the residual norm

$$\|r_0 - AV_k y\|_2 = \|\beta e_1 - \bar{H}_k y\|_2.$$

Orthonormality justifies the equality. GMRES solves the smaller least-squares problem to minimize the actual residual over this search space. When the new norm is zero the subspace is already invariant, and no further independent direction is available through another ordinary Arnoldi step.

Background · Krylov subspace

$$\mathcal{K}_k(A, b) = \operatorname{span}(b, Ab, \ldots, A^{k-1}b)$$

The starting vector determines which parts of the spectrum are visible. Orthogonalisation avoids storing nearly dependent raw powers.

DERIVATION AND APPLICATIONS

Calculate exactly when the first GMRES step stalls

For one Krylov step the only available correction is a multiple of the initial residual $r = b - Ax$. The resulting change in the residual is parallel to Ar. Assume a real nonsingular A and $r \neq 0$, which guarantees $Ar \neq 0$. To choose the coefficient α, expand the squared residual norm,

$$\|r - \alpha Ar\|_2^2 = r^T r - 2\alpha (Ar)^T r + \alpha^2 \|Ar\|_2^2.$$

The derivative vanishes at the coefficient of the orthogonal projection of r onto Ar. Hence the new residual is perpendicular to the direction in which it was allowed to change,

$$\alpha_* = \frac{(Ar)^T r}{\|Ar\|_2^2}, \qquad (Ar)^T (r - \alpha_* Ar) = 0.$$

Let θ denote the angle between r and Ar. At the minimum the squared length removed is exactly the squared length of that projection. The remaining fraction of residual is

$$\frac{\|r_{\text{new}}\|}{\|r\|} = \sqrt{1 - \frac{[(Ar)^T r]^2}{\|Ar\|^2 \|r\|^2}} = |\sin\theta|.$$

Thus equal or opposite directions allow an exact one-step solve, while perpendicular directions deliver no change. This is purely geometric and applies without symmetry. It also clarifies why restarting after one step can be costly: later basis vectors could expose a direction inaccessible on the first line.

WORKED EXAMPLE

Compute the residual reduction for a shear

Use $A = \left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$ and $r = (0, 1)^T$. Then $Ar = (2, 1)^T$, $\alpha_* = 1/5$, and $r_{\text{new}} = (-0.4, 0.8)^T$. Its norm is

$$\sqrt{0.8} = 0.894.$$

The dot product with Ar is $-0.8 + 0.8 = 0$, an independent check that the selected coefficient minimizes the residual along this line.

SECOND APPLICATION

Detect stagnation despite a nonsingular operator

Take $A = \operatorname{diag}(2, -1)$ and $r = (1, \sqrt{2})^T$. The dot product $r^T Ar = 2 - 2 = 0$ forces the first coefficient to zero. A second direction completes a two-dimensional space because $Ar = (2, -\sqrt{2})^T$ is independent of r. In fact $A^{-1} r = (0.5, -\sqrt{2})^T = -0.5 r + 0.5 Ar$. The two-direction correction solves the problem exactly although a one-step restart would repeat the same state.

A residual as a polynomial

Consider a square nonsingular matrix and an initial residual $r_0 = b - Ax_0$. A correction in the span of $r_0, Ar_0, \ldots, A^{k-1} r_0$ is a polynomial in A applied to r_0. After subtracting the action of A, the new residual is $p_k(A) r_0$ with $p_k(0) = 1$. This constraint matters because the zero polynomial would certify

convergence before the calculation began. Generalised minimal residual iteration, or GMRES, minimises the Euclidean norm over all such residuals. The spaces are nested, so the exact-arithmetic residual norms cannot increase during a cycle.

Replacing large products by small coordinates

Start with $v_1 = r_0/\beta$, where $\beta = \|r_0\|_2$. For each new direction form $w = Av_j$, subtract its projections onto the existing orthonormal columns, and normalise what remains. Each projection coefficient occupies one row of column j of a small matrix. The final normalising length occupies its first subdiagonal. This Arnoldi construction gives

$$AV_k = V_{k+1}\bar{H}_k.$$

If the length vanishes, the current subspace is invariant and no further normalisation is possible.

$$\|b - A(x_0 + V_k y)\|_2 = \|V_{k+1}(\beta e_1 - \bar{H}_k y)\|_2 = \|\beta e_1 - \bar{H}_k y\|_2.$$

The second equality uses orthonormality. Solve this small rectangular least-squares problem, for example by updating triangular factors using plane rotations. The resulting residual is orthogonal to the range of AV_k. Repeated projection may be necessary when rounding spoils orthogonality. A freshly calculated physical residual can test the small model's estimate before reporting convergence.

One eigenvalue and two directions

Take $A = \left(\begin{smallmatrix} 1 & 4 \\ 0 & 1 \end{smallmatrix}\right)$, $x_0 = 0$, $b = e_2$. The first correction is αe_2, so the squared residual is $16\alpha^2 + (1-\alpha)^2$. Differentiation gives $\alpha = 1/17$ and length $4/\sqrt{17}$, about 0.9701. Almost the entire initial residual survives despite every eigenvalue being one. The second direction comes from the component of Ae_2 perpendicular to e_2, namely $4e_1$. With both directions available the exact solution $(-4, 1)^T$ is accessible, and its residual vanishes.

For a diagonal matrix equal to the identity, a single step would suffice for every right-hand side. Here a nonzero off-diagonal entry couples two coordinates and the minimal polynomial is $(z-1)^2$. Counting distinct eigenvalues alone misses that coupling.

Paying for a long memory

Each new direction needs storage and projections against earlier directions. Restarting after a fixed number releases memory, but discards those directions. A new cycle still starts from the best current iterate. It lacks the previous cycle's simultaneous freedom to alter its old coefficients. Record matrix-vector products and the residual after each restart when comparing costs. On the example, two directions retained together already achieve the exact answer, providing a direct check on a proposed implementation.

Read a preconditioned residual

Left preconditioning with M^{-1} makes GMRES minimise $\|M^{-1}(b - Ax)\|$ instead. If $M = \mathrm{diag}(1, 1000)$, a physical residual $(0, 1)^T$ has preconditioned norm 0.001. A small number in the stopping display could therefore leave the second physical balance comparatively inaccurate. Right preconditioning solves $AM^{-1}y = b$, then forms $x = M^{-1}y$; for exact preconditioner actions its minimisation still uses the original residual. Compare work and physically meaningful residuals when assessing either choice, especially if the equations use different units.

How restarting can erase a necessary direction

GMRES minimises the full residual norm over its current correction space. A one-step restart retains only the first correction direction generated from the new residual. If that direction makes no angle with the gradient of the residual objective in the appropriate sense, progress can cease even for an invertible, perfectly conditioned matrix. To see the geometry, suppose the current residual is r. The one-dimensional trial residual is $r - \alpha Ar$. Minimising its squared norm for real arithmetic yields

$$\alpha = (Ar)^T r / \|Ar\|^2.$$

Thus no reduction occurs when Ar is perpendicular to r. Restarting immediately recreates precisely the same subproblem.

Unrestarted GMRES can still add Ar and further powers to its Krylov space. In exact arithmetic it can therefore escape one-step stagnation. This distinction matters for storage allocation: reducing the basis length can change the convergence mechanism itself. Merely multiplying the allowed number of restart cycles cannot recover directions if every cycle begins and ends with the same residual. Conversely increasing the restart length helps only if the additional space contains directions that produce a lower residual.

Compare restart lengths for a quarter-turn operator

Set $A = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $b = (1, 0)^T$, beginning at zero. Its transpose times itself is the identity, so its two-norm condition number is one. Yet $Ab = (0, 1)^T$ is perpendicular to the initial residual b. The residual after trial step αb is $(1, -\alpha)^T$, whose squared length is $1 + \alpha^2$. The best step is exactly zero. GMRES restarted every step thus leaves the residual norm at one forever.

With a two-vector basis use corrections $x = \alpha b + \beta Ab = (\alpha, \beta)^T$. The residual is $(1 + \beta, -\alpha)^T$, which vanishes for $\alpha = 0, \beta = -1$. The exact solution is reached. In polynomial language no degree-one real polynomial satisfying $p(0) = 1$ can vanish at both eigenvalues $+i, -i$. The degree-two choice $p(z) = 1 + z^2$ does so because $A^2 = -I$. Both formulations explain the same measurement without attributing the failure to singularity or loss of precision.

For a nonsingular real skew-symmetric matrix the one-step numerator vanishes for every real r, since a scalar $r^T Ar$ equals its own negative after transposition. This family therefore offers a repeatable stress test for short-restart implementations. A preconditioner or a shift that alters the search geometry can

change the result, but any claimed gain should be checked using the residual of the original equations and a comparable count of matrix-vector products.

There is also a simple residual recomputation check at convergence: for the two-step state $(0, -1)^T$, multiplying by the original matrix gives $(1, 0)^T$ exactly. If a code reports the same state after one iteration, it has either used a different search direction or mislabelled its iteration count. Exposing the current basis length alongside the total matrix-vector count makes a restart experiment easier to reproduce.

For complex arithmetic, use the real or complex least-squares minimisation appropriate to the actual correction space; the perpendicularity argument above assumes real data.

Change parameters · Explore the worked calculation

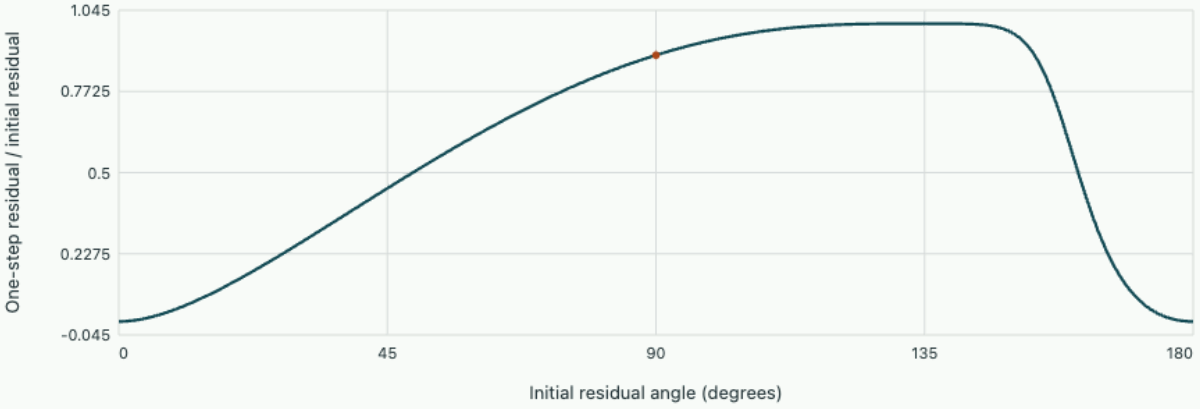
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Shear coupling **2**

Initial residual angle (degrees) **90**



One-dimensional GMRES correction | Selected residual

The shear matrix has both eigenvalues equal to one. Selected first-step residual factor 0.89443. At most two exact-arithmetic directions solve this nonsingular two-variable problem.

Study the full derivation

When stable eigenvalues allow large transient growth · Matrix computations and spectral accuracy

Two damped states exchange energy in only one direction. Its eigenvalues predict ultimate decay, while a short-time response can grow by a large factor. Compute the matrix exponential, its gain and a resolvent-based sensitivity check.

Problem sheet 5 · Graphs and search spaces

Predict elimination fill and compute minimum-residual updates for two systems with identical eigenvalues.

Choose an elimination order

A four-vertex star has centre 1, joined only to leaves 2, 3 and 4. Its symmetric positive definite matrix has generic nonzero edge weights. Count newly introduced edges if the centre goes first. Give an ordering avoiding fill and explain why it works.

▼ **Worked solution**

Eliminating 1 connects each pair of its surviving neighbours. The three unordered pairs are (2,3), (2,4) and (3,4), none previously present. They occupy three new positions in one triangular factor's remaining graph.

In order (2,3,4,1) each eliminated leaf has a single surviving neighbour. There is no pair to connect, so no fill occurs. Positive definiteness ensures diagonal pivots exist without changing this structural ordering.

Explore the related calculation with adjustable inputs. Inspect the effective couplings obtained by eliminating an interior variable. The same substitution generates the Schur operator in the stated constrained system.

Two minimum-residual steps

Take $A = \left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$, $b = e_2$ and $x_0 = 0$. Compute the first GMRES iterate and its residual norm. Show that the second unrestarted step can give zero residual.

▼ **Worked solution**

Along $x = \alpha e_2$ the squared residual is $4\alpha^2 + (1-\alpha)^2$. Its derivative is $10\alpha - 2$, so $x_1 = e_2/5$ and $r_1 = (-2/5, 4/5)^T$. Its norm is $2/\sqrt{5}$.

The directions $e_2, Ae_2 = (2,1)^T$ span the plane. Their combination can realise $x_\star = (-2,1)^T$; multiplication gives

$$Ax_\star = (0,1)^T.$$

Thus the second minimum is exactly zero.

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.

Does the load matter?

Reuse the preceding matrix with $b = e_1$. Determine the Krylov dimension and first residual. Generalise your result to loads $(c, 0)^T$, and explain the discrepancy with the previous problem.

▼ Worked solution

Now $Ae_1 = e_1$, so every power of A produces the same direction. Choosing $x = e_1$ solves the system in one step. For $c \neq 0$, the dimension is one and $x = ce_1$ already solves it. A zero load has zero initial residual.

The earlier load occupied the generalised eigenvector direction, so Ae_2 added an independent component. Matrix size and eigenvalues agree in both tests, but the minimal polynomial relative to the starting residual changes.

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 11 · Iterative refinement and backward error

Recover accuracy from an approximate factorisation and determine when a small residual genuinely certifies a linear solve.

FROM FIRST PRINCIPLES

Derive a correction using the observed imbalance

For a computed solution $\widehat{x}$ of $Ax = b$, evaluate $r = b - A\widehat{x}$. Assume A is nonsingular. Subtract the equation for $\widehat{x}$ from the exact equation to identify the missing correction,

$$A(x_* - \widehat{x}) = r.$$

If the matrix factors are already available, a correction solve uses those factors at much lower cost than a fresh factorization. In floating point suppose the solve acts like an exact solve with a matrix M close to A. Ignoring rounding in the residual and final addition for the moment, the update gives

$$x_{k+1} = x_k + M^{-1} r_k,$$
$$e_{k+1} = (I - M^{-1}A)e_k,$$

where $e_k = x_k - x_*$. A norm of the error propagation matrix below one gives a contraction. Accurate residual evaluation matters because its own rounding error becomes new correction data. A tiny change in the iterate can mean either that the equations are satisfied or that small corrections are being rounded away. The scaled residual distinguishes these cases.

Background · Residual

$$r = b - A\hat{x}$$

A small residual tests the computed equation. Converting it into an error estimate needs information on the inverse or the troublesome singular directions.

DERIVATION AND APPLICATIONS

Show why a componentwise backward error survives a change of units

Equations and unknowns often have very different numerical scales. A componentwise backward error measures how much each coefficient and measured load must change relative to its own magnitude to make a computed state exact. For fixed $\widehat{x}$, ask for perturbations satisfying $(A + \Delta A)\widehat{x} = b + \Delta b$ with $|\Delta A| \le \eta |A|$ and $|\Delta b| \le \eta |b|$, entrywise. The residual then obeys

$$r = b - A\widehat{x} = \Delta A\,\widehat{x} - \Delta b, \qquad |r_i| \le \eta\big(|A||\widehat{x}| + |b|\big)_i.$$

Each row can attain any imbalance between these two extremes by choosing signs for its coefficient and load perturbations and scaling them together. Therefore the smallest admissible common relative tolerance is

$$\eta_* = \max_i \frac{|r_i|}{(|A||\widehat{x}| + |b|)_i}.$$

A zero denominator means that row has zero imbalance for the specified state; assign its ratio zero. Now change equation units with a positive diagonal D and coordinate scales with another positive diagonal S. Use $A' = DAS,\ b' = Db,\ \widehat{x}' = S^{-1}\widehat{x}$. Direct substitution gives

$$r' = Dr,$$
$$|A'||\widehat{x}'| + |b'| = D\big(|A||\widehat{x}| + |b|\big).$$

Every row ratio is unchanged. The plain norm of a residual can instead change by orders of magnitude under exactly the same re-expression. A small componentwise error means a nearby model fits; the forward uncertainty also depends on the conditioning of that model.

WORKED EXAMPLE

Compare two differently scaled balances

For $A = \operatorname{diag}(2, 1000)$, $b = (2, 1000)^T$, and $\widehat{x} = (1.01, 0.999)^T$, the residual is $(-0.020, 1)^T$. Row ratios are

$$0.020/4.020 = 0.00498,$$
$$1/1999 = 0.000500.$$

The first equation sets the backward error despite the larger absolute residual in the second.

SECOND APPLICATION

Change coordinate and equation units simultaneously

For a scalar balance $3x = 12$, test $\widehat{x} = 3.9$. Its backward error is

$$0.3/(11.7 + 12) = 0.01266.$$

Express $x = 100x'$ and multiply the equation by 0.01. The transformed equation is $3x' = 0.12$, with $\widehat{x}' = 0.039$, residual 0.003, and denominator 0.237. Their ratio again gives 0.01266, verifying invariance directly.

A correction equation

Suppose a large matrix has already been factored at modest precision. Let x be the resulting approximation and recompute $r = b - Ax$ using the original coefficients. The exact missing correction solves $Ad = r$. Reusing the approximate factors is cheap because the expensive elimination has finished. If their effective matrix is M, one refinement step solves $Md = r$ and updates $x \leftarrow x + d$. Forming the residual from the factors themselves could erase the very discrepancy being measured.

With $e = x_\star - x$, the residual equals Ae in exact arithmetic. Substitution yields

$$e_{\mathrm{new}} = (I - M^{-1}A)e.$$

Thus a sufficient convergence condition in a chosen induced norm is $\|I - M^{-1}A\| < 1$. The size of the initial error affects how many corrections are needed, while the quality of the approximate inverse controls the contraction. The availability of more accurate residual arithmetic matters once cancellation in $b - Ax$ competes with the true defect.

Diagonal model for imperfect factors

Take $A = \mathrm{diag}(3, 7)$, $b = (3, 7)^T$, and use $M = \mathrm{diag}(3, 8)$ throughout. The first approximate solve gives

$$x_0 = (1, 0.875)^T.$$

Its residual is $(0, 0.875)^T$, and the inexpensive correction is $(0, 0.109375)^T$. Hence the second component becomes 0.984375. The corresponding errors in that coordinate decrease from $1/8$ to $1/64$. Each further update divides the error by eight because $1 - 7/8 = 1/8$. The exact first component stays fixed.

If the incorrect factor were 3 instead of 8 in the second equation, its error multiplier would be $1 - 7/3 = -4/3$. Iterating would alternate across the solution with increasing amplitude. Extra residual accuracy could reveal that failure promptly but could never turn this correction matrix into a contraction.

Asking which equation was solved

A backward error asks for the smallest perturbation of the input for which the computed vector is exact. For relative, componentwise changes to both coefficients and load, $|\Delta A| \le \eta |A|$ and $|\Delta b| \le \eta |b|$, the rowwise correction must absorb r_i. The triangle inequality bounds its capacity by $\eta(|A||x| + |b|)_i$. Choosing perturbation signs independently within each row attains this capacity, giving

$$\eta = \max_i \frac{|r_i|}{(|A||x| + |b|)_i}.$$

A row with zero denominator and zero residual contributes zero. In the example's second row the initial backward error is $0.875/(7 \cdot 0.875 + 7) = 1/15$, even though its forward error is one eighth of the true

component. After one correction the denominator is 13.890625 and the residual is 0.109375. The resulting backward error is $1/127$. Uniformly changing that row's units leaves both ratios unchanged.

A stopping decision

Record the residual obtained using the original data and monitor the size of each accepted correction. Once the residual is comparable to accumulated rounding in its evaluation, more corrections may stagnate. A small backward error then certifies compatibility with nearly identical equations. Converting it to a forward accuracy statement also needs sensitivity information about those equations and the particular output of interest.

Separate data accuracy from solve accuracy

Suppose the second load in the diagonal example was measured as 7.00 ± 0.07. Even an exact linear solve then gives a second coordinate known only to about 0.01. The initial solve error of 0.125 matters relative to that scale. Two corrections already reduce this error to about 0.00195. Further corrections improve the numerical solution of the nominal input but cannot resolve variation in the measured load. Stating both uncertainties clarifies the practical value of the last digits.

Track residual error through the correction equation

Let the approximate correction solve use a nonsingular matrix M. Suppose residual evaluation returns $\hat{r} = Ae + \delta r$, where δr includes accumulated rounding, and the state update has an additional absolute error δx. One correction yields $e_{\mathrm{new}} = (I - M^{-1}A)e - M^{-1}\delta r - \delta x$ if $e = x_\star - x$. A contraction $\|I - M^{-1}A\| \le q < 1$ shrinks the first term, but every iteration injects new errors. If their combined effect is bounded by d, induction gives

$$\|e_k\| \le q^k \|e_0\| + d(1 - q^k)/(1 - q).$$

This inequality separates an initial geometric improvement from a possible long-run floor. A residual-evaluation bound must use scales of the terms being subtracted, such as $|b| + |A||x|$, rather than the size of the tiny mathematical residual alone. Also distinguish errors in evaluating the residual from the final rounding of $x + d$: very small corrections may be accurate but cannot change a stored state if they fall below its local spacing. Computing residuals in higher precision helps only when that residual error is a limiting contribution.

Resolve a correction lost in a cancellation

Consider the decimal thought experiment $1000x_1 + x_2 = 1001,\ x_1 + x_2 = 2$, whose exact state is $(1, 1)^T$. Store the candidate $(1.00, 0.990)^T$ and evaluate terms using three significant decimal digits with rounding to nearest. Even if the original input coefficients and load are kept exactly, evaluating the first row naively makes $1000 + 0.990$ round to 1.00×10^3; the rounded representation of the right-hand side

1001 is also 1.00×10^3. A residual assembled entirely in that low precision therefore sees a first component of zero. The second component, $2 - 1.99 = 0.01$, remains visible.

Computing the residual using the original data with adequate precision gives $(0.01, 0.01)^T$. Subtracting these two correction equations yields $999 d_1 = 0$, hence $d_1 = 0,\ d_2 = 0.01$, recovering the exact state in one exact correction solve. If instead the low-precision residual $(0, 0.01)^T$ is used with the same exact correction matrix, subtraction yields $999 d_1 = -0.01$, giving $d_1 \approx -0.00001001,\ d_2 \approx 0.01001001$. Part of the estimated correction is now allocated to a coordinate that was already exact.

The example uses an exact correction solve to isolate the residual effect. Low-precision factors introduce a separate perturbation. If 1001 had first been irreversibly stored as 1000 in the only surviving copy of the problem, later high-precision arithmetic could not restore the intended equation. Preserving the original coefficients and loads is therefore part of the method whenever refinement is expected to recover accuracy relative to higher-precision input data.

One can estimate which digit was lost without rerunning the solve. The absolute first-row summands have total magnitude about one thousand, while their desired residual is one hundredth. The computation is extracting a quantity roughly one hundred thousand times smaller than its large terms. By comparison the corresponding factor in the second row is only a few hundred. Precision can therefore be inadequate in one equation long before it limits the others.

Change parameters · Explore the worked calculation

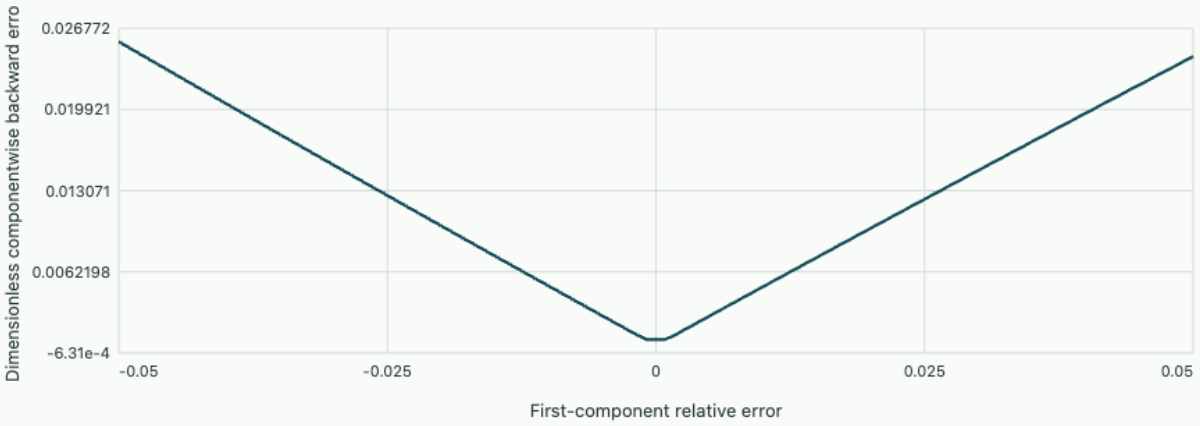
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Second equation scale **500**

Second-component relative error **-0.001**



Maximum componentwise backward error

Second raw residual magnitude 0.5, second row backward error 0.00050025. The scale cancels from the row-normalised diagnostic although the raw residual changes.

Chapter 12 · Eliminating constraints with Schur complements

Eliminate auxiliary variables, recover constraint forces and account for the sign and conditioning of a reduced system.

FROM FIRST PRINCIPLES

Eliminate one block of unknowns before solving the rest

A coupled model can have different types of unknowns, such as internal and boundary displacements. Write its linear system in block form with a nonsingular first block K. The first row expresses one group in terms of the other,

$$Ku + Cv = f,$$
$$u = K^{-1}(f - Cv).$$

Substitute into the remaining balance $Bu + Dv = g$. Collect all terms that still depend on v to obtain a smaller system and a recovery step,

$$(D - BK^{-1}C)v = g - BK^{-1}f,$$
$$Ku = f - Cv.$$

The new coefficient matrix is the Schur complement of K. The term in the right-hand side is needed even when no external load acts directly on the retained variables. In practice, notation such as $K^{-1}C$ denotes solves with columns of C or an operator action; explicit inversion is unnecessary. If the full matrix and K are both nonsingular, so is the Schur complement. Very small reduced eigenvalues can signal a weakly constrained remaining quantity.

Background · Schur complement

$$S = D - CA^{-1}B$$

Eliminate the first block of a partitioned system by solving with A. Forming an explicit inverse is unnecessary and may discard useful structure.

DERIVATION AND APPLICATIONS

Update a solved system after adding one spring or penalty

Suppose factors of a symmetric positive definite matrix A already exist and a small model change adds one positive stiffness k acting on the relative coordinate $u^T x$. The added energy is $\frac{1}{2}k(u^T x)^2$, whose

gradient adds $ku(u^T x)$ to the original balance. Introduce its scalar force $t = ku^T x$. Solving the old matrix with the new internal force gives

$$(A + kuu^T)x = b,$$
$$x = x_0 - zt,$$
$$Ax_0 = b,$$
$$Az = u.$$

Use the constitutive equation for t to eliminate the full state and solve a one-dimensional Schur complement. The scalar $u^T z = u^T A^{-1} u$ is positive for nonzero u. Hence the denominator is safely nonzero for a positive added stiffness,

$$(k^{-1} + u^T z)t = u^T x_0,$$
$$x = x_0 - z\frac{k\,u^T x_0}{1 + k\,u^T z}.$$

The new relative displacement follows by applying u^T again,

$$u^T x = \frac{u^T x_0}{1 + k\,u^T A^{-1} u}.$$

This explicitly shows how the added stiffness reduces extension. If the old extension is zero, adding the spring changes neither the old force balance nor the solution. As k becomes arbitrarily large the extension tends to zero while the internal force approaches a finite constraint reaction. The original factors can serve many choices of stiffness when the location u stays fixed.

WORKED EXAMPLE

Retrofit a link between two supported nodes

Take $A = \operatorname{diag}(2, 4)\,\mathrm{N/m}$, load $b = (2, 0)^T\,\mathrm{N}$, and $u = (1, -1)^T$. Initially $x_0 = (1, 0)^T\,\mathrm{m}$; the unit response $z = (0.5, -0.25)^T\,\mathrm{m/N}$ gives

$$u^T z = 0.75\,\mathrm{m/N}.$$

Adding $k = 4\,\mathrm{N/m}$ gives $t = 1\,\mathrm{N}$, $x = (0.5, 0.25)^T\,\mathrm{m}$ and relative extension $0.25\,\mathrm{m}$.

SECOND APPLICATION

Determine the stiffness needed for a target extension

A different model gives old relative displacement $6.00\,\mathrm{mm}$ and relative compliance $0.0200\,\mathrm{m/N}$. Cutting the extension to $2.00\,\mathrm{mm}$ requires $1 + k(0.0200) = 3$, hence $k = 100\,\mathrm{N/m}$. The new force is $0.200\,\mathrm{N}$, producing a compliance displacement $0.0200(0.200) = 4.00\,\mathrm{mm}$, exactly the required reduction.

Two coupled sets of equations

A constrained equilibrium may contain displacements u and reaction coefficients λ. Suppose the energy Hessian K is symmetric positive definite and the constraint matrix B has independent rows. With applied force f and prescribed values g, stationarity of $u^T K u/2 - f^T u + \lambda^T (Bu - g)$ gives $Ku + B^T \lambda = f$ and $Bu = g$. The full symmetric matrix is indefinite. Minimising the quadratic energy under the constraints still gives a unique displacement because the energy curves upward along every permissible direction.

Deriving a reduced operator

Solve the first equation formally for $u = K^{-1}(f - B^T \lambda)$. Inserting this expression into the constraint gives a smaller positive system,

$$S\lambda = BK^{-1} f - g,$$

$$S = BK^{-1} B^T.$$

The literal lower diagonal block left by elimination is $-S$. Multiplying both sides by minus one produces the positive form written above. For any nonzero vector z, full row rank implies $B^T z \neq 0$, hence

$$z^T S z = (B^T z)^T K^{-1} (B^T z) > 0.$$

This proof explains why redundant constraints create a singular reduced matrix. Even in that case the physical displacement can remain unique while reactions assigned to duplicate equations become ambiguous.

OneSimulation · How stiff must a link be to act like a constraint?

Two grounded springs with stiffnesses 2 and 8 N/mm carry the force vector $(10, 0)^T$ N. Before coupling them, the gap between their displacements is 5 mm. A positive link reaction λ reduces that gap according to $g = g_0 - S\lambda$. The scalar Schur complement $S = BK^{-1}B^T = 0.625$ mm/N measures the gap closed by a unit reaction. A real link of stiffness k_c obeys $g = \lambda / k_c$. Eliminating the gap gives

$$\lambda = \frac{g_0}{S + 1/k_c}.$$

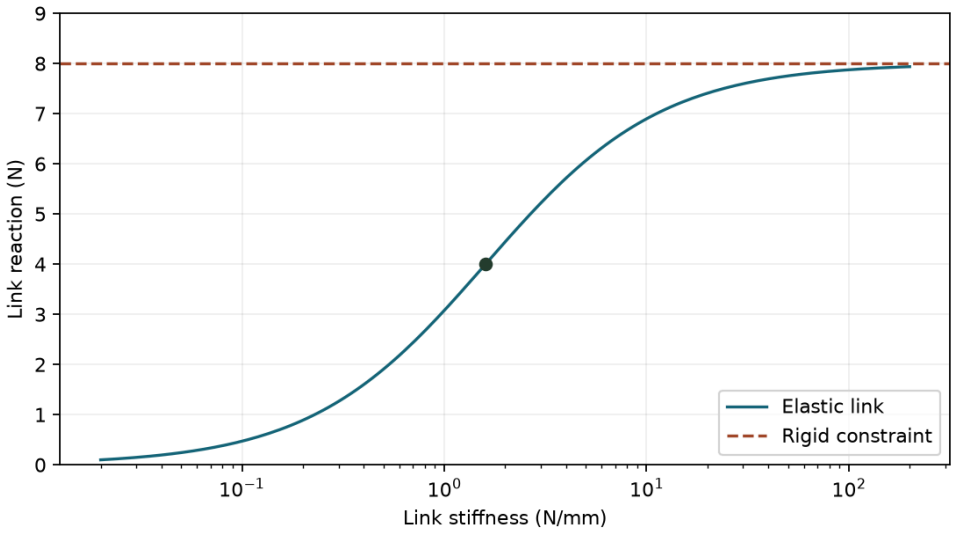
The stiffness of the structure as seen through the link is $1/S = 1.6$ N/mm. Matching this value transfers half the limiting reaction. Reaching 90% requires nine times this stiffness, or 14.4 N/mm. The code checks the reduced force against the full coupled displacement solution over four decades of link stiffness.

```
"""Find the reaction transmitted by a compliant link between two supports."""
import matplotlib.pyplot as plt
import numpy as np
from onesim.numerics.linear.matrices.dense_linear_algebra import solve_dense
```

```
support = np.diag([2.0, 8.0])  # N/mm
load = np.array([10.0, 0.0])  # N
gap = np.array([1.0, -1.0])  # positive when the first support moves farther
free_gap = gap @ solve_dense(support, load)
compliance = gap @ solve_dense(support, gap)  # mm/N
link_stiffness = np.geomspace(0.02, 200, 150)  # N/mm
reaction = free_gap / (compliance + 1 / link_stiffness)
rigid_reaction = free_gap / compliance

# Check the reduced equation against the full two-displacement system.
for stiffness, force in zip(link_stiffness, reaction):
    displacement = solve_dense(support + stiffness * np.outer(gap, gap), load)
    assert np.isclose(stiffness * (gap @ displacement), force)
print(f"Half the rigid reaction at {1/compliance:.2f} N/mm; rigid limit {rigid_reaction:.2f} N")

fig, ax = plt.subplots(figsize=(6.8, 3.8), constrained_layout=True)
ax.semilogx(link_stiffness, reaction, color="#176879", label="Elastic link")
ax.axhline(rigid_reaction, ls="--", color="#a4492b", label="Rigid constraint")
ax.plot(1/compliance, rigid_reaction/2, "o", color="#273d30")
ax.set(xlabel="Link stiffness (N/mm)", ylabel="Link reaction (N)", ylim=(0, 9))
ax.grid(alpha=0.2)
ax.legend()
fig.savefig("linear-algebra.png", dpi=160)
```



The bend in the curve occurs where the link and the two supports contribute equal compliance. The high-stiffness plateau is the 8 N reaction of the rigid constraint.

Download the Python example

Two springs sharing a displacement

Let $K = \operatorname{diag}(2, 8)$ in N/m and apply $f = (10, 0)^T$ N. A rigid link demands $u_1 - u_2 = 0$, so $B = (1, -1)$ and $g = 0$. The uncoupled displacement is $(5, 0)^T$ m. The reduced compliance is $1/2 + 1/8 = 5/8$ m/N, giving $\lambda = 5/(5/8) = 8$ N. Recovering the displacement gives $u = (1, 1)^T$ m. The first spring carries 2 N and the other carries 8 N. Independently, the common displacement under a total stiffness of 10 N/m is exactly one metre.

Prescribing a gap $g = 0.5$ m changes the reduced right-hand side to 4.5 m. The reaction drops to 7.2 N and the displacements become 1.4 and 0.9 m. Their difference verifies the new constraint. This calculation also identifies the multiplier sensitivity to a gap error as $-1/S$, here -1.6 N/m.

Solving without an explicit inverse

Applying S to a trial vector takes a multiplication by B^T, a solve with K, and a multiplication by B. The reduced matrix may be dense even with local springs because the stiffness solve communicates through the structure. Reuse a factorisation when the load changes. With approximate inner solves, control their errors relative to the requested outer accuracy. Check force balance and compatibility independently after reconstructing u. A tiny reduced residual obtained with inaccurate inner solves need not imply that the reconstructed state satisfies the original load equations.

A duplicate constraint as a diagnostic

Append a second row $2B$ to the rigid-link constraint and prescribe a compatible right-hand side $2g$. The new reduced matrix is $S\begin{pmatrix}1 & 2\\ 2 & 4\end{pmatrix}$, which has determinant zero. Only the combination $\lambda_1 + 2\lambda_2$ affects the applied reaction in the mechanical equations. In the zero-gap example every pair with that combination equal to 8 N gives the same displacement and resultant link force. A solver's complaint about a singular constraint block can thus expose a duplicated physical statement, even when a force measurement or displacement plot looks perfectly reasonable.

Condense internal degrees of freedom while preserving energy

Partition the degrees of freedom of a symmetric positive definite stiffness matrix into interior coordinates i and retained boundary coordinates b. For a specified boundary displacement, the interior balance is

$$K_{ii}u_i = f_i - K_{ib}u_b.$$

Eliminating it produces the boundary operator $S = K_{bb} - K_{bi}K_{ii}^{-1}K_{ib}$ and the effective boundary load $f_b - K_{bi}K_{ii}^{-1}f_i$. This elimination differs physically from setting the interior displacement to zero: the interior has been allowed to respond to the prescribed boundary value and applied load.

With zero load, insert $u_i = -K_{ii}^{-1}K_{ib}u_b$ into the full strain energy. Completing the square in u_i shows that the minimum energy at fixed boundary coordinates is $u_b^T S u_b/2$. Since the full stiffness is positive definite, any nonzero boundary vector yields positive minimum energy. This gives a physical proof that the condensed operator is positive definite. A free-floating component may instead have rigid-body modes; then the same proof establishes only nonnegativity, and some imposed displacement patterns cause no stress.

Reduce a spring chain and preserve the internal load

Two springs connect a fixed left support to an observable endpoint through an internal joint. Their stiffnesses are 6 and 3 N/mm. Let the internal and boundary displacements be u_i, u_b, and apply forces 3 and 6 N at those nodes. The strain energy is $[6u_i^2 + 3(u_b - u_i)^2]/2$. Differentiating gives

$$K_{ii} = 9,\ K_{ib} = K_{bi} = -3,\ K_{bb} = 3.$$

Condensing yields $S = 3 - 9/9 = 2$ N/mm and effective boundary load $6 - (-3)(3/9) = 7$ N. Thus $u_b = 3.5$ mm. Recovering the internal state gives

$$u_i = (3 + 3u_b)/9 = 1.5\,\mathrm{mm}.$$

The first spring therefore carries $6(1.5) = 9$ N, and the second carries

$$3(3.5 - 1.5) = 6\,\mathrm{N}.$$

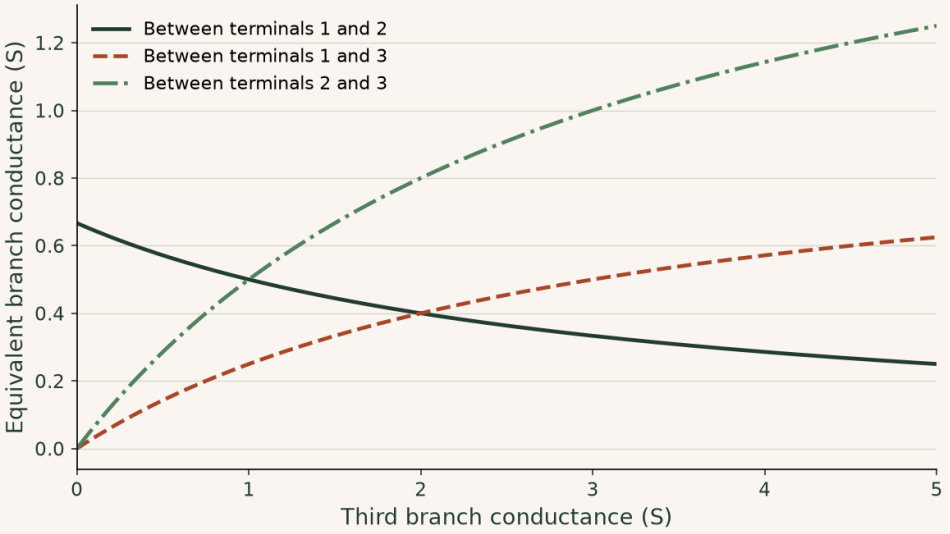
Their difference is the internal force of 3 N and the support reaction balances the total applied load of 9 N. Simply using the well-known series stiffness 2 N/mm with the endpoint force 6 N would yield 3 mm, missing the transmitted effect of the interior load. Both condensed stiffness and condensed right-hand side are necessary in a reduced model.

A local finite-element assembly can repeat this calculation for many independent element interiors, retaining only interface coordinates for the global solve. To compute internal stress afterward, keep enough local factor information to reconstruct each eliminated state. The boundary displacement alone does not directly encode the internal load distribution; two differently loaded elements can have the same endpoint displacement and different internal stresses. This distinction matters whenever reduction is justified using one observable but results are later requested for another.

Under a change in the prescribed end condition, the same reduced law can be used in the opposite direction. Hold the endpoint at $u_b = 0$ with the internal 3 N force still present and remove the applied endpoint load. The recovered internal displacement is $1/3$ mm. The computed endpoint force required to hold it fixed is $-3u_i = -1$ N. A boundary law containing only the series stiffness would predict zero force for this imposed zero displacement.

Read the coupling introduced by eliminating an interior node

Three terminals join a central node through conductances $1, 2, g$ S. The central potential is their conductance-weighted average. Substitute it into each terminal current. The Schur complement introduces a direct pairwise conductance $g_i g_j / \sum g_\ell$.



At g = 0 the first two resistors are in series, giving 2/3 S. At g = 3 the first-to-second coupling is only 1/3 S and two previously absent edges have appeared. Eliminating a single degree of freedom can make a sparse matrix locally dense.

▶ **Read values from the plot**

Change parameters · Read the coupling introduced by eliminating an interior node

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.

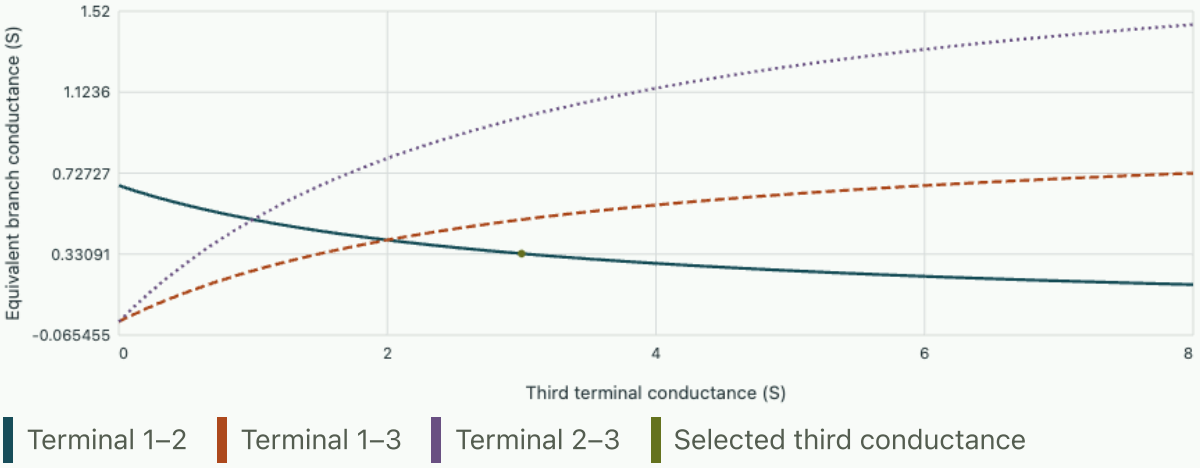
Open this calculation with adjustable inputs

Try the calculation

First terminal conductance (S) 1

Second terminal conductance (S) 2

Selected third conductance (S) 3



At g_3 = 3 S the eliminated 1–2 branch has conductance 0.33333 S. All three couplings come from the same Schur complement.

Study the full derivation

When stable eigenvalues allow large transient growth · Matrix computations and spectral accuracy

Two damped states exchange energy in only one direction. Its eigenvalues predict ultimate decay, while a short-time response can grow by a large factor. Compute the matrix exponential, its gain and a resolvent-based sensitivity check.

Problem sheet 6 · Accurate constrained solves

Track the correction error, distinguish residual definitions, and recover physical forces from a constrained quadratic model.

Reuse an inaccurate scalar factor

Solve $5x = 10$ using the approximate factor 6 for the initial solution and every correction. Derive the error after k corrections and count how many bring absolute error below 0.002.

▼ **Worked solution**

The starting value $10/6$ leaves $e_0 = 1/3$. The correction adds $5e_k/6$, so $e_{k+1} = e_k/6$ and $e_k = 1/(3 \cdot 6^k)$.

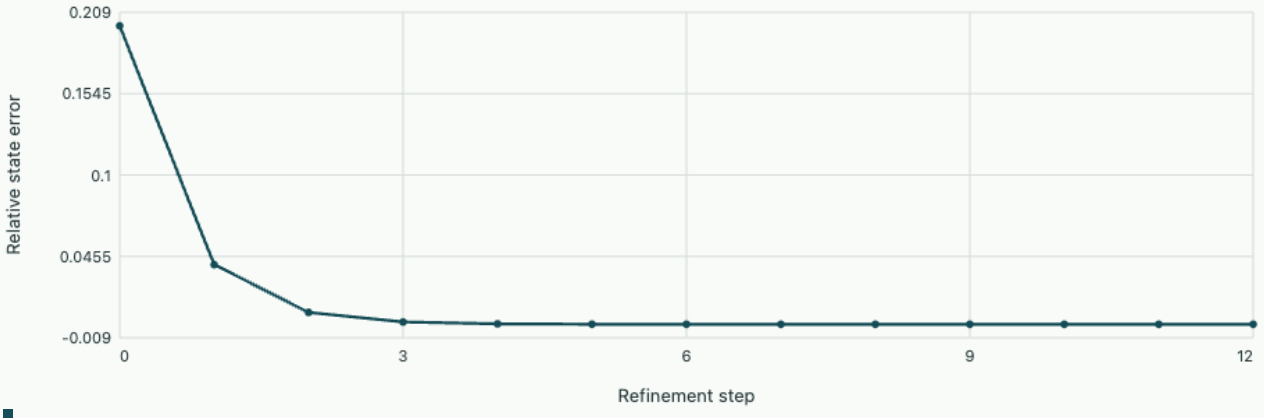
Two corrections give $1/108 \approx 0.00926$. Three give $1/648 \approx 0.00154$, below the threshold. Residuals must use the original factor 5 because the initial value solves the approximate equation exactly.

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

Approximate factor relative error	**0.25**
Initial state error	**0.2**



Residual correction with approximate factor

Using approximate scalar factor â = a(1 + ε), each correction multiplies the error by ε/(1 + ε) = 0.2. Accurate residual formation permits improvement only when this factor has magnitude below one.

Scale the units

For $A = \operatorname{diag}(1, 1000)$, $b = (1, 1000)^T$, $x = (0.99, 1)^T$, compute the infinity norm of the residual and the maximum componentwise backward error. Repeat when the first row and its load are multiplied by 100.

▼ **Worked solution**

Initially $r = (0.01, 0)^T$, hence its infinity norm is 0.01. The maximum row ratio is

$$0.01/(0.99 + 1) = 1/199.$$

Rescaling changes the first residual to 1 and its denominator to 199. The maximum absolute residual increases hundredfold, whereas the backward error remains $1/199$. The two systems describe exactly the same solution.

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.

Prescribe a weighted displacement

Minimise $(u_1^2 + 4u_2^2)/2$ subject to $u_1 + 2u_2 = 3$. Using $Ku + B^T\lambda = 0$, determine the reduced stiffness equation, the multiplier and the stored energy. Compare that energy with the admissible state $(3, 0)$.

▼ **Worked solution**

With $K = \operatorname{diag}(1, 4)$, $B = (1, 2)$, the scalar compliance is $1 + 4/4 = 2$. The reduced system is $2\lambda = -3$. Back-substitution gives $u = (1.5, 0.75)^T$, satisfying the prescribed value.

The energy is

$$(2.25 + 4 \cdot 0.5625)/2 = 2.25.$$

The competing state has energy 4.5. Positive definiteness on the constraint line guarantees that the stationary state is its unique minimum.

Explore the related calculation with adjustable inputs. Inspect the effective couplings obtained by eliminating an interior variable. The same substitution generates the Schur operator in the stated constrained system.

Reconstruct a displacement measured almost twice in the same

direction

Two gauges measure b = Ax with A having rows (1, 0) and (1, 0.01). With b = (1.000, 1.010) find x. Each reading has independent standard deviation 0.001. Estimate the standard deviation of each displacement and their covariance. Evaluate a proposed third gauge measuring only the second displacement.

▼ Worked solution

The first equation gives $x_1 = b_1$ and differencing gives $x_2 = 100(b_2 - b_1) = 1$. The solution is (1, 1). A residual of zero follows automatically for an invertible square system, even though the second displacement is obtained by subtracting two close readings.

The covariance follows by propagating the independent reading errors. $\mathrm{Var}(x_1) = 10^{-6}$ and $\mathrm{Var}(x_2) = 10^4(2 \times 10^{-6}) = 0.020$, so the standard deviations are 0.001 and 0.1414. The covariance is $-100 \times 10^{-6} = -0.0001$. An upward perturbation of the first gauge increases x_1 but decreases the reconstructed x_2.

A third independent reading of x_2 with the same standard deviation carries much more precision along the weak direction. Fuse the old x_2 estimate with this reading using inverse variance: the new variance is $1/(1/0.020 + 1/10^{-6}) \approx 9.9995 \times 10^{-7}$. In the joint fit solve the overdetermined system by QR. Repeating the original two directions could improve repeatability, but would retain the poor geometry.

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 · Why a positive energy has a Cholesky factor

Start with the work stored in two coupled springs and discover a triangular square root of a matrix by completing the square.

From positive energy to two triangular solves

If displacements x store elastic energy $x^T Ax/2$, then a fixed structure with no free modes has $x^T Ax > 0$ for every nonzero x. Equilibrium under load f minimises $x^T Ax/2 - f^T x$. Differentiation gives $Ax = f$ for symmetric A. Can the energy itself tell us which operations are safe in a solver?

Write the first coordinate as s and the remaining coordinates as z. Expand the quadratic form using first entry a, coupling column b, and trailing block C. The cross term is $2sb^T z$. Adding and subtracting $(b^T z)^2/a$ gives

$$x^T Ax = a(s + b^T z/a)^2 + z^T(C - bb^T/a)z.$$

Positivity with $x = e_1$ proves $a > 0$. For any nonzero z, choose $s = -b^T z/a$. The remaining term must then be positive, so the reduced matrix is positive definite. The argument repeats down to a scalar. Each completed square supplies one row of a triangular factor and one positive pivot. In matrix notation $A = LL^T$, with $L_{11} = \sqrt{a}$ and $L_{2:n,1} = b/\sqrt{a}$. Solve $Ly = f$ forward, then $L^T x = y$ backward. Neither inverse need be formed.

Worked equilibrium and a change of load

Use $A = \left(\begin{smallmatrix} 4 & 2 \\ 2 & 5 \end{smallmatrix}\right)$ kN/m. The first diagonal factor is 2, the off-diagonal factor is 1, and the remaining pivot is $5 - 1^2 = 4$. Thus $L = \left(\begin{smallmatrix} 2 & 0 \\ 1 & 2 \end{smallmatrix}\right)$ in consistent numerical units. For $f = (2, 3)^T$ kN the intermediate result is

$$y = (1, 1)^T.$$

Back substitution gives $x = (0.25, 0.5)^T$ m. Check the second row directly: $2(0.25) + 5(0.5) = 3$.

Now apply $f = (0, 2)^T$ to the same structure. The intermediate vector changes to $(0, 1)^T$ and the displacements to $(-0.25, 0.5)^T$. The same two square roots serve both load cases. This matters for many load cases on one stiffness matrix.

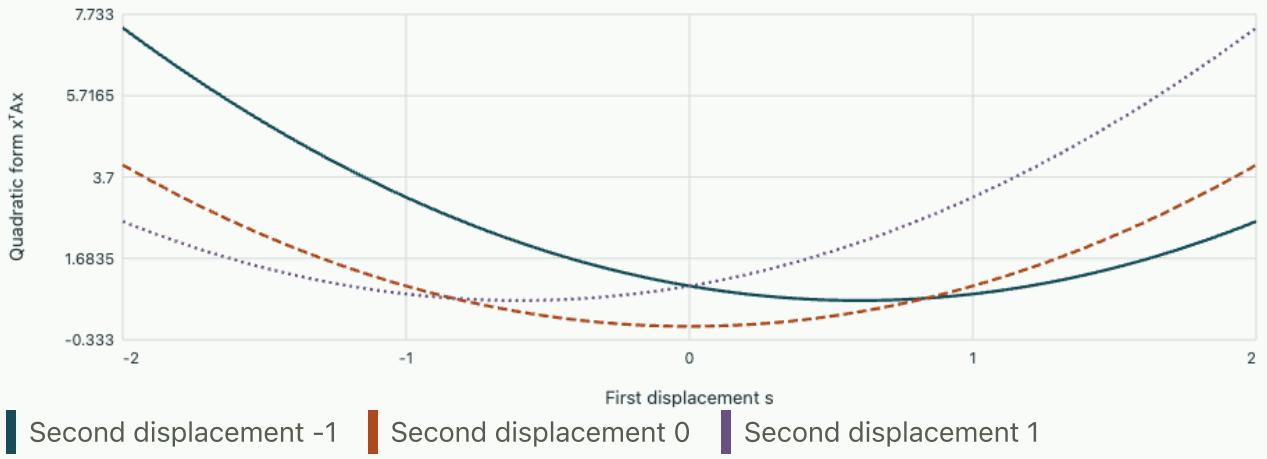
Change the coupling

For the dimensionless matrix with diagonal 1 and off-diagonal c, predict which load direction becomes soft as c approaches one. The plot follows elimination along the first displacement for three fixed values of the second. Read the minimum on each slice before changing c.

Open this experiment online

Try the calculation

Coupling c **0.6**



Second pivot 0.64. For load (1, –1), first displacement 2.5.

The second pivot is $1 - c^2$. Under the antisymmetric load $(1, -1)^T$ the displacement is $(1, -1)^T/(1 - c)$. A modest change in the input coupling can therefore have a large physical effect even with an exact factorisation.

Tutorial on pivots and supports

A free two-node spring has matrix $k\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$. Where does the recurrence stop and which physical motion does it reveal?

▼ **Worked solution**

Elimination gives the second pivot $k - k^2/k = 0$. Translating both endpoints by the same amount stores zero energy. Fixing one displacement removes that unconstrained motion.

For diagonal entries 9 and 4, what range of coupling b admits positive pivots? Interpret the endpoints.

▼ **Worked solution**

The first pivot is 9. The second is $4 - b^2/9 > 0$, or $|b| < 6$. At equality the second completed square vanishes. There is a displacement with zero energy even though each diagonal term separately is positive.

Chapter 14 · Removing unequal scales before an iteration

Derive the change of variables behind diagonal preconditioning, then measure what it can and cannot repair in a two-spring system.

Coordinates in the energy

Suppose one displacement is measured in millimetres and the other in metres. A steepest-descent direction depends on the Euclidean lengths of the coordinates. The physical minimum does not. Let A be symmetric positive definite, $D = \operatorname{diag}(A)$, and introduce $y = D^{1/2}x$. Substituting $x = D^{-1/2}y$ in the energy produces

$$\widetilde{A} = D^{-1/2} A D^{-1/2},$$

$$\tilde{b} = D^{-1/2} b.$$

The diagonal of the new matrix is one. Its off-diagonal entry is the original coupling divided by the geometric mean of the two direct stiffnesses. This is a dimensionless measure of interaction. The transformed gradient step maps back to $x_{j+1} = x_j + \alpha D^{-1}(b - A x_j)$.

Choosing a common step

For any eigenvector of the transformed matrix the iteration multiplies the error by $1 - \alpha\lambda$. With eigenvalues in $[m, M]$, equalise the two extreme absolute factors by taking $1 - \alpha m = -(1 - \alpha M)$. This gives $\alpha = 2/(M + m)$ and worst-case contraction $(M - m)/(M + m)$. If m approaches zero the direction barely moves, however accurately the linear algebra is performed.

Two quite different causes of a slow solve

For $A = \operatorname{diag}(100, 1)$, an optimal unscaled step shrinks the error by only $99/101$. Diagonal scaling gives the identity. One step of length one then solves both components exactly. This is the effect of unlike units or stiffness scales.

For a unit-diagonal matrix with coupling 0.95, scaling leaves the matrix unchanged. Its eigenvalues are 1.95 and 0.05. Each optimal step can retain 95% of the error. The nearly free opposing motion is part of the problem itself. A preconditioner that represents the coupling, or a different physical constraint, is needed for faster convergence in that direction.

Compare convergence factors

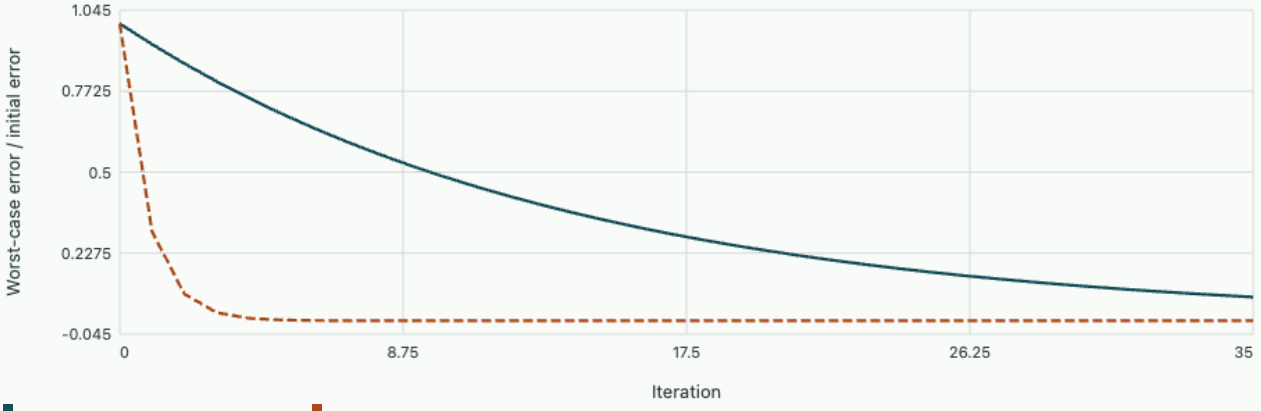
Vary the stiffness ratio and coupling independently. The curves are worst-case error factors raised to the iteration count. They give bounds attained by an extreme eigenmode with this particular fixed optimal step.

Open this experiment online

Try the calculation

Direct stiffness ratio **25**

Dimensionless coupling **0.3**



Optimal fixed-step contraction 0.93026 before scaling and 0.3 afterwards.

Tutorial on interpreting convergence

If a bound decreases from 1 to 10^{-3} in k steps, how do you determine k from a known factor 0.8?

▼ **Worked solution**

Take logarithms in $0.8^k \le 10^{-3}$. Because $\log(0.8)$ is negative, division reverses the inequality. The smallest integer is 31.

Why should a conjugate-gradient implementation use a positive definite preconditioner?

▼ **Worked solution**

In transformed variables CG minimises an energy for a symmetric positive definite operator. A positive definite preconditioner admits an invertible square root and transfers this property by congruence. An indefinite matrix does not define the corresponding norm.

Read the university treatment

Track which linear system must be solved at each iteration. Compare the cost of a preconditioner with the reduction in the iteration count. Driscoll and Braun · Preconditioning

Chapter 15 · How a moderately accurate mode can give a precise frequency

Use the Rayleigh quotient to estimate an eigenvalue, understand its stationary error and check a proposed mode with a residual.

A frequency estimate from a trial shape

For $Kq = \lambda M q$ with symmetric K and positive definite mass M, a trial vector v fixes a direction but leaves an amplitude free. Substituting $q(t) = a(t)v$ in kinetic and elastic energy gives effective mass $v^T M v$ and stiffness $v^T K v$. Their ratio $R(v)$ estimates the squared frequency. Rescaling v has no effect. For $M = I$, differentiating $v^T K v / (v^T v)$ gives a gradient proportional to $Kv - Rv$. Its zeros are precisely the eigenvectors.

Angular error and the residual

Take exact orthonormal modes u_1, u_2 and a trial $v = u_1 \cos\theta + u_2 \sin\theta$. Orthogonality removes all cross terms. Then

$$R(v) - \lambda_1 = (\lambda_2 - \lambda_1)\sin^2\theta,$$
$$\|Kv - R(v)v\|_2 = |\lambda_2 - \lambda_1|\,|\sin\theta\cos\theta|.$$

The eigenvalue error is second order in a small angular error. The residual is first order. For general dimension write $v = \sum c_i u_i$ with unit norm. The squared residual is $\sum |c_i|^2 (\lambda_i - R)^2$, so at least one eigenvalue lies within one residual norm of R. Identifying which mode requires more information about the spectrum.

Two measurements

For squared frequencies 4 and 9 in s⁻² and a 10-degree error in the first trial mode, $R = 4 + 5(0.03015) = 4.1508$ s⁻². The estimated angular frequency is 2.0373 rad/s. The relative error in frequency is about half the relative error in its square at small errors.

For eigenvalues 4 and 4.01, a 45-degree mixture gives a residual of just 0.005. This very small number is consistent with a large uncertainty in shape. If the two eigenvalues were identical, every combination in the eigenspace would be an exact mode.

Tilt the trial vector

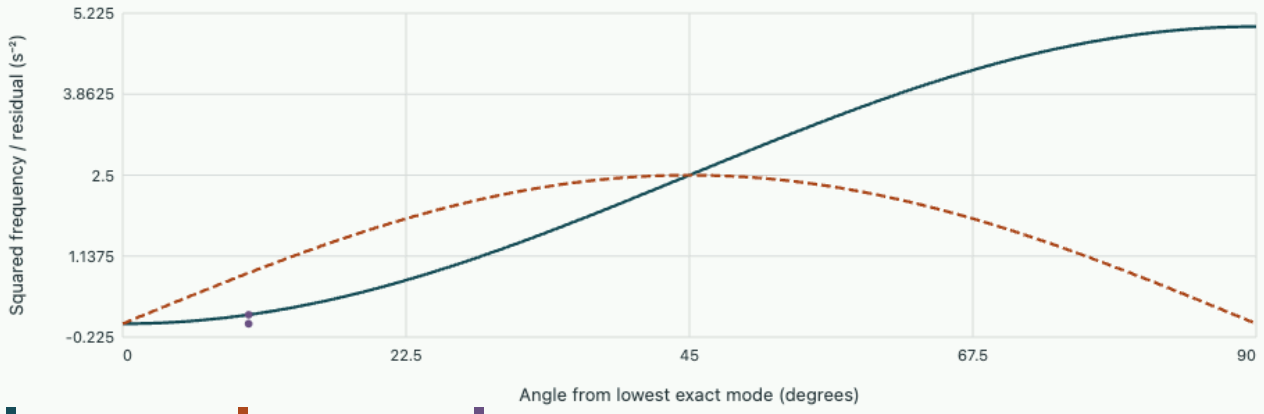
The plot shows the quotient error and residual across trial angles for a chosen spectral gap. Move the angle marker, halve a small angle and compare the two responses.

Open this experiment online

Try the calculation

Separation of squared frequencies (s⁻²) 5

Trial angle (degrees) 10



Quotient error Residual norm Selected angle

Selected angle gives quotient error 0.15077 s⁻² and residual 0.85505 s⁻².

Tutorial on frequency estimates

For eigenvalues 2 and 10, a trial produces $R = 2.08$. Assuming the two-mode model, estimate its angular error.

▼ **Worked solution**

The mixture fraction is

$$(2.08 - 2)/(10 - 2) = 0.01.$$

Hence $|\sin\theta| = 0.1$, or about 5.74 degrees near the lower mode.

A trial gives $R = 5$ and residual zero. Does it prove that 5 is the lowest eigenvalue?

▼ **Worked solution**

It proves that the trial vector is an exact eigenvector for 5. To certify the lowest eigenvalue one also needs a lower bound or a description of the remaining spectral subspace.

Read the university treatment

Compare how quickly an eigenvector direction converges with the rate for its Rayleigh-quotient eigenvalue estimate. Driscoll and Braun · Power iteration

Chapter 16 · How to rotate two point clouds into agreement

Reduce a nonlinear-looking alignment problem to two scalar sums, distinguish reflections from rotations and recognise a geometry that contains no information about angle.

Eliminate translation before fitting orientation

A microscope images the same markers twice. Marker i has coordinates p_i in the first image and q_i in the second. Seek a proper planar rotation R and translation t minimising $\sum_i \|q_i - Rp_i - t\|^2$. Differentiating with respect to t gives

$$t = \bar{q} - R\bar{p}.$$

Substituting this expression centres both sets. Write $P_i = p_i - \bar{p}$ and $Q_i = q_i - \bar{q}$. Orthogonality of R makes $\sum \|P_i\|^2 + \|Q_i\|^2$ independent of the angle. Minimising distance is therefore equivalent to maximising $\sum Q_i^T R P_i$.

Recover the angle and its conditioning

Insert the rotation matrix with first column $(\cos\theta, \sin\theta)$. Collect the cosine and sine terms to obtain $C\cos\theta + S\sin\theta$, with $C = \sum P_i \cdot Q_i$ and $S = \sum (P_{ix}Q_{iy} - P_{iy}Q_{ix})$. The maximum occurs at $\operatorname{atan2}(S, C)$. Unlike an ordinary arctangent, this keeps the quadrant and handles $C = 0$. The objective curvature at the maximum is $-\sqrt{C^2 + S^2}$. Small curvature signals that very different angles fit almost equally well.

Two fits with identical measurement precision

For centred markers $(-1, 0), (1, 0)$ and observations $(0, -1), (0, 1)$, $C = 0, S = 2$. The result is 90 degrees. Adding the displacement $(3, 2)$ to every observation leaves the fitted angle unchanged and recovers precisely that translation. If both markers coincide then all centred P_i vanish and no angle can be identified.

Now separate the original markers by $2a$ and shift one observed marker by a transverse measurement error ϵ. At zero true rotation the connecting segment tilts by $\arctan(\epsilon/(2a))$. The same 0.1 mm error means about 2.86° for $a = 1$ mm but only 0.29° for $a = 10$ mm. A low spatial residual alone is an incomplete calibration certificate.

Vary the calibration baseline

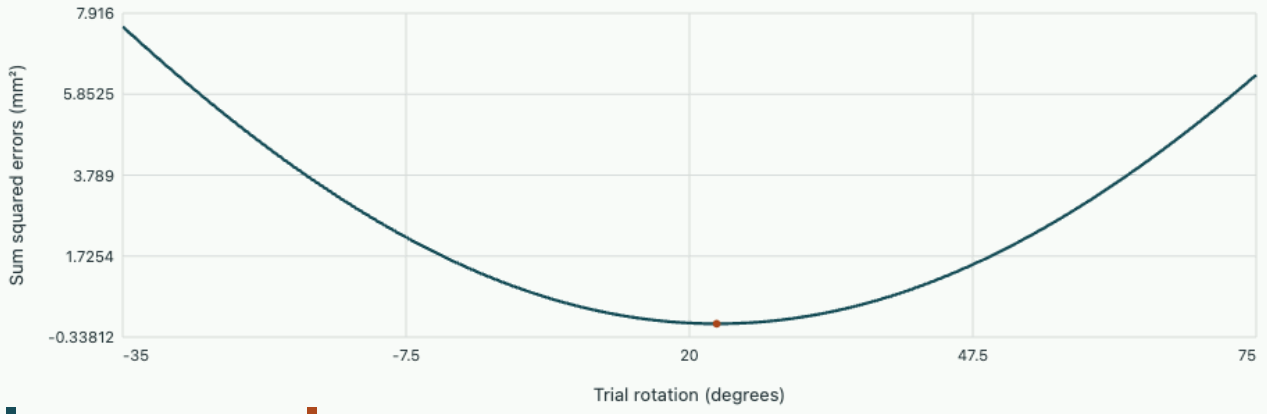
The true angle is fixed at 20 degrees. A vertical error is added to the positive marker in the second image. The plotted squared distance is computed after centring. Locate its minimum and examine how broad it is for a short or long baseline.

Open this experiment online

Try the calculation

Marker half-baseline (mm) **2**

Vertical error of right marker (mm) **0.2**



Squared fit residual Optimal angle

Recovered angle 22.645°, error 2.6449°. Sum squared residuals 0.0026457 mm².

Tutorial on what the fit measures

If an unconstrained matrix maps a chiral triangle to its mirror with zero error, can a proper planar rotation do the same?

▼ **Worked solution**

A rotation has determinant +1 and preserves signed area. A reflection reverses signed area. A non-collinear labelled triangle cannot match its mirror exactly under a proper rotation and translation.

Why can one ignore weights for equally precise markers but not for a blurred outlier?

▼ **Worked solution**

Independent isotropic errors with equal variance give a common quadratic weight. Different variances require weighted centroids and weighted C, S sums. A very noisy marker should contribute less to the fitted orientation.

Chapter 17 · Reading an effective resistance from a singular graph matrix

Assemble Kirchhoff's equations from oriented edge differences, remove the arbitrary reference voltage and relate a two-terminal measurement to minimum dissipation.

One matrix for current balance

Give each resistor edge an arbitrary orientation. A row of the incidence matrix B has −1 at its tail and +1 at its head. The difference Bv measures an oriented voltage drop. With diagonal conductances G, the dissipated power is

$$P = v^T B^T G B v.$$

Differentiating half this quadratic minus injected-current work $j^T v$ gives $Lv = j$ with $L = B^T G B$. Changing an edge orientation changes a sign twice and leaves L unchanged.

A constant voltage lies in the nullspace because $B\mathbf{1} = 0$. On a connected graph the converse follows from $v^T L v = \sum_e g_e (\Delta v_e)^2$: zero energy forces equal voltage across every edge. The current injections must sum to zero. Set one reference voltage to zero or constrain the mean to remove this unobservable constant.

Compute a three-node measurement in two ways

Connect terminals A and C by a 6 Ω resistor. A second path passes through B with 2 Ω between A and B and 4 Ω between B and C. Inject 1 A at A, remove 1 A at C and ground C. Kirchhoff at B gives $(V_B - V_A)/2 + V_B/4 = 0$, so $V_B = 2V_A/3$. At A, $V_A/6 + (V_A - V_B)/2 = 1$, giving $V_A = 3$ V and $V_B = 2$ V. Effective resistance is 3 Ω.

Each parallel path carries 0.5 A, and dissipates $0.5^2(6) = 1.5$ W. Total heat equals $I(V_A - V_C) = 3$ W. Opening the direct edge doubles the effective resistance. Opening an edge that bridges two otherwise disconnected subgraphs can instead make it impossible to pass any steady current between the terminals.

Challenge a network calibration

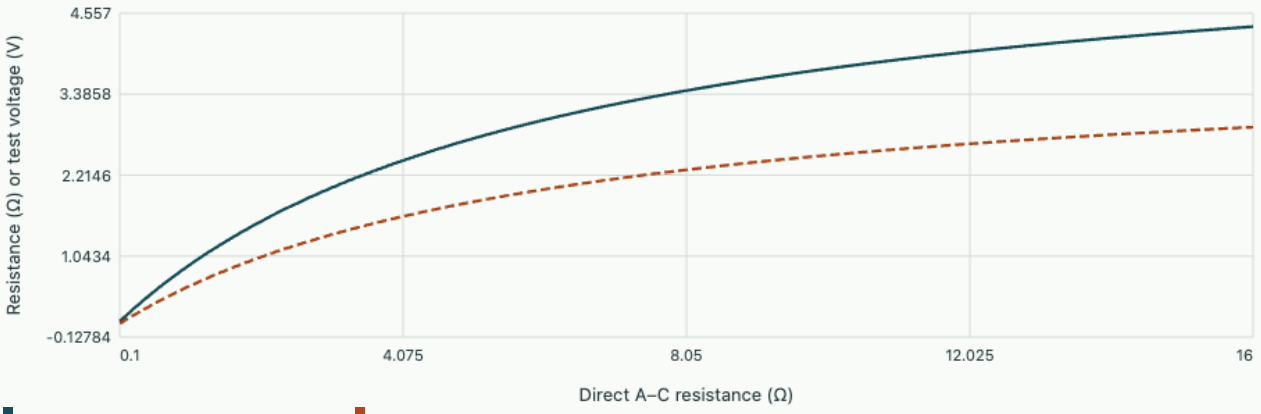
Change the two series resistors. The independent variable is the direct A–C resistance. The graph compares the terminal resistance with the middle-node voltage under a 1 A test current. In these numerical units volts and ohms have equal numbers; each curve's legend identifies the physical quantity.

Open this experiment online

Try the calculation

A–B resistance (Ω) 2

B–C resistance (Ω) 4



Effective resistance (Ω) Voltage at B for 1 A (V)

With direct edge 6 Ω, effective resistance 3 Ω and interior voltage 2 V.

Tutorial on identifiability in a circuit

Can a single terminal-resistance reading isolate the two series resistor values?

▼ **Worked solution**

It determines only their sum if the direct edge is known. Measuring V_B under the same known 1 A injection determines the division within that sum.

Why does minimising dissipation with prescribed terminal voltages yield Kirchhoff balance at interior nodes?

▼ **Worked solution**

A variation of any free voltage contributes twice its net current departure to the first variation of total power. At the minimum all such coefficients vanish. Terminal voltages are fixed and their reaction currents are the measured injections.

Study the full derivation

Solve a singular network without inventing a ground current · Matrix computations and spectral accuracy

A floating resistor network has unique voltage drops but no absolute voltage reference. Derive its nullspace, test charge balance, and distinguish adding a computational gauge from changing the physical boundary conditions.

Chapter 18 · Counting vibration modes without calculating their shapes

A long chain of beads is tested on a shaker. The instrument can sweep only a narrow frequency interval, and the operator wants to know which resonance lies inside it. Turn a mass and spring model into a counting problem. Work out why a sign in an elimination step can identify a frequency that has never been computed directly.

Start with one displacement per bead

Number the moving beads from one to n. A fixed support at either end has zero displacement. Each neighbouring pair is joined by an identical linear spring of stiffness k, including the two end springs. Write down the force on bead i by extending each spring separately. The left spring pulls in proportion to $u_{i-1} - u_i$, the right spring in proportion to $u_{i+1} - u_i$. The diagonal coefficient two comes from two neighbours, including a fixed support when the bead is at an end. Removing an end spring would change the physical model and its first frequency.

$$m_i \ddot{u}_i + k(2u_i - u_{i-1} - u_{i+1}) = 0,$$
$$u_0 = u_{n+1} = 0.$$

Set $u_i(t) = v_i \cos(\omega t)$ and cancel the nonzero time factor. The resulting equation is $Kv = \lambda M v$ with $\lambda = \omega^2$. The stiffness matrix has a quadratic energy interpretation: multiplying by a displacement on either side gives the sum of all squared spring extensions times k. This sum is positive unless all displacements vanish, because the end supports pin the chain. The diagonal mass matrix is also positive. Thus a frequency of zero in this experiment would call for an explanation.

Directly dividing each row by its own mass produces $M^{-1}K$, which need not look symmetric. Instead set $z = M^{1/2} v$. Multiplying by $M^{-1/2}$ on the left gives $Az = \lambda z$ with $A = M^{-1/2} K M^{-1/2}$. The diagonal entries are $a_i = 2k/m_i$; the neighbouring entries are $b_i = -k/\sqrt{m_i m_{i+1}}$. Both have units of inverse time squared. Increasing one bead mass changes two neighbouring off-diagonals as well as a diagonal. Changing only the diagonal in a program would simulate a different chain.

Derive the shift count by completing squares

Pick a real test value s in squared-frequency units. In an orthonormal eigenbasis the quadratic form of $A - sI$ is $\sum_j (\lambda_j - s) y_j^2$. The largest dimension of a subspace on which this form is strictly negative counts eigenvalues below the test value. An invertible change of coordinates maps a negative subspace bijectively onto another negative subspace of the same dimension. The largest dimension stays fixed. This is the part of the law of inertia that our computation needs.

Now look at the same form in the original bead coordinates. Its first terms are $d_1 z_1^2 + 2 b_1 z_1 z_2 + (a_2 - s) z_2^2$, where $d_1 = a_1 - s$. Complete the square in z_1. The coefficient of the transformed first square is d_1. The coefficient left for z_2^2 is reduced by b_1^2 / d_1. Only the next coordinate is touched because the matrix is tridiagonal. Repeat the same operation until no cross term remains.

$$d_1 = a_1 - s,$$

$$d_{i+1} = a_{i+1} - s - \frac{b_i^2}{d_i}.$$

Each square completion is an invertible shear when its pivot is nonzero. Consequently the number of negative pivots equals the number of negative eigenvalues of the shifted matrix. This gives a practical count with linear work per shift and only one saved pivot. It also explains why the signs of the unmodified diagonal alone cannot work. Every spring couples two neighbouring trial directions; elimination carries the contribution of that coupling forward.

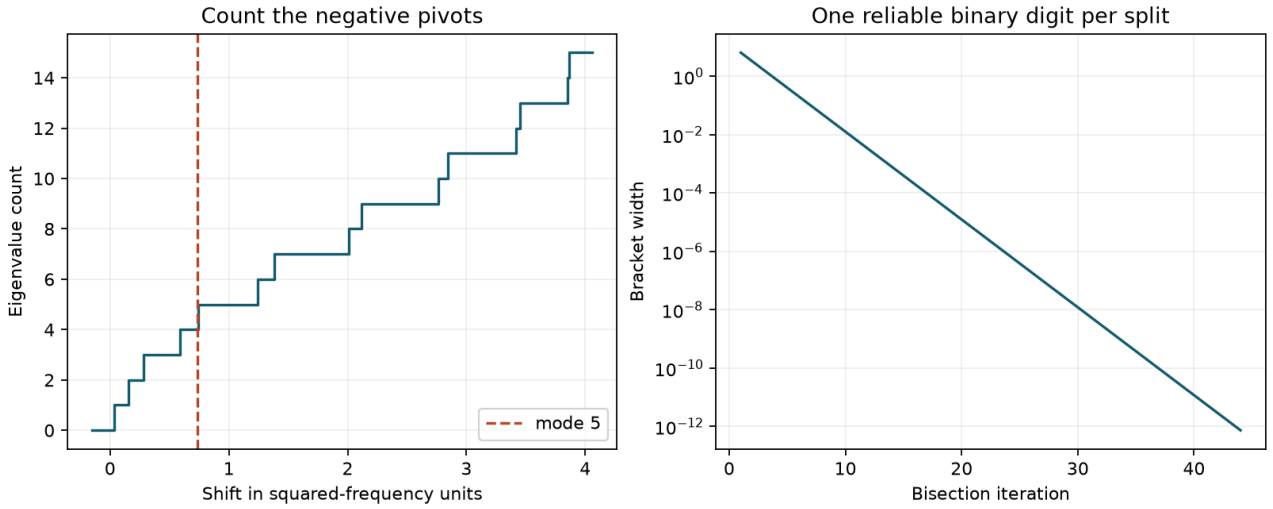
Find the bracket for a requested resonance

Put $k = m_i = 1$ in a three-bead chain. Its diagonal is $2, 2, 2$ and both off-diagonals equal -1. At $s = 1$ the second pivot is exactly zero, so choose $s = 1.2$ first to follow ordinary arithmetic. The pivots are $0.8, -0.45, 3.0222\ldots$. There is one eigenvalue below 1.2. At $s = 2.4$ they are $-0.4, 2.1, -0.87619\ldots$, giving two eigenvalues below 2.4. The second resonance in squared-frequency units lies between these two shifts. The uncertainty in angular frequency is the difference between their square roots.

For a longer chain it would be inconvenient to choose the first two trial shifts by luck. In an eigenvector choose a component of maximum absolute value. Divide its matrix row by that component. The sum of the neighbouring contributions is bounded in magnitude by the sum of the neighbouring absolute off-diagonals. Every eigenvalue therefore falls inside at least one interval centred on a diagonal entry with that row sum as radius. The minimum left endpoint and the maximum right endpoint supply an enclosing interval. A small outward extension avoids ties at the first or last bound.

To select the j-th eigenvalue in ascending order, halve the interval. If the count at its midpoint is less than j, the desired eigenvalue is on the upper side. Otherwise keep the lower side. Once the interval is initialised, the reduction factor does not depend on the accuracy of an initial frequency guess. After r splits the width in exact arithmetic equals its original width divided by 2^r. A small eigenvalue can still have a relatively wide frequency interval after taking square roots, so apply the stopping criterion in the quantity your instrument actually reports.

Separate the numerical count from an experimental identification



A heavy central bead moves the jumps of the count. Bisection follows one chosen jump. Its interval closes at a predictable rate even when two neighbouring resonances look close on the staircase.

For the uniform chain try $v_i = \sin(i\theta)$. This trial shape vanishes at the left support. The right support requires

$$\theta = j\pi/(n+1).$$

Substitution into an interior row and the sum formula for neighbouring sines give

$$\lambda_j = 4(k/m)\sin^2[j\pi/(2(n+1))].$$

For three unit beads the spectrum is approximately 0.58579, 2 and 3.41421. This furnishes an independent reference for all the mode counts, including the mode with an exactly representable eigenvalue. The notebook also compares with a dense symmetric solve which can be repeated using OneSimulation's linear algebra backend.

Double the mass of the middle bead and think about the second uniform mode before solving again. Its displacement vector is proportional to $(1, 0, -1)$. There is a node at the heavier bead. Both Kv and Mv are unchanged for this particular vector, so its squared frequency remains two. The lowest mode has substantial displacement there and generally moves. The operator can use this contrast experimentally: a frequency that hardly shifts under a known local mass change constrains the possible mode shape even when that shape was never recorded.

The count itself describes the modelled spectrum. A measured frequency response can hide a resonance if the shaker or the sensor sits exactly on a modal node. Conversely, a broad peak may contain two lightly separated modes. To compare with an experiment one needs the forcing and measurement directions as well as the frequency count. In the symmetric ideal chain, replacing a centre sensor with two off-centre sensors can expose a mode that a centre sensor cannot see. The counting algorithm has no term for this observation geometry.

What a zero pivot means on a computer

A zero intermediate pivot need not be an eigenvalue of the entire matrix. It may be an eigenvalue of a smaller leading block. Approaching the troublesome shift from one side resolves the ambiguous endpoint convention. If instead one jumped directly to an arbitrary large replacement pivot, its coupling contribution and sometimes its sign count would be wrong. The notebook uses a small negative pivot in unit-scaled problems, as in a conventional safeguarded Sturm count. An exact eigenvalue at the shift is then counted on the lower side. It tests that convention on the known three-bead chain.

Floating-point arithmetic places another limit on what it means to halve a bracket. Eventually the rounded midpoint equals one of its endpoints. At that stage further iterations cannot separate floating-point trial values. A production interval solver also handles extreme scaling, split tridiagonal blocks and tiny off-diagonals, and can report an interval containing a cluster. The notebook is deliberately a transparent solver for moderate, nondimensional examples. Its final consistency test tolerates a few units of roundoff so that the dense reference can be rounded differently.

Open editable notebook Download .ipynb

Tutorial on counts, brackets and experimental nodes

1. A four-second sweep is centred at an angular frequency of 1.4 per second. Does the three-bead example put the second natural frequency between 1.40 and 1.43 per second?

▼ **Work through the solution**

Square both positive bounds to get 1.96 and 2.0449 in inverse seconds squared. The count changes from one to two across them because the middle eigenvalue is 2. Its frequency is the square root of two, approximately 1.41421 per second. A finite measurement time still limits how sharply an instrument can identify this peak.

---

2. A starting interval in squared frequency is 8 units wide. How many halvings guarantee a width below one thousandth of a unit, and what part of the cost grows with bead count?

▼ **Work through the solution**

Thirteen halvings give

$$8/8192 = 0.0009765625.$$

Twelve leave 0.001953125. Each test visits every diagonal and each connecting off-diagonal once. The number of halvings follows from the desired absolute accuracy; the work of each count grows with the length of the chain.

---

3. Break the spring between beads 7 and 8 of a fixed-end chain without adding a support. How should the inertia procedure and the stiffness entries change?

▼ Work through the solution

Remove the coupling and subtract that spring's stiffness from each of the two adjacent unscaled diagonal entries before mass normalisation. The two components now have an independently free end and a fixed outer support. Count the shifted pivots separately in the two blocks and add the counts. Simply setting the off-diagonal to zero while keeping both diagonals would add artificial restoring forces at the cut.

Further reading · LAPACK Working Note 172 on Sturm counts and floating-point arithmetic for symmetric tridiagonal matrices

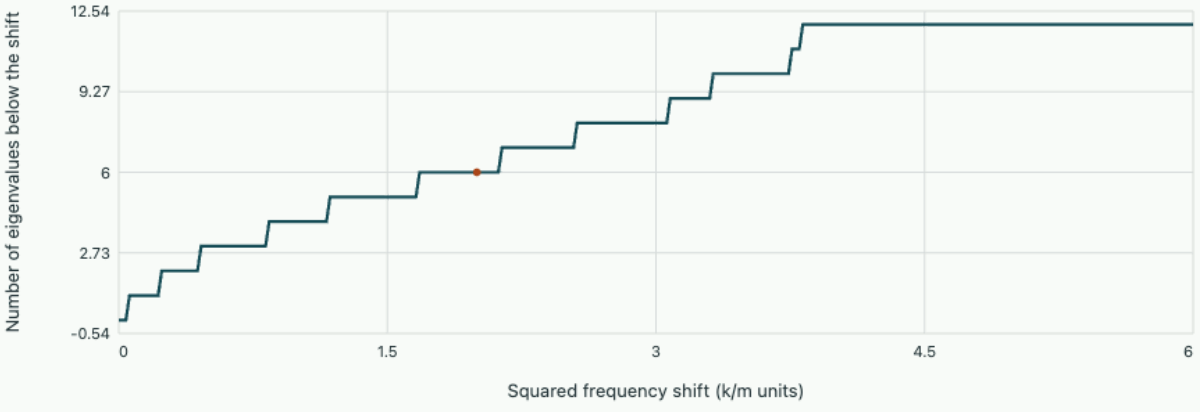
Change parameters · Counting vibration modes without calculating their shapes

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

Beads	12
Central mass / other masses	2
Selected squared frequency	2



LDL^T sign count Selected shift

There are 6 modes below the selected shift. Near a zero pivot this illustration uses a small negative pivot, corresponding to the upper-side limit of the counting step.

Problem sheet 7 · Project · Can two displacement sensors distinguish soft and stiff modes?

A pair of spring-mounted platforms forms a small inverse experiment. Use energy, conditioning, factorisation and modal residuals to decide what can be inferred from a measured displacement.

Assemble the stiffness from a physical experiment

Each platform is attached to a wall by a spring of stiffness $k = 1$ N/m. A coupling spring between platforms has stiffness $c = 49.5$ N/m. Start with the extensions of all three springs, write twice the energy and identify K. Predict the two eigenvectors before computing its eigenvalues.

▼ **Work through the solution**

The extension of the coupling is $x_2 - x_1$. Thus $2U = k(x_1^2 + x_2^2) + c(x_2 - x_1)^2$ and the diagonal entries are $k + c = 50.5$, with off-diagonals $-c = -49.5$. In-phase motion leaves the coupling untouched and has eigenvalue 1. Opposing motion stretches it and has eigenvalue 100. The 2-norm condition number is 100.

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.

Solve one load and verify the result independently

Apply $f = (1, 0)$ N. Solve in modal coordinates, then calculate the elimination pivots and the diagonal of the Cholesky factor and recover the same displacements by a separate route. Which part of x is most difficult to infer from imperfect displacement measurements?

▼ **Work through the solution**

The load is half the sum of $(1, 1)$ and $(1, -1)$. Divide by their stiffnesses to get $x = (0.505, 0.495)$ m. The first pivot is 50.5, the second is

$$\det K / 50.5 = 100 / 50.5 = 1.980198.$$

Eliminating the first row yields this second pivot with right-hand side $49.5/50.5$. These elimination pivots are the diagonal of D in an LDL^T factorization. The positive-diagonal Cholesky factor has their square roots on its diagonal, about 7.106 and 1.407. Back substitution agrees. The difference of the displacements is only 0.01 m. Estimating the large coupling stiffness from that difference is sensitive to small sensor errors.

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.

Change the energy and test the diagnosis

Divide K by 50.5 to give a matrix with unit diagonal. Enter the new off-diagonal into the chapter on positive pivots to the nearest slider value and compare the minimum across slices. Next use the scaling experiment with equal diagonal stiffnesses and the absolute coupling. Explain the residual slow direction.

▼ **Work through the solution**

The normalised coupling is about -0.9802. Its second pivot is $1 - 0.9802^2 \simeq 0.0392$. The almost-flat valley runs along the in-phase motion. Flipping the sign of the second coordinate changes the sign of the coupling and preserves eigenvalues. Diagonal preconditioning cannot separate the scales because the diagonal terms are already equal; a basis that separates the two physical modes can.

Transfer the conclusion to a reported frequency

A unit-mass trial mode is tilted 3° away from the soft mode. Predict its Rayleigh quotient and the frequency bias. Would a two-decimal-place frequency match identify the trial shape?

▼ **Work through the solution**

For eigenvalues 1 and 100 the quotient is $1 + 99\sin^2(3^\circ) \simeq 1.2712$ s⁻². The angular frequency is about 1.1275 rad/s, an error close to 12.8%. At an independently verified angular frequency of 1.00 rad/s to the nearest 0.01, this two-mode model requires $1 + 99\sin^2\theta < 1.005^2$, hence $|\theta| < 0.58^\circ$ up to the arbitrary eigenvector sign. The reading bounds the tilt but does not determine its value or sign. A smaller eigenvalue gap would weaken that bound.

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 · Can an eigensolver certify the pitch of a continuous string?

A taut string gives a known continuum spectrum. Use its kinetic and elastic energies to construct two different matrix pencils. Both can pass a residual test to machine precision while assigning visibly different pitches to the same oscillation.

Build both matrices from a single element

A straight string of length $L = 1$ m, tension $S = 40$ N and mass per unit length $\rho = 0.0040$ kg/m has fixed ends. Divide it into N intervals of length $h = L/N$. Approximate transverse displacement on each interval by a straight line between neighbouring nodal displacements u_i, u_{i+1}. Use this shape for both velocity and slope. Which local mass entries would be lost by assigning half the segment mass to each end before squaring the interpolated velocity?

Integrate the kinetic and elastic quadratic forms

Derive the two 2×2 matrices and explain which physical motion makes the elastic energy vanish.

▼ **Work through the solution**

The interpolated velocity at local coordinate $\xi \in [0, 1]$ is $(1 - \xi)\dot{u}_i + \xi \dot{u}_{i+1}$. The squared shape functions each integrate to $1/3$ and their product integrates to $1/6$. The slope is constant, $(u_{i+1} - u_i)/h$. Comparing the integrals with half a matrix quadratic form gives

$$M_e = \frac{\rho h}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad K_e = \frac{S}{h}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

A common translation costs no elastic energy and a uniform velocity gives exactly the total element mass. Replacing M_e by its diagonal of row sums produces two equal masses $\rho h/2$. This lumped version also gets the uniform-motion energy right but gives more kinetic energy for velocities of opposite sign at the nodes. Assemble contributions at shared nodes and eliminate fixed endpoint degrees of freedom after applying the local row sum. Summing rows only after deleting endpoint columns silently changes the end-adjacent mass.

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.

Diagonalise a sparse pencil without forming an inverse

The interior stiffness row applies $S(2u_j - u_{j-1} - u_{j+1})/h$. Set $u_j = \sin(j\theta)$ so

$$u_{j-1} + u_{j+1} = 2\cos\theta\, u_j.$$

What does each mass row do to the same nodal vector, and which wavenumbers obey both boundary conditions?

Find both discrete dispersion relations

Impose $u_0 = u_N = 0$, discard the zero vector and express the angular frequencies directly.

▼ **Work through the solution**

Use $\theta = m\pi/N$ for $m = 1, \dots, N-1$. The consistent mass row multiplies the nodal sine by $\rho h(2 + \cos\theta)/3$; the lumped row multiplies it by ρh. Inserting $u(t) = ve^{i\omega t}$ into $M\ddot{u} + Ku = 0$ gives

$$Kv = \omega^2 Mv.$$

With $c = \sqrt{S/\rho} = 100$ m/s the results are

$$\omega_{\mathrm{L}} = \frac{2c}{h}\sin(\theta/2),$$

$$\omega_{\mathrm{C}} = \omega_{\mathrm{L}}\sqrt{\frac{3}{2 + \cos\theta}},$$

$$\omega_{\mathrm{exact}} = c\,m\pi/L.$$

The consistent mass includes the exact kinetic energy of each piecewise-linear trial displacement. For this conforming space its Rayleigh–Ritz eigenvalues bound the exact squared frequencies from above. The lumped result is lower because $\sin u < u$ for positive u. Lumping changes the inner product, so the same variational upper-bound argument cannot simply be transferred to it. For a nonuniform mesh the sine vectors need not diagonalise either matrix. A symmetric generalised eigensolver or a Cholesky reduction $M = BB^T$, $B^{-1}KB^{-T}w = \lambda w$ retains the appropriate symmetry. Recover $v = B^{-T}w$ and normalise in the mass inner product.

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.

Distinguish algebraic and discretisation error

Audit a computed fourth harmonic

Use eight intervals. Both sine vectors give tiny relative residual $\|Kv - \lambda Mv\|/[(\|K\| + |\lambda|\|M\|)\|v\|]$ in their own equations. Report the predicted frequency with each mass. Explain how an experimenter could spot the error without an exact continuum reference.

▼ **Work through the solution**

Here $\theta = \pi/2$. The continuum fourth harmonic is 200.0 Hz. The lumped value is $800\sqrt{2}/(2\pi) = 180.06$ Hz, and the consistent value is larger by $\sqrt{3/2}$, giving 220.53 Hz. Each algebraic residual tests whether the vector solves the matrix equation supplied to the solver. Neither residual tests the geometric approximation used to build that equation. Refine while keeping the same physical mode number and

track the same number of interior zeros. Comparing the highest available harmonic after every refinement compares different physical modes.

For a smooth fixed mode the frequency ratios expand as $1 - \theta^2/24 + O(\theta^4)$ for lumping and $1 + \theta^2/24 + O(\theta^4)$ for consistent mass. Doubling the number of intervals quarters the leading bias. Their mean cancels the leading term for this uniform constant-coefficient string. Such cancellation is a consequence of these specific expansions; on a different mesh, solve or measure the convergence order again. Checking mass orthogonality between the reported modes gives a separate test that an eigensolver has returned distinct vectors, particularly when modes approach a cluster.

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.

Measure the usable part of a spectrum

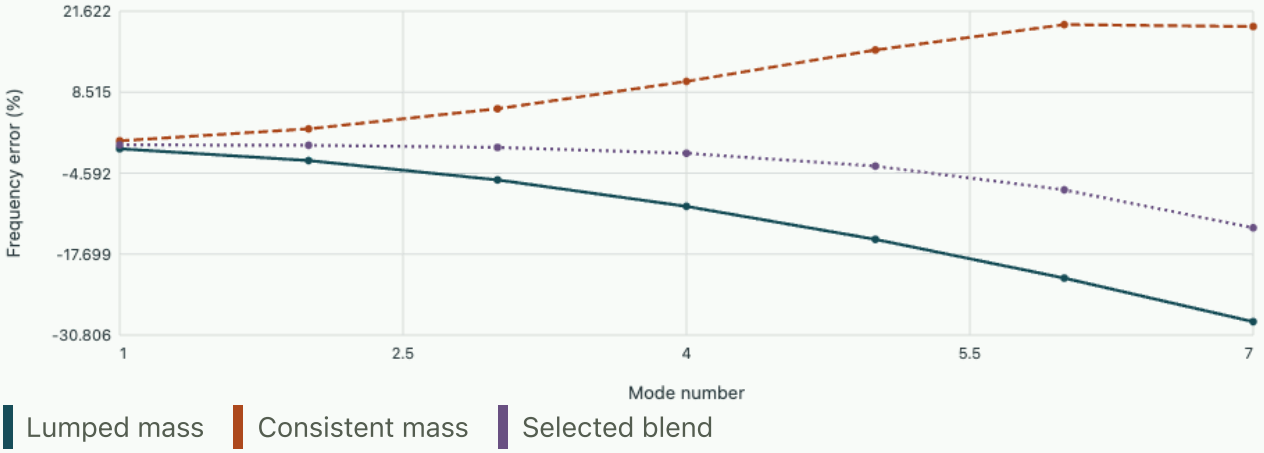
The experiment plots the two percentage errors against mode number over the actual discrete spectrum. Move the interval count while keeping track of mode 4. Look for a mode with about 1% frequency error and estimate how many points per wavelength the mesh must supply. The optional blended mass $M_\alpha = (1-\alpha)M_L + \alpha M_C$ gives another quantitative test of the asymptotic cancellation. Blending mass matrices differs from averaging the two eigenfrequencies at finite h.

Open this experiment online

Try the calculation

Mesh intervals **8**

Consistent-mass fraction **0.5**



Mode 4: exact 200.00 Hz, selected blend 197.25 Hz. The two endpoints are fixed.

Predict a dispersion-cancelling blend

Expand the blended mass multiplier and find the value of α cancelling the leading smooth-mode frequency error. Why can it still give misleading high harmonics?

▼ **Work through the solution**

The mass multiplier divided by ρh is

$$1 - \alpha(1 - \cos\theta)/3 = 1 - \alpha\theta^2/6 + \cdots .$$

Its inverse square root contributes $+\alpha\theta^2/12$ to the frequency ratio. Multiplying by the lumped sine ratio gives leading coefficient $(2\alpha - 1)/24$, which vanishes at $\alpha = 1/2$. At eight intervals mode 4 has a blended frequency $180.06/\sqrt{5/6} = 197.25$ Hz. The improvement is substantial but finite. As the mode approaches the grid scale, the discarded higher powers of θ cease to be small; exact solution of the blended pencil cannot restore a sub-element waveform.

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.

Further reading

- MIT 2.092, lecture 21 · solving the generalised eigenvalue problem

Notebook practice · Counting vibration modes without calculating their shapes

Matrix computations and spectral accuracy

Trace how algebraic structure, representation and numerical precision affect the solution of linear problems. Verify each conclusion with an independent diagnostic.

1. ## When stable eigenvalues allow large transient growth

   Two damped states exchange energy in only one direction. Its eigenvalues predict ultimate decay, while a short-time response can grow by a large factor. Compute the matrix exponential, its gain and a resolvent-based sensitivity check.

2. ## How a random sketch finds a dominant subspace

   An instrument returns a matrix too large to factor densely but cheap enough to multiply by a few vectors. Build a low-rank approximation by probing its range, explain what a random draw can miss and design a residual check using fresh probes.

3. ## Solve a singular network without inventing a ground current

   A floating resistor network has unique voltage drops but no absolute voltage reference. Derive its nullspace, test charge balance, and distinguish adding a computational gauge from changing the physical boundary conditions.

4. ## Refitting a sensor as observations arrive or disappear

   A calibration line receives one new observation at a time. Update its least-squares factorisation using rotations, quantify the influence of a deleted observation and detect when removing a sample destroys identifiability.

5. ## Which resistor can carry its current through another path?

   A short circuit diagram leads to a matrix that describes both voltage differences and sensitivity to losing an edge. Change one conductance and follow where the current goes.

6. ## How a computer can lose a sensor direction

   Two sensors report almost the same quantity. Trace a perturbation through orthogonalisation, a Gram matrix and a rectangular SVD, then decide which computed digits the data support.

7. ## Recovering a spectrum from three-term recurrences

   A large vibrating system supplies products with its stiffness operator. Derive the short recurrence implied by symmetry and use a residual and a moment test to interpret its small projected spectrum.

Matrix computations and spectral accuracy · 1. When stable eigenvalues allow large transient growth

Two damped states exchange energy in only one direction. Its eigenvalues predict ultimate decay, while a short-time response can grow by a large factor. Compute the matrix exponential, its gain and a resolvent-based sensitivity check.

Solve the evolution one component at a time

Consider dimensionless time t and state $x = (x_1, x_2)^T$ with $\dot{x} = Ax$, $A = \left(\begin{smallmatrix} -1 & K \\ 0 & -2 \end{smallmatrix}\right)$, for real $K \geq 0$. The second component decays independently; its transient drives the first. A physical example is a linearised pair of relaxation processes with one readout feeding another. Choose state units so the Euclidean norm represents the amplification of interest. If the coordinates measure unrelated quantities such as volts and kelvins, the unweighted norm changes if one simply changes units. A physical energy or covariance supplies a more appropriate weighted norm.

Starting from $x(0) = (u, v)$, integrate the second row to find

$$x_2 = e^{-2t} v.$$

The first row obeys $(e^t x_1)' = K e^{-t} v$, hence

$$e^{tA} = \begin{pmatrix} e^{-t} & K(e^{-t} - e^{-2t}) \\ 0 & e^{-2t} \end{pmatrix}.$$

Both eigenvalues are negative and the whole matrix exponential tends to zero. Yet for $v = 1, u = 0$ the driven component reaches $K/4$ at $t = \ln 2$, since $e^{-t} - e^{-2t}$ is maximal there. With $K = 12$, the state at this instant is $(3, 0.25)$ and its norm is 3.010. This is a lower bound on the maximum possible gain over unit input vectors and on the maximum over all times. The peak of the driven component need not coincide exactly with the peak of the total norm.

Find the input that grows most at a chosen time

For $M = e^{tA}$, the squared output norm is $x_0^T M^T M x_0$. Its maximum over unit vectors is the largest eigenvalue of $M^T M$, hence the gain is the largest singular value of M. For the triangular matrix above denote its entries by a, b, d on the first diagonal, off-diagonal and second diagonal. The characteristic polynomial of $M^T M$ has trace $s = a^2 + b^2 + d^2$ and determinant $a^2 d^2$. Its larger eigenvalue is $[s + \sqrt{s^2 - 4a^2 d^2}]/2$. This direct formula lets a simulation compare a specified initial condition with the optimal combination at each time. The optimal right singular vector varies with time; a plot of the optimal curve is an envelope of different preparations.

Another test asks whether any instantaneous squared norm can increase at $t = 0$. Its derivative is $x_0^T (A^T + A) x_0$. The largest eigenvalue of the symmetric part $H = (A + A^T)/2$ is $(-3 + \sqrt{1 + K^2})/2$. It is positive precisely when $K > \sqrt{8}$. If H is negative semidefinite every instantaneous state has a nonpositive norm derivative and the Euclidean semigroup is contractive. This criterion concerns energy growth in a specified norm and is distinct from asymptotic stability. At $K = 12$

the eigenvector of H giving fastest initial growth is also not generally the input giving largest long-time response.

Link a small singular value to sensitivity of eigenvalues

For z away from either eigenvalue, direct triangular inversion gives an off-diagonal resolvent entry $K/[(z+1)(z+2)]$. Even at complex points visibly far from the eigenvalues the resolvent norm can be large. To see what that means, let $\sigma_{\min}(zI - A) = \varepsilon$ with unit right and left singular vectors v, u. Then $(zI - A)v = \varepsilon u$. The rank-one perturbation $E = \varepsilon u v^*$ makes $(A + E)v = zv$, while its spectral norm is precisely ε. Conversely any perturbation moving an eigenvalue to z must have norm at least $\sigma_{\min}$. Therefore a contour of $1/\|(zI - A)^{-1}\|$ maps the perturbation required for an eigenvalue there. This is the pseudospectrum construction. Physical perturbations may have additional sparsity or sign constraints, so unstructured contours describe the allowed full matrix perturbations of the stated norm.

Compare one preparation with the worst-case envelope

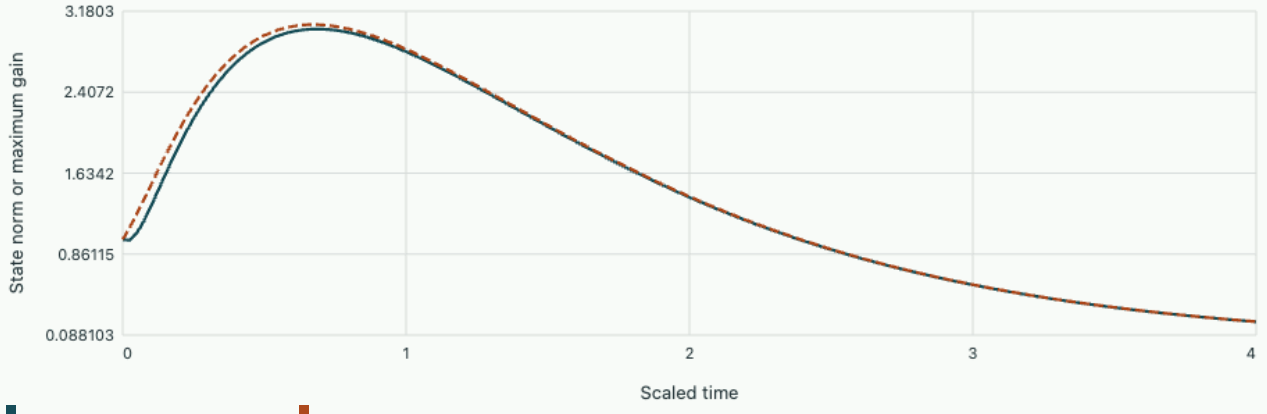
The experiment evaluates the exact triangular evolution and the largest singular value at each time. Change the coupling first, then rotate the unit initial input. At zero coupling both state components decay. At $K = 12$ with initial excitation only in the second coordinate the output norm exceeds three before decaying. Explain why the envelope can exceed every single displayed preparation at different times and why changing the input does not alter either eigenvalue.

Open this experiment online

Try the calculation

Coupling K **12**

Initial angle (degrees) **90**



At t = ln 2, selected norm 3.010, optimal gain 3.051. Both eigenvalues stay −1 and −2.

How a detector can miss the strongest transient

The apparatus often measures a scalar trace $y = c^T x$. With preparation $x(0) = (0, 1)^T$ and a detector aligned with the first coordinate, $y = K(e^{-t} - e^{-2t})$. A detector aligned with the second coordinate instead records e^{-2t}, whose graph decreases immediately. Rotating the detector leaves the evolution untouched. At $t = \ln 2$ and $K = 12$, two orthogonal recordings therefore give 3 and 0.25. Their sum of squares recovers 9.0625, so their joint norm reproduces the transient concealed in the second recording alone.

Given an input direction b and a detector c, the measurable input-to-output amplitude is $c^T e^{tA} b$. The optimal detector for that input is parallel to $e^{tA} b$ by Cauchy–Schwarz. Maximising also over the preparation recovers the singular-value experiment above. Thus a discrepancy between a large theoretical bound and a small experimental trace can arise from either the excitation or the readout. An experiment that rotates only one of them cannot by itself test the unrestricted bound.

Tutorial · verify three distinct amplification diagnostics

Find the exact growth at one time

With $K = 8$, compare the driven component for input $(0, 1)$ and the maximum Euclidean gain at $t = \ln 2$.

▼ **Work through the solution**

The matrix entries are

$$a = 0.5, b = 2, d = 0.25.$$

The specified input gives norm $\sqrt{4.0625} = 2.016$; the driven component itself is 2. The optimum follows from $s = 4.3125$ and $a^2 d^2 = 0.015625$, giving gain about 2.076. Selecting the unit right singular vector of the larger eigenvalue attains it. A direct check is to rotate a unit input around the circle and find the maximum by sampling increasingly fine angles.

Construct a structured perturbation placing an eigenvalue at zero

Keep both diagonal entries fixed and alter only A_{21} by δ. Determine the value needed for a zero eigenvalue and compare it with changing only a diagonal entry.

▼ **Work through the solution**

The determinant becomes $2 - K\delta$. Thus $\delta = 2/K$, or $1/6$ for $K = 12$. Changing only the first diagonal entry would require a shift of one, and changing only the second a shift of two. The smaller lower-left change closes a positive feedback loop. It need not be the smallest unrestricted spectral-norm

perturbation, which is obtained from the singular value of $-A$, but it explicitly demonstrates which physical feedback is implicated.

Decide which conclusion survives nondimensionalisation

Replace the coordinates by $y = (x_1/10, x_2)$. How do the eigenvalues, coupling and interpretation of a Euclidean gain change?

▼ **Work through the solution**

The similar matrix DAD^{-1}, with $D = \mathrm{diag}(0.1, 1)$, has diagonal entries $-1, -2$ and off-diagonal $K/10$. The dynamics represent the same trajectories in different units; the eigenvalues agree. The bare Euclidean norm in y now downweights the original first component, so it can show much less gain. To recover the original norm use $y^T D^{-T} D^{-1} y$. A statement about growth must name the norm and scales, especially when assessing whether a stable numerical model can amplify perturbations beyond a physical tolerance.

Further reading

- Oxford · Trefethen, Pseudospectra of linear operators
- Course investigation · Arnoldi and eigenvalue residuals

Matrix computations and spectral accuracy · 2. How a random sketch finds a dominant subspace

An instrument returns a matrix too large to factor densely but cheap enough to multiply by a few vectors. Build a low-rank approximation by probing its range, explain what a random draw can miss and design a residual check using fresh probes.

Start with the approximation that would be best if its subspace were known

Let A be an $m \times n$ real data matrix. It can represent a collection of spatial snapshots, one per column. A reduced rank-k model stores k spatial patterns and a small coefficient matrix. If a matrix Q has orthonormal columns spanning a proposed range, then the closest approximation with columns in that range in Frobenius norm is QQ^TA: each snapshot is independently orthogonally projected. Store $B = Q^TA$; its singular-value decomposition is small when Q has relatively few columns. The distinction between finding the subspace and factoring within it is important. A perfect small SVD cannot recover a spatial direction missing from Q.

If $A = U\Sigma V^T$ is exactly rank k, choose $n \times k$ test matrix Ω and form $Y = A\Omega = U_k\Sigma_k V_k^T\Omega$. If the last $k \times k$ factor is invertible, Y spans exactly the range of A. Independent Gaussian columns make singularity a zero-probability event in exact arithmetic, because the rotated coefficients are again a continuous Gaussian matrix. This statement does not imply every draw has a well-conditioned factor, nor that a single probe spans every near-rank-k matrix. Orthogonalising Y with a numerically reliable QR yields the candidate basis Q.

Derive the effect of gap and starting alignment

Consider rank reduction of $A = \operatorname{diag}(10, 1)$ with one unit probe $\omega = (\cos\theta, \sin\theta)$. Its image is $(10\cos\theta, \sin\theta)$, so the angle ϕ between the sampled direction and dominant left singular vector obeys $\tan\phi = \tan\theta/10$ whenever $\cos\theta \neq 0$. A typical mixed direction is pulled toward the dominant range. A nearly perpendicular probe can still give a poor direction even with a substantial singular gap. For $\theta = 45^\circ$, $\tan\phi = 0.1$. The residual $(I - qq^T)A$ has only one nonzero singular value because its left range is one-dimensional. With $q = (\cos\phi, \sin\phi)$, its norm squared is $100\sin^2\phi + \cos^2\phi$, giving $\sqrt{200/101} = 1.407$. The optimal rank-one error is the discarded singular value 1.

One power step forms $Y = (AA^T)A\Omega$, raising singular values to the third power in the sampled direction while the approximation is still formed for the original A. The angle ratio shrinks to $\tan\theta/1000$; the residual norm is then about 1.00005 for the same 45° probe. Additional steps require more passes over data. In finite precision reorthogonalise between multiplications so weak required directions are not erased by a large gap. If the target includes both leading and moderate directions, blindly powering a block without intermediate QR can make the very reduced subspace one wants numerically rank deficient.

Independently test an approximation after building it

Add p extra columns to Ω to reduce the chance of poorly exciting a dominant combination. An oversampled Q has $k + p$ columns; project to $B = Q^T A$, take its dominant k small left singular vectors and map them back through Q. For an exactly rank-k input the extra columns need appropriate numerical rank handling. In the nearly low-rank case they provide room to represent part of the tail and improve the final cut. Cost depends on access: computing $A^T Q$ or performing every additional power pass can dominate on streamed or distributed data, even though the final dense SVD is tiny.

Freeze the fitted approximation $\tilde{A}$ and consider an independent random sign vector g with entries equally likely ± 1. Write $R = A - \tilde{A}$. Since $\mathbb{E}[gg^T] = I$,

$$\mathbb{E}\|Rg\|^2 = \operatorname{tr}(R^T R) = \|R\|_F^2.$$

Multiple independent checks estimate residual energy and its sampling spread without explicitly assembling R. This estimates Frobenius norm squared; the maximum directional error is a different objective. A single small response does not certify a small spectral norm or guarantee missed directions are unimportant to a particular downstream measurement. Test the physically relevant load or measurement directions as well.

Background · Why must the test vector be independent of the sketch?

Build the sampled range Q by orthogonalising $A\Omega$. With exact arithmetic every sampled column already lies in the range of Q. Applying the residual to those same vectors gives

$$(I - QQ^T)A\Omega = 0.$$

For example, one column sampled from the two-dimensional identity spans only a line. It reconstructs its own training vector exactly while the unit vector perpendicular to it has error one. A fresh isotropic probe has, on average, nonzero energy in every missing direction. For a standard Gaussian vector g independent of Q, the mean squared diagnostic is $\mathbb{E}\|Rg\|^2 = \|R\|_F^2$, with $R = (I - QQ^T)A$. This identity follows by expanding $g^T R^T R g$ and using $\mathbb{E}[gg^T] = I$. One draw is still a noisy diagnostic, especially when only a single direction is missing.

Which starting direction defeats a random range probe?

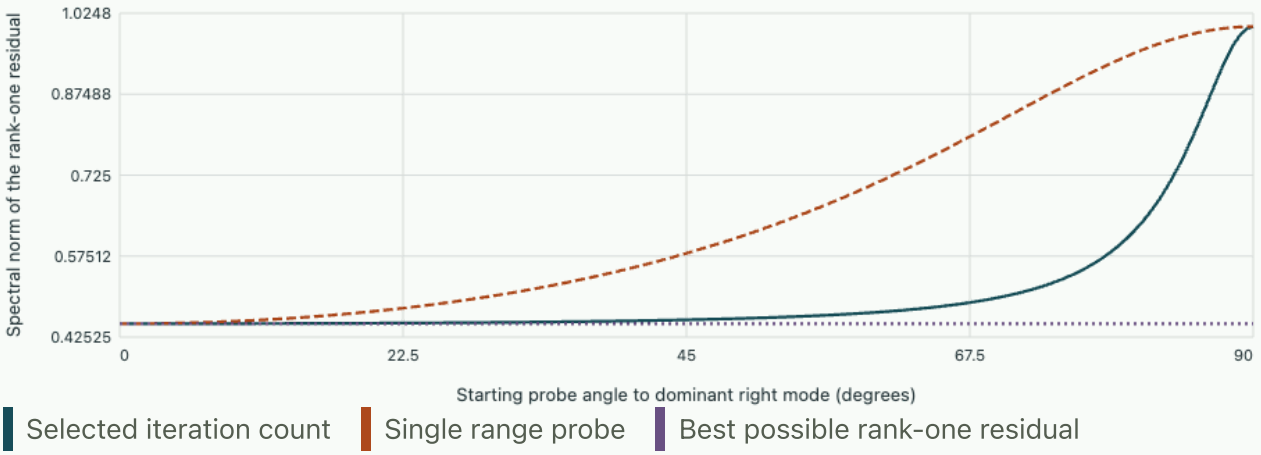
For a diagonal matrix with singular values 1 and r, rotate a test vector through the dominant right singular direction. Each power iteration multiplies the tangent of its output angle by r². Set r to 0.45, compare zero and two iterations, then inspect the exactly orthogonal probe at 90°. Explain why its residual stays at one.

Open this experiment online

Try the calculation

Singular-value ratio σ_2/σ_1 **0.45**

Power iterations **1**



At 60°, residual 0.47105 against the best possible 0.45. A probe orthogonal to the leading mode gives residual one at every iteration count.

Tutorial · check an apparently good random approximation

Identify a failure when the test vector is reused

A matrix A is approximated by the orthogonal projection onto the range of $A\Omega$. You test residuals only on the original columns of Ω and find zeros. What does this certify?

▼ **Work through the solution**

By construction $QQ^TA\Omega = A\Omega$, up to roundoff if the full sampled range was kept. Thus the test is nearly tautological and does not measure error on independent directions. If Ω has a missing dominant right singular direction, all these residuals can vanish while the leading reconstruction is poor. Use fresh random probes and targeted physical snapshots held out during fitting. If a final rank truncation is applied within Q, the original probe residual need no longer vanish exactly but remains dependent on the fitted subspace.

Calculate a deliberately poor draw

Use $A = \mathrm{diag}(10, 1)$ and $\tan\theta = 100$. Compare the one-pass angle and residual with one power iteration.

▼ **Work through the solution**

Initially $\tan\phi = 10$, so $\sin^2\phi = 100/101$ and the residual norm squared is $(10000+1)/101$, giving about 9.95. With one power step $\tan\phi = 0.10$, giving 1.407 instead. The bad starting direction had a small but nonzero overlap, so powering can recover it. An exactly perpendicular deterministic probe remains perpendicular after every power step. Oversampling adds independent opportunities to see a missed direction.

Design a stopping criterion for snapshot energies

Independent residual probes give squared norms $1.0, 1.4, 0.8, 1.6, 1.2$ in scaled energy units, while independently known $\|A\|_F^2 = 100$. Estimate the residual energy fraction and its sampling uncertainty. Does it imply a 1.2% relative error on every snapshot?

▼ **Work through the solution**

The sample mean is 1.2 and sample standard deviation $\sqrt{0.10} = 0.316$, so the standard error is about 0.141 for independent probes. The estimated fraction of total squared Frobenius norm is 1.20%, with approximate standard error 0.14 percentage points. The relative Frobenius norm itself is about $\sqrt{0.012} = 0.110$. Individual snapshots may fare much better or worse; the tests estimate a global aggregate. Five samples give only a rough uncertainty estimate, particularly if residual energy is concentrated in a few directions.

Further reading

- Caltech · Halko, Martinsson and Tropp on randomised matrix decompositions
- Course investigation · singular values and low-rank error

Matrix computations and spectral accuracy · 3. Solve a singular network without inventing a ground current

A floating resistor network has unique voltage drops but no absolute voltage reference. Derive its nullspace, test charge balance, and distinguish adding a computational gauge from changing the physical boundary conditions.

Build the conductance matrix as an energy Hessian

Orient each of the m connections of a connected graph and form an incidence matrix B with one row per edge, containing $+1$ at its upstream node and -1 at its downstream node. For node voltages v, differences Bv determine currents $i_e = g_e(Bv)_e$ in edges of conductance $g_e > 0$. Conservation at nodes reads $B^T i = b$, where externally injected currents b are positive into the network. Eliminating edge currents yields $Lv = b,\ L = B^T G B$ with $G = \mathrm{diag}(g_e)$. Reversing an edge changes a row of B and its current sign together; L and physical voltage predictions are invariant.

The identity $v^T L v = \sum_e g_e (v_i - v_j)^2 \geq 0$ gives the nullspace directly. Equality forces equal voltage along each edge and hence constant voltage on a connected graph. Thus $\ker L = \mathrm{span}\{\mathbf{1}\}$ and its range is $\mathbf{1}^\perp$. A solution exists exactly when total injection is zero. On a graph with disconnected components the current sum must vanish separately on each component; one global cancellation can hide two incompatible islands. This is a compatibility condition on the physical data, distinct from the gauge freedom that persists for already compatible data.

Compare grounding, projection and a bordered solve

If b is compatible, choosing one node potential as zero removes one variable and one redundant conservation equation. The remaining principal submatrix is positive definite because a zero-energy vector extending with zero at the reference must vanish everywhere. An alternative chooses mean-zero voltage without favouring one node. Introduce a Lagrange multiplier λ and solve

$$\begin{pmatrix} L & \mathbf{1} \\ \mathbf{1}^T & 0 \end{pmatrix} \begin{pmatrix} v \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}.$$

The last equation fixes the gauge. Multiplying the first block by $\mathbf{1}^T$ gives

$$\lambda = (\sum b_i)/n.$$

For consistent data $\lambda = 0$, so the bordered solution satisfies the original equations exactly and is the Moore–Penrose minimum-norm solution. For inconsistent data the solver effectively removes a uniform current λ from every node. This may be sensible as a least-squares projection under an explicitly chosen equal-error model, but it is not the same physical network as a returning current concentrated at an actual external lead. Retain and interpret the multiplier.

Another formulation works directly in the zero-mean subspace with projector $P = I - \mathbf{1}\mathbf{1}^T/n$. Iterative methods can project the right side and reproject iterates to remove accumulated constant drift. A preconditioner must preserve the intended subspace or be followed by the appropriate projection. The

gauge-invariant observable $v_i - v_j$ is always preferable for comparing solvers with different references. Comparing raw voltages node by node can report a large error between mathematically equivalent physical solutions. A separate edge-current balance and dissipated power check validate more than one reference-dependent vector residual.

Verify a triangle by current conservation and work

Connect nodes 1–2 and 2–3 with 1 S each, and 1–3 with 0.5 S. Inject 1 A at node 1 and remove it at node 3. The direct branch has 2 Ω and the two-edge indirect branch also totals 2 Ω, so 0.5 A flows along each path. The terminal difference is 1 V. Taking $v_3 = 0$ gives $v = (1, 0.5, 0)$ V, while subtracting their mean yields $(0.5, 0, -0.5)$ V. Both produce Joule power $2(0.5^2/1) + (0.5^2/0.5) = 1$ W. The terminal injection work is $b^T v = 1$ W in either gauge because the sum of b vanishes.

A guard against concealed modelling error is to solve first with all conservation rows, reporting the incompatible component of b; only then introduce the intended reference. If measured injected currents are $1.00, 0, -0.97$ A and node 3 is grounded by deleting its equation, the computed branch currents imply that node 3 actually returns 1.00 A. The measured 0.03 A mismatch is thereby absorbed entirely by the chosen reference. The bordered projection instead distributes 0.01 A equally among the nodes. A calibrated return-lead model or a covariance-weighted correction can choose between such interpretations. A small residual after row deletion alone cannot tell whether the three measured currents were physically consistent.

What changes when a floating voltmeter is re-zeroed?

Connect three nodes by a triangle. The 1–3 and 2–3 conductances are each 1 S. Inject the selected current at node 1 and remove it at node 3. The graph sweeps a common reference-voltage shift. Use the oriented drop across branch 1–2 to predict its current, then compare the total injected power with the displayed dissipation.

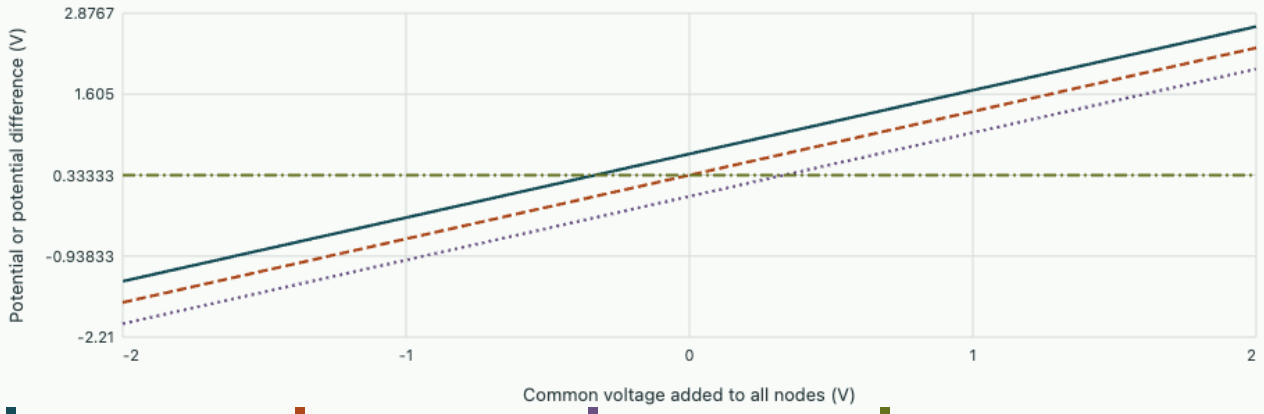
The downloadable OneSimulation circuit solves the three resistors and current source twice, choosing a different reference node. The two absolute voltage reports change. Their branch drops must agree.

Open this experiment online

Try the calculation

Conductance between nodes 1 and 2 (S) **1**

Current from node 1 to node 3 (A) **1**



Branch current I_{12} = 0.33333 A. Total Joule dissipation $I(V_1–V_3)$ = 0.66667 W at every gauge.

Download the OneSimulation model

In your OneSimulation environment, run the model with these inputs.

```
python networks-and-calibration.py network --conductance 1 --current 1
```

Tutorial · audit gauges, disconnected graphs and sensitivities

Find an effective resistance without choosing a reference

For connected L, point injection pair $b = I(e_a - e_b)$ and minimum-norm inverse L^+, derive the effective resistance and verify its sign.

▼ Work through the solution

The gauge-fixed solution is $v = L^+ b$, so terminal voltage divided by I is $R = (e_a - e_b)^T L^+ (e_a - e_b)$. Positive definiteness on the zero-sum subspace gives nonnegative R, strictly positive for distinct connected nodes. In the example it is 1 Ω, independently confirmed by the parallel branches. Summing injected power $I^2 R$ matches edge dissipation.

Determine how much one connection is worth

What is the first derivative of the triangle's effective resistance with respect to the direct conductance? Compare a prediction for raising it by 0.01 S with a direct circuit reduction.

▼ **Work through the solution**

For unit injection, differentiating $Lv = b$ on the fixed gauge subspace gives

$$dR = -v^T(dL)v.$$

A single conductance change has $dL = (e_i - e_j)(e_i - e_j)^T dg$, so

$$dR/dg = -(v_i - v_j)^2.$$

The direct drop is 1 V per ampere, giving -1 in compatible units. Prediction is $R \simeq 0.990\ \Omega$; exact parallel conductance becomes 1.01 S, so $R = 0.99010\ \Omega$. A positive conductance increment cannot increase effective resistance here.

Diagnose an island and an unobservable offset

Remove every link between two internally connected components carrying net injections $+0.10, -0.10$ A. Is one ground constraint enough? What if both component sums are corrected to zero?

▼ **Work through the solution**

The disconnected graph has two independent constant null vectors, one per component. Although the total current is zero, both component compatibility tests fail. No voltage gauge can make the missing inter-island return path appear. If both sums become zero, two independent gauge constraints still remain necessary for unique node voltages. Potential differences within either component are uniquely determined; differences across the two floating components remain arbitrary unless a physical reference or coupling is introduced.

Further reading

- ETH Zürich · numerical methods for CSE, sparse systems and positive definite forms
- Course investigation · constrained systems and Schur complements

Matrix computations and spectral accuracy · 4. Refitting a sensor as observations arrive or disappear

A calibration line receives one new observation at a time. Update its least-squares factorisation using rotations, quantify the influence of a deleted observation and detect when removing a sample destroys identifiability.

Rotate away a new observation while preserving the objective

Suppose A has full column rank p, observations b, and an orthogonal factorisation has already transformed the upper part of the problem into $Rx = z$ with upper-triangular R. Lower transformed rows carry residuals but no coefficients. Because the transformations are orthogonal,

$$\|Ax - b\|^2 = \|Rx - z\|^2 + s^2,$$

where s^2 is the accumulated minimum residual sum of squares. Appending row a^T and response y requires minimising $\|Rx - z\|^2 + (a^T x - y)^2 + s^2$. One can update the small augmented triangle instead of reloading every original sample. This is particularly useful when storing raw observations is expensive and fits need refreshing repeatedly.

Write the new row directly below $[R\ z]$. At column j, combine diagonal value r with bottom value q using a Givens rotation with $c = r/\operatorname{hypot}(r,q),\ s = q/\operatorname{hypot}(r,q)$. Multiplication of the two selected rows by $\left(\begin{smallmatrix} c & s \\ -s & c \end{smallmatrix}\right)$ gives a nonnegative diagonal $\sqrt{r^2+q^2}$ and zero in the bottom column. Apply the same rotation to every remaining column and to the right side. If both inputs vanish choose the identity and investigate numerical rank separately. A robust hypotenuse evaluation avoids overflow when $r^2 + q^2$ itself exceeds the available range even though its square root is representable.

After sweeping all p columns the bottom row is zero in coefficients and has transformed response γ. The new triangular solve gives the updated fit and the minimum residual sum rises by γ^2. Each step preserves a squared norm, so this identity is a direct independent audit of online accumulation. Weighted least squares fits into the same calculation by scaling the row and response by the square root of its weight first. Dynamic reweighting of historical observations is a different operation and generally requires preserved information about individual rows.

Use one-dimensional corrections for interpretation

The rank-one update also has a useful algebraic form without explicitly computing a dense inverse. Set $G = A^T A$, $w = G^{-1}a$ evaluated by two triangular solves with R, and $h = a^T w$. After adding a point the normal matrix becomes $G + aa^T$. Subtracting the old normal equations from the new ones gives $(G + aa^T)\delta x = a(y - a^T x)$, so $\delta x = w(y - a^T x)/(1 + h)$, since $(G + aa^T)w = (1 + h)a$. The increment to residual sum of squares is $(y - a^T x)^2/(1 + h)$. This formula for diagnostics is consistent with the orthogonal update and shows explicitly how a surprising response interacts with geometric leverage.

As an example fit a line $y = c + dt$ at times $t = -1, 0, +1$ with readings $0, 1, 2$. The exact line is $c = d = 1$, with $G = \operatorname{diag}(3, 2)$. A new observation at $t = 2$ reports $y = 4$ rather than predicted 3. Here $a = (1, 2)^T$, $w = (1/3, 1)^T$, $h = 7/3$. The corrected coefficients are $c = 1.1,\ d = 1.3$, and residual

sum rises by $1/(1+7/3) = 0.30$. Substitution back into all four data gives residuals $0.2, -0.1, -0.4, 0.3$; their squared sum is indeed 0.30. The update has moved a previously exact historical fit because the new reading changed the estimated slope and intercept.

Distinguish a safe deletion from rank loss

For deleting a row already included in the fitted G, use that row's in-sample leverage $\ell = a^T G^{-1} a$. It lies between zero and one as a diagonal element of the orthogonal projection $A(A^T A)^{-1} A^T$. If $\ell < 1$, the remaining-data fit changes by $-G^{-1} a\, r/(1-\ell)$, where $r = y - a^T x$ is the residual while the point was present. The leave-one-out prediction residual is $r/(1-\ell)$. A point with small in-sample residual can therefore be highly influential when nearly essential to identification. If $\ell = 1$, deletion loses rank, and a formula dividing by $1-\ell$ is warning of a structural change. Downdating a triangular factor should use a well-conditioned dedicated downdate or refactor with rank detection when close to this boundary. Squaring a poorly conditioned design into G just for convenience can lose accuracy.

Refit the line after one influential observation

Begin with three calibration pairs on y = x at x = –1, 0 and 1. Place a fourth sensor farther from the centre and choose its deviation from the old line. Predict its training residual and the error made when the entire fourth observation is omitted. Move it farther out with a fixed nonzero offset. Which residual approaches zero and which retains the full offset? Set the offset to zero and check whether a high-leverage point must also have a large residual.

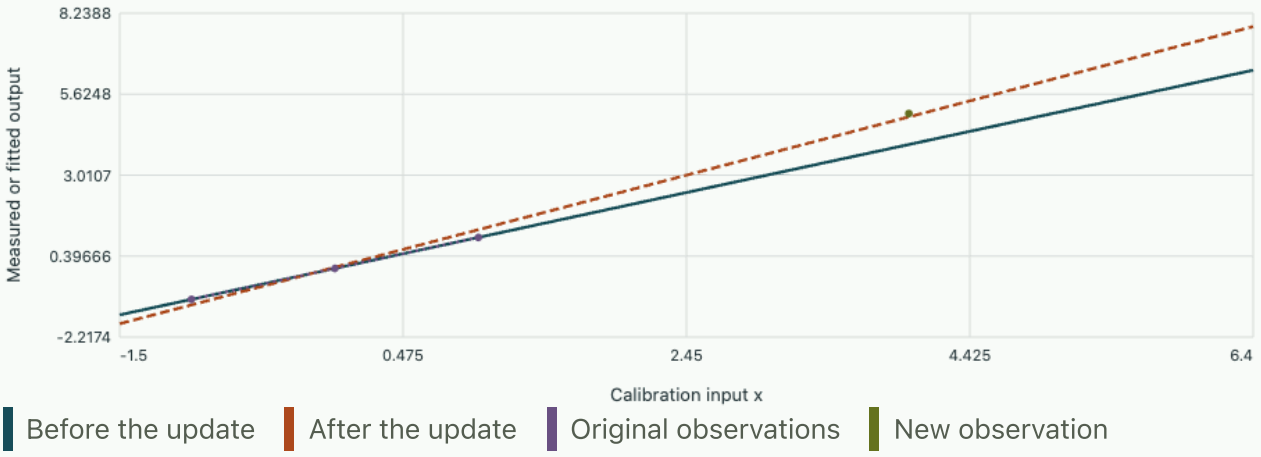
The OneSimulation model fits the full augmented observation matrix with its dense least-squares solver. Compare this independent refit with the one-row update on screen.

Open this experiment online

Try the calculation

New observation coordinate 4

New observation's offset from y = x 1



Updated slope 1.2143, intercept 0.035714. New row leverage 0.89286; deleted prediction residual 1 versus training residual 0.10714.

Download the OneSimulation model

In your OneSimulation environment, run the model with these inputs.

```
python networks-and-calibration.py line --coordinate 4 --offset 1
```

Tutorial · inspect an update from three angles

Perform a complete scalar QR update

Three readings of a constant have average 2 and residual sum 0.50. Add value 5 at equal weight. Find the updated average, residual increment and scalar triangle.

▼ **Work through the solution**

Before appending,

$$R = \sqrt{3},\ z = 2\sqrt{3}.$$

With appended row 1 the new diagonal is 2. The rotated right side is $(\sqrt{3}/2)(2\sqrt{3}) + (1/2)5 = 5.5$, giving average 2.75. The bottom response is $(-1/2)(2\sqrt{3}) + (\sqrt{3}/2)5 = 1.5\sqrt{3}$, contributing 6.75 to sum of squares and giving new total 7.25. The mean from the raw sum is also $(6 + 5)/4 = 2.75$.

Reverse the four-point line update

For the newly appended point in the line example, infer its in-sample leverage from the original append leverage, then recover the original leave-one-out residual.

▼ **Work through the solution**

Adding its own row changes leverage from h to $\ell = h/(1+h) = 0.70$, obtained by solving $(G + aa^T)^{-1}a = w/(1+h)$. Its current fitted value is $1.1 + 2(1.3) = 3.7$, giving residual 0.3. Dividing by $1 - \ell = 0.3$ recovers the predicted deletion residual 1.0. Confusing h with ℓ would incorrectly put a leverage larger than one inside a leave-one-out formula.

Detect a geometry where a zero residual is inevitable

Fit a line with many observations at $t = 0$ and exactly one at $t = 1$. Show why the latter's in-sample residual can be zero for any value and why its fit quality cannot validate the slope.

▼ **Work through the solution**

All $t = 0$ samples determine the intercept by their mean. The unique nonzero time then fixes the slope to hit its own response exactly, so its residual is always zero and its in-sample leverage is one. Removing it leaves only one time coordinate and the slope unidentifiable. An extra independent observation at a distinct time, or a justified physical slope prior, is needed to test predictive slope accuracy. Increased repetitions at $t = 0$ reduce intercept noise without repairing this missing diagnostic.

Further reading

- ETH Zürich · numerical methods for CSE, least squares and orthogonal factorisation
- Course investigation · orthogonal reduction for least squares
- Separate algebraic solver error from mesh error

Matrix computations and spectral accuracy · 5. Which resistor can carry its current through another path?

A short circuit diagram leads to a matrix that describes both voltage differences and sensitivity to losing an edge. Change one conductance and follow where the current goes.

Ground a node and derive the electrical leverage

Nodes A, B and C form a triangular loop. Node D connects only to C. Every edge is an ideal linear resistor. Orient an edge from node i to node j and give it an incidence vector b_e with a +1 at i and a −1 at j. Its current under nodal voltages v is $g_e b_e^T v$, where conductance g_e is measured in siemens. Kirchhoff's current law sums all outgoing currents into the net injection f. It gives

$$Lv = f,$$
$$L = \sum_e g_e b_e b_e^T.$$

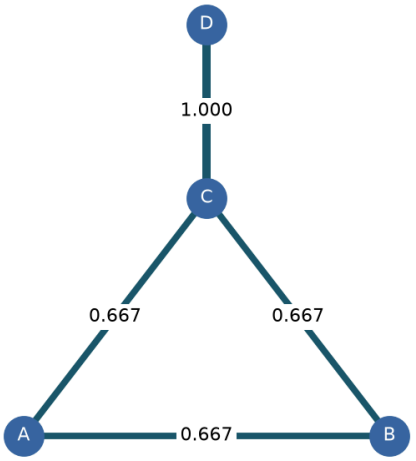
Adding a constant to every potential preserves all currents. Fix the potential of D at zero and remove its row and column. For a connected network the resulting grounded matrix is positive definite. Inject one ampere at one end of an edge and withdraw it at the other. With grounded coordinates the voltage difference per unit current is the effective resistance $R_e = b_e^T L_g^{-1} b_e$. The fraction of this test current that flows through the tested edge is $\ell_e = g_e R_e$. A bridge carries the full ampere. A parallel path reduces its share.

There is also a global check on the calculation. The sum of all the leverages is $\operatorname{tr}(L_g^{-1} \sum_e g_e b_e b_e^T) = n - 1$ on a connected graph with n nodes. Each leverage depends only on the graph and its conductances. The actual current in a specified operating condition also depends on the chosen source and sink.

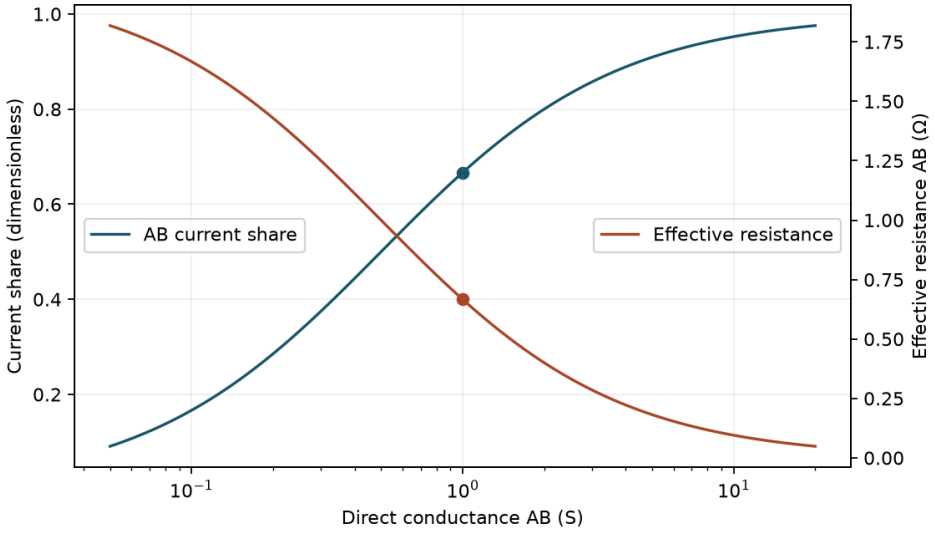
Solve the triangle by hand

With 1 S on every branch, the detour A–C–B consists of two 1 Ω resistors in series. Inject 1 A at A and extract it at B. The detour's 0.5 S combines with the direct 1 S, so the measured drop is $1/1.5 = 2/3$ V. The direct current is 2/3 A. The dead-end C–D branch carries zero current in this experiment. When tested between its own endpoints, however, C–D carries the entire test current. The four leverages are consequently $2/3, 2/3, 2/3, 1$ and sum to three.

Double A–B's conductance to 2 S. Its leverage rises to $2/2.5 = 0.8$, while the voltage needed for the same 1 A drops to 0.4 V. This is why the two traces in the experiment have opposite slopes. Rerun with a weaker bridge. Its resistance increases, yet its leverage stays equal to one.



Each label gives the share of a separate one-ampere test between that edge's endpoints. The pendant branch remains at unity for any positive conductance.



The current-share curve tends to one as the direct branch dominates the detour. The voltage per ampere falls from two ohms toward zero.

Exercise · remove the direct branch

How much does the terminal resistance rise when the 1 S direct edge is removed? What changes if the bridge is removed instead?

▼ **Work through the solution**

For A and B the sole remaining connection consists of two 1 Ω resistors, so resistance rises from 2/3 Ω to 2 Ω, an increase of 4/3 Ω. Removing C–D instead isolates D. Measurements within the triangle remain finite, but a test injecting current at D and withdrawing it at C no longer has a finite DC voltage solution.

Exercise · differentiate the measured voltage

Under a fixed one-ampere A-to-B test, find the sensitivity of the terminal resistance to the direct conductance at 1 S.

▼ **Work through the solution**

For variable g the resistance is $1/(g+0.5)$. Differentiation gives $-1/(g+0.5)^2 = -4/9$ Ω/S at the chosen point. The magnitude equals the square of the voltage drop per ampere across that branch. Differentiating $L_g v = f$ at fixed injection gives the same result for any branch, $\partial(f^T v)/\partial g_e = -(b_e^T v)^2$ for the unit test.

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

Direct AB conductance (S) 1

Bridge CD conductance (S) 1



Resistance is normalised by 1 Ω on this shared dimensionless axis. Selected AB resistance 0.66667 Ω, leverage 0.66667. Bridge resistance 1 Ω and bridge leverage 1; the four edge leverages sum to 3.

Open editable notebook

Download .ipynb

Matrix computations and spectral accuracy · 6. How a computer can lose a sensor direction

Two sensors report almost the same quantity. Trace a perturbation through orthogonalisation, a Gram matrix and a rectangular SVD, then decide which computed digits the data support.

Subtracting a projection in finite precision

Let A be a real m-by-n matrix, with $m \geq n$ and independent columns. The first unit vector is $q_1 = a_1/\|a_1\|_2$. To construct another, classical Gram–Schmidt computes all coefficients $r_{ij} = q_i^T a_j$ against the original column before subtracting the sum. Modified Gram–Schmidt instead updates a running residual v after each projection, computing the next coefficient from that residual.

$$v^{(0)} = a_j,$$

$$r_{ij} = q_i^T v^{(i-1)},$$

$$v^{(i)} = v^{(i-1)} - q_i r_{ij}.$$

In exact arithmetic the two formulas agree by orthogonality. In floating point, a component along an earlier q_i can survive subtraction. Dividing by the norm of a nearly cancelled column amplifies that component. For the three columns of the matrix with first row $(1, 1, 1)$ and lower block δI_3, the difference of two original columns has size proportional to δ. Computing both of the third column's coefficients before either subtraction carries the error of previously computed directions into both coefficients. Sequential projection removes some of that contamination. A second pass projects the surviving residual again. Monitor both $\|A - QR\|/\|A\|$ and $\|I - Q^T Q\|$. A small first diagnostic alone permits poor orthogonality.

Why a reflector gives a backward-error result

Given nonzero x, set $v = x + \operatorname{sign}(x_1)\|x\|e_1$, using positive sign at zero. Then $H = I - 2vv^T/(v^T v)$ is orthogonal and Hx has only its first entry nonzero. The sign ensures the construction of v_1 adds magnitudes. Assume normal-range floating-point arithmetic with unit roundoff u and $mu \ll 1$. A length-m dot product has absolute error bounded by $\gamma_m |v|^T |x|$, with $\gamma_m = mu/(1 - mu)$, by multiplying the factors $1 + \delta_i$, $|\delta_i| \leq u$, along the sum. Cauchy–Schwarz bounds this by $\gamma_m \|v\|\|x\|$.

Apply a stored, normalised reflector as a dot product and an axpy. Each column incurs a perturbation of order mu times its norm, including normalisation error. The exact reflector for the stored direction has norm one. Pull each local error backwards through the preceding exact orthogonal transformations and add. In Frobenius norm this yields

$$A + \Delta A = Q\widehat{R}, \qquad \|\Delta A\|_F \leq Cmnu\,\|A\|_F + O(u^2),$$

where C accommodates the implementation's fixed number of elementary operations. This explains the nature of the bound without claiming an implementation-independent sharp constant. A tiny singular value may still have large relative uncertainty under this absolute perturbation.

Turn a rectangular matrix into a bidiagonal one

Apply a left reflector to column 1 and a right reflector to row 1, starting its operation in column 2. Repeat on the trailing block. Each transformation preserves earlier zeros. This yields $A = UBV^T$ with orthogonal factors and upper bidiagonal B. As a concrete first step take rows $(3, 0), (4, 5), (0, 12)$. Multiplying the first two rows by the rotation with rows $(3/5, 4/5), (-4/5, 3/5)$ leaves rows $(5, 4), (0, 3), (0, 12)$. A second left rotation combines $3, 12$, giving diagonal entries $5, \sqrt{153}$ and superdiagonal 4. Two columns make the right step trivial here. A third column would require eliminating the new first row's tail.

For a bidiagonal matrix one can carry out shifted QR on $B^T B$ implicitly. If the first diagonal and off-diagonal entries of that tridiagonal matrix are a, b, select a right rotation from $(a - \mu, b)$. Applying it to B introduces a subdiagonal entry. A left rotation eliminates that entry and moves the disturbance to the next column pair. Repeat to the bottom and accumulate both sets of rotations. Vanishing superdiagonals split independent blocks. This obtains singular values without explicitly storing an already rounded Gram matrix. For two almost parallel sensor columns $(1, \delta, 0)^T, (1, -\delta, 0)^T$ the singular values are $\sqrt{2}, \sqrt{2}|\delta|$. The entries $1 \pm \delta^2$ of the Gram matrix can round to the same number.

OneSimulation experiment · Resolve the smaller singular value

Change the separation grid. Which lost singular value can a direct SVD still measure?

Download the executable notebook · Python source

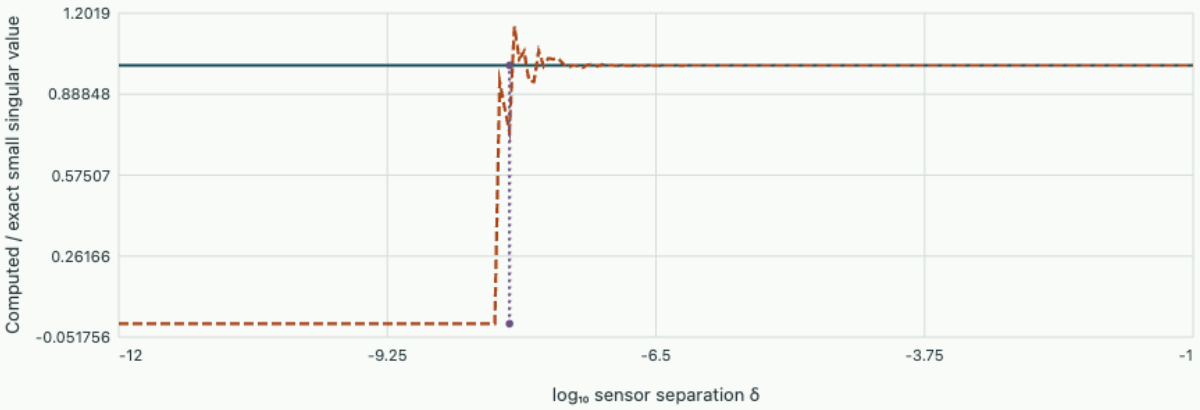
Adjust the experiment in the browser. Move the selected separation across the loss threshold and switch between binary32 and binary64 arithmetic. The Gram estimate evaluates the rounded difference $(1 + \delta^2) - (1 - \delta^2)$; the stable comparison uses the exact orthogonality of the two nonzero rows. It is an analytic benchmark, while the notebook below runs the rectangular SVD and OneSimulation eigenvalue solver. Locate the threshold near $\delta \sim \sqrt{u}$, then explain why changing arithmetic precision cannot recover information already lost in noisy measurements.

Open the live experiment

Try the calculation

Selected $\log_{10}\,\delta$ **-8**

Arithmetic bits **64**



Stable formula from orthogonal rows

Rounded Gram-matrix difference

Selected separation

At δ = 1e-8, exact σmin = 1.4142e-8, condition number = 100000000. The 64-bit Gram calculation retains 0.74506 of the small singular value. The reference curve uses the analytic row-orthogonality formula, not an implemented SVD.

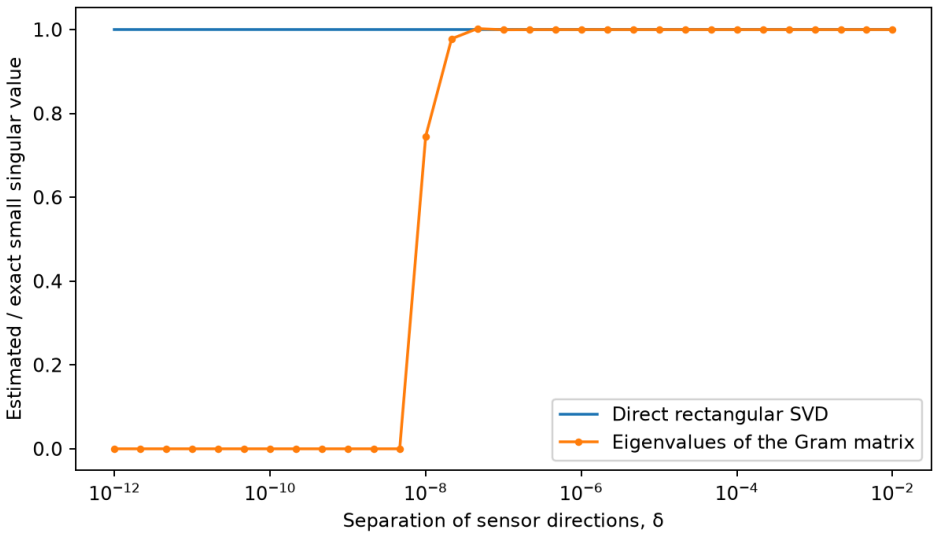
```
"""Measure loss of the weaker of two almost parallel sensor directions."""
import matplotlib.pyplot as plt
import numpy as np
from onesim.numerics.linear.matrices.dense_linear_algebra import symmetric_eigenvalues

separations = np.logspace(-12, -2, 31)
direct, normal = [], []
for delta in separations:
    matrix = np.array([[1.0, 1.0], [delta, -delta], [0.0, 0.0]])
    reference = np.sqrt(2.0) * delta
    direct.append(np.linalg.svd(matrix, compute_uv=False)[-1] / reference)
    smallest = symmetric_eigenvalues(matrix.T @ matrix)[0]
    normal.append(np.sqrt(max(smallest, 0.0)) / reference)

assert np.allclose(direct, 1.0)
assert normal[0] == 0.0
print(f"At delta={separations[0]:.0e}: direct/reference={direct[0]:.3f}, Gram/reference=
{normal[0]:.3f}")
fig, ax = plt.subplots(figsize=(7, 4), layout="constrained")
ax.semilogx(separations, direct, label="Direct rectangular SVD")
ax.semilogx(separations, normal, "o-", ms=3, label="Eigenvalues of the Gram matrix")
ax.set(xlabel="Separation of sensor directions, δ", ylabel="Estimated / exact small singular value")
ax.legend()
fig.savefig("numerics-small-singular-value.png", dpi=170)
```

Output from the reference run

```
At delta=1e-12: direct/reference=1.000, Gram/reference=0.000
```



A Gram-matrix eigenvalue collapses as the separation of two sensor directions decreases.

Tutorial on orientation and detectability

Reconstruct the product for the bidiagonal example

▼ **Work through the solution**

The trace of B^TB is $25 + 16 + 153 = 194$, equal to the sum of squares of the six input entries. Its determinant is $25(153) = 3825$. Trace and determinant determine the two squared singular values independently.

Predict a failure threshold and compare two noise sources

▼ **Work through the solution**

When δ^2 is of order u, rounding an explicitly formed Gram matrix threatens its weaker eigenvalue. Perturbations in the original data of norm 10^{-5} already permit errors of that size in singular values by the variational characterisation. A stable algorithm cannot certify a sensor separation much smaller than instrument uncertainty.

Trefethen and Bau · QR, rounding error and numerical SVD

Matrix computations and spectral accuracy · 7. Recovering a spectrum from three-term recurrences

A large vibrating system supplies products with its stiffness operator. Derive the short recurrence implied by symmetry and use a residual and a moment test to interpret its small projected spectrum.

Let symmetry remove old directions

Assume $A = A^T$, choose $\|q_1\| = 1$, and initialise $\beta_0 = 0, q_0 = 0$. Arnoldi projects Aq_j against every available basis vector. For $i < j - 1$, Aq_i already lies in the span through q_{i+1}, so $q_i^T A q_j = (Aq_i)^T q_j = 0$. For $i = j - 1$ the product is the previously computed β_{j-1}. Thus the only new diagonal coefficient is $\alpha_j = q_j^T A q_j$.

$$w = Aq_j - \beta_{j-1} q_{j-1} - \alpha_j q_j,$$

$$\beta_j = \|w\|,$$

$$q_{j+1} = w / \beta_j.$$

A zero β_j signals an invariant subspace in exact arithmetic. Stacking the identities gives $AQ_k = Q_k T_k + \beta_k q_{k+1} e_k^T$ where T_k is real symmetric tridiagonal. A normalised eigenpair (θ, z) of T_k gives a Ritz vector $Q_k z$ and residual norm $\beta_k |z_k|$. In an eigenbasis of A, that residual is at least the distance from θ to the nearest eigenvalue. To bound an angle to a particular eigenvector also separate its eigenvalue from its neighbours.

Read the first two measurements

Use $A = \operatorname{diag}(1, 2, 4)$ and $q_1 = (1, 1, 1)^T / \sqrt{3}$. Direct averaging gives $\alpha_1 = 7/3$ and $\beta_1 = \sqrt{14}/3$. The second direction is $(-4, -1, 5)^T / \sqrt{42}$, giving $\alpha_2 = 59/21$. The projected pair captures only part of the original three-dimensional spectrum. The first quadrature node $7/3$ already gives the exact weighted average $q_1^T A q_1$, while its square misses $q_1^T A^2 q_1 = 7$ by $14/9$, exactly β_1^2.

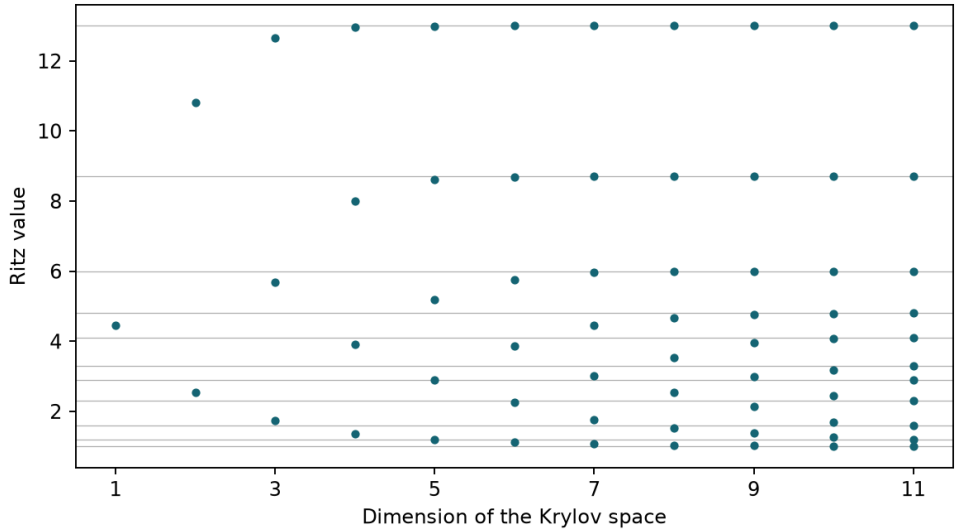
More generally, $q_1^T f(A) q_1$ integrates f against masses $|v_i^T q_1|^2$ at the eigenvalues. For degrees up to $2k - 1$, paths through the tridiagonal recurrence cannot step outside the first k directions and return to the first one in time. Hence

$$q_1^T p(A) q_1 = e_1^T p(T_k) e_1 = \sum_i |z_{1i}|^2 p(\theta_i).$$

This is the Gauss quadrature rule for that spectral measure. It also explains why a mode orthogonal to the starting vector remains invisible regardless of the iteration count.

Why can a computed tridiagonal matrix contain a duplicate mode?

A rounding perturbation can reintroduce a converged eigendirection into later vectors. The three-term recurrence removes only the last two components explicitly. The reintroduced component can grow until a second Ritz vector reports the old mode. Monitor the Gram matrix of stored basis vectors and residuals in the original space. Selective or full reorthogonalisation trades extra products and storage for independent directions.



Each column shows the projected eigenvalues at one subspace dimension. Horizontal lines mark the exact spectrum. The isolated upper mode settles early. The starting vector has nonzero overlap with every mode.

Tutorial on invisible modes and misleading accuracy

Start from a different vector for the same operator

▼ **Work through the solution**

With $q_1 = (1, 0, 0)^T$, the first residual vanishes. The one-node quadrature is exact for every polynomial for this particular observation, despite both eigenvalues 2 and 4 remaining undetected.

Derive the quadrature weights for a two-mode experiment

▼ **Work through the solution**

With equal projections on two distinct eigenvectors, the measure has two masses of one half. The first recurrence coefficient is their arithmetic mean and the first off-diagonal coefficient is half their absolute

difference. The second diagonal equals the first. Diagonalising that two-by-two matrix returns both eigenvalues and weights without prior knowledge of the modes.

Trefethen and Bau · Krylov projections and spectral quadrature

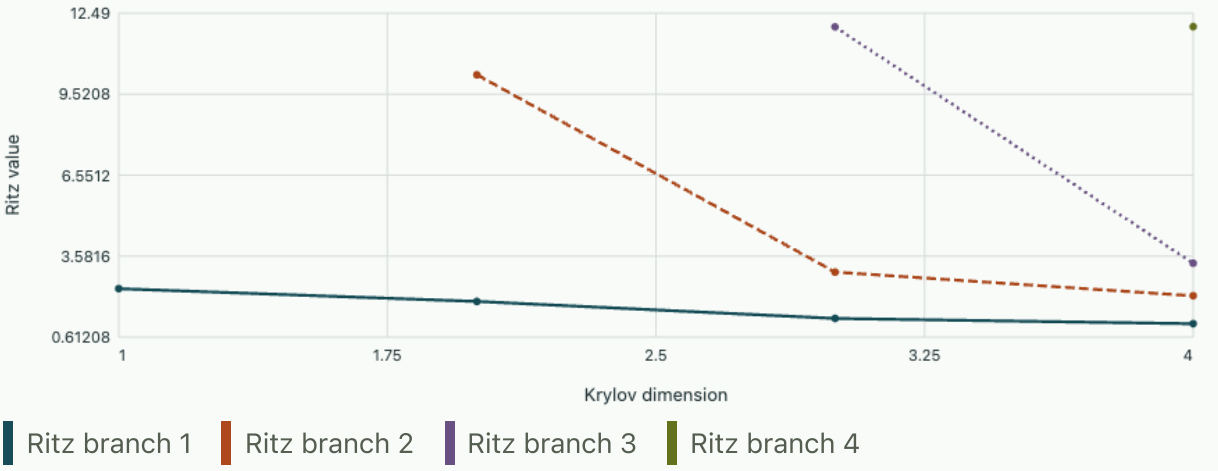
Change parameters · Recovering a spectrum from three-term recurrences

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

Separated largest eigenvalue	12
Starting overlap with highest mode	0.4
Krylov dimension	4



Exact spectrum: 1, 1.4, 2, 2.6, 3.4, 12. Full reorthogonalisation suppresses duplicate directions. Zero initial highest-mode overlap makes that eigenvalue invisible.

Operators derived from a physical balance

Construct the matrix of a constrained network, a mechanical mode or a spatial field. Determine its nullspace, symmetry and natural inner product before choosing a factorisation.

1. How a continuum balance becomes a finite-element stiffness matrix
2. Finding a displacement when no load acts in its direction
3. Coupled oscillators and normal coordinates
4. Estimating a cantilever frequency from energy
5. Incompressibility and discrete pressure modes
6. Pure flux boundaries, compatibility and the missing constant
7. Why two finite-element wave solvers propagate the same mode at different speeds
8. A converged field solver can still give the wrong voltage
9. Measure tunnelling from a four-site spectrum
10. How many lattice sites remember the colour of a neighbour?

How a continuum balance becomes a finite-element stiffness matrix

Engineering mechanics

A single bar element can predict an end displacement exactly and still miss the stress field inside. The difference follows directly from the balance it is asked to satisfy.

Turn a differential balance into work against a test displacement

Let a straight small-strain bar of length L have area A, Young's modulus E, displacement $u(x)$ and distributed axial load $q(x)$ in N/m. Tensile axial resultant is $N = EAu'$. Force balance on a slice gives $N' + q = 0$. Clamp the left end and prescribe a right-end force P. Multiply the balance by any smooth virtual displacement v satisfying $v(0) = 0$, integrate along the bar, and integrate the derivative of N by parts.

$$\int_0^L EA\,u'v'\,dx = P\,v(L) + \int_0^L qv\,dx.$$

The boundary term at the clamp disappears because that displacement is prescribed. The unknown reaction can later be recovered by equilibrium. This formulation requires only first derivatives of trial displacements. Piecewise linear displacements are therefore admissible even though their slopes may jump at internal nodes.

Compute the stiffness rather than memorising it

On an element from $x = 0$ to h, interpolate the two nodal displacements linearly:

$$u = (1 - x/h)u_1 + (x/h)u_2.$$

Let the test displacement have the same two basis functions. A virtual change of the first or second node gives the corresponding row of the balance. Both derivatives are constant with opposite signs.

$$B = \begin{bmatrix} -1/h & 1/h \end{bmatrix}, \qquad K_e = \int_0^h B^T EAB\,dx = \frac{EA}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

For a constant distributed load, each basis function integrates to $h/2$, so each node receives equivalent load $qh/2$. The two rows of stiffness sum to zero. A uniform translation stores no strain energy. At a shared node the neighbouring virtual-work contributions add because both elements refer to the same unknown displacement.

Worked example · compare a nodal value with the stress in a cell

Take a 1 m bar with $EA = 2.0$ MN, uniform $q = 1200$ N/m and free right end. Exact balance integrated from $N(L) = 0$ gives

$$N = q(L - x).$$

Integrating strain from the clamp produces $u = q(Lx - x^2/2)/(EA)$. The free-end displacement is 0.300 mm. With a single element the free row of the discrete system has stiffness EA/L and load $qL/2$, giving the same displacement. Yet its piecewise constant axial force is

$$EAu(L)/L = 600\,\mathrm{N}.$$

The exact force ranges from 1200 N at the clamp to zero at the free end. The numerical force represents the mean over the cell.

Worked example · see what doubling resolution changes

Divide the bar into two equal elements. The middle node receives 600 N from both halves together, while the free node receives 300 N. The right element carries 300 N. Equilibrium of the middle node then puts 900 N in the left element. Their extensions are $900(0.5)/(2 \times 10^6) = 0.225$ mm and 0.075 mm. Nodal displacements agree with the exact quadratic at both nodes. The largest one-sided stress-resultant error within either element is now 300 N instead of 600 N. Comparing only tip displacements would hide this improvement.

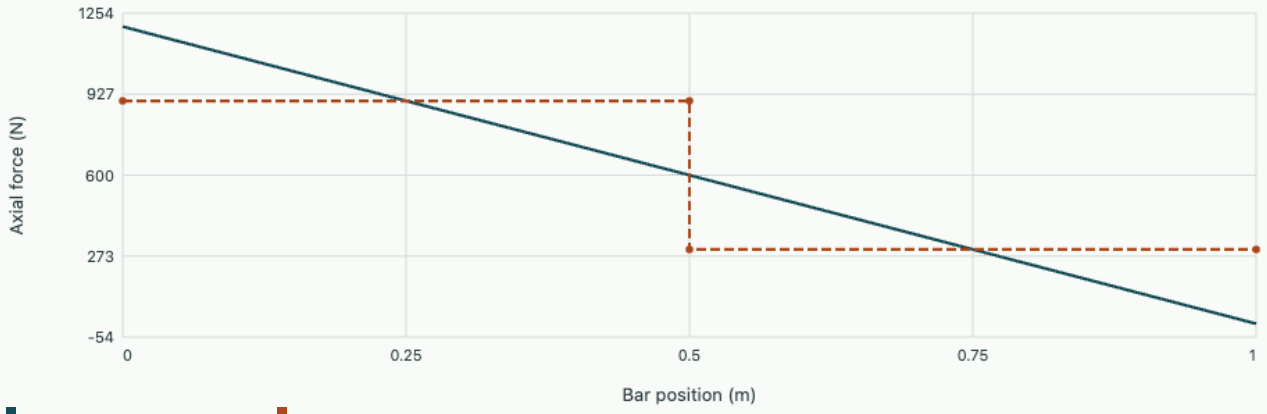
Double the number of cells until the force error is below 100 N for the baseline load. Inspect the values on both sides of an element boundary and explain why agreement in tip displacement alone could have ended this refinement study too early.

Open this experiment online

Try the calculation

Number of elements 2

Distributed load (N/m) 1200



Exact equilibrium Element force

The tip displacement is 0.300 mm on every mesh here. Maximum axial-force error is 300.0 N with 2 elements.

A related OneSimulation calculation

Compare the bar's nodal displacement with a OneSimulation truss assembled from axial members. Run the area-doubling case and check the predicted inverse-area change in displacement.

Run the truss stiffness comparison in OneSimulation

Tutorial · find the support reaction from the discarded row

For the two-element calculation, recover the left reaction using the assembled loads and displacement. Then derive an expression for the maximum axial-force error with n equal elements. Why can two different mesh studies give two different conclusions?

▼ **Evaluate the clamp row and the local profile**

The left internal nodal force is -900 N. The left external distributed-load share is $+300$ N. Equilibrium requires an additional reaction -1200 N. On an element of width L/n the exact linearly varying force changes by qL/n; its constant midpoint approximation misses an endpoint by $qL/(2n)$. Tip displacement and maximum local force answer different engineering questions. Mesh convergence must therefore name its measured quantity.

Further reading · A. F. Bower, Applied Mechanics of Solids, §8.1, finite-element equations for elastic solids

Finding a displacement when no load acts in its direction

Engineering mechanics

A displacement gauge measures the lateral travel of a truss joint under a vertical load. A force balance determines the bar forces, but cannot directly give the sideways motion. Introduce a dummy load, differentiate the stored energy and recover the gauge reading with its sign.

Link an external force coordinate to stored strain energy

For a linear elastic pin-jointed truss, neglect support movement and temperature changes and assume small displacements. Bar i has constant axial force N_i, length L_i, modulus E_i and area A_i. Integrating its force against its extension gives

$$U_i = N_i^2 L_i/(2E_i A_i).$$

Superposition makes each force linear in the external loads. A small increment of load changes the equilibrium configuration. For a linear elastic conservative system the complementary energy equals U and its derivative with respect to an independent applied force gives displacement along that force.

$$U = \sum_i \frac{N_i^2 L_i}{2E_i A_i},$$

$$\delta_Q = \frac{\partial U}{\partial Q} = \sum_i \frac{N_i n_i L_i}{E_i A_i},$$

$$n_i = \frac{\partial N_i}{\partial Q}.$$

The derivative n_i is the member force produced per unit trial load in the requested direction, with all other independent loads held fixed. Solve this trial equilibrium even if $Q = 0$ physically. Setting $Q = 0$ before differentiating erases its entire contribution. As a dimensional check, n_i is dimensionless when Q and N_i both use newtons, and the sum has dimensions of length.

Compute the two-bar flexibility matrix

A joint hangs from two overhead pins by straight bars slanting up-left and up-right, each at 45° to horizontal. Positive joint coordinates are rightward and downward. Let V be the applied downward force and H the rightward force; bars carry tension positive. Equilibrium gives

$$N_L = (V + H)/\sqrt{2},\ N_R = (V - H)/\sqrt{2}.$$

Define bar axial compliances $c_i = L_i/(E_i A_i)$. Insert the forces into the energy and differentiate with respect to H and V independently:

$$\begin{bmatrix} \delta_H \\ \delta_V \end{bmatrix} = \frac{1}{2} \begin{bmatrix} c_L + c_R & c_L - c_R \\ c_L - c_R & c_L + c_R \end{bmatrix} \begin{bmatrix} H \\ V \end{bmatrix}.$$

The cross coefficients are equal because the mixed derivatives of the stored energy commute. This is Maxwell–Betti reciprocity. A cross displacement can be negative even though both direct compliances and stored energy are positive. The eigenvalues of this particular flexibility matrix are c_L, c_R, corresponding to the two independent bar-extension directions.

Predict a sideways gauge reading and reverse the test

Take $c_L = 0.0020,\ c_R = 0.0040$ mm/N. At $V = 100$ N and $H = 0$, both members carry 70.71 N. The vertical displacement is 0.300 mm, and the horizontal displacement is −0.100 mm. The more compliant right bar stretches farther; the joint moves left and down to accommodate both bar extensions. Half the vertical load times its displacement gives 15 N mm of strain energy. Summing the two bar energies independently gives the same result.

Apply instead $H = 100$ N and $V = 0$, assuming the slender right bar remains stable under compression. The vertical reading is −0.100 mm and horizontal motion is +0.300 mm. Thus the lateral response to the vertical test predicts the vertical response to a lateral test of equal magnitude. Equalizing bar areas would remove the cross response if the lengths and moduli matched. In practice a pin with clearance or a cable that goes slack can break the assumed linear bilateral response.

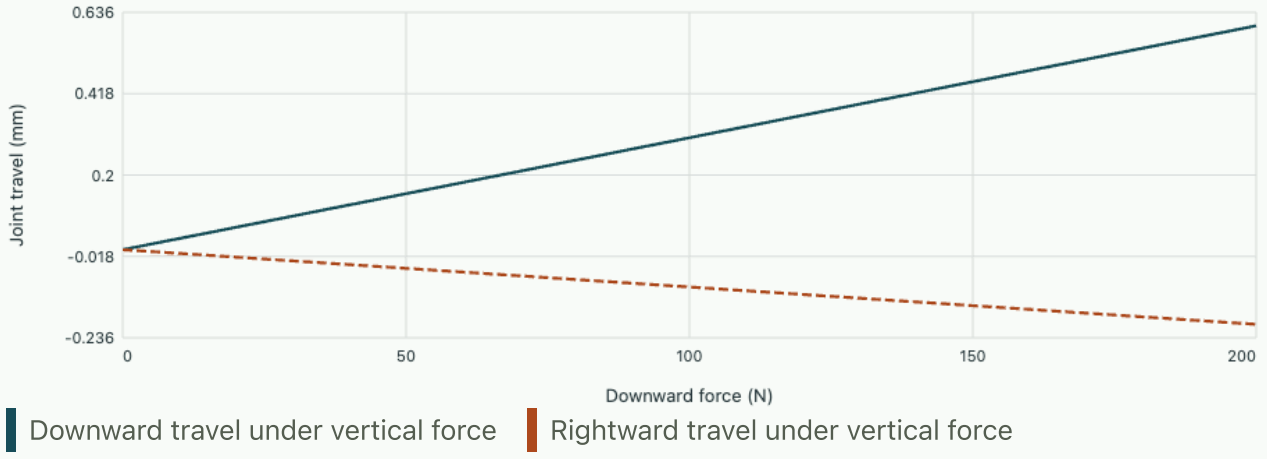
Exchange the left and right compliances. Before looking at the new graph, mark which readout should reverse sign and which should remain equal. Then set the two values alike and identify the symmetry that removes cross-coupling.

Open this experiment online

Try the calculation

Left bar compliance (µm/N) 2

Right bar compliance (µm/N) 4



At 100 N down the joint travels 0.300 mm down and -0.100 mm right. By reciprocity the latter also gives downward travel under 100 N right.

A related OneSimulation calculation

Use the course truss to test the same flexibility scaling on a larger assembly. Double every member area at fixed loads and geometry, then compare the joint displacement with the separately calculated chord force.

Run the OneSimulation truss experiment

Tutorial · identify the two member compliances from one load case

A 200 N vertical test gives +0.50 mm down and +0.10 mm right. Estimate each compliance and predict horizontal travel under simultaneous 60 N rightward and 100 N downward loading. How can one additional gauge reading test reciprocity?

▼ **Show the identification and loading prediction**

Direct compliance is $0.50/200 = 0.0025$ mm/N and cross compliance is $0.10/200 = 0.0005$ mm/N. Adding and subtracting gives $c_L = 0.0030,\ c_R = 0.0020$ mm/N. The predicted horizontal motion is

$$0.0025(60) + 0.0005(100) = 0.200\,\mathrm{mm}.$$

A 200 N rightward test should give +0.10 mm down. A disagreement after changing load direction suggests checking seating and the linear-elastic domain before assigning a single compliance to each bar.

Further reading · David Roylance, MIT 3.11 · truss displacements, strain energy and Castigliano's theorem

Coupled oscillators and normal coordinates

Machine dynamics

Derive and interpret the two modes of a coupled machine, then decide which mode an applied load can excite.

FROM FIRST PRINCIPLES

Find natural modes from the free force balance

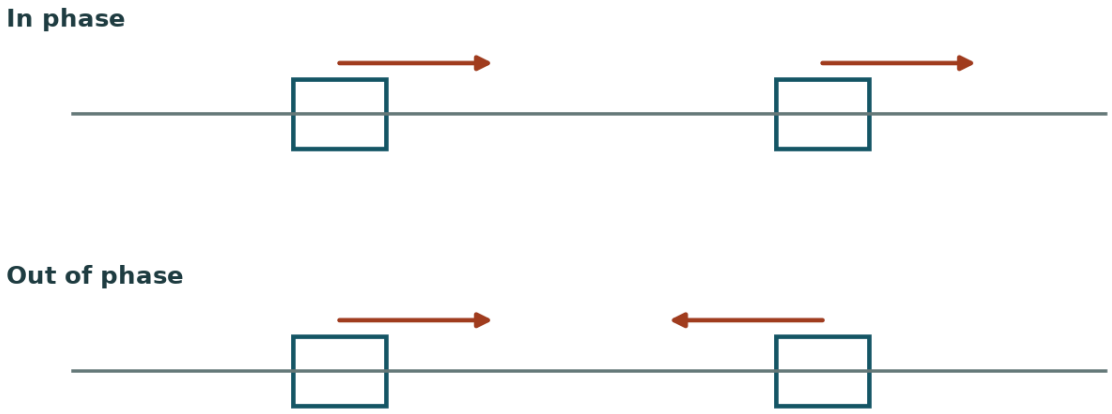
For small motions of a conservative system with several coordinates, linear inertia forces are represented by a positive-definite mass matrix M and linear restoring forces by a symmetric stiffness matrix K. Collect the free balance and try a motion in which the coordinate amplitudes stay in one fixed ratio as each oscillates at frequency ω.

$$M\ddot{\mathbf{x}} + K\mathbf{x} = 0,$$
$$\mathbf{x} = \boldsymbol{\phi}\cos(\omega t) \implies (K - \omega^2 M)\boldsymbol{\phi} = 0.$$

A nonzero shape $\boldsymbol{\phi}$ can solve these homogeneous equations only when the coefficient matrix is singular. The determinant condition gives candidate squared frequencies, and the nullspace at each frequency gives a mode shape up to scale.

$$\det(K - \omega^2 M) = 0,$$
$$K\boldsymbol{\phi}_j = \omega_j^2 M\boldsymbol{\phi}_j.$$

For identical masses linked by a spring, a common translation leaves that link unstretched and an opposite motion stretches it from both ends. This observation identifies common and differential shapes even before evaluating the determinant. Initial displacement and velocity select their amplitudes; the eigenproblem identifies the motions the structure can support.



Which spring deforms in each mode?
Assume equal end springs k, equal masses m and a central coupling spring k^c. A common displacement leaves the coupling spring at its original length. In the lower drawing each mass moves a distance x towards the other, compressing the central spring by 2x. Half a cycle later the same spring extends by 2x. The ratios of restoring force to displacement are k and k + $2k^c$, giving squared frequencies k/m and (k + $2k^c$)/m.

DERIVATION AND APPLICATIONS

Quantify the splitting of two weakly coupled modes

When two machines have the same uncoupled natural frequency, adding a weak connecting spring splits that frequency into two. Consider identical sliders of mass m each attached to ground by stiffness k and linked to each other by stiffness k_c. Neglect damping and choose displacements in the same direction from equilibrium. The coupling spring responds only to the displacement difference. Adding and subtracting the two force balances therefore gives separate equations for $s = (x_1 + x_2)/2$ and $d = (x_1 - x_2)/2$.

$$m\ddot{s} + ks = 0,$$

$$m\ddot{d} + (k + 2k_c)d = 0.$$

Release slider one at A and slider two at zero, both from rest. The two modal coordinates then both start at $A/2$. Write $\omega_s = \sqrt{k/m}$ and $\omega_d = \sqrt{(k + 2k_c)/m}$ and combine their solutions.

$$x_1 = \tfrac{A}{2}(\cos \omega_s t + \cos \omega_d t),$$

$$x_2 = \tfrac{A}{2}(\cos \omega_s t - \cos \omega_d t).$$

Introduce $\bar{\omega} = (\omega_s + \omega_d)/2$ and $\Delta\omega = \omega_d - \omega_s$. The sum-to-product identities separate the fast motion from slow modulation. When the split is small, the modulation resembles a repeated transfer of motion between the sliders.

$$x_1 = A\cos(\bar{\omega}t)\cos(\Delta\omega t/2),$$
$$x_2 = A\sin(\bar{\omega}t)\sin(\Delta\omega t/2).$$

The first minimum of the slider-one displacement envelope occurs at $t = \pi/\Delta\omega$. Its instantaneous velocity there need not vanish because the envelope itself is changing. The envelope timing remains identifiable from repeated peaks at weak coupling.

WORKED EXAMPLE

Time the exchange of motion

For $m = 1.00$ kg, $k = 100$ N/m and $k_c = 10.5$ N/m, the two frequencies are 10.0 and 11.0 rad/s. The first envelope minimum appears at 3.14 s and the first full recovery of its magnitude at 6.28 s. Releasing the first slider by 8.00 mm sets both modal initial amplitudes to 4.00 mm. Releasing both sliders together instead excites only the slower common motion.

Change the inputs and inspect this calculation below

SECOND APPLICATION

Estimate coupling from a beat measurement

A second pair has mass 2.00 kg per slider and measured lower frequency 25.0 rad/s. Consecutive magnitude maxima of one slider's envelope occur 5.00 s apart. Thus

$$\Delta\omega = 2\pi/5.00 = 1.257\,\mathrm{rad/s},$$
$$k_c = m[(25 + 1.257)^2 - 25^2]/2 = 64.4\,\mathrm{N/m}.$$

The independently inferred grounding stiffness is $m\omega_s^2 = 1250$ N/m, making the connector about 5.2 percent as stiff.

Interactive calculation · Change the inputs of this example

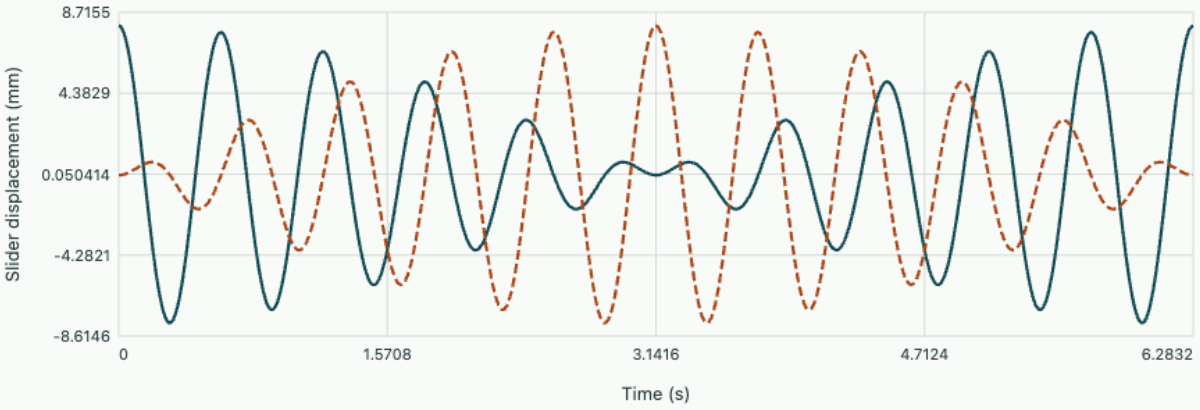
Use the adjustable quantities to test the two worked calculations above. The plot recomputes this derivation's balance or constitutive relation; its numerical readout reports the quantities needed for the forward calculation and the inverse or design question. Keep the fixed assumptions stated in the readout when comparing with the example values.

Open this live calculation

Try the calculation

Coupling stiffness (N/m) **10.5**

Initial release (mm) **8**



Frequencies 10.000 and 11.000 rad/s; first envelope minimum at 3.142 s. Equal masses1 kg; grounded stiffness100 N/m.

Two masses allow two independent shapes

Two identical instruments sit on separate spring mounts and are connected by a flexible link. Pull both by the same amount and the link keeps its natural length. Pull them in opposite directions and the link stretches twice as much as either displacement. The two tests reveal distinct restoring mechanisms. A model with one equivalent spring and the sum of both masses would discard one of them.

Let x_1, x_2 be horizontal displacements from equilibrium. Each mass is m, each ground spring has stiffness k, and the coupling spring has stiffness k_c. With light springs, small displacements and no damping, the kinetic energy is $m(\dot{x}_1^2 + \dot{x}_2^2)/2$ and the elastic energy is $k(x_1^2 + x_2^2)/2 + k_c(x_1 - x_2)^2/2$. The coupling energy contains the difference of displacements because it is the extension between its ends.

Separate space from time

Taking the derivative of elastic energy with respect to each coordinate gives $m\ddot{x}_1 + (k + k_c)x_1 - k_c x_2 = 0$ and the symmetric second equation. Look for $\mathbf{x} = \boldsymbol{\phi}\cos\omega t$. Nonzero modal shapes solve $(\mathbf{K} - \omega^2\mathbf{M})\boldsymbol{\phi} = 0$. For this pair the determinant equation is

$$(k + k_c - m\omega^2)^2 - k_c^2 = 0.$$

Taking both signs of the square root gives

$$\omega_1^2 = k/m,$$

$$\omega_2^2 = (k + 2k_c)/m.$$

Substitution back yields shapes proportional to $(1, 1)$ and $(1, -1)$. For the first, the connector transmits no incremental force. For the second, the connector acts strongly. Reversing both signs of a plotted eigenvector or multiplying it by a nonzero number describes the same physical mode.

Predict a release test at both instruments

Take $m = 5.0$ kg, $k = 2000$ N/m and $k_c = 1500$ N/m. The two angular frequencies are 20.0 and 31.62 rad/s, or 3.18 and 5.03 Hz. Pull only the first instrument 4.0 mm and release from rest. The initial displacement vector equals $2(1, 1) + 2(1, -1)$ mm. Each modal amplitude starts at its corresponding value with zero initial velocity. Hence

$$x_1(t) = 2\cos(20t) + 2\cos(31.62t),$$

$$x_2(t) = 2\cos(20t) - 2\cos(31.62t)$$

in millimetres. At zero time the second displacement vanishes while the connector exerts 6.0 N on its mass. Its initial acceleration is 1.2 m/s² toward the first instrument’s displacement. This apparent transfer of motion needs neither forcing nor damping: modal displacements add and cancel differently at the two observation points.

Distinguish available modes from visible modes

Define average and relative coordinates $q_s = (x_1 + x_2)/2$ and $q_a = (x_1 - x_2)/2$. Adding or subtracting the original equations decouples them. Equal external forces excite only q_s; equal and opposite forces excite only q_a. Taking the signal $(x_1 + x_2)/2$ likewise hides the antisymmetric mode. A missing spectral peak can therefore reveal the symmetry of a load or sensor arrangement, even though the corresponding eigenmode still exists.

If the two masses differ, the simple sum and difference cease to diagonalise the equations. The general route is to form the physical mass and stiffness matrices and solve the symmetric generalised eigenproblem. Distinct modes of an undamped conservative system are orthogonal with respect to mass,

$\boldsymbol{\phi}_i^T \mathbf{M} \boldsymbol{\phi}_j = 0$. This weighted orthogonality is what permits a measured motion or distributed load to be projected into separate modal coordinates.

Predict which mode responds to adding rotary inertia

For a unit mode shape ϕ, the quotient $\omega^2 = \phi^T K \phi / (\phi^T M \phi)$ compares stored elastic energy with inertia. Increasing inertia at a coordinate matters most when that coordinate has appreciable motion in the mode. A frame with three tip coordinates offers a way to check this reasoning beyond the symmetric two-mass illustration.

Compare two changes to the same frame: double only the tip mass from 200 to 400 kg, then restore it and double only the tip rotary inertia from 60 to 120 kg m². Predict whether either change can raise a natural frequency for the same stiffness and compute the fractional shift of the third mode in each case. Use the negligible-member-mass row throughout. Check the frame frequency measurements.

Derive bounce and pitch modes of a supported machine

A long machine supported on two mounts can translate and rotate as a nearly rigid body even if there is negligible frame bending. Let the mass centre move vertically by z and rotate by small angle θ, with positive rotation raising a point to the right. Vertical spring deflections relative to ground are $z - a\theta$ on the left and $z + b\theta$ on the right. Measure displacements from static equilibrium. The kinetic energy is $(m\dot{z}^2 + I_G\dot{\theta}^2)/2$, and the spring energy is $[k_L(z - a\theta)^2 + k_R(z + b\theta)^2]/2$.

Differentiating potential energy gives stiffness terms $K_{zz} = k_L + k_R,\ K_{\theta\theta} = k_L a^2 + k_R b^2,\ K_{z\theta} = -k_L a + k_R b$. The mass matrix is diagonal because translation was measured at the centre of mass. A vertical translation produces a moment whenever the weighted mount locations fail to balance, and a pure rotation then also produces a vertical force by reciprocity. If $k_L a = k_R b$ the modes are pure translation and pure pitch. Equal spring rates alone do not guarantee decoupling when the lever arms differ.

For free motion solve $(K - \omega^2 M)\phi = 0$. In mixed coordinates, metres and radians label different displacement entries, so a numerical component ratio needs units or a stated length scale. Measured vertical motions at two actual sensor locations provide an unambiguous observable mode shape.

Separate a heaving mode from a rocking mode

Let $m = 200$ kg, $I_G = 40\,\mathrm{kg\,m^2}$, $a = b = 0.50$ m and mount rates $k_L = 30000,\ k_R = 50000\,\mathrm{N/m}$. The stiffness entries are 80000 N/m, 20000 N m/rad and 10000 N in the mixed term. Setting the determinant to zero gives $(80000 - 200\lambda)(20000 - 40\lambda) - 10000^2 = 0$ with $\lambda = \omega^2$. Dividing through by 8000 yields

$$\lambda^2 - 900\lambda + 187500 = 0.$$

The squared frequencies are 327.5 and 572.5 s⁻², giving 2.88 and 3.81 Hz.

For the lower mode the first row implies

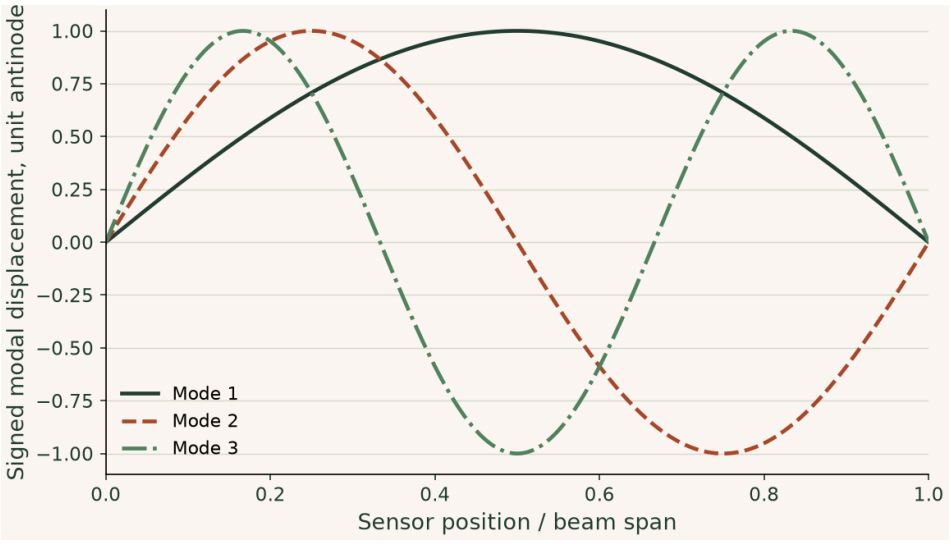
$$\theta/z = -(80000 - 200 \times 327.5)/10000 = -1.45\,\mathrm{rad/m}.$$

If mass-centre motion has 1 mm amplitude, the mount-point displacements are 1.725 and 0.275 mm, in phase but strongly unequal. The higher mode gives $\theta/z = +3.45\,\mathrm{rad/m}$, so the left point now moves oppositely to the right. An accelerometer placed close to a mode's nodal point can scarcely detect it although that mode remains physically present. For the low mode the vertical node lies about $1/1.45 = 0.69$ m to the right of the centre, beyond the right mount. Two well-separated measurement points make the rotational contribution identifiable.

A static load experiment can identify the stiffness coupling before a modal test. Apply a known vertical force F at the centre and no external moment. The pitch-equilibrium equation gives $\theta = -K_{z\theta}z/K_{\theta\theta} = -0.50z$ rad per metre for these mounts. Substitution into the vertical balance leaves effective static stiffness $80000 - 10000^2/20000 = 75000\,\mathrm{N/m}$. Under 150 N centre load, predicted motion is 2.00 mm and pitch −0.0010 rad, giving left and right mount readings of 2.50 and 1.50 mm. A centre gauge alone would only indicate an apparent softening, while the unequal pair exposes the rotational coupling that later appears in the two mode shapes.

Find a sensor position that can see a mode

Use the first three bending modes of an ideal simply supported uniform beam as a continuous analogue of coupled oscillators. A local displacement transducer reports a modal coordinate multiplied by the signed shape at the measurement location.



A midspan sensor misses the second mode even if excitation and damping would produce a large second modal coordinate. At one third span the third mode is invisible. Several sensor locations help distinguish a missing mode from a measurement node.

▶ **Read values from the plot**

Interactive calculation · change the inputs of this example

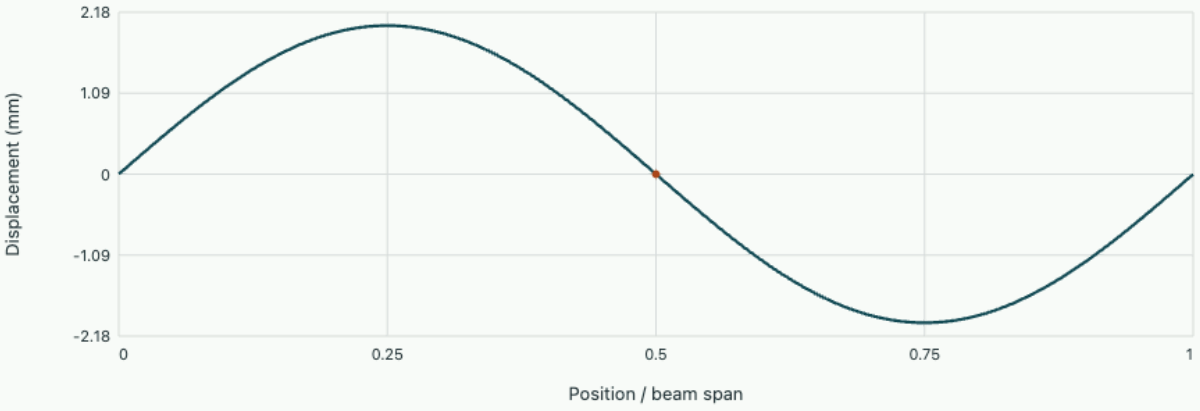
Move the sensor to a node and then change mode number. A zero local reading does not imply a stationary beam.

The curves below are recalculated from this section's equations in your browser. The fixed reference plot above remains available for comparison.

Open this live calculation

Try the calculation

Mode number	2
Sensor position / span	0.5
Modal amplitude (mm)	2



Selected mode Sensor reading

The sensor reads 0.000 mm. Mode 2 has 1 internal displacement nodes. A zero reading can conceal nonzero modal energy.

Estimating a cantilever frequency from energy

Machine dynamics

A clamped beam carries a small sensor at its tip. Its precise mode shape is unknown. Use the static deflection shape to approximate the fundamental frequency and see why concentrating a small mass at the tip has a large effect.

Choose a displacement field that respects the support

Write $w(x,t) = q(t)\phi(s)$ with $s = x/L$, taking q as tip displacement. Choose the shape from a uniform cantilever under a static transverse tip force. The bending moment varies linearly from root to tip. Integrating curvature twice with $w(0) = w'(0) = 0$, then normalizing at $s = 1$, gives

$$\phi = s^2(3 - s)/2.$$

The shape satisfies both essential clamp conditions. Its free-tip shear need not match the exact unloaded vibration mode. The energy formulation can use such admissible trial shapes.

Background · Rayleigh quotient

$$R(u) = u^{\mathsf{T}} K u / (u^{\mathsf{T}} M u)$$

For symmetric stiffness and positive mass this estimates squared angular frequency. A kinematically admissible displacement is required for a useful structural estimate.

Compute effective stiffness and mass from the same field

The strain energy density of a slender beam in bending is $\frac{1}{2} EI w_{xx}^2$. Its translational kinetic energy is $\frac{1}{2}\rho A w_t^2$ per unit length. Changing variables from x to s introduces different powers of L into the two integrals. Include a point mass M at the tip, neglecting its rotary inertia.

$$U = \frac{EIq^2}{2L^3}\int_0^1 (\phi'')^2\, ds,$$

$$K = \frac{\dot{q}^2}{2}\left[\rho A L \int_0^1 \phi^2\, ds + M\right].$$

Since $\phi'' = 3(1 - s)$, the stiffness integral is 3. Squaring $s^2(3 - s)/2$ and integrating $(9s^4 - 6s^5 + s^6)/4$ gives $33/140$. The Euler–Lagrange equation for q is consequently a single oscillator equation with

$$\omega_R^2 = \frac{3EI/L^3}{(33/140)\rho A L + M}.$$

For M=0 this gives the dimensionless coefficient $\sqrt{140/11} = 3.5675$. Solving the exact unloaded cantilever condition $\cosh\beta\cos\beta = -1$ gives

$$\beta_1^2 = 3.5160.$$

The estimate is high by about 1.5%. The Rayleigh minimum principle explains that direction of error for conservative systems with admissible trial shapes. Restricting motion to a single shape restricts the set over which the energy quotient can attain its minimum.

Choose a sensor after estimating its dynamic effect

Take $EI = 10$ N m², $L = 0.50$ m and total beam mass 0.40 kg. The tip stiffness is 240 N/m and the distributed effective mass is only 0.0943 kg. The estimate gives 8.03 Hz without a sensor and 6.49 Hz with a 0.05 kg tip sensor. The sensor weighs one eighth of the beam, yet it has over half the effective mass associated with the chosen tip coordinate.

Place the same sensor at half-span instead. There $\phi(0.5) = 0.3125$. Its kinetic energy contributes $M\phi(0.5)^2 = 0.00488$ kg to the effective mass, so the estimate rises to 7.83 Hz. The sensor measures a smaller local motion as it perturbs the frequency less. Compare that trade-off at different attachment positions with the plot.

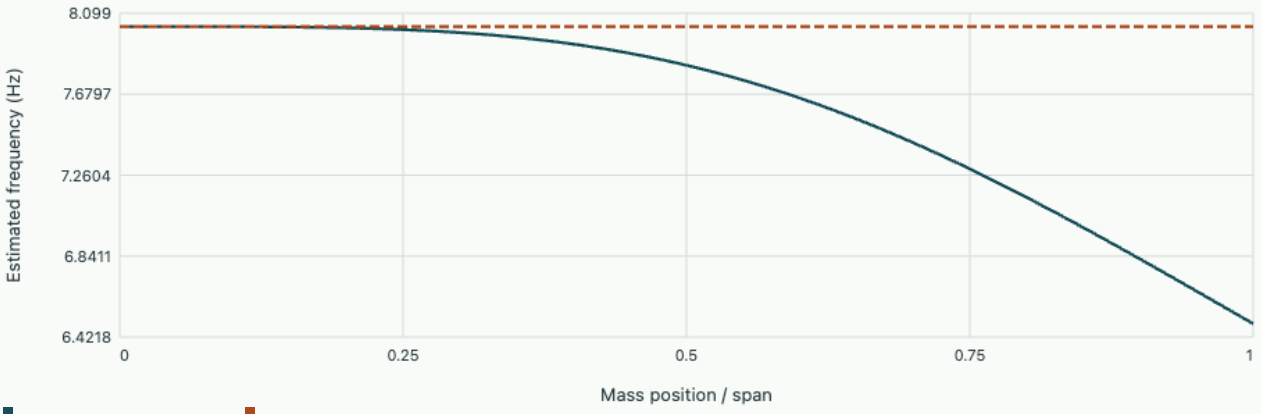
Move along the position axis with a fixed sensor mass. Identify the interval in which frequency is most sensitive to changing attachment position. Double the sensor mass and check whether the change in frequency doubles at the tip.

Open this experiment online

Try the calculation

Attached mass (kg) **0.05**

Bending rigidity (N m²) **10**



The estimate is 6.49 Hz with the sensor at the tip and 7.83 Hz at half-span. The bare-beam estimate is 8.03 Hz.

A related OneSimulation calculation

The course frame model has several modes and an independently variable payload rotary inertia. Compare a pure translation-mass change with a rotary-inertia change; the scalar point-mass estimate above addresses the first mechanism.

Open the OneSimulation frame eigenvalue experiment

Tutorial · estimate an unknown attached mass

The same bare beam is fitted with a small tip part and resonates at 6.0 Hz. Estimate the attached mass. Give one reason the inferred mass could differ from a scale reading even when the frequency measurement is accurate.

▼ **Show the mass estimate**

The required effective mass is

$$240/(2\pi \cdot 6)^2 = 0.1689\,\mathrm{kg}.$$

Subtract 0.0943 kg for the beam to obtain about 0.0746 kg. Clamp compliance would reduce the effective stiffness and frequency, which this fit would interpret as extra tip mass. The trial shape and omission of rotary inertia also limit the estimate for a part with appreciable extent.

Further reading · Bower, Applied Mechanics of Solids · Rayleigh–Ritz methods

Incompressibility and discrete pressure modes

Numerical partial differential equations

Derive the velocity–pressure block system and use a periodic mesh to detect a pressure pattern that exerts no discrete force.

FROM FIRST PRINCIPLES

Eliminate velocity to expose the pressure equation

In a mixed incompressible discretization, velocity responds to applied force and pressure while pressure enforces zero discrete divergence. Let A be a positive definite velocity block after applying boundary conditions, B the discrete divergence, and use B^{T} with the corresponding weak pressure sign convention. With forcing f and prescribed divergence g the block equations are

$$Au + B^{\mathsf{T}}p = f,$$
$$Bu = g.$$

Eliminate velocity by applying the inverse action of A to the first equation and substituting into the second. The resulting pressure Schur system is

$$Sp = BA^{-1}f - g,$$
$$S = BA^{-1}B^{\mathsf{T}}.$$

Since $p^{\mathsf{T}}Sp = (B^{\mathsf{T}}p)^{\mathsf{T}}A^{-1}(B^{\mathsf{T}}p)$, a pressure vector is invisible precisely when its discrete gradient vanishes. A constant pressure often has this property by physics; additional oscillatory pressure null modes can indicate incompatible spaces. Fixing a pressure reference removes the arbitrary constant but does not cure those extra modes. Once pressure is obtained, substitute it back to recover a velocity satisfying the original continuity equation.

DERIVATION AND APPLICATIONS

Follow a pressure Fourier mode through two discrete gradients

The divergence and pressure-gradient operators must recognize every nonconstant pressure that is meant to constrain the velocity. Consider an even number of uniformly spaced periodic cells along a line, with pressure p_j at cell centres. A one-dimensional Fourier mode $p_j = e^{ij\theta}$ is a convenient algebraic test; the constant mode is the expected arbitrary pressure reference. Suppose a colocated method uses a centred pressure gradient at the same locations:

$$(G_c p)_j = (p_{j+1} - p_{j-1})/(2h) = i[\sin\theta/h]e^{ij\theta}.$$

The pressure Schur operator in a simplified projection with identity velocity mass is $G_c^* G_c$. Its eigenvalue is the squared magnitude of that gradient multiplier. It vanishes both at $\theta = 0$ and at the checkerboard

$\theta = \pi$. To compare, place normal velocity at faces and difference its adjacent cell pressures:

$$(G_s p)_{j+1/2} = (p_{j+1} - p_j)/h = \frac{2i\sin(\theta/2)}{h} e^{i(j+1/2)\theta}.$$

Combining the face gradient with its adjoint yields two different pressure operators on the same grid:

$$\lambda_c(\theta) = \sin^2\theta/h^2,$$
$$\lambda_s(\theta) = 4\sin^2(\theta/2)/h^2.$$

At long wavelengths both agree with k^2 for physical wavenumber $k = \theta/h$. At the alternating mode, staggered differences detect adjacent pressure differences even though a centred two-cell difference cancels them. This calculation isolates one spatial compatibility issue; other geometries and velocity boundary conditions require the full divergence-gradient pairing to be checked.

WORKED EXAMPLE

Compare a resolved oscillation with a checkerboard

On a grid with $h = 0.100$ m, choose $\theta = \pi/2$. The two Schur eigenvalues are 100 and 200 m⁻². At $\theta = \pi$ they become zero and 400 m⁻² respectively. A unit-amplitude alternating pressure consequently produces zero centred forces, although its differences across neighbouring faces have magnitude 20.0 per metre.

SECOND APPLICATION

Test the effect of setting the pressure mean to zero

An even periodic grid with pressure values $+3, -3, +3, -3$ Pa has zero arithmetic mean. Fixing that mean does nothing to its spurious mode in the centred scheme. With face velocities the pressure gradients alternate between -60 and $+60$ Pa/m for $h = 0.100$ m. A compatible projection sees these forces and excludes this pattern from the nullspace. Removing an arbitrary constant addresses a different degeneracy.

Pressure enforces a constraint

For a steady incompressible viscous fluid, one simplified model is $-\mu\Delta u + \nabla p = f$, $\nabla \cdot u = 0$, with prescribed wall velocity. Multiplying momentum by a velocity test function vanishing at the wall and integrating by parts gives

$$\mu \int \nabla v : \nabla u - \int p\, \nabla \cdot v = \int f \cdot v.$$

Testing incompressibility with a pressure test function completes a coupled weak formulation. Adding a constant to pressure leaves its gradient unchanged, so use a reference pressure or impose zero mean in a closed domain.

Choose discrete velocity and pressure spaces, and define B as the negative weak divergence. After eliminating prescribed velocities, the block equations are

$$Au + B^T p = f,\ Bu = 0,$$

with A positive definite under these boundary conditions. Eliminating velocity yields the pressure operator $BA^{-1}B^T$. Its quadratic form is $(B^T p)^T A^{-1}(B^T p)$. A nonconstant pressure whose discrete gradient vanishes therefore survives as a spurious null mode, even if the velocity block is exceptionally well conditioned.

Test a checkerboard

On an even periodic one-dimensional grid use the collocated centred gradient $(p_{j+1} - p_{j-1})/(2h)$. Choose $p_j = (-1)^j$. The two neighbouring pressures match, so the gradient is zero at every velocity point. The pressure has zero mean and amplitude one, making it independent of the constant pressure ambiguity. For values $1, -1, 1, -1$ and $h = 0.25$, the result is still the identically zero vector. A pressure solver applied to such a discretisation can retain arbitrary amounts of this alternating pattern without changing the velocity residual.

Placing velocities on the faces instead leads to a face gradient $(p_{j+1} - p_j)/h$. On the same four-cell grid its magnitudes are all 8 with alternating signs. This layout couples the previously invisible pressure variation to the velocity unknowns. The one-dimensional example diagnoses an algebraic pairing issue. In higher dimensions the geometry, boundaries and full selection of discrete spaces also affect stability.

A uniform coupling requirement

A family of stable mixed spaces requires a lower bound, independent of mesh size, on the largest discrete divergence pairing a normalised pressure can obtain with a normalised velocity test. This is the discrete inf-sup condition, with pressures measured after removing their constant ambiguity. It excludes both exact checkerboards and nearly invisible modes that become worse as the mesh is refined. A standard finite-element pairing on suitable triangulations combines continuous quadratic velocities with continuous linear pressures. Other pairings need a separate stability argument or stabilising terms.

Check discrete divergence and momentum, then inspect pressure on grid-scale patterns. A smooth velocity, or even a small residual in both block equations, provides very little information about the amplitude of a pressure component on which those equations have no sensitivity.

Give pressure a reference without changing forces

Replacing a four-cell periodic pressure vector $(2, 4, 6, 8)^T$ by its zero-mean representative $(-3, -1, 1, 3)^T$ leaves every face difference unchanged. The mean removal fixes one arbitrary pressure level. Adding the checkerboard $(1, -1, 1, -1)^T$ also preserves its mean, yet changes the individual face differences. This distinction makes a practical diagnostic: verify a constant shift has no physical effect,

then test whether a nonconstant alternating shift produces a physically nonzero force in the chosen velocity space. Passing the former test alone cannot reveal uncoupled pressure coordinates.

Project a tentative velocity onto discrete incompressibility

After an explicit force update a tentative velocity v^* may have nonzero discrete divergence. Suppose M is a symmetric positive definite velocity mass matrix and the desired constraint is $Bv = 0$. Find the divergence-free velocity closest to v^* in the kinetic metric by minimising $(v - v^*)^T M (v - v^*)/2$. Introducing a multiplier p gives

$$M(v - v^*) + B^T p = 0.$$

Applying B to the resulting corrected velocity yields

$$BM^{-1}B^T p = Bv^*.$$

Here p acts as a pressure impulse; conversion to a physical pressure depends on the time-step and scaling conventions.

The correction $M^{-1}B^T p$ is mass-orthogonal to every admissible velocity w, since $w^T B^T p = (Bw)^T p = 0$. Hence the norm of the tentative state splits into squared norms of the corrected velocity and removed component. This projection never increases kinetic energy in homogeneous constraints. A null pressure mode with $B^T p = 0$ does not affect the corrected velocity; fixing a gauge can select a unique multiplier. If the discrete gradient and divergence are assembled inconsistently, the orthogonality and corresponding energy argument may fail even if a pressure equation is solved accurately.

Follow an impulse through three velocity unknowns

Choose $M = \mathrm{diag}(1, 2, 1)$, and let $Bv = (v_1 - v_2, v_2 - v_3)^T$. Constraint-compatible velocities have all three components equal. The tentative velocity is $(0, 0, 2)^T$, with weighted momentum two and kinetic energy two. Its divergence vector is $(0, -2)^T$. Multiplying $BM^{-1}B^T$ gives diagonal entries 1.5 and off-diagonal entries −0.5. Solving this two-by-two system yields

$$p = (-0.5, -1.5)^T.$$

Recover $B^T p = (-0.5, -1, 1.5)^T$, then

$$M^{-1}B^T p = (-0.5, -0.5, 1.5)^T.$$

Subtracting gives $v = (0.5, 0.5, 0.5)^T$, now satisfying both constraints. Its weighted momentum is still two because the constant velocity vector is admissible and therefore mass-orthogonal to the correction. The final kinetic energy is 0.5. The removed component has energy $[0.25 + 2(0.25) + 2.25]/2 = 1.5$,

exactly the energy difference. Both conservation of uniform momentum and removal of the incompatible component have been checked numerically.

With prescribed boundary fluxes the target constraint is generally inhomogeneous. The closest admissible velocity then lies in an affine space; the preceding zero-centred energy decrease need no longer hold because boundaries can do work. Transforming to fluctuations about a boundary-compatible lifting can restore an orthogonality statement for the correction. Before interpreting a pressure projection's apparent dissipation or energy injection, specify the boundary conditions and the metric used in its derivation.

For the same tentative state an ordinary unweighted Euclidean projection would set all velocities to their arithmetic mean $2/3$. That also makes both discrete differences zero, but the weighted momentum becomes $8/3$, exceeding its initial value two. A check on incompressibility alone would accept this physically different state. Choosing a mass matrix in the pressure correction is therefore consequential whenever cell masses, staggered volumes or basis-function weights differ. Evaluating both the discrete divergence and the physically weighted momentum provides information that neither result supplies on its own.

The impulse components give an additional equation-level diagnostic because applying their discrete gradient must reproduce the actual velocity correction.

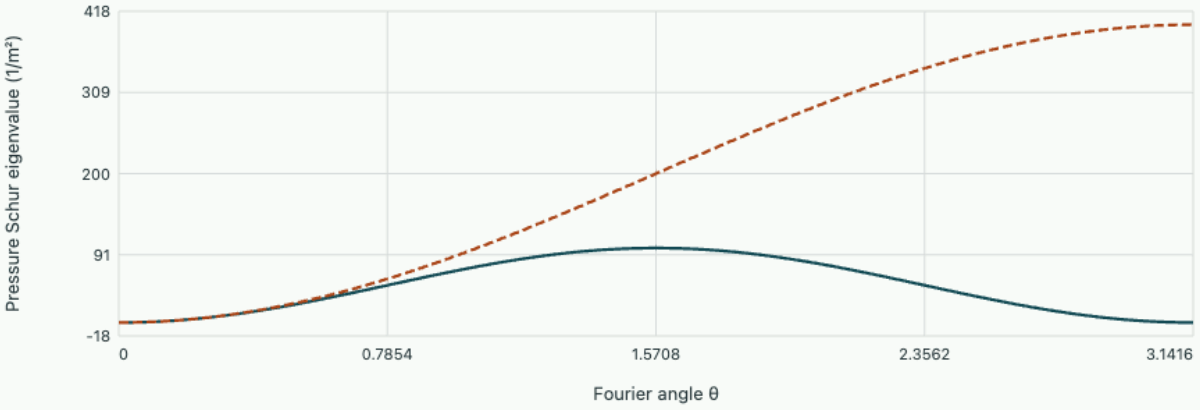
Change parameters · Explore the worked calculation

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

Grid spacing h (m) **0.1**



Collocated centred gradients Staggered face gradients

At θ = π the centred symbol vanishes while the staggered symbol is 400 m⁻². Removing the pressure mean does not remove this alternating zero-mean mode.

Pure flux boundaries, compatibility and the missing constant

Numerical partial differential equations

If a steady diffusion equation prescribes only fluxes, a solver can return a singular-matrix warning for two different physical reasons. Derive the mass balance and the nullspace separately, then decide when choosing a reference value resolves the problem.

Test the continuum equation with a constant

On $0 < x < 1$ take a unit conductivity, positive rightward flux $q = -u'$, and a steady source f with $q' = f$. Prescribe the rightward flux at both endpoints as q_L, q_R. Integrating the divergence equation gives

$$q_R - q_L = \int_0^1 f \, dx.$$

The outward unit normal is negative at the left endpoint, so the net outward flow is $q_R - q_L$, even when both recorded sensor values are positive. In several dimensions the same requirement follows by integrating the divergence and applying the divergence theorem.

$$\int_0^1 u'v' \, dx = \int_0^1 fv \, dx + q_L v(0) - q_R v(1).$$

Testing the weak equation with $v = 1$ makes its left side identically zero. This recovers the flux balance without differentiating an approximate field. For every compatible right-hand side, adding any constant to u leaves its gradient and both prescribed fluxes unchanged. The magnitude of a potential is undetermined, while potential differences and flux can be well defined. Each disconnected region with pure-flux boundaries has its own compatibility equation and arbitrary constant.

Background · Compatibility condition

$$\int_\Omega f \, dV = -\int_{\partial\Omega} \partial_n u \, dA$$

For minus the Laplacian of u equal to f with pure Neumann data, this integral condition is required. A constant still has to be fixed separately.

Separate the left nullspace test from the gauge condition

Take cell-centred finite volumes with equal width h and face flux $q_{i+1/2} = -(u_{i+1} - u_i)/h$. The balance of each cell is $q_{i+1/2} - q_{i-1/2} = hf_i$, with prescribed values replacing the two exterior faces. Add all rows. Interior face contributions cancel in equal and opposite pairs. The resulting symmetric matrix A has

a constant right null vector and its range is orthogonal to that vector. Let b be the integrated loads including boundary terms and $w_i = h$ the integration weights. A mean value $\bar{u}$ on this unit interval can be imposed by the augmented system

$$\begin{bmatrix} A & w \\ w^T & 0 \end{bmatrix} \begin{bmatrix} u \\ \lambda \end{bmatrix} = \begin{bmatrix} b \\ \bar{u} \end{bmatrix}, \qquad \lambda = \frac{\mathbf{1}^T b}{\mathbf{1}^T w}.$$

For compatible data $\lambda = 0$. For incompatible data the first block solves for a deliberately modified source $f_i - \lambda$. Choosing a mean makes the augmented system solvable either way; checking the multiplier identifies what it solved. Pinning a node and silently dropping one conservation equation can conceal the same mismatch as an artificial reaction at the pin.

Predict four cell values and audit an inconsistent measurement

Set $f = 2,\ q_L = 0,\ q_R = 2$. Exact integration gives

$$q = 2x,\ u = -x^2 + C.$$

For a continuous mean of zero $C = 1/3$. With four equal cells the centre coordinates are 1/8, 3/8, 5/8 and 7/8. The centroid samples of $-x^2$ have mean −0.328125, so the discrete mean-zero solution is $0.3125, 0.1875, -0.0625, -0.4375$. Adjacent differences give interior face fluxes 0.5, 1.0 and 1.5. The difference between the discrete centre values and the continuous mean-zero samples is the constant $-h^2/12$; every interior face gradient is exact.

Now change only the recorded right flux to 2.4. The original source supplies 2 units per unit time and the boundaries demand 2.4. Thus $\lambda = -0.40$ and the augmented system supplies an effective uniform source of 2.4. Alternatively, keeping the original source and the left measurement forces a right flux of 2. The plot shows both reconstructions and tells which data were modified. Adjusting only the mean translates either field without changing its slope. In a transient storage model, imbalance would change the mean in time instead of permitting a steady solution.

Set a nonzero flux mismatch, then vary the mean and the cell count independently. Observe which differences can be removed by a vertical translation. Check that the integrated imbalance remains the same under refinement.

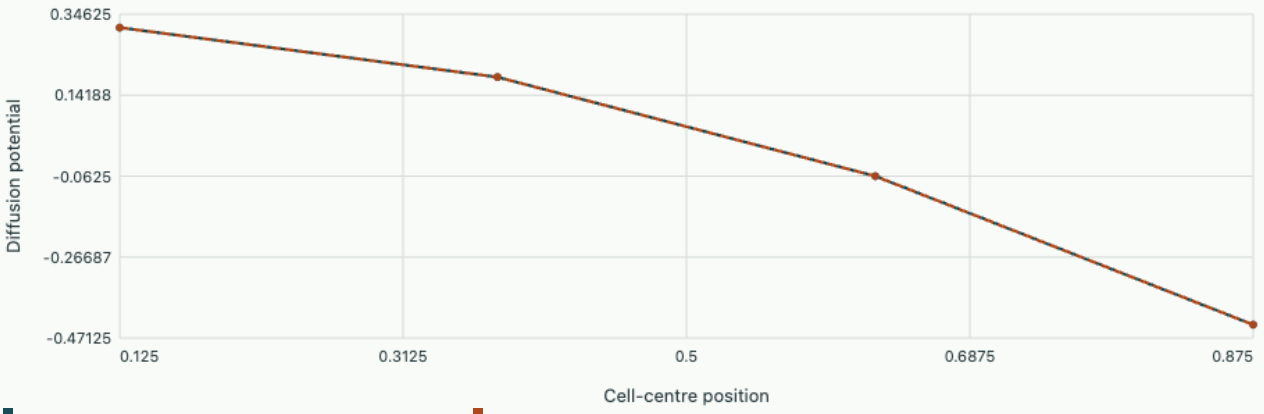
Open this experiment online

Try the calculation

Prescribed discrete mean **0**

Right flux minus 2 **0**

Number of cells **4**



Keep original source and left flux Enforce both fluxes with adjusted source

Global source minus measured outflow is 0.000. The data are compatible. The two reconstructions coincide. Both curves are shifted to a discrete mean of 0.000.

Tutorial · locate the bias and choose a comparison observable

A unit-length domain receives $f(x) = 1 + 2x$. A left sensor reports rightward flux 0.3 and a right sensor reports 2.1. Compute the integrated residual. If both flux sensors are trusted, find the constant source adjustment implied by the augmented system. If only the source and left sensor are trusted, predict the right measurement. What quantity can two solvers compare without agreeing on a gauge?

▼ **Show the diagnostic**

The source integral is 2 and net outflow is

$$2.1 - 0.3 = 1.8.$$

Their difference is +0.2. The multiplier is +0.2, so the adjusted source would be $0.8 + 2x$. Keeping the original source instead predicts

$$q_R = 0.3 + 2 = 2.3.$$

Differences $u(x_b) - u(x_a)$, gradients and fluxes are gauge independent. A single imbalance equation cannot tell which instrument is wrong. Independent source or flux measurements are needed to choose between these reconstructions.

Further reading · FEniCS Project · Poisson equation with pure Neumann boundary conditions and a mean-value multiplier

Why two finite-element wave solvers propagate the same mode at different speeds

Numerical partial differential equations

Build the element mass matrix by integrating shape functions and track the effect of diagonal lumping on the wave spectrum.

From a continuum wave to nodal masses

For $u_{tt} = c^2 u_{xx}$, multiply by a test function and integrate by parts. On a periodic line the boundary contributions cancel. A piecewise-linear element of width h uses shape functions $1-\xi, \xi$, with $\xi \in [0,1]$. Integrating products of shape functions gives

$$M_e = (h/6)\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Integrating derivative products gives

$$K_e = (c^2/h)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Row-sum lumping replaces each mass row by its sum on the corresponding diagonal, $h/2$. This preserves the mass associated with a uniform velocity but changes the response to a velocity that varies within an element.

Background · Numerical dispersion

$$c_{\mathrm{ph,h}} = \omega_h(k)/k$$

A discretised wave can retain its amplitude and still propagate at the wrong speed. Compare phase and group errors over the actual resolved wavelengths.

Diagonalise the assembled periodic problem

A periodic nodal mode has $u_j = e^{ij\theta}$. Each neighbour contributes $e^{\pm i\theta}$ relative to node j. Thus K multiplies by $2c^2(1-\cos\theta)/h$, consistent M by $h(2+\cos\theta)/3$, and lumped M by h. Dividing stiffness by mass gives the discrete squared frequency. Relative to the exact frequency $c\theta/h$,

$$\frac{\omega_L}{c\theta/h} = \frac{2\sin(\theta/2)}{\theta},$$

$$\frac{\omega_C}{c\theta/h} = \frac{2\sin(\theta/2)}{\theta}\sqrt{\frac{3}{2+\cos\theta}}.$$

Both tend to one for long waves; the leading quadratic errors have opposite signs. This is a spatial effect, present even with exact time integration.

Two wavelengths in one mesh

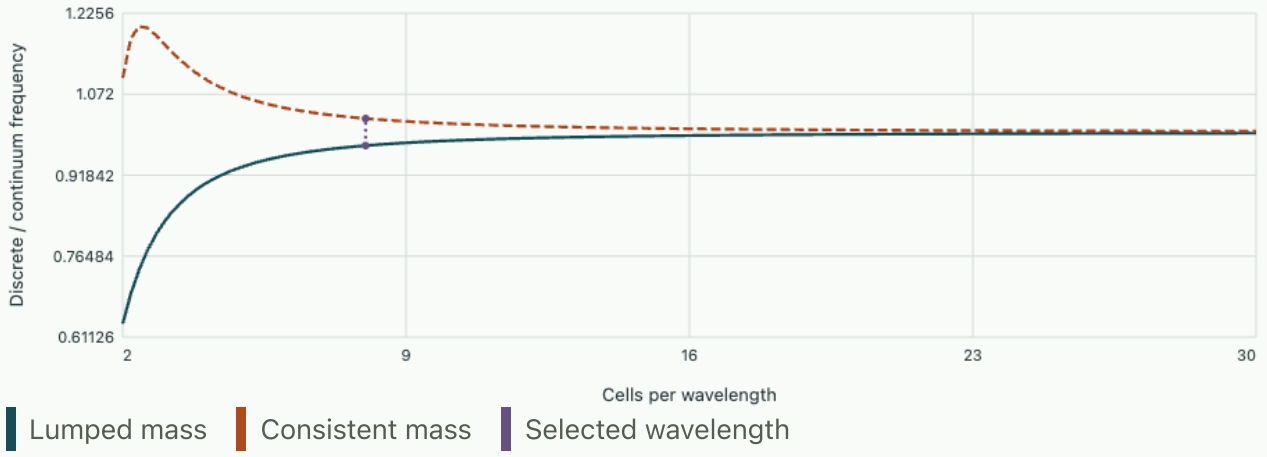
At four cells per wavelength $\theta = \pi/2$. The lumped ratio is $2\sqrt{2}/\pi \approx 0.9003$; the consistent ratio is about 1.1027. One wave solver lags by roughly ten per cent and the other leads by roughly ten per cent. At eight cells per wavelength the ratios move much closer to one, but repeated propagation over many wavelengths can still accumulate visible phase error.

At the shortest grid mode $\theta = \pi$, the consistent frequency is $\sqrt{12}c/h$ and the lumped frequency $2c/h$. A central-difference time integrator stable for $h_t\omega \le 2$ therefore requires a smaller step for the consistent mass. Accuracy at one wavelength and stability at the grid scale are different design criteria.

Open this experiment online

Try the calculation

Cells per selected wavelength 8



At 8 cells per wavelength the lumped ratio is 0.9745 and the consistent ratio 1.0259.

Tutorial on interpreting a convergence test

Which mass gives the exact kinetic energy for a linear velocity varying from v to $-v$ on one element?

▼ **Worked solution**

The continuum energy with unit density is $hv^2/6$. The consistent matrix reproduces it by exact integration. The lumped kinetic energy is $hv^2/2$, because it assigns both end masses the full endpoint speed.

How would you isolate spatial dispersion from time stepping in a measurement of a cavity eigenfrequency?

▼ **Worked solution**

Solve the generalised eigenproblem $Kq = \omega^2 Mq$ directly for increasingly fine meshes or converge time steps on each fixed mesh. Then compare with the continuum frequency and report both h and the time tolerance.

A converged field solver can still give the wrong voltage

Numerical field laboratory

Verify a two-dimensional electrostatic calculation with two independent charge patterns. Use symmetry to distinguish a transposed output, a coarse mesh and a poorly converged linear solve.

Choose a voltage before choosing a solver

A very long grounded tube has a square cross section $0 \le x, y \le L$, with $L = 1$ m. The contents are uniform along the axis. A model of its transverse charge density should predict what a voltage probe placed inside would measure. Three numerical statements that can sound similar are actually quite different: the discrete equations have a small residual, the grid solution approximates the continuous solution, and the physical source describes the tube. We can test the first two in a laboratory problem where the last is specified exactly.

Start by prescribing a voltage that satisfies all four walls. Introduce $s_m(x) = \sin(m\pi x/L)$ and choose $V_*(x, y) = [s_1(x) + g s_2(x)] s_1(y)$ volts. The parameter g is the signed amplitude of the second spatial pattern relative to the first. A probe at the centre sees exactly 1 V for every g, since $s_2(L/2) = 0$. A calculation checked at that point alone could entirely lose the second pattern and still pass. At $x = L/4$ on the centreline, $V_* = 1/\sqrt{2} + g$; at $3L/4$ it is $1/\sqrt{2} - g$. Their difference is $2g$ volts. First write down all three expected readings for $g = 0.5$.

Manufacture the source by differentiating the field

The physical problem uses Gauss's law and the electrostatic relation $\mathbf{E} = -\nabla V$. For uniform vacuum permittivity they yield

$$-\nabla^2 V = \rho/\epsilon_0.$$

A product $s_m(x) s_1(y)$ returns the factor $(m^2 + 1)\pi^2/L^2$ when acted on by the negative Laplacian. Applying this separately to the two patterns gives

$$\rho/\epsilon_0 = (\pi/L)^2 [2 s_1(x) + 5 g s_2(x)] s_1(y) \ \mathrm{V}.$$

The right side has dimensions V/m² after including the factor L^{-2}; multiplication by permittivity gives C/m³. When g becomes large enough, part of the imposed density changes sign. That is permissible for a manufactured source. The calculation represents a solution of the stated Poisson problem. Judging whether a particular apparatus could maintain that charge-transport model.

Uniqueness gives the reference solution special status. If there were a different voltage for the same walls and charge, their difference w would satisfy Laplace's equation with zero boundary value. Integrate $w(-\nabla^2 w)$ over the square. Integration by parts leaves $\int |\nabla w|^2$, because w is zero on the boundary. The integral must vanish; w is constant and its boundary makes that constant zero. This is why agreement with the manufactured voltage is a test of the solver and its input, rather than a choice between two physical solutions.

Background · How can the same source test a second boundary condition?

A smooth manufactured solution that vanishes on every edge tests a homogeneous Dirichlet solve. A code that incorrectly sets every requested boundary voltage to zero may still pass. Add a harmonic lift to the known field on a square of width L. For a constant permittivity the charge depends on the Laplacian, so the exact source is unchanged.

$$\widetilde{V}(x,y) = V(x,y) + V_0\Big(1 + \frac{x}{L}\Big),$$

$$-\epsilon\nabla^2\widetilde{V} = -\epsilon\nabla^2 V.$$

Prescribe the lifted values on all four sides and solve again using the same charge samples. The difference of the computed solutions should be the linear lift, which a centred second difference represents exactly on a uniform grid. A discrepancy points first to boundary-value handling or to the way known values enter the right-hand side. This comparison isolates that error from the truncation error on the original sine modes.

Solve the discrete problem on paper as a second reference

Let $h = L/n$. The standard negative Laplacian has centre coefficient $4/h^2$ and four neighbouring coefficients $-1/h^2$. Substitute the samples of $s_m(x)s_1(y)$. In the x direction the sum of forward and backward neighbours is twice the centre value times $\cos(m\pi/n)$. Collect the corresponding y term. The resulting positive grid eigenvalue is

$$\lambda_{h,m} = \frac{4}{h^2}\Big[\sin^2\frac{m\pi}{2n} + \sin^2\frac{\pi}{2n}\Big].$$

The continuum eigenvalue is $\lambda_m = (m^2+1)(\pi/L)^2$. If the source is the continuous source sampled at the nodes, the discrete voltage coefficient is multiplied by $\lambda_m/\lambda_{h,m}$. Since $\sin z < z$ for positive z, each factor exceeds one on a resolving grid. The faster spatial oscillation has the larger relative bias. The resulting total pointwise error can still change sign because the spatial modes themselves have signs.

This provides three independent calculations: a continuous closed-form field, an exact solution of the chosen difference equations and a numerical linear solve. Agreement of the latter two isolates assembly, indexing and algebraic errors. Comparing either with the continuous solution measures spatial discretisation error. Tightening an iterative tolerance indefinitely cannot remove a difference between $\lambda_{h,m}$ and λ_m.

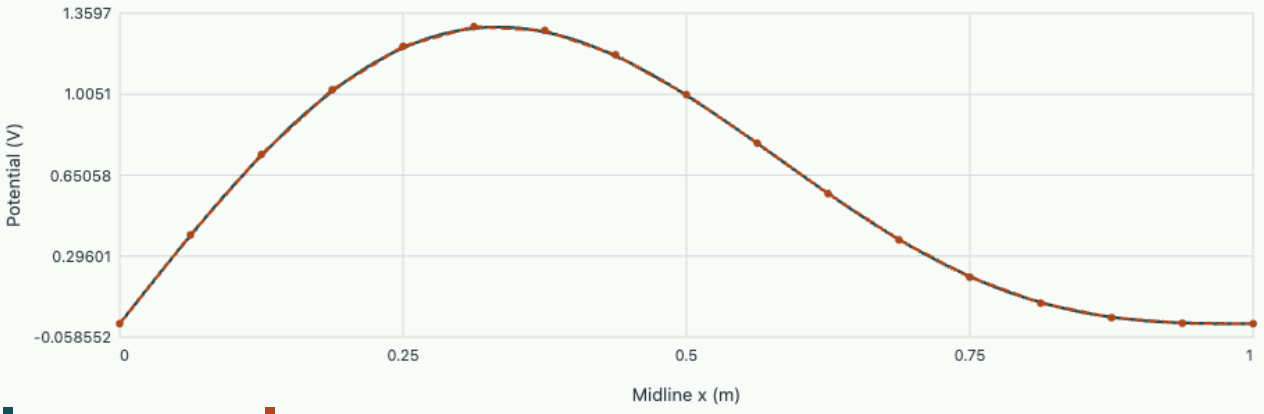
Run the full field and compare genuinely different observations

Open the live experiment

Try the calculation

Cells per side **16**

Signed second-mode amplitude **0.5**



Continuous field Solved grid field

Midline maximum error 0.0077708 V. Quarter-point difference 1.011 V. 2 CG step(s), relative residual 1.45e-13.

Download the OneSimulation model

In your OneSimulation environment, run the model with these inputs.

```
python field-laboratory.py modes --cells 16 --contrast 0.5
```

The browser constructs the right-hand side over the two-dimensional interior and applies conjugate gradients to the five-point operator from an all-zero initial field. The two-mode right-hand side belongs to a two-dimensional invariant subspace, so exact-arithmetic convergence takes at most two iterations. In the special case $g = 0$, only one mode is excited. This unusually small iteration count is a property of the test and says very little about a rough source on the same grid.

Download the model and run the displayed case. Its CSV contains the midline $y = L/2$, ordered by increasing x. For $n = 16,\ g = 2$ the value at $x = 0.1875$ m is about 2.42542 V on the grid and 2.40333 V analytically. If the centre is correct but the 0.1875 m reading is about 0.55736 V, check the orientation of the sampled data array. Keeping x fixed at its midpoint instead of y happens to annihilate the entire second mode.

A convergence report a second person can reproduce

Set $g = 0.5$. Record the centre voltage, the quarter-point voltage difference, the largest nodal error and the relative algebraic residual for $n = 8, 16, 32$. Use even grids divisible by four if reading exact quarter points. Estimate an order from two successive maximum errors using $p = \log(E_n/E_{2n})/\log 2$. For this smooth problem the leading difference between $\sin^2 z$ and z^2 is proportional to z^4, giving a second-

order spatial error at fixed physical modes. If g or the physical source profile changes during refinement, the ratio no longer tests that claim.

An experiment that quotes only the centre voltage has poor sensitivity to g. An experiment quoting only the algebraic residual may perfectly solve the wrong mesh problem. An experiment that compares each plot to itself after smoothing lacks an independent reference. The three diagnostics above address these difficulties with quantities small enough to check by hand.

Extend the verification to a genuinely three-dimensional field

Ground all six faces of a cube with side length L=1 m. A potential that varies in every coordinate is

$$V_*(x, y, z) = A \sin(\pi x/L) \sin(\pi y/L) \sin(\pi z/L),$$

$$\rho = 3\epsilon(\pi/L)^2 V_*.$$

Each second derivative contributes −(π/L)² times the potential. Adding the three derivatives gives the source above; every boundary vanishes because one sine is zero. This is a manufactured source for checking a numerical method. In the supplied calculation ε=1 F/m and A=1 V; scaling both charge and permittivity together leaves the voltage unchanged.

On a uniform nodal grid, the seven-point negative Laplacian has centre coefficient 6/h² and six neighbour coefficients −1/h². The same sine addition identity used in two dimensions gives

$$\lambda_h = \frac{12}{h^2} \sin^2\left(\frac{\pi h}{2L}\right),$$

$$V_h = \frac{3(\pi/L)^2}{\lambda_h} V_*.$$

The browser evaluates this discrete eigenmode directly. The spatial view places a coloured slice inside the cube; its height is the physical z coordinate. Moving that slice reveals the interior field while the companion line plot compares the grid values with the exact three-dimensional solution. The numerical potential is slightly too large because λh is smaller than its continuum limit.

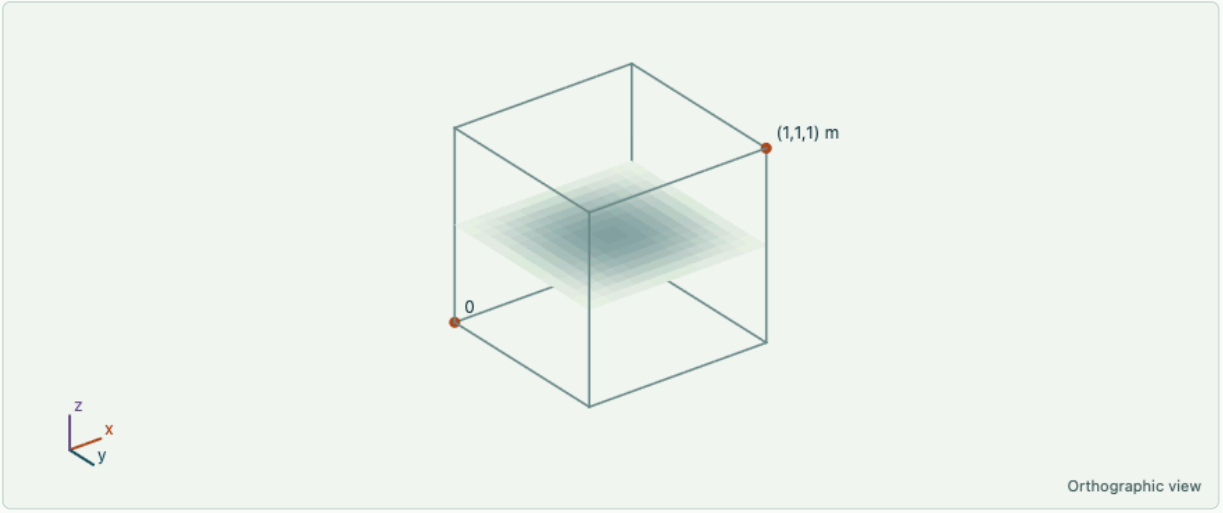
Try it · Rotate a 3D field, move its slice and refine its mesh

Move the plane from the centre to a grounded face and predict the change. Then compare 8 and 16 cells per direction: the maximum nodal error should decrease by approximately four. Doubling A doubles the absolute error while leaving the relative error almost unchanged. Rotation changes the viewpoint only.

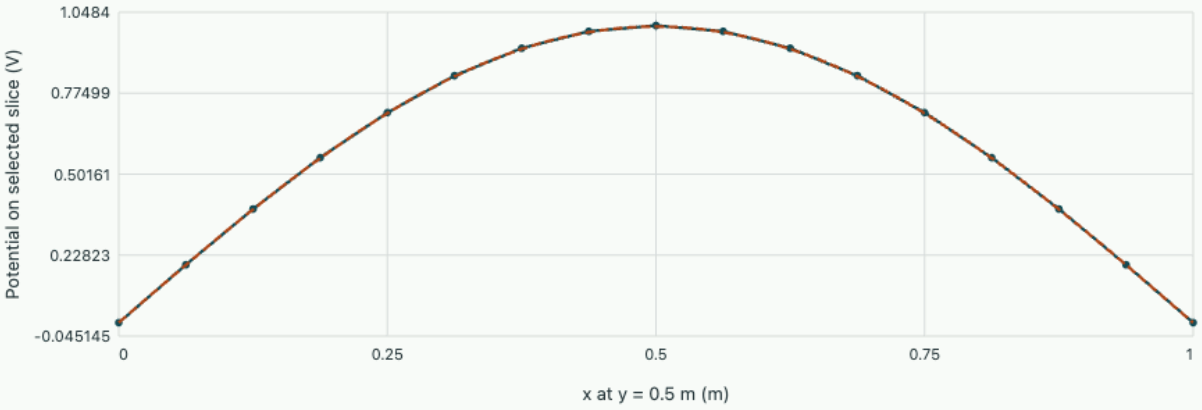
Open the live 3D experiment

Try the calculation

Cells per direction	16
Source amplitude (V)	1
Slice position (% of depth)	50



A slice through the three-dimensional potential, at z = 0.500 m. Colour is potential: pale = 0 V, dark green = 2.03 V; slice values range from 0 to 1.003 V. Grid-face colours average their four nodal values. Axes: x = position (m); y = position (m); z = position (m).



Seven-point FDM · nodal slice Exact 3D solution · same slice

16³ cells; selected grid plane z = 0.500 m. Maximum nodal error in the whole cube 3.219e-3 V (0.322%). All six boundary potentials are zero.

Download the OneSimulation model

In your OneSimulation environment, run the model with these inputs.

```
python poisson-cube-3d.py --cells 16 --amplitude 1 --slice 50
```

Download the 3D OneSimulation notebook · Download Python script

The notebook assembles and solves the full 3D OneSimulation finite-difference model. It reports an RMS error at cell centres; the browser reports a maximum error at nodes for its stated seven-point operator. These are different diagnostics. Compare each against its own continuous reference and repeat with a finer mesh.

```
"""Solve a grounded three-dimensional cube with OneSimulation and verify the field."""
import argparse
from dataclasses import dataclass
import numpy as np
import matplotlib.pyplot as plt
from onesim.cad.d3 import Box
from onesim.meshing import StructuredHexahedralMesher
from onesim.numerics import Numerics
from onesim.numerics.linear import SparseDirect
from onesim.numerics.fdm import FiniteDifference, CentralDifference, CentralLaplacian
from onesim.physics.electrostatics import FixedPotential, PoissonEquation
from onesim.simulation import Problem, Simulation

parser = argparse.ArgumentParser(description=__doc__)
parser.add_argument('--cells', type=int, default=16)
parser.add_argument('--amplitude', type=float, default=1.0, help='Reference potential amplitude in
V')
parser.add_argument('--slice', type=float, default=50, help='Depth as a percentage of the cube')
args = parser.parse_args()
if not 8 <= args.cells <= 24 or not np.isfinite(args.amplitude) or args.amplitude <= 0 or not 0 <=
args.slice <= 100:
    parser.error('Use 8–24 cells, a finite positive amplitude, and a slice between 0 and 100
percent')

@dataclass(frozen=True)
class Permittivity:
    unit = 'F/m'
    def evaluate(self, **state):
        return 1.0

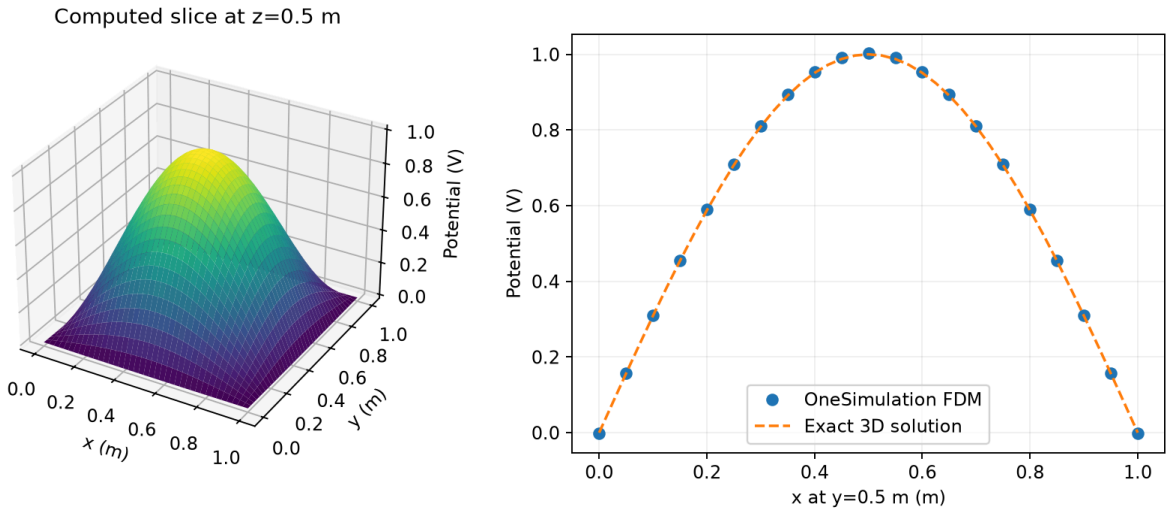
cube = Box(length=1, width=1, height=1)
exact = lambda x, y, z: args.amplitude * np.sin(np.pi*x) * np.sin(np.pi*y) * np.sin(np.pi*z)
field = PoissonEquation(domain=cube.volume, permittivity=Permittivity(),
                        charge_density=lambda x, y, z: 3*np.pi**2*exact(x,y,z))
problem = Problem(name='Three-dimensional manufactured potential', geometry=cube, physics=(field,),
                  boundary_conditions=(FixedPotential(physics=field, boundary=cube.boundaries.all,
value=0.0),))
mesh = StructuredHexahedralMesher(cells=(args.cells,)*3).generate(cube)
simulation = Simulation(problem=problem, mesh=mesh, numerics=Numerics(
    spatial=FiniteDifference(gradient=CentralDifference(), diffusion=CentralLaplacian()),
    linear_solver=SparseDirect()))
simulation.validate().require_valid()
result = simulation.solve_stationary()
potential = result.field(field.potential)
centres = mesh.coordinates[mesh.cells].mean(axis=1)
computed = potential.evaluate(centres)
reference = exact(*centres.T)
error = np.sqrt(np.mean((computed-reference)**2))
z = round(args.cells*args.slice/100)/args.cells
print(f'{args.cells}³ cells; source amplitude {args.amplitude:g} V; slice z={z:.3f} m')
print(f'Cell-centre RMS error against the exact 3D solution: {error:.6e} V')

x = np.linspace(0,1,41)
X,Y = np.meshgrid(x,x)
points = np.column_stack([X.ravel(),Y.ravel(),np.full(X.size,z)])
V = potential.evaluate(points).reshape(X.shape)
line = np.column_stack([x,np.full_like(x,.5),np.full_like(x,z)])
```

```
fig = plt.figure(figsize=(10,4),layout='constrained')
ax = fig.add_subplot(121,projection='3d')
surface = ax.plot_surface(X,Y,V,cmap='viridis',linewidth=0)
ax.set(xlabel='x (m)',ylabel='y (m)',zlabel='Potential (V)',title=f'Computed slice at z={z:g} m')
ax = fig.add_subplot(122)
ax.plot(x,potential.evaluate(line),'o',markevery=2,label='OneSimulation FDM')
ax.plot(x,exact(x,.5,z),'--',label='Exact 3D solution')
ax.set(xlabel='x at y=0.5 m (m)',ylabel='Potential (V)')
ax.grid(alpha=.2); ax.legend()
fig.savefig('poisson-cube-3d.png',dpi=170)
assert np.isfinite(error) and error/args.amplitude < .02
```

Measured output at the default settings:

```
16³ cells; source amplitude 1 V; slice z=0.500 m
Cell-centre RMS error against the exact 3D solution: 1.138076e-03 V
```



Actual OneSimulation output. Here the left panel plots potential as height above the selected x–y plane; the right panel compares the same field along y=0.5 m.

Tutorial

A colleague sets $g = -0.4$ and reports quarter-point voltages 1.107 V and 0.307 V on a very fine grid. Which change in the recorded setup would explain the observation?

▼ Work through the solution

For the stated sign, the expected pair in increasing x is 0.3071 V then 1.1071 V. Reversing the x axis interchanges the pair while preserving the centre and the symmetric mode. Reversing the sign of the second source mode produces the same pair. A field measurement with a labelled coordinate direction distinguishes these possibilities.

You prescribe the discrete eigenvalue as the source amplitude during every refinement. The returned solution agrees with your analytic curve at all nodes to machine precision. Which part of the verification

has disappeared?

▼ **Work through the solution**

This tests the algebra of the discrete eigenmode and the linear solve. It changes the continuum source on every grid and cancels exactly the spatial discretisation error you intended to measure. Rerun with a wavelength and source amplitude specified in physical units independently of the grid.

A second probe has independent voltage noise of standard deviation 5 mV. Estimate the standard deviation of g inferred from the difference of the two quarter-point readings, if both probes have the same independent noise.

▼ **Work through the solution**

The difference has variance $2(0.005)^2$ V² and equals $2g$ volts. Dividing by two gives standard deviation $0.005/\sqrt{2} \simeq 0.00354$ in the dimensionless amplitude. A shared additive offset cancels; a gain error common to both readings does not.

Measure tunnelling from a four-site spectrum

Computational quantum experiments

Four localised orbitals form a minimal double well. Raise the inner two site energies and ask why two states concentrated in the outer wells still acquire different energies. Measure the splitting in the spectrum, predict an oscillation in time and reproduce the levels independently with OneSimulation.

Construct an operator before drawing an energy diagram

Label the sites $1, 2, 3, 4$ in order. The end orbitals have on-site energy zero, the middle two have energy $V > 0$ and every neighbouring pair has real hopping matrix element $-t$ with $t > 0$. The four occupation states form an orthonormal basis for one particle. Equal diagonal energies of the ends do not make their symmetric and antisymmetric combinations degenerate when there is a virtual path through the interior. There are no direct links between sites one and three or one and four. Keeping track of every allowed step will fix the power of t in the final answer.

$$H = \begin{pmatrix} 0 & -t & 0 & 0 \\ -t & V & -t & 0 \\ 0 & -t & V & -t \\ 0 & 0 & -t & 0 \end{pmatrix}.$$

Reflection maps $1 \leftrightarrow 4$ and $2 \leftrightarrow 3$, and commutes with H. In each parity sector use the normalised outer and inner basis vectors $(|1\rangle \pm |4\rangle)/\sqrt{2}$ and $(|2\rangle \pm |3\rangle)/\sqrt{2}$. The diagonal outer element vanishes in both sectors; the hopping inside the middle pair shifts its even and odd energies to $D_e = V - t$ and $D_o = V + t$. The outer–inner matrix element remains $-t$. This reduces the four-dimensional problem to two symmetric two-by-two matrices without neglecting any state.

Eliminate the barrier amplitudes and estimate the splitting

In a sector with interior energy D, the secular equation is

$$E(E - D) = t^2.$$

The lower root is $E_-(D) = [D - \sqrt{D^2 + 4t^2}]/2$, or the numerically better $-2t^2/(D + \sqrt{D^2 + 4t^2})$. For $D \gg t$ the shift is approximately $-t^2/D$. It is more negative for the even state because $D_e < D_o$. Subtract the two shifts before approximating their difference.

$$\Delta E = E_-(D_o) - E_-(D_e) \simeq t^2 \left(\frac{1}{V - t} - \frac{1}{V + t} \right) \simeq \frac{2t^3}{V^2}.$$

A single outward excursion produces an order-t^2/V shift in both end states. Only a three-hop path connects the different ends, so their splitting is smaller. Measuring an overall displacement of two spectral lines and measuring the distance between them test different physical mechanisms.

Predict a time trace and define what is measured

Denote the two low-energy eigenstates by $|e\rangle, |o\rangle$. Prepare $(|e\rangle + |o\rangle)/\sqrt{2}$. In the corresponding opposite superposition $(|e\rangle - |o\rangle)/\sqrt{2}$, the occupation at time τ equals $\sin^2(\Delta E\,\tau/2\hbar)$. Its first maximum is at $\pi\hbar/\Delta E$. These superpositions include interior weight, especially at low barrier. A detector on site four alone measures a different projector. Preparing bare site one excites both high and low doublets and can generate additional fast oscillations.

For $V/t = 4$, the two lower eigenvalues in units of t are $(3 - \sqrt{13})/2$ and $(5 - \sqrt{29})/2$, about −0.30278 and −0.19258. Their separation is about $0.11019t$. The high-barrier approximation gives $0.125t$. With $t = 1\,\mathrm{meV}$, $\hbar = 0.65821\,\mathrm{meV\,ps}$ and the first complete transfer in the dressed basis takes about 18.8 ps. If a measurement integrates over many such oscillations, it may return one half despite perfectly coherent evolution.

Compare spectra from three independent constructions

Use the parameter controls to vary the middle barrier and hopping. The plot shows the exact splitting and its leading virtual-hopping prediction over a range of V/t. The printed result gives an actual transfer time. The downloadable OneSimulation model builds the original four-site Hamiltonian and solves its eigenvalue problem by contour integration. Compare all four eigenvalues with the two parity blocks before interpreting their small difference. A wrong sign on one edge of this open chain can be removed by a local change of basis and may leave its spectrum unchanged, so also check any phase-sensitive observable used for preparation.

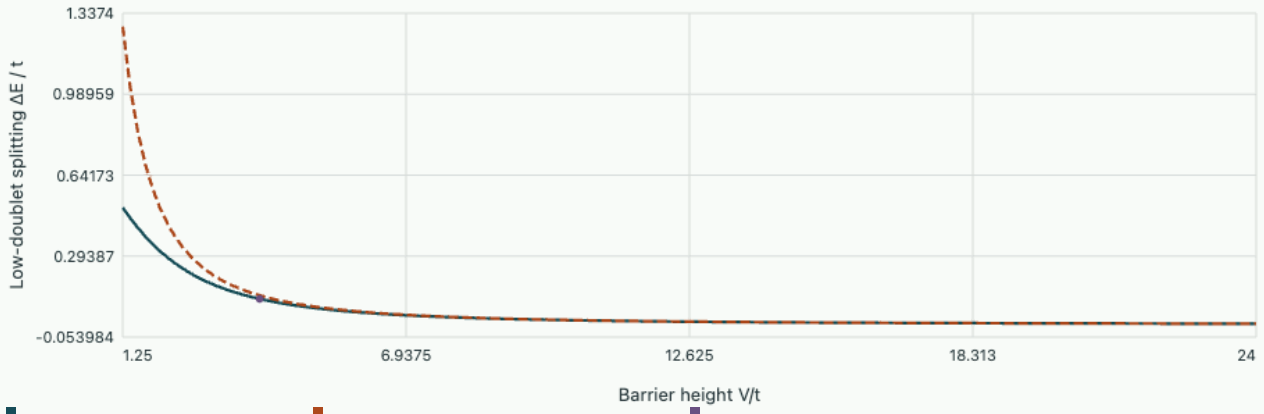
Keep the hopping fixed at 1 meV and compare two barriers. Start with a 4 meV barrier and raise it to 8 meV. Record both splittings and estimate the change in the first transfer time before reading off the prediction.

Open this experiment online

Try the calculation

Middle site energy V (meV) 4

Hopping t (meV) 1



Exact parity sectors | Three-hop approximation | Selected barrier

Splitting 0.11019 meV. First transfer between dressed end states at 18.766 ps.

Download the OneSimulation model

In your OneSimulation environment, run the model with these inputs.

```
python spectra-and-orbits.py wells --barrier 4 --hopping 1
```

Tutorial

Hold V fixed and halve t in the high-barrier limit. How do the common shift, splitting and transfer time change?

▼ **Work through the solution**

The leading common shift scales as t^2/V and falls by four. The splitting scales as t^3/V^2 and falls by eight. The transfer time grows eightfold. Halving both V and t instead keeps their ratio fixed and halves every eigenvalue.

Explain why a small static asymmetry between the end wells can suppress almost complete transfer while increasing their measured frequency splitting.

▼ **Work through the solution**

In a two-state reduction, detuning δ and coupling J give separation $\sqrt{\delta^2 + 4J^2}$. Starting in one localised state, the maximum transfer probability is $4J^2/(\delta^2 + 4J^2)$. Both the frequency and the contrast are needed to infer the coupling.

How could one check that a contour solver has found both members of a closely spaced doublet rather than printing one twice?

▼ Work through the solution

Check the rank of the independent projected vectors and the residual norm for each vector. Then exploit reflection parity or solve in each symmetry sector independently. Repeating a low residual for two nearly parallel vectors does not establish a two-dimensional eigenspace.

The continuous barrier calculation provides a different model in which barrier width eventually produces an exponential dependence.

How many lattice sites remember the colour of a neighbour?

Counting, correlations and thermal measurements

A chain of three-state spins gains energy when adjacent colours agree. A weak field favours one colour over the other two. Use a transfer matrix to calculate both free energy and the distance over which correlations survive. Recover the same numbers from OneSimulation eigenvalues without relying on symmetry reduction.

Find a local weight whose product counts each energy once

Each site has colour $a = 0, 1, 2$. A matching pair contributes energy $-J$, with $J > 0$, and a site of colour zero earns an additional $-H$. In a periodic ring each site belongs to two adjacent bonds. Give each of those bonds half the local field energy so the product of their weights counts each field term exactly once. Define $K = J/(k_B T)$ and $h = H/J$.

$$M_{ab} = \exp\left[K\left(\delta_{ab} + \frac{h}{2}(\delta_{a0} + \delta_{b0})\right)\right],$$

$$Z_N = \operatorname{tr}(M^N).$$

To see the trace, write the partition function as a sum over $a_1, \ldots, a_N$ of the product $M_{a_1 a_2} \cdots M_{a_N a_1}$. Summing an intermediate index is matrix multiplication; closing the initial and final index is the trace. This is exact even for a short periodic chain defined by that bond count. The transfer matrix contains no time step: multiplying it advances spatial position.

Compare the leading eigenvalue with the next one

Without bias the matrix has diagonal e^K and all other entries one. The uniform vector has eigenvalue $e^K + 2$. Every vector with components summing to zero has eigenvalue $e^K - 1$, twice degenerate. Thus $Z_N = (e^K + 2)^N + 2(e^K - 1)^N$. The thermodynamic free energy per site is $-k_B T \ln(e^K + 2)$. To measure a connected spatial correlation, insert a diagonal colour observable at sites separated by r transfer operations. The uniform contribution cancels from the connected part; each nonuniform mode contributes a factor $(\lambda_j/\lambda_1)^r$ when its overlap is nonzero. The longest correlation length in sites is

$$\xi^{-1} = \ln(\lambda_1/|\lambda_2|).$$

At $k_B T = J$ and zero bias, $\lambda_1 = 4.71828,\ \lambda_2 = 1.71828$ and $\xi = 0.9900$ sites. At low temperature e^K is large and $\xi \simeq e^K/3$. A finite chain can then appear ordered simply because it is shorter than its correlation length. There is no finite positive temperature at which these two one-dimensional eigenvalues cross.

Keep track of a mode that a chosen detector cannot see

At $h \neq 0$, colours one and two remain interchangeable. Their difference vector is an eigenmode of eigenvalue $e^{K} - 1$. In the orthonormal basis comprising colour zero and the sum of one and two, the remaining block is

$$\begin{pmatrix} e^{K(1+h)} & \sqrt{2}e^{Kh/2} \\ \sqrt{2}e^{Kh/2} & e^{K}+1 \end{pmatrix}.$$

Its eigenvalues follow from the trace and the square root of the squared diagonal difference plus four times the squared off-diagonal element. Sort all three eigenvalues. A detector sensitive only to whether the spin is zero is invariant under interchange of colours one and two. It has no overlap with their odd difference mode, even if that mode sets the mathematically longest length. A resolved measurement of colours one and two can see it. Quoting a correlation length from eigenvalues requires specifying which observation can actually couple to that eigenvector.

Change temperature and examine the spectral gap

Vary reduced temperature and reduced field H/J independently. The plot compares the longest colour-correlation length to the exact zero-field result. At a chosen point the downloadable model constructs the full three-by-three matrix and computes its modes using OneSimulation's contour eigensolver. It reports free energy per site and all three eigenvalues. In this small system the direct reduction above is available as an independent check. The same matrix construction generalises to more states even when no colour symmetry reduces the size.

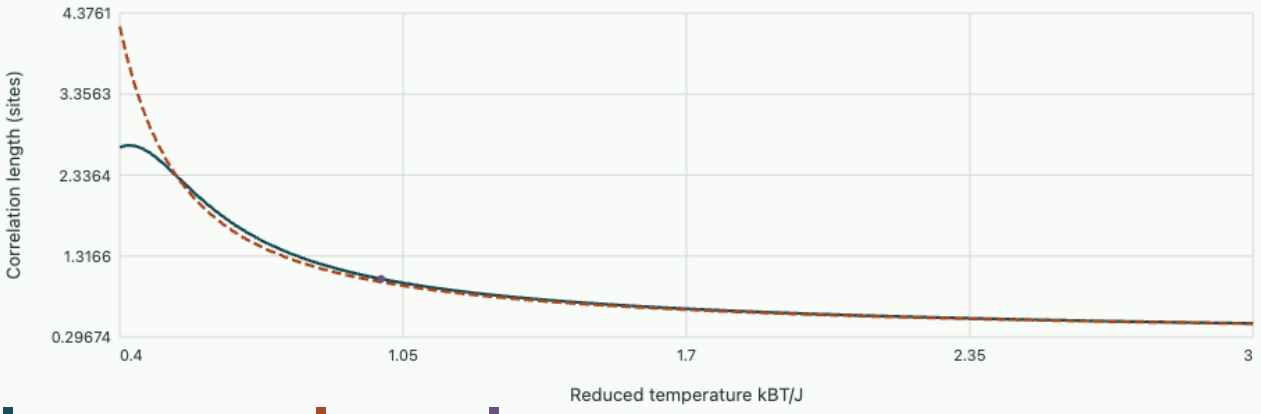
Compare two chains at the same temperature. Set the colour bias to zero and apply a positive bias of 0.30 J. Use the transfer-matrix eigenvalue ratio to explain the change in the longest correlation length.

Open this experiment online

Try the calculation

Reduced temperature kBT/J **1**

Reduced colour field H/J **0.15**



Selected colour bias | Zero bias | Selected temperature

Leading eigenvalue 4.9881, second 1.8884, longest length 1.0295 sites.

Download the OneSimulation model

In your OneSimulation environment, run the model with these inputs.

```
python spectra-and-orbits.py chain --temperature 1 --bias 0.15
```

Tutorial

At $k_B T = J,\ h = 0$ what fraction of the asymptotic leading contribution to a connected correlation survives after three lattice sites?

▼ Work through the solution

The factor is $(1.71828/4.71828)^3 \simeq 0.0483$. The correlation observable sets its initial amplitude; the eigenvalue ratio sets this fraction.

Why does squaring a single bond weight to represent the effect of two bonds give the wrong two-step transition amplitude?

▼ Work through the solution

Two steps can pass through any of the three intermediate colours. Matrix multiplication sums all possibilities $M_{ab}M_{bc}$ over b. Squaring the single number M_{ac} loses those intermediate configurations.

Suppose only the leading eigenvalue is reliably resolved. Which thermodynamic-limit prediction is possible and which remains undetermined?

▼ **Work through the solution**

One can obtain free energy per site, and with its parameter dependence the derivatives giving energy or magnetisation. A connected spatial length requires a subleading eigenvalue and overlap of the chosen observable. A precise leading free energy alone cannot give that numerical length.

Compare this transfer calculation with the two-state Ising chain experiment.

Ill conditioning and information

Inspect the small singular directions in actual observation operators. Distinguish a good residual from a well-determined physical parameter.

1. Weighted least squares and residual diagnostics
2. Singular systems and the amplification of noise
3. Regularisation, resolution and the price of stability
4. Adjoint gradients through a forward solver
5. Choosing regularisation when the true answer is hidden
6. How model error masquerades as a fitted physical parameter
7. Capacitance matrices, floating nodes and controlled approximation bounds
8. Reconstructing a source from noisy voltage measurements
9. Which part of a sample changes a four-electrode impedance measurement?
10. Scattering from a central potential and what angular data can recover

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,

$$b = L^{-1}y,$$
$$B = L^{-1}A,$$
$$\operatorname{Cov}(L^{-1}\epsilon) = I.$$

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

$$\sqrt{0.10^2 + 0.20^2} = 0.224\,\text{g}.$$

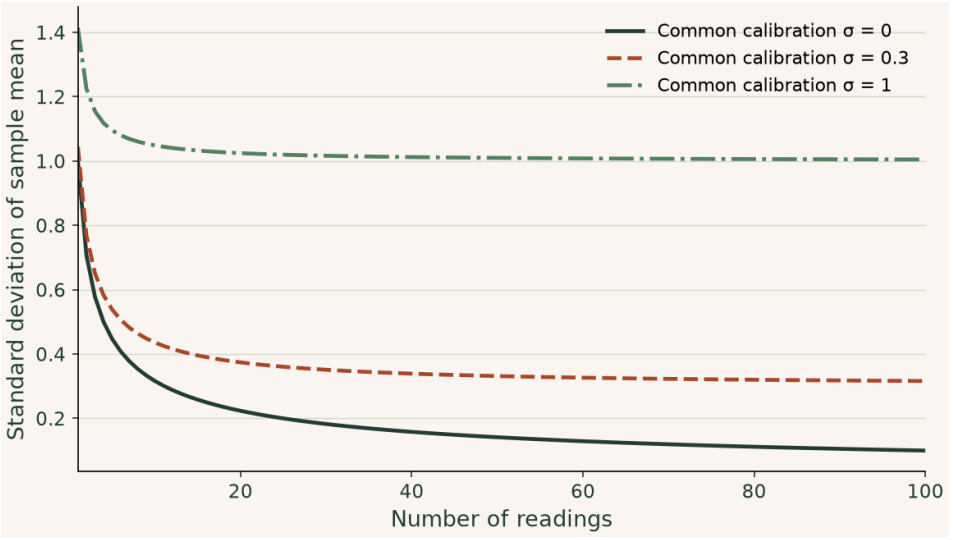
The observed difference of 0.50 g is about 2.24 such standard deviations. That discrepancy is relevant to the model for individual readings and averaging windows. It supplies no direct evidence for or against a common calibration offset of a particular size, because that offset has cancelled algebraically.

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 $1/\sqrt{N}$ law misses this limiting uncertainty.

▶ **Read values from the plot**

Change parameters · Find a measurement floor that averaging cannot remove

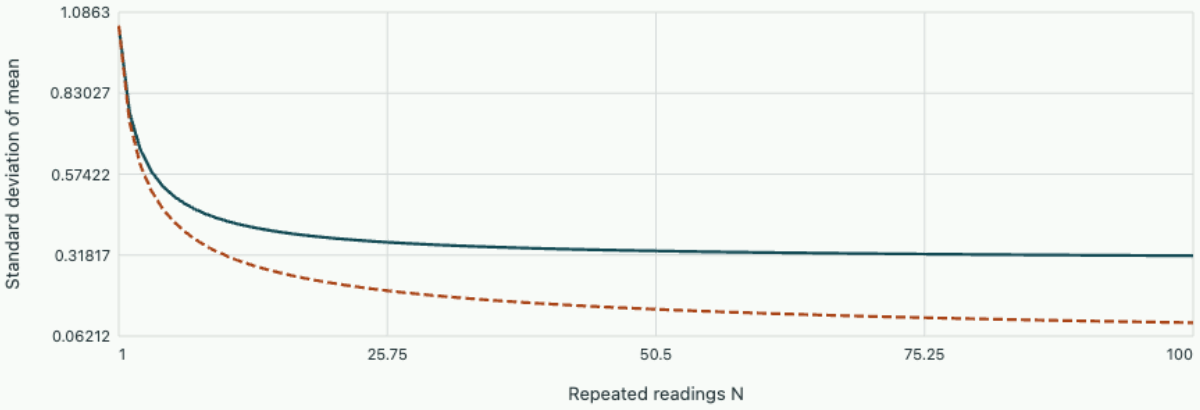
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Shared calibration SD **0.3**

Independent reading SD **1**



Correct common-offset covariance Incorrect independent-offset assumption

The limiting standard deviation is the shared calibration SD 0.3. Repeating the same instrument does not remove its common unknown offset.

Singular systems and the amplification of noise

Inverse problems and least squares

Identify which patterns survive a forward measurement and which become dominated by uncertainty when inverted.

FROM FIRST PRINCIPLES

Identify the measured pattern hidden in each singular direction

A forward measurement $y = Ax + \epsilon$ is easiest to interpret in orthonormal directions where A acts by simple scaling. For a singular triplet with positive σ_j, $Av_j = \sigma_j u_j$ and $A^T u_j = \sigma_j v_j$. Project the complete data equation onto its output direction,

$$u_j^T y = \sigma_j v_j^T x + u_j^T \epsilon.$$

Recovering the coefficient in the state divides the measured projection by σ_j. Thus the reconstruction error in that direction is

$$v_j^T(\widehat{x} - x) = \frac{u_j^T \epsilon}{\sigma_j}.$$

When the data are whitened with unit noise covariance, the projected noise has standard deviation one and the recovered coefficient has noise standard deviation $1/\sigma_j$. A zero singular value hides the corresponding state component completely. The numerical rank of a matrix and the set of coefficients that can be inferred at useful precision can therefore differ even for an exactly known forward model.

DERIVATION AND APPLICATIONS

Derive the singular spectrum of a local averaging instrument

An instrument that averages neighbouring locations transmits broad patterns and attenuates rapid changes. Consider a periodic one-dimensional array of N equally spaced cells. At every cell the detector returns a symmetric three-point average, with response coefficient $0 \le a \le 1/2$. With wraparound indices the noiseless map is

$$(Ax)_j = a x_{j-1} + (1 - 2a) x_j + a x_{j+1}.$$

Translation symmetry suggests using normalized Fourier patterns $v_k(j) = e^{2\pi i k j / N} / \sqrt{N}$. Neighbouring samples of one mode differ only by a phase. Substitution yields its eigenvalue,

$$Av_k = \lambda_k v_k,$$
$$\lambda_k = 1 - 2a + 2a\cos(2\pi k/N) = 1 - 4a\sin^2(\pi k/N).$$

The averaging matrix is symmetric and the Fourier modes are orthonormal, so the singular value of each mode is $|\lambda_k|$. A negative eigenvalue reverses the measured phase of a real sinusoidal pattern but has the same noise amplification as a positive one of equal magnitude. A zero destroys the mode. For independent equal-variance data errors of variance s^2, an inverse on a nonzero mode gives

$$\mathrm{E}\,|\delta x_k|^2 = \frac{s^2}{|\lambda_k|^2}.$$

The periodic boundary assumption supplies the Fourier eigenvectors; a finite array with distinct edge rules generally has a different spectrum. Preservation of the constant mode, $\lambda_0 = 1$, checks that the normalized averaging kernel conserves total signal.

WORKED EXAMPLE

Identify an invisible checkerboard pattern

For $N = 8$ and $a = 1/4$, the alternating vector corresponds to $k = 4$ and has $\lambda_4 = 0$. No improvement in readout precision can infer its coefficient from these data. The quarter-cycle mode $k = 2$ instead has $\lambda_2 = 1/2$; noise of standard deviation 0.020 in its measured coefficient becomes 0.040 in the inferred amplitude.

SECOND APPLICATION

Change the optical mixing and forecast a contrast limit

A second array has $a = 0.20$ and even N, giving alternating-mode gain $1 - 4(0.20) = 0.20$. A true normalized contrast coefficient 0.30 appears as 0.060 in the data. At coefficient noise deviation 0.030 that is only a two-standard-deviation response. To infer the same state coefficient with deviation 0.050, the measurement deviation must fall to $0.20(0.050) = 0.010$.

Rotate the coordinates to expose the experiment

For a whitened matrix $B = U\Sigma V^{\mathsf{T}}$, the unit vector v_j is mapped to $\sigma_j u_j$. Parameter patterns v_j and measurement patterns u_j are orthonormal in their respective coordinates. A small singular value means a full unit change in a particular parameter combination makes only a small change in the data. This conclusion depends on scaling: metres or millimetres for one coordinate change the geometry. Scale each parameter by a physically meaningful range before interpreting a singular vector as a weak process.

For nonzero singular values the least-squares minimum-norm inverse takes the form

$$\hat{x} = \sum_{\sigma_j > 0} \frac{u_j^{\mathsf{T}} b}{\sigma_j} v_j.$$

Left-singular directions orthogonal to the range contribute to the residual and cannot be fitted. Right-null directions contribute nothing to the data and are assigned zero by the minimum-norm convention. For a full-column-rank matrix the largest response of the inverse to a unit data perturbation has norm $1/\sigma_{\min}$. A relative worst-case bound involves $\kappa_2 = \sigma_{\max}/\sigma_{\min}$; the actual perturbation direction can be far less damaging.

A blur that almost removes the contrast

Let a sensor mix two unknown levels with $A = \left(\begin{smallmatrix} 0.505 & 0.495 \\ 0.495 & 0.505 \end{smallmatrix}\right)$. It passes the average direction unchanged and attenuates the difference direction by 0.01. For $x = (2, 0)$ it predicts $y = (1.01, 0.99)$. Add the modest perturbation $(0.01, -0.01)$. Inverting the perturbed data exactly gives $(3, -1)$: a 0.01 perturbation per sensor produced a unit change in each reconstructed level. In contrast, perturbation $(0.01, 0.01)$ just raises both fitted levels by 0.01. Both inversions have zero residual. Compare the recovered patterns with independently known precision to decide which interpretation is supportable.

Why finer discretisation can make an inverse problem harder

An integrating or blurring instrument smooths rapidly oscillating parameter fields. Discretising more finely introduces additional oscillatory patterns, each with very small image under the forward map. For a convolution on a periodic grid the Fourier modes themselves diagonalise the operation: the response to each frequency is the Fourier transform of the kernel there. In a Gaussian blur this multiplier falls rapidly with frequency. Formally dividing by it at every resolved frequency amplifies high-frequency noise. Finer output sampling cannot automatically recover what the original acquisition smoothed away.

Plot σ_j, $|u_j^{\mathsf{T}} b|$, and their ratio against index. For recoverable modes the projected signal can be distinguished from the noise floor. When the singular values keep decreasing but the data coefficients fluctuate at roughly constant noise magnitude, division begins to generate structure from noise. The transition can be gradual and projections are random, especially in a short record. Use noise estimates or replicated observations before choosing a cutoff.

Distinguish numerical rank from usable resolution

A singular value large enough to represent in floating point may still give a parameter uncertainty too large for the scientific question. Conversely several poorly measured parameters may admit a well-measured sum. In whitened coordinates the uncertainty of a linear functional follows from how it projects onto v_j. A functional orthogonal to small-singular-value directions can remain stable without the full reconstruction being stable. The next chapter makes the decision to suppress these directions explicit and quantifies the bias it creates.

Ask whether a continuum solution has finite energy

A finite matrix has only finitely many singular values, so a formally inverted vector can always be printed after zero modes have been removed. For a compact smoothing operator on a space of square-integrable functions, there can be infinitely many positive singular values tending to zero. Write the data coefficients as b_j in the left singular basis, with no data component orthogonal to the closure of the range. Solving mode by mode forces the state coefficients to be $x_j = b_j/\sigma_j$. Orthogonality gives

$$\|x\|^2 = \sum_j |b_j/\sigma_j|^2.$$

A square-integrable solution exists exactly when that last series converges; an arbitrary additional nullspace component can then be added if the operator is not injective.

This range condition explains why diminishing data amplitudes are insufficient evidence of a reconstructible state. They must diminish sufficiently fast relative to the strength with which the instrument transmits each mode. The converse is constructive: if the squared quotient series converges then the partial reconstructed states converge in norm, and continuity of the forward operator takes their limit to the original compatible data. A finite truncation conceals failure of this condition because every partial sum is finite.

Follow two spectra as the reconstruction grid grows

Take singular values $\sigma_j = 1/j$, and compare data sets with coefficients j^{-2} and $j^{-3/2}$, respectively. Both data sets have finite squared norm. In the first, the reconstructed coefficients are $1/j$ and the squared state norm is $\sum j^{-2}$, which converges. Beyond a cutoff N, its tail is bounded by $\int_N^\infty x^{-2}dx = 1/N$. Keeping 100 modes therefore leaves state error at most 0.10 in this particular normalisation. A corresponding forward residual can be much smaller because applying the operator divides each tail term by an additional j.

In the second data set, the state coefficients are $1/\sqrt{j}$. Their squared norm through mode N is the harmonic sum. From 100 to 200 retained modes it increases by about $\log 2$, and the same approximate increase recurs on every doubling. Meanwhile the squared data discrepancy after truncation is $\sum_{j>N} j^{-3}$, bounded by $1/(2N^2)$. At 100 modes the data residual is at most 0.0071. The residual can become arbitrarily small although no finite-energy exact state exists for this idealised data set.

In a measured spectrum, only finitely many coefficients are observed and their far tail includes instrument noise. Exact membership in an infinite-dimensional range cannot be proven by drawing a straight line through a short log plot. The continuum calculation instead suggests two independent checks during refinement: inspect whether reconstruction norms stabilise, and compare data coefficients to an independently estimated noise floor before dividing them by small singular values.

For the first sequence even an excellent forward residual is a relatively optimistic diagnostic: the squared forward tail behaves like N^{-3}, while the squared state tail behaves only like N^{-1}. Comparing norms without specifying whether the operator has already smoothed the discrepancy makes the two convergence rates appear contradictory.

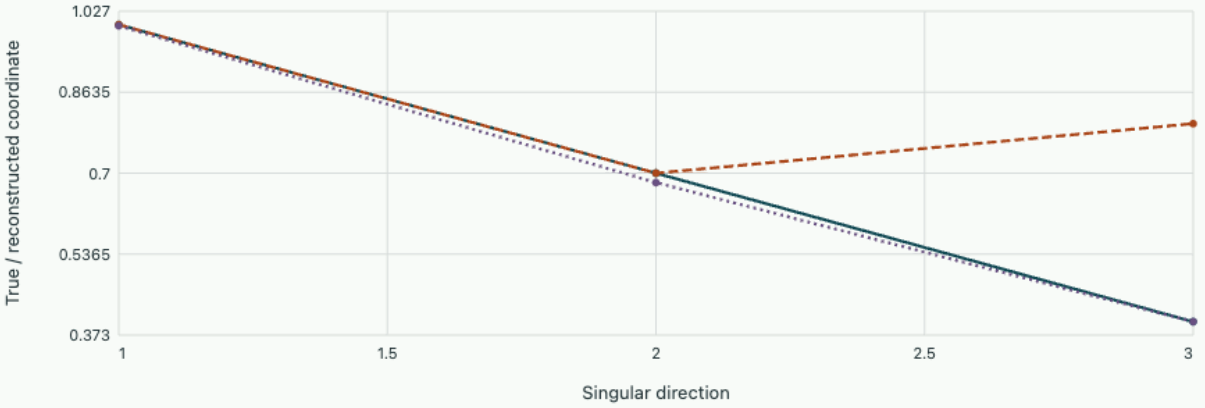
Change parameters · Singular systems and the amplification of noise

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

Weak singular value	**0.05**
Noise in weak data coordinate	**0.02**
Regularisation μ	**0.05**



Weak-mode noise is amplified by 20 before regularisation. The regularised weak-mode resolution gain is 0.5.

Regularisation, resolution and the price of stability

Inverse problems and least squares

Choose which features to recover, derive the associated filter and quantify the bias introduced by that choice.

FROM FIRST PRINCIPLES

Derive how a quadratic penalty filters small singular values

Suppose whitened data b are modeled by Ax, and large reconstructed coefficients are implausible. Penalize their squared norm with strength $\lambda^2 > 0$. The minimizer balances the measured residual and penalty,

$$\min_x \|Ax - b\|^2 + \lambda^2 \|x\|^2.$$

In singular-vector coordinates let $z_j = v_j^T x$ and $\beta_j = u_j^T b$, where $Av_j = \sigma_j u_j$. Orthogonality makes the objective a sum of scalar terms $(\sigma_j z_j - \beta_j)^2 + \lambda^2 z_j^2$, plus a constant from unmodeled data directions. Setting the derivative of each term to zero gives

$$z_j = \frac{\sigma_j}{\sigma_j^2 + \lambda^2} \beta_j.$$

At $\sigma_j \gg \lambda$ this is close to the ordinary inverse. At a tiny singular value it avoids dividing the noise by that tiny number, at the price of attenuating the true coefficient. A null component is set to zero by this penalty. A different physical reference state or a derivative penalty imposes different information and changes what the reconstruction preserves.

Background · Tikhonov estimate

$$\arg\min_m \{\|Gm - y\|^2 + \alpha \|Lm\|^2\}$$

The regularisation operator selects which departures are costly. Compare values of alpha only after fixing how both terms are scaled.

DERIVATION AND APPLICATIONS

Predict a withheld observation without repeating every regularized fit

Suppose a linear model has independent equal-variance readings y_i, design rows a_i^T, and a fixed positive quadratic penalty $\lambda^2\|x\|^2$. One way to compare penalty strengths is prediction on withheld readings. Define $H = A^TA + \lambda^2 I$, fit all rows to obtain $\widehat{x}$, and let $\widehat{x}_{-i}$ be the fit with row i removed while retaining the same penalty. Adding and subtracting the omitted row in the smaller normal equations yields

$$H\widehat{x} = A^Ty,$$
$$H\widehat{x}_{-i} = A^Ty - a_i(y_i - a_i^T\widehat{x}_{-i}).$$

Subtract the second identity from the first. Multiplying by $a_i^TH^{-1}$ shows that the shift in the fitted prediction is proportional to the withheld residual. The constant of proportionality h_i is its regularized leverage,

$$a_i^T(\widehat{x} - \widehat{x}_{-i}) = h_i(y_i - a_i^T\widehat{x}_{-i}),$$
$$h_i = a_i^TH^{-1}a_i.$$

Writing $r_i = y_i - a_i^T\widehat{x}$ and rearranging gives the leave-one-out residual and corresponding validation score,

$$y_i - a_i^T\widehat{x}_{-i} = \frac{r_i}{1-h_i},$$
$$\mathrm{CV}(\lambda) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{r_i}{1-h_i}\right)^2.$$

A point with large leverage helps set its own fitted value, so its ordinary residual can be misleadingly small. The formula assumes rowwise omission makes sense for the data. If nearby observations share noise or the intended prediction is a whole new experiment, withholding complete independent groups answers a different question.

WORKED EXAMPLE

Check the shortcut against two explicit refits

Fit a constant to $y = (1, 3)^T$ with $\lambda^2 = 1$. The all-data estimate is $4/3$, and both leverages are $1/3$. Dividing the ordinary residuals by $2/3$ gives withheld errors $-0.5, 2.5$, hence $\mathrm{CV} = 3.25$. Fitting only 3 yields $3/2$; fitting only 1 yields $1/2$, confirming both predictions.

SECOND APPLICATION

Detect a high-leverage reading

For another scalar fit take sensitivities $1, 3$, data $2, 4$ and $\lambda^2 = 1$. The estimate is $14/11$, and the second reading has ordinary residual $2/11$ and leverage $9/11$. Its withheld error is

$$(2/11)/(2/11) = 1.$$

Directly using only the first reading gives $x = 1$, which predicts 3 at the second station rather than the observed 4.

Two ways to control the unstable inverse

A truncated SVD keeps only the first k singular vectors ordered by decreasing singular value. All retained coefficients are inverted exactly; the others are set to zero. Its noiseless reconstruction is a projection onto the retained right-singular subspace. The choice is transparent but a hard cutoff can place two comparably informative directions on opposite sides of the threshold. In some spatial reconstructions sharp spectral cutoffs also introduce oscillations.

A smoother choice is zero-order Tikhonov regularisation. In scaled parameters and whitened measurements minimise $\|Bx - b\|^2 + \lambda^2 \|x\|^2$ for $\lambda > 0$. Differentiation gives

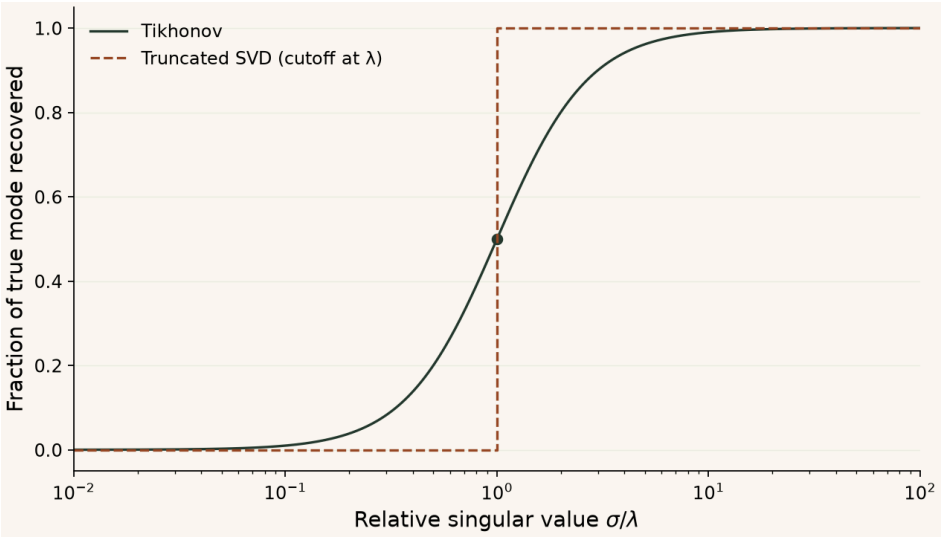
$$(B^{\mathsf{T}} B + \lambda^2 I) x_\lambda = B^{\mathsf{T}} b.$$

Solve with an augmented QR or with the SVD. Each coefficient now carries a filter:

$$x_\lambda = \sum_j \frac{\sigma_j}{\sigma_j^2 + \lambda^2} (u_j^{\mathsf{T}} b) v_j,$$

$$f_j = \frac{\sigma_j^2}{\sigma_j^2 + \lambda^2}.$$

If $b = Bx$ exactly, the estimate of $v_j^{\mathsf{T}} x$ is the true coefficient times f_j. Where $\sigma_j = \lambda$ the noiseless amplitude is halved. For a very small singular value, the noise gain now approaches zero; the price is that a real fine-scale feature in the same direction is also suppressed.



Noiseless resolution of an individual singular mode. A value of one half means that the inversion retains half of the true amplitude, even with perfect data.

Calculate the bias and variance for one component

Suppose a scalar instrument reads $b = 0.1x + \epsilon$, true $x = 2$, with noise standard deviation 0.1. Direct inversion gives an unbiased estimate with standard deviation 1. At $\lambda = 0.1$ the filter is $1/2$; the estimated mean is 1 and the standard deviation is 0.5. Mean squared error is bias squared plus variance: 1.25 here, compared with 1 for direct inversion. For true $x = 0$ the same regularisation would reduce mean squared error from 1 to 0.25. Stabilising an estimate does not guarantee a smaller error for every possible signal. The importance assigned to plausible signals is part of the modelling decision.

Turn a source condition into a noise-dependent accuracy bound

The bias–variance calculation describes a particular signal and a noise distribution. Suppose instead we know only a deterministic data-error bound $\|b^\delta - Bx^\dagger\|_2 \le \delta$. To derive a uniform reconstruction guarantee we must also specify which signals qualify. One useful assumption is the source condition $x^\dagger = B^T w,\ \|w\|_2 \le R$, in scaled coordinates. In singular coordinates a coefficient of $x^\dagger$ is then σ_j times a coefficient of w. Thus arbitrarily large features cannot hide in arbitrarily weak directions for the given R. This assumption also excludes a component in the null space.

Write $\alpha = \lambda^2$ for the penalty coefficient. Split the reconstruction error into the data perturbation passed through the regularised inverse and the loss produced by regularising exact data. In a mode of singular value σ, the noise multiplier is $\sigma/(\sigma^2 + \alpha)$. The noiseless bias is $-\alpha/(\sigma^2 + \alpha)$ times the true coefficient, or $-\alpha\sigma/(\sigma^2 + \alpha)$ times the source coefficient. Because $(\sigma - \sqrt{\alpha})^2 \ge 0$, the first multiplier is bounded by $1/(2\sqrt{\alpha})$. The same inequality bounds the magnitude of the second by $\sqrt{\alpha}/2$. Orthogonality and the triangle inequality now give

$$\|x_\alpha^\delta - x^\dagger\|_2 \le \frac{\delta}{2\sqrt{\alpha}} + \frac{R\sqrt{\alpha}}{2}.$$

If $R, \delta > 0$, the two terms balance at $\alpha = \delta / R$, giving the bound $\sqrt{R\delta}$. This is an a priori parameter choice based on stated uncertainty. For a family of increasingly resolved operators, keeping R controlled matters. If the model class admits large coefficients in directions of vanishing sensitivity, the same claim has no uniform basis.

When bias and data error reinforce one another

Take $B = \operatorname{diag}(1, 0.1)$ and $w = (0, 1)^T$, so $R = 1$, the true signal is $(0, 0.1)^T$, and exact data are $(0, 0.01)^T$. Suppose the second measurement is perturbed downwards by 0.01. Then

$$\delta = 0.01, \ \alpha = 0.01.$$

With exact data, reconstruction would retain half the weak coefficient, producing 0.05 and bias -0.05. The data perturbation is multiplied by $0.1/(0.01 + 0.01) = 5$, adding another -0.05. The reported signal is zero, and the error 0.1 attains the theoretical bound. Here the vulnerable direction has precisely $\sigma = \sqrt{\alpha}$. On another spectrum the achievable error may be much smaller.

Choose what the penalty expresses

For an expected background x_0 and a derivative operator D, a suitable criterion may use $\|D(x - x_0)\|^2$. A first-difference penalty discourages adjacent jumps while leaving constants unpenalised; a second-difference penalty instead leaves affine trends unpenalised on a simple uniform open grid. The minimiser is unique if $\ker B \cap \ker D = \{0\}$ for $\lambda > 0$. Check this condition before interpreting a smooth result. Penalising a field whose true behaviour has a sharp interface spreads the interface; inspect the response to a synthetic sharp feature before claiming a physical transition width.

Choose a strength from measurements, then report resolution

If a bound δ for the whitened data error is known, the discrepancy principle seeks a fit whose residual norm is comparable to δ. Smaller λ usually fits data more closely; larger values raise residual and reduce the penalised norm. A noise level estimated in the original physical units must be transformed consistently. If model error alone exceeds the assumed discrepancy, no amount of careful parameter tuning makes the noise model correct. Other choices include validation on genuinely independent observations and examining the residual-versus-penalty curve across several decades.

Report the chosen operator, units, strength and selection criterion. Alongside a reconstruction, show how its resolution changes under a modest change of regularisation. A feature that vanishes under a noise-consistent alternative requires more evidence. A predicted quantity can remain stable despite substantial uncertainty in the underlying image; testing that quantity directly can be the scientifically relevant validation.

Per Christian Hansen develops the singular-value picture and parameter choice further in Discrete Inverse Problems.

When a specified discrepancy determines a unique strength

Consider the zero-background quadratic penalty and express the data in left singular coordinates b_j. Substituting the filtered solution back into the forward model gives residual coefficient $\lambda^2 b_j/(\sigma_j^2+\lambda^2)$, with the sign chosen as data minus prediction. The data component $b_\perp$ orthogonal to the range of the matrix survives unchanged for every strength. Orthogonality therefore expresses the squared discrepancy as $\|b_\perp\|^2+\sum_j |b_j|^2[\lambda^2/(\sigma_j^2+\lambda^2)]^2$, with the sum taken over positive singular values. This equation allows a statement more definite than the heuristic that smoothing worsens the fit.

For each positive singular value the quantity in brackets increases strictly from zero to one as λ grows from zero to infinity. Thus the residual norm increases continuously from the irreducible distance $\|b_\perp\|$ to the full data norm. There is a unique positive finite strength for every target strictly between these endpoints provided at least one measured singular component is nonzero. A target below the first endpoint is infeasible. A target near the upper endpoint can permit a nearly zero reconstruction. The exact same reasoning for general penalties needs care when some physical modes receive no penalty.

Allocate a discrepancy after accounting for an unmodelled reading

A scaled instrument produces $b=(3,1)^T$ for a single unknown x, with predictions $Bx=(2x,0)^T$. The second channel, which should be zero according to the stated model, already consumes one unit of residual norm. Set a total target of $\sqrt{1.36}$ and use the zero-background criterion. Squaring the norm leaves $1.36-1=0.36$ for the squared discrepancy in channel one. The required nonnegative first-coordinate discrepancy is therefore 0.6. Its prediction is $3-0.6=2.4$, so the reconstructed coefficient must be 1.2.

Check by solving the actual penalised problem. The normal equation is

$$(4+\lambda^2)x=6.$$

Substituting $x=1.2$ gives $4+\lambda^2=5$, hence $\lambda=1$. The noiseless resolution factor for that coefficient is

$$4/(4+1)=0.8.$$

This calculation identifies the amplitude tradeoff associated with the selected discrepancy. It also shows why setting the target to 0.9 could never work: the second channel alone contributes one unit for every value of x.

Suppose a second independent reading of the missing channel also repeatedly lies near one, despite a nominal uncertainty of only 0.05 units. The persistent signed defect would be evidence to inspect the forward model or reference subtraction. Treating that behaviour as an excuse to enlarge the noise budget can produce an attainable discrepancy while systematically oversmoothing the channel that does measure x. Inspect residuals by physical observation as well as their aggregate norm before attributing a discrepancy-selected strength to the uncertainty of the instrument.

In this example the formula for arbitrary strength is $x = 6/(4 + \lambda^2)$. Differentiating it shows that near $\lambda = 1$ a strength increase of 0.1 initially lowers the coefficient by about 0.048. This local sensitivity is another interpretable number to publish if the noise target is itself estimated from a short calibration series.

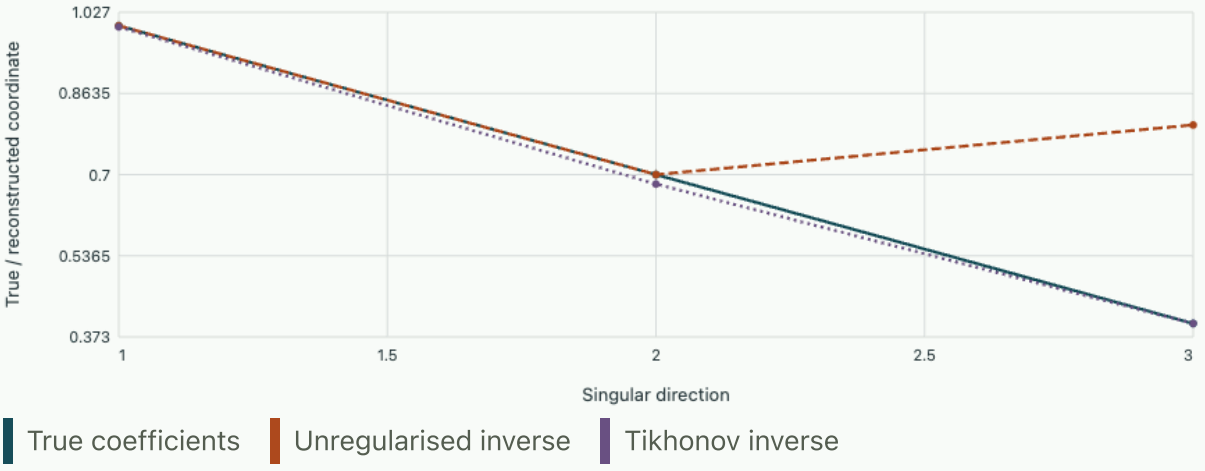
Change parameters · Regularisation, resolution and the price of stability

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

Weak singular value	**0.05**
Noise in weak data coordinate	**0.02**
Regularisation μ	**0.05**



Weak-mode noise is amplified by 20 before regularisation. The regularised weak-mode resolution gain is 0.5.

Adjoint gradients through a forward solver

Inverse problems and least squares

Differentiate a scalar objective with respect to many parameters without recomputing a separate perturbed solution for each one.

FROM FIRST PRINCIPLES

Differentiate a scalar objective through a linear state equation

Suppose the state u solves $K(p)u = f(p)$ for a given parameter vector, with nonsingular K, and a scalar objective $J(u, p)$ measures mismatch. A small parameter perturbation changes the state through both the load and the operator. Differentiating the balance gives

$$K\,du = df - (dK)u,$$
$$dJ = J_u^T du + J_p^T dp.$$

To obtain a scalar derivative without computing every state sensitivity, solve the transpose system $K^T\lambda = J_u$ once. Substitution eliminates du,

$$dJ = J_p^T dp + \lambda^T[df - (dK)u].$$

Each parameter then contributes through its derivatives of f, K, J. The transpose is required even if the forward solver is more familiar; it moves the scalar objective's weights backward through the linear response. A directional finite difference of the complete objective is an independent check on a computed adjoint gradient.

Background · Adjoint sensitivity

$$J^{\mathsf{T}} r$$

This operation maps a data-space vector back to parameter space. For a discretised physical model the chosen inner products determine the correct adjoint.

DERIVATION AND APPLICATIONS

Propagate a terminal measurement backward through continuous time

Consider a scalar state satisfying $\dot{u} = a(t)u + p\,b(t)$ over $0 \le t \le T$, with known initial state independent of p. A sensor observes $u(T)$, giving loss $J = \frac{1}{2}[u(T) - y]^2$. Assume continuous

coefficients and exact evolution for this calculation. The forward derivative $w = \partial u / \partial p$ satisfies an inhomogeneous initial-value problem, and the objective derivative depends on its endpoint,

$$\dot{w} = a(t)w + b(t),$$
$$w(0) = 0,$$
$$\frac{dJ}{dp} = [u(T) - y]w(T).$$

Instead of solving for every possible forcing sensitivity separately, introduce a terminal weight $\lambda(T) = u(T) - y$. Choose its evolution to cancel the term proportional to w in the derivative of λw. The product rule gives

$$\dot{\lambda} = -a(t)\lambda,$$
$$\frac{d}{dt}(\lambda w) = \lambda b(t).$$

Integrate over the interval. The initial product vanishes because $w(0) = 0$, and the terminal product equals the desired gradient. Thus

$$\frac{dJ}{dp} = \int_0^T \lambda(t) b(t)\, dt.$$

One backward weight suffices for many scalar parameters entering through different forcing shapes. A parameter-dependent initial condition would add $\lambda(0)\partial_p u(0)$. A discrete time integrator approximating the continuous equations also introduces integration error; differentiating the actual discrete steps gives the derivative of the computed loss at that finite step size.

WORKED EXAMPLE

Reconstruct a sensitivity in a decaying reservoir

For $a = -1,\ b = 1,\ T = \ln 2,\ u(0) = 0,\ p = 4$, the terminal state is

$$4(1 - e^{-T}) = 2.$$

With observation $y = 1.5$, the terminal adjoint is 0.5. Earlier weights are $\lambda(t) = 0.5 e^{-(T-t)}$, so

$$dJ/dp = 0.5(1 - e^{-T}) = 0.25.$$

Direct differentiation of $\frac{1}{2}(p/2 - 1.5)^2$ at four confirms the result.

SECOND APPLICATION

Compare two controls observed only at the endpoint

For $\dot{u} = -u$ between two instantaneous unit-gain input pulses, consider their amplitudes as independent parameters. A pulse at time t_i contributes $e^{-(T-t_i)}$ to terminal-state sensitivity. At $T = 3$ with terminal residual 2, the gradient for a pulse at $t = 1$ is $2e^{-2} = 0.271$; for a pulse at $t = 2$ it is

$$2e^{-1} = 0.736.$$

Later forcing has greater leverage because it has less time to decay.

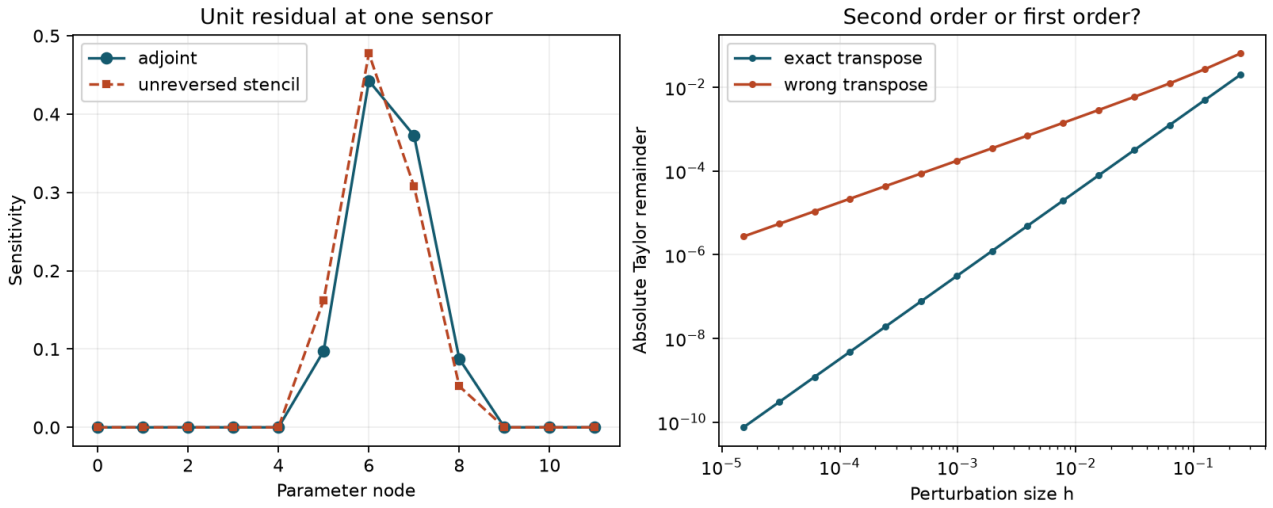
Does the gradient see where the instruments were actually placed?

A periodic ring of parameter nodes is mapped to fewer instrument readings in two operations. First a narrow, slightly asymmetric sensor spreads each nodal signal to its neighbours. Write this as a matrix S. Next the instruments sample between nodes. For an instrument a fraction f of the way from node j to $j + 1$, the corresponding row of B has entries $1 - f$ and f at those nodes. The complete forward model is $A = BS$. It is rarely necessary to assemble this product on a large mesh, but a small explicit example makes it possible to audit every entry.

Take Euclidean inner products so the conventions are clear. An adjoint must satisfy $(Av)^T w = v^T A^T w$ for every parameter direction v and sensor residual w. Moving each matrix across the scalar product yields

$$A^T w = S^T (B^T w).$$

The interpolation transpose scatters a sensor residual back to both neighbouring nodes with exactly the original weights. The smoothing transpose then exchanges the left and right coefficients of the stencil. Applying the forward smoothing again can look plausible and still shift the inferred sensitivity footprint.



Reusing a slightly asymmetric forward stencil produces a plausible-looking sensitivity. Its first-order remainder betrays the error when the objective is perturbed in an independent parameter direction.

To test the gradient choose the merit function $J(m) = \frac{1}{2}\|Am - y\|^2$. Its directional derivative is $v^T A^T (Am - y)$. Expand the quadratic objective exactly to find

$$J(m + hv) - J(m) - hv^T A^T (Am - y) = h^2 \|Av\|^2 / 2.$$

Halving h divides the remainder by four until rounding intervenes. Substituting a wrong derivative leaves an extra term linear in h. The notebook checks both the observed scaling and the coefficient of the quadratic term at one finite step. Several independent directions help avoid a chance cancellation in a single test.

With a fraction $f = 0.35$, one unit of observation residual is split into 0.65 and 0.35 before reverse smoothing. Its total remains one because the row weights sum to one, but conservation of that sum is a weak test: using the wrong asymmetric stencil also preserves it. If the underlying inversion uses volume-weighted mesh or measurement inner products, those weights must be included consistently; the plain transpose in this experiment applies to its explicitly chosen Euclidean metric.

The pictured footprint is obtained by sending a single residual backwards. It can also be reconstructed element by element by perturbing one parameter node forwards and recording just that sensor. The two pictures should agree for every node. This costs more forward evaluations, but on a small test ring it checks the location of a discrepancy directly.

Suppose the continuous forward code records a sensor at fraction 0.35 while its adjoint implementation rounds that sensor to the nearest node. Can an exact solution of the adjoint linear system repair the discrepancy?

▼ Work through the solution

The error occurs in the observation operator itself. For a unit residual the true transpose must assign 0.65 and 0.35, whereas the rounded version assigns one and zero. Accurate solves propagate these different

inputs accurately. An isolated dot-product test of the observation operator followed by a test of the full composition locates the discrepancy.

Change parameters · Explore this notebook example

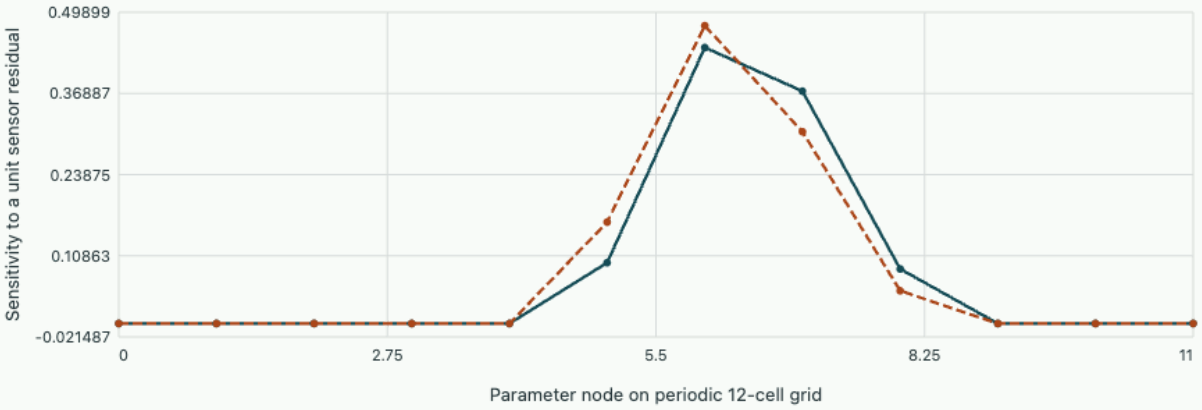
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Sensor interpolation fraction **0.35**

Left smoothing weight **0.25**



The sensitivity sums to 1. Asymmetric smoothing must reverse its direction in the adjoint; the mismatch vanishes only at symmetric weights.

Open editable notebook Download .ipynb

The state is also a function of the parameters

Suppose a discretised forward model solves $K(p)u = f(p)$. Data enter through Hu, and observations have been whitened. The objective is $J = \frac{1}{2}\|Hu - y\|^2 + R(p)$. A direct perturbation of a material parameter changes the predicted state even when that parameter never appears in the observation operator. For fixed H, the state derivative contribution is $g_u^{\mathsf{T}} du$, with $g_u = H^{\mathsf{T}}(Hu - y)$. Differentiating the state equation gives

$$K\,du = df - (dK)u.$$

Solving this equation once per parameter would be prohibitively expensive for a fine material field.

Transpose the solve

Define an adjoint state by $K^{\mathsf{T}}\lambda = g_u$. Then

$$g_u^{\mathsf{T}} du = \lambda^{\mathsf{T}} K\,du = \lambda^{\mathsf{T}}[df - (dK)u].$$

For each parameter component the gradient is

$$\frac{\partial J}{\partial p_j} = \frac{\partial R}{\partial p_j} + \lambda^{\mathsf{T}}\left(\frac{\partial f}{\partial p_j} - \frac{\partial K}{\partial p_j}u\right).$$

This uses one forward and one transposed solve for one scalar objective, independent of the number of gradient components. Forming the contractions for all components still costs work, often local to mesh elements. If sensors or boundary conditions also depend explicitly on parameters, their derivative terms must be included. A symmetric diffusion operator may reuse a factorisation; an advection operator need not equal its transpose.

Work a two-cell example

Consider $K = \begin{pmatrix} p+1 & -1 \\ -1 & 2 \end{pmatrix}$, $f = (1, 0)$, and observe the second component with value $y = 0.4$. At $p = 1$, $u = (2/3, 1/3)$, so the residual is $-1/15$. The adjoint load is $(0, -1/15)$; solving gives $\lambda = (-1/45, -2/45)$. Only the first diagonal entry of K depends on p. With no regulariser the derivative is $-\lambda_1 u_1 = 2/135$. Check independently: $u_2 = 1/(2p+1)$, whose derivative at one is $-2/9$. Multiplying by the residual gives the same positive gradient. Increasing p lowers an already underpredicted second reading.

Test the computed gradient in a direction

Pick a physically scaled perturbation v. A correct gradient gives a linear residual $J(p + hv) - J(p) - h\,g^{\mathsf{T}}v$ of order h^2 for a smooth problem. Halving h should roughly quarter this remainder across a useful range. At very small h, forward-solver tolerances and floating-point cancellation can dominate. Repeat with directions touching different components and with the same boundary-condition handling used in production. A close fit to the observed data is a poor gradient test when residuals nearly vanish: a wrong data gradient can then look small by accident. Evaluate at a parameter away from the best fit, solve the state again for each perturbation and compare the resulting objective values. To detect a missing transpose, include a nonsymmetric forward operator in the test. The

symmetric two-cell example above can expose a sign error but cannot distinguish the forward matrix from its transpose.

Reverse a sequence of state updates

Suppose the forward solver evolves $u_{n+1} = G_n(u_n, p)$ for a fixed number of steps, and an objective sums observation losses $\ell_n(u_n, p)$ from time zero to the final time. The direct chain rule carries a state derivative with respect to every parameter through the entire trajectory. To reverse the calculation, start with $\lambda_N = \partial_{u_N} \ell_N$ as a column vector. A perturbation du_n affects both the loss at time n and the following state by $G_{n,u} du_n$. Moving the latter action across a dot product gives the recursion $\lambda_n = \partial_{u_n} \ell_n + G_{n,u}^T \lambda_{n+1}$.

The parameter gradient is the sum of explicit derivatives of all losses, the initial term $(\partial_p u_0)^T \lambda_0$, and the contributions $G_{n,p}^T \lambda_{n+1}$ from every step. Each reversed step needs the state information used by its forward counterpart. Storing the whole trajectory makes that information available immediately. Recomputing segments between stored checkpoints trades extra forward evaluations for memory. These formulas differentiate the actual discrete transitions, so an adaptive solver with parameter-dependent step rejection or event switches needs a careful interpretation of the differentiated computation.

Make the transpose observable in a two-step calculation

Take a two-coordinate forward transition $G(u, p) = Mu + p e_2$, with $M = \left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$, $u_0 = 0$, and $p = 1$. Two steps give

$$u_1 = (0, 1)^T,$$

$$u_2 = (2, 2)^T.$$

Observe the first coordinate only at the end, with target value one. The quadratic loss is one half and its final state derivative is

$$\lambda_2 = (1, 0)^T.$$

One reverse step gives

$$\lambda_1 = M^T \lambda_2 = (1, 2)^T.$$

The two parameter contributions are $e_2^T \lambda_2 = 0$ and $e_2^T \lambda_1 = 2$, so the gradient is two.

The zero contribution from the last injection has a physical meaning: the injected second coordinate can influence the first only during a subsequent transition. The first injection has time to do that through the off-diagonal element two. Using M in place of M^T in the backwards calculation would instead keep $\lambda_1 = (1, 0)^T$, predicting zero total sensitivity. This example exercises a distinction that disappears when every state matrix is symmetric.

For a scalar parameter perturbation h, solving both steps again gives terminal first coordinate $2(1+h)$ and loss $J(1+h) = (1+2h)^2/2$. The remainder after subtracting the base loss and the predicted linear term is exactly $2h^2$. For $h = 0.1$ and 0.05 the remainders are 0.02 and 0.005. The erroneous zero gradient instead gives 0.22 and 0.105. The two sequences have very different observed convergence factors even though they use identical forward trajectories and observation data.

More generally an individual parameter may enter the map at several times with different signs. The corresponding gradient terms must be accumulated rather than overwritten during the backward pass. In this example overwriting after the earliest step happens to return the correct value because the last contribution is zero. Adding a simultaneous observation of the final second coordinate makes both contributions nonzero and provides a complementary regression case for the time-index accumulation.

For a terminal loss summing both coordinate discrepancies, the adjoint load contains both sensor contributions before the first reverse transition.

Change parameters · Adjoint gradients through a forward solver

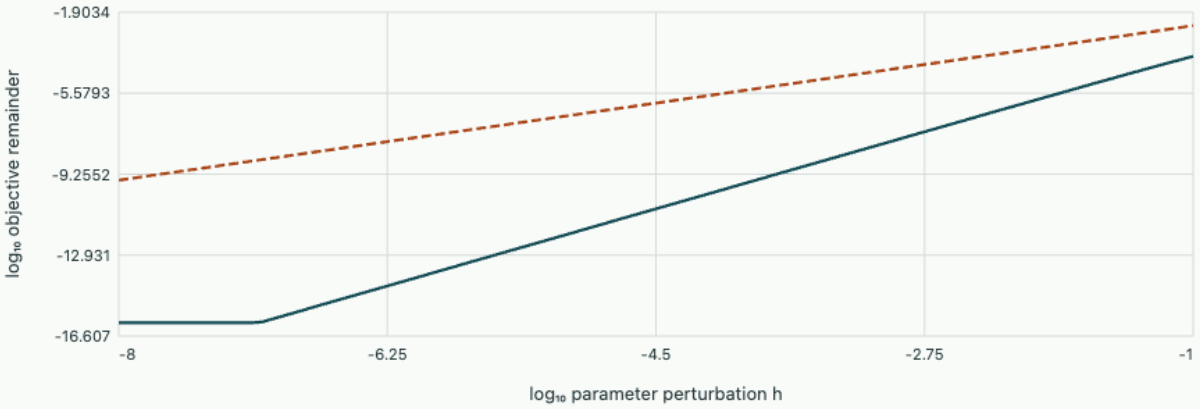
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Material parameter p **1**

Observed second state **0.4**

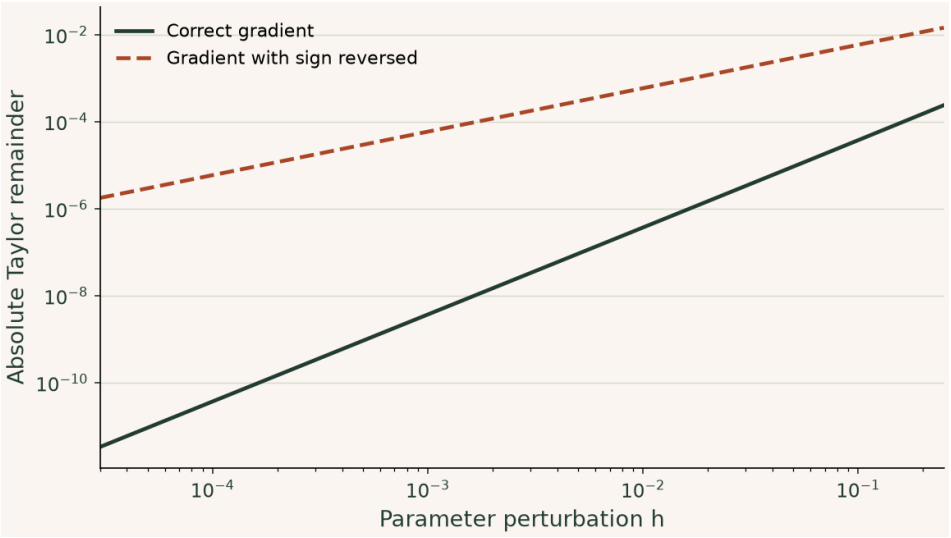


Subtract correct adjoint gradient Subtract wrong-sign gradient

Analytic and adjoint derivative = 0.014815. A correct gradient leaves a quadratic remainder until cancellation dominates; at an exact fit this sign test loses sensitivity.

Recognise an incorrect gradient by the rate of its remainder

Verify the reduced gradient for $J(p) = (1/(p+2) - 0.6)^2/2$ at $p = 1$. For each perturbation solve the scalar equilibrium again and subtract the original objective and the candidate linear term. The sign of the state sensitivity is essential.



In the range shown the Taylor remainder falls quadratically with the correct gradient and approximately linearly with the reversed one. The use of an analytic scalar model isolates this diagnostic; tolerances in a larger forward solve can mask the small-h regime.

▶ **Read values from the plot**

Choosing regularisation when the true answer is hidden

Inverse problems and least squares

Derive two parameter choices from different information, then see how an excellent fit to sensor data can restore the noise in a weakly visible mode.

Each observation competes with a penalty

In orthogonal singular coordinates, let $d_i = \sigma_i x_i + \eta_i$. For a candidate $\lambda > 0$, minimise $\sum_i (\sigma_i x_i - d_i)^2 + \lambda^2 \sum_i x_i^2$. Setting each derivative to zero gives

$$\widehat{x}_i = \sigma_i d_i / (\sigma_i^2 + \lambda^2).$$

Predicted data are $H_i d_i$, where $H_i = \sigma_i^2/(\sigma_i^2 + \lambda^2)$. These factors range from zero to one. Their sum is the effective number of fitted data directions.

If an independently obtained bound for the whitened noise norm is δ, compare $\sum_i [(1 - H_i) d_i]^2$ with δ^2. The residual rises as λ increases. A discrepancy choice seeks the largest smoothing consistent with the stated noise level. If there is no crossing, either the search range, model or noise estimate needs examination.

A prediction criterion without a noise bound

For a linear smoother $\widehat{d} = H d$, deleting observation i gives the leave-one-out residual $r_i/(1 - H_{ii})$ under the usual fixed-design ridge fit. Generalised cross-validation replaces the individual leverages by their mean. This produces the computable quantity

$$\mathrm{GCV}(\lambda) = \frac{\|d - \widehat{d}\|^2 / n}{[1 - \operatorname{trace}(H)/n]^2}.$$

The denominator penalises fits that spend nearly one degree of freedom per datum. The criterion estimates predictive performance under assumptions on noise and design. A selected value need not minimise the error in the unknown parameter, and highly unequal leverages can defeat the approximation.

A sensor that barely sees a mode

Take sensitivities $1, 0.01$, true parameters $(1, 1)$ and data $(1, 0.02)$. Unregularised inversion returns $(1, 2)$ with zero residual. A perturbation of one hundredth in the second observation doubled the second parameter. With $\lambda = 0.01$ the second gain is $\sigma/(\sigma^2 + \lambda^2) = 50$, so the same datum yields 1. This exact agreement with the hidden truth is accidental.

For $\lambda = 0.1$, the second estimate drops to about 0.0198. A weak direction has been almost removed. Meanwhile the first estimate remains about 0.9901. A single residual norm does not tell you which parameter direction was suppressed.

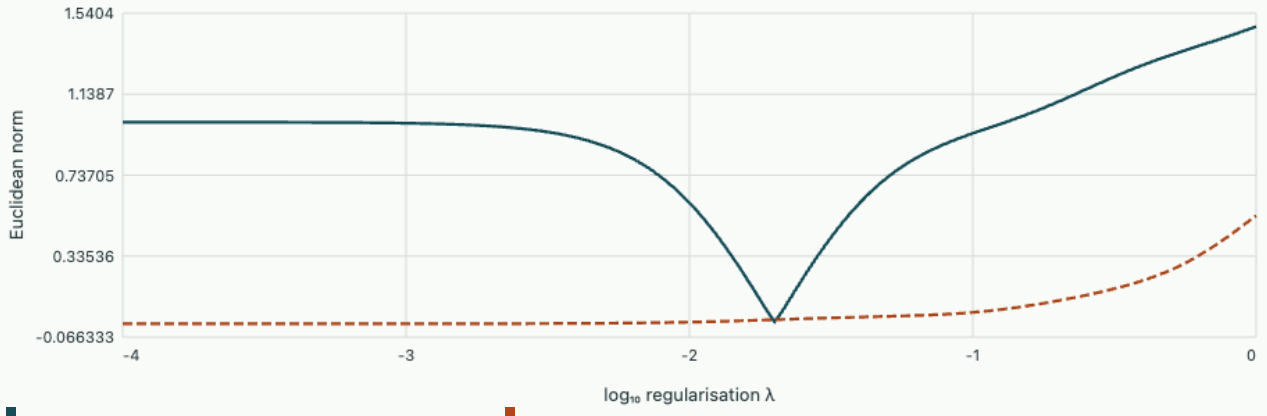
Inspect an entire regularisation path

Three synthetic sensors have sensitivities $1, 0.2, 0.02$, true parameters all equal to one, and an adjustable perturbation only in the last sensor. This deliberately uneven noise makes it possible to test when the GCV assumptions are plausible. Inspect parameter error and data residual over a logarithmic range of λ. The result reports the best GCV value on the sampled grid and the residual at that choice.

Open this experiment online

Try the calculation

Error in weakest observation **0.02**



Grid minimum GCV 0.0046083 at λ = 0.035481, giving residual 0.030988. Actual data noise norm 0.02.

Tutorial on choosing a criterion

If all H_i equal a common $h < 1$, can GCV distinguish different values of h for fixed data?

▼ **Worked solution**

The numerator contains $(1-h)^2\|d\|^2/n$ and the denominator $(1-h)^2$. They cancel exactly. There is no identifiable minimum from this score.

If residuals remain much larger than a well-calibrated noise norm even at negligible regularisation, should increasing the penalty help?

▼ **Worked solution**

Increasing the penalty shrinks the fit and usually increases residual. Check the forward model, its range and data preparation, including the uncertainty units used in whitening.

Read the university treatment

Compare the information each stopping rule needs about measurement noise with the information available in the three-sensor experiment. Per Christian Hansen · Stopping rules for noisy data

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

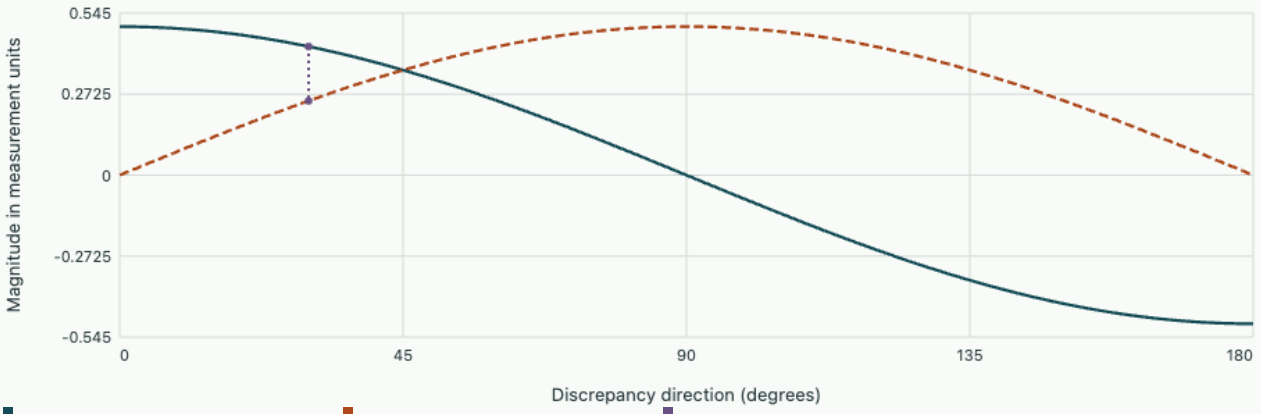
In a two-reading instrument, the model sensitivity is the unit horizontal vector. Vary the magnitude and direction of the omitted contribution. Observe parameter bias and visible residual separately, and locate cases with the same discrepancy norm but completely different fit diagnoses.

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

Capacitance matrices, floating nodes and controlled approximation bounds

Electrostatic field theory and measurement

A multi-electrode device generally has no single capacitance until the other terminals have been connected or isolated. An energy identity determines the matrix that relates voltages to charges, and a complementary flux identity can certify estimates even when the exact field is unknown.

Build the matrix from unit-voltage experiments

Consider n isolated conductor surfaces embedded in a connected linear dielectric, plus a grounded reference conductor or decay relative to a reference at infinity. There is no free charge in the dielectric. Define the weighting function $w_j(\mathbf{x})$ by putting conductor j at 1 V, all other conductors and the reference at zero, then dividing by 1 V. For arbitrary voltages V_j, linearity and uniqueness give

$$\phi(\mathbf{x}) = \sum_j V_j w_j(\mathbf{x}).$$

The charge on conductor i is an integral of displacement leaving its metal into the dielectric. Using a normal $\mathbf{n}$ pointing out of the dielectric into the metal gives

$$Q_i = \int_{S_i} \epsilon \partial_n \phi \, dA.$$

Therefore

$$Q_i = \sum_j C_{ij} V_j,$$

$$C_{ij} = \int_{S_i} \epsilon \, \partial_n w_j \, dA = \int_\Omega \nabla w_i \cdot \epsilon \nabla w_j \, dV.$$

The last equality follows by multiplying the charge-free field equation for w_j by w_i and integrating by parts. The boundary integral selects just conductor i because its weighting function is unity there and zero on every other conductor. The volume expression proves reciprocity $C_{ij} = C_{ji}$ for a symmetric dielectric tensor. The diagonal terms are positive, since they integrate the squared fields of each independent one-volt experiment. In an ordinary connected isotropic configuration a positive voltage on j with $i \neq j$ grounded puts negative induced charge on i; the corresponding mutual coefficient C_{ij} is nonpositive.

The stored energy is

$$U = \tfrac{1}{2} \sum_{ij} V_i C_{ij} V_j.$$

Positive material response makes the reduced matrix with an effective reference positive definite under usual connected boundary conditions. If all conductors including the external closing conductor are retained without fixing a reference, changing all voltages by a common constant produces no field. The

full matrix then has a zero common-voltage mode and each row sums to zero. Removing the grounded reference from the matrix removes this gauge null direction; its remaining rows need not sum to zero. Treating both statements as universal would create a spurious constraint on measured terminal data.

Convert a field matrix to a circuit measurement

A two-port device might have reduced matrix in picofarads $C = \begin{pmatrix} 12 & -3 \\ -3 & 8 \end{pmatrix}$ relative to its grounded enclosure. With conductor 1 at 2 V and conductor 2 at 0 V, charges are 24 pC and -6 pC. The grounded enclosure takes the compensating -18 pC. Reciprocity predicts that holding conductor 1 at 0 V and conductor 2 at 2 V puts -6 pC on conductor 1. The diagonal 12 pF includes both the coupling to conductor 2 and the path to the enclosure. In an equivalent lumped model the positive pair capacitance between 1 and 2 is 3 pF, while the two separate couplings to ground are 9 and 5 pF. The sign of a Maxwell matrix off-diagonal is opposite to a positive capacitor drawn between two nodes.

Now disconnect conductor 2 after neutralising it. Its total charge stays zero while its voltage becomes an unknown. The second matrix row demands $-3V_1 + 8V_2 = 0$. Thus $V_2 = 3V_1/8$, giving apparent capacitance seen at terminal 1 equal to $12 - 9/8 = 10.875$ pF. Grounding conductor 2 would instead yield 12 pF. The change is a measurable effect of the prepared state, not a change in the dielectric constant of the material. If conductor 2 already carries charge Q_2^0, its voltage acquires a constant offset Q_2^0/C_{22} in addition to the response to V_1.

For several floating nodes write their block equation as $Q_f = C_{fa}V_a + C_{ff}V_f$. If the reference fixes the potential freedom, solve $V_f = C_{ff}^{-1}(Q_f - C_{fa}V_a)$ and substitute into the active-node block. The differential active capacitance is the Schur complement $C_{aa} - C_{af}C_{ff}^{-1}C_{fa}$. It is symmetric and positive for a positive full reduced matrix. With fixed floating charges the constant term shifts active charges but leaves this differential matrix unchanged in a linear medium. An actual cable measurement that results from leaving a neighbouring shield unconnected should be compared to this elimination, not to the original diagonal.

Obtain an upper capacitance bound by trying a voltage

For a two-terminal charge-free dielectric at prescribed voltage difference V, every admissible continuous trial potential satisfying the electrode voltages gives an energy at least as large as the exact electrostatic energy. This follows by writing the trial as exact plus a perturbation vanishing on fixed-voltage electrodes and expanding the energy. The mixed term integrates to zero because the exact displacement is divergence free. If exterior walls are insulated, the exact normal displacement vanishes there and the trial need not be assigned a fixed potential on them. Thus $C \le 2U_{\mathrm{trial}}/V^2$. The fields must genuinely arise as a gradient of one trial voltage. Guessing unrelated components of an electric field does not provide this upper bound.

For a capacitor of plate area A, separation d, and scalar permittivity varying in the gap, choose a linear voltage along the plate normal. The trial energy gives

$$C/A \le \langle \epsilon \rangle / d,$$

where the average is over the gap volume divided by Ad. For dielectric columns that run parallel to the applied field with no change of material along the field direction, the linear potential can be exact. For layers stacked normal to the field it usually overestimates capacitance. The same arithmetic average thus serves as the actual result or a strict upper bound depending on geometry, even when the material fractions are identical.

A complementary trial with a divergence-free displacement

Fix total plate charges $\pm Q$ instead, taking no volume free charge. An admissible displacement trial must have the prescribed flux into each electrode, no flux through insulating side walls, continuous normal component across free-charge-free interfaces, and zero divergence inside. It need not correspond to an exact potential-gradient electric field. Consider $U_D[\mathbf{D}] = \frac{1}{2}\int \mathbf{D}\cdot\epsilon^{-1}\mathbf{D}\,dV$. Write a trial as $\mathbf{D}_* + \delta\mathbf{D}$. The cross term is $\int(-\nabla\phi_*)\cdot\delta\mathbf{D}\,dV = 0$: integration by parts gives zero bulk divergence and each conductor has constant voltage multiplying zero net flux change. Thus $U_D \geq Q^2/(2C)$ and

$$\frac{Q^2}{2U_D^{\mathrm{trial}}} \leq C \leq \frac{2U_V^{\mathrm{trial}}}{V^2}.$$

Unlike two competing estimates, these inequalities identify which side contains the exact answer. For the same parallel-plate geometry choose constant displacement $D_z = Q/A$. It yields $C/A \geq 1/[d\langle 1/\epsilon\rangle]$, the harmonic-mean lower bound. For layers stacked normal to the plates that trial is exact. For columns parallel to the field the upper bound is exact. An arbitrarily shaped two-phase composite satisfies both under the stated plate and insulating-side conditions. This gives a strong physical check before solving detailed microstructure on a mesh.

Use measured numbers to rule out a material model

Take equal volume fractions of relative permittivities 2 and 8 in a cell of area 100 mm² and thickness 1.0 mm. The harmonic mean is 3.2, arithmetic mean 5.0. Since $\epsilon_0 A/d = 0.8854$ pF, the admissible capacitance interval is 2.83–4.43 pF. A measured 3.3 pF is consistent with many internal geometries and does not determine how the two materials are arranged. A calibrated value of 5.2 pF would require examining fringing, area and thickness, parallel cable capacitance, dispersive or conducting response, and the assumed constituent data. Merely refining an interior mesh cannot produce a static value outside the proven interval if it faithfully solves the same ideal model.

For a field computation use the matrix reciprocity relation as one diagnostic and the two variational bounds as another. An electrostatic finite-element potential may produce a reliable energy estimate while numerical differentiation at a sharp electrode edge produces poor charge density. Integrating boundary fluxes gives each terminal charge, but reconstructing C_{ij} by the mixed energy integral often has a different error sensitivity. Comparing both over a grid sequence can indicate whether the error is concentrated at edges or in the field volume. Symmetrising a noisy measured matrix by averaging its transpose removes an antisymmetric discrepancy but cannot establish positivity or explain an inaccurate common scale.

Construct a complete measurable matrix without numerical differentiation

Three thin spherical conductors have radii $a < b < c$. The outer one is tied to the reference potential. Homogeneous material of permittivity ϵ_1 fills the inner annulus $a < r < b$ and another material ϵ_2 fills $b < r < c$. Spherical symmetry gives each annulus its own independent radial displacement field fixed by its electrode voltage difference. Define the ordinary two-sphere gap capacitances $K_1 = 4\pi\epsilon_1 ab/(b-a)$,

$$K_2 = 4\pi\epsilon_2 bc/(c-b).$$

The inner conductor takes $K_1(V_1 - V_2)$. The middle one has two faces, carrying the opposite inner gap charge and an outer gap charge $K_2 V_2$. Thus $Q_1 = K_1 V_1 - K_1 V_2$ and $Q_2 = -K_1 V_1 + (K_1 + K_2)V_2$. Every sign and every term can be read from a physical face.

In vacuum choose $a = 10$ mm, $b = 20$ mm, $c = 40$ mm. Then $K_1 = 2.225$ pF and $K_2 = 4.451$ pF. Driven at 3.0 V with the middle conductor grounded, the inner sphere carries 6.68 pC and the outer gap has zero field. With the middle conductor floating and neutral, $V_2 = K_1 V_1/(K_1 + K_2) = 1.0$ V. The inner sphere now carries 4.45 pC. The inner and outer faces of the floating sphere carry exactly opposite charges, although the conductor has a nonzero voltage. Integrating a current meter attached to the conductor during charging need not tell you each face's charge separately.

Now electrically connect the inner two conductors. Their common voltage sets the inner field to zero. The combined conductor has capacitance K_2 to the exterior, independent of a. This third preparation is distinct from grounding or floating the middle shell. Before calculating an experimental 'device capacitance', draw the electrical connections as well as the geometric cross section. In a larger matrix a driven conductor and a driven guard at the same voltage should have their charges combined only if the experimental ammeter actually measures their total current.

One can also recover the two constituent permittivities from just two carefully selected scalar measurements. Ground the middle shell to determine K_1 directly. Then float it neutral and measure $K_{\mathrm{ser}} = K_1 K_2/(K_1 + K_2)$, giving $K_2 = K_{\mathrm{ser}} K_1/(K_1 - K_{\mathrm{ser}})$. If K_2 is very large, the floating and grounded readings nearly coincide and their difference becomes noise sensitive. An alternative is to drive inner and middle together and measure the outer-gap current directly. The algebra selects a better experimental protocol instead of hiding ill-conditioning in a formula.

Tighten a variational bound without solving the entire field

Take the stacked two-layer capacitor with no space charge, equal layer thickness and isotropic permittivities $\epsilon_1 = 2\epsilon_0, \epsilon_2 = 8\epsilon_0$. A single linear trial produces the arithmetic-mean upper bound. Insert one interior voltage degree of freedom at a point that coincides with the material interface. The resulting piecewise-linear trial space contains the exact solution. Minimising its energy makes the normal displacements equal, reproducing the harmonic-mean capacitance exactly. A global quadratic trial with the same number of extra parameters will usually perform worse because its derivative is continuous while the exact normal field jumps.

To see this quantitatively write $x = z/d$ on the unit interval and $V(x) = V_0[x + \eta x(1-x)]$, with the lower half at relative permittivity 2 and upper half at 8. This trial fixes both endpoints. Its dimensionless capacitance upper estimate is

$$c(\eta) = \int_0^1 \epsilon_r(x)[1 + \eta(1-2x)]^2 dx.$$

Integrating gives

$$c(\eta) = 5 - 3\eta + (5/3)\eta^2.$$

The best quadratic trial has $\eta = 0.9$ and $c = 3.65$, compared with 5.00 for the straight line and 3.20 for the exact layered solution. The unrefined constant-displacement lower trial already supplies 3.20, so the two-field comparison certifies an interval before any field-map visualisation.

Why did the smooth trial still overestimate by 14%? At its endpoints it predicts slopes 1.9 and 0.1 times the mean field magnitude, whereas the exact field has slopes 1.6 and 0.4 in the two halves. Optimising a volume energy distributes a global smooth approximation across both layers. It has no mechanism to jump at the known interface, so an error persists even with its one coefficient optimised perfectly. The example explains why fitting material interfaces in a mesh, or using basis functions that admit correct normal derivative jumps, can matter more than merely increasing polynomial degree.

In an irregular two-dimensional composite an exactly divergence-free complementary trial can be constructed from a stream function Ψ: take $\mathbf{D} = (\partial_y \Psi, -\partial_x \Psi)$ in Cartesian coordinates. Its divergence vanishes identically. Total flux through an oriented path is a difference of Ψ at its endpoints, so carefully chosen boundary values enforce required electrode charges and zero leakage on insulating sides. Optimising $\int \mathbf{D} \cdot \epsilon^{-1} \mathbf{D}/2$ over such trial functions tightens the lower capacitance bound. Optimising the independent voltage trial tightens the upper one. The difference between their resulting bounds quantifies what the chosen approximation spaces still fail to resolve.

What an AC bridge measures when a nominally floating terminal leaks

A real high-impedance lead is neither perfectly floating nor an ideal ground at every frequency. Add a leakage resistance R from conductor 2 of the two-port example to its grounded enclosure. Keep the reduced static capacitance matrix C fixed, linear and quasistatic at the frequencies of interest. Let terminal 1 be driven sinusoidally and terminal 2 have no other external drive. Using $e^{i\omega t}$ here for circuit phasors, charge continuity at terminal 2 gives

$$i\omega(C_{21}V_1 + C_{22}V_2) + V_2/R = 0.$$

The outward conductive current plus the rate of charge stored in the electrode must vanish. Solving for the active terminal admittance gives

$$Y_{\mathrm{in}} = i\omega C_{11} + \frac{\omega^2 C_{12}C_{21}}{1/R + i\omega C_{22}}.$$

At low frequency the leakage holds terminal 2 close to ground, so the leading measured capacitance is C_{11}. At high frequency the floating condition is recovered and $Y_{\mathrm{in}}/(i\omega) \to C_{11} - C_{12}C_{21}/C_{22}$. The effective admittance also has a positive real part, since the driven voltage on terminal 2 supplies Joule heat in the leakage resistance. This term persists even if the dielectric itself is lossless. An AC capacitance meter reporting a frequency-dependent loss tangent might therefore be detecting a boundary preparation rather than intrinsic material loss.

Take $C_{22} = 8$ pF and $R = 100$ MΩ. The crossover $1/(2\pi R C_{22})$ is 199 Hz. At this frequency the quadrature capacitance is halfway between 12.0 and 10.875 pF, namely 11.4375 pF. The corresponding loss conductance divided by ω is 0.5625 pF. An independent verification uses the port voltage ratio at crossover,

$$|V_2/V_1| = |C_{21}|/(\sqrt{2}C_{22}) = 0.265.$$

Compute dissipated power as $|V_2|^2/(2R)$ for peak phasor amplitudes and compare it with $\mathrm{Re}\, Y_{\mathrm{in}} |V_1|^2/2$; the two results must coincide.

The crossover shifts inversely with the leakage resistance while its two limiting capacitances stay fixed in this simple model. Measuring several shunt resistances or adding a known external one provides a direct way to test this interpretation. If a neighbouring shield is connected to a voltage follower instead, its response includes the follower's own gain and phase; static floating-node elimination is insufficient. Extending the field-derived capacitance matrix by experimentally known terminal circuits makes it possible to distinguish changes in electrostatic geometry from changes in what the surrounding measurement wiring enforces.

Close the charge balance at a floating conductor

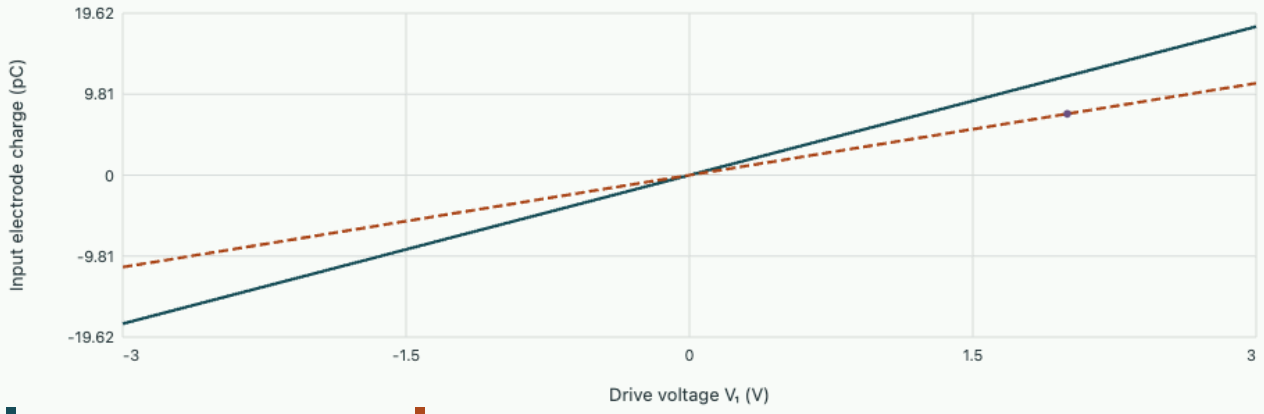
Compare voltages obtained by fixing a conductor potential and by prescribing its net charge. Check the sign of the off-diagonal capacitance terms.

Open this experiment online

Try the calculation

Mutual capacitance (pF) 4

Drive voltage V_1 (V) 2



Second electrode grounded Second electrode neutral and floating

Selected floating charge on electrode 1

Neutral electrode 2 floats to 1.143 V. Grounded input capacitance 6 pF, floating input capacitance 3.714 pF. Floating-case field energy 7.429 pJ.

Tutorial · distinguish a floating port and an inadmissible trial

What changes after disconnecting a biased guard?

Using the 12, 8 and −3 pF coefficients, start with terminal 1 at 0 V and terminal 2 at 1 V. Disconnect terminal 2, then raise terminal 1 to 2 V. Find the new floating voltage, both charges and the change of active charge.

▼ **Work through the solution**

Initially $Q_2 = 8$ pC, which stays fixed. The new condition is $8 = -3(2) + 8V_2$, giving $V_2 = 1.75$ V. Terminal 1 initially held -3 pC; finally it holds $12(2) - 3(1.75) = 18.75$ pC. Its change is 21.75 pC, or 10.875 pF times the 2 V increment. This reproduces the neutral-floating differential result but a different absolute charge. A charge amplifier reset at the initial state observes the change, whereas a device measuring absolute node potential distinguishes the stored preparation.

Can a lower voltage-trial estimate be trusted?

A potential approximation for the two-phase cell returns 2.70 pF while a constant-displacement trial gives 2.83 pF. The implementer reports the smaller number because its potential looked smoother. Diagnose the inconsistent pair.

▼ **Work through the solution**

For admissible fields on the same problem the flux trial gives a lower capacitance bound and the potential trial an upper one. The reversed interval cannot result from arbitrary coarseness alone. Check that the voltage trial satisfies the exact terminal difference and that all physical material volumes and energy factors are included. Check that the displacement trial has the correct area-normalised total charge and satisfies flux continuity. Finally examine numerical quadrature and unit conversion. Comparing the plots of the two fields does not fix an inadmissible constraint. If both are corrected, their gap gives an actual a posteriori uncertainty interval even without an analytic solution.

Further reading

- Measure a variational energy gap
- Change the geometry and permittivity of a layered capacitor
- Griffiths, 5th ed., §§2.4–2.5 and 4.4 · capacitance and field energy

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

$$y_j = \sum_m H_m f_m e^{ik_m x_j} + e_j,$$

$$H_m = e^{-|k_m|h}\,\operatorname{sinc}(k_m w/2),$$

$$\operatorname{sinc} t = \frac{\sin t}{t}.$$

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

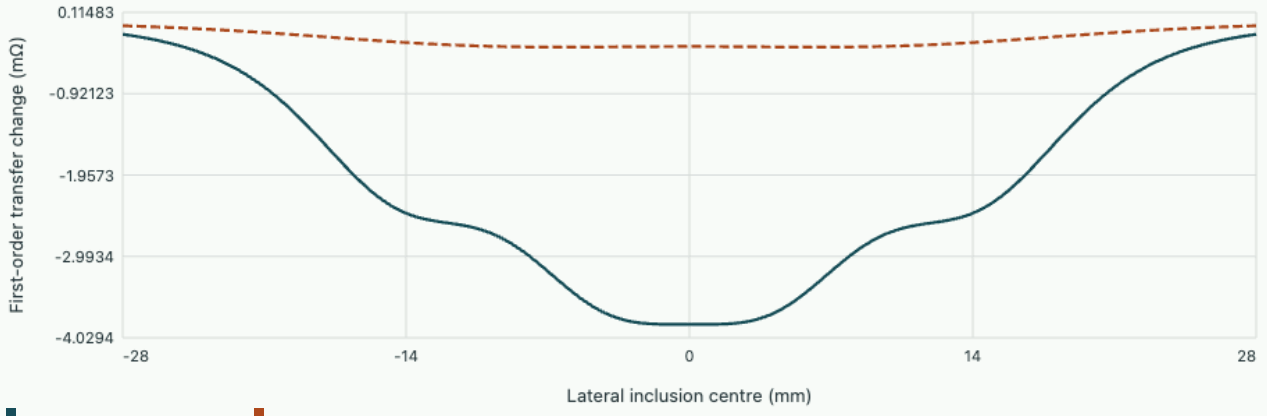
Compare two contact configurations and their sensitivity to a local material change. An accurate potential at the electrodes does not necessarily resolve the interior.

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

Which part of a sample changes a four-electrode impedance measurement?

Electromechanics, optics and measurement

Driving one electrode pair and measuring another yields a transfer impedance whose spatial sensitivity can be positive, negative or nearly zero. A second field solution supplies the sensitivity without perturbing every element of a numerical mesh.

Distinguish the current terminals from the voltmeter terminals

Consider a bounded two-phase electrical sample in a benign laboratory characterisation cell. Drive harmonic current $+I$ through electrode a and collect $-I$ through b. Two independently selected electrodes c, d read the difference $U_c - U_d$ with negligible voltmeter current. The measured transfer impedance is

$$Z_{cd,ab} = (U_c - U_d)/I.$$

A four-terminal measurement reduces some lead-voltage errors, but finite-area electrodes still alter the field and contact processes can affect it. Reversing either pair reverses the signed transfer reading. Interchanging both pairs leaves it unchanged in a reciprocal linear sample.

Adopt phasors $e^{i\omega t}$ and a local complex admittivity $\gamma = \sigma + i\omega\epsilon$, with real positive scalar σ, ϵ for this first model. The total current density is $\mathbf{J} = -\gamma\nabla u$. In the electroquasistatic regime $\nabla \cdot (\gamma\nabla u) = 0$ away from injected current at the boundary. Magnetic induction and propagation are negligible at the chosen cell dimensions and frequencies; checking electrode and cable dynamics remains necessary even when the sample itself is quasistatic. Perfectly conducting measurement electrodes are equipotentials. Inactive electrodes have zero *net* current, which allows charge redistribution over their finite surfaces; a strictly insulating uncovered wall instead requires zero normal current locally.

Use an outward normal $\mathbf{n}$ from the sample. Define positive terminal current into the sample by $I_i = \int_{S_i} \gamma \partial_n u \, dA$; it equals minus the outward conduction-plus-displacement flux. The pair currents sum to zero, and only voltage differences matter, so fix one reference or a mean voltage. Integrate by parts against any test voltage v constant on each ideal electrode, giving

$$\int_\Omega \gamma \, \nabla u \cdot \nabla v \, dV = \sum_i I_i \, v_i.$$

This is a complex bilinear identity for reciprocity and linear sensitivity. Real average heating is obtained separately by conjugating the physical fields when computing power. Confusing those two products can give an incorrect phase or an incorrect sign for an imaginary admittivity perturbation.

Perturb the sample while keeping injected currents fixed

Let ψ_{ab} be the potential per unit current driven from a to b, and ψ_{cd} the solution when current is instead driven between the measurement electrodes. Using each as test function for the other shows directly

that $Z_{cd,ab} = Z_{ab,cd}$. Now change material admittivity by a small distribution $\delta\gamma$ and let the drive field change by $\delta\psi_{ab}$. Holding all currents and electrode geometry fixed, differentiate the weak equation:

$$\int_\Omega \gamma \nabla \delta\psi_{ab} \cdot \nabla v \, dV = -\int_\Omega \delta\gamma \, \nabla\psi_{ab} \cdot \nabla v \, dV.$$

Choose $v = \psi_{cd}$ and apply the unperturbed measurement-pair equation to $\delta\psi_{ab}$. Its left side becomes the change in voltage difference per drive current, hence

$$\delta Z_{cd,ab} = -\int_\Omega \delta\gamma(\mathbf{x}) \, \nabla\psi_{ab}(\mathbf{x}) \cdot \nabla\psi_{cd}(\mathbf{x}) \, dV.$$

The minus sign has an immediate check in a conducting bar: increasing conductivity lowers its resistance at fixed current. The sensitivity density is the overlap of the drive and reciprocal fields, not simply the magnitude of the drive field. If the fields are locally perpendicular a small isotropic admittivity perturbation is invisible to first order in this channel. If the fields are opposed the local real sensitivity can have the opposite sign to nearby regions. For tensor perturbations replace the product by $-\nabla\psi_{cd} \cdot \delta\boldsymbol{\gamma} \nabla\psi_{ab}$. Reciprocity in the above form requires symmetric material tensors.

A one-dimensional experiment tests the sign, normalisation and scale

Take a uniform real-conductivity bar of length 20 mm, cross-section $A = 100$ mm² and conductivity 0.20 S/m. Assume end electrodes cover its faces and ideal voltage taps select the full length in this limiting four-wire experiment. The impedance is

$$L/(\sigma A) = 1000\ \Omega.$$

The drive and reciprocal gradients per ampere have magnitude $1/(\sigma A)$. Increase conductivity in a full cross-section segment 2.0 mm long by 0.020 S/m. The first-order formula predicts

$$\delta R = -\ell \delta\sigma / (\sigma^2 A) = -10.0\ \Omega.$$

Direct series integration of $1/(\sigma A)$ gives a change from 100.0 to 90.91 Ω in the segment, or $-9.09\ \Omega$.

The 0.91 Ω difference is consistent with truncating at first order for a 10% conductivity change. Recompute at half the material perturbation: the exact change becomes $-4.762\ \Omega$, only 0.238 Ω away from the first-order $-5.00\ \Omega$. The residual decreases by about four when the step halves. Repeating at several positive and negative perturbations and taking a centred difference provides a sharper test of the adjoint code. Varying only the mesh resolution with a fixed large material contrast cannot remove a genuine nonlinear remainder.

The same one-dimensional geometry also demonstrates an absolute identifiability limit. Interchanging two equal-length full-cross-section inserts with different conductivities leaves the end-to-end resistance unchanged, since the integral of reciprocal conductivity is independent of ordering. Perfect measurement

precision cannot recover the order. A voltage tap between the two inserts or an alternative contact geometry is needed to break that symmetry.

Interpret real and imaginary voltage readings together

Even if each local σ, ϵ is real, the fields ψ in an inhomogeneous sample usually carry spatially varying phase. A change $\delta\sigma$ can therefore change both the real and imaginary transfer impedance, as can $i\omega\delta\epsilon$. Treat a complex reading as two real measurements with their calibrated covariance. To first order one finite-element cell with piecewise-constant parameters contributes coefficients $-\int_{\text{cell}} \nabla\psi_{ab} \cdot \nabla\psi_{cd}\, dV$ for $\delta\sigma$ and $i\omega$ times that coefficient for $\delta\epsilon$. The units of the two unknowns differ, so rescale with plausible physical parameter changes before comparing singular values.

As a simple frequency check choose a uniform engineered dielectric-conductor composite with $\sigma = 0.010$ S/m and $\epsilon_r = 1000$. At $f = \sigma/(2\pi\epsilon) \simeq 180$ kHz, $\gamma = \sigma(1 + i)$. The same bar geometry then gives $Z = (L/A)/\gamma = 10.0(1 - i)$ kΩ, a phase of -45°. At low frequency the phase tends to zero and the real impedance to 20.0 kΩ. A constant cable delay produces phase proportional to frequency independent of the sample parameters, so record a calibrated reference at each frequency before attributing an observed phase change solely to the specimen.

Derive a spatial sensitivity that can be checked without a mesh

Consider a homogeneous, thick conducting calibration block with a flat, electrically insulating top except at four small contacts on the same line. Coordinates are x, y along the surface and $z > 0$ into the block. If a unit current enters through a point at the surface, it spreads over hemispheres of radius r. Uniform normal current through a hemisphere requires

$$J_r = 1/(2\pi r^2).$$

The radial field is J_r/σ; integrating from r to infinity gives potential $1/(2\pi\sigma r)$. The coefficient differs from a point source in the interior where an entire sphere carries the current. Subtract a corresponding solution for the terminal extracting the current to satisfy conservation.

Let the drive pair be at $x = \mp a$ and the sense pair at $x = \mp c$, with positive terminal at negative x in each case. At surface points separated from the point contacts the baseline transfer impedance becomes

$$Z_0 = \frac{1}{\pi\sigma}\left(\frac{1}{a - c} - \frac{1}{a + c}\right) \qquad (0 < c < a).$$

For $a = 15$ mm, $c = 5$ mm and $\sigma = 0.20$ S/m this is 79.58 Ω. A 1 mA real current gives 79.58 mV. The same number follows on interchanging the drive and measurement pairs. The large individual contact potentials of an ideal point source are singular at the driving contact itself, yet the transfer potential evaluated at a distinct contact is finite. A two-terminal impedance obtained by setting $c = a$ is divergent

in this point-contact approximation. To make such a measurement calculable one must assign physical electrode radii, contact laws and current distributions.

Place a small spherical inclusion of radius 1.0 mm at $x = y = 0$ and depth z, with conductivity just 0.020 S/m above background. At the centreline the vertical field components cancel in each paired solution. Differentiating the two point-source potentials horizontally gives a gradient magnitude $a/[\pi\sigma(a^2 + z^2)^{3/2}]$ for a drive half-spacing a; replace a by c for the reciprocal measurement field. The first-order change in impedance is the product of the two signed gradients times minus the conductivity perturbation and the sphere volume. The centre sample approximates the volume integral when both background gradients vary slowly over the sphere.

At $z = 8$ mm the result is -3.86 mΩ, a $-3.86\,\mu\mathrm{V}$ perturbation at 1 mA, about 48 parts per million of baseline. Increasing depth to 16 mm leaves only -0.320 mΩ. A point midway between one source and one sensing contact can experience a very different gradient alignment from a point on the common centreline. This is why a general inclusion shift should be calculated by differentiating the vector field before taking the dot product. Replacing each gradient by its magnitude would forbid the experimentally possible positive changes. A point inclusion moved twice as deep does not, in general, change the answer by one universal inverse power. Each field samples a different combination of depth and contact separation.

Retain unknown contact impedance as a separate physical parameter

A metal current terminal is nearly equipotential internally but can have an interface voltage drop at the sample. Let U_i be its metal potential, u the potential on the adjacent sample, $j_i = \gamma \partial_n u$ positive into the body and z_i an area-specific contact impedance in Ω m². A simple local interface law is $U_i - u = z_i j_i$, together with $\int j_i \, dA = I_i$. At an inactive metal terminal only this integral is zero. Current can enter on one side and return on the other if the sample potential varies appreciably across its face. A high-impedance voltmeter does not by itself make that metal patch equivalent to an absent electrode.

For sample and electrode test fields (v, W), integration by parts and the contact law give

$$B((u,U),(v,W)) = \int_\Omega \gamma \nabla u \cdot \nabla v \, dV + \sum_i \int_{S_i} \frac{(U_i - u)(W_i - v)}{z_i} \, dA = \sum_i I_i W_i.$$

Here the current and potential gauges are chosen consistently on both drive and reciprocity solves. The contact product is as symmetric as the bulk product, so finite reciprocal contact impedances retain equality upon drive/sense exchange. If the contact changes, differentiating $1/z_i$ introduces a minus, and the minus from differentiating the full field equations cancels it. The additional first-order sensitivity is $+\sum_i \int \delta z_i j_{i,d} j_{i,m} \, dA$, with both current densities normalised to unit terminal current. In a uniform two-terminal contact of area A, the two normalised currents are $1/A$ and one contact contributes $+\delta z/A$, as an ordinary series contact should. This independent limit fixes both the positive sign and the square-metre convention.

For a controlled reconstruction, form one column per candidate sample perturbation and another per uncertain contact coefficient. Weight measured channels by their noise covariance. Two nearly

proportional columns then warn that bulk variation and degrading contact cannot be separated by those channels over that bandwidth. Reciprocity catches lead swaps and some drift, but a reciprocal change in a contact can pass this check. Repeat with a different electrode pattern, or calibrate against a homogeneous reference block at the same frequency. To verify a numerical sensitivity, perturb the contact alone by equal positive and negative amounts and compare the centred impedance difference with the surface integral. Then perturb the sample alone and compare with the volume integral. A successful combined test could otherwise conceal equal and opposite implementation errors.

Move a weak inclusion beneath four surface contacts

The block is treated as a homogeneous half-space except for a 1 mm radius sphere with a positive conductivity perturbation. Each plotted point uses the two unperturbed analytical vector gradients at the sphere centre. The second curve repeats the calculation at twice the selected depth. At 8 mm and 5 mm sensing half-spacing the predicted centre perturbation is −3.86 mΩ against a baseline of 79.58 Ω. Look for lateral positions where the sign changes before interpreting increased conductivity as a uniformly smaller transfer impedance.

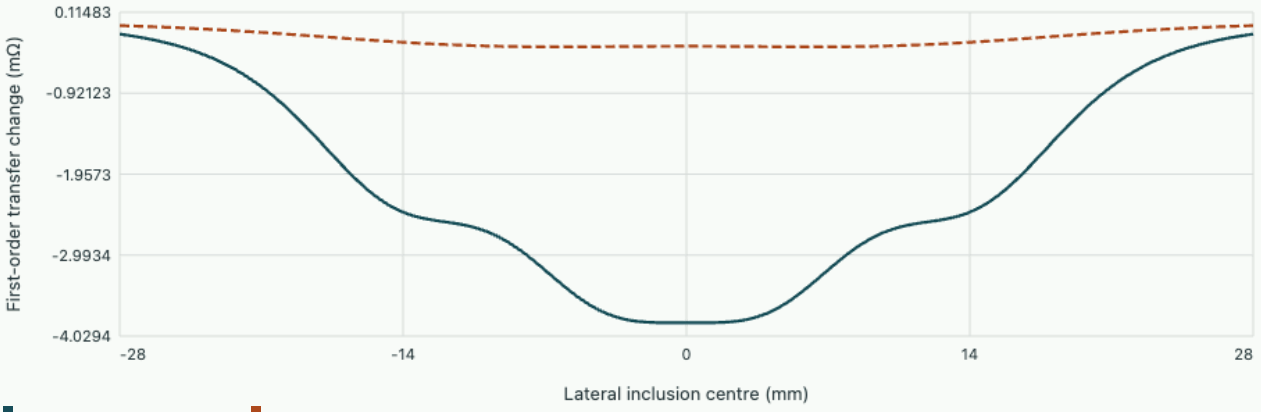
Widen the sensing pair and predict how both the transfer baseline and the local sensitivity will respond. A larger transfer voltage does not guarantee a proportionately larger fractional change for a chosen deep inclusion. The ideal point-contact fields vary steeply close to electrodes. When the inclusion radius approaches the distance over which the gradients change, volume quadrature supplies a better approximation than evaluating a single point. A finite laboratory specimen also needs boundaries sufficiently distant compared with the patterned contact spacing.

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.

Tutorial · bound a perturbation and design another observation

Predict a transfer change without solving again

A field computation for unit drive and unit reciprocal current in a small inclusion gives approximately parallel real field gradients of magnitudes 1500 and 400 V/(m A). The inclusion volume is 2 mm³. Its conductivity rises by 0.010 S/m. Predict the resistance change and the observed voltage change at a 5 mA drive. What changes if only the measurement electrodes are reversed?

▼ **Work through the solution**

The overlap is $6.0 \times 10^{5}\ \mathrm{V}^2/(\mathrm{m}^2\mathrm{A}^2)$. Multiplying by $2 \times 10^{-9}\ \mathrm{m}^3$ and the material perturbation gives $\delta Z = -1.2 \times 10^{-5}\ \Omega$, or -12 μΩ. At 5 mA the transfer voltage changes by -60 nV. Reversing only c, d reverses the reciprocal field and the transfer change. The result assumes nearly uniform overlap inside the inclusion and small perturbation. Subdividing the inclusion or independently perturbing the forward model tests these assumptions differently. A voltmeter's microvolt accuracy would be inadequate for the proposed single measurement.

Do more electrodes resolve two unknown defects?

Two calibrated real channels have sensitivity rows proportional to $(1, 0.8)$ and $(0.8, 1)$ for two dimensionless defect amplitudes. Which combination is measured poorly? Would two additional measurements that simply reverse the existing electrode signs supply a new material constraint?

▼ **Work through the solution**

The common pattern has gain 1.8 and the difference pattern gain 0.2; their singular-value ratio is nine in these units. Distinguishing which defect changed is harder than detecting a common material change. Sign reversals duplicate the same rows up to a minus sign. They can cancel offsets or improve precision by averaging independent noise, but do not increase the rank of the noise-free sensitivity matrix. A drive/measurement pair with opposing sensitivities to the two spatial regions would directly strengthen the weak difference direction.

Further reading

- Derive electrostatic reciprocity from a boundary-value problem
- Check complex field conventions with an independent power balance
- Somersalo, Cheney and Isaacson · existence and uniqueness for the complete electrode model

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 $[m_1^2 + m_2^2 + 2m_1 m_2\cos\chi]/(m_1 + m_2)^2$. 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 $\sin i = b/a,\ \sin r_i = b/(n a)$. 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 $2a \cos r_i/(n v_\infty)$, because the interior chord has offset b/n and speed $n v_\infty$. 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

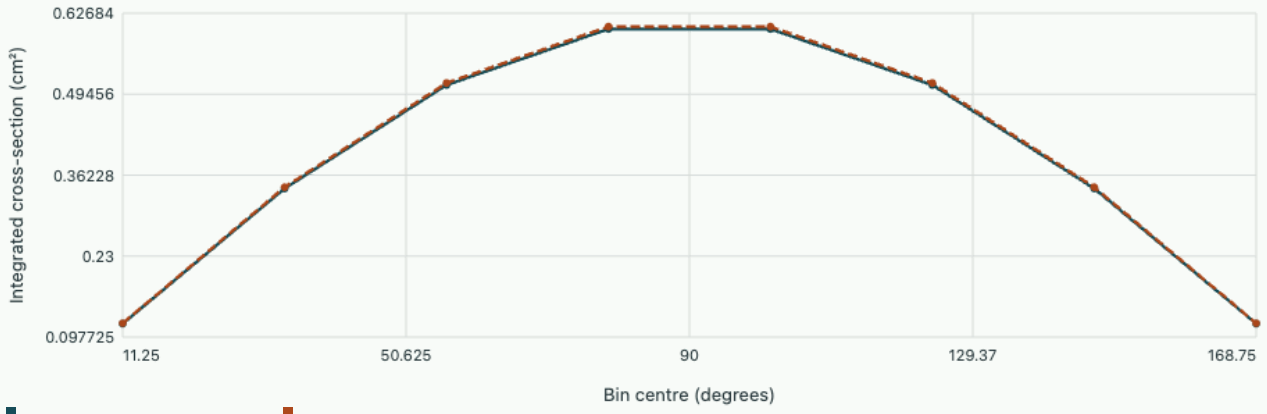
Bin the exactly solvable hard-sphere reflection in polar angle and compare the integrated cross-section with the incoming geometric area.

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.

Applied study

Separate residual, solution error and conditioning when solving linear systems.

Can a small residual hide a large error?

After discretising steady heat conduction, the unknown nodal temperatures form a vector x. Heat balance at the nodes gives $Ax = b$. Substitute an approximate solution $\hat{x}$ back into this equation. The mismatch $r = b - A\hat{x}$ is the residual. The temperature error $e = x - \hat{x}$ satisfies $Ae = r$. Even an exactly evaluated residual still has to pass through A^{-1} to tell us about temperature.

For nonsingular A, the Euclidean norm gives

$$\|e\|_2 \le \|A^{-1}\|_2 \|r\|_2.$$

Since $\|b\|_2 \le \|A\|_2 \|x\|_2$, a nonzero right-hand side gives

$$\frac{\|e\|_2}{\|x\|_2} \le \kappa_2(A) \frac{\|r\|_2}{\|b\|_2}, \qquad \kappa_2(A) = \|A\|_2 \|A^{-1}\|_2.$$

The condition number measures how strongly a problem can amplify a relative perturbation. The bound covers the worst direction. An error in a different direction may have a much smaller amplification.

Work through a poorly resolved direction

Take $A = \operatorname{diag}(1, 10^{-4})$, $b = (1, 10^{-4})^T$. The exact solution is $(1, 1)^T$. The trial vector $(1, 0)^T$ gives a residual of magnitude 10^{-4}, nearly 10^{-4} relative to b. Yet one component is missing altogether and the relative solution error is

$$1/\sqrt{2} = 0.707.$$

Along the second coordinate, multiplying by A suppresses an error by a factor of ten thousand.

Scale the second equation by 10^4. The exact solution stays the same and the scaled matrix is the identity. The new residual of the same trial vector is $(0, 1)^T$. Always say which system and norm a reported stopping test uses. Temperature differences from a direct solve on a small mesh provide an independent check in physical units.

Choose an iteration from the matrix structure

For a symmetric positive definite matrix, solving $Ax = b$ minimises the strictly convex quadratic $J(x) = x^T A x / 2 - b^T x$. Its negative gradient at $\hat{x}$ is r. Conjugate gradients (CG) searches in the span of $r_0, Ar_0, \ldots$, using search directions orthogonal in the A-inner product. It minimises the energy error $\|e\|_A = \sqrt{e^T A e}$ over the growing trial space.

In exact arithmetic CG terminates in at most as many iterations as there are distinct eigenvalues. A positive definite preconditioner M seeks a more favourable spectrum for $M^{-1/2}AM^{-1/2}$. For a moderate problem a direct factorisation gives a convenient reference. Both methods solve a discrete system, whose mesh error persists even when the algebraic equations are solved to roundoff.

Problem class

Set a defensible tolerance

A condition estimate bounds $\kappa_2(A)$ by 5000. How small should the relative residual be to guarantee relative solution error at most 1%? If the solver reports 10^{-5}, what can you conclude?

▼ **Show the calculation**

Require $5000\|r\|/\|b\| \le 0.01$, or a relative residual at most 2×10^{-6}. A residual of 10^{-5} guarantees at most 5%. The actual error can still be below 1% because the bound may be loose.

Check whether CG applies

Consider two interior temperatures in a uniform line with fixed ends. Their matrix, up to a positive factor, is $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$. Identify its two modes, condition number and maximum number of exact-arithmetic CG steps.

▼ **Show the calculation**

Multiplication gives eigenvalues 1 for $(1,1)^T$ and 3 for $(1,-1)^T$. Both are positive, and the matrix is symmetric. Thus $\kappa_2 = 3$ and two search directions suffice. The alternating mode has a larger stiffness because the temperature changes more sharply between neighbours.

ONESIMULATION LABORATORY

Separate algebraic error from mesh error

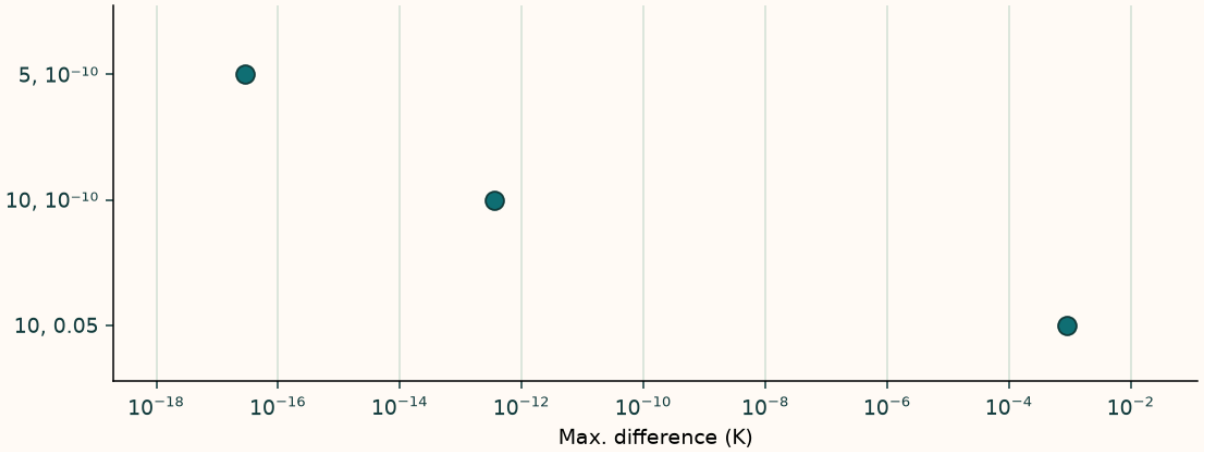
Compare recorded OneSimulation runs on this page. To test your own inputs, download the Python model and run the model in an activated OneSimulation environment.

A unit cube has conductivity 1 W/(m·K), uniform heating 1 W/m³ and zero temperature rise on every face. The script compares CG and a direct solver on the same trilinear finite-element mesh. Predict the effects of first refining the mesh and then relaxing only the CG tolerance. Download the Python experiment · OneSimulation linear solvers.

```
python linear-algebra.py
python linear-algebra.py --cells 10
python linear-algebra.py --cells 10 --rtol 0.05
```

Solver comparison on the heated cube. The difference uses all nodes.

Cells per axis, CG tolerance	Direct peak (K)	CG peak (K)	Max. difference (K)
5, 10^{-10}	0.05483074	0.05483074	2.8×10^{-17}
10, 10^{-10}	0.05708900	0.05708900	3.6×10^{-13}
10, 0.05	0.05708900	0.05620782	8.8×10^{-4}



On the ten-cell mesh, relaxing the CG tolerance increases the peak solver-to-solver difference to about 0.00088 K. The logarithmic axis resolves the many orders of magnitude.

The refined mesh includes the cube centre as a node, whereas the five-cell mesh samples the peak off-centre. The change between direct nodal maxima therefore includes a sampling effect. The final two rows use identical nodes, isolating the effect of the CG stopping test. Differences near roundoff can vary between implementations.

Change parameters · Change the laboratory inputs

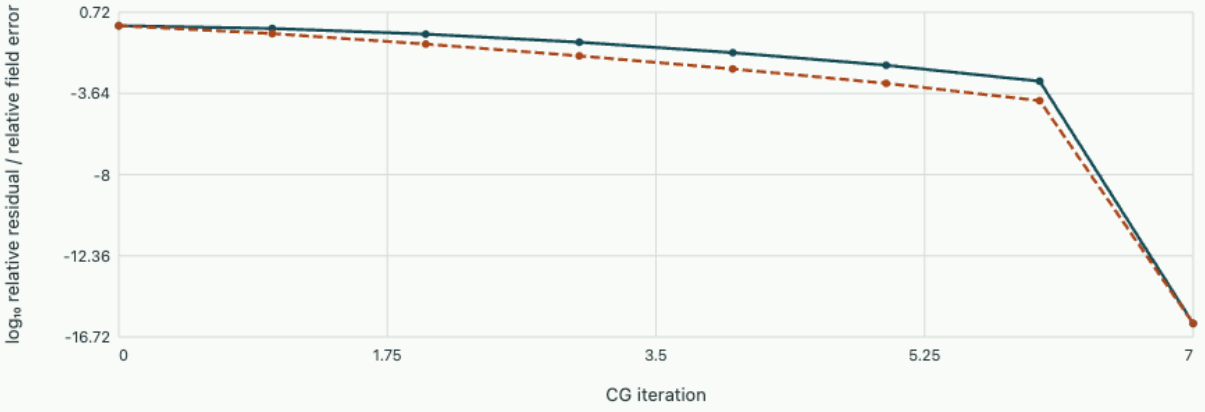
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Cube mesh intervals 6

$\log_{10}$ relative CG tolerance -5



For −Δu = 1 in a grounded unit cube: reference peak 0.053966; selected CG peak 0.053966; maximum difference 0. This browser companion uses a seven-point finite-difference grid; the linked OneSimulation laboratory uses trilinear finite elements, so compare convergence behaviour rather than identical nodal values.

Problem sheet 3 · Bound the effect of stopping a linear solve early

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.

Condition number

$$\kappa(A) = \|A\|\|A^{-1}\|$$

For an invertible square matrix choose a compatible norm. Conditioning describes data sensitivity, while stability refers to the algorithm processing the data.

Derivation · Conditioning and backward error

Krylov subspace

$$\mathcal{K}_k(A, b) = \operatorname{span}(b, Ab, \ldots, A^{k-1}b)$$

The starting vector determines which parts of the spectrum are visible. Orthogonalisation avoids storing nearly dependent raw powers.

Derivation · Arnoldi iteration and GMRES

Orthogonal projector

$$P = QQ^{\mathsf{T}}$$

Columns of Q must be orthonormal in the stated real inner product. Least squares removes the component of the data orthogonal to the fitted range.

Derivation · Least squares and orthogonal factorisation

Residual

$$r = b - A\hat{x}$$

A small residual tests the computed equation. Converting it into an error estimate needs information on the inverse or the troublesome singular directions.

Derivation · Iterative refinement and backward error

Schur complement

$$S = D - CA^{-1}B$$

Eliminate the first block of a partitioned system by solving with A. Forming an explicit inverse is unnecessary and may discard useful structure.

Derivation · Eliminating constraints with Schur complements

Singular value

$$Av_i = \sigma_i u_i$$

Input and output directions can live in different spaces. A small positive singular value can cause large inverse amplification.

Derivation · Singular values and minimum-norm solutions

Further reading

- Follow the sequence from orthogonality and projection to conditioning and backward error. Use the same least-squares problem to distinguish sensitivity of the data from stability of a factorisation. Trefethen and Bau · Numerical Linear Algebra
- Chapters 2, 3 and 8 connect factorisations, least squares and Krylov methods through executable examples. Driscoll and Braun · Fundamentals of Numerical Computation

Compare specific topics with the university sources