Numerical partial differential equations

Discretise field equations with conservation and boundary data, resolve interfaces and pressure constraints, and estimate error in a chosen output.

Read the subject in depth

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

- ### Spatial discretisation and field equations

  Derive a discrete field problem from a continuous equation, choose appropriate boundary data and test the convergence of quantities a measurement could resolve.

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. ### Boundary data, uniqueness and singular sources

   Derive the continuous problem before assembling a discrete row. Test what happens to uniqueness, source flux and interface voltage under a change of boundary data.

2. ### Heat, momentum and wave transport

   Test the same spatial method on diffusion, a propagating pulse and an elastic mode. Compare dissipation, dispersion and boundary residuals on a common footing.

How the course fits together

1. **Chapters 1–4**

   Turn flux balance on a cell into neighbouring unknowns. The weak form imposes the same balance against test functions. Check whether face fluxes cancel, then add storage to predict a field evolving in time.

2. **Chapters 5–8**

   Track the numerical speed of waves and verify a discretisation using a known manufactured field. Interfaces require matching flux, and a tensor coefficient makes the flux direction depend on material orientation.

3. **Chapters 9–12**

   Advection, pressure constraints and widely separated mesh scales require different discretisations and solvers. End by defining the particular output whose error matters and weighting residuals accordingly.

Contents

1. Chapter 1 · Flux differences for an elliptic balance
2. Chapter 2 · Weak forms and finite elements
3. Problem sheet 1 · Flux balances and virtual work
4. Chapter 3 · Finite volumes and moving fronts
5. Chapter 4 · Time-dependent diffusion
6. Problem sheet 2 · Fronts and diffusive modes
7. Chapter 5 · Wave speeds and numerical dispersion
8. Chapter 6 · Verification, refinement and local error
9. Problem sheet 3 · Wave resolution and verification
10. Chapter 7 · Interfaces, contact resistance and Robin data
11. Chapter 8 · Anisotropic diffusion and mesh alignment
12. Problem sheet 4 · Material interfaces and anisotropy
13. Chapter 9 · Cell Péclet number and monotonicity
14. Chapter 10 · Incompressibility and discrete pressure modes
15. Problem sheet 5 · Transport weights and pressure spaces
16. Chapter 11 · Multigrid and coarse-grid correction
17. Chapter 12 · Adjoints and errors in a chosen output
18. Problem sheet 6 · Transfers and observation error
19. Chapter 13 · Pure flux boundaries, compatibility and the missing constant
20. Chapter 14 · Building a local residual estimator and refining where it matters
21. Chapter 15 · What a periodic grid does to a decaying Fourier mode
22. Chapter 16 · Why two finite-element wave solvers propagate the same mode at different speeds
23. Chapter 17 · Moving a cell average without inventing or losing material
24. Chapter 18 · When diffusion destabilises a uniform state
25. Problem sheet 7 · Project · Verify a heat solver before interpreting its temperature
26. Problem sheet 8 · Resolve an outflow layer by fitting the numerical face flux
27. Spatial discretisation and field equations
28. Boundary data, uniqueness and singular sources
29. Heat, momentum and wave transport
30. Applied study
31. Notation and assumptions
32. Further reading

Chapter 1 · Flux differences for an elliptic balance

Assemble a conservative boundary-value problem from face fluxes, trace the effect of boundary conditions and establish a discrete energy identity.

FROM FIRST PRINCIPLES

Derive a cell equation by integrating the physical balance

Consider steady one-dimensional conduction in a rod with area $A(x)$, conductivity $k(x) > 0$ and heating per unit length $s(x)$. Define heat current F positive to the right. Conservation in any finite interval requires that the current leaving equals the current arriving plus generation. Integrating over cell i fixes a discrete equation without yet approximating a derivative:

$$F_{i+1/2} - F_{i-1/2} = \int_{x_{i-1/2}}^{x_{i+1/2}} s(x)\,dx.$$

Each shared face must use the same current in its two adjacent cells. To close a face between constant-material regions with no intervening source, integrate Fourier's law from one cell centre to the other. The heat current is constant on that interval, so it can be taken outside the integral. Defining the face resistance $R_{i+1/2}$ makes the sign of the temperature drop explicit:

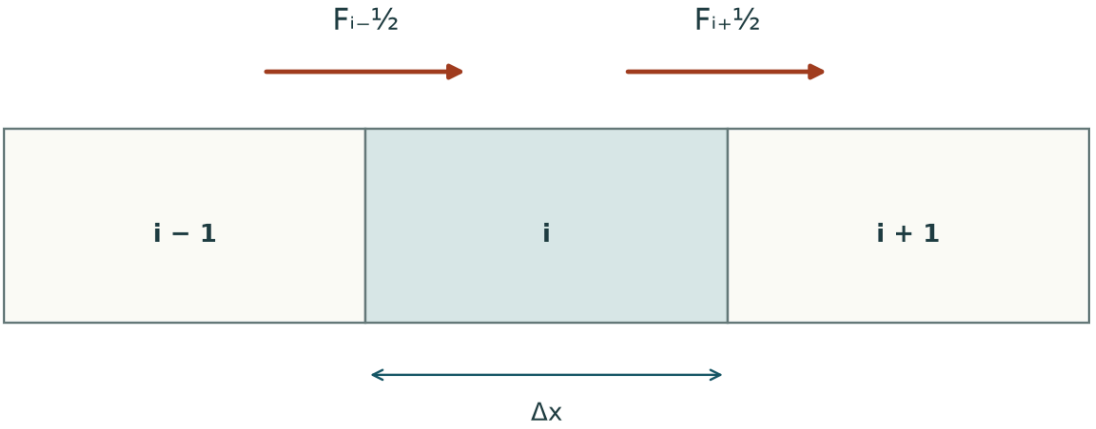
$$\begin{aligned} R_{i+1/2} &= \int_{x_i}^{x_{i+1}} \frac{dx}{k(x)A(x)}, \\ T_{i+1} - T_i &= -F_{i+1/2}\,R_{i+1/2}, \\ F_{i+1/2} &= \frac{T_i - T_{i+1}}{R_{i+1/2}}. \end{aligned}$$

On a uniform homogeneous mesh this reduces to $kA(T_i - T_{i+1})/h$. Across a material change the half-cell resistances must instead be added. Substituting face currents into the cell balances produces the matrix system. Boundary temperatures, prescribed currents or film laws determine which additional boundary equations close it.

Background · Conservative face flux

$$|C_i|\dot{u}_i + \sum_{f \subset \partial C_i} F_{if} = |C_i| s_i$$

Use the same face flux with opposite orientations in neighbouring cells. Interior contributions then cancel in the global balance.



Why an internal face must enter its neighbours with opposite signs
A positive rightward flux brings material into cell i through its left face and removes material through its right. Dividing the difference by Δx gives the rate of change of the cell average. Summing over cells cancels each shared flux once; only the two outside faces survive.

DERIVATION AND APPLICATIONS

Why a flux-only boundary requires a compatibility condition

An insulated or flux-controlled steady diffusion calculation can fail even when each individual face law is correct. Consider a one-dimensional bar of unit area, positive conductivity k, and cell sources s_i per unit length. Define $F_{i+1/2}$ as heat current positive to the right everywhere, including both boundaries. At steady state the current leaving the right face minus the current entering the left face equals the total source in that cell:

$$F_{i+1/2} - F_{i-1/2} = s_i \Delta x_i.$$

Summing over cells cancels every shared interior current. No detail of the face-gradient approximation can change the remaining global condition:

$$F_R - F_L = \sum_i s_i \Delta x_i.$$

With only prescribed fluxes, the discrete gradient also cannot determine a temperature offset. For a connected linear system with positive face transmissibilities, its assembled quadratic form is a sum of squared temperature differences:

$$\mathbf{T}^{\mathsf{T}} K \mathbf{T} = \sum_{\text{interior faces}} g_{ij}(T_i - T_j)^2,$$

$$K\mathbf{1} = 0.$$

The nullspace consists precisely of constant temperatures. A compatible right-hand side gives a family of solutions; prescribe one reference temperature or a weighted mean to select a member. An incompatible right-hand side gives no steady solution. Physically, the imbalance would accumulate energy. Removing one algebraic equation or pinning a temperature without checking global balance can inadvertently create an unreported heat sink.

WORKED EXAMPLE

Reconstruct temperatures from compatible flux data

Take three equal cells with source totals $1, 2, 1$ W, left incoming current $F_L = 2$ W and right outgoing current $F_R = 6$ W. Internal face currents are successively 3 and 5 W. With both conductances equal to 2 W/K, the temperature drops are 1.5 and 2.5 K from left to right. For a prescribed arithmetic mean of 300 K the temperatures are 301.833, 300.333 and 297.833 K. Adding a common constant would leave every face measurement unchanged.

SECOND APPLICATION

Diagnose an unbalanced insulated test

Seal both ends so $F_L = F_R = 0$ but retain the same sources. Their total is 4 W. If the combined heat capacity is 200 J/K, the heat-capacity-weighted mean temperature must rise at $4/200 = 0.0200$ K/s. A steady solver reporting fixed temperatures under those conditions should be checked for a hidden imposed temperature or a dropped residual. The missing four watts have a direct transient interpretation.

> **Start from this chapter.** Start with the difference of two face fluxes. The coefficient varies in space, so leaving its derivative out changes the manufactured source. Use the supplied solve to check the refined nodal error.
>
> OneSimulation laboratory · Measure spatial convergence on a known field

Place fluxes where conservation uses them

Consider steady conduction on $0 < x < L$ with heat flux $q = -k(x)T'$ and volumetric generation f. Taking unit cross-sectional area, steady energy balance gives $q' = f$. On uniform nodes separated by h, integrate the balance over a control segment centred on x_i:

$$q_{i+1/2} - q_{i-1/2} \simeq h f_i.$$

A shared face must have the same numerical flux in both adjacent equations, otherwise summing local equations can create an artificial internal source or sink.

Use $q_{i+1/2} = -k_{i+1/2}(T_{i+1} - T_i)/h$. Then row i of the discrete operator has diagonal $(k_{i-1/2} + k_{i+1/2})/h^2$ and off-diagonals $-k_{i\pm1/2}/h^2$. For positive k the off-diagonals are negative and the diagonal balances them. With zero source, an unconstrained interior value is a weighted average of

its neighbours. This relation explains why a source-free discrete conductor with fixed end temperatures cannot develop an interior temperature larger than both ends.

Compute the face law across an interface

For smooth k an arithmetic average of neighbouring nodal coefficients provides a second-order face approximation on a uniform grid. A sharp material interface calls for a different physical calculation. Consider two half-segments in series with constants k_L, k_R and no interface storage or contact resistance. The same flux crosses both halves. Adding their temperature drops gives $T_L - T_R = q[h/(2k_L) + h/(2k_R)]$, hence $k_{\text{eff}} = 2/(1/k_L + 1/k_R)$. This is a harmonic average for equal half-widths.

With $k_L = 1,\ k_R = 9$ W/(m·K) and a 10 K drop over a 0.1 m pair, the effective conductivity is 1.8 W/(m·K) and the flux is 180 W/m². Using the arithmetic mean would give 500 W/m², severely underestimating the resistance of the poorer conductor. In the limit $k_L \to 0^+$, the physical flux tends to zero while that arithmetic estimate remains finite. For unequal layer thicknesses use the actual layer resistances rather than either unweighted average.

Examine uniqueness and the role of the boundary

For homogeneous fixed temperatures and an interior vector u extended by zero at both ends, grouping adjacent row contributions gives

$$h \sum_i u_i (A_h u)_i = \sum_{\text{faces}} \frac{k_{i+1/2}}{h} (u_{i+1} - u_i)^2.$$

Each term is nonnegative. Vanishing energy requires a constant vector, and the fixed ends force that constant to be zero. Thus the interior matrix is symmetric positive definite. This reasoning justifies CG and gives a uniqueness argument that does not depend on inspection of a particular numerical matrix. With only prescribed heat fluxes a spatial constant instead remains undetermined. A solution then requires overall heat compatibility and one temperature reference or a mean constraint.

As a check, take constant $k = 1$, $L = 1$, zero end temperatures and generation 2. The exact profile $T = x(1 - x)$ has central value 0.25. Central differences recover its second derivative exactly, so all uniform grid nodes satisfy the balance without truncation error. This successful check verifies signs, boundary handling and the quadratic patch. To measure a convergence order use a smooth field with a nonzero leading truncation term, such as a sine profile, and compare at common physical positions.

Background · Can fixing one node hide an incompatible heat balance?

With specified flux at both ends, a steady source must exactly equal the net heat leaving the domain. After moving boundary terms to the right, the discrete equations have the form $KT = b$, where K is symmetric and its rows sum to zero. Consequently

$$\mathbf{1}^T b = \mathbf{1}^T K T = 0.$$

Pinning a temperature removes the undetermined additive constant. If the vector b fails this sum test, however, the pinned solve creates a reaction at that node. Physically that reaction is an unmodelled heat supply or heat sink. After solving, restore the original equation at the pinned node and inspect its residual. It exposes exactly the load that would have to be supplied to maintain the supposed steady state.

Why fixing one boundary value controls the solution

Suppose the left end of a conducting rod is held at zero temperature and the right end is insulated. A finite system of flux balances may be invertible on one mesh. We would also like a bound which survives refinement. In the continuous problem, use the space $V = \{v \in H^1(0, L) : v(0) = 0\}$. Its members have square-integrable first derivatives and a well-defined endpoint value. Every admissible difference between two temperature fields belongs to this same space. The fundamental theorem of calculus and Cauchy–Schwarz give

$$|v(x)|^2 = \left| \int_0^x v'(s)\, ds \right|^2 \le x \int_0^x |v'(s)|^2\, ds, \qquad \|v\|_2 \le \frac{L}{\sqrt{2}} \|v'\|_2.$$

The second inequality follows by integrating the first and bounding its inner integral by the full derivative norm squared. Thus a small temperature gradient forces the temperature itself to be small in this particular space. Take a possibly discontinuous conductivity with $0 < k_0 \le k(x) \le k_1$ almost everywhere and a square-integrable heat source f. Testing the balance and integrating once gives $a(u, v) = \ell(v)$ for every $v \in V$, with

$$a(u, v) = \int_0^L k u' v'\, dx,$$

$$\ell(v) = \int_0^L f v\, dx.$$

The upper bound on conductivity gives

$$|a(u, v)| \le k_1 \|u'\|_2 \|v'\|_2.$$

The lower bound gives $a(v, v) \ge k_0 \|v'\|_2^2$, called coercivity. The endpoint inequality bounds $|\ell(v)|$ by $L\|f\|_2\|v'\|_2/\sqrt{2}$. These are the hypotheses of the Lax–Milgram theorem on the Hilbert space with

norm $\|v'\|_2$. Consequently the weak solution exists and is unique. For two loads, subtract the weak equations and test with their solution difference. Cancelling one derivative norm gives

$$\|(u-\tilde{u})'\|_2 \le L\|f-\tilde{f}\|_2/(\sqrt{2}\,k_0).$$

Every conforming discrete subspace inherits the same estimate.

How large a change can a load error produce?

In scaled units let $L=1,\ k=2,\ f=3$. Integration of $-2u''=3$, first from the insulated end and then from the fixed end, gives

$$u' = 1.5(1-x),$$

$$u = 1.5x - 0.75x^2.$$

Its derivative norm is

$$1.5/\sqrt{3} = 0.8660.$$

The theorem guarantees at most $3/(2\sqrt{2}) = 1.0607$. Raising the source everywhere by 0.12 changes that norm of the difference by $0.06/\sqrt{3} = 0.03464$, within the guaranteed 0.04243. The estimate also applies to nonuniform perturbations of the same load norm. If both ends are insulated, constants have zero derivative and can be added freely. A net heat input then rules out steady balance altogether.

Practise on sheet 1 · A longer rod with a clamped or insulated left end. Further reading · Heidelberg University · Lecture Notes on Scientific Computing with Partial Differential Equations · §5.5 Abstract Existence Theory, Theorem 5.22 (Lax–Milgram), and §6.3 Mixed Problem.

Eliminate boundary values with the correct sign

With four intervals on a unit rod, fixed left temperature 10, fixed right temperature 20 and zero source, multiply each row by h^2/k. The three unknowns then satisfy a tridiagonal system with diagonal two, off-diagonal minus one and load $(10,0,20)^T$. Moving a fixed value to the right changes its sign, producing positive boundary contributions here. Solving gives the linearly spaced interior temperatures 12.5, 15 and 17.5. Every discrete face reports the same flux for constant k.

This deliberately simple case isolates boundary elimination and sign conventions. Reversing only one load sign can generate a plausible-looking smooth curve with entirely wrong end-to-end transport. After checking the values, also sum all interior balances and verify cancellation of the internal face contributions. The resulting global statement includes only boundary fluxes and any source integrated over the interior control volumes.

Prove a discrete maximum principle from the face balances

Let positive conductances $G_{i-1/2}$ and $G_{i+1/2}$ connect each internal node to its neighbours, with both end temperatures specified. In the absence of internal sources its balance reads $G_{i-1/2}(u_i - u_{i-1}) + G_{i+1/2}(u_i - u_{i+1}) = 0$. Dividing by the sum of the conductances makes u_i a convex combination of its two neighbours. If a strict global maximum occurred inside the domain above both prescribed boundary values, its neighbours would each have to equal this maximum. Repeating along the connected grid would carry the same value to a boundary, contradicting the assumption.

The same reasoning applied to the negative field bounds the minimum. With a nonnegative internal heat source s_i on the right, an interior minimum below both boundaries would have both outgoing temperature differences nonpositive. Their positive weighted sum cannot equal a positive source, and if the sum is zero, propagation again carries the minimum to the boundary. A positive source may raise an interior maximum while preserving the lower comparison bound. Negative face conductances or a disconnected component invalidate the argument; symmetry and a small algebraic residual do not by themselves restore it.

Compare two thermal loads without inverting their matrices

Connect internal nodes 1 and 2 to end temperatures zero and ten, using consecutive conductances 1, 2 and 4 W/K. With no source the equations are

$$3u_1 - 2u_2 = 0,$$
$$-2u_1 + 6u_2 = 40.$$

The first gives $u_1 = 2u_2/3$, and the second gives

$$u_2 = 60/7,\ u_1 = 40/7.$$

Both values lie between the boundary temperatures. The common leftward heat flow has magnitude $40/7$ W, checked independently from each of the three temperature drops times its conductance.

Add a heat source of 7 W at node one. The difference v between new and old solutions has zero boundary values and satisfies

$$3v_1 - 2v_2 = 7,\ \ -2v_1 + 6v_2 = 0.$$

The second relation gives $v_1 = 3v_2$, so

$$v_2 = 1,\ v_1 = 3.$$

Both internal states warmed, consistent with the comparison principle applied to the difference. The additional outflow at the left boundary is $1 \cdot 3 = 3$ W and at the right is $4 \cdot 1 = 4$ W, accounting for the full applied source.

This provides two independent implementation tests. Conservation checks that the boundary power change equals the injected power. Monotonicity checks the signs of all changes in the field. A sign error in a face contribution can violate one of these checks even while the computed algebraic residual for the incorrectly assembled matrix is tiny. Both tests use physically accessible quantities and remain meaningful when no exact closed-form solution of a complex geometry is available.

The argument extends to a connected graph, since a weighted average at a maximising node requires equality along every positive-conductance edge. This observation is useful for testing an unstructured implementation. Deliberately disconnecting an entirely insulated component produces a free constant there, so boundary comparisons cannot control that component until a temperature reference or independent constraint is supplied.

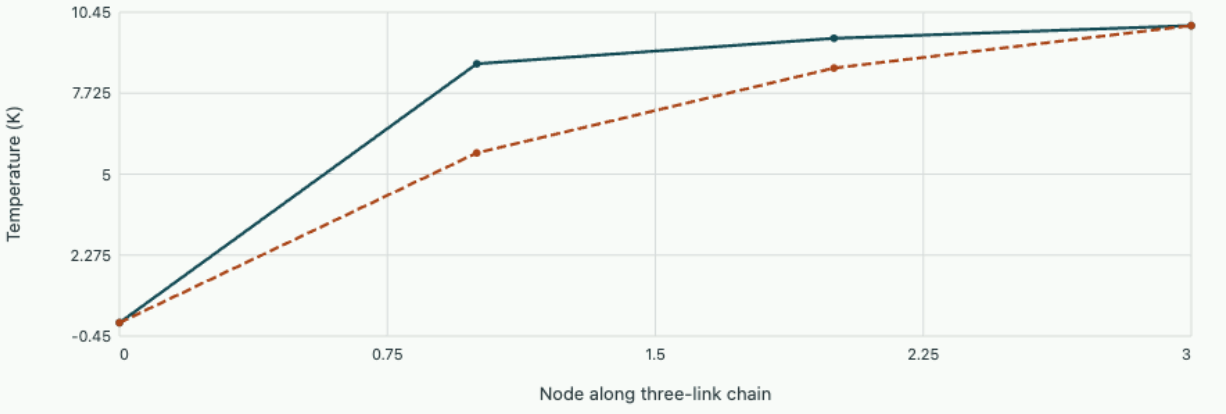
Change parameters · Flux differences for an elliptic balance

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

Open this calculation with adjustable inputs

Try the calculation

Middle conductance (W/K)	2
Source at first interior node (W)	7
Right boundary temperature (K)	10



Additional boundary outflow 7 W equals the inserted 7 W. Both interior temperature increments are nonnegative.

Study the full derivation

Choose the physically admissible wave in a conservative update · Spatial discretisation and field equations

A discontinuous scalar flow can satisfy integral conservation in more than one way. Derive the shock speed, use a convex entropy to distinguish a compression from an expansion and build a finite-volume flux from the resulting local solution.

Chapter 2 · Weak forms and finite elements

Derive a variational statement by integration by parts, assemble element matrices for a one-dimensional rod and explain what the Galerkin solution minimises.

FROM FIRST PRINCIPLES

Obtain the weak diffusion problem by moving one derivative

For a stationary diffusion problem $-\nabla \cdot (k\nabla u) = f$, suppose temperature is prescribed on part of the boundary and the outward flux $q_n = -k\nabla u \cdot \mathbf{n}$ on the remainder. A classical second derivative may not exist across a material interface. Instead multiply the conservative equation by a smooth test function v that vanishes where u is prescribed. The divergence theorem moves one derivative from the flux onto v:

$$\int_\Omega k\nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega - \int_{\Gamma_N} v q_n \, d\Gamma.$$

Approximate u with continuous element shape functions N_j that are differentiable inside each element. Choose each allowable shape function as the test function. Linearity then yields element contributions

$$K_{ij}^{(e)} = \int_e k\nabla N_i \cdot \nabla N_j \, d\Omega,$$

$$f_i^{(e)} = \int_e f N_i \, d\Omega.$$

Neighbouring elements contribute to the same global unknown at a shared node. Adding their contributions enforces weak equilibrium. The flux sign in the boundary term follows from an outward-positive definition; switching to prescribed inward heat supply reverses that sign in the load vector.

Background · Weak form

$$\int k\nabla u \cdot \nabla v \, dV = \int f v \, dV + \text{boundary terms}$$

Integration by parts lowers the derivative requirement. Essential and natural boundary data enter the discrete system differently.

DERIVATION AND APPLICATIONS

How a Robin boundary enters both the stiffness matrix and the

load

A convecting boundary couples an unknown surface temperature to an ambient temperature. Unlike a prescribed heat source it contributes to both sides of the assembled equations. Consider steady diffusion $-\nabla \cdot (k \nabla T) = s$ in a two-dimensional region of unit thickness, with $k > 0$. On a straight boundary segment impose the empirical film law $-k \nabla T \cdot \mathbf{n} = h(T - T_\infty)$ using the outward normal. Multiply the field equation by a test function v vanishing on prescribed-temperature boundaries, integrate, and apply the divergence theorem:

$$\int_\Omega k \nabla v \cdot \nabla T \, dA - \int_{\partial\Omega} v k \nabla T \cdot \mathbf{n} \, ds = \int_\Omega v s \, dA.$$

Substitution on the film boundary Γ places the unknown surface term on the left and the ambient contribution on the right:

$$\int_\Omega k \nabla v \cdot \nabla T \, dA + \int_\Gamma h v T \, ds = \int_\Omega v s \, dA + \int_\Gamma h v T_\infty \, ds$$

with any other natural boundary terms added separately. To assemble one linear edge of length ℓ, use edge functions $N_1 = 1 - \xi$, $N_2 = \xi$, $\xi \in [0, 1]$. Insert the nodal interpolation for T and choose each N_i as test function. Constant h, T_∞ give

$$K_\Gamma = \frac{h \ell}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad f_\Gamma = \frac{h \ell T_\infty}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The off-diagonal term is present because both nodes influence temperature across the same edge. In the limit $h = 0$ the boundary is insulated. A very large film coefficient constrains the temperature toward ambient but can introduce a large stiffness contrast.

WORKED EXAMPLE

Assemble a convecting edge by hand

Use a one-metre out-of-plane thickness, $\ell = 0.300$ m, $h = 20.0$ W/(m² K) and $T_\infty = 290$ K. Then K_Γ has diagonals 2.00 W/K and off-diagonals 1.00 W/K. The load is $(870, 870)$ W. At nodal temperatures 300 and 310 K the film residual $K_\Gamma T - f_\Gamma$ is $(40, 50)$ W. The sum 90 W agrees with $h \ell [(300 + 310)/2 - 290]$.

SECOND APPLICATION

Test a constant solution before using the boundary routine

Set both temperatures to 290 K, the ambient value. Each row has sum 3.00 W/K, so every residual is $3(290) - 870 = 0$. If the code includes the ambient load but forgets the matrix, it predicts 1740 W of

heating at thermal equilibrium. This constant-field patch test detects sign errors and omission of the surface matrix independently of the volume mesh.

Reduce the differentiability required by equilibrium

A rod is fixed at $x = 0$, pushed by a distributed load f and by an end force t at $x = L$. Axial displacement u satisfies $-(EAu')' = f$, $u(0) = 0$ and $EAu'(L) = t$. The product EA is positive. A trial profile assembled from linear pieces has jumps in its derivative, so differentiating it twice pointwise is unsuitable. Instead multiply the strong equation by a test displacement v vanishing at the fixed end and integrate by parts once.

The resulting equation is

$$\int_0^L EA\, u'v'\, dx = \int_0^L fv\, dx + tv(L).$$

Only first derivatives are required. The fixed displacement is imposed directly on trial and test spaces; the end traction appears in the load functional. This clarifies why prescribing both force and displacement independently at the same end can overconstrain a second-order problem. The test functions represent allowable virtual displacements, and the integral identity expresses equality of internal and external virtual work.

Background · Why can the derivative jump while the displacement stays continuous?

At a bonded interface with no concentrated force, displacement and axial force are continuous. If EA jumps, the strain u' can jump inversely even though the product EAu' is continuous. For a continuous piecewise linear displacement, the weak first derivative is the ordinary piecewise constant slope. The squared-gradient energy is therefore finite.

A displacement jump is different. Replacing a jump of size d by a straight ramp over width δ costs strain energy $EAd^2/(2\delta)$ for constant positive stiffness, which diverges as the ramp shrinks. A standard conforming displacement space encodes this distinction. If the mechanics really allow slip, the interface needs its own model and a space which permits the two traces to differ.

Assemble one element and connect it to its neighbour

On an element from x_i to x_{i+1} of length h, use shape functions $N_1 = 1 - s,\ N_2 = s$ with $s = (x - x_i)/h$. Their physical derivatives are $-1/h,\ 1/h$. For constant EA, integrating products $EAN_a'N_b'$ gives the matrix $(EA/h)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$. A constant distributed load contributes $fh(1,1)^T/2$. Summing element contributions for a shared global node enforces balance there without prescribing the internal traction manually.

Take $L = 1,\ EA = 1,\ f = 2,\ t = 0$ in consistent units and use two equal elements. After eliminating the fixed displacement, the stiffness matrix is $\left(\begin{smallmatrix} 4 & -2 \\ -2 & 2 \end{smallmatrix}\right)$ and the load is $(1, 0.5)^T$. Solving gives

$$u(0.5) = 0.75,\ u(1) = 1.$$

The exact solution is $2x - x^2$ and agrees at both nodes. Inside the first element at $x = 0.25$, the linear interpolant gives 0.375, whereas the exact value is 0.4375. Nodal agreement can coexist with a substantial interior error.

Assemble a physical triangle from a reference element

A line element has one length scale. A triangle can stretch, rotate and shear, so the element area alone cannot determine its diffusion matrix. Start with the reference vertices $(0,0), (1,0), (0,1)$ and the linear shape functions $1 - \xi - \eta, \xi, \eta$. Send the reference triangle to a physical cell with vertices P_0, P_1, P_2 by $x = P_0 + J(\xi, \eta)^T$, where the columns of J are $P_1 - P_0$ and $P_2 - P_0$. For a nondegenerate triangle, the chain rule and area transformation give

$$\nabla_x N_i = J^{-T} \nabla_{\xi,\eta} \widehat{N}_i,$$
$$dx\, dy = |\det J|\, d\xi\, d\eta.$$

To see why the inverse transpose occurs, take an arbitrary displacement $d\widehat{x}$. Both gradients must describe the same change in a shape function. Thus $(\nabla_x N_i)^T J\, d\widehat{x} = (\nabla \widehat{N}_i)^T d\widehat{x}$ for every such displacement. For unit isotropic conductivity, the transformed gradients are constant and the physical stiffness entries reduce to

$$K_{ij} = |T| \nabla N_i \cdot \nabla N_j,$$

with $|T| = |\det J|/2$. A constant cell load f contributes $f|T|/3$ at each vertex because each linear shape function has mean $1/3$.

A sheared cell and an exact affine-field check

Take $P_0 = (0,0), P_1 = (2,0), P_2 = (1,1)$. The columns of J are $(2,0)$ and $(1,1)$, so the cell area is 1. Applying J^{-T} to the three reference gradients gives $(-1/2, -1/2), (1/2, -1/2), (0,1)$. Their dot products and the load $f = 3$ yield

$$K = \begin{pmatrix} 1/2 & 0 & -1/2 \\ 0 & 1/2 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix}, \qquad b = (1,1,1)^T.$$

Every row sums to zero because a constant field has zero gradient. For a more discriminating test use $u = x + 2y$, with vertex values $U = (0,2,3)^T$. Weighting the three gradients by these values gives $(1,2)$. The physical integral of its squared gradient equals $1(1^2 + 2^2) = 5$, and direct multiplication gives $U^T K U = 5$. A transposed mapping error can pass the constant test and fail this second one.

Practise on sheet 1 · Two triangles of the same area. Further reading · Heidelberg University · Lecture Notes on Scientific Computing with Partial Differential Equations · §7.4 P_k Finite Elements and §7.6 Construction of the Finite Element Stiffness Matrix.

Identify the energy that is best approximated

Let $a(u,v) = \int EA\, u'v'$ and denote the external work by $\ell(v)$. The discrete solution u_h satisfies $a(u_h, v_h) = \ell(v_h)$ for every test function from its finite-dimensional space. The continuous solution satisfies the same equation for every such test, so $a(u - u_h, v_h) = 0$. This Galerkin orthogonality means the error is perpendicular to the trial space in the energy inner product. For any other admissible discrete w_h, write $u - w_h = (u - u_h) + (u_h - w_h)$ and expand its squared energy norm. The cross term vanishes and $\|u - w_h\|_a^2 = \|u - u_h\|_a^2 + \|u_h - w_h\|_a^2$.

Thus the conforming Galerkin solution is the best energy approximation in its space for this symmetric linear problem and exact integration. The statement concerns a specific norm and model. It does not guarantee pointwise optimal stresses or carry unchanged into nonsymmetric transport problems. In the rod example the computed strain is constant on each element, replacing a linearly varying exact strain. Refinement improves this derivative field even when a table of nodal displacements already looks exact.

Recover a support reaction

The reaction at a prescribed-displacement node is hidden if one inspects only the reduced interior system. Assemble all element internal forces Ku_h before discarding constrained rows, then subtract the full external load vector. The entries at free degrees of freedom should be near zero; the entry at a prescribed degree of freedom gives the reaction in the chosen sign convention.

In the uniformly loaded rod, the first element strain is

$$(0.75 - 0)/0.5 = 1.5.$$

Its internal nodal force at the left is -1.5, while the first node receives an external distributed-load share of $+0.5$. Their difference -2 is the support reaction. It balances the total applied load $2 \cdot 1 = 2$. Looking only at the derivative in the first element would have reported magnitude 1.5, missing that load share. A reaction from assembled balance and a stress recovered by pointwise differentiation are related but distinct quantities on a finite mesh.

Obtain actual norm errors from the Galerkin property

For $-u'' = f$ on the unit interval with homogeneous Dirichlet values, the elementary energy inner product is $a(u,v) = \int u'v'$. Let $I_h u$ be the continuous piecewise-linear interpolant of a sufficiently smooth exact solution at mesh nodes. On each element its slope is the average of u'. Every linear finite-element test function v_h has constant derivative on that element. Consequently

$$\int_E (u - I_h u)' v_h' = v_h' [u - I_h u]_{\partial E} = 0.$$

Summing proves global energy orthogonality.

The Galerkin solution satisfies the same orthogonality, and positive definiteness gives uniqueness. Thus $u_h = I_h u$ in this one-dimensional, constant-coefficient problem with exactly integrated load and matching boundary values. Nodal errors vanish, although the field and derivative between nodes can still be inaccurate. Variable coefficients inside an element generally destroy the simple telescoping argument because the factor multiplying v_h' is no longer constant. The result also should not be transferred to multidimensional triangulations without proof.

Compare nodal, field and flux accuracy for one solution

Choose $u = x(1 - x)$, so $f = 2$. On an element starting at a write $s = x - a \in [0, h]$. Subtract the straight interpolant from the parabola to get $u - u_h = s(h - s)$, vanishing at both nodes and attaining $h^2/4$ at the element centre. Its derivative is $h - 2s$; the exact and numerical slopes agree at the midpoint but differ by magnitude h at each endpoint taken from that element. Testing only the midpoint slope or the nodal values would entirely conceal these other discrepancies.

Compute the squared energy error on this element:

$$\int_0^h (h - 2s)^2 ds = h^3/3.$$

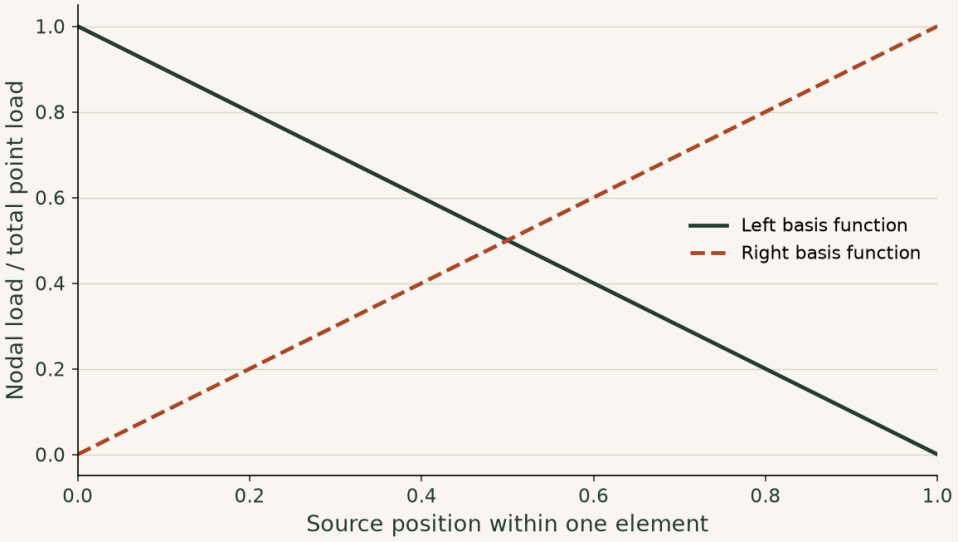
With $1/h$ equal elements, the global energy error is $h/\sqrt{3}$. Similarly the element squared field error is $\int_0^h s^2 (h - s)^2 ds = h^5/30$, giving global L^2 error $h^2/\sqrt{30}$. For $h = 1/4$, the two errors are approximately 0.14434 and 0.01141; the maximum field error is 0.015625. Halving the mesh width halves the energy error but divides the field error by four. All these measurements concern the very same numerical solution whose nodal error is exactly zero.

For a physical diffusion flux with constant conductivity one, flux error equals minus derivative error. A cell-averaged flux test is also special here: the numerical slope equals the exact average slope on each element for the same reason used in the proof. To verify local flux resolution, compare at more than a midpoint or a cell average and retain the measure used for convergence. A reported order is a statement about an observable and a norm as well as a finite-element space.

These formulas also identify a misleading manufactured test: if an automated comparison stores only nodal differences, the estimated order is undefined because both coarse and fine errors vanish. Adding independently evaluated quadrature points inside elements exposes the field error immediately. To check a flux implementation, use a quadrature rule capable of resolving the squared linear derivative error; a one-point midpoint rule would again report zero despite the nonzero exact energy norm.

Distribute a moving point source in a weak form

A one-dimensional Poisson problem has a concentrated source $Q\delta(x-a)$. Its weak load against a linear hat function is $Q\phi_i(a)$, so the two adjacent degrees of freedom share it in proportion to their shape functions.



At 30% of the element length the split is 70% left, 30% right. The loads sum to Q and reproduce the correct first moment. Placing the whole source at the nearest node preserves total load but jumps abruptly at mid-element.

▶ **Read values from the plot**

Change parameters · Distribute a moving point source in a weak form

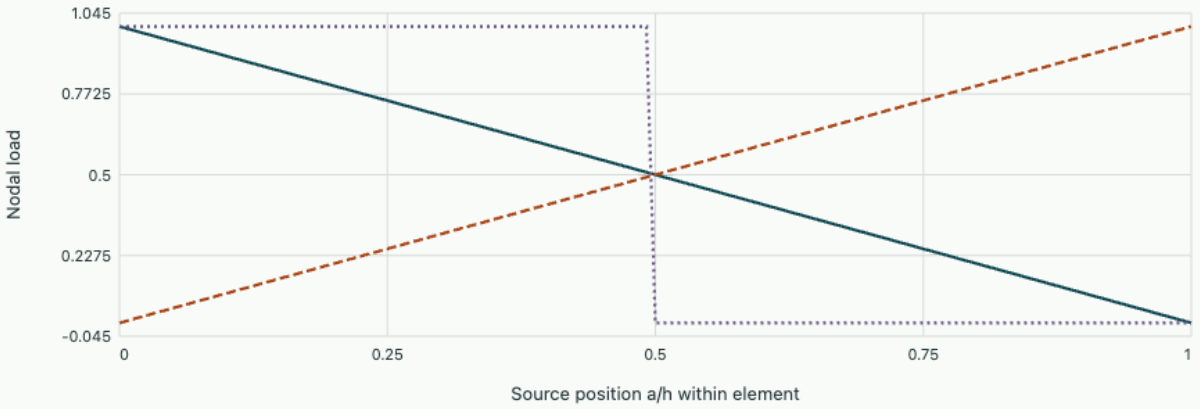
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Source strength Q **1**

Selected element coordinate a/h **0.3**



Selected loads 0.7 and 0.3 sum to Q = 1 and reproduce the source first moment. Nearest-node placement preserves only the total.

Study the full derivation

Bound diffusion error by balancing a reconstructed flux · Spatial discretisation and field equations

A temperature field from a finite-element calculation looks smooth. Reconstruct a heat flux which exactly balances the source, then derive an upper bound for the true energy error without knowing the exact temperature.

Problem sheet 1 · Flux balances and virtual work

Choose the correct conductance at an interface, test pure-flux compatibility and assemble a rod with unequal stiffnesses.

Preserve heat flow across unequal layers

A wall has two layers, 0.02 m at $k = 2$ W/(m·K) and 0.06 m at $k = 12$ W/(m·K). The outer faces are held at 350 K and 320 K, respectively. Find the common flux, the interface temperature and the effective conductivity over the full thickness.

▼ **Worked solution**

The resistances per unit area are $0.02/2 = 0.01$ and $0.06/12 = 0.005$ m²·K/W. Total resistance is 0.015, so the flux magnitude is $30/0.015 = 2000$ W/m² from hot to cold. The first drop is $2000(0.01) = 20$ K, giving an interface temperature of 330 K.

The effective conductivity is $0.08/0.015 = 5.333$ W/(m·K). Equal-width averaging would ignore the different thicknesses. Substituting this conductivity across the full wall reproduces the same flux.

Explore the related calculation with adjustable inputs. Vary series resistances and observe their fractions of the total drop. Divide by the stated bulk resistance when converting dimensional values to a comparison.

Check a pure-flux problem

On $0 \le x \le 2$, the steady flux obeys $q' = 3$ in consistent units. The prescribed endpoint fluxes, both positive along increasing x, are $q(0) = 2$ and $q(2) = 7$. Is a steady state possible? Correct the right flux. For $k = 1$, find a compatible temperature with mean temperature zero.

▼ **Worked solution**

Integrating requires

$$q(2) - q(0) = \int_0^2 3\,dx = 6.$$

The supplied difference is 5, so no steady solution satisfies both data. Take $q(2) = 8$, giving $q = 2 + 3x$.

Because $q = -T'$, $T = C - 2x - 1.5x^2$. Its average over the interval is $C - 2 - 2 = C - 4$, hence $C = 4$. The mean fixes the otherwise arbitrary temperature offset while leaving flux unchanged.

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.

Assemble two rods in series

Two equal elements of unit length have $EA = 2$ and $EA = 6$ in consistent units. Fix the left node, apply a unit tensile load on the right and omit distributed load. Assemble the reduced matrix, solve for displacements and verify the axial forces and strain energy.

▼ **Worked solution**

Assembly gives

$$K = \begin{pmatrix} 8 & -6 \\ -6 & 6 \end{pmatrix},$$

$$b = (0, 1)^T.$$

The second row gives $u_2 - u_1 = 1/6$. The first then gives $2u_1 = 1$, hence

$$u_1 = 1/2,\ u_2 = 2/3.$$

The forces are $2(1/2) = 1$ and $6(1/6) = 1$. The stored energies sum to $2(1/2)^2/2 + 6(1/6)^2/2 = 1/3$, also $u^T b/2 = (2/3)/2$. This verifies both continuity of axial force and consistency of the assembled work.

Explore the related calculation with adjustable inputs. Check a small conservative conductance system under a changed load; the same assembled balance form applies to the rod stiffness calculation.

A longer rod with a clamped or insulated left end

For a rod of length 2, use $k = 4$, a unit distributed load, zero temperature on the left and insulation on the right. Compute the exact derivative norm and its coercivity bound. Then insulate the left end too. Which compatibility condition fails, and what further information would be needed if it held?

▼ **Worked solution**

The flux condition gives $u' = (2 - x)/4$, so $\|u'\|_2^2 = (1/16)\int_0^2 (2 - x)^2 dx = 1/6$, giving 0.40825. The load norm is $\sqrt{2}$. The estimate gives

$$2\sqrt{2}/(\sqrt{2}\,4) = 0.5.$$

The anchored solution therefore satisfies the bound.

Testing a two-insulated-end balance with the constant test function 1 requires

$$\int_0^2 f\,dx = 0.$$

The present load integrates to 2. A zero-net load would restore compatibility, but $u + C$ would solve for every constant C. A prescribed mean temperature, for example, selects one representative.

Explore the related calculation with adjustable inputs. Change the propagation length scale and mesh resolution to compare shooting sensitivity with a boundary-value discretisation.

Revisit chapter 1 · Why fixing one boundary value controls the solution.

Two triangles of the same area

Stretch the reference triangle to vertices $(0, 0), (4, 0), (0, 1/4)$. Find the gradients associated with its two axis vertices and their diagonal stiffness entries. Compare their ratio with the reference triangle, and explain why a fixed area does not ensure comparable directional stiffness.

▼ Worked solution

Here $J = \mathrm{diag}(4, 1/4)$ has determinant one. The physical gradients are $(1/4, 0)$ and $(0, 4)$. Multiplying their squared lengths by the cell area $1/2$ gives $K_{11} = 1/32$ and $K_{22} = 8$.

The ratio $K_{22}/K_{11} = 256$. On the reference triangle both entries are $1/2$. A unit change in nodal value over the short altitude gives a much larger derivative. Area fixes the product of stretches, while gradient amplification depends on inverse stretches in each direction. This ratio describes these entries. The unrestricted element also has a constant null mode.

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

Revisit chapter 2 · Assemble a physical triangle from a reference element.

Chapter 3 · Finite volumes and moving fronts

Derive an upwind transport update by conservation, interpret the Courant number as a fraction of cell contents and determine what conservation says about front speed.

FROM FIRST PRINCIPLES

Derive a conservative update from storage in a cell

Consider the one-dimensional conservation law $u_t + \partial_x f(u) = 0$ with no distributed source. Let $\bar{u}_i$ denote the average over cell i of width Δx. The integral of u over the cell is what the discrete variable stores. Integrate first over the cell and then over one time step. Only fluxes crossing the two faces can change its inventory:

$$\Delta x(\bar{u}_i^{n+1} - \bar{u}_i^n) = -\int_{t_n}^{t_{n+1}} [f_{i+1/2}(t) - f_{i-1/2}(t)]\, dt.$$

Replace each time-averaged face flux by one numerical function $\hat{f}$ of neighbouring states, using the same value from both sides. For constant positive advection speed a, the state reaching a face comes from upstream. With piecewise-constant cell data and $C = a\Delta t/\Delta x \le 1$ this gives

$$\hat{f}_{i+1/2} = a\bar{u}_i^n,$$
$$\bar{u}_i^{n+1} = (1 - C)\bar{u}_i^n + C\bar{u}_{i-1}^n.$$

The update conserves total mass on a periodic grid because every internal face cancels. Nonlinear fluxes also need a rule selecting the physically admissible interface state, especially when a jump separates two different initial cell values.

Background · Courant number

$$C = |a|\Delta t/\Delta x$$

For scalar advection this compares a transported distance with a cell width. An admissible range depends on the spatial method and time integrator jointly.

DERIVATION AND APPLICATIONS

Quantify the quadratic quantity lost by an upwind update

A conservative finite-volume method for advection preserves the sum of cell contents under periodic boundary conditions. It need not preserve their contrast. Consider the constant-speed transport equation $u_t + au_x = 0$ with $a > 0$, uniform cells of width Δx and periodic indexing. Exact translation

preserves $\int u^2\,dx$. The first-order upwind scheme transports a fraction $C = a\Delta t/\Delta x$ of a cell's content across its right boundary per step:

$$u_i^{n+1} = (1 - C)u_i^n + Cu_{i-1}^n.$$

For $0 \le C \le 1$ this is a convex combination and cannot create a value outside its two inputs. To see precisely how much variation is removed, square one update and complete a difference square:

$$[(1 - C)x + Cy]^2 = (1 - C)x^2 + Cy^2 - C(1 - C)(x - y)^2.$$

Sum over all cells. Periodicity allows the shifted sum of old squared values to be relabelled. The exact discrete identity, with $E^n = (\Delta x/2)\sum_i (u_i^n)^2$, becomes

$$E^{n+1} - E^n = -\frac{\Delta x}{2} C(1 - C) \sum_i (u_i^n - u_{i-1}^n)^2.$$

This loss vanishes for spatially constant data and for $C = 1$, when the update shifts by one complete cell. Strong cell-to-cell jumps dissipate much more of the quadratic quantity than smooth profiles sampled with the same grid. Total mass still telescopes exactly; the two balance laws diagnose different aspects of accuracy.

WORKED EXAMPLE

Follow one step on four periodic cells

Take $\Delta x = 1$, $C = 1/2$ and initial averages $(1, 1, 0, 0)$. The updated values are $(0.5, 1, 0.5, 0)$. Both arrays sum to two. The quadratic energy falls from 1.00 to 0.750. There are two unit jumps at the start, and the loss formula gives $(1/2)(1/2)(1/2)(2) = 0.250$, accounting for the full change.

SECOND APPLICATION

Use an alternating pattern as an instability detector

For $(1, -1, 1, -1)$ each neighbour has opposite sign, so one step multiplies the amplitude by $1 - 2C$. At $C = 1/2$ all variation disappears in one step although the exact wave would keep its amplitude. At $C = 1.10$ the factor is -1.20 and energy grows by 1.44. The sign reversal in $C(1 - C)$ therefore predicts growth at this resolved wavelength. Conservation of the zero mean still holds in both experiments and alone cannot detect the instability.

Start from this chapter. Construct a shared flux on each interior face so neighbouring cell balances use opposite signs. Examine how resolving the source wavelength changes the computed field at fixed residual tolerance.

> OneSimulation laboratory · Measure spatial convergence on a known field

Store quantities over cells

Consider a passive concentration satisfying $u_t + (au)_x = 0$ with positive constant speed a on a periodic line. Let $\bar{u}_i = h^{-1} \int_{x_{i-1/2}}^{x_{i+1/2}} u \, dx$ be its cell mean. Integrating the conservation law in space and time shows that every change in cell inventory comes from fluxes through its two faces. If neighbouring cells use the same face flux, summing all updates cancels every internal transfer exactly. On a periodic line the total stays fixed, even for a discontinuous concentration.

For positive speed, material reaching the right face during a short step comes from the cell to its left. Approximate the outgoing state by that cell's mean, giving $F_{i+1/2} = a\bar{u}_i$. Forward Euler in time yields

$$\bar{u}_i^{n+1} = (1 - C)\bar{u}_i^n + C\bar{u}_{i-1}^n,$$

where $C = a\Delta t/h$. For $0 \le C \le 1$ this is a convex average, preserving the range of neighbouring values. At $C = 1$ the entire array shifts exactly by one cell per step. If $a < 0$, information comes from the other side of the face and the upwind direction must reverse.

Identify the source of artificial spreading

Start with one cell mean equal to one and zeros everywhere else. With $C = 1/2$, one step gives adjacent means $1/2, 1/2$. A second step gives $1/4, 1/2, 1/4$, while the total remains one. Exact advection over two such steps shifts the original pulse by one cell with no change in shape. The scheme has conserved mass yet broadened the pulse. Peak height or variance therefore tests information not supplied by a global balance.

For smooth u, Taylor-expand the upwind update in space and time. Replace the leading u_{tt} using the exact transport equation, $u_{tt} = a^2 u_{xx}$. The leading modified equation is

$$u_t + au_x = [ah(1 - C)/2]u_{xx} + \cdots.$$

The bracket acts as numerical diffusivity. It is positive for the monotone range and vanishes at $C = 1$. This local smooth expansion helps interpret observed spreading, but the Taylor expansion itself does not hold across a jump.

Choose a flux when speed depends on the state

For a nonlinear conservation law $u_t + f(u)_x = 0$, characteristics can intersect and create a discontinuity. Enclose a moving jump in a thin control volume. The net flux imbalance fixes its speed $s = (f(u_R) - f(u_L))/(u_R - u_L)$. For Burgers flux $f = u^2/2$ with left state 2 and right state 0, this gives $s = 1$. The faster characteristics on the left overtake the slower ones to form a shock. Reversing

these states produces a spreading fan under the entropy condition even though the algebraic jump formula alone would still supply a speed.

A numerical face flux must be consistent when the two cells agree and must incorporate information from both sides when a nonlinear wave can travel either way. A conservative method that converges can reproduce the weak jump balance, while selection of a physically admissible solution also requires an entropy criterion. For practical assessment, monitor total inventory, motion of the front, overshoot and the number of cells over which it is smeared. The four measurements address distinct properties, and reducing one defect need not reduce all of them.

Move the control surface instead of the grid

Across a front moving with speed s, flux in the front's own frame is $f(u) - su$. The jump relation enforces equality of this quantity on both sides. For the Burgers jump from 2 to 0 with speed one, the left flux in the moving frame is $2 - 1 \cdot 2 = 0$, and the right flux is also zero. The laboratory flux differs across the front because a moving interface is sweeping conserved material through stationary space.

This distinction helps when checking a simulation by counting material in a region that tracks a front. Differentiating an integral over moving endpoints adds boundary-motion terms. In a periodic fixed window the update should conserve total amount without those extra terms. In a moving measurement window, physical transport through the window and motion of the window both influence the recorded inventory, even under a perfectly conservative scheme.

Distinguish a conservative jump from the physical weak solution

For inviscid Burgers flow $u_t + (u^2/2)_x = 0$, integrating conservation across a moving jump gives the Rankine–Hugoniot speed $s = [f]/[u] = (u_L + u_R)/2$. It uses only balance, so a discontinuity with that speed satisfies the weak conservation equation for either orientation of its left and right states. To select the physically relevant solution one needs an entropy condition. For a compressive shock characteristic speeds enter it from both sides: $u_L > s > u_R$. If $u_L < u_R$, the characteristics separate and a continuous rarefaction fan fills the gap.

For example the entropy $\eta(u) = u^2/2$ has entropy flux $q(u) = u^3/3$, because $q' = \eta' f'$. A jump produces distributional entropy $[q] - s[\eta]$, which must be nonpositive for a dissipative solution. At equal-magnitude opposite states the entropy jump $[\eta]$ is zero, so orientation alone changes the sign of the entropy production. A conservative scheme can preserve every cell total correctly yet converge to an inadmissible expansion shock if it fails this additional condition.

Choose the correct interface flux for two opposite Riemann data

Take $u_L = -1,\ u_R = 1$. Both endpoint fluxes equal one half, and the jump-speed formula gives zero. Holding this discontinuity fixed would conserve the cell totals. Its entropy production, however, is

$$1/3 - (-1/3) = 2/3 > 0.$$

The admissible Riemann solution is the self-similar fan: at coordinate $\xi = x/t$, $u = -1$ for $\xi \le -1$, $u = \xi$ between minus one and one, and $u = 1$ for $\xi \ge 1$. At the interface $\xi = 0$, the flux is therefore zero. Using an arithmetic average of the two endpoint fluxes would instead maintain the inadmissible jump.

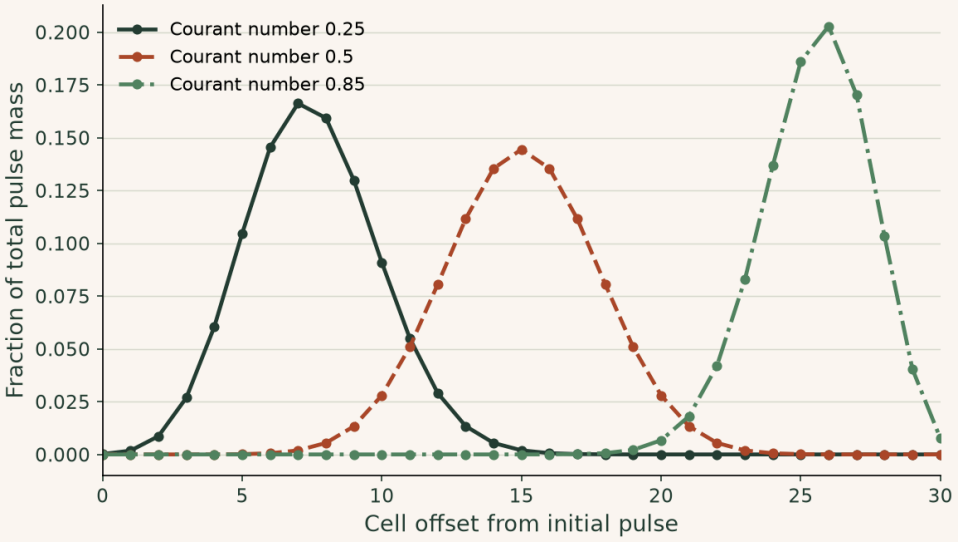
For a finite-volume update with $\Delta t/\Delta x = 0.5$ and initially constant states on either side, the neighbouring left cell changes from -1 to $-1 - 0.5(0 - 0.5) = -0.75$. The right cell changes from one to $1 - 0.5(0.5 - 0) = 0.75$. Their sum remains zero. A direct integration of the fan gives the same cell averages: the disturbance occupies half of each cell and linearly brings that half from its far-state value toward zero. Thus both conservation and the local wave pattern can be verified independently.

Reverse the initial states to $u_L = 1,\ u_R = -1$. The zero-speed jump now has entropy production $-2/3$, and both families of characteristics enter it. Its correct interface flux is one half. The same numerical update then leaves both neighbouring cell averages unchanged, exactly as required for a stationary shock. Testing both orientations is more revealing than testing a single conserved jump: they have identical endpoint flux magnitudes and the same Rankine–Hugoniot speed but different physically admissible interface states.

A further integration test shifts the interface by precisely one cell and reruns the update. On a uniform periodic grid with unchanged physical states, the new solution should be a one-cell translation of the old one. A difference between the two calculations could reveal an indexing error in the face flux even when a global conservation diagnostic remains identically zero.

Measure numerical spreading in a conservative transport step

Start with all mass in one cell. For positive velocity the explicit upwind update keeps fraction $1 - C$ and transfers fraction C to the next cell each step. Thirty updates produce a binomial distribution on an unbounded uniform grid.



Each profile sums to one. Its centre moves 30C cells and its variance is $30C(1-C)$ cells^2. Conservation and the correct mean travel distance can coexist with artificial broadening of a pulse that should translate without changing shape.

▶ **Read values from the plot**

Change parameters · Measure numerical spreading in a conservative transport step

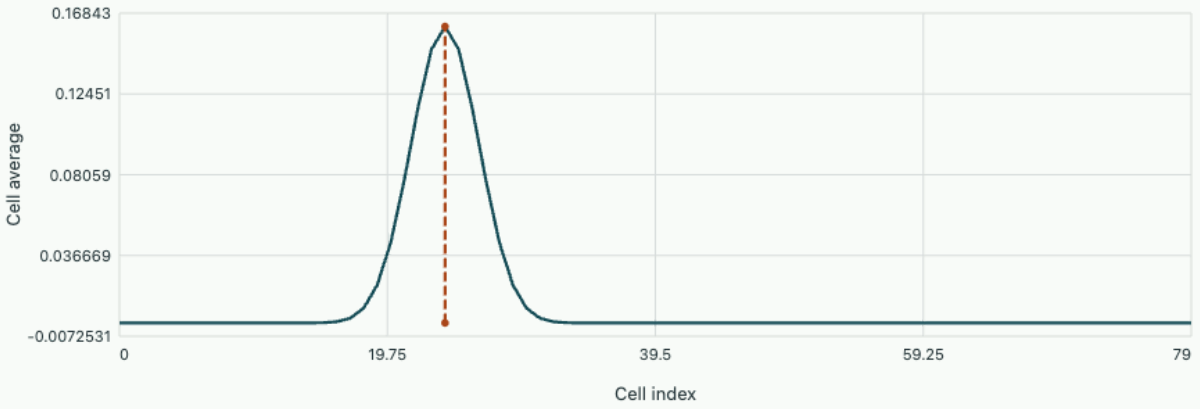
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Courant number C **0.5**

Number of steps **24**



Total cell sum 1, centroid 24, added variance 6 cells².

Study the full derivation

Choose the physically admissible wave in a conservative update · Spatial discretisation and field equations

A discontinuous scalar flow can satisfy integral conservation in more than one way. Derive the shock speed, use a convex entropy to distinguish a compression from an expansion and build a finite-volume flux from the resulting local solution.

Chapter 4 · Time-dependent diffusion

Derive the diffusion time-step restriction from two independent arguments and compare explicit and implicit decay of spatial modes.

FROM FIRST PRINCIPLES

Derive the explicit heat step and its grid-dependent time scale

Take the heat equation $T_t = \alpha T_{xx}$ with constant positive diffusivity on a uniform one-dimensional grid. Fourier conduction through each side of a cell produces a temperature update proportional to the difference of the two neighbour gradients. With forward Euler timing, collect those fluxes to obtain

$$T_i^{n+1} = T_i^n + r(T_{i-1}^n - 2T_i^n + T_{i+1}^n),$$
$$r = \alpha \Delta t / h^2.$$

When $0 \le r \le 1/2$ the new temperature is a convex average of three old ones, preserving bounds on a source-free grid. To check whether a higher step could still be stable for all periodic grid disturbances, substitute $T_j^n = g^n e^{ij\theta}$, where j is the node index. The amplification simplifies to

$$g = 1 + r(e^{-i\theta} - 2 + e^{i\theta}) = 1 - 4r \sin^2(\theta/2).$$

The alternating mode at $\theta = \pi$ gives $g = 1 - 4r$, so $|g| \le 1$ again requires $r \le 1/2$. Halving spatial spacing therefore quarters the largest explicit time step for this discretization. Implicit stepping changes this restriction but still has to resolve any transient relevant to the measurement.

DERIVATION AND APPLICATIONS

Can an energy-stable heat step create an alternating transient?

A diffusion mode should decay rather than oscillate. To distinguish decay of a norm from preservation of sign, examine the finite-element or finite-volume semidiscrete system $M\dot{u} + Ku = 0$, with symmetric positive definite mass matrix M and thermal stiffness K positive semidefinite. An eigenvector satisfying $Kv = \lambda M v$ has scalar amplitude $y(t)$ obeying $y' = -\lambda y$. A generalized θ step applied to this mode gives

$$[1 + \theta z] y^{n+1} = [1 - (1-\theta) z] y^n,$$
$$z = \Delta t \, \lambda.$$

For Crank–Nicolson take $\theta = 1/2$. The factor $G = (1 - z/2)/(1 + z/2)$ has $|G| \le 1$ for all $z \ge 0$ but changes sign when $z > 2$. An initial grid-scale component can therefore alternate in time. To test a remedy, replace the first full step by two half backward-Euler steps. Their combined factor is

$$G_{\text{start}} = (1 + z/2)^{-2}.$$

Thereafter return to the second-order Crank–Nicolson update. After n further full steps the mode amplitudes, with and without the initial damping, are

$$y_{\text{damped}} = G_{\text{start}} G^n y^0,$$
$$y_{\text{plain}} = G^{n+1} y^0.$$

For small z, both startup strategies agree through first order, while for large z the two half steps strongly suppress the unresolved mode. This argument explains a startup treatment for nonsmooth initial data; sign preservation of every physical grid value is a stronger matrix property than sign preservation of one mode.

WORKED EXAMPLE

Check damping of a near-grid-scale eigenmode

Take $\lambda = 2000$ s⁻¹ and $\Delta t = 0.0100$ s. Then $z = 20$. A single Crank–Nicolson step leaves $-9/11 = -0.818$ of the amplitude; two backward-Euler half steps leave $1/121 = 0.00826$. The exact amplitude factor is

$$e^{-20} = 2.06 \times 10^{-9}.$$

Both discrete results are bounded, but the startup treatment removes almost all of a mode that the time grid cannot resolve.

SECOND APPLICATION

Check the cost for a physically slow component

For the same mesh and step, a slow mode with $\lambda = 2.00$ s⁻¹ has $z = 0.0200$. Crank–Nicolson gives $0.99/1.01 = 0.980198$, the startup gives $1/1.01^2 = 0.980296$ and the exact factor is 0.980199. The slow component changes by less than 10^{-4} in this comparison. A timestep must still resolve a transient that is itself part of the desired output.

Discretise space and time separately

A heat pulse in a thin insulated rod smooths over time according to $u_t = Du_{xx}$ for constant positive diffusivity D. Use a uniform grid of spacing h, and initially examine interior points or periodic boundaries so the stencil is unchanged everywhere. Replacing the second derivative by a centred difference gives a system of ordinary differential equations for the nodal temperatures. The time method still has to resolve the fastest rate that this spatial operator can generate.

An explicit Euler step yields $u_i^{n+1} = r u_{i-1}^n + (1-2r) u_i^n + r u_{i+1}^n$ with $r = D\Delta t / h^2$. If $0 \le r \le 1/2$, the update is a convex combination of three temperatures. The maximum cannot rise and the minimum cannot fall in a source-free interior step. Starting with one hot node at 1 and neighbours at zero, $r = 0.4$ gives a centre value 0.2 and neighbours 0.4 after a single step. At $r = 0.6$ the centre would become negative although the continuum heat equation preserves nonnegativity.

Identify the spatial modes that set the limit

Insert $u_j^n = g^n e^{\mathrm{i} j \theta}$, where i is the imaginary unit and j indexes the grid. The two neighbouring values contribute factors $e^{\pm \mathrm{i}\theta}$, yielding $g = 1 - 4r \sin^2(\theta/2)$. Long waves have $\sin^2(\theta/2) \simeq \theta^2/4$, so their decay approaches the continuum rate as the mesh is refined. The alternating grid mode $\theta = \pi$ has the most negative amplification and enforces $r \le 1/2$ for bounded magnitude.

At the limiting value $r = 1/2$ the alternating mode flips sign without losing amplitude. A strict margin damps it. For $D = 10^{-4}$ m²/s and $h = 2$ mm, the maximum explicit step is

$$h^2/(2D) = 0.020\,\mathrm{s}.$$

Refining to 1 mm reduces that limit to 0.005 s. The mesh gives a sharper temperature profile yet can cost four times more steps over the same physical interval, besides the increased number of spatial degrees of freedom.

Compare backward Euler and Crank–Nicolson

Writing the semidiscrete system as $u' = -A_h u$, backward Euler solves $(I + \Delta t A_h) u^{n+1} = u^n$. Crank–Nicolson averages the old and new diffusive fluxes, giving $(I + \Delta t A_h/2) u^{n+1} = (I - \Delta t A_h/2) u^n$. For a Fourier mode put $s = 4r \sin^2(\theta/2)$. Their factors are $1/(1+s)$ and $(1 - s/2)/(1 + s/2)$, respectively. Both remain bounded at arbitrarily large s, but Crank–Nicolson may alternate for under-resolved sharp features.

With $s = 20$, backward Euler retains only $1/21$ per step. Crank–Nicolson yields $-9/11$, retaining much more of the fast component and reversing its sign. For a smooth low-frequency mode Crank–Nicolson generally gives more accurate temporal decay because it is second order. Decide whether early sharp transients or later smooth behaviour is the relevant physical quantity. To identify spatial and temporal error separately, refine Δt first on a fixed mesh, then refine space with time error kept well below spatial error. Merely halving both at once cannot reveal which term dominated the original discrepancy.

Apply a prescribed heat input without confusing it with temperature

An insulated periodic rod with uniform heating rate s measured in K/s has an average temperature that must rise by $s\Delta t$ each step. Sum the centred second differences over the periodic indices: each nodal

value appears with coefficients $1, -2, 1$, which cancel. The source alone changes the mean. For $s = 3$ K/s and a step of 0.02 s the numerical average must increase by 0.06 K, independently of the diffusivity and initial profile.

A code can pass this global test while still mishandling short-wave damping, since an alternating error has zero spatial mean. Pair a mean-temperature test with a nonconstant mode-decay test. Use a homogeneous periodic run for isolates of diffusion, and a uniform temperature with prescribed source for the storage term. A failure in one test then identifies a different part of the mathematical balance than a failure in the other.

Prove decay of a discrete thermal energy

Let a closed semidiscrete diffusion model be $M\dot{u} + Ku = 0$, with M symmetric positive definite and K symmetric positive semidefinite. Assume $K\mathbf{1} = 0$, expressing zero flux for a uniform temperature. Backward Euler satisfies $M(u^{n+1} - u^n) + \Delta t K u^{n+1} = 0$. Multiplying by the constant vector and using symmetry gives

$$\mathbf{1}^T M u^{n+1} = \mathbf{1}^T M u^n.$$

Thus the capacity-weighted sum, proportional to stored sensible heat for constant capacities and a fixed reference, is conserved for any step size.

Now multiply by u^{n+1}. The identity $a^T M(a-b) = [\|a\|_M^2 - \|b\|_M^2 + \|a-b\|_M^2]/2$ yields the exact balance

$$\tfrac{1}{2}\|u^n\|_M^2 - \tfrac{1}{2}\|u^{n+1}\|_M^2 = \Delta t\,(u^{n+1})^T K u^{n+1} + \tfrac{1}{2}\|u^{n+1} - u^n\|_M^2.$$

Both terms are nonnegative. The quadratic quantity therefore decays although total sensible heat is conserved. Subtracting the conserved mean first yields the same decay statement for temperature variation around equilibrium. The first term represents spatial smoothing measured at the new time, and the second reflects the numerical time update. This derivation gives a stability certificate on nonuniform grids without assuming Fourier modes.

Audit two cells with different heat capacities

Choose capacities 1 and 3 J/K, conductance 2 W/K, and initial temperature deviations $(4, 0)^T$ K. The conserved sensible heat is 4 J, so the final equilibrium in a closed system must be one kelvin above reference. At $\Delta t = 0.5$ s the matrix $M + \Delta t K$ has rows $(2, -1)$ and $(-1, 4)$, and the right-hand side is $(4, 0)^T$. The new temperatures are $(16/7, 4/7)^T$. Checking their weighted sum gives $16/7 + 3(4/7) = 4$, as required. The unweighted average changes, which is physically appropriate for unequal capacities.

The quadratic quantity at the start is $4^2/2 = 8$. After the step it is

$$[(16/7)^2 + 3(4/7)^2]/2 = 152/49.$$

Its drop is $240/49$. The conductive contribution is $(0.5)(2)(12/7)^2 = 144/49$; the time-update contribution is

$$[(-12/7)^2 + 3(4/7)^2]/2 = 96/49.$$

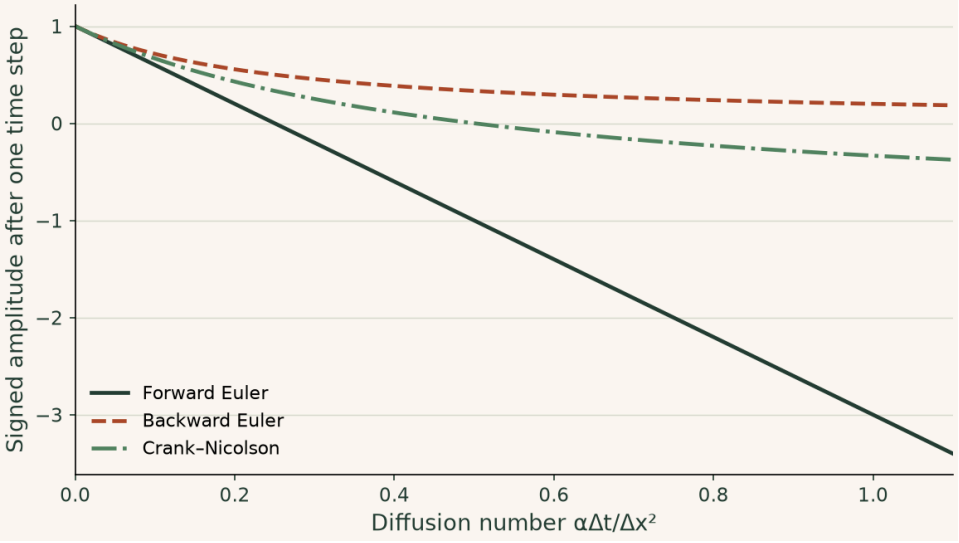
Their sum recovers the measured drop exactly. Neither contribution equals a loss of the 4 J of sensible heat. The quadratic measure depends on squared temperatures and has different physical units.

A positive heat supply or a prescribed exterior temperature introduces a boundary or source work term into these identities. Applying the closed-system decay test to an externally heated experiment would therefore diagnose the wrong physical model. Verify which conserved or dissipated quantity is actually implied by the capacity matrix, boundary conditions and source terms before comparing numerical integrators.

The conserved mean suggests another check after many large steps. The steady state must have zero conductance dissipation, which for two connected cells requires equal temperatures. Together with the weighted heat total this fixes both temperatures at one kelvin above reference. A scheme that instead approaches the unweighted initial average of two would be converging to the wrong conserved quantity despite producing a visually smooth, uniform endpoint. The measured quadratic variation about the correct mean tends to zero; the uncentred squared quantity tends to a positive constant because the remaining uniform state still has temperature above the chosen reference.

What happens to a grid-scale temperature ripple?

The alternating spatial mode on a uniform heat grid has discrete Laplacian eigenvalue $-4/\Delta x^2$. Substitute this mode into each method and compare the amplitude after one step. The plot isolates the shortest resolvable pattern.



Forward Euler reverses the mode above 1/4 and becomes unstable above 1/2. Crank–Nicolson stays bounded for every plotted step but also reverses the mode above 1/2. Smooth initial data can hide this high-frequency diagnostic.

▶ **Read values from the plot**

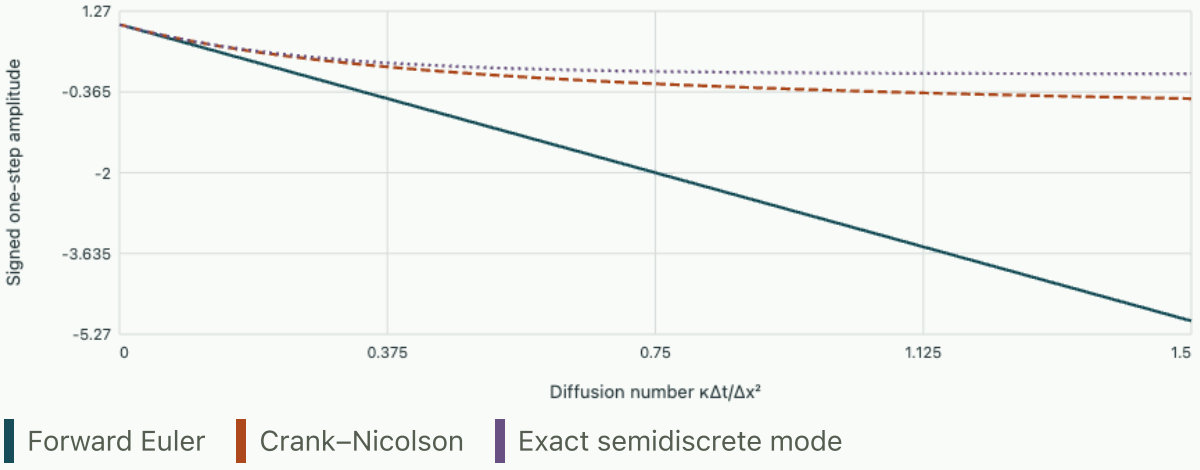
Change parameters · What happens to a grid-scale temperature ripple?

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

Highest diffusion number	**1.5**
Mode fraction of Nyquist	**1**



Directional Laplacian magnitude times Δx^2 is 4. Forward Euler is stable up to r = 0.5 and nonnegative only up to r = 0.25.

Problem sheet 2 · Fronts and diffusive modes

Track conservation on a periodic mesh, distinguish shocks from expanding fans and choose a time step for an explicit diffusion solver.

Advect a marked volume

On a periodic ring of four cells take means $(1, 0, 0, 0)$, positive speed and Courant number $C = 1/4$. Compute two first-order upwind updates, verify inventory and compare the maximum with that from an exact translation of a square pulse by half a cell.

▼ Worked solution

The first update is $(3/4, 1/4, 0, 0)$. Applying the same convex average gives $(9/16, 6/16, 1/16, 0)$. Both arrays sum to one, so the total conserved amount is one cell width times this sum.

After an exact shift of half a cell, the two overlapping cell averages are $1/2, 1/2$ and the other two vanish, although the pointwise pulse height is still one. The numerical maximum of $9/16$ is larger than the exact cell-average peak at this instant, even as the numerical support is more spread out. A peak comparison must specify whether it uses point values or cell means.

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.

Identify a shock and a fan

For $f(u) = u^2/2$, consider left and right states 3 and 1. Compute the discontinuity speed, compare it with the two characteristic speeds and locate its position after 0.4 s from the origin. Reverse the states and identify the edges of the entropy-admissible solution.

▼ Worked solution

Fluxes are 4.5 and 0.5, giving $s = (0.5 - 4.5)/(1 - 3) = 2$. Since the characteristic speeds $f'(u)$ are 3 on the left and 1 on the right, both enter the discontinuity in its moving frame. Its position is $0.4(2) = 0.8$.

With left 1 and right 3 the characteristics move apart and an expansion fan fills the region between $x = 0.4$ and $x = 1.2$. A single jump satisfying the algebraic jump balance would have the wrong entropy behaviour 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.

Allocate a time budget on a refined grid

On a periodic rod $D = 2 \times 10^{-5}$ m²/s and $h = 1$ mm. Find the explicit diffusion step limit. To reach 1 s with $r = 0.4$, how many steps are needed? Repeat after halving spatial spacing and compare the total number of nodal updates on a fixed length.

▼ **Worked solution**

The stability limit is

$$h^2/(2D) = 0.025\,\mathrm{s}.$$

For $r = 0.4$, use $0.4h^2/D = 0.020$ s, or 50 steps. Halving h quarters the step to 0.005 s, giving 200 steps.

There are about twice as many spatial nodes on the refined line. The combination of twice as many nodes and four times as many steps requires about eight times as many local updates, neglecting boundary and setup costs.

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 · Wave speeds and numerical dispersion

Derive the leapfrog dispersion relation, use the Courant condition to distinguish boundedness from accurate travel time and initialise a standing-wave calculation consistently.

FROM FIRST PRINCIPLES

Derive the numerical wave dispersion relation by advancing a single mode

Consider a one-dimensional homogeneous wave with speed c, uniform spatial grid h and time increment Δt. Centred second differences approximate both accelerations and spatial curvatures using the values on either side of their evaluation point. The governing equation $u_{tt} = c^2 u_{xx}$ becomes

$$u_j^{n+1} - 2u_j^n + u_j^{n-1} = C^2(u_{j+1}^n - 2u_j^n + u_{j-1}^n),$$
$$C = c\Delta t/h.$$

To determine what the grid does to one travelling component, insert $u_j^n = e^{i(kjh - \omega n\Delta t)}$ and divide by the value at (j, n). In both directions the pair of opposite phases combines to a cosine. Using $2 - 2\cos z = 4\sin^2(z/2)$ then gives

$$\sin^2(\omega\Delta t/2) = C^2 \sin^2(kh/2).$$

Long wavelengths recover $\omega \simeq c|k|$. Short waves can propagate at a different phase speed even in a perfectly homogeneous physical medium. Requiring a real ω for every resolved k gives $C \leq 1$. This derivation concerns interior grid modes; domain boundaries can introduce reflected waves as a separate numerical effect.

DERIVATION AND APPLICATIONS

Derive a conserved energy for the centred wave update

A time-stepping method can support a numerical travelling wave while transporting its phase at a slightly incorrect speed. For a second diagnostic, ask whether total discrete vibration energy remains bounded. Consider a uniform one-dimensional periodic grid for $u_{tt} = c^2 u_{xx}$, unit mass density and no forcing. Denote the forward spatial difference by D and the half-step velocity by $v^{n+1/2} = (u^{n+1} - u^n)/\Delta t$. Summation by parts on this grid gives $\langle v, D^{\mathsf{T}} D u\rangle = \langle Dv, Du\rangle$ using the grid inner product. The centred step can be written

$$v^{n+1/2} - v^{n-1/2} = -c^2 \Delta t D^{\mathsf{T}} D u^n.$$

Take the inner product with $(v^{n+1/2} + v^{n-1/2})/2$. The left side becomes a difference of two kinetic energies. On the right use $\Delta t(v^{n+1/2} + v^{n-1/2}) = u^{n+1} - u^{n-1}$ and move D to the other factor:

$$E^{n+1/2} = E^{n-1/2},$$

$$E^{n+1/2} = \tfrac{1}{2}\|v^{n+1/2}\|^2 + \tfrac{1}{2}c^2\langle Du^{n+1}, Du^n\rangle.$$

The potential term couples consecutive time levels. Writing it in terms of the midpoint displacement $\bar{u} = (u^{n+1} + u^n)/2$ identifies when the invariant bounds the solution:

$$E^{n+1/2} = \tfrac{1}{2}\|v\|^2 - \tfrac{1}{8}c^2\Delta t^2\|Dv\|^2 + \tfrac{1}{2}c^2\|D\bar{u}\|^2.$$

Since $\|Dv\| \le 2\|v\|/\Delta x$, the kinetic part is bounded below by $(1 - C^2)\|v\|^2/2$, with $C = c\Delta t/\Delta x$. The invariant controls nonconstant waves for $C < 1$. An exact but indefinite conservation law alone would not guarantee a bounded numerical solution.

WORKED EXAMPLE

Read a timestep from a material and grid

Take $c = 100$ m/s, $\Delta x = 0.0500$ m and $\Delta t = 0.400$ ms. Then $C = 0.800$ and the bound preserves at least $1 - C^2 = 0.360$ of the usual half-step kinetic term even in the worst spatial mode. The customary limiting step from this argument is 0.500 ms.

SECOND APPLICATION

Show why equality of successive energies can mislead

Take the alternating grid mode for which $\|Dv\| = 2\|v\|/\Delta x$, and choose adjacent displacements whose midpoint is zero. At $C = 1.20$ the invariant for this state is

$$E = (1 - 1.44)\|v\|^2/2 = -0.220\|v\|^2.$$

Negative energy here is a property of the numerical quadratic form. A solver can preserve it exactly and still exhibit amplitude growth.

Distinguish transport of phase from transport of energy

A taut string with small transverse displacements and fixed tension obeys $u_{tt} = c^2 u_{xx}$. The physical speed c is independent of wavelength in this ideal one-dimensional model. A discrete solver can transport long waves at nearly the correct speed while short waves arrive late. In a pulse comprising many wavelengths, this artificial dependence on frequency distorts its shape even in a stable computation. A phase comparison at a probe provides information beyond monitoring maximum amplitude.

Use centred second differences in both time and space, with $C = c\Delta t/h$. Solving explicitly for the new level gives

$$u_i^{n+1} = 2u_i^n - u_i^{n-1} + C^2(u_{i-1}^n - 2u_i^n + u_{i+1}^n).$$

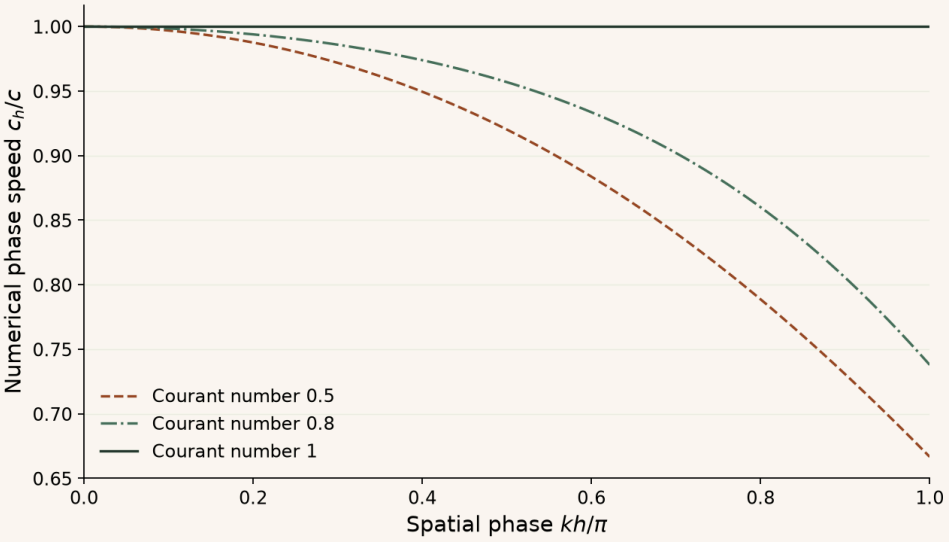
To initialise, expand $u(\Delta t) = u(0) + \Delta t\, v(0) + (\Delta t)^2 c^2 u_{xx}(0)/2 + \cdots$. Replacing $u_{xx}(0)$ by the same spatial difference gives a first level consistent with the initial velocity and acceleration. Simply copying the starting displacement to level one can introduce a spurious launch velocity.

Derive the discrete dispersion relation

Take a travelling mode $u_i^n = \exp[\mathrm{i}(i\theta - n\omega_h \Delta t)]$ with spatial phase $\theta = kh$. Inserting it makes the temporal second difference proportional to $-4\sin^2(\omega_h \Delta t/2)$ and the spatial second difference proportional to $-4\sin^2(\theta/2)$. Thus on the positive-frequency branch

$$\sin(\omega_h \Delta t/2) = C \sin(kh/2).$$

A real ω_h exists for all resolvable real spatial phases only when $C \le 1$. If $C > 1$, the shortest wave makes the right-hand side exceed one; a temporal amplification root then lies outside the unit circle. Even when stable, the ratio of numerical to physical phase speed is $2\arcsin[C\sin(\theta/2)]/(C\theta)$. A fine spatial sample makes θ small and the ratio tends to one. At $C = 1$ the ideal uniform one-dimensional interior stencil reproduces the wave phase on the resolved branch. Boundaries, material transitions and multiple dimensions require separate analysis.



Centred one-dimensional wave discretisation. A short wave remains stable for the illustrated Courant numbers while accumulating substantial phase error in two of the runs.

Estimate a travel-time error

Take $C = 0.5$ and four points per wavelength, so $\theta = \pi/2$. Then

$$\omega_h \Delta t = 2\arcsin[0.5\sin(\pi/4)] = 0.72273\,\mathrm{rad}.$$

The physical advance is

$$C\theta = \pi/4 = 0.78540\,\mathrm{rad}.$$

The numerical phase speed is about 0.920 of the true speed. After ten physical periods, its accumulated phase is about 9.20 cycles. A stable solution can therefore be nearly one oscillation behind while retaining a visually plausible amplitude.

Group velocity is $d\omega_h/dk$, obtained by differentiating the dispersion relation. On this branch $v_g/c = \cos(\theta/2)/\sqrt{1 - C^2\sin^2(\theta/2)}$, about 0.756 in the four-point example. A narrow-band packet's envelope and an individual crest then propagate at different numerical speeds. When interpreting a probe trace, say whether the reported arrival corresponds to a crest, an envelope or a threshold. Repeat the calculation at additional points per wavelength using the same Courant number to isolate spatial-phase improvement from changes in the time integration.

Interpret the grid's travelling domain of dependence

In one explicit wave step, a node uses only itself and its two nearest neighbours. After n steps its numerical domain of dependence extends at most n grid cells to either side. The continuous wave at the same time can depend on initial data a distance $cn\Delta t$ away. For convergence that physical domain must fit inside the numerical one, giving $c\Delta t \le h$. This reproduces the Courant condition from geometry without inserting a Fourier mode. It is a necessary information-speed test; by itself it does not establish stability or accuracy of an arbitrary scheme.

As an example, with $c = 340$ m/s and $h = 0.01$ m, the step bound is about 29.4 microseconds. A 20 microsecond step satisfies it but need not give an accurate answer for a mode represented by only three or four points per wavelength. Time-of-flight information and wavelength resolution are related to different discretisation choices. Document both when comparing two wave calculations.

Derive a reflected wave from a discrete outgoing condition

For a one-dimensional wave travelling right at speed c, the outgoing continuum condition is $u_t + cu_x = 0$ at the right boundary. It annihilates an exact right-travelling waveform. A numerical boundary update using forward time and backward space gives

$$u_N^{n+1} = (1 - C)u_N^n + Cu_{N-1}^n,$$

where $C = c\Delta t/h$. A wave satisfying the interior discrete dispersion relation need not satisfy this one-sided relation. Even when the continuum boundary is perfectly absorbing, a mismatch of discrete phase relations can send energy back into the computational domain.

Let $z = e^{-i\omega\Delta t}$, with the incident and reflected grid waves represented by $e^{i\theta(j-N)}z^n$ and $Re^{-i\theta(j-N)}z^n$. At the boundary their sum is $1 + R$ times z^n. Substitution in the outgoing update and grouping incident and reflected terms gives

$$R = -\frac{z - 1 + C - Ce^{-i\theta}}{z - 1 + C - Ce^{i\theta}}.$$

This amplitude ratio assumes a homogeneous grid, a propagating frequency and distinct incident and reflected branches. It compares each wave at the boundary itself. Signals measured farther inward acquire additional phase and travel-time factors and can interfere if both arrive in the same observation window.

Calculate a reflection generated by numerical mismatch

Use the centred wave scheme with $C = 1/2$ and grid angle $\theta = \pi/2$, four cells per wavelength. The interior relation gives $\sin(\omega \Delta t/2) = (1/2)\sin(\pi/4) = 1/(2\sqrt{2})$, hence $z = 0.75 - i\sqrt{7}/4$. The incident mismatch in the numerator is $0.25 - i(\sqrt{7}/4 - 0.5)$; the reflected denominator is $0.25 - i(\sqrt{7}/4 + 0.5)$. Squaring the magnitudes gives $(3 - \sqrt{7})/4$ and $(3 + \sqrt{7})/4$, respectively. Therefore $|R| = \sqrt{(3 - \sqrt{7})/(3 + \sqrt{7})} \approx 0.2505$. Roughly one quarter of the incident amplitude is reflected solely by this discrete boundary treatment.

For incident and reflected branches on the same symmetric grid, their group-speed magnitudes match. The corresponding reflected energy fraction is $|R|^2 \approx 0.0627$, about six percent. That is small enough for a short experiment to appear satisfactory until the pulse returns to a sensor. Testing only the incident travel time would miss this boundary error. Compare a time window before reflection with the expected window for the return journey.

At $C = 1$ and $0 < \theta < \pi$, the right-moving interior root is $z = e^{-i\theta}$. The numerator vanishes exactly and the boundary update transmits that branch without reflection. This exact compatibility is special to this one-dimensional grid, this step size and this outgoing condition. Varying the propagation angle in two dimensions, changing an interface spacing or using a higher-order interior stencil changes the phase relation that the boundary operator needs to match.

There is a practical way to isolate this reflection in a pulse computation. Record incident and outgoing velocity or flux at the same interior sensor, and choose the duration so that the reflected signal arrives after the incident pulse has passed. Normalise amplitudes in the same physical variable, and avoid dividing a pressure peak by a displacement peak. A broadband pulse contains many values of the grid angle, so its returned waveform can change shape even when each monochromatic component obeys the calculated reflection coefficient.

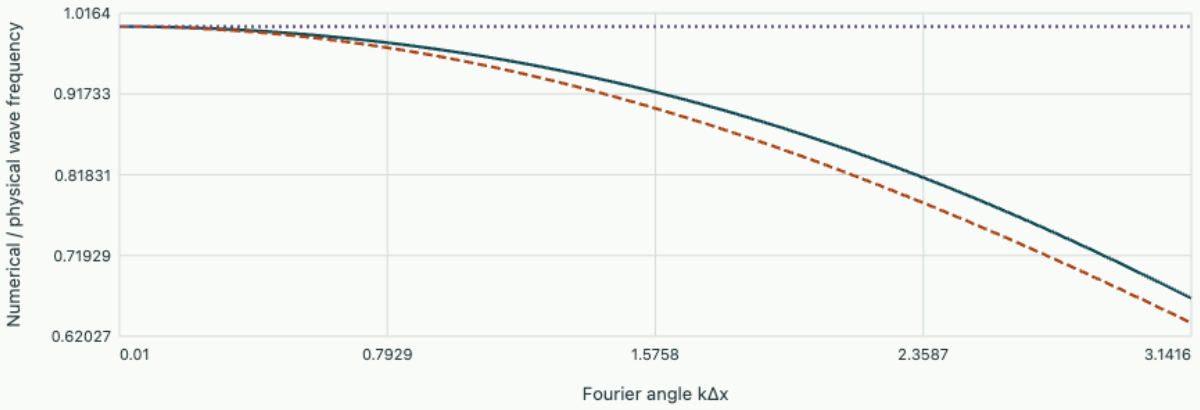
Change parameters · Wave speeds and numerical dispersion

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

Courant number ν **0.5**



The formula follows sin(ωh/2) = ν sin(kΔx/2). At ν = 1, this one-dimensional constant-speed scheme has exact resolved-mode phase propagation.

Study the full derivation

How much of an outgoing wave reflects from the numerical boundary? · Spatial discretisation and field equations

A finite grid truncates an effectively open medium. Derive an exact one-dimensional radiation condition, predict the angular reflection of its two-dimensional extension and design a pulse experiment that separates boundary reflection from interior dispersion.

Chapter 6 · Verification, refinement and local error

Use a manufactured solution to verify an implementation, estimate observed convergence and identify when adaptivity is needed for the physical quantity of interest.

FROM FIRST PRINCIPLES

Derive the observed order and an extrapolated continuum output

Verification compares the discretized problem against a known solution or controlled refinement. Choose one scalar output Q, the same geometry and the same physical observation point on each mesh. Assume a smooth leading error $Q_h = Q + Ch^p + o(h^p)$ on geometrically similar meshes, with refinement ratio $r > 1$. Subtract successive estimates to remove the unknown continuum output:

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

Their ratio approaches r^p, provided the leading coefficient is nonzero. Solve for the order, then use the finest pair to remove the leading error term:

$$p_{\text{obs}} = \frac{\ln\left|(Q_h - Q_{h/r})/(Q_{h/r} - Q_{h/r^2})\right|}{\ln r},$$
$$Q_{\text{ext}} = Q_{h/r^2} + \frac{Q_{h/r^2} - Q_{h/r}}{r^{p_{\text{obs}}} - 1}.$$

The ratio is meaningful only when a consistent leading error dominates. Differential changes of solver tolerance, geometry or timestep across the meshes can contaminate it. A manufactured solution independently tests discretization and boundary terms by inserting a chosen exact field into the differential operator and supplying the resulting source.

Background · Observed convergence order

$$p_{\text{obs}} = \log(e_h/e_{h/2})/\log 2$$

Errors must refer to the same observable and norm. Cancellation, singular boundaries and a reference solution of insufficient accuracy can distort the result.

DERIVATION AND APPLICATIONS

How the measured norm changes an observed convergence rate

A verification study may report very different rates for the same discrete field. Choose a fixed one-dimensional domain of length one, divided into cells of width h. Suppose the solution is smooth away from an interface, where a nearby discontinuity, boundary approximation or reconstruction produces a localized error. Let the error magnitude be Ah^{α} on a fixed number m of cells near the interface, and at most Bh^{p} elsewhere. The count m stays fixed under refinement. For a finite $q \geq 1$ use the physically weighted norm instead of an unweighted vector norm:

$$\|e_h\|_q^q = \sum_i h|e_i|^q \lesssim mA^q h^{q\alpha+1} + B^q h^{qp}.$$

As h decreases the term with the smaller exponent decays more slowly. If the indicated leading magnitudes are actually attained, the corresponding rate of the norm is

$$\|e_h\|_q = O(h^{\min(\alpha+1/q,p)}),$$
$$\|e_h\|_\infty = O(h^{\min(\alpha,p)}).$$

The extra $1/q$ comes from the shrinking physical measure of the affected region. It disappears in a maximum norm. A single signed integral of error can give still another result because opposite local errors cancel:

$$\left|\sum_i he_i\right| \leq \sum_i h|e_i| = \|e_h\|_1.$$

A convergence statement therefore requires the reported norm, physical location, and mesh sequence. If an interface stays at a changing subcell position as the grid is refined, the leading coefficient can vary; observed slopes from just two grids can then be misleading.

WORKED EXAMPLE

Calculate three norms for a displaced step

Let one cell of a piecewise-constant reconstruction differ from the exact field by 1.00, with zero error elsewhere. At $h = 1/64$, the L^1 error is $1/64 = 0.015625$, the L^2 error is $\sqrt{1/64} = 0.125$ and the maximum error is 1.00. Refining to $h = 1/256$ gives 0.003906, 0.0625 and 1.00. The measured rates are one, one half and zero for precisely the same displacement.

SECOND APPLICATION

Compare cancellation with local accuracy

Suppose two adjacent cells at $h = 1/100$ have errors $+0.500$ and -0.500. The signed integral discrepancy is zero; the L^1 error is 0.0100 and the L^2 error is

$$\sqrt{0.005} = 0.0707.$$

Checking only total inventory would certify the output but miss the incorrect distribution. A local measurement or absolute-error norm verifies a different and often more demanding quantity.

ONESIMULATION LABORATORY

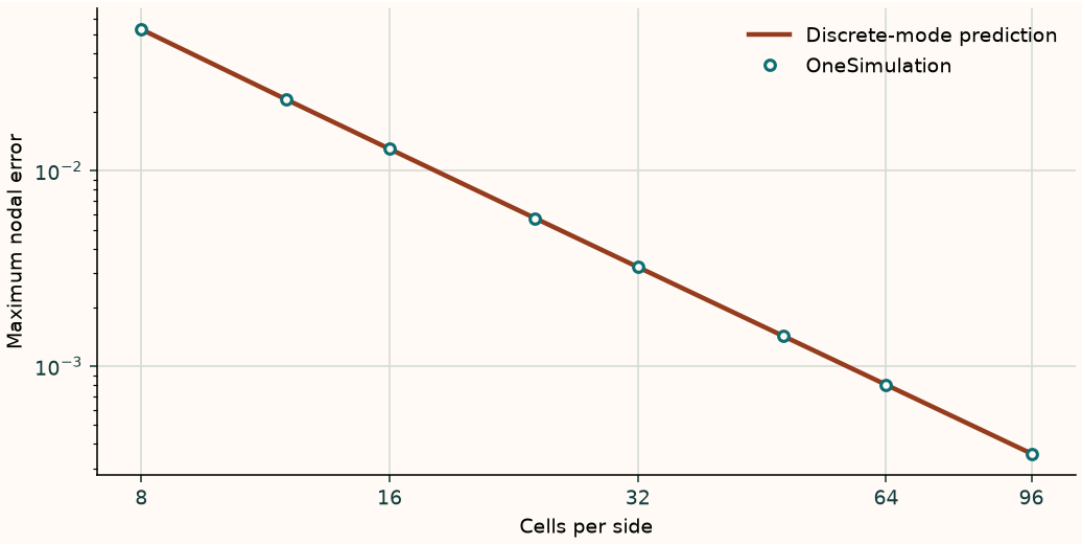
Predict grid error from the wavelength

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 $f = 2a^2 u$ 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
```

> **Start from this chapter.** Predict an error ratio of four for second-order spatial accuracy when both grid spacings halve. Compare three meshes at the same sine mode, then test a shorter wavelength.
>
> OneSimulation laboratory · Measure spatial convergence on a known field

Build a test with a known answer

Consider $-\nabla \cdot (D\nabla u) + \kappa u = f$ on the unit square. Select the exact field $u_\star = \sin(\pi x)\sin(\pi y)$ and prescribe its value at the boundary. Now choose $D = 1 + x$, $\kappa = 2$, using coordinates normalised to a

metre. Differentiation gives

$$\Delta u_\star = -2\pi^2 u_\star,$$
$$u_{\star,x} = \pi \cos(\pi x)\sin(\pi y).$$

The required source is therefore

$$f = [2(1+x)\pi^2 + 2]u_\star - \pi\cos(\pi x)\sin(\pi y).$$

The derivative of D is essential: inserting only the local value of D times the Laplacian specifies a different differential equation. Manufacturing a source tests whether the code solves the stated model and boundary conditions. Experimental validation asks a separate question about fidelity to physical observations. A polished plot resembling the exact field does not substitute for a numerical error measure with defined sample positions and units.

Read a three-grid study

On successively halved mesh spacings suppose the error in one norm follows $E_h = Ch^p + o(h^p)$. Taking ratios gives

$$p \simeq \log_2(E_h / E_{h/2}).$$

Three grids provide two such estimates, indicating whether a consistent asymptotic trend is present. A single ratio may result from cancellation between terms of different order. State the norm, grid location, geometry representation, time step if applicable, and the residual tolerance. Comparing a cell average on one grid with a nodal peak on another can produce a meaningless rate.

For the variable-diffusion square, maximum nodal errors at 24, 48 and 96 intervals per axis are approximately $1.33345 \times 10^{-3}, 3.33763 \times 10^{-4}, 8.34283 \times 10^{-5}$. The observed orders are approximately 1.998 and 2.000. The finest-grid algebraic residual is already far below the field error in its respective normalisation. Tightening the iteration tolerance alone will not remove the leading spatial defect. Raising the sine mode while holding the mesh fixed increases derivatives of the exact solution and tests a different resolution regime. Inspect the grid and mode experiment.

Decide where further resolution matters

A manufactured sine is smooth everywhere. Actual domains may contain a re-entrant corner, a material jump or a narrow source. Derivative bounds used in theoretical convergence estimates can fail locally there. Uniformly halving a two-dimensional mesh roughly quadruples the cell count while spending most new cells away from a localised feature. Local refinement attempts to concentrate new degrees of freedom where they matter to a chosen output. Which cells merit refinement depends on the equation, boundary data and the output being measured.

For a conforming finite-element approximation, one class of error indicators combines an element residual with jumps in normal flux across interior faces. The residual tests whether each element satisfies the differential balance, while the jumps measure disagreements between neighbouring reconstructed fluxes. In a constant-coefficient Poisson example, a piecewise linear solution can have zero elementwise Laplacian even when it bends globally; the slope changes then appear at interelement faces. An indicator using only the interior derivative would miss this information. Mesh-length factors and coefficient weights are necessary to combine the terms in a specific norm.

Finish a refinement study by tracking the physical observable itself. If the objective is integrated heat flow, a changing maximum pointwise gradient near one corner may have little influence on that integral. Conversely, apparently stable domain averages can hide an unresolved peak. Report observed change with resolution and an independent balance or analytic check, specifying exactly which quantity has been verified.

Choose cells to refine from residuals and flux jumps

A mesh-convergence table needs several complete solutions. To spend the next refinement locally, examine the equation that the present field still fails to satisfy. For $-u'' = f$ with zero Dirichlet endpoints, let u_h be its conforming piecewise-linear Galerkin solution. Integrating the error equation by parts on each cell K gives

$$\int_0^1 (u - u_h)' v' \, dx = \sum_K \int_K (f + u_h'') v \, dx - \sum_z J_z v(z),$$

$$J_z = u_h'(z^-) - u_h'(z^+).$$

The sum over z includes interior nodes only. Even when the cell derivative is constant and $u_h'' = 0$ in each open cell, its jumps measure failures of flux continuity. Dropping these terms would miss part of the weak residual. Galerkin orthogonality lets us substitute $v - I_h v$ for v, where $I_h v$ is a local conforming approximation. Local interpolation and trace estimates supply the mesh-length factors in an indicator. One conventional allocation on a uniform mesh is

$$\eta_K^2 = h_K^2 \int_K (f + u_h'')^2 \, dx + \frac{h_K}{2} \sum_{z \in \partial K \text{ interior}} J_z^2.$$

The half shares a node between its two neighbors. An upper bound on the energy error involves the square root of the summed indicators and a mesh-independent constant. Efficiency estimates can also involve unresolved variations in the load. The indicator is a refinement criterion, not a measured cell error.

Spending one refinement near a source

On the two half-intervals let $f = 0$ on the left and $f = 4$ on the right. The central hat function integrates to $1/4$ on the right, so the interior load is 1. Its stiffness is 4. The central temperature is $1/4$. The cell slopes are $1/2$, $-1/2$, so $J = 1$. The jump contributes $1/4$ to each indicator squared. The

source contributes zero on the left and $(1/2)^2(16/2) = 2$ on the right. Thus the right cell accounts for $2.25/2.5 = 90\%$ of the total. A bulk-marking rule targeting at least 75% marks this one cell.

An exact check explains the distinction between indicator and error. Integrating the differential equation gives the exact slope $1/2$ throughout the left cell and $1/2 - 4(x - 1/2)$ on the right. The true squared energy error is zero on the left and $\int_0^{1/2}(1 - 4s)^2 ds = 1/6$ on the right. The indicator assigns some weight to the neighbor because both cells share a flux jump. Its largest contribution nevertheless identifies the curved part of the field.

Practise on sheet 3 · Refine a three-cell source patch. Further reading · Heidelberg University · Lecture Notes on Scientific Computing with Partial Differential Equations · §9.2 Residual-based a-posteriori Error Estimator and §9.3 Local Mesh Adaptation.

Check refinement at an identical location

Suppose a displacement is reported at the centre of a square. On an even grid that point may coincide with a node, whereas on an odd grid it can lie at the centre of a cell. Taking the maximum over mesh nodes produces a different sample on each grid. A refinement curve can then combine field discretisation error with the movement of the evaluation point. Choose a fixed physical point and evaluate each solution there with its own consistent interpolation, or compare a well-defined integral over the same physical region.

Similarly, compare gradients on compatible supports. Nodal averages of element stresses can look smoother than the individual element values, yet smoothing can alter extrema and convergence rates. Record the native support and any recovery operator. This makes an apparent improvement reproducible and gives students enough information to explain a discrepancy between two otherwise credible simulation reports.

Separate space and time with a crossed refinement study

Suppose a reported output has asymptotic error $Q(h, k) - Q_\star = Ah^p + Bk^q$ plus smaller terms, with spatial spacing h and time step k. Refining both coordinates in a fixed relation examines only one line across this two-dimensional error surface. Opposite signed contributions can cancel on that line. A genuine spatial order check uses several h at fixed sufficiently small k, and a temporal check uses several k at controlled h. When exact values are available, the individual error signs provide particularly clear evidence of cancellation.

A small crossed design with coarse and fine settings in both directions already offers a structural test. Denote its four outputs by $Q_{cc}, Q_{cf}, Q_{fc}, Q_{ff}$, with the first subscript indicating the spatial mesh. For an additive leading error the spatial change $Q_{cc} - Q_{fc}$ equals $Q_{cf} - Q_{ff}$. Their difference, $Q_{cc} - Q_{cf} - Q_{fc} + Q_{ff}$, measures a mixed contribution or departure from the leading separated model. A small difference supports that model over the tested settings but cannot prove that the assumed orders are correct; additional refinement levels are needed to estimate orders from the data.

Recover a hidden temporal bias from four runs

Take a manufactured output $Q = 10 + 2h^2 - k$. The benchmark value is ten. Choose $h_c = 0.2,\ h_f = 0.1,\ k_c = 0.1,\ k_f = 0.05$. The four outputs are respectively $9.98,\ 10.03,\ 9.92,\ 9.97$. At either time step, halving the spatial mesh reduces the output by 0.06. At either spatial mesh, halving the time step increases the output by 0.05. The mixed difference is zero. Refining both together moves the value by only −0.01, concealing spatial and temporal changes five or six times larger.

If the orders are independently known as $p = 2,\ q = 1$, the fine-grid spatial bias is

$$(Q_{cf} - Q_{ff})/(2^2 - 1) = 0.02.$$

The fine-time bias is

$$(Q_{fc} - Q_{ff})/(2^1 - 1) = -0.05.$$

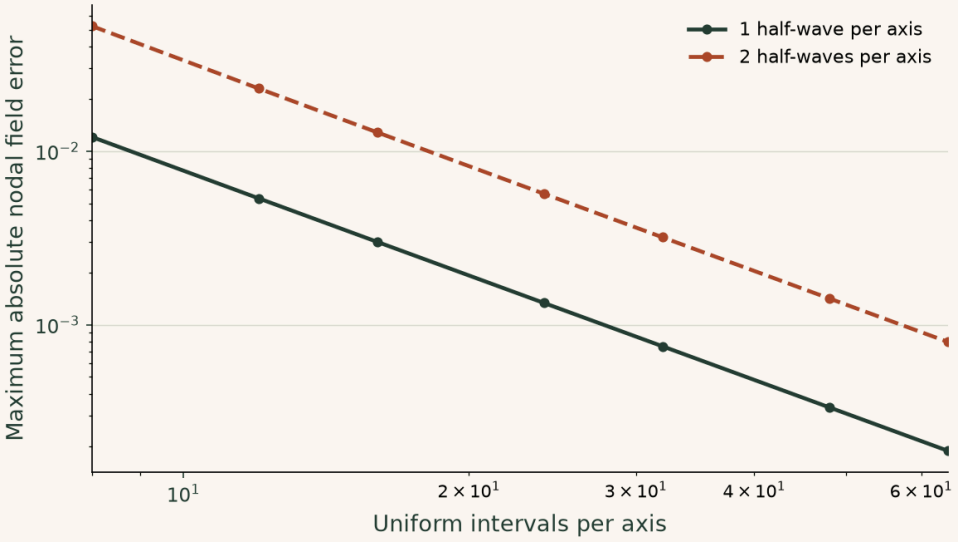
Subtracting both from the fine result gives $9.97 - 0.02 + 0.05 = 10$, exactly recovering the limit in this idealised expansion. Choosing instead $k = 2h^2$ on every paired run would return exactly ten for every mesh, including a very coarse one. A convergence plot along that line would demonstrate cancellation rather than independent verification.

Explicit stability restrictions can make unrestricted crossing impractical. If a coarse time step is unstable on the finest spatial grid, choose a stable common range or use an implicit reference to isolate spatial error. Keep output locations, boundary representation and integration tolerances consistent across the crossed study, since changes in those other approximations can also enter the mixed difference.

A fifth run can interrogate the assumption that the fine-grid spatial change is asymptotic. At $h = 0.05,\ k = 0.05$ the manufactured output is 9.955. The change from 9.97 is 0.015, one quarter of the earlier spatial change 0.06 at the same time step. This independent ratio supports $p = 2$ for the tested family; it does not distinguish a true zero cross term from a smaller mixed term below the numerical resolution.

Resolve a second spatial frequency on the same grid

Set the exact field on the square to $\sin(m\pi x)\sin(m\pi y)$ and manufacture the corresponding source for diffusion coefficient $1 + x$ and reaction coefficient 2. Run the downloadable OneSimulation model with m = 1 and 2 on each grid. The boundary data agree, while the source and the exact interior curvatures change.



Both series approach a slope of minus two on the log–log axes. At fixed grid count the higher-frequency solution is less accurate. Doubling its grid count restores the points per wavelength; compare those shifted entries to see how closely that simple resolution argument predicts the actual errors.

▶ **Read values from the plot**

Download the OneSimulation model (Python)

Change parameters · Resolve a second spatial frequency on the same grid

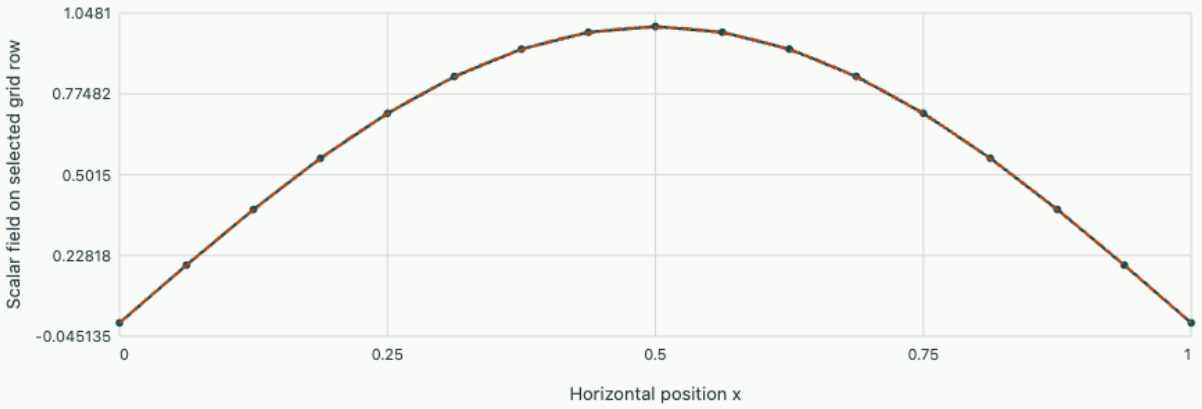
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Mesh intervals per axis **16**

Sine half-waves per direction **1**



Variable-coefficient five-point solve | Manufactured continuum field

For −∇·[(1+x)∇u] + 2u = f, maximum nodal error 0.0030008, relative algebraic residual 4.2371e-11. Slice y = 0.5; 17 Jacobi-preconditioned CG iterations. Refinement measures discretisation after the algebraic solve has converged.

Study the full derivation

Bound diffusion error by balancing a reconstructed flux · Spatial discretisation and field equations

A temperature field from a finite-element calculation looks smooth. Reconstruct a heat flux which exactly balances the source, then derive an upper bound for the true energy error without knowing the exact temperature.

Problem sheet 3 · Wave resolution and verification

Prescribe the first time level of a vibrating string, measure a convergence order and decide what a residual reveals under grid refinement.

Release a mode from rest

For a unit string with fixed ends take initial shape $u_i^0 = \sin(\pi i/N)$ and initial velocity zero. Derive the factor relating u_i^1 to u_i^0 for the centred wave scheme with Courant number C. Evaluate it for $N = 4,\ C = 0.5$.

▼ **Worked solution**

The central second difference of the sine is $-4\sin^2(\pi/(2N))u_i^0$ before dividing by h^2. The first-step Taylor formula therefore gives

$$u_i^1 = [1 - 2C^2\sin^2(\pi/(2N))]u_i^0.$$

At $N = 4, C = 0.5$, the factor is

$$1 - 0.5\sin^2(\pi/8) = 0.92678.$$

Copying the initial displacement instead would omit the physical acceleration produced by the curved shape.

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.

Quantify a grid-dependent observable

A computed outflow on meshes $h, h/2, h/4$ is 2.080, 2.020 and 2.005 W. Assuming an asymptotic $Q_h = Q + ch^p$, estimate p, the extrapolated flux and the fine-grid error. Can those figures alone demonstrate conservation to the same tolerance?

▼ **Worked solution**

The successive differences are 0.060 and 0.015 W, whose ratio is 4. Thus

$$2^p = 4,\ p = 2.$$

Extrapolate $Q = 2.005 + (2.005 - 2.020)/3 = 2.000$ W, giving estimated fine-grid error +0.005 W.

A separate conservation check must sum inflows, outflows, sources and storage. A sequence of outflow values can converge cleanly even under a method that systematically loses heat elsewhere.

Explore the related calculation with adjustable inputs. Refine the grid on a known solution and keep the small linear-solver residual separate from the spatial field error.

Relate flux jumps to a residual indicator

A unit rod is approximated by two equal linear elements with nodal values $(0, 1, 1)$ and $k = 1$. Compute each element's flux, the flux jump at their interface, and the elementwise second derivative. In a source-free steady model, explain what a check based only on that second derivative would miss.

▼ **Worked solution**

The slopes are 2 and 0, giving fluxes –2 and 0 in the positive coordinate direction. The right-element flux minus the left-element flux is therefore 2. Second derivatives vanish inside both linear pieces.

In a source-free steady conductor the same flux must cross the interface from either element. An interior derivative check would report zero inside each element while ignoring the change at the joining node. An indicator including interface flux mismatch detects this loss of balance.

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.

Refine a three-cell source patch

Split the unit interval into thirds and set $f = 0, 9, 0$ on successive cells, with zero temperature at both ends. Assemble the two interior values, calculate all three indicators and find a smallest set that accounts for 80% of the sum of indicator squares.

▼ **Worked solution**

The middle cell contributes $9(1/3)/2 = 3/2$ to each endpoint. The interior matrix is $3\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$, so symmetry gives

$$u_1 = u_2 = 1/2.$$

The slopes are $3/2, 0, -3/2$. Both jumps have magnitude $3/2$.

Each attached jump contributes $(1/6)(3/2)^2 = 3/8$ to a cell. The middle source term contributes $(1/3)^2(81/3) = 3$. Thus the indicator squares are $3/8, 15/4, 3/8$, totaling $9/2$. The middle fraction is

$$(15/4)/(9/2) = 5/6.$$

Marking just that cell exceeds the target.

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

Revisit chapter 6 · Choose cells to refine from residuals and flux jumps.

Predict the error before making a finer grid

In the variable-diffusion laboratory the mode-one errors on 24, 48 and 96 cells per side are 1.33345×10^{-3}, 3.33763×10^{-4} and 8.34283×10^{-5}. Estimate the last observed order, then predict the error at 192 cells and the number of interior unknowns. Decide whether reducing only the algebraic residual tolerance is likely to produce the same improvement.

▼ **Work through the solution**

For $E_N \approx C N^{-p}$, the last doubling gives

$$p = \log_2(E_{48}/E_{96}) \simeq 2.00.$$

Another doubling predicts $E_{192} \simeq 2.09 \times 10^{-5}$. Eliminating prescribed boundary nodes leaves $191^2 = 36\,481$ interior values. The observed rate is a checkable prediction for this smooth fixed mode. A smaller residual on the 96-cell mesh solves the same discrete equations more accurately. The reported residual is already small, and refinement is what changes their truncation error.

Use the OneSimulation laboratory to check your answer

Chapter 7 · Interfaces, contact resistance and Robin data

Derive a face flux that transports the same heat through two different materials and through a convective boundary.

FROM FIRST PRINCIPLES

Derive an interface flux from two half-cell temperature drops

Two neighbouring cell centres can lie in materials with sharply different conductivities. Consider a planar interface between them, normal to a one-dimensional steady heat flow. Let d_L, d_R be the distances from the centres to the interface, and suppose there is no interface source, storage or contact resistance. Steady conservation requires one heat flux q, positive to the right, and perfect contact requires one interface temperature T_I. Applying Fourier's law to each half gives

$$q = k_L(T_L - T_I)/d_L = k_R(T_I - T_R)/d_R.$$

Solve each expression for its temperature drop, add the two drops and eliminate T_I. The conductance follows from the actual material distances:

$$q = \frac{T_L - T_R}{d_L/k_L + d_R/k_R},$$

$$k_{\text{eff}} = \frac{d_L + d_R}{d_L/k_L + d_R/k_R}.$$

For equal distances this is the harmonic mean; only equal conductivities recover the same result as the arithmetic mean. If one material approaches zero conductivity with positive thickness, the flux approaches zero. A finite thermal contact resistance adds another temperature drop qR_c, while a heat-generating interface alters the flux-continuity condition itself.

DERIVATION AND APPLICATIONS

Enforce the flux jump at a heat-generating interface

Two materials in series carry the same conductive current only if their common interface stores and generates no energy. Consider two homogeneous slabs with an idealized thin heater between them. Let T_L, T_R be prescribed outer surface temperatures, R_L, R_R the positive thermal resistances per unit area in m² K/W, and s the heater release in W/m². Assume perfect thermal contact, steady conduction and negligible thickness of the heater. Define the interface temperature as T_I and flux q positive to the right.

Fourier's law integrated through each slab relates its own constant flux to its temperature drop. The pillbox balance across the heater requires the right flux to exceed the left one by the generated heat:

$$q_L = (T_L - T_I)/R_L,$$
$$q_R = (T_I - T_R)/R_R,$$
$$q_R - q_L = s.$$

Substitute the two face laws into the balance. With $g_j = 1/R_j$ as the two conductances per area, the interface temperature and individual fluxes are

$$T_I = \frac{g_L T_L + g_R T_R + s}{g_L + g_R},$$
$$q_R = g_R(T_I - T_R),$$
$$q_L = g_L(T_L - T_I).$$

Alternatively eliminate T_I to express the fluxes as through-flow plus a split source. The resistances determine the fraction of added heat directed to each boundary:

$$q_L = \frac{T_L - T_R - sR_R}{R_L + R_R},$$
$$q_R = \frac{T_L - T_R + sR_L}{R_L + R_R}.$$

Setting $s = 0$ restores a single common flux. Equal outside temperatures send more generated heat toward the lower resistance. For an interface at a mesh face the two cells therefore need distinct physical face fluxes linked by the stated jump; assigning exactly the same flux while adding no source anywhere violates energy balance.

WORKED EXAMPLE

Determine both heat flows through an instrumented wall

Let $T_L = 320$ K, $T_R = 300$ K, $R_L = 0.100$, $R_R = 0.200$ m² K/W and $s = 150$ W/m². The interface sees conductances 10.0 and 5.00 W/(m² K) and reaches 323.33 K. The left flux is -33.3 W/m² and the right flux 116.7 W/m². Heat leaves on both sides, including toward the hotter external surface. Their oriented difference equals 150 W/m².

SECOND APPLICATION

Select a heater setting that makes one sensor read zero

Keeping the same temperatures and layers, impose $q_L = 0$. Then

$$T_I = T_L = 320\,\mathrm{K},$$
$$s = (T_L - T_R)/R_R = 100\,\mathrm{W/m^2}.$$

All generated heat exits on the right. Reducing s to zero instead gives 66.7 W/m² from left to right through both layers. A zero left-side flux in the heated experiment therefore does not imply an insulating left layer.

Touch two long solids: which temperature does the interface acquire?

Two long blocks, initially at temperatures $T_L > T_R$, are pressed into ideal thermal contact at $x = 0$. Denote conductivity by k, volumetric heat capacity by C, and diffusivity by $\alpha = k/C$. At early times the penetration distance on each side is small compared with its length. Model the blocks as half-spaces with a common interfacial temperature T_*. This transient contact problem has no fixed layer thickness to insert into a pair of steady thermal resistances.

In the right half-space the similarity coordinate is $x/(2\sqrt{\alpha_R t})$. Substituting a function F of this coordinate into the heat equation gives

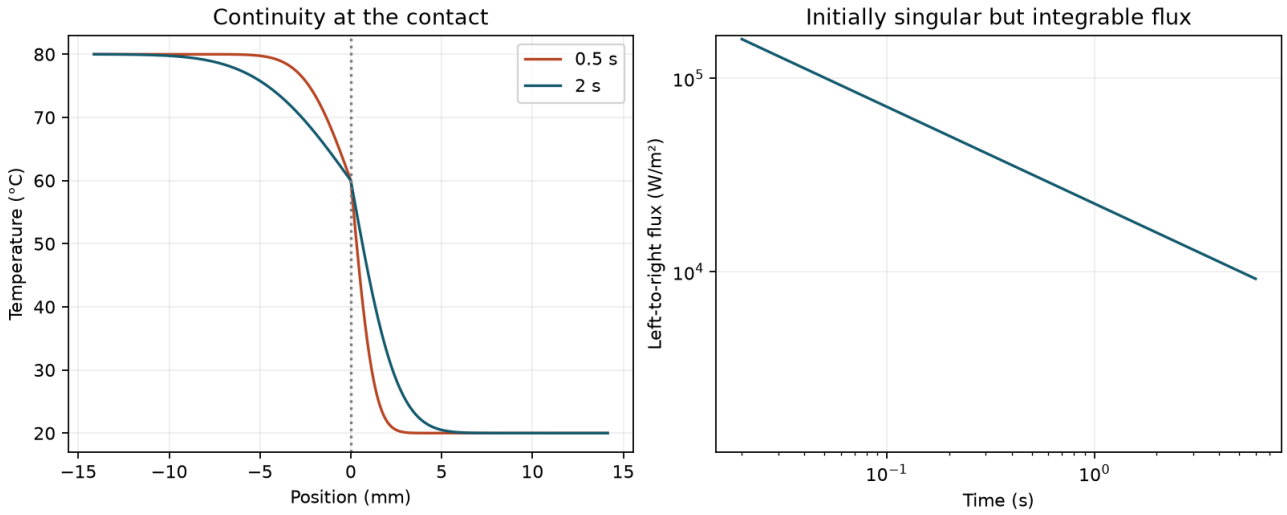
$$F'' + 2\eta F' = 0.$$

Its bounded solution matching a fixed surface and a uniform distant temperature has an error-function shape. Thus $T_R(x,t) = T_* + (T_R - T_*)\,\mathrm{erf}[x/(2\sqrt{\alpha_R t})]$. The left solution uses distance $-x$, its own diffusivity and the initial left temperature. These profiles share their interface temperature by construction; conservation still requires their heat fluxes to match.

Differentiate the error functions at zero and apply Fourier's law. The left flux equals $e_L(T_L - T_*)/\sqrt{\pi t}$, and the right flux is $e_R(T_* - T_R)/\sqrt{\pi t}$, both oriented from left to right. Here $e = \sqrt{kC}$ is thermal effusivity. Equating the fluxes gives

$$T_* = (e_L T_L + e_R T_R)/(e_L + e_R).$$

The interfacial temperature is independent of time within this idealisation, while the penetration layers broaden continuously.



The two temperature gradients differ because flux continuity weights each gradient by its own conductivity. The common temperature depends on heat capacity as well as conductivity.

For the notebook's $k_L = 4,\ k_R = 1\,\mathrm{W/(m\,K)}$ and equal capacities of $10^6\,\mathrm{J/(m^3\,K)}$, the effusivity ratio is two. Between 80 and 20 degrees Celsius the common temperature is 60 degrees Celsius. The initially infinite ideal flux is integrable: the energy transferred by time t equals $2tq(t)$ per unit contact area. The notebook independently integrates temperature loss and gain throughout the two bodies and checks this balance. A real surface resistance, a finite approach time or the far surface of a finite specimen would eventually invalidate different parts of the idealisation.

A finite-volume scheme offers a sensitive check of this benchmark. On a sufficiently long computational domain, sum nodal or cell-average energies on each side at the same time. Internal face fluxes cancel in the total by conservation. A scheme may pass that global check while partitioning energy incorrectly between materials. Compare both individual energy changes, the early flux integrated over the first step and the temperature nearest the interface. This combination tests the material weighting and the time integration independently.

Keep both initial temperatures and conductivities fixed, but multiply the right-hand volumetric heat capacity by four. Predict T_* and the right-hand penetration depth at the same time.

▼ **Work through the solution**

The right effusivity doubles and now equals the left one. The interface becomes 50 degrees Celsius. At fixed k_R the right diffusivity falls to one quarter, so the penetration distance proportional to $\sqrt{\alpha_R t}$ halves. A larger local capacity can draw more flux while confining the altered temperature to a narrower region.

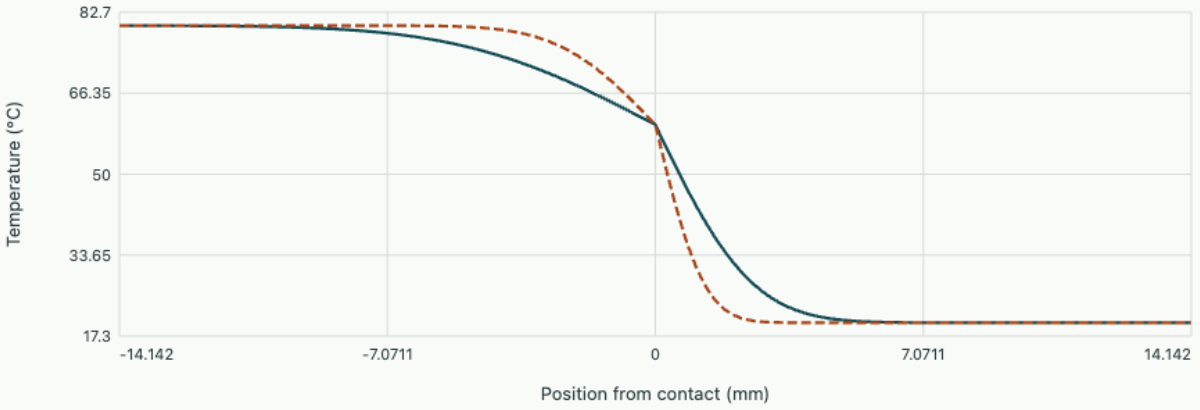
Change parameters · Explore this notebook example

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

Left conductivity (W/m/K)	4
Right conductivity (W/m/K)	1
Elapsed time (s)	2



Common interface temperature 60 °C; heat flux 15958 W/m². Both semi-infinite materials have volumetric heat capacity 10⁶ J/m³/K; contact resistance is zero.

Open editable notebook Download .ipynb

One flux through resistances in series

Let $q = -kT'$ be heat flux in the positive coordinate direction, with unit cross-sectional area and no internal source. Between two cell centres separated by an interface, let the left and right centre-to-interface distances be d_L, d_R. Piecewise constant conductivities $k_L, k_R > 0$ make the equilibrium temperature linear in each segment. If the interface has perfect contact, its temperature T_f is continuous. Integrating Fourier's law on each side gives

$$T_L - T_f = q\,d_L/k_L,$$
$$T_f - T_R = q\,d_R/k_R.$$

Adding eliminates the unknown interface temperature.

$$q = \frac{T_L - T_R}{d_L/k_L + d_R/k_R}.$$

For equal distances this amounts to using the harmonic mean conductivity over the full centre spacing. Inserting the same signed q as the outflow of the left cell and the inflow of the right cell produces cancellation when their balances are added. If a material interface stores energy or supplies heat independently, a flux jump belongs in the balance and this steady series calculation must be modified.

Measure the temperature drop

Take $d_L = d_R = 0.01$ m, $k_L = 2,\ k_R = 10$ W/(m K) and $T_L = 30,\ T_R = 18$ °C. The series resistance per unit area is $0.005 + 0.001 = 0.006$ m²K/W. The flux is 2000 W/m² and the interface temperature is $30 - 2000(0.005) = 20$ °C. The lower conductivity carries the larger temperature drop despite carrying the same flux. Using the arithmetic mean 6 with the centre difference over 0.02 m instead predicts 3600 W/m².

Introduce a contact resistance $R_c = 0.002$ m²K/W. Its temperature jump is qR_c; the total resistance rises to 0.008, reducing the heat flux to 1500 W/m². The two sides of the contact now have temperatures 22.5 and 19.5 °C. Imposing a single interface temperature on this model would remove the physically measured 3 K jump.

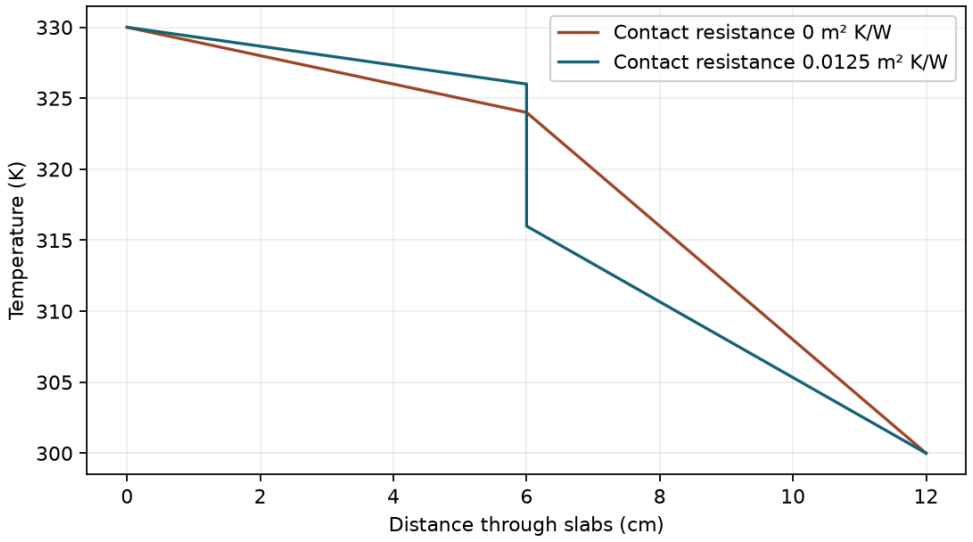
OneSimulation · Resolve a temperature jump at imperfect contact

Two slabs are each 0.06 m thick, with conductivities 12 and 3 W/(m K). Their outer surfaces are held 30 K apart. A contact resistance R_c permits two distinct temperatures at the shared physical location. The flux stays continuous. The face law and the exact reference for the whole assembly are

$$T_{f,L} - T_{f,R} = qR_c,$$
$$q = \frac{T_L - T_R}{L_1/k_1 + R_c + L_2/k_2}.$$

The code adds R_c only to the resistance between the two cells adjacent to the contact. OneSimulation solves the resulting balance system. The flux falls from 1200 to 800 W/m² for $R_c = 0.0125$ m² K/W. Reconstructing the two face temperatures gives 326 K and 316 K. Connecting only the cell-centre temperatures would smear this 10 K jump over one mesh cell. Refining an ideal-contact model would converge to the wrong interface law.

```
"""Find the temperature jump at an imperfect contact between two slabs."""
import matplotlib.pyplot as plt
import numpy as np
from onesim.numerics.linear.matrices.dense_linear_algebra import solve_dense

n = 12  # even number of equal, cell-centred control volumes
length = 0.12  # m
dx = length / n
k = np.repeat([12., 3.], n // 2)  # W/(m K)
left, right = 330., 300.  # K
centres = (np.arange(n) + 0.5) * dx
fig, ax = plt.subplots(figsize=(6.8, 3.8), constrained_layout=True)
for contact, colour in [(0.0, "#a4492b"), (0.0125, "#176879")]:  # m² K/W
    resistance = np.r_[dx/(2*k[0]), dx/(2*k[:-1]) + dx/(2*k[1:]), dx/(2*k[-1])]
    resistance[n//2] += contact
    g = 1 / resistance
    matrix = np.diag(g[:-1]+g[1:]) - np.diag(g[1:-1], 1) - np.diag(g[1:-1], -1)
    rhs = np.zeros(n)
    rhs[0], rhs[-1] = g[0]*left, g[-1]*right
    temperature = solve_dense(matrix, rhs)
    fluxes = -np.diff(np.r_[left, temperature, right]) / resistance
    q = fluxes[0]
    assert np.allclose(fluxes, (left-right)/(length/2/k[0]+contact+length/2/k[-1]))
    hot_face = temperature[n//2-1] - q*dx/(2*k[0])
    cold_face = temperature[n//2] + q*dx/(2*k[-1])
    assert np.isclose(hot_face-cold_face, q*contact)
    x = np.r_[0, centres[:n//2], length/2, length/2, centres[n//2:], length]
    y = np.r_[left, temperature[:n//2], hot_face, cold_face, temperature[n//2:], right]
    ax.plot(100*x, y, color=colour, label=f"Contact resistance {contact:g} m² K/W")
    print(f"Flux {q:.0f} W/m²; interface jump {hot_face-cold_face:.1f} K")
ax.set(xlabel="Distance through slabs (cm)", ylabel="Temperature (K)")
ax.legend()
ax.grid(alpha=0.2)
fig.savefig("partial-differential-equations.png", dpi=160)
```



The vertical segment reconstructs both surface temperatures at the same location. All control-volume faces carry the same steady flux, including the imperfect contact.

Download the Python example

Apply a convective boundary

At an exterior face the outward heat flux obeys $q = h_c(T_s - T_\infty)$, a Robin condition. Eliminate the unknown surface temperature as above to obtain

$$q = (T_P - T_\infty)/(d/k + 1/h_c).$$

If h_c tends to zero, the boundary becomes insulating. With very large h_c, the surface approaches ambient temperature while the finite centre-to-face resistance remains. For $d = 0.01$ m, $k = 2$, $h_c = 100$ W/(m² K), $T_P - T_\infty = 15$ K, the predicted flux is 1000 W/m² and the surface is 10 K above ambient. Multiplying by face area converts this flux to the power used in a cell energy balance. Mixing up these two quantities makes the result depend incorrectly on mesh area.

A jump located inside a cell

If the materials change at a known position inside a segment, the fractions of that segment matter. Moving the interface so that 75 percent of the length has conductivity 2 and 25 percent has conductivity 10 produces effective conductivity $1/(0.75/2 + 0.25/10) = 2.5$ W/(m K). Giving each material half the segment instead would give $10/3$ W/(m K). The harmonic rule depends on actual layer thicknesses. A mesh-refinement study that snaps the interface to different positions at every refinement can mix geometric representation error with flux-discretisation error.

Derive effective transport parallel and normal to layers

A fine laminate of isotropic materials can have a direction-dependent macroscopic conductivity. Consider two perfectly bonded planar layers occupying thickness fractions θ and $1-\theta$, with conductivities k_1, k_2. Assume steady conduction, no internal sources, and boundary conditions that permit a one-dimensional field for the loading orientation being tested. For transport normal to the layers, continuity forces equal flux. Adding their temperature drops across a total thickness L gives

$$k_\perp = [\theta/k_1 + (1-\theta)/k_2]^{-1}.$$

For transport parallel to the layers the imposed longitudinal gradient is common and the area-weighted fluxes add, giving $k_\parallel = \theta k_1 + (1-\theta)k_2$.

The arithmetic and harmonic expressions are thus both correct, but apply to different geometries of imposed transport. Their ordering follows without a conductivity table: multiplying the two positive weighted sums $\theta k_1 + (1-\theta)k_2$ and $\theta/k_1 + (1-\theta)/k_2$ gives $1 + \theta(1-\theta)(k_1 - k_2)^2/(k_1 k_2) \geq 1$. Equality requires identical conductivities or a missing layer. Spatial averaging performed before identifying the flux direction would erase precisely the distinction that determines the coarse physics.

Predict two experiments on the same material sample

Suppose a repeated laminate consists of 40 percent material with conductivity 1 and 60 percent with conductivity 10 W/(m K). Then $k_{\parallel} = 0.4 + 6 = 6.4$ W/(m K) and $k_{\perp} = 1/(0.4 + 0.06) = 50/23 \approx 2.174$ W/(m K). A temperature drop of 4 K over 0.02 m imposes gradient magnitude 200 K/m. Perfectly normal loading produces flux magnitude about 434.8 W/m², whereas the same gradient parallel to the layers yields 1280 W/m². The larger flux is consistent with the fast material providing a continuous path in the latter orientation.

For the normal experiment the low-conductivity phase occupies 0.008 m and accounts for a drop $434.8(0.008)/1 \approx 3.478$ K. The high-conductivity phase occupies 0.012 m and drops only $434.8(0.012)/10 \approx 0.522$ K. Adding recovers the imposed four kelvin. For the parallel experiment individual flux densities are 200 and 2000 W/m². Averaging by their area fractions gives

$$0.4(200) + 0.6(2000) = 1280\,\mathrm{W/m^2}.$$

These local checks use the same two material properties and explain the difference in the observed coarse responses.

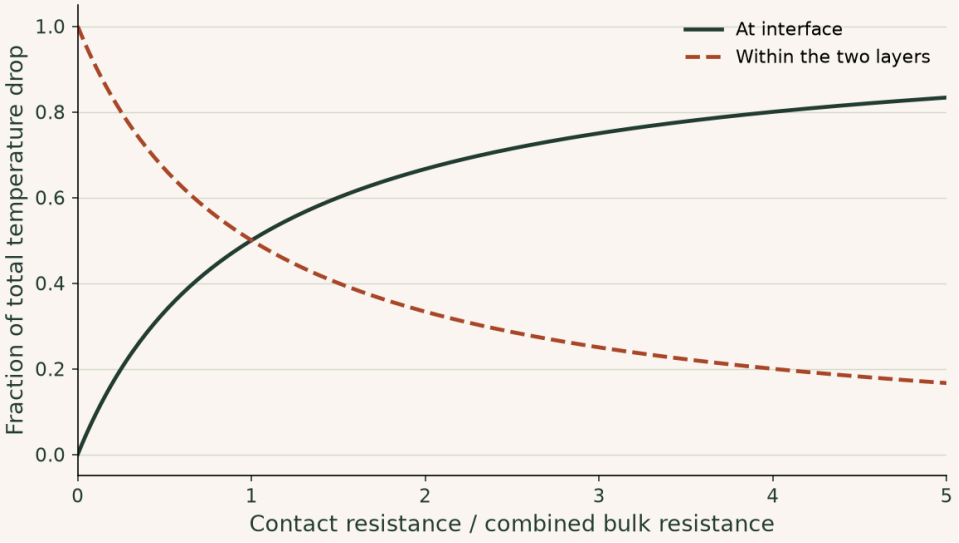
Using the effective tensor assumes the observation scale spans representative layers. A sensor within one layer may respond to its local temperature slope instead. Finite contact resistance adds a temperature jump in the normal direction; if the number of interfaces per macroscopic length changes, the effective normal resistance changes even at identical material fractions. Measuring both orientations can distinguish geometric or contact effects that a single scalar fitted conductivity would conceal.

Rotating a laminate specimen through ninety degrees interchanges the roles of these two effective values without changing its volume fractions. This makes an orientation experiment particularly discriminating: a fit that happened to reproduce the first orientation by adjusting a scalar conductivity predicts the same flux for the rotated specimen, whereas the fine-layer balance predicts a factor of about $6.4/(50/23) = 2.94$ between them.

Repeating the test over multiple layer spacings also probes the assumption of negligible contact resistance.

Separate the temperature jump from the bulk gradient

A steady one-dimensional stack has two positive bulk conduction resistances in series with a contact resistance R_c. Flux is common to all three. Scale R_c by the sum of bulk resistances and predict where probes would measure the imposed end-to-end temperature difference.



When contact resistance equals the bulk sum, half the end-to-end drop occurs at a geometrically thin interface. A numerical scheme that enforces temperature continuity there can converge smoothly to the wrong interface model.

▶ **Read values from the plot**

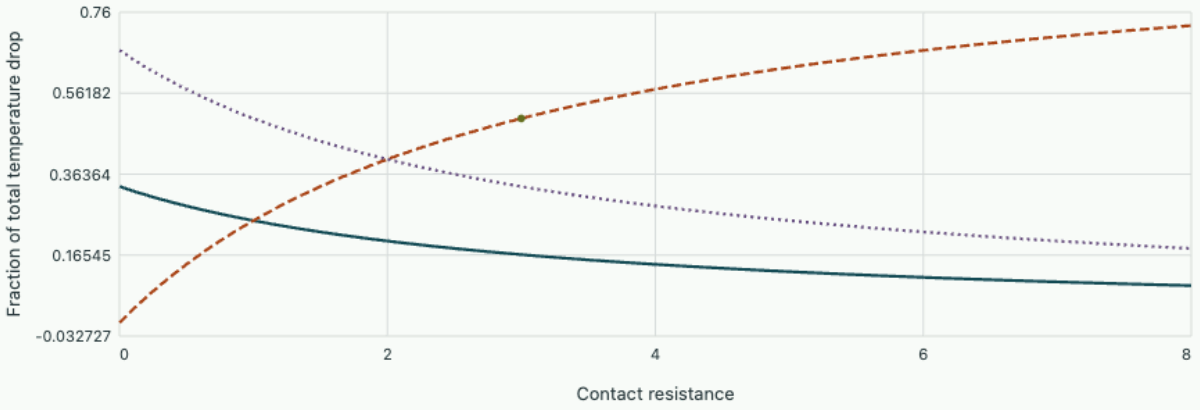
Change parameters · Separate the temperature jump from the bulk gradient

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 bulk resistance	1
Second bulk resistance	2
Selected contact resistance	3



At selected contact resistance, the jump carries 50% of the imposed drop. The three fractions always sum to one.

Study the full derivation

Bound diffusion error by balancing a reconstructed flux · Spatial discretisation and field equations

A temperature field from a finite-element calculation looks smooth. Reconstruct a heat flux which exactly balances the source, then derive an upper bound for the true energy error without knowing the exact temperature.

Chapter 8 · Anisotropic diffusion and mesh alignment

Relate a diffusion tensor to principal directions and test whether a face-flux approximation can represent a rotated material.

FROM FIRST PRINCIPLES

Derive the anisotropic diffusion operator from a rotated material law

Choose the principal axes of a homogeneous medium. Along them the measured flux response is diagonal, so heat conduction in direction 1 has coefficient $k_1 > 0$ and in direction 2 has $k_2 > 0$. Let orthogonal matrix R map principal-axis components into laboratory coordinates. Gradients rotate back with R^{T}, while the calculated flux rotates forward with R. The material law in laboratory axes is therefore

$$\mathbf{q} = -K\nabla T, \qquad K = R\begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix}R^{\mathsf{T}}.$$

Insert this vector law into local steady conservation $\nabla \cdot \mathbf{q} = s$, with s a positive volumetric source. Expanding for constant K gives the mixed derivative explicitly:

$$-K_{xx}T_{xx} - 2K_{xy}T_{xy} - K_{yy}T_{yy} = s.$$

Positive principal conductivities make $\nabla T^{\mathsf{T}} K \nabla T$ nonnegative, the basic energy property that a spatial discretization should respect. In a rotated material, the flux need not be perpendicular to an isotherm. A face formula using only the difference of two temperatures separated along the face normal may therefore omit a real transverse-gradient contribution.

DERIVATION AND APPLICATIONS

When a two-point face flux is exact for anisotropic diffusion

A discretization can preserve conservation and still send heat in the wrong direction. Consider a planar face with area A, outward unit normal $\mathbf{n}$ from cell P toward cell N, and centre displacement $\mathbf{d} = \mathbf{x}_N - \mathbf{x}_P$. Let the constant material tensor K be symmetric positive definite. For a linear temperature field $T = T_0 + \mathbf{g}\cdot\mathbf{x}$, the continuum face rate and measured neighbour difference are

$$F = -A\mathbf{n}^{\mathsf{T}}K\mathbf{g} = -A(K\mathbf{n})\cdot\mathbf{g},$$
$$T_N - T_P = \mathbf{d}\cdot\mathbf{g}.$$

A candidate two-point law is

$$F_2 = -\tau(T_N - T_P).$$

For it to reproduce the exact flux for every possible gradient, the two vectors multiplying $\mathbf{g}$ must be identical. This gives a joint mesh-material condition:

$$\tau \mathbf{d} = AK\mathbf{n}.$$

On a grid aligned with a principal axis and with cell centres aligned along the normal, this relation gives the familiar transmissibility. Otherwise split $AK\mathbf{n}$ into a component parallel to $\mathbf{d}$ and a remainder $\mathbf{c}$ perpendicular to it. Taking a scalar product with $\mathbf{d}$ yields

$$\tau = A(K\mathbf{n}) \cdot \mathbf{d} / |\mathbf{d}|^2,$$
$$\mathbf{c} = AK\mathbf{n} - \tau \mathbf{d},$$
$$F = -\tau(T_N - T_P) - \mathbf{c} \cdot \mathbf{g}.$$

The correction requires a second gradient component, usually reconstructed from more cells. Using one corrected flux with opposite signs for the two neighbours still preserves conservation. The derivation checks consistency for an exactly linear field; monotonicity and stability require separate tests.

WORKED EXAMPLE

Test a Cartesian face in a tilted material

Let $A = 1.00$ m², $\mathbf{d} = (0.200, 0)$ m, $\mathbf{n} = (1, 0)$ and K have entries $(K_{xx}, K_{xy}, K_{yy}) = (4, 1, 2)$ W/(m K). For $\mathbf{g} = (0, 3.00)$ K/m the neighbour difference vanishes, yet the exact outward flux is -3.00 W. The decomposition gives $\tau = 20.0$ W/K and $\mathbf{c} = (0, 1)$ in compatible units. The entire flux therefore comes from the reconstructed transverse gradient.

SECOND APPLICATION

Choose a skew cell arrangement that needs no correction

Keep the same tensor and normal but set $\mathbf{d} = (0.200, 0.0500)$ m. It lies along $(4, 1) = K\mathbf{n}$, so $\tau = 20.0$ W/K and $\mathbf{c} = 0$. For a gradient $(2.00, -1.00)$ K/m the neighbour temperature rise is 0.350 K, giving -7.00 W. Direct evaluation $-[4(2) + 1(-1)]$ agrees. Geometric orthogonality to the physical face is not required when the tensor itself tilts the relevant direction.

Begin with an energy and a constitutive law

In a conducting sheet with directional properties, heat flux is

$$\boldsymbol{q} = -K\nabla T,$$

where K is symmetric positive definite. With constant volumetric capacity, conservation gives a diffusion operator proportional to $\nabla \cdot (K \nabla T)$. For insulating boundaries and no sources, multiplying by T and integrating by parts gives a decay term $-\int \nabla T^T K \nabla T$. It is nonpositive because the eigenvalues of K are positive. This energy test checks the sign of a discrete spatial operator even when flux and gradient do not point along the same line.

Let the principal conductivities be 4 and 1. Rotating the principal axes through 45 degrees gives

$$K = R\,\mathrm{diag}(4,1)R^T = \begin{pmatrix} 2.5 & 1.5 \\ 1.5 & 2.5 \end{pmatrix}.$$

Along the diagonal direction $(1,1)/\sqrt{2}$, multiplying by K scales the gradient fourfold; along the other diagonal the scale is one. Along the Cartesian x direction a transverse flux component appears. The off-diagonal entries express a real constitutive response and cannot be dropped by refining the same incorrect stencil.

A patch test for a face flux

Use the exact affine temperature $T(x,y) = y$, whose gradient is $(0,1)^T$. The flux is $(-1.5, -2.5)^T$. A vertical face with outward normal e_x therefore carries -1.5 units of flux per unit length. A two-point formula based only on temperatures at the left and right cell centres sees identical temperatures and returns zero. Both fields have zero divergence in the interior because each flux is constant, so checking interior residuals alone would miss the incorrect boundary transport. Compare individual face fluxes as well as cell residuals in a linear patch test.

To recover the normal flux one can reconstruct the full gradient using neighbouring data, or choose a weak form integrating $(\nabla v)^T K \nabla T$ over each element. On a triangle with affine basis functions the gradients are constant, so constant K yields an element matrix $A_{ij} = |E| \nabla \phi_i^T K \nabla \phi_j$. Symmetry is immediate and the row sums vanish because the basis functions sum to one.

Distinguish length scales from transport axes

For constant K, a Fourier mode $e^{i \xi \cdot x}$ decays with factor $e^{-t \xi^T K \xi}$ when capacity is one. At unit wave number, a mode along the fast principal axis falls to $e^{-4} \approx 0.0183$ after one unit of time; a perpendicular one retains $e^{-1} \approx 0.368$. This test excites the two conductivities separately. On an elongated mesh the mesh dimensions also affect the fastest discrete time scale, so the same physical tensor can require different explicit step sizes on two grids.

Rotate the temperature, keeping the material fixed

The directional heat transport along a unit vector n is controlled by $n^T K n$ when the imposed temperature gradient is parallel to n. For the worked tensor and $n = (1,0)^T$, this effective conductivity is 2.5 even though the full flux has a transverse component. If the specimen instead has insulated long sides, then its transverse gradient can adjust until transverse flux vanishes. Solving $1.5 T_x + 2.5 T_y = 0$

gives $T_y = -0.6T_x$ and effective longitudinal flux $-1.6T_x$. The same material therefore gives different one-dimensional reductions under different boundary constraints.

Construct a positive directional stencil for a rotated tensor

For constant $K = \left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$, the diffusion operator is $au_{xx} + 2bu_{xy} + cu_{yy}$. On a square grid, the centred second difference along the lattice diagonal $(1, 1)$, divided by h^2, approximates $(\partial_x + \partial_y)^2 u$, including the required mixed derivative with positive sign. For $b \geq 0$ one can therefore combine the three directional second differences with coefficients $a - b, c - b, b$ along the horizontal, vertical and positive diagonal directions. Expanding these terms recovers the desired operator for every quadratic field.

If all three coefficients are nonnegative, a local maximum has a nonpositive spatial diffusion update. This sign property supports a discrete maximum principle and, with a sufficiently small explicit time step, a positive averaging update. For negative b, the opposite diagonal changes the mixed derivative sign and the analogous coefficients involve $|b|$. Positive definiteness of the continuum tensor only requires $ac > b^2$, which is weaker than both $a \geq |b|$ and $c \geq |b|$. A physically valid material can therefore defeat this particular compact positive stencil.

Reach a strongly tilted direction using a wider stencil

Set $a = 1,\ b = 1.5,\ c = 4$. The determinant $4 - 2.25 = 1.75$ is positive and both eigenvalues are positive. Yet the compact three-direction construction would assign horizontal coefficient -0.5. To see the numerical problem, give the centre value zero, both horizontal neighbours value one and all diagonal and vertical neighbours zero. The centre is a minimum, but the compact operator returns $-1/h^2$. An explicit positive time step would drive that minimum below zero, contradicting the smoothing behaviour expected from a monotone diffusion discretisation.

Consider instead a second difference along the longer lattice vector $(1, 2)$. Dividing by h^2 with coordinate increments $h, 2h$ approximates $u_{xx} + 4u_{xy} + 4u_{yy}$. Multiplying it by 0.75 produces the desired mixed coefficient three. The remaining horizontal and vertical coefficients are $1 - 0.75 = 0.25$ and $4 - 3 = 1$, both positive. At an interior node this wider stencil has all positive neighbour couplings. On the quadratic patch $u = xy$ the exact operator value is three; the long diagonal evaluates the centred difference as four, and its weighted contribution is exactly three while both axial differences vanish.

The longer reach introduces a boundary question: a point one row from the edge may lack a neighbour two rows away. Replacing it ad hoc with a nearest available value can destroy consistency or positivity. A boundary treatment, a mesh aligned to the material axes or a compatible finite-element formulation must address that issue separately. Selecting a positive interior stencil alone does not prove a complete domain discretisation monotone.

For the wider stencil an explicit update at an interior node has centre coefficient $1 - 2(0.25 + 1 + 0.75)\Delta t/h^2$ for unit heat capacity. Preserving nonnegative weights therefore

requires $\Delta t \leq h^2/4$ in the units of this example. Positive face couplings are necessary for this argument, while a compatible time step supplies the remaining condition.

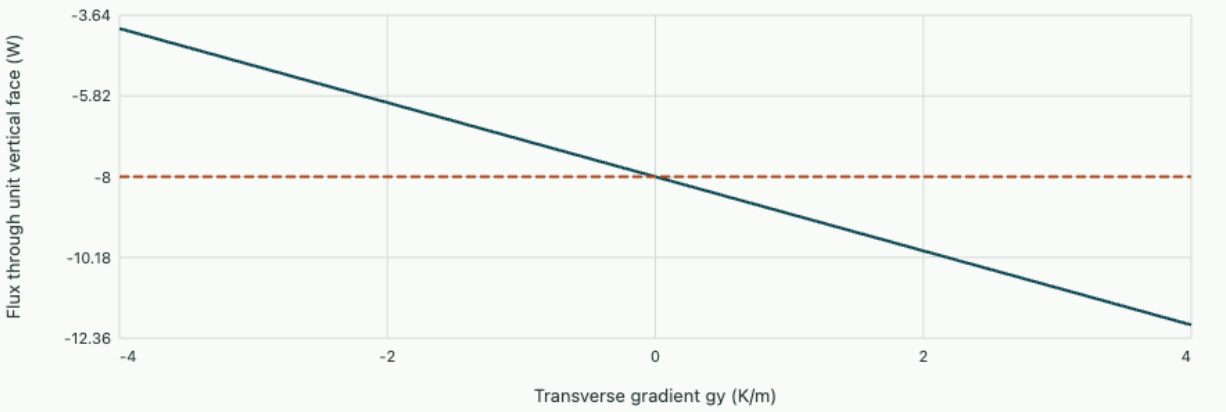
Change parameters · Explore the worked calculation

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

Tensor cross term Kxy	1
Neighbour offset dy (m)	0
Longitudinal gradient gx (K/m)	2



Full tensor flux Uncorrected two-point flux

Correction coefficient Kxy – τdy = 1. It vanishes when neighbour displacement aligns with K n, at dy = 0.05 m. Tensor entries Kxx = 4, Kyy = 2 keep all offered tensors positive definite.

Problem sheet 4 · Material interfaces and anisotropy

Recover face temperatures and fluxes, and use a linear temperature field to test a rotated conductivity tensor.

Add an interface contact

Two cell centres at 45 and 15 °C each lie 0.02 m from a contact. The conductivities are 4 and 2 W/(m K), and the contact resistance is 0.005 m²K/W. Find flux, both surface temperatures and the conducted power over 0.03 m².

▼ **Worked solution**

The resistances are $0.02/4 = 0.005,\ 0.005,\ 0.02/2 = 0.010$ m²K/W in series. Total 0.020 gives $q = 30/0.020 = 1500$ W/m² toward the colder cell.

The hot-side face is $45 - 1500(0.005) = 37.5$ °C. The cold-side face is $37.5 - 7.5 = 30$ °C, and the last drop is 15 K. Power is $1500(0.03) = 45$ W.

Explore the related calculation with adjustable inputs. Vary series resistances and observe their fractions of the total drop. Divide by the stated bulk resistance when converting dimensional values to a comparison.

Identify a convective resistance

The centre temperature is 50 °C, ambient is 20 °C and outward flux is 1200 W/m². The half-cell resistance is 0.01 m²K/W. Determine surface temperature and boundary heat-transfer coefficient. Test the prediction after halving the centre-to-face distance while keeping centre and ambient values fixed.

▼ **Worked solution**

The surface is $50 - 1200(0.01) = 38$ °C. Boundary resistance is $(38 - 20)/1200 = 0.015$ m²K/W, giving $h_c = 66.67$ W/(m²K).

The new total resistance is $0.005 + 0.015 = 0.020$, so flux is 1500 W/m² and surface temperature is 42.5 °C. The material and boundary laws were held fixed; the unchanged centre value now belongs to a different physical location.

Explore the related calculation with adjustable inputs. Vary series resistances and observe their fractions of the total drop. Divide by the stated bulk resistance when converting dimensional values to a comparison.

Test a rotated flux

Take $K = \left(\begin{smallmatrix} 3 & -1 \\ -1 & 3 \end{smallmatrix}\right)$, $T = 2x + y$. Find both flux components and eigenvalues. Which diagonal plane-wave direction decays faster, and what normal flux should an east-facing face record?

▼ **Worked solution**

$K\nabla T = (5, 1)^T$, so the flux is $(-5, -1)^T$. The principal conductivities are 2 along $(1, 1)$ and 4 along $(1, -1)$.

With equal wave numbers and unit capacity, the second diagonal decays twice as rapidly in its exponent. An east-facing face has normal e_x and records –5. Dropping the off-diagonal coefficient would instead give –6.

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

Notebook practice · Touch two long solids: which temperature does the interface acquire?

Chapter 9 · Cell Péclet number and monotonicity

Derive the sign restriction for centred transport, identify a numerical boundary layer and quantify the diffusion introduced by upwinding.

FROM FIRST PRINCIPLES

Derive the Péclet restriction for a centred steady stencil

Take steady one-dimensional transport with constant speed $a > 0$, diffusivity $D > 0$ and no interior source. Conservation requires constant total flux $J = au - Du'$. On a uniform mesh of spacing h, approximate the advective face value by the arithmetic mean and the face gradient by a difference. The right face flux becomes

$$J_{i+1/2} = a(u_i + u_{i+1})/2 - D(u_{i+1} - u_i)/h.$$

Equating the two neighbouring face fluxes and collecting the centre coefficient gives a relation among the three consecutive values:

$$2u_i = (1 + \mathrm{Pe}_h/2)u_{i-1} + (1 - \mathrm{Pe}_h/2)u_{i+1},$$
$$\mathrm{Pe}_h = ah/D.$$

For $|\mathrm{Pe}_h| \le 2$ the centre value is a positive weighted average of its neighbours. Above that threshold one weight becomes negative and the discrete solution can overshoot even when the continuum solution is monotone. This restriction is on the local cell Péclet number, not on aL/D for the whole physical domain. Smaller cells, an upwind flux or an exponentially fitted flux can change the discrete behaviour.

DERIVATION AND APPLICATIONS

Distinguish an imposed inlet flux from an imposed inlet concentration

With diffusion, a concentration measurement alone does not specify the total material entering a channel. Consider steady one-dimensional advection at constant speed $a > 0$ and diffusion coefficient $D > 0$, constant area, no reaction, and coordinate $0 < x < L$. Concentration is $u(x)$. The net species flux is the sum of advection and diffusive backflow. Conservation makes it constant throughout the interval:

$$J = au - Du',$$
$$J' = 0.$$

Suppose the upstream supply delivers a specified total flux $J = au_f$, where u_f is the incoming feed concentration used to normalize the supply rate. Prescribe an outlet concentration $u(L) = u_L$

maintained by a reservoir. Subtract the uniform particular solution u_f. The resulting first-order homogeneous equation for the excess has a convenient outlet-referenced solution:

$$(u-u_f)' = (a/D)(u-u_f),$$
$$u(x) = u_f + (u_L - u_f)e^{-a(L-x)/D}.$$

At the inlet the interior concentration can differ from the feed because diffusion also contributes to the prescribed mass rate. Define $\mathrm{Pe} = aL/D$ and evaluate both terms at $x = 0$:

$$u(0) = u_f + (u_L - u_f)e^{-\mathrm{Pe}},$$
$$-Du'(0) = -a[u(0) - u_f].$$

For large Péclet number the outlet affects only a layer of thickness D/a. For small Péclet number the outlet reservoir can affect the whole channel. Imposing both $u(0) = u_f$ and the specified total flux at the inlet would impose more conditions than this problem needs. A numerical method should record which physical boundary condition it approximates.

WORKED EXAMPLE

Compute a moderate-Péclet-number channel profile

Take $L = 1.00$ m, $a = 0.0200$ m/s, $D = 0.0100$ m²/s, $u_f = 1.00$ and $u_L = 0$ kg/m³. The inlet concentration is $1 - e^{-2} = 0.8647$ kg/m³ and the midpoint is $1 - e^{-1} = 0.6321$ kg/m³. Inlet advection carries 0.01729 kg/(m² s) and diffusion supplies 0.00271 kg/(m² s), giving the prescribed total 0.0200.

SECOND APPLICATION

Quantify a boundary-condition error

If a code instead fixes $u(0) = 1.00$ and $u(L) = 0$, its exact constant flux for those Dirichlet data satisfies

$$u(0) = (J/a)(1 - e^{-\mathrm{Pe}}).$$

Hence $J = 0.0200/(1 - e^{-2}) = 0.0231$ kg/(m² s), about 15.7% too large for the original feed problem. Grid refinement can converge cleanly to that incorrect-for-purpose value because the error lies in the boundary model.

Which length sets the balance?

Model a steady scalar on $[0, L]$ with $au' - \kappa u'' = 0$, constant speed $a > 0$ and diffusivity $\kappa > 0$. Advection and diffusion balance over length $\ell = \kappa/a$. On a mesh with spacing h, the cell Péclet number $P = ah/\kappa$ measures how many layer lengths fit into one cell. This definition uses the full cell spacing; some conventions divide it by two. Stating the convention matters when quoting a threshold.

Write the centred interior equation in flux form. Its coefficients multiplying u_{j-1}, u_j, u_{j+1} are $-\kappa/h^2 - a/(2h), 2\kappa/h^2, -\kappa/h^2 + a/(2h)$. For $P \le 2$, the neighbour coefficients are nonpositive and the row sum is zero, so each interior value is a weighted average of its neighbours. For $P > 2$, one weight changes sign. The discrete maximum principle can fail even though the centred derivative is formally second-order accurate on smooth resolved fields.

Watch a node cross below the boundary range

Take $L = 1, a = 1, \kappa = 0.1$, $u(0) = 0, u(1) = 1$ and two equal cells. Then

$$h = 0.5, P = 5.$$

At the only interior node the coefficients are $-1.4, 0.8, 0.6$. The numerical value is -0.75, outside the prescribed boundary range. The exact solution is $(e^{ax/\kappa} - 1)/(e^{aL/\kappa} - 1)$, increasing and close to zero except near the outlet. At $x = 0.5$ it is $1/(e^5 + 1) \approx 0.00669$. The coarse grid does not resolve a physical layer of width 0.1.

Using the upstream difference $(u_j - u_{j-1})/h$ changes the coefficients to $-2.4, 2.8, -0.4$; the computed midpoint is $1/7$. The value lies in range but the layer is excessively broad. A monotone method can replace an oscillation with a large smoothing error. Plot the magnitude of the profile at fixed physical locations while refining to distinguish those outcomes.

Explain the extra diffusion

Expand $u_{j-1} = u_j - hu'_j + h^2u''_j/2 + \cdots$. The upstream derivative is $u'_j - hu''_j/2 + O(h^2)$. The leading modified equation therefore contains effective diffusivity $\kappa + ah/2$. In the example the artificial addition is 0.25, larger than the physical diffusivity 0.1. Reducing h below 0.2 restores the coefficient-sign requirement for the centred formula, while resolving the layer shape accurately generally needs several cells across ℓ. A convergence study should report both bounds compliance and the actual layer width or flux.

Fit the face flux to a constant-coefficient cell

For smooth constant coefficients the total flux $J = au - \kappa u'$ is constant in a source-free cell. Solve this first-order equation between endpoints separated by h. The exact relation is $u_R = e^P u_L + (J/a)(1 - e^P)$, hence $J = a(e^P u_L - u_R)/(e^P - 1)$. At large positive P, the flux approaches au_L, the value carried from upstream. As P tends to zero, expansion gives a diffusive difference $-\kappa(u_R - u_L)/h$ plus the centred advective contribution $a(u_L + u_R)/2$. This exponential fitting uses additional local knowledge of the differential equation. It reproduces the constant-coefficient cell relation exactly, but variable material properties and sources require new local approximations.

Allocate grid points to a singularly perturbed outflow layer

For steady positive advection and small diffusivity the outflow layer shape is approximately $e^{-(L-x)/\ell}$, with physical thickness scale $\ell = \kappa/a$. A uniform grid that resolves that scale at a fixed number of points may waste most of its cells where the solution is almost flat. A simple layer-adapted strategy divides the domain into a coarse interval and a fine interval ending at the outflow boundary. Choose the fine-region width τ by specifying how small the layer should already be at their interface.

If $e^{-\tau/\ell} = N^{-2}$ for a grid with N total cells, then $\tau = 2\ell \log N$. Capping this width at $L/2$ prevents a separate-layer construction when the physical layer is no longer thin. Place half the cells on each subinterval. This produces a piecewise-uniform grid, often called a Shishkin mesh for this boundary-layer setting. The amplitude threshold motivates its construction; it is not itself a proof of a particular numerical convergence rate. Such a proof must also control truncation, stability and matching across the grid transition.

Compare the information captured by sixty-four cells

Consider $L = 1$ m, $a = 1$ m/s and $\kappa = 0.002$ m²/s. Then $\ell = 0.002$ m. A uniform mesh of 64 cells has spacing 0.015625 m, almost eight physical layer lengths per cell. The first interior point back from the outflow has layer amplitude about $e^{-7.8125} \approx 0.000405$. Most of the transition from nearly zero to one takes place in a single cell. A method may fit an acceptable boundary value while giving a poor near-wall shape or slope.

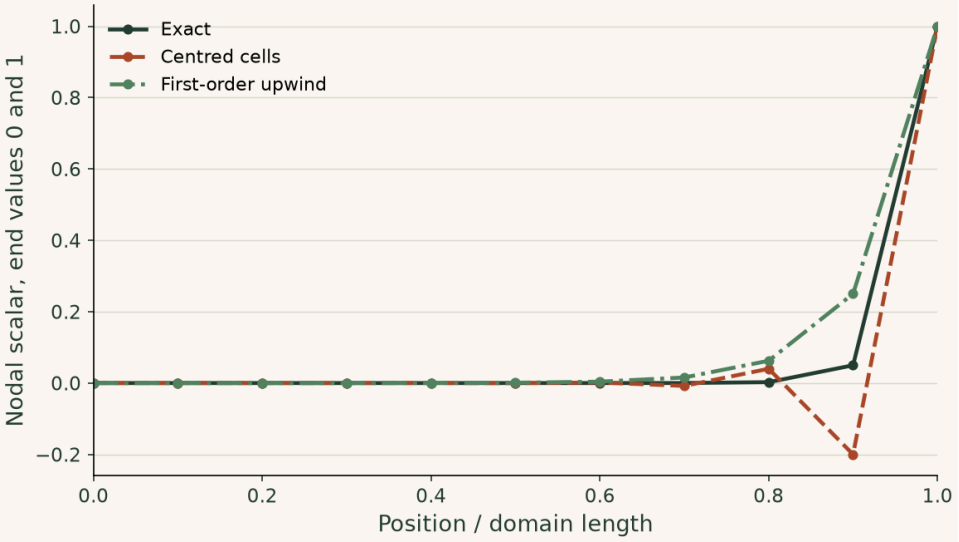
The two-region design instead places its transition $\tau = 0.004 \log 64 \approx 0.01664$ m before the end. With 32 fine cells, their width is approximately 0.000520 m, so the fine cell Péclet number is about 0.260. At the last interior point the layer amplitude is now $e^{-0.260} \approx 0.771$. Several distinct intermediate values are available near the wall. At the coarse-to-fine transition the amplitude is $64^{-2} \approx 0.000244$, just as designed.

The coarse cells on the remaining 0.98336 m have width about 0.03073 m and cell Péclet number about 15.4. Their small expected layer amplitude does not license an unstable or nonmonotone coarse discretisation. Use a spatial scheme whose behaviour is controlled there, and check flux continuity through the changing cell sizes. If advection reverses, the thin layer moves to the opposite boundary; retaining the old mesh would concentrate cells in the wrong region. In time-dependent transport, moving layers may similarly call for adaptive tracking instead of a fixed partition.

The wall derivative supplies a more demanding check than a prescribed boundary value. For the exponential layer its exact slope at outflow has magnitude $1/\ell = 500$ per metre. A backward difference across the final fine cell gives approximately $(1 - 0.771)/0.000520 = 440$ per metre. Even this substantially improved mesh still underestimates the wall slope by about twelve percent, which matters if wall flux is the reported observable.

Expose a monotonicity failure on ten cells

For constant positive advection and diffusion solve $u'' - \mathrm{Pe}\, u' = 0$ with end values zero and one. At cell Péclet number three, the centred recurrence has ratio $(1 + \mathrm{Pe}_h/2)/(1 - \mathrm{Pe}_h/2) = -5$; the upwind ratio is four.



The centred solution overshoots the allowed interval by oscillating below zero near the outflow. Upwinding gives a monotone but overly broad layer. Neither the small algebraic residual nor a monotone curve alone establishes spatial accuracy.

▶ **Read values from the plot**

Change parameters · Expose a monotonicity failure on ten cells

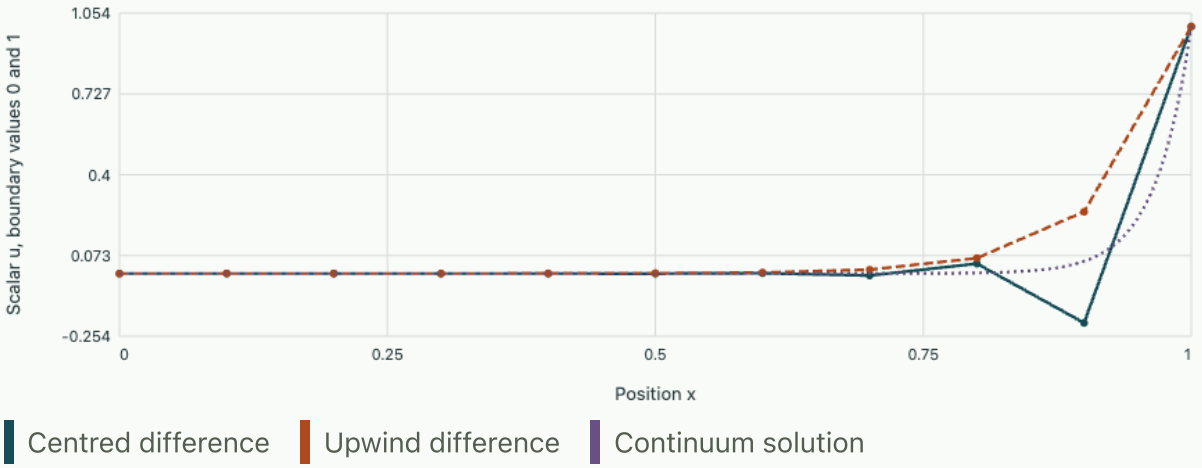
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Cell Péclet number 3

Mesh cells 10



Total Péclet number 30. The centred minimum is -0.2. The monotonicity threshold is cell Pe = 2.

Study the full derivation

Choose the physically admissible wave in a conservative update · Spatial discretisation and field equations

A discontinuous scalar flow can satisfy integral conservation in more than one way. Derive the shock speed, use a convex entropy to distinguish a compression from an expansion and build a finite-volume flux from the resulting local solution.

Chapter 10 · Incompressibility and discrete pressure modes

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.

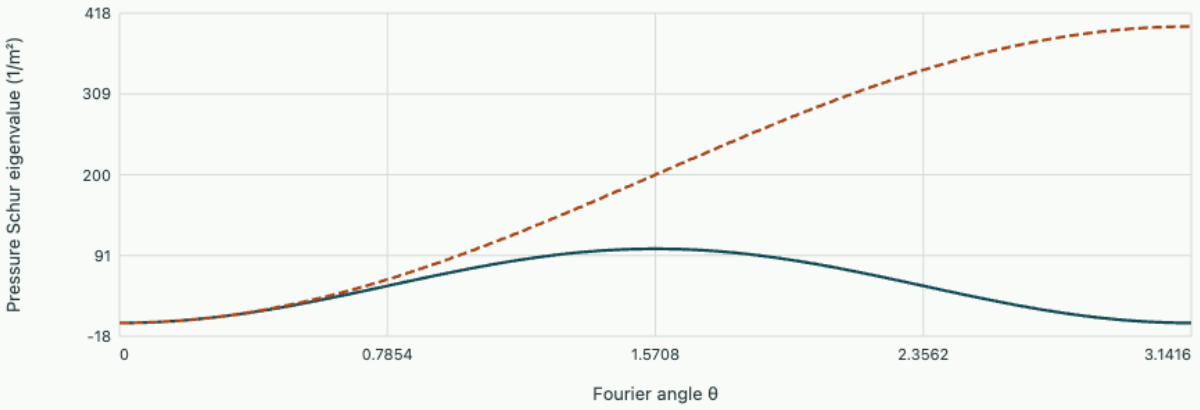
Change parameters · Explore the worked calculation

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.

Problem sheet 5 · Transport weights and pressure spaces

Locate the onset of a sign error and test which discrete pressure differences actually drive a velocity.

Resolve a layer

Let $a = 3$ m/s, $\kappa = 0.06$ m²/s on a 1 m interval. Find the minimum number of equal cells for the centred operator's nonpositive neighbour coefficients, and the number giving five cells per physical layer length. Estimate the upwind artificial diffusivity at each resolution.

▼ **Worked solution**

The layer length is

$$\ell = \kappa/a = 0.02\,\mathrm{m}.$$

The sign test requires $h \le 2\ell = 0.04$ m, hence at least 25 cells. Five cells across ℓ require $h \le 0.004$ m or 250 cells.

The upwind addition $ah/2$ is 0.06 m²/s at 25 cells and 0.006 at 250. The monotonicity threshold and satisfactory physical resolution describe quite different demands.

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 transport direction

On an equally spaced grid take $a < 0$. Derive the upstream discrete derivative, its leading artificial diffusivity in $au' - \kappa u''$, and the boundary near which a thin outflow layer forms for fixed unequal Dirichlet values.

▼ **Worked solution**

With flow toward decreasing x, upstream information lies to the right, giving $(u_{j+1} - u_j)/h = u'_j + hu''_j/2 + O(h^2)$.

Multiplying by a negative speed gives the modified coefficient $-(\kappa + |a|h/2)u''$. On $[0, L]$ inflow is at L and outflow at zero. The term $e^{ax/\kappa}$ decays away from zero, locating a possible outflow boundary layer there.

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.

Find the pressure nullspace

For an even periodic grid with six pressure nodes, solve $p_{j+1} - p_{j-1} = 0$ for all j. Remove the constant ambiguity and give the dimension of the remaining nullspace. Repeat the argument for five nodes.

▼ Worked solution

The relation is

$$p_{j+2} = p_j.$$

Six nodes split into independent even and odd cycles, giving two arbitrary amplitudes. Removing the constant leaves one alternating mode.

With five nodes, increments of two cycle through every index, so all entries agree. Only the constant remains and zero mean removes it. This finite odd-grid observation by itself does not establish a mesh-independent mixed stability 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.

Chapter 11 · Multigrid and coarse-grid correction

Combine smoothing with a coarse residual equation and compute which components a two-grid correction can remove exactly.

FROM FIRST PRINCIPLES

Derive the coarse-grid correction as a residual equation

Let a linear discretization satisfy

$$A_h u_h = b_h.$$

For an iterate $\tilde{u}_h$ define error $e_h = u_h - \tilde{u}_h$ and residual $r_h = b_h - A_h \tilde{u}_h$. Subtracting the equations gives an exact identity:

$$A_h e_h = r_h.$$

A coarse mesh cannot represent arbitrary fine error. Suppose the error remaining after smoothing is approximated by $P e_H$, with prolongation P from coarse to fine. Choose coarse coefficients so the remaining residual is orthogonal to the coarse test directions; with restriction $R = P^{\mathsf{T}}$ this condition becomes

$$\begin{aligned} R(r_h - A_h P e_H) &= 0, \\ A_H e_H &= R r_h, \\ A_H &= R A_h P. \end{aligned}$$

Update the fine iterate by $\tilde{u}_h \leftarrow \tilde{u}_h + P e_H$. If the coarse solve is exact, this removes any error initially in the range of P. Components outside that range must be reduced by smoothing or another correction. This separation explains why simply smoothing a fine mesh for longer can be inefficient when long-wavelength error survives each relaxation sweep.

DERIVATION AND APPLICATIONS

Choose a Jacobi weight for the errors a coarse grid cannot resolve

A multigrid smoother should damp oscillations left invisible or ambiguous on the coarse mesh. Consider the uniform one-dimensional discretization of $-u''$ with stencil $(-1, 2, -1)/h^2$, away from boundaries or on a periodic grid. Let the error be a Fourier mode $e_j = e^{ij\theta}$. The Jacobi diagonal is $D = (2/h^2)I$. Inserting the Fourier mode into the discrete operator and dividing by D gives

$$(D^{-1}Ae)_j = \tfrac{1}{2}(2 - e^{i\theta} - e^{-i\theta})e_j = (1 - \cos\theta)e_j.$$

One weighted Jacobi step $\tilde{u} \leftarrow \tilde{u} + \omega D^{-1}(b - A\tilde{u})$ therefore multiplies the error by a real factor. With standard factor-two coarsening, call the interval $\pi/2 \le |\theta| \le \pi$ high frequency:

$$g_\omega(\theta) = 1 - \omega(1 - \cos\theta), \qquad g_\omega \in [1 - 2\omega, 1 - \omega] \quad (\omega > 0).$$

The maximum absolute value on this interval lies at an endpoint. To minimize that worst factor, set the two endpoint values equal in magnitude and opposite in sign. The solution is

$$\begin{aligned} 1 - \omega &= -(1 - 2\omega), \\ \omega &= 2/3, \\ \max_{\text{high}} |g_\omega| &= 1/3. \end{aligned}$$

An ordinary undamped Jacobi sweep ($\omega = 1$) leaves the checkerboard oscillating with unchanged magnitude. The optimal weight here depends on the operator, dimensionality and coarsening convention. It should not automatically be transferred to anisotropic grids or different stencils.

WORKED EXAMPLE

Calculate how many smoothing sweeps are needed

A pure high-frequency error initially has amplitude 0.900 K. With $\omega = 2/3$, three sweeps reduce it to at most $0.900/27 = 0.0333$ K. For the checkerboard the factor is $-1/3$, so after the third sweep the error has reversed sign. Four sweeps bound the amplitude by 0.0111 K. A residual norm insensitive to sign still sees the decay.

SECOND APPLICATION

Observe why smoothing alone is slow

At $\theta = \pi/16$, the same smoother gives

$$g = 1 - (2/3)(1 - \cos(\pi/16)) = 0.9872.$$

After ten sweeps a unit component still has amplitude $0.9872^{10} = 0.879$. A high-frequency component over the same ten sweeps is at most $3^{-10} = 1.69 \times 10^{-5}$. Coarse-grid correction targets the slowly decaying component that these sweeps leave behind.

Two kinds of persistent error

For a uniform Dirichlet grid let A_h be the positive three-point approximation to $-u''$. A smooth algebraic error changes slowly between adjacent nodes. Its local residual can be modest even when its amplitude is appreciable. High-frequency errors have large second differences and respond rapidly to a

suitable relaxation sweep. Weighted Jacobi uses $u \leftarrow u + \omega D^{-1}(f - A_h u)$, with diagonal $D = 2I/h^2$. On a sine mode of grid angle θ its error multiplier is $1 - \omega(1 - \cos\theta)$.

For $\omega = 2/3$, this multiplier lies between $-1/3$ and $1/3$ for angles between $\pi/2$ and π. Low angles have multipliers near one, so repeated sweeps alone cost increasingly many iterations as the grid is refined. Those surviving smooth errors can often be represented on a coarser mesh. The purpose of restricting a residual is to solve an error equation for that representation.

Derive the correction

Let P interpolate coarse values to the fine grid and R restrict a fine vector. Choose the Galerkin coarse operator $A_H = R A_h P$. Given current residual $r = f - A_h u$, solve $A_H e_H = R r$ and add $P e_H$. If the true error is precisely Pv, the right-hand side equals $A_H v$. An exact coarse solve recovers v and removes the entire error in one correction. For general errors the new residual satisfies $R r_{\text{new}} = 0$; high-frequency components may remain on the fine grid.

Compute one correction by hand

Take $h = 1/4$ with three interior fine nodes and one interior coarse node at the midpoint. Linear interpolation is $P = (1/2, 1, 1/2)^T$, and full weighting gives

$$R = P^T/2.$$

The fine operator is 16 times the tridiagonal matrix with diagonal 2 and off-diagonal –1. Multiplication gives $A_h P = (0, 16, 0)^T$ and $A_H = 8$. Suppose $u = 0$ and the exact fine solution is P, so the load is $(0, 16, 0)^T$. The restricted residual is 8. The coarse correction returns the coefficient one and removes the error exactly.

For error $e = (1, 0, -1)^T$, the fine residual is $32e$, whose restriction is zero. The coarse grid cannot correct this component. Its grid angle is $\pi/2$, so one weighted Jacobi sweep leaves $e/3$. These two vectors demonstrate why combining the two operations matters. Repeating the idea recursively gives a V-cycle. With a fixed number of smoothing steps per level, geometrically shrinking grids have total work proportional to the finest number of unknowns, provided coarse solvers and transfers also have linear cost. Good convergence rates still depend on smoothing and approximation properties for the actual operator.

Coarse linear solves do not have to be exact at every level

In the three-node example with fine error P, suppose an approximate coarse solve returns 0.9 instead of one. The correction leaves $0.1P$, so the restricted residual falls from 8 to 0.8. A recursive V-cycle effectively uses one approximate correction of this kind on intermediate levels; only the smallest system is usually solved directly. Measure the residual reduction per cycle across a family of grids. A roughly

constant work per unknown is valuable only if the number of cycles required for a fixed reduction also remains controlled.

Predict a smoothing failure in anisotropic diffusion

For the operator $-\alpha u_{xx} - \beta u_{yy}$ on an equal-spacing Cartesian grid, weighted point Jacobi has error multiplier

$$g(\theta_x, \theta_y) = 1 - \omega \frac{\alpha(1-\cos\theta_x) + \beta(1-\cos\theta_y)}{\alpha+\beta}.$$

This follows by substituting a Fourier error mode into the tridiagonal second differences in each direction and dividing by the diagonal $2(\alpha+\beta)/h^2$. Algebraic smoothness depends on the coefficients, while representability on a coarse mesh depends on geometric sampling. Under strong anisotropy a mode highly oscillatory in the weak direction can produce little residual relative to the diagonal and therefore survive point relaxation. A standard coarse interpolation space may also fail to represent it.

A line smoother solves all strongly coupled unknowns along a row together. Its block diagonal retains the horizontal second differences and the diagonal part of the vertical coupling. Modes constant along that row see no horizontal stiffness after applying the block, so the scale of their correction is set by the weaker vertical coupling rather than the large point diagonal. Alternatively, semicoarsening retains fine resolution in a direction in which troublesome oscillations must remain representable. The correct choice depends on the orientation of the anisotropy relative to the actual mesh.

Evaluate an error mode ignored by the ordinary two-grid cycle

Take a periodic grid with an even number of rows, $\alpha = 100,\ \beta = 1,\ \omega = 2/3$, and error $e_{ij} = (-1)^j$, constant along each row. Its discrete horizontal curvature is zero and its vertical curvature has eigenvalue $4/h^2$. Point Jacobi divides by $202/h^2$, giving $g = 1 - 4/(303) = 299/303 \approx 0.9868$. Ten sweeps still leave about 0.876 of the error amplitude. Geometrically the mode is as oscillatory as the grid permits, yet pointwise smoothing barely changes it.

Full weighting in the vertical direction uses weights $1/4, 1/2, 1/4$. Applied to the alternating signs it gives zero, so the restricted residual also vanishes and an ideal coarse solve receives no information about this error. This calculation identifies a concrete failure of the usual assumption that all components missed by coarsening have already been strongly damped. A high-quality coarse linear solve cannot repair a zero coarse right-hand side.

For a horizontal line Jacobi block the horizontal part contributes zero on a row-constant vector, leaving block eigenvalue $2/h^2$. The preconditioned error-mode eigenvalue is now $(4/h^2)/(2/h^2) = 2$, and a damped line sweep has multiplier $1 - (2/3)2 = -1/3$. Ten such sweeps reduce the amplitude to $1/59049$. Each line sweep costs more than a point sweep, so the practical comparison also depends on

implementation and geometry. The dramatic contrast in this mode calculation explains which coupling the extra solve is buying and why iteration counts alone are insufficient for benchmarking.

For diagonal anisotropy the identity of a strongly coupled row can change as the material rotates across the grid. A line relaxation designed only for horizontal coupling may then lose its advantage in another subdomain. Compare mode or patch tests aligned with several local principal directions before generalising the success of the horizontal example.

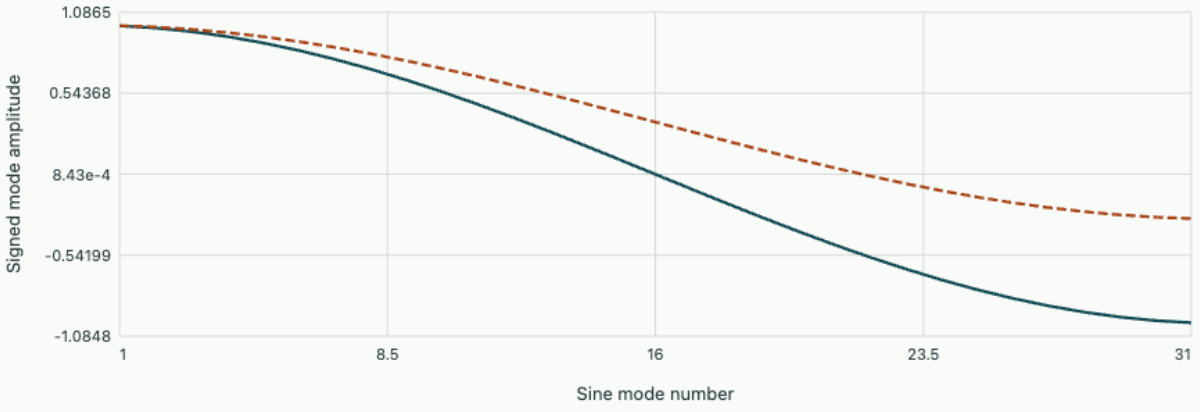
Change parameters · Explore the worked calculation

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

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.

Chapter 12 · Adjoints and errors in a chosen output

Weight equation residuals by the response of a chosen observation and distinguish algebraic error from mesh error.

FROM FIRST PRINCIPLES

Derive an output error from an adjoint solve

Consider a nonsingular discrete linear problem $Au = b$ with approximate solution $\tilde{u}$. A user may care about one output $J(u) = c^{\mathsf{T}}u$ rather than every entry of u. The unknown error $e = u - \tilde{u}$ satisfies $Ae = r$ with residual $r = b - A\tilde{u}$. Directly computing e would require another solve for a complete field. To weight the known residual for this output, select a vector z whose transposed action reproduces the output coefficients:

$$A^{\mathsf{T}}z = c,$$
$$c^{\mathsf{T}}e = z^{\mathsf{T}}Ae.$$

Substitution yields an exact algebraic output identity and a norm bound:

$$J(u) - J(\tilde{u}) = z^{\mathsf{T}}r, \qquad |J(u) - J(\tilde{u})| \le \|z\|_2\|r\|_2.$$

Large entries of z identify rows where a residual matters strongly to this particular observable, including possible sign cancellation between rows. A different output requires a different adjoint. The identity concerns error in solving the stated algebraic system. To estimate error relative to a continuum solution, discretization residuals and a suitable continuous or enriched adjoint must also enter.

DERIVATION AND APPLICATIONS

Weight a distributed residual for one point-temperature measurement

A local output need not be equally sensitive to residuals everywhere on the domain. Consider $-u'' = f$ on $0 < x < L$ with homogeneous Dirichlet boundaries and constant unit diffusivity. Suppose a smooth reconstructed numerical field $\tilde{u}$ satisfies both boundary values. Define the residual $r = f + \tilde{u}''$ so that the error $e = u - \tilde{u}$ obeys $-e'' = r$. Seek the error in one output $J(u) = u(\xi)$ at an interior location ξ.

Introduce a weight z that solves $-z'' = \delta(x - \xi)$ with zero end values. Away from ξ the weight is linear, and integrating the equation through the observation point determines its slope jump:

$$z(0) = z(L) = 0,$$
$$z'(\xi^-) - z'(\xi^+) = 1.$$

Continuity at ξ and the endpoint conditions determine both linear pieces. They give an explicit tent-shaped influence function

$$z(x) = \begin{cases} x(L-\xi)/L, & x \le \xi, \\ \xi(L-x)/L, & x \ge \xi. \end{cases}$$

Multiply the error equation by z and integrate by parts on the two intervals. The exterior terms vanish and the unit slope jump isolates the requested point value:

$$J(u) - J(\tilde{u}) = e(\xi) = \int_0^L z(x)\, r(x)\, dx.$$

A residual near the chosen output can have a larger effect than an equal residual near a prescribed boundary. Residuals of opposite signs can cancel for this observation without producing a globally accurate field. For piecewise linear finite-element fields, the distributional second derivative also contains jumps of the first derivative; omitting those terms would invalidate the identity.

WORKED EXAMPLE

Calculate a local temperature error

Take $L = 1$ m and $\xi = 0.250$ m. Suppose $r = 8.00$ K/m² on $0 < x < 0.250$ m and zero elsewhere. The relevant weight is $0.750x$, so the point error is

$$8(0.750)(0.250)^2/2 = 0.1875\,\mathrm{K}.$$

The positive residual means the exact temperature at the observation exceeds the reconstructed value.

SECOND APPLICATION

Move the same residual nearer the far boundary

Put the same 8.00 K/m² on $0.750 < x < 1.00$ m instead. Now the weight is $0.250(1-x)$, yielding $8(0.250)(0.250)^2/2 = 0.0625$ K. Both residuals have identical L^1 and L^2 magnitude; the first affects this particular sensor three times as strongly. Moving the sensor would change the ranking without changing either residual field.

A residual has a direction

Let $Au = b$ be an invertible linear discrete model and let $Q(u) = c^T u$ report a scalar output, for example a weighted mean temperature. For a computed approximation $\tilde{u}$, set $r = b - A\tilde{u}$. The error is $A^{-1}r$, and hence $Q(u) - Q(\tilde{u}) = c^T A^{-1} r$. Define the adjoint or dual solution by $A^T z = c$. Then the exact output identity is

$$Q(u) - Q(\tilde{u}) = z^T r.$$

A single adjoint gives the output effect of many possible residuals. A large residual where the observation has low sensitivity can contribute less than a small residual elsewhere. Opposite contributions can cancel in this particular output, even though they represent substantial state errors.

Compute the weight of each row

Take $A = \left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$, $b = (1, 0)^T$, and observe the second unknown with $c = (0, 1)^T$. Then $u = (2/3, 1/3)^T$ and $z = (1/3, 2/3)^T$. For $\tilde{u} = (0.6, 0.3)^T$, the residual is $(0.1, 0)^T$. Its weighted sum is $1/30$, exactly the missing part $1/3 - 0.3$. If a solve produced residual $(0, 0.1)^T$ instead, the same residual norm would double the output error. This output cares twice as much about the second row as about the first.

For residual $(0.2, -0.1)^T$, the signed estimate is exactly zero. The sum of absolute contributions is $2/15$, an upper bound insensitive to cancellation. An analyst interested only in the selected output can accept a calculation whose state vector another observer would reject. Specify the observation before choosing stopping and refinement criteria.

Move from an algebraic error to a mesh error

An accurately solved finite-element system has almost zero algebraic residual in its own discrete test space. It can still differ from the continuum solution. Let $a(u, v) = \ell(v)$ be the continuum weak equation and define $R(v) = \ell(v) - a(u_h, v)$. For a linear observation the continuum dual satisfies $a(v, z) = Q(v)$, giving $Q(u) - Q(u_h) = R(z)$. Galerkin orthogonality means $R(z_h) = 0$ for every test function on the same mesh, hence $R(z) = R(z - z_h)$.

Estimating this term requires a dual approximation richer than the test space used for the primal solve, or an independent reconstruction. Integrating element residuals by parts can expose both cell defects and jumps in normal flux. Weight these local contributions by the unresolved dual variation to discover which regions can affect the output strongly. Check an estimated correction against actual output changes on a refined mesh; a poor dual approximation can misidentify important regions even when the primal residuals have been evaluated exactly.

Compare a point observation with an average

In the two-unknown example, replace the second-node observation by the mean $c = (1/2, 1/2)^T$. The constant vector is an eigenvector of A with eigenvalue one, so the adjoint becomes $z = c$. A purely antisymmetric residual $(0.1, -0.1)^T$ makes the mean output error zero. For the earlier second-node output the same residual gives $-1/30$. A mesh designed to resolve an average can therefore allocate effort differently from one designed for a single sensor. Whenever a reported error changes markedly

between observables, inspect the dual weights before attributing the change to a flaw in the physical simulation.

Turn primal and dual approximation errors into an output bound

For a symmetric coercive bilinear form a, let the exact state u satisfy $a(u, v) = \ell(v)$, and the exact dual z satisfy $a(v, z) = Q(v)$ for a linear output Q. With conforming Galerkin approximations on the same trial space, $a(u - u_h, v_h) = 0$ for every discrete test. Consequently $Q(u) - Q(u_h) = a(u - u_h, z - z_h)$. Cauchy–Schwarz in the energy inner product gives an absolute bound by $\|u - u_h\|_a \|z - z_h\|_a$. Accuracy of either the state or the observation sensitivity can separately make the output error small.

When the primal and dual energy errors both decrease proportionally to h, their product guarantees an $O(h^2)$ output error under these assumptions. It does not guarantee second order for every possible observation. A point evaluation in a higher-dimensional continuum problem may be unbounded on the energy space or yield a dual solution with a strong singularity. Then the dual error may converge much more slowly. Likewise nonconforming approximation, inexact integration or inconsistent boundary terms introduce additional residual terms that are absent from the simple Galerkin cancellation.

A case where the product bound is exact

Let $a(v, w) = v^T A w$ with $A = \operatorname{diag}(2, 8)$, exact state $u = (1, 2)^T$, and reduced approximation space spanned by $s = (1, 1)^T$. The Galerkin coefficient is

$$(s^T A u)/(s^T A s) = (2 + 16)/10 = 1.8.$$

Choose output $Q(v) = (6, 8)v$; its exact dual is $z = (3, 1)^T$ since $Az = (6, 8)^T$. Projecting that dual onto the same space gives coefficient $(6 + 8)/10 = 1.4$. The primal and dual errors are therefore $(-0.8, 0.2)^T$ and $(1.6, -0.4)^T$.

Their energy inner product is

$$2(-0.8)(1.6) + 8(0.2)(-0.4) = -3.2.$$

Direct evaluation gives $Q(u) = 22$ and $Q(u_h) = 25.2$, confirming the signed error exactly. Squared energy errors are

$$2(0.8)^2 + 8(0.2)^2 = 1.6,$$

$$2(1.6)^2 + 8(0.4)^2 = 6.4.$$

Their square-root product is 3.2, attaining the absolute bound because the two error vectors are antiparallel. The bound cannot determine the sign; the signed dual residual can.

Replace the observation by $Q_2(v) = (2, 8)v$. Its exact dual is s, already representable, so $z - z_h = 0$ and the output error vanishes even though the state error is unchanged. Indeed $Q_2(u) = 18 = Q_2(u_h)$. A mesh judged solely by that particular observable could look exact. A different sensor weighing the same simulation states produces a demonstrably nonzero output error, which is why adaptive refinement objectives should name their actual observations.

If the approximate dual used the same one-dimensional space and one simply evaluated the primal residual on it, the reported correction would be exactly zero for either output. The first observation shows that such a number can be badly misleading. An enriched dual must include a direction outside the primal test space; in this two-coordinate example, adding either coordinate direction gives the full space and recovers the signed error exactly. On a fine physical mesh full enrichment may be too expensive, making local higher-order reconstruction or an independently refined dual mesh attractive. The product bound then clarifies which additional error in the dual approximation needs to be controlled.

It also identifies the observation for which refining the primal mesh alone is unnecessary.

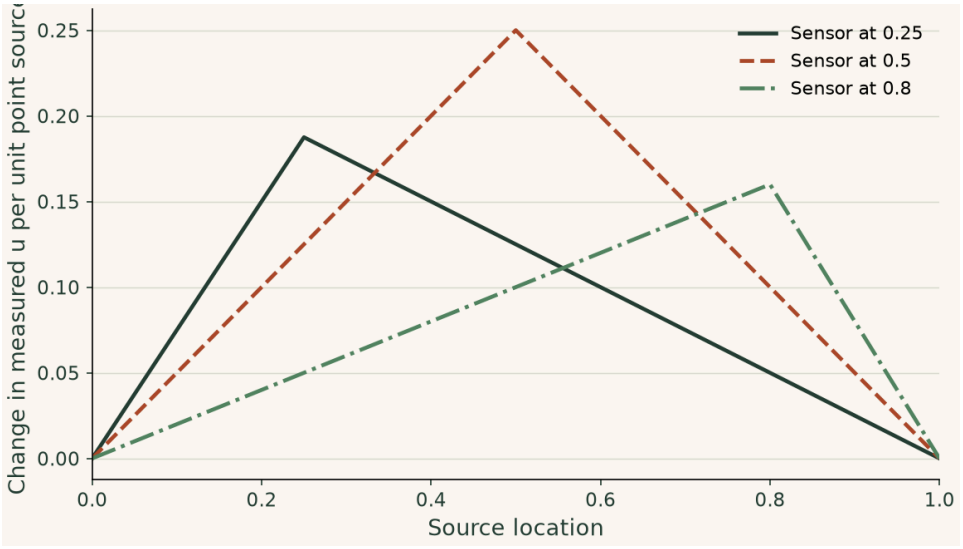
Bound the uncertainty of an approximate dual correction

A practical enriched dual $\tilde{z}$ may itself have numerical error. In the symmetric positive definite algebraic setting the difference between true output error and its estimated correction is $(z - \tilde{z})^T r$, where $r = b - A u_h$. With dual residual $r_z = c - A\tilde{z}$, this equals $r_z^T A^{-1} r$. Cauchy–Schwarz bounds its magnitude by $\|r_z\|_{A^{-1}} \|r\|_{A^{-1}}$. This identity explains why a small ordinary residual norm can be insufficient for either the primal or dual solve: directions associated with small eigenvalues receive greater weight in the inverse-energy norm.

In the preceding example choose $\tilde{z} = (2.5, 1)^T$. The primal residual is $(-1.6, 1.6)^T$; the approximate correction is therefore $-4 + 1.6 = -2.4$, moving the coarse output 25.2 to 22.8. The actual output is 22, leaving error −0.8 after correction. The dual residual is $(1, 0)^T$, with squared inverse-energy norm $1/2$. The squared primal residual norm in the same metric is 1.6. Their product of norms is $\sqrt{0.8} \approx 0.8944$, containing the observed remaining discrepancy. For larger problems, estimating these norms can itself require a separate bound or preconditioner. Reporting only the corrected value without assessing the accuracy of the dual can make a visually improved output appear more certain than the calculation warrants.

Which source errors matter to a point measurement?

For $-u'' = f$ with zero values at both ends of the unit interval, seek the sensitivity of $u(a)$ to a point load. The adjoint is piecewise linear, vanishes at the endpoints, is continuous at a and has derivative jump minus one.



For a sensor at 0.25, an equal positive source error at 0.25 has three times the effect of one at 0.75. Opposite-signed local residuals may cancel in the target output, even when their unweighted norm is appreciable.

▶ **Read values from the plot**

Change parameters · Which source errors matter to a point measurement?

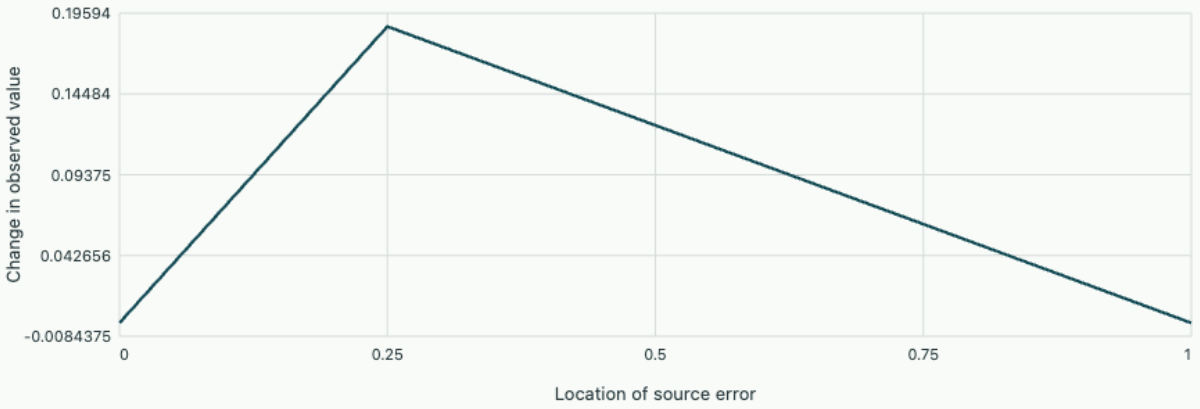
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Sensor coordinate a **0.25**

Point-source error strength **1**



Adjoint Green sensitivity

Sensitivity peaks at the sensor: 0.1875. Source errors at opposite signs can cancel in this output while their norm remains large.

Problem sheet 6 · Transfers and observation error

Track which error can survive a grid transfer and compute an exact adjoint correction for a nonsymmetric system.

A smooth mode survives smoothing

Weighted Jacobi for a Dirichlet Laplacian uses $\omega = 2/3$. Compute the multiplier at grid angles $\pi/3$ and $2\pi/3$, and the error left by two sweeps. Explain how this affects the choice of a coarse-grid space.

▼ **Worked solution**

For $\pi/3$ the multiplier is $1 - (2/3)(1 - 1/2) = 2/3$, leaving $4/9$ after two sweeps. At $2\pi/3$ it is $1 - (2/3)(1 + 1/2) = 0$, so the error vanishes in one sweep.

Smoother errors can persist while oscillatory components decay. A coarse interpolation space should approximate those persistent smooth components. A coarse solve alone cannot remove arbitrary fine-grid error.

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

Test exact coarse correction

Use $A = \left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$, $P = (1, 1)^T$, $R = P^T/2$. Find the coarse operator and the coarse correction for true error $(3, 1)^T$. Verify the restricted residual afterward.

▼ **Worked solution**

$AP = (1, 1)^T$ and $RAP = 1$. The initial residual is

$$A(3, 1)^T = (5, -1)^T.$$

Its restriction is 2, giving coarse correction $P\,2 = (2, 2)^T$.

The remaining error is $(1, -1)^T$, with residual $(3, -3)^T$. Restricting gives zero even though the fine error remains. An exact coarse solve eliminated only the represented 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.

Use the transpose

Let $A = \left(\begin{smallmatrix} 2 & 1 \\ 0 & 3 \end{smallmatrix}\right)$, $c = (1, 0)^T$, and suppose a computed state's residual is

$$r = (0, 0.3)^T.$$

Compute the observation error with an adjoint, verify it by solving $Ae = r$, and explain the sign.

▼ **Worked solution**

Solving $A^T z = c$ gives $2z_1 = 1,\ z_1 + 3z_2 = 0$, so $z = (1/2, -1/6)^T$. Its dot product with r is –0.05.

The error equation gives $e_2 = 0.1,\ 2e_1 + 0.1 = 0$, hence

$$c^T e = -0.05.$$

A positive second-row defect forces a negative first-component change through the positive off-diagonal coupling. Using A instead of its transpose for the dual would miss that effect.

Explore the related calculation with adjustable inputs. Inspect the Taylor remainder with the correct and wrong-sign adjoint. The explorer uses a two-state equilibrium; apply the stated scalar or nonsymmetric operator to obtain this exercise's exact gradient.

Decide if another grid refinement answers the engineering question

A steady diffusion solver predicts a dimensionless mean sensor value of 0.42800, 0.42200 and 0.42050 on meshes with 16, 32 and 64 cells per side. The linear solver reports relative residuals below 10^{-10}. Estimate the converged value and the error on the finest grid. A measurement reads 0.4170 with uncertainty ±0.0010. Can the discrepancy be attributed to mesh error alone?

▼ **Worked solution**

Assume the same smooth solution, geometry, sensor integration and refinement ratio on all three grids. The ratio of successive differences is 0.0060/0.0015 = 4.0, giving observed order p = 2. Extrapolate the fine estimate by adding (0.42050–0.42200)/(2²–1) = –0.00050. The resulting estimate is 0.42000.

The fine-grid spatial error magnitude is therefore about 0.00050. A relative algebraic residual near 10^{-10} describes the discrete equations and is in different units from the absolute sensor error. It cannot replace this grid calculation. Repeating the study with a local peak instead of a mean can give a different observed order.

The measurement interval spans 0.4160 to 0.4180. Even the extrapolated prediction exceeds its upper end by 0.0020. Under the assumed error model mesh refinement alone appears too small to explain that gap. Test coefficient uncertainty, boundary data and sensor averaging separately. The numerical experiment in chapter 6 supplies a known exact field for checking the discretisation independently of the physical measurement.

Explore the related calculation with adjustable inputs. Refine the grid on a known solution and keep the small linear-solver residual separate from the spatial field error.

Chapter 13 · Pure flux boundaries, compatibility and the missing constant

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.

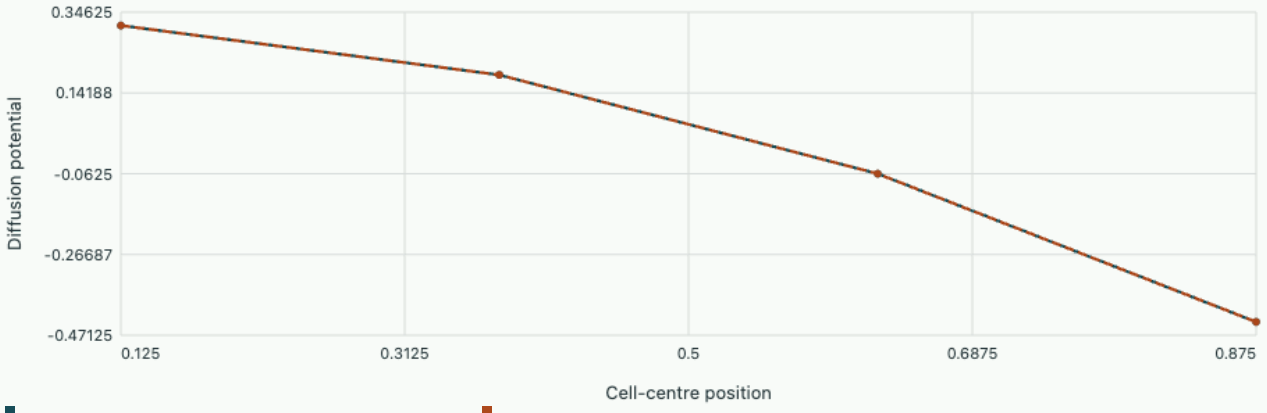
Open this experiment online

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

Study the full derivation

A Fourier Poisson solver and the trouble caused by unresolved products · Spatial discretisation and field equations

In a periodic medium the constant part of a source decides whether a Poisson problem has any solution. Derive the Fourier solver, separate spectral and finite-difference symbols, and see how a nonlinear product can inject a false low-frequency source.

Chapter 14 · Building a local residual estimator and refining where it matters

A finite-element solution can satisfy the assembled equations to machine precision while still missing curvature between nodes. Integrate the residual of the computed field by parts on each element and use the resulting cell and interface terms to decide where to spend the next degrees of freedom.

Ask what the discrete solution does to an unresolved test function

Consider $-u'' = f$ on a unit interval with zero Dirichlet data. Use continuous piecewise-linear finite elements with exact integration on a possibly nonuniform mesh. The weak solution satisfies $a(u, v) = \int u'v' = \int fv$ for every test function vanishing at the endpoints. For $e = u - u_h$, subtract the computed field to obtain

$$a(e, v) = R(v) = \int fv - \int u_h' v'.$$

The algebraic system makes $R(v_h) = 0$ for all functions in the chosen finite-element space. Tests with more local structure expose what is missing.

$$R(v) = \sum_K \int_K (f + u_h'')v \, dx + \sum_{i \text{ interior}} J_i v(x_i),$$

$$J_i = u_h'(x_i^+) - u_h'(x_i^-).$$

The sign follows by adding the right boundary term of the left cell and the left boundary term of the right cell. On a linear cell $u_h'' = 0$ classically. Its piecewise constant gradient jumps at interior nodes. In diffusion with varying coefficient k the corresponding jump is in ku_h', since continuity of normal flux is the conservation condition. A discontinuous material coefficient can cause an exact gradient jump without any flux imbalance.

Scale the cell and node terms before marking

Subtract the finite-element interpolant or a local quasi-interpolant from v inside the residual. Galerkin orthogonality permits this subtraction and bounds the remaining test magnitude by local length times its gradient. This leads to a computable residual indicator. One common one-dimensional choice with a symmetric edge length h_i equal here to the smaller neighbouring cell is

$$\eta_K^2 = h_K^2 \int_K f^2 \, dx + \frac{1}{2} \sum_{i \in \partial K \text{ interior}} h_i J_i^2,$$

$$\eta^2 = \sum_K \eta_K^2.$$

Sharing the node contribution prevents double counting. Under the usual mesh and coefficient assumptions such estimates bound the energy error up to constants; their absolute numbers need not

equal the measured error. Data oscillation can matter if the source is approximated or oscillates within a cell. To mark by bulk contribution, sort η_K^2 in descending order and take cells until their sum exceeds a chosen fraction of the global squared estimator. Bisect marked cells, transfer the coefficient data, assemble and solve again. Reusing the old gradient jumps after refinement would skip the effect of the new trial space.

Work through the first two-element decision

Let $f = 1$ on the left half and $f = 9$ on the right. With two equal cells and zero end values there is a single interior hat unknown. Diagonal stiffness is 4 and the integrated hat load is

$$1(0.25) + 9(0.25) = 2.5.$$

The middle value is therefore 0.625. The slopes are 1.25 and –1.25, giving $J = -2.5$. Since $h = 0.5$, the two cell terms are 0.125 and 10.125. Each side receives $(1/2)(0.5)(2.5)^2 = 1.5625$ from the node. The right cell contributes $11.6875/(1.6875 + 11.6875) = 87.38\%$ of the squared indicator. A 50% bulk mark selects only that cell.

For this particular problem the finite-element solution equals the piecewise-linear interpolant of the true nodal values. This follows by testing the derivative of the interpolation error against a constant basis gradient on each element. For constant f on each element the derivative error is therefore a line of slope –f with zero mean. Its exact squared energy norm is $\sum f_K^2 h_K^3/12$. Splitting the right half into two equal cells reduces that half's contribution by a factor of four. Total energy error falls from 0.924 to 0.470 with one extra degree of freedom. Splitting only the left cell leaves about 0.920. Run the loop and compare the indicator's ranking with this independent verification quantity.

Run the mesh from zero to three refinement rounds with a 50% bulk mark. Record error against free values. Repeat with a 90% mark and compare equal computational costs where available. Which choice targets the source localization most strongly?

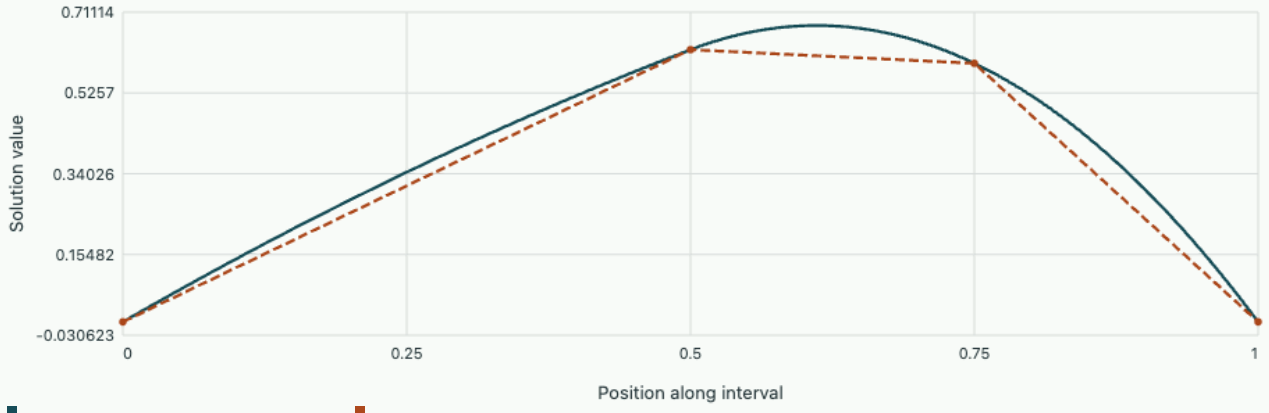
Open this experiment online

Try the calculation

Refinement rounds **1**

Right-hand source **9**

Bulk marking fraction **0.5**



After 1 refinement round there are 3 cells and 2 free values. Exact energy error is 0.4705. The global residual indicator is 2.0963.

Tutorial · predict the benefit of two candidate splits

Consider a left cell of length 0.50 with $f = 1$ and a right cell of length 0.125 with $f = 9$. Which single split would remove more exact squared energy error? Show how stiffness residual and source localization answer distinct questions.

▼ **Show the local comparison**

Bisection removes three quarters of that cell's $f^2h^3/12$. Before division by 12 the left weight is $1(0.5)^3 = 0.125$; the right weight is

$$81(0.125)^3 = 0.1582.$$

Refining the small right cell still yields about 27% more improvement. Conversely the assembled nodal residual could already be zero on either mesh after a direct solve. Reducing the algebraic tolerance further would improve neither piecewise-linear profile. This calculation is a verification benchmark made possible by the manufactured piecewise-constant data.

Further reading · Joachim Schöberl, Interactive Finite Elements · a posteriori error estimation by residuals

Study the full derivation

Bound diffusion error by balancing a reconstructed flux · Spatial discretisation and field equations

A temperature field from a finite-element calculation looks smooth. Reconstruct a heat flux which exactly balances the source, then derive an upper bound for the true energy error without knowing the exact temperature.

Chapter 15 · What a periodic grid does to a decaying Fourier mode

A periodic diffusion solver predicts the shape of a sampled wave while its amplitude decays too slowly. Fourier analysis separates spatial differentiation error, an unresolved physical mode and errors caused by the time integrator.

Diagonalise the differential operator on the periodic interval

Use a dimensionless coordinate $0 \le x < 2\pi$ and solve $u_t = \kappa u_{xx}$ with periodic values and periodic derivatives. The complex exponentials e^{imx}, with integer m, satisfy both conditions and are eigenfunctions of the second derivative. Put $u(x,t) = A_m(t)e^{imx}$ and cancel the common exponential. The scalar amplitude equation is

$$A_m' = -\kappa m^2 A_m.$$

Thus every cosine or sine mode decays by $\exp(-\kappa m^2 t)$, with no phase shift, while the m=0 mean is conserved.

On a uniform grid $x_j = 2\pi j/N$, the samples of m and $m + \ell N$ agree for every integer ℓ. A discrete Fourier transform returns coefficients in one representative band. For the real cosine test with even N we can fold a positive physical mode into $k = \min(r, N-r)$, $r = m \bmod N$. A Fourier spectral Laplacian multiplies that coefficient by $-k^2$. For a resolved mode it differentiates the trigonometric interpolant exactly, up to arithmetic. For unresolved data it differentiates the aliased representative; no operation using only those samples can identify the original m.

Apply the centred difference stencil to a single wave

The standard second difference is $(u_{j+1} - 2u_j + u_{j-1})/h^2$, $h = 2\pi/N$. Put the grid wave $u_j = e^{ikjh}$ into the stencil. Factor out the common phase and simplify with $e^{i\xi} + e^{-i\xi} = 2\cos\xi$. This gives the symbol and its small-h expansion:

$$\frac{e^{ikh} - 2 + e^{-ikh}}{h^2} = -\frac{4}{h^2}\sin^2(kh/2) = -k^2 + \frac{k^4 h^2}{12} + O(k^6 h^4).$$

The effective squared wavenumber is smaller than k^2 for every nonconstant mode in the resolved band, so this spatial approximation decays too slowly if the time equation is integrated exactly. The error is modest when $kh \ll 1$, but at the Nyquist mode $kh = \pi$ the ratio is

$$4/\pi^2 = 0.4053.$$

A stable time integrator cannot repair this spatial attenuation error without additional modeling. The experiment exponentiates both discretizations' symbols exactly, isolating space and sampling here.

Measure separate errors with the same sampled initial data

Choose $\kappa = 0.5,\ m = 3,\ N = 16$. The true squared wavenumber is 9 and the difference stencil gives 8.006. Starting at unit amplitude, at $t = 0.20$ exact and spectral amplitudes are 0.40657; finite differences give 0.44906. Doubling N to 32 changes the difference result to 0.41716. The absolute amplitude error falls from 0.04249 to 0.01059, a ratio of 4.01 consistent with the second-order expansion. This is a spatial-convergence study because time evolution introduces no truncation error.

Now take $m = 12,\ N = 16$. Every sampled cosine equals the low mode $k = 4$ at $t = 0$. At $t = 0.02$, the continuum amplitude for the true physical mode is $e^{-1.44} = 0.237$, whereas the spectral grid predicts

$$e^{-0.16} = 0.852.$$

A spectral computation can therefore agree exactly with its trigonometric interpolant and disagree strongly with the evolution of the intended signal. Sampling on a finer grid or obtaining prior bandwidth information resolves that ambiguity. The mean mode is particularly revealing when a nonconstant physical cosine aliases to a constant.

Start at mode 3 and successively double N from 8 to 32. Compute two amplitude-error ratios at the same time. Then use mode 12 on 16 points to observe an alias, and increase N until the Fourier grid represents the intended mode.

Open this experiment online

Try the calculation

Periodic grid points	16
Physical cosine mode	3
Diffusivity	0.5



The grid represents mode 3 as 3. At t = 0.20, amplitudes are 0.40657 for the physical mode, 0.40657 for the Fourier grid and 0.44906 for the difference stencil.

Tutorial · resolve a damping rate to a target tolerance

For a mode well inside the grid bandwidth, estimate the number of samples per wavelength needed for a relative error below 1% in the decay rate with centred differences. Test the estimate for $m = 2$. How could the time integrator contaminate a measurement of this spatial error?

▼ **Show the resolution estimate**

Writing p for samples per wavelength gives

$$kh = 2\pi/p.$$

The relative rate error is approximately $(kh)^2/12$. Requiring at most 0.01 gives $p \gtrsim 18.14$; try 19 samples per wavelength, or $N = 38$ for $m = 2$. The exact ratio $[\sin(\pi/19)/(\pi/19)]^2$ is about 0.991. A forward-Euler step would replace exponential propagation by $1 - \kappa k_{\mathrm{eff}}^2 \Delta t$. Its repeated-step error could partly mask the underdamping of the spatial stencil. Compare with the exponential of the discrete symbol before attributing a measured rate to space alone.

Further reading · Lloyd N. Trefethen, Spectral Methods in MATLAB · periodic grids, Fourier differentiation and downloadable programs

Check the discrete eigenvalue with OneSimulation

A spatial sine also diagonalises the five-point Dirichlet Laplacian on a square. Use the same modified wave-number argument in both coordinate directions. Run `python poisson-mode.py --mode 2 --cells 8 16 32` and compare the observed nodal error with the discrete eigenvalue prediction in the CSV. This is a steady forced problem whose spatial operator can be checked independently of any time integrator. Download the OneSimulation verification model.

Study the full derivation

A Fourier Poisson solver and the trouble caused by unresolved products · Spatial discretisation and field equations

In a periodic medium the constant part of a source decides whether a Poisson problem has any solution. Derive the Fourier solver, separate spectral and finite-difference symbols, and see how a nonlinear product can inject a false low-frequency source.

Chapter 16 · Why two finite-element wave solvers propagate the same mode at different speeds

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

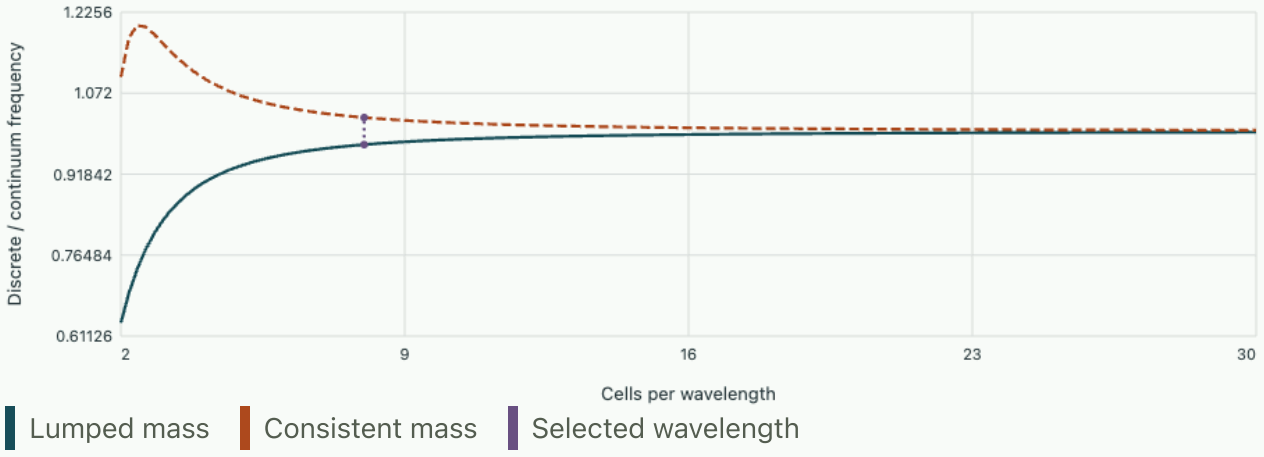
At four cells per wavelength $\theta = \pi/2$. The lumped ratio is $2\sqrt{2}/\pi \approx 0.9003$; the consistent ratio is about 1.1027. One wave solver lags by roughly ten per cent and the other leads by roughly ten per cent. At eight cells per wavelength the ratios move much closer to one, but repeated 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 $h_t\omega \leq 2$ 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.

Chapter 17 · Moving a cell average without inventing or losing material

Derive a conservative transport step from cell overlap and measure how a monotone scheme spreads a pulse despite conserving its total mass exactly.

Transport is a translation

For constant $a > 0$, the solution to $u_t + a u_x = 0$ is $u(x,t) = u_0(x - at)$. Consider periodic cells of width h. Over one step Δt, a cell translated back through distance $a\Delta t$ overlaps part of its original cell and part of its left neighbour. If the initial reconstruction is constant in each cell and $C = a\Delta t/h$ lies between zero and one, the new average is

$$u_j^{n+1} = (1 - C)u_j^n + C u_{j-1}^n.$$

Summing all cells telescopes on a periodic grid, preserving total material $h\sum_j u_j$. Each new value is a convex combination, so the maximum cannot increase and the minimum cannot fall. For $C > 1$ this particular two-cell overlap ceases to describe the translated interval.

Derive the spreading instead of reading it off a plot

Start from one unit of material in one cell and follow its distribution on an effectively unbounded grid. After one step its index increment behaves like a Bernoulli choice: zero with weight $1 - C$, one with weight C. Repeated convolution gives a binomial profile. Its mean index after n steps is nC and variance $nC(1 - C)$. Multiplying by h^2 and using $n = t/\Delta t$ gives a physical variance $ah(1 - C)t$. Equating this with $2D_{\mathrm{num}}t$ gives

$$D_{\mathrm{num}} = ah(1 - C)/2.$$

This growth describes numerical spreading of a translated shape even though the scheme conserves mass.

Two ways to reach the same physical time

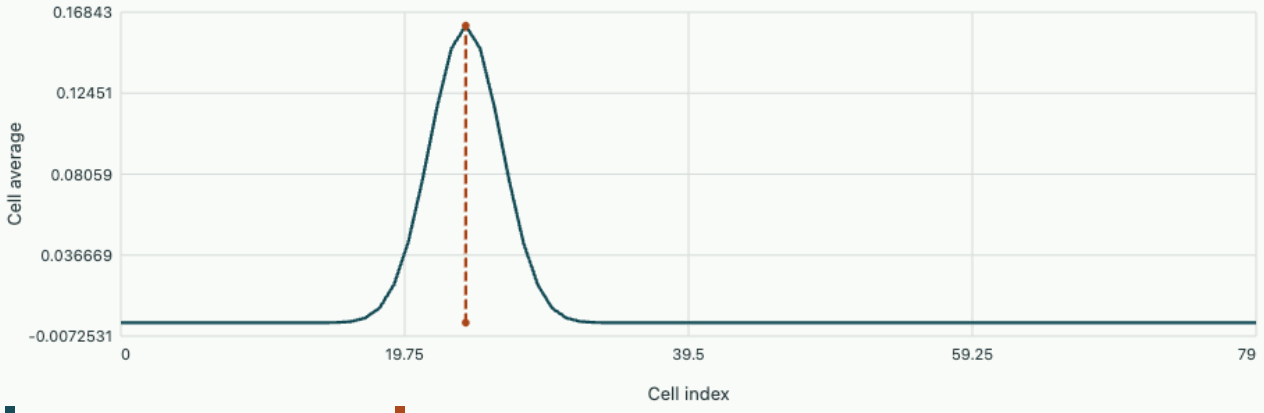
A pulse should move 12 cells. Taking 12 steps with $C = 1$ shifts data by exactly 12 indices. Taking 24 steps with $C = 0.5$ has the same mean displacement but an added variance of six cell widths squared. In a source-free transport test, peak height and width reveal an error hidden from the total-mass budget.

Open this experiment online

Try the calculation

Courant number C **0.5**

Number of steps **24**



Total cell sum 1, centroid 24, added variance 6 cells².

The experiment starts with one full cell on a ring of 80 cells. Up to 40 steps avoid wraparound across the displayed pulse for the chosen starting index. The blackboard binomial calculation predicts centroid and spread independently of the plotted recurrence.

Tutorial on interpreting artificial diffusion

If $C = 0.25$ and the pulse has moved ten cells on average, what is its added variance?

▼ Worked solution

Mean travel requires $n = 40$ steps. Variance in cell units is $40(0.25)(0.75) = 7.5$. Multiplying by h^2 gives the physical variance.

Is a larger time step always the cure for pulse spreading?

▼ Worked solution

For constant transport on this uniform grid $C = 1$ happens to align the translation with cell boundaries. Multiple speeds, nonuniform meshes and source physics generally preclude such alignment everywhere. A higher-order bounded reconstruction can reduce spreading while preserving conservation.

Read the university treatment

Use the advection examples to compare convergence in a norm with conservation of an integral and propagation of an extremum. Randall LeVeque · Finite Difference Methods for Ordinary and Partial Differential Equations

Chapter 18 · When diffusion destabilises a uniform state

Two scalar signals on a ring respond locally to each other and exchange signal with their neighbours. A single well-mixed cell returns to its reference state after a small disturbance. Can coupled cells amplify a spatial pattern? Derive a condition for this surprising outcome and test how faithfully a finite grid represents its wavelength.

Begin with a stable, spatially uniform experiment

Let u and v be small dimensionless departures from a reference operating point. They represent signed deviations, so their components can be negative without implying a negative absolute material quantity. Choose local time evolution $u_t = u + v,\ v_t = -3u - 2v$. A positive u tends to increase itself immediately; the second channel subsequently feeds back on the first. To determine whether a perturbation persists one must examine both channels together. Looking only at the positive diagonal coefficient in the first equation would give a wrong stability conclusion.

Write the state as a column and call the coefficient matrix J. The characteristic polynomial of a two-by-two matrix is $\mu^2 - \operatorname{tr}(J)\mu + \det J$. Here the trace is minus one and the determinant is one. The two roots are $(-1 \pm i\sqrt{3})/2$. An isolated cell therefore has damped oscillatory motion for every initial perturbation; no normal mode has a positive growth rate. Some observables can still grow transiently in mixtures of modes. The present experiment asks about exponential growth of a mode over time.

Allow spatial exchange by adding $D_u u_{xx}$ and $D_v v_{xx}$ on a periodic ring of dimensionless circumference L. Both diffusion coefficients are positive. Without coupling each field's Fourier component would decay at a rate proportional to the square of its wavenumber. The system changes qualitatively when the two diffusion rates differ because the path that restores local stability involves both fields. A spatially varying signal in one channel can be smoothed before it delivers the same feedback as in a spatially uniform experiment.

Reduce the field equation to one matrix per wavelength

Periodicity imposes $k_m = 2\pi m / L$ for integer m. Write a normal mode as $w \exp(i k_m x + \mu t)$. Differentiating the spatial factor twice gives $-k_m^2$. Define $q = k_m^2$. Substitution shows that μ must be an eigenvalue of $J - qD$, where D is diagonal with entries D_u, D_v. Distinct modes do not interact in this spatially homogeneous linear model. One may analyse each mode separately and superpose the results.

$$\tau(q) = -1 - (D_u + D_v)q,$$

$$\delta(q) = 1 - (D_v - 2D_u)q + D_u D_v q^2.$$

The trace τ is negative for every $q \geq 0$. Complex conjugate roots therefore have negative real parts. A positive eigenvalue must be real and accompanied by a negative one, so the determinant must be negative. Conversely, negative determinant ensures real eigenvalues of opposite signs. The instability question has become a test of the sign of a quadratic function on nonnegative arguments. Both at zero and at arbitrarily large q the determinant is positive. If a growing mode exists, it must occupy an intermediate range.

Put $b = D_v - 2D_u$ and $c = D_u D_v$. The minimum occurs at $q = b/(2c)$, which is positive only if $b > 0$. Its value is $1 - b^2/(4c)$. Thus the continuous dispersion curve contains an open interval of growing wavenumbers if $b > 0$ and $b^2 > 4c$. At equality one neutral wavenumber appears; a neutral mode neither grows nor establishes instability. The physically allowed mode numbers of the finite ring must still place some squared wavenumber inside the interval. A condition derived for a continuum of wavelengths alone can overpredict what fits in a short apparatus.

$$q_{\pm} = \frac{b \pm \sqrt{b^2 - 4c}}{2c}, \qquad q_- < k_m^2 < q_+.$$

Calculate the unstable band before running a simulation

Choose $D_u = 0.02, D_v = 1$. Then

$$b = 0.96,\ c = 0.02.$$

The two roots of the determinant are about 1.065 and 46.93. On a ring of circumference 2π, $k_m = m$, so the continuum equation amplifies modes two through six. The uniform mode still decays. Mode one has $q = 1$, just below the lower edge, while mode seven has $q = 49$, just above the upper edge. Both are valuable tests because a moderate numerical error in the dispersion relation could reverse their classification.

The largest real part of the two eigenvalues is $\mathrm{Re}[\tau + \sqrt{\tau^2 - 4\delta}]/2$. Choose the square root with nonnegative real part. Where the discriminant is negative the real part of the square root is zero; treating a negative discriminant as an error and dropping those wavenumbers would omit damped oscillatory physics. At $q = 4$, the determinant is -2.52 and the trace is -5.08. The positive rate is approximately 0.4553. During one dimensionless time unit an initially pure growing eigenmode at that wavenumber is multiplied by about 1.58.

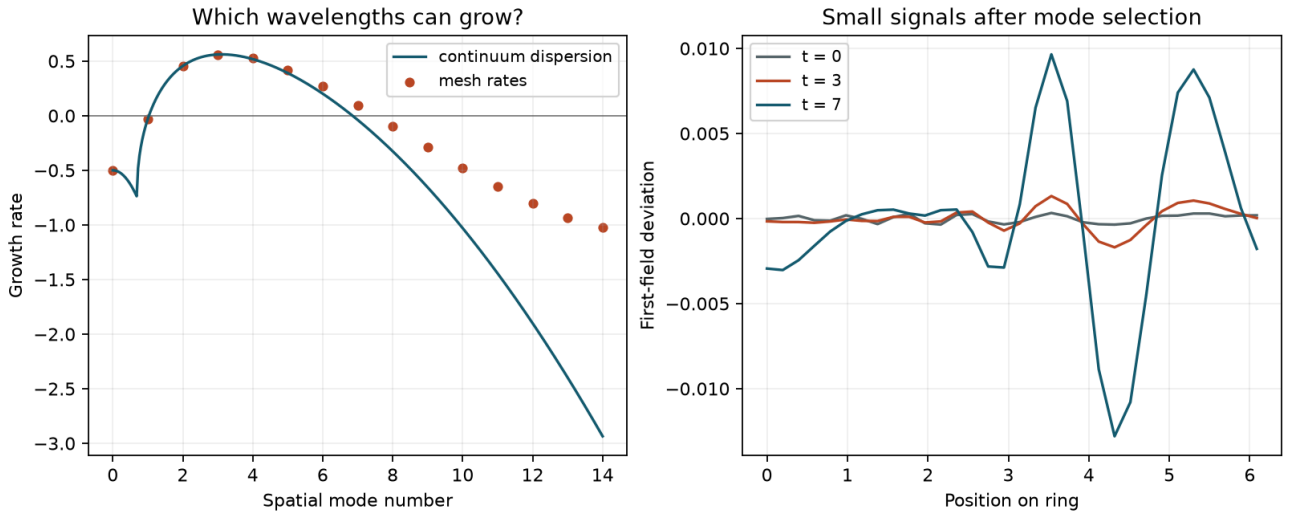
Derive what a particular spatial difference sees

Now discretise the ring with N equally spaced points of spacing $h = L/N$. The centred second difference of a discrete Fourier mode multiplies it by $(e^{ikh} - 2 + e^{-ikh})/h^2$. Using the cosine half-angle identity gives a real, nonpositive factor. Compare it with the exact factor $-k^2$. They agree to leading order at small kh but separate substantially near the grid scale.

$$q_h(k) = \frac{4}{h^2}\sin^2(kh/2) = k^2 - \frac{h^2 k^4}{12} + O(h^4 k^6).$$

For each sampled mode the finite grid uses $J - q_h D$. The same algebraic determinant test applies, with a different argument. On the 32-point default grid, mode seven has q_h about 41.8 rather than 49. Its grid growth rate is positive even though its continuum rate is negative. This discrepancy survives any

refinement of the time integrator on that fixed grid. The early growth has a reproducible numerical explanation and can be diagnosed without waiting for a long transient simulation.



The points sample the discrete operator. A point above zero whose continuum counterpart is below zero is a spatial-discretisation error even when the time evolution of the grid is calculated exactly.

A meaningful refinement experiment must follow a fixed physical mode as N increases. Following the highest mode of each grid instead changes its physical wavelength. Similarly, changing circumference while keeping the mode number fixed changes the wavelength and therefore the physical growth rate. The notebook exposes both controls so the two experiments can be distinguished. Large-scale modes can appear converged while the classification at the upper band edge remains sensitive to cell count.

Relate an eigenvalue to a picture of a pattern

Construct a numerical experiment in which the expected time dependence is actually known. For each of several low spatial modes, use the leading eigenvector of the two-channel finite-grid matrix and multiply by a small, controlled initial amplitude. Evolve its coefficient by its complex exponential, then take real parts to obtain real sensor signals. The resulting trace has no time-stepping error. An independently assembled full block matrix, including the periodic connection between the first and last cells, supplies another way to recover the largest growth rate. The optional OneSimulation eigensolver applies to this complete matrix directly.

A visible crest in a total field is not a single Fourier coefficient. It can move or change height through interference even when each included mode is decaying. To measure a rate compare the amplitude of one spatial Fourier component at two times. If the initial field contains both eigenvectors at that same wavenumber, the coefficient may not yet show one pure exponential. Preparing a known eigenvector or observing later, while the state remains within the linear range, makes the inference interpretable.

There is also an upper limit to what this calculation predicts. A linear mode with positive rate grows exponentially without saturation. Once its amplitude is comparable to the scale used to linearise the original response, nonlinear terms could shift the wavelength selection, mix modes or produce a

bounded state. The notebook chooses small initial amplitudes and finite observation times precisely to isolate onset. It cannot infer the final pattern simply by extending the linear growth curve indefinitely.

Set a resolution requirement from the measured quantity

Near the upper band edge the question can be posed in terms of a tolerance on q. A finite-grid prediction of a positive rate changes sign precisely when its effective q_h crosses the edge. If the target mode has q barely above that edge, estimate whether the leading truncation bias $h^2 q^2/12$ exceeds the distance to it. This supplies an inexpensive grid-design calculation before allocating a time integrator. Because the expansion assumes kh is small, verify the final decision with the exact sine symbol.

Away from an edge the sign can already be reliable while the predicted doubling time is inaccurate. A positive growth rate corresponds to a doubling time of $\ln 2/\mu$; relative errors are strongly amplified near neutral stability if measured against a very small rate. Record both the sign and an absolute uncertainty on the rate when comparing computations near onset. A single label such as stable or unstable does not communicate that sensitivity.

Open editable notebook Download .ipynb

Tutorial on modes and finite domains

1. Give a short proof that equal positive diffusion coefficients cannot create an instability in this model.

▼ **Work through the solution**

If $D_u = D_v = d$, the spatial block is $J - dqI$. Its eigenvectors remain those of the stable local system, and every eigenvalue shifts left by dq. All real parts are at most minus one half. This argument works for any locally stable constant matrix with identical scalar diffusion in every channel.

---

2. Keep the two default diffusion coefficients but change the ring circumference to π. Which positive continuum mode numbers grow?

▼ **Work through the solution**

The allowed wavenumbers are $k_m = 2m$. The inequality becomes $1.065 < 4m^2 < 46.93$. It admits $m = 1, 2, 3$, corresponding to wavenumbers 2, 4 and 6. Comparing mode numbers alone between the two rings would hide this wavelength correspondence.

---

3. A Fourier amplitude rises from $2 \cdot 10^{-4}$ to $8 \cdot 10^{-4}$ in four time units. Estimate the growth rate and describe an additional measurement needed before identifying a physical instability.

▼ **Work through the solution**

A pure mode gives $\mu = \log(4)/4 \simeq 0.3466$ per time unit. Confirm that the same physical wavenumber grows on refined grids and that its rate approaches the continuum dispersion calculation. A positive finite-grid rate by itself cannot distinguish a real band from the extra mode admitted by a coarse Laplacian.

Further reading · Duke University, linear stability analysis of two coupled reaction–diffusion equations

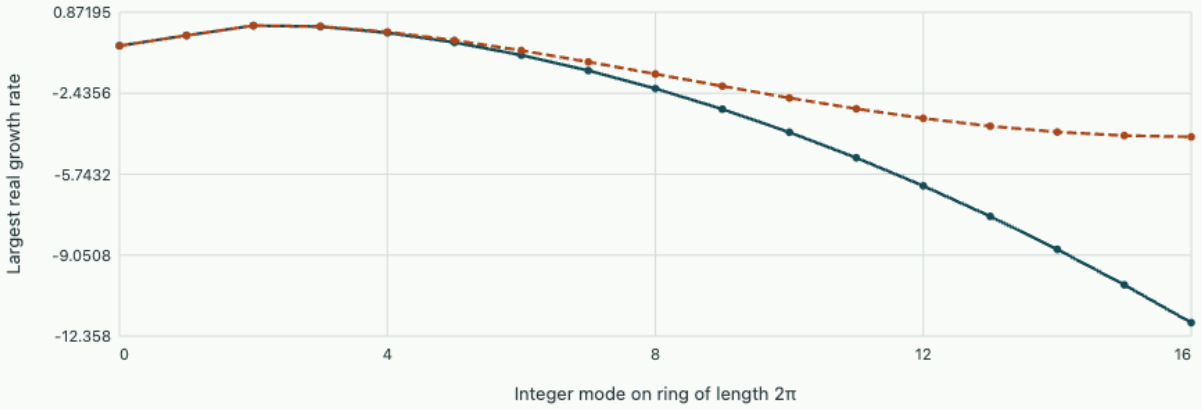
Change parameters · When diffusion destabilises a uniform state

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

Diffusion Du	0.05
Diffusion Dv	1
Periodic grid cells	32



The spatially uniform mode has real growth −0.5. A positive plotted value identifies a diffusion-driven instability; refining the grid changes which wavelength the numerical model selects.

Problem sheet 7 · Project · Verify a heat solver before interpreting its temperature

A laboratory plate can be heated internally and cooled at its boundary. Check global feasibility, the spatial error between nodes and the wavelength-dependent decay of periodic modes.

Can the thermal problem have a steady solution?

A one-dimensional bar of unit length and cross section obeys $-u'' = 2$ in scaled units. Prescribe outward conductive fluxes $q_L = 0.5$, $q_R = 1.2$. Integrate the equation before solving. What physical quantity changes with time under these fluxes? Correct the right flux while keeping the left.

▼ **Work through the solution**

Outward flux is $-u'n$. Total outward flux must equal the integrated source, 2. The prescribed sum is 1.7, leaving a positive heating imbalance of 0.3. With finite heat capacity the mean temperature rises. A compatible right flux is 1.5. A steady calculation with purely Neumann data then needs a reference temperature. Use the Neumann compatibility experiment. Its left sensor reports zero rightward flux. For this comparison set $v = u - 0.5x$, which shifts the rightward flux by +0.5 at both ends. The original measurement then corresponds to the slider setting “Right flux minus 2” of −0.30, and the corrected data correspond to zero.

Fix the reference and check signs

For the compatible data, find the quadratic solution with zero spatial mean. Verify both fluxes and show which part a gradient measurement cannot recover.

▼ **Work through the solution**

Integrating twice gives $u = -x^2 + 0.5x + C$ because at the left $q_L = u'(0) = 0.5$. At the right $-u'(1) = 1.5$. The mean is $-1/3 + 1/4 + C$, so $C = 1/12$. Adding any constant leaves every flux and gradient unchanged. A measured mean supplies information absent from the differential equation.

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.

Explain the missing inter-element curvature

Interpolate that exact quadratic linearly on uniform elements of width h. Derive the peak pointwise interpolation error on one element and the jump in its piecewise-constant derivative between adjacent elements. Predict what happens when h halves. Compare the two errors with the residual and refinement example, which uses unequal sources on the two halves and reports a global energy norm.

▼ **Work through the solution**

On $[a, b]$, $u - I_h u = -(x - a)(x - b)$, which peaks at $h^2/4$. The slope of each secant equals the exact derivative at its midpoint. Moving to the next element changes this slope by $-2h$. Halving h quarters the peak value error and halves the jump magnitude. The adaptive experiment measures the gradient error in an energy norm. Uniformly halving all element widths would halve that norm for this quadratic. Selective refinement can give a different global ratio. Machine-small residual in the assembled linear system alone measures neither the field error nor the derivative jump.

Test an independent transient boundary condition

Now join the ends periodically. For $u_t = u_{xx}$ on a unit ring, a sine of mode m decays at rate $(2\pi m)^2$. Derive the semi-discrete finite-difference rate and evaluate its ratio to the true rate at eight points per wavelength.

▼ **Work through the solution**

A Fourier mode on the grid acquires factor $-4\sin^2(\pi m h)/h^2$ under the centred second difference. At eight points per wavelength, the ratio is $[\sin(\pi/8)/(\pi/8)]^2 \approx 0.94964$. Numerical decay is about 5% too slow even with exact time integration. Compare with the modal diffusion experiment with 16 nodes and mode 2. The frequency and domain scaling differ, but the decay-rate ratio depends only on points per wavelength. Periodic and flux tests isolate distinct boundary and discretisation errors.

Problem sheet 8 · Resolve an outflow layer by fitting the numerical face flux

A dilute solute moves down a channel and diffuses against its mean flow. Fixed concentrations at the ends generate a thin downstream layer. Construct the cell flux directly from the differential equation, then ask how a solver can return nearly the right total throughput while still predicting an impossible local concentration.

Solve the physical flux problem before discretising

Let $x \in [0,1]$, positive constant speed a, and constant diffusivity D. Steady conservation gives $\partial_x J = 0$ with $J = au - Du_x$. Impose $u(0) = 1$, $u(1) = 0$. Find $u(x)$ and the flux. Where must the sharp gradient occur when a/D is large?

▼ **Work through the solution**

Constant flux turns the equation into $u' - (a/D)u = -J/D$. Multiplication by $e^{-ax/D}$ and integration gives a constant plus an exponential. The boundary values yield

$$u(x) = [1 - e^{-(a/D)(1-x)}]/[1 - e^{-a/D}],$$

$$J = a/[1 - e^{-a/D}].$$

Far upstream for large a/D the concentration stays near one and most of the flux is advective. At the downstream wall the concentration is zero, but the diffusive term carries the same positive flux. The layer thickness scale is D/a, a length. Constant throughput requires convection and diffusion to trade roles across the layer.

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.

Derive the Scharfetter–Gummel face relation

Take a segment from x_i to x_{i+1} of length h with given endpoint concentrations u_i, u_{i+1}. Solve the same constant-coefficient equation across that segment and express the face flux in a form with a well-defined zero-velocity limit.

▼ **Work through the solution**

Define $P = ah/D$ and $B(P) = P/(e^P - 1)$ with $B(0) = 1$. Integrating between the endpoints gives

$$u_{i+1} = e^P u_i - (J/a)(e^P - 1).$$

Solving for the common flux and using $B(-P) = e^P B(P)$ yields

$$J_{i+1/2} = (D/h)[B(-P)u_i - B(P)u_{i+1}].$$

At $P = 0$ it reduces to the centred diffusive flux. For large positive P, $B(P)$ becomes small while $B(-P)$ approaches P; the advective part then uses the upstream value. Both coefficients stay positive for real P, supporting monotone solutions of this one-dimensional problem.

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.

Numerical evaluation requires more than inserting the formula literally. Near zero use $P/\operatorname{expm1}(P)$ or the series $1 - P/2 + P^2/12 + \cdots$. At very large positive P, use $B(P) = Pe^{-P}/(1 - e^{-P})$ to avoid exponential overflow. The identity $B(-P) = B(P) + P$ avoids a second evaluation. Equal face fluxes across adjacent cells impose local conservation: whatever leaves one control volume enters its neighbour. For constant properties and no internal source, the exact local solution on each segment matches the analytic global nodal values, even when few grid points lie within the layer.

Diagnose a nearly correct throughput with impossible concentrations

Set $a = 1$, $D = 1/24$ and use eight equal segments. Compare the penultimate nodal concentration from exponential fitting, centred advection and first-order upwinding. Why can global throughput hide the error?

▼ **Work through the solution**

The cell Péclet number is three. For any two-point constant flux relation with constant coefficients the homogeneous nodal variation is r^i. Exponential fitting gives $r = e^P$; centred advection gives

$$r = (1 + P/2)/(1 - P/2) = -5;$$

upwinding gives

$$r = 1 + P = 4.$$

With these Dirichlet values the solution is $(r^N - r^i)/(r^N - 1)$. At node seven the three values are approximately $0.9502,\ 1.2000,\ 0.7500$. The centred approximation exceeds the inflow concentration and alternates around it because r is negative. Upwinding preserves the admissible interval but makes the layer broader. Exponential fitting gives the true nodal value for this local model.

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.

For each scheme, a constant concentration transports advective flux au, while its r^i homogeneous part has zero net flux. The numerical throughput consequently equals $ar^N/(r^N - 1)$. With eight cells the centred scheme predicts about 1.00000256, remarkably close to the true 1.00000000004, despite its local overshoot of twenty percent. A scale measuring total output could validate the wrong spatial

solver. Report a pointwise quantity, a monotonicity check and flux conservation in addition to throughput if the concentration near the outlet matters.

Test fitted nodes and resolve the inside of a cell

Vary the number of segments and the global Péclet number in the live graph. Under which interpretation is exponential fitting exact, and how could a plotted straight line through its last two values mislead an experimenter?

▼ Work through the solution

Under the specified homogeneous coefficients, zero sources and exact boundary data, the fitted scheme reproduces the analytic solution at every node and returns its correct constant flux. It does not make a piecewise-linear drawing between those nodes the exact local profile. With eight cells at global Péclet number 24, the midpoint of the last cell is $1/16$ from the outlet. The exact concentration is about $1 - e^{-1.5} = 0.777$; linear interpolation between 0.9502 and zero gives only 0.4751. Reconstruction inside the cell requires the exponential solution with the two endpoint values, or more resolution. Distinguish the face flux actually used in the method from a derivative estimated by visually differentiating the drawn polyline.

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.

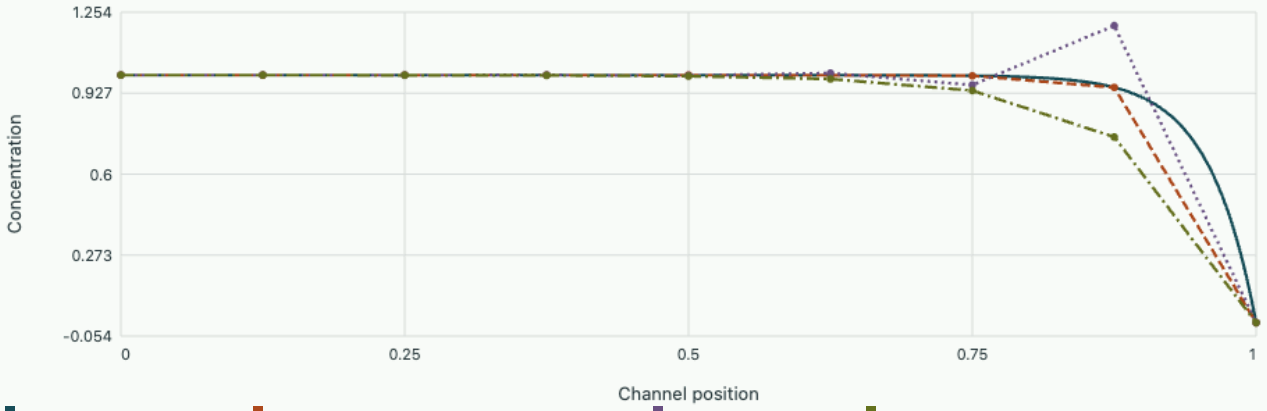
Try cell Péclet numbers just below and above two. The centred solution changes from a very sharp monotone profile to a sign-alternating boundary correction. The fitted flux changes continuously through the same transition. If coefficients vary significantly within a cell or a source deposits mass inside it, one needs a corresponding local model, source integration or refinement. An exact result for one constant-coefficient test helps isolate the subsequent approximation. It cannot establish accuracy for every heterogeneous transport problem.

Open this experiment online

Try the calculation

Segments 8

Global Péclet number 24



Analytic profile · Fitted nodes joined linearly · Centred flux · Upwind flux

Cell Péclet number 3.00. At the penultimate node, fitted 0.95021, centred 1.20000, upwind 0.75001.

The Sesame semiconductor solver documents exponential fitting of its carrier fluxes.

Notebook practice · When diffusion destabilises a uniform state

Spatial discretisation and field equations

Derive a discrete field problem from a continuous equation, choose appropriate boundary data and test the convergence of quantities a measurement could resolve.

1. ## Choose the physically admissible wave in a conservative update

   A discontinuous scalar flow can satisfy integral conservation in more than one way. Derive the shock speed, use a convex entropy to distinguish a compression from an expansion and build a finite-volume flux from the resulting local solution.

2. ## Bound diffusion error by balancing a reconstructed flux

   A temperature field from a finite-element calculation looks smooth. Reconstruct a heat flux which exactly balances the source, then derive an upper bound for the true energy error without knowing the exact temperature.

3. ## A Fourier Poisson solver and the trouble caused by unresolved products

   In a periodic medium the constant part of a source decides whether a Poisson problem has any solution. Derive the Fourier solver, separate spectral and finite-difference symbols, and see how a nonlinear product can inject a false low-frequency source.

4. ## How much of an outgoing wave reflects from the numerical boundary?

   A finite grid truncates an effectively open medium. Derive an exact one-dimensional radiation condition, predict the angular reflection of its two-dimensional extension and design a pulse experiment that separates boundary reflection from interior dispersion.

5. ## Can a stable finite-element heat step turn a cold node negative?

   Start a conducting rod with one positive temperature pulse. Assemble the heat capacity from first principles and determine when an explicit step respects both decay and the sign of every temperature.

6. ## Solving two-dimensional diffusion one direction at a time

   Derive a second-order alternating-direction step, identify the boundary data required halfway through it and distinguish splitting error from mesh and time errors.

7. ## Following the error from one reference triangle to an entire mesh

   Derive a first-order energy error estimate and a second-order displacement estimate for a conforming linear triangular discretisation. Track the assumption on element shape at the point where it enters the proof.

8. Which waves are allowed to enter a numerical boundary?

Diagonalise linear acoustics, impose one physical condition per incoming wave and compare diffusive and dispersive transport schemes on the same spatial Fourier mode.

Spatial discretisation and field equations · 1. Choose the physically admissible wave in a conservative update

A discontinuous scalar flow can satisfy integral conservation in more than one way. Derive the shock speed, use a convex entropy to distinguish a compression from an expansion and build a finite-volume flux from the resulting local solution.

Integrate across a moving discontinuity

Consider inviscid Burgers' equation $u_t + (u^2/2)_x = 0$. Where u is smooth its characteristics move at speed u. For left state $u_L > u_R$, faster left characteristics catch slower right ones and the profile steepens. If discontinuity position is $x = st$, integrate the conservation law across a narrow moving control volume. The storage change from the moving boundary gives $s(u_R - u_L) = f(u_R) - f(u_L)$, so for unequal states $s = (u_L + u_R)/2$. The value is a conservation-law jump speed; applying the pointwise product $u\,u_x$ across a discontinuity without defining it could predict another speed.

The same jump calculation formally allows $u_L < u_R$, although a smooth monotone connection then spreads. To identify the physical weak solution, add a small positive viscosity νu_{xx} and consider its limiting dissipation. For convex entropy $\eta = u^2/2$ with entropy flux $q = u^3/3$, multiplication of the viscous equation by u gives

$$\eta_t + q_x = \nu \eta_{xx} - \nu u_x^2.$$

As viscosity vanishes the entropy production is nonpositive in distributions. Across a sharp jump the coefficient is

$$[q] - s[\eta] = (u_R - u_L)^3/12,$$

where brackets mean right minus left. It is negative for a compression shock and positive for an expansion shock. Thus the latter satisfies mass conservation but violates this entropy criterion.

Construct the rarefaction from its self-similar coordinate

For a left-to-right increase seek $u(x,t) = U(\xi)$ with $\xi = x/t$. Substituting into the smooth equation gives $(U - \xi)U' = 0$, so each region is constant or $U = \xi$. Matching both states yields u_L for $\xi < u_L$, ξ for $u_L \le \xi \le u_R$, and u_R beyond. This continuous fan is the physically selected expansion. With $u_L = -1, u_R = 1$, the origin lies inside the fan and $u(0,t) = 0$. A stationary expansion jump would misleadingly carry flux $1/2$ across that interface rather than the physical zero.

In a finite-volume scheme cell averages are updated by shared interface fluxes:

$$\bar{u}_i^{n+1} = \bar{u}_i^n - (\Delta t/\Delta x)(F_{i+1/2} - F_{i-1/2}).$$

Solve the local Riemann problem between the piecewise-constant left and right cell states and take $F = f(U(0))$. For a compression use the left state if $s > 0$ and the right if $s < 0$; either flux agrees for a stationary shock. For an expansion reaching across zero take zero; if the expansion moves wholly right

use its left state, and if wholly left use its right. These cases give the scalar Godunov flux. The update telescopes internally because both neighbouring cells use exactly the same physical interface flux.

Distinguish shock location from smoothness and numerical thickness

A first-order Godunov discretisation under its CFL restriction is monotone and avoids spurious oscillations, but a moving discontinuity generally spreads over several cells. A centred pointwise difference may look sharper for smooth data while producing oscillations or the wrong propagation under discontinuous data. Test shock speeds from integrated cell mass or a consistent location reconstruction, compare fan areas and measure L^1 error under refinement. The pointwise maximum error right at a jump can remain large even as the location and integral error converge. A visually thin front is not by itself an accuracy certificate.

For data $u_L = 1, u_R = -0.5$ the shock moves right at 0.25. At $t = 0.25$ it lies at $x = 0.0625$. In the interval $[-1, 1]$ with unchanged boundary states, the total mass begins at 0.5 and grows by $[f(1) - f(-0.5)]t = (0.5 - 0.125)0.25 = 0.09375$, so the exact integral is 0.59375. A conservative numerical calculation can achieve this integral essentially to roundoff while retaining a visibly smeared shock. Conservation, entropy selection and resolved spatial profile are three independent questions.

Switch between a shock and a transonic fan

The experiment solves the finite-volume update on $[-1, 1]$ with constant exterior states, a time step chosen for maximum CFL 0.8 and output time 0.25. Compare the conservative cell averages with the exact self-similar profile. Reverse both initial states to turn a compression into an expansion, then put the fan across zero. Increasing cell count changes numerical thickness, while the mass should follow the boundary flux difference at every resolution. Equal left and right states are a separate constant-preservation test.

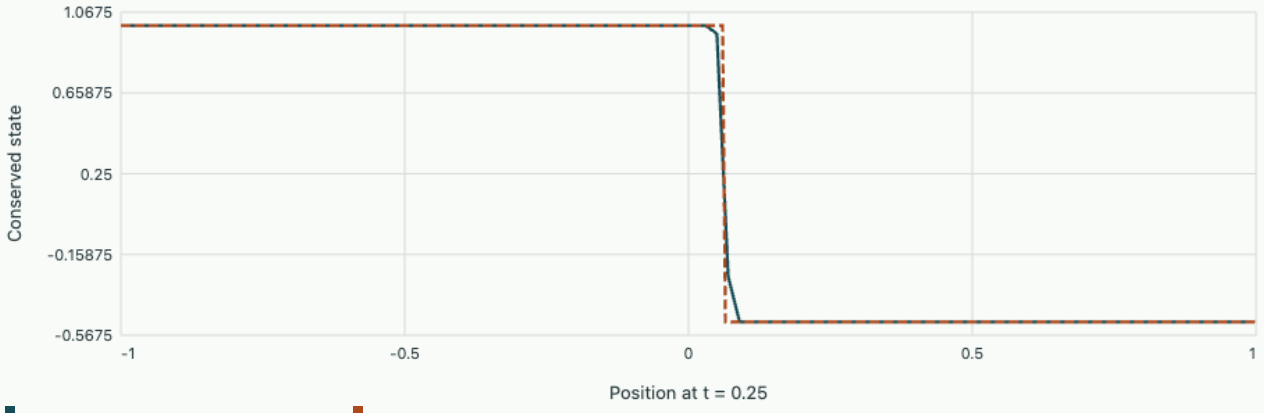
Open this experiment online

Try the calculation

Left state 1

Right state -0.5

Cells 100



Godunov cell averages Exact entropy solution

Shock speed 0.250. Time steps 16, maximum CFL 0.781, integral 0.59375.

Infer shock position from an integral measurement

A blurred profile does not require an arbitrary definition of its centre. Suppose the plateau values on $[-L, L]$ are measured reliably and a single discontinuity lies at X. Its integrated state is $M = (L + X)u_L + (L - X)u_R$. The equivalent sharp-interface position is therefore

$$X = [M - L(u_L + u_R)]/(u_L - u_R).$$

With $L = 1,\ u_L = 1,\ u_R = -1/2$, a cell-sum reading $M = 0.59375$ gives $X = 0.0625$. A mass uncertainty of 0.003 gives a position uncertainty of 0.002 in the same coordinates. The inference becomes ill-conditioned as the plateau contrast tends to zero.

This provides a mesh-independent diagnostic when numerical viscosity smears the jump. Track both an area-based position and a measure of thickness, for example the distance between the two intermediate levels one-quarter and three-quarters across the jump. A conservative solver may retain a highly accurate area-based position while thickness changes strongly under refinement. In a rarefaction the same scalar area exists, but two real moving edges delimit a resolved fan. Replacing the fan by one jump would throw away precisely the information the entropy condition supplied.

Tutorial · compare conservation and entropy

Calculate one update next to a moving shock

With states 1 and -0.5, take $\Delta t/\Delta x = 0.5$. Find the interfacial flux and the new values in the two cells adjacent to the discontinuity, assuming homogeneous neighbours farther out.

▼ **Work through the solution**

The shock speed is positive, so Godunov flux at the jump is $f(1) = 0.5$. The adjacent left cell sees equal incoming and outgoing fluxes and stays 1. The adjacent right cell sees incoming 0.5 and outgoing 0.125, so becomes

$$-0.5 + 0.5(0.375) = -0.3125.$$

This is exactly the average associated with movement of the shock through one eighth of that right cell: $0.125(1) + 0.875(-0.5) = -0.3125$.

Check a conservative expansion jump

For $-1 \to 1$, show that a stationary discontinuity satisfies mass conservation but give its entropy violation. Which profile should occupy $x = 0.10$ at $t = 0.25$?

▼ **Work through the solution**

Both end fluxes equal 0.5, so the formal discontinuity speed is zero. Its entropy coefficient is $(1 - (-1))^3/12 = 2/3 > 0$, contradicting dissipation. The rarefaction has $\xi = x/t = 0.4$ inside its fan, hence $u = 0.4$ there. A scheme retaining a stationary sharp expansion jump can pass a mass check and still fail this physical verification.

Set a stable step and predict a conserved integral

On 100 uniform cells in $[-1, 1]$ take $u_L = -0.25, u_R = 0.75$. What step gives maximum CFL 0.75? What mass is expected at time 0.20?

▼ **Work through the solution**

Here $\Delta x = 0.02$ and maximum speed is 0.75, giving $\Delta t = 0.020$. Initial mass is 0.5. Boundary flux difference is $0.03125 - 0.28125 = -0.25$; over time 0.20 the mass falls to 0.45. Both fan edges are still far from the domain ends. For longer runs boundary data and outgoing-wave treatment would alter this budget.

Further reading

- LeVeque · finite-volume methods for hyperbolic problems
- Clawpack Riemann book · Burgers shocks, rarefactions and entropy
- Course investigation · conservative flux updates
- Measure spatial convergence against an exact field

Spatial discretisation and field equations · 2. Bound diffusion error by balancing a reconstructed flux

A temperature field from a finite-element calculation looks smooth. Reconstruct a heat flux which exactly balances the source, then derive an upper bound for the true energy error without knowing the exact temperature.

Identify the natural error measure in a diffusion problem

Consider $-\nabla \cdot (k\nabla u) = f$ on a bounded domain, with homogeneous Dirichlet temperature for clarity and strictly positive scalar conductivity $k(\mathbf{x})$ bounded above and below. The weak form is $\int k\nabla u \cdot \nabla w = \int fw$ for every test w vanishing on the boundary. A conforming approximate temperature v satisfies the prescribed boundary trace but may not solve the equation. Its energy error $\|\nabla(u-v)\|_k^2 = \int k|\nabla(u-v)|^2$ weights local gradient errors according to their physical conductivity. This differs from pointwise temperature error, total dissipation error or the absolute error of one designated sensor.

Define true heat flux $q = -k\nabla u$. It obeys $\nabla \cdot q = f$. Suppose a reconstructed flux σ obeys the same divergence exactly and has the prescribed normal flux on any Neumann boundary, with normal component continuous across interior mesh faces. Such fluxes belong naturally to a space often written $H(\mathrm{div})$. Conforming temperature and equilibrated flux are separate requirements: a continuous piecewise-linear temperature gives generally discontinuous normal gradient flux across faces, even if all global nodal equations were solved accurately.

Prove an upper bound by cancelling a cross term

Take the weighted constitutive defect $\sigma + k\nabla v = (\sigma - q) + k\nabla(v-u)$. Expand its squared k^{-1} norm. The cross term is $2\int(\sigma-q)\cdot\nabla(v-u)$. Integrating by parts gives a volume term proportional to $\nabla\cdot(\sigma-q)=0$ and a boundary term zero where $v-u=0$ or the normal flux difference is prescribed zero. Interior normal continuity prevents an extra jump term. Hence

$$\|\sigma + k\nabla v\|_{k^{-1}}^2 = \|\sigma - q\|_{k^{-1}}^2 + \|\nabla(v-u)\|_k^2.$$

The left side can be calculated from the approximation and reconstruction. It bounds the unknown energy error from above, with equality precisely when the reconstructed flux equals the true flux. A divergence-compatible but inaccurate σ still certifies an upper bound, albeit possibly a loose one. This identity is an instance of the Prager–Synge hypercircle principle. With approximated data or inexact equilibration, extra data oscillation and residual terms must be bounded before calling an indicator a guaranteed bound.

Background · How does a nonzero flux imbalance enter the certificate?

Take heat flux $q = -k\nabla u$, so that $\nabla \cdot q = f$. Suppose a reconstructed flux $\hat{q}$ leaves a source residual $r = f - \nabla \cdot \hat{q}$, and the approximate and exact temperatures agree on the entire Dirichlet boundary. Testing the residual equation with their difference $e = u - u_h$ gives

$$\int k|\nabla e|^2 = \int re - \int (\hat{q} + k\nabla u_h) \cdot \nabla e.$$

With $k \geq k_{\min} > 0$ and a Poincaré constant C_P such that $\|e\| \leq C_P \|\nabla e\|$, the energy norm is bounded by $\|k^{-1/2}(\hat{q} + k\nabla u_h)\| + C_P\|r\|/\sqrt{k_{\min}}$. The second term is the price of imbalance. If source data have been projected onto simpler functions during reconstruction, include their residual against the actual source before calling the result a bound.

Carry out the entire construction on two elements

On $[0, 1]$ let $k = 1,\ f = 2,\ u(0) = u(1) = 0$. The exact $u = x(1 - x)$ is available for checking although unnecessary for the estimator. The conforming piecewise-linear interpolant on two half-elements has slopes $+0.5, -0.5$. Any flux $\sigma(x) = 2x + C$ exactly balances the source. Choose C to minimise $\int(\sigma + v')^2 dx$. Its derivative is $2\int(2x + C + v')dx = 2(1 + C)$, since $v(1) - v(0) = 0$. Thus $C = -1$, giving $\sigma = 2x - 1 = q$. The estimator is $\sqrt{1/12} = 0.288675$: on each half interval the defect is a linear ramp from -0.5 to $+0.5$ or its translated equivalent. Direct integration of $(u' - v')^2$ confirms equality.

If instead choose $C = -0.8$, source equilibrium is still exact but the flux carries a uniform 0.2 offset. Its squared error contribution is 0.04, giving bound $\sqrt{1/12 + 0.04} = 0.3512$. Choosing an equilibrated flux means satisfying local balance; optimising its closeness to the constitutive flux sharpens the estimate. The true temperature error is unchanged in both comparisons. In more dimensions mixed finite elements or local face-flux reconstructions enforce shared normal flux degrees of freedom; local least-squares choices can improve sharpness without requiring a global exact solution.

Use element contributions without confusing local and global bounds

The squared estimator splits as $\sum_K \int_K k^{-1}|\sigma + k\nabla v|^2$. Large element contributions can guide refinement and give a reproducible stopping rule for global energy error. The exact orthogonality proof, however, integrated over the full domain. Boundary terms between elements may fail to vanish on one element individually, so one local contribution need not be a strict independent upper bound on that element's actual error. For a particular output of interest, e.g. a mean temperature in a small sensor region, use the dual problem or a proven continuous inequality to translate energy uncertainty into output uncertainty.

Vary an equilibrated flux while holding the solution fixed

The source and exact quadratic temperature are those of the worked one-dimensional example. The plot compares the exact flux, the raw elementwise flux from the continuous linear temperature and a reconstructed flux with the correct divergence and a freely chosen constant offset. Changing this constant is permitted because the original problem specifies boundary temperatures. The mean energy-error contribution from an offset appears in quadrature with the true temperature-gradient error, exactly as in the derived identity. For two elements and offset 0.20 the bound is about 0.3512 while the actual energy error is 0.2887.

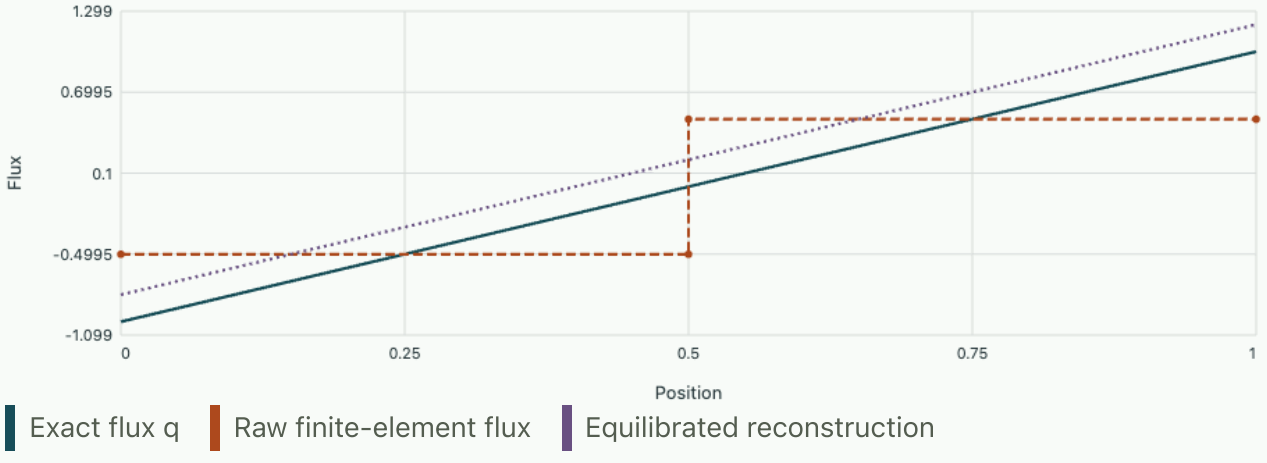
Increase the number of cells with the offset held fixed. The raw staircase becomes more accurate, but the reconstruction bound eventually plateaus near the absolute offset because the reconstructed flux still contains its deliberate uniform error. Set the offset to zero to recover the sharp certificate on each mesh. Notice that a constant flux offset preserves the source balance in this Dirichlet test; it would generally violate a specified Neumann flux at a boundary. A valid estimator depends as much on the boundary conditions as on the differential equation.

Open this experiment online

Try the calculation

Uniform elements **2**

Flux offset from optimum **0.2**



True energy error 0.28868, guaranteed reconstructed bound 0.35119. The offset does not change the exact divergence.

Tutorial · verify an error certificate rather than a residual plot

Predict refinement on a uniform mesh

Use the same quadratic exact solution and N equal elements. Find the energy error of the nodal interpolant and verify the two-element value.

▼ **Work through the solution**

On each interval v' is the average of the linear exact derivative $1 - 2x$. The derivative difference is a line of slope -2 centred at the midpoint. Its squared integral is

$$4h^3/12 = h^3/3.$$

Summing N terms gives error $h/\sqrt{3}$. At $h = 1/2$ this equals $1/\sqrt{12}$. Halving the element size halves the energy error even though the nodal temperature error for this interpolation is exactly zero at the nodes.

Diagnose a spuriously perfect estimator

An engineer chooses $\sigma = -k\nabla v$ directly and gets a zero constitutive defect. Why does this not prove $v = u$ for the two-element example?

▼ **Work through the solution**

This piecewise-constant flux has zero derivative inside each element and a jump at the middle node. It fails the required divergence equation $f = 2$ as a distributed source and fails normal continuity. The missing volume and interface cross terms invalidate the orthogonality argument. A zero defect certifies a true solution only alongside both primal admissibility and flux equilibrium with the stated boundary conditions.

Bound the error of a mean temperature

On the unit interval with zero endpoint error, use $\|w\|_2 \le \|w'\|_2/\pi$ to bound the domain-mean error from the sharp two-element certificate. Compare with direct integration.

▼ **Work through the solution**

Cauchy–Schwarz on a unit-length domain gives

$$|\int (u - v)| \le \|u - v\|_2 \le 0.288675/\pi = 0.0919.$$

The exact integral of u is $1/6$; the triangular interpolant has area $1/8$, giving actual mean error $1/24 = 0.04167$, within the bound. An error certificate for the gradient can yield a conservative output bound through an independently justified inequality; its units and domain constants cannot simply be omitted.

Further reading

- INRIA · Vohralík on flux reconstruction and guaranteed error bounds
- Course investigation · output error and the dual problem
- Measure spatial convergence against an exact field

Spatial discretisation and field equations · 3. A Fourier Poisson solver and the trouble caused by unresolved products

In a periodic medium the constant part of a source decides whether a Poisson problem has any solution. Derive the Fourier solver, separate spectral and finite-difference symbols, and see how a nonlinear product can inject a false low-frequency source.

Find compatibility and gauge before dividing by a wave number

On a periodic interval of length 2π, solve $-u'' = f$. Integrating over one period gives $0 = \int f$ because periodic derivative values agree at the ends. This is an existence condition. If a solution exists, adding a constant changes neither field gradient nor source, so specify a mean u_0 separately to fix uniqueness. These two roles sometimes get conflated in numerical code that simply sets the Fourier coefficient at zero to zero. Setting $u_0 = 0$ fixes a gauge; deleting nonzero f_0 removes a physical mean source and changes an incompatible model. Check the measured source mean before applying a projected solve.

For $f(x) = \sum_k f_k e^{ikx}$, take $u_k = f_k/k^2$ for $k \neq 0$. For a length L in physical units replace k by $2\pi j/L$. A compatible source $f = \cos 3x + 0.2 \sin x$ gives mean-zero $u = \cos 3x/9 + 0.2 \sin x$. Higher spatial frequencies are smoothed by the inverse Laplacian; data noise at low but nonzero wavenumber can affect the inferred field more strongly than visually rapid fluctuations of equal source amplitude. A compatible constant reference level of potential cannot be inferred from this equation even with noiseless source data.

Compare an exact spectral solve with a discrete operator

On N equispaced nodes, an FFT diagonalises any periodic translation-invariant discrete convolution, but the diagonal symbol depends on the operator being solved. A spectral second derivative uses k^2 for $-u''$, while a second-order periodic difference stencil uses $\lambda_h(k) = 4\sin^2(kh/2)/h^2$, $h = 2\pi/N$. The same FFT can invert either symbol. Verify a solver by applying the matching discrete operator back to its result. A Fourier result with tiny residual for the finite-difference stencil is a precise solution of that stencil; it does not thereby reproduce the continuum Fourier amplitude exactly.

At $N = 16, k = 4$, $kh/2 = \pi/4$, hence $\lambda_h = 2/h^2 = 12.97$, compared with continuum $k^2 = 16$. A unit cosine source has spectral potential amplitude 0.0625; the finite-difference inverse gives approximately 0.0771. Both can pass a discrete residual at machine precision in their appropriate systems. At fixed physical mode the symbols approach one another as N increases; at mode number constantly near the Nyquist limit their relative discrepancy need not vanish. A convergence test keeping k/N fixed follows different physical sources at each refinement.

Identify aliased modes before applying the inverse

For N samples, Fourier modes with indices differing by an integer multiple of N coincide at all nodes. Suppose a source includes v^2, where $v = \cos(5x)$ on a 12-point grid. Squaring exactly yields

$[1+\cos(10x)]/2$. But sampled $\cos 10x = \cos 2x$, so the grid reports an apparent low $k = 2$ component. Subtracting the known mean gives a compatible source, but the resulting potential amplitude becomes $0.5/2^2 = 0.125$ at the wrong spatial frequency rather than $0.5/10^2 = 0.005$ at the true one. The inverse Laplacian magnifies the false low-mode contribution, transforming a sampling error into a large field error.

To obtain the product projected on a retained Fourier band, zero-pad the input Fourier coefficients on a sufficiently enlarged grid, transform to space, multiply there, transform back and discard unretained output modes. For modes $|k| \le K$, a quadratic product extends to $2K$. To ensure none of its high modes folds back into the retained range, choose multiplication grid size strictly greater than $3K$, or handle endpoint/Nyquist modes explicitly. This is the usual three-halves padding rule relative to a base grid around $2K$. Different FFT normalisations require scaling when padding and truncating. Simply padding already multiplied samples cannot recover frequency information that has already aliased.

For a deliberately filtered small problem one can instead drop the top third of the spectrum on a fixed grid before nonlinear multiplication, retaining only $|k| < N/3$. This avoids erroneous feedback into the retained modes but also discards legitimate high-frequency information. Choosing a finer physical grid can be preferable when the physical quantity of interest depends on those modes. Dealiasing corrects a numerical representation of the chosen projected equations; it does not supply a physical closure for unresolved turbulence, interfaces or material fluctuations.

Follow one unresolved harmonic through the Poisson inverse

The simulation multiplies a cosine of mode m by itself, removes its exact mean and inverts the periodic Poisson operator. The continuum curve retains the true product mode 2m. The unpadded curve uses whichever harmonic the coarse point samples represent, while the third curve shows the correctly projected solution restricted to the original resolvable Fourier band. At 12 samples and m equal to 5, coarse multiplication creates a false mode two whose potential amplitude is 25 times the desired continuum amplitude. The resolved-band projection correctly eliminates that harmonic rather than inventing a low mode.

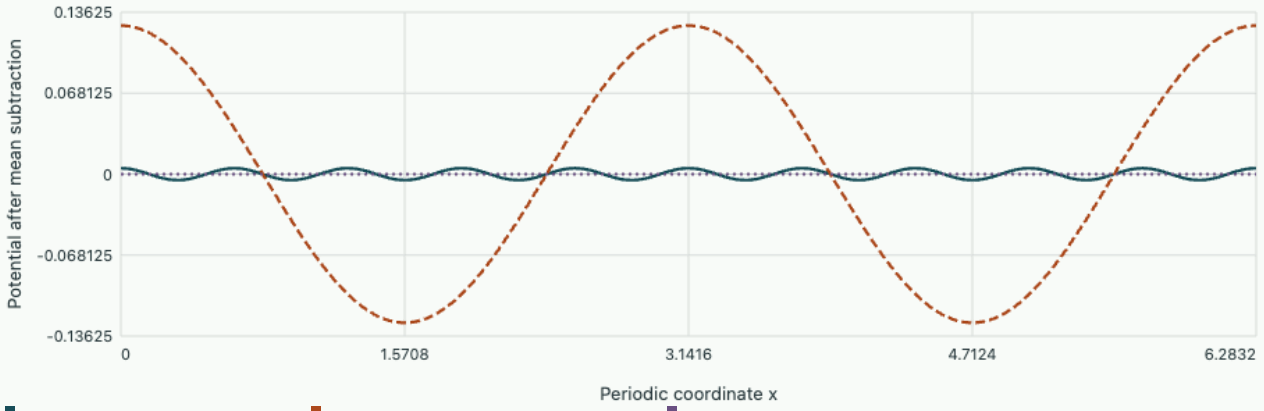
Increase N until the true product is representable. The three answers then agree, including the phase of the spatial variation. The control limits the actual input frequency to a representable non-Nyquist input and reports the mode used, so changing the grid cannot silently inject an impossible input. A projected curve equal to zero means none of this particular product survives the retained band. It does not mean the continuum product or the underlying fine-scale field vanishes.

Open this experiment online

Try the calculation

Base grid samples N 12

Requested input mode m 5



Continuum solution | Unpadded grid product | Resolved-band projection

Representable input m = 5, product mode 10, coarse alias 2. Continuum amplitude 0.00500, coarse amplitude 0.12500.

Tutorial · verify mean source, discrete response and nonlinear products

Detect an impossible boundary-value problem

A periodic source has $f(x) = 0.03 + \sin x$. A program reports $u = \sin x$ and a machine-small residual after dropping the zero Fourier mode. Evaluate the actual residual and explain its interpretation.

▼ **Work through the solution**

For the reported solution $-u'' = \sin x$, so the original pointwise residual $-u'' - f = -0.03$ everywhere and its integral is -0.06π. There is no periodic solution to the original equation. The programme exactly solved the mean-subtracted problem. An external sink or a physically justified zero-mean correction could make such a model appropriate, but the removed source must be stated and checked.

Predict the grid error of a low physical mode

For $k = 1$, expand $\lambda_h(k)$ and find the leading relative error in the finite-difference solution amplitude for a fixed unit cosine source.

▼ **Work through the solution**

Using $\sin^2 z = z^2 - z^4/3 + \cdots$ gives

$$\lambda_h = k^2 - k^4 h^2/12 + O(h^4).$$

Inversion yields

$$1/\lambda_h = k^{-2}[1 + k^2 h^2/12 + O(h^4)].$$

The discrete potential amplitude is too large to leading order. Refining at the same k by halving h should reduce the amplitude error by roughly four. This statement differs from finding the residual of the already solved stencil.

Choose a padding size from the desired output band

Input modes occupy $|k| \le 5$, and only output modes $|k| \le 5$ of their quadratic product will be kept. Compare multiplication grids with 14 and 16 points.

▼ Work through the solution

With 14 points a possible product $k = 10$ aliases to -4, contaminating the retained range. With 16 points all modes $6, \ldots, 10$ which wrap do so outside $[-5, 5]$; for example 10 wraps to -6. Thus 16 is sufficient for retained coefficients. It is not sufficient to represent and report the complete $k = 10$ output wave itself, which needs a larger bandwidth. Distinguish accuracy of a projected nonlinear evolution from recovery of every generated physical harmonic.

Further reading

- ETH Zürich · numerical methods for CSE, Fourier methods and filtering
- MIT · Steven Johnson on FFT differentiation and the Nyquist mode
- Course investigation · consistency of spatial discretisations
- Measure spatial convergence against an exact field

Spatial discretisation and field equations · 4. How much of an outgoing wave reflects from the numerical boundary?

A finite grid truncates an effectively open medium. Derive an exact one-dimensional radiation condition, predict the angular reflection of its two-dimensional extension and design a pulse experiment that separates boundary reflection from interior dispersion.

Factor the one-dimensional wave equation

For a constant wave speed $c > 0$,

$$u_{tt} - c^2 u_{xx} = (\partial_t - c\partial_x)(\partial_t + c\partial_x)u = 0.$$

A right-moving waveform $F(x - ct)$ satisfies $u_t + cu_x = 0$ and a left-moving waveform satisfies $u_t - cu_x = 0$. At the right boundary impose $u_t + cu_x = 0$ to transmit a purely outgoing right-moving wave; at the left boundary reverse the normal sign. This exact radiation condition requires no knowledge of F for the constant-coefficient 1D medium. A Dirichlet end instead forces a reflected amplitude minus the incident one, while a Neumann end reflects with the same sign. Both have unit reflected energy for an ideal undamped wave.

A separate energetic argument tests the sign. The interior energy $E = \frac{1}{2}\int (u_t^2 + c^2 u_x^2)dx$ has derivative $dE/dt = c^2[u_t u_x]_{\text{left}}^{\text{right}}$. At a transmitting right end $u_x = -u_t/c$, giving outward energy loss $-cu_t^2$. Imposing the opposite sign there would permit energy injection. A dissipative boundary need not absorb perfectly: it can still reflect some of the incoming wave. The energy identity tests passivity, while a pulse-amplitude experiment tests reflection separately. An initially localised pulse that has not yet reached any boundary is a useful independent check of the interior solver.

Predict reflection as the incidence direction changes

For a flat boundary normal to $+x$ in two dimensions, the local condition $u_t + cu_x = 0$ is exact only for normal incidence. Consider incident wave $e^{i(k_x x + k_y y - \omega t)}$ and reflected wave of amplitude R with k_x reversed. Use $\omega = c\sqrt{k_x^2 + k_y^2}$ and place the boundary at $x = 0$. Substitution into the condition gives $(-\omega + ck_x) + R(-\omega - ck_x) = 0$, hence

$$R = \frac{ck_x - \omega}{ck_x + \omega} = \frac{\cos\theta - 1}{\cos\theta + 1},$$

where θ is measured from the outward normal. The reflected normal-energy fraction is $|R|^2$ because incident and reflected magnitudes and normal speed factors match. Normal incidence gives zero. At 60°, $R = -1/3$ and power reflection $1/9$; at grazing incidence its magnitude approaches one. An unspecified claim of a “nonreflecting boundary” has no definite quantitative meaning without frequency, angle, medium and discretisation.

An ideal exterior transparent condition can be expressed after transforming in tangential direction and time: $\partial_x \hat{u} = ik_x \hat{u}$ with $k_x = \sqrt{(\omega/c)^2 - k_y^2}$ on the outgoing or decaying branch. The square root

shows why a single local first derivative cannot cover all tangential frequencies exactly. Expanding the square root near normal incidence motivates higher-order absorbing boundary conditions; other methods use a carefully matched absorbing layer. The present first-order formula provides a baseline against which such alternatives can be tested. An evanescent component with imaginary normal wavenumber requires the decaying branch; interpreting it as a propagating beam with a real incidence angle would be incorrect.

Measure boundary error independently of grid dispersion

An interior finite-difference solver usually has a numerical dispersion relation $\omega_h(k_x, k_y)$. A discrete version of the continuum boundary condition need not be exactly matched even for a strongly resolved normally incident wave, particularly if boundary and interior stencils have different dispersion errors. To isolate the two effects, first propagate the same pulse on a much longer or periodic domain until it passes the observation probe and before any distant return. Record that reference. Then repeat with the open boundary placed closer and compare at the same probe over the predicted return window. Project signals onto spatial directions or separate incident and reflected time windows to estimate amplitude and energy.

For a probe distance D from a flat boundary and a narrow incident beam at angle θ, normal group speed is $c\cos\theta$. The time between incident and reflected peaks measured on a normal coordinate is roughly $2D/(c\cos\theta)$, provided the probe or array also samples the reflected beam's tangential position. A point probe on an oblique pencil beam need not see both pulses at all; an extended plane-wave excitation or a probe array is more suitable for estimating the stated reflection coefficient. Run a sequence of grid refinements at fixed physical angle and wavelength. Reflection approaching the analytical $|R|$ indicates continuum boundary-model error; reflection shrinking toward zero at normal incidence indicates discretisation mismatch.

Measure a boundary reflection across incidence angles

The plotted energy fractions follow directly by matching an incoming and a reflected plane wave at a flat boundary. One curve uses the physically correct wave speed in the first-order condition; the other lets the specified boundary speed differ from the interior. Use the angle slider to inspect the signed amplitude as well as its squared magnitude. At matched speed and sixty degrees the amplitude is minus one third and the reflected energy is about 11.11 percent. At normal incidence the matched boundary absorbs the incident plane wave completely.

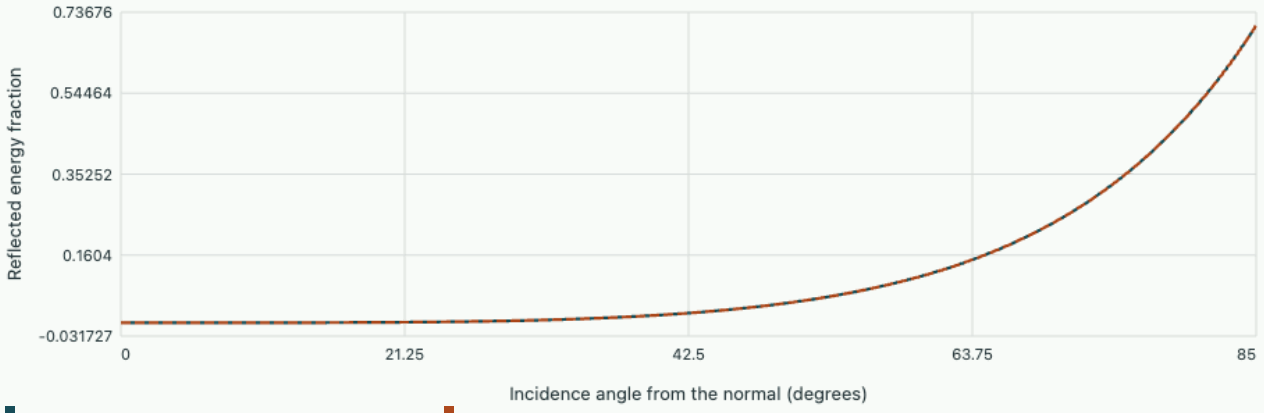
Choose a boundary speed larger than the interior speed. There is then an oblique angle at which the mismatch exactly compensates the projected normal phase in this local equation. It obeys cosine of the angle equal to the reciprocal speed ratio. For a ratio 1.25 this is about 36.87 degrees. Matching this one direction sacrifices perfect normal incidence; it is not an omnidirectional improvement. The graph concerns the continuous condition. An actual spatial discretisation can add reflection through numerical dispersion, which the longer-domain pulse comparison in this unit is designed to detect.

Open this experiment online

Try the calculation

Boundary to physical speed ratio **1**

Inspected angle (degrees) **60**



Selected first-order boundary Speed-matched first-order boundary

At 60° the selected signed amplitude is -0.3333, reflected energy 11.11%.

Tutorial · identify the return signal

Determine the admissible incidence angle for one-percent reflected energy

Solve $|R|^2 \leq 0.01$ for a real outgoing incidence angle $0 \leq \theta < 90^\circ$.

▼ Work through the solution

The amplitude criterion is $(1-\cos\theta)/(1+\cos\theta) \leq 0.1$, so

$$\cos\theta \geq 0.9/1.1 = 0.81818.$$

This gives $\theta \lesssim 35.1^\circ$. Requiring one-percent amplitude instead would impose a stricter angle. A specification should state whether it concerns a field amplitude or a quadratic energy measurement.

Calculate reflection from an incorrectly tuned 1D boundary

The physical medium has speed c but the implemented condition uses $u_t + c_b u_x = 0$ with $c_b = 1.20c$. Find the normally reflected amplitude and energy fraction.

▼ Work through the solution

The right-travelling dispersion gives

$$k_x = \omega/c.$$

Repeating the boundary substitution yields

$$R = (c_b/c - 1)/(c_b/c + 1) = 0.20/2.20 = 0.0909,$$

with energy reflection 0.00826. Its phase sign differs from the negative reflection arising for oblique incidence with a correctly tuned speed. A reflected waveform's sign can thus help diagnose a mismatch rather than reporting only its energy.

Design a time window to catch a weak return

A plane pulse propagates at 200 m/s and a line of probes is 3.0 m inside the boundary. Estimate normal-incidence and 60° return delays. Which comparisons are needed when a 2% amplitude is measured at normal incidence?

▼ **Work through the solution**

The delays are

$$2(3)/200 = 0.030\,\mathrm{s},$$
$$2(3)/(200 \times 0.5) = 0.060\,\mathrm{s}.$$

The continuum boundary predicts zero at normal incidence, so check interior numerical dispersion, boundary stencil, angular spread of the finite-width excitation, signal extraction and any physical material change near the boundary. A larger reference domain and refinement at fixed wavelength distinguish numerical return from pulse tails. A 2% amplitude corresponds to 0.04% reflected energy under matching impedances, not 2% energy.

Further reading

- Engquist and Majda · absorbing boundary conditions for numerical wave simulation
- Course investigation · numerical propagation of waves
- Measure spatial convergence against an exact field

Spatial discretisation and field equations · 5. Can a stable finite-element heat step turn a cold node negative?

Start a conducting rod with one positive temperature pulse. Assemble the heat capacity from first principles and determine when an explicit step respects both decay and the sign of every temperature.

Integrate the heat equation against a hat function

Take a uniform rod of length one with the two endpoint temperatures held at zero. With constant diffusivity $\kappa > 0$ the evolution is $u_t = \kappa u_{xx}$. Approximate $u(x,t)$ as a linear combination of continuous piecewise-linear hat functions. Multiply by an interior hat and integrate. Integration by parts transfers one spatial derivative onto each hat. The boundary term vanishes. The nodal temperature vector satisfies

$$M\dot{u} = -Ku.$$

On an element of length h the two hats are $1 - x/h$ and x/h in local coordinates. Integrating their products gives diagonal entries $h/3$ and off-diagonal entries $h/6$ in the element mass matrix. Their constant derivatives give diagonal stiffness κ/h and off-diagonal stiffness $-\kappa/h$. After assembly and imposing Dirichlet values, the interior mass has diagonal $4h/6$ and neighbouring entries $h/6$. The stiffness has diagonal $2\kappa/h$ and neighbours $-\kappa/h$.

Forward Euler advances with $G = I - \Delta t M^{-1}K$. On an eigenvector of $K\phi = \lambda M\phi$ it multiplies the amplitude by $1 - \Delta t\lambda$. A bound $\Delta t\lambda_{\max} \le 2$ therefore ensures no growth in the mass norm. Preserving nonnegative temperatures for every nonnegative initial vector requires every entry of G to be nonnegative. Each column is the response to a single positive nodal impulse, so the latter condition can be tested directly.

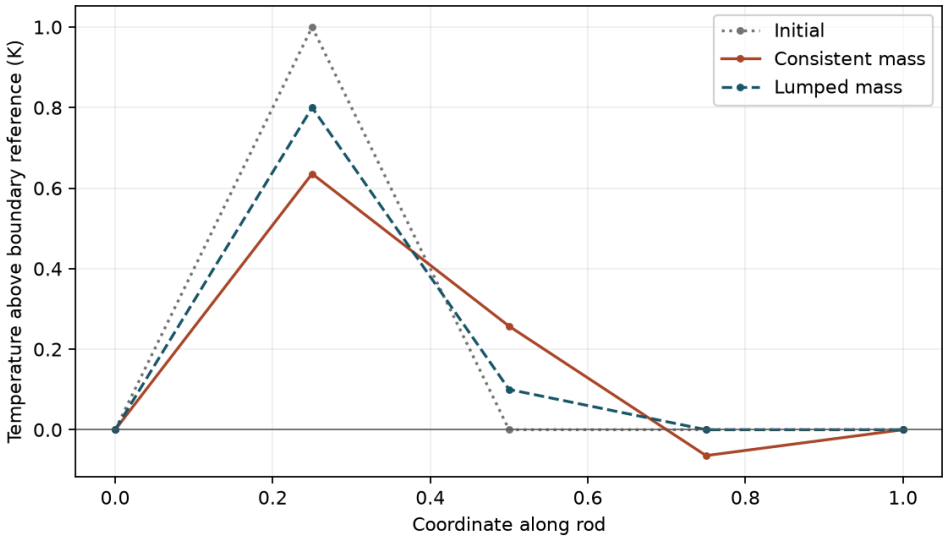
One alternative assigns half of each element's total heat capacity to each endpoint, before removing boundary degrees of freedom. Each interior mass then equals h and the explicit stencil has weights $r, 1-2r, r$, with $r = \kappa\Delta t/h^2$. These coefficients are all nonnegative for $0 \le r \le 1/2$. The first and last interior equations also couple to the fixed zero boundary temperatures.

Compute the remote temperature after one update

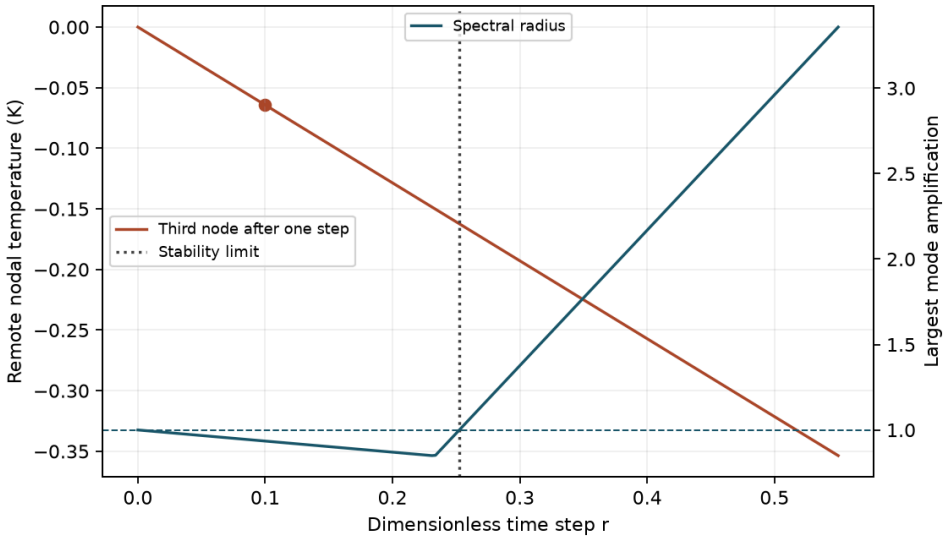
With three interior nodes, $h = 1/4$, initially set the left interior node to 1 K above the boundary reference and both others to zero. Inverting the consistent mass in this three-by-three system gives the first column of $(h^2/\kappa)M^{-1}K$ as $(51, -36, 9)/14$. At $r = 0.1$ the interior temperatures after one step become $(0.63571, 0.25714, -0.06429)$ K. The cold node two edges away immediately drops below the lower reference.

The largest dimensionless eigenvalue is about 7.9223. The chosen step gives $r\lambda_{\max}h^2/\kappa \approx 0.7922$, comfortably inside the linear stability interval. Lumping instead gives $(0.8, 0.1, 0)$ K. The negative consistent-mass entry shrinks linearly when the step shrinks. For this impulse its sign persists for every positive first step. With OneSimulation installed locally, the notebook also constructs a physical line mesh and obtains the mass and stiffness by integrating its hat functions at mapped quadrature points.

Compare the assembled result against the explicit tridiagonal entries. The experiment shows why reducing the time step can satisfy a stability test without restoring a discrete maximum principle.



Line segments are the linear finite-element interpolants. Both ends have a fixed zero temperature and participate in the heat balance.



The orange trace is the one-step temperature at the third interior node. The vertical line is the exact stability threshold for this mesh. Crossing it eventually produces amplification greater than one.

Exercise · keep the lumped temperatures nonnegative

Suppose $h = 0.02$ m and $\kappa = 8 \times 10^{-5}$ m²/s. Find the greatest time step allowed by nonnegative weights. What stencil results at that step?

▼ Work through the solution

The condition gives $\Delta t = h^2/(2\kappa) = 2.5$ s. The central coefficient vanishes and each adjacent value receives weight one half. At the first interior node the fixed boundary value supplies one of these two

terms. A step slightly above this threshold gives a negative central coefficient and can drive an isolated positive impulse below zero.

Exercise · derive the largest mode on successively finer meshes

An interior sine wave has neighbouring sum $2\cos(\theta)$ times its own nodal value. Use this fact to find the limiting stability condition for forward Euler with consistent mass as the number of uniform elements grows.

▼ **Work through the solution**

The mass eigenvalue is $h(2+\cos\theta)/3$ and the stiffness eigenvalue is $2\kappa(1-\cos\theta)/h$. Their ratio, multiplied by h^2/κ, is $6(1-\cos\theta)/(2+\cos\theta)$. The largest allowed angle tends to π, giving 12. The limiting stable step therefore has $r \le 1/6$. A positivity test still inspects the entries of the update matrix separately.

Change parameters · Explore this notebook model

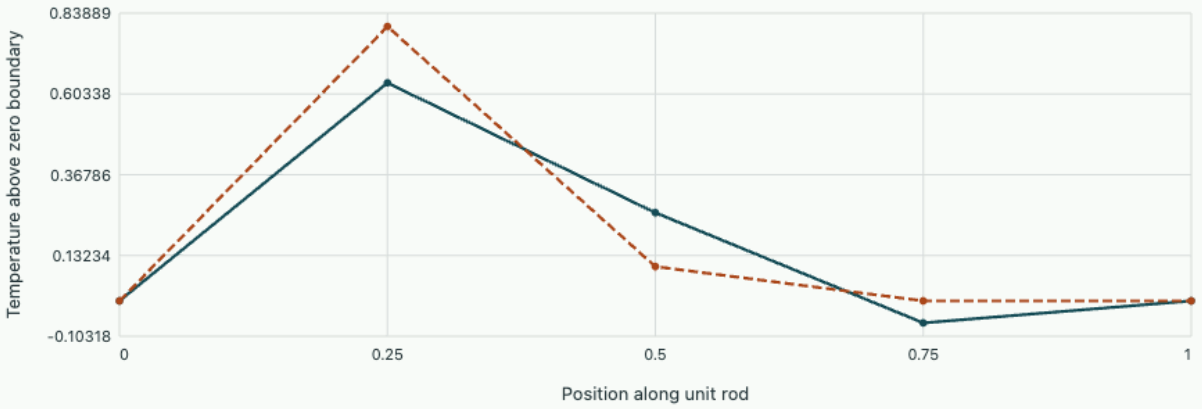
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Interior nodes **3**

Diffusion number r **0.1**



Consistent-mass stability limit r = 0.25245. Third interior temperature -0.064286. Spectral stability and preservation of nonnegative temperature are separate properties.

Open editable notebook Download .ipynb

Spatial discretisation and field equations · 6. Solving two-dimensional diffusion one direction at a time

Derive a second-order alternating-direction step, identify the boundary data required halfway through it and distinguish splitting error from mesh and time errors.

Factor the two-dimensional heat operator

For $u_t = \kappa(u_{xx} + u_{yy})$ on a rectangle, first impose homogeneous Dirichlet data. Centred second differences on a Cartesian mesh give $\dot{U} = (X + Y)U$, where X, Y are the directional diffusion matrices. For constant diffusivity and these boundaries the Kronecker factors commute and are symmetric negative definite. Put $s = h/2$. Multiplying out a factored Crank–Nicolson step introduces $s^2 XY$ on both sides:

$$(I - sX)(I - sY)U^{n+1} = (I + sX)(I + sY)U^n.$$

Applied to a smooth solution the added defect is $s^2 XY(U^{n+1} - U^n) = O(h^3)$ at fixed mesh. The Peaceman–Rachford realisation uses line solves

$$(I - sX)U^* = (I + sY)U^n, \qquad (I - sY)U^{n+1} = (I + sX)U^*.$$

For a shared eigenmode with directional eigenvalues $\lambda_x, \lambda_y \le 0$, the amplification factor is the product of their midpoint factors. Each has modulus at most one and the product matches the exponential through order two. This argument establishes unconditional stability for the separable case. For highly damped modes a midpoint factor can change sign. Stability alone does not guarantee monotonicity or accurate damping.

Give the intermediate grid a mathematically consistent boundary

An intermediate variable solves an algebraic factor equation. It has no general obligation to equal the physical wall data at the midpoint time. For inhomogeneous boundaries choose a smooth lifting w whose trace equals the specified boundary values at all times, and evolve $v = u - w$ with zero trace and forcing $f + \kappa\Delta w - w_t$. For $u = a(t)x$ with prescribed trace on all sides, $w = a(t)x$ and $f = a'(t)x$ give identically zero lifted field. A nonzero computed field in this manufactured case isolates a boundary or source inconsistency.

For a smooth lifted source g, add $sg(t_n + s)$ to both line-equation right sides. Expansion of the first stage and substitution into the second yields $hg + h^2(X + Y)g/2$ through second order, along with the midpoint approximation to the source time variation. Mixing time levels at the wall can instead inject a first-order defect. Check that spatial derivatives of the lifted source stay resolved when taking a simultaneous time–space limit.

OneSimulation experiment · Propagate one heat mode in two directions

Double the spatial resolution while holding the final time fixed. Keep the grid-dependent reference in the comparison.

Download the executable notebook · Python source

Adjust the heat field and rotate its surface. Take initial data $\sin(\pi x)\sin(q\pi y)$ on the unit square with zero wall values and $\kappa = 1$. Its directional grid eigenvalues are $\lambda_x = -4\sin^2(\pi\Delta x/2)/\Delta x^2$ and $\lambda_y = -4\sin^2(q\pi\Delta x/2)/\Delta x^2$. Each time step multiplies its amplitude by

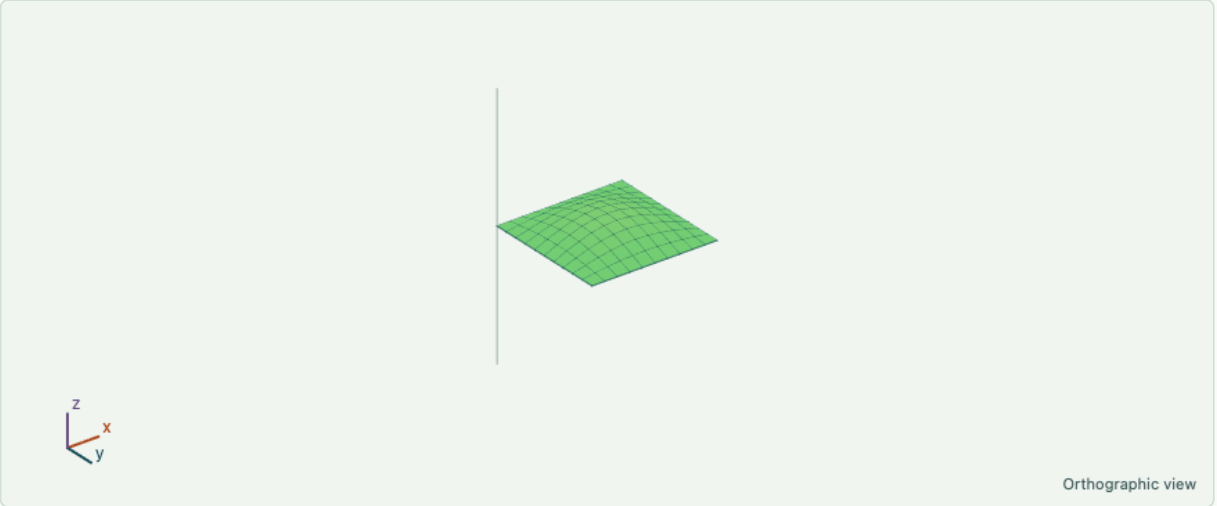
$$G = \frac{1 + h\lambda_x/2}{1 - h\lambda_x/2}\,\frac{1 + h\lambda_y/2}{1 - h\lambda_y/2}.$$

The browser evaluates G^N, exactly the mode reduction of the two matrix solves in the OneSimulation notebook. The rotatable surface shows the numerical field; its vertical coordinate is temperature, so this remains a two-dimensional spatial problem. The plot compares an antinode slice with both $e^{(\lambda_x+\lambda_y)t}$ and the continuum amplitude $e^{-\pi^2(1+q^2)t}$. First double the number of time steps on one fixed mesh, then refine the mesh. Choose a higher vertical mode and very few steps to see stable but inaccurate damping, sometimes with a reversed sign. A vertical reference from –1 to +1 keeps amplitude loss visible as the field cools.

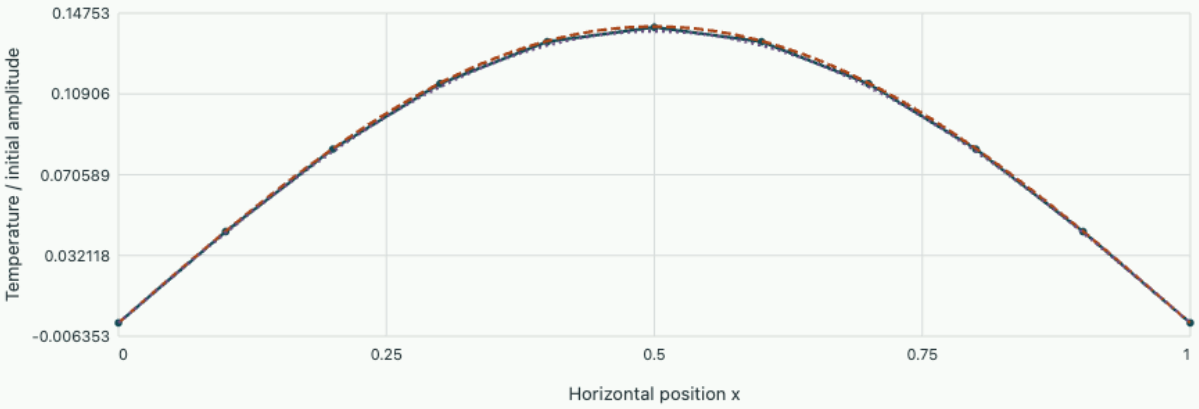
Open the live experiment

Try the calculation

Interior grid points per direction	9
Final time	0.1
Time steps	8
Vertical mode number q	1



Three-dimensional graph of a two-dimensional heat equation: height and colour show temperature, not a third spatial coordinate. The mesh is the ADI numerical solution. Rotate it to see the q vertical lobes; the fixed −1 to +1 height reference makes decay visible. Axes: x = x / L; y = y / L; z = u / u_0.



ADI grid mode, antinode slice
Exact time evolution on this grid
Continuum heat equation

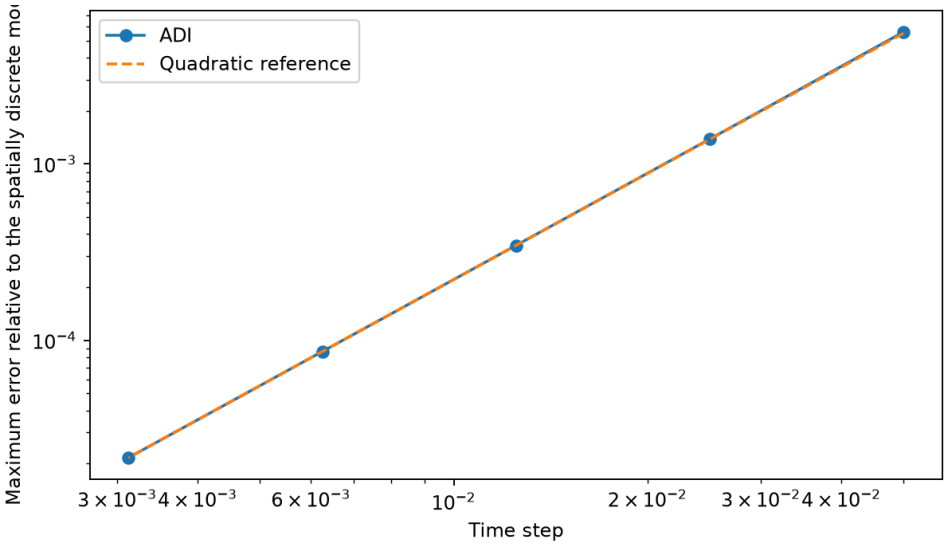
ADI amplitude = 0.14083. Temporal amplitude error = 0.00034519; spatial amplitude error = 0.0022661. The slice is at y = 1/(2q) = 0.5. Diffusivity and side length are one.

```
"""Measure the temporal order of an ADI heat step against an exact grid mode."""
import matplotlib.pyplot as plt
import numpy as np
from onesim.numerics.linear.matrices.dense_linear_algebra import solve_dense
```

```
m = 9
spacing = 1/(m+1)
d = (np.diag(-2*np.ones(m)) + np.diag(np.ones(m-1), 1) + np.diag(np.ones(m-1), -1))/spacing**2
sine = np.sin(np.pi*spacing*np.arange(1, m+1))
initial = np.outer(sine, sine)
decay = -4*np.sin(np.pi*spacing/2)**2/spacing**2
end = 0.1
ref = initial*np.exp(2*decay*end)
steps, errors = np.array([2, 4, 8, 16, 32]), []
for count in steps:
    plus, minus = np.eye(m) + end*d/(2*count), np.eye(m) - end*d/(2*count)
    state = initial.copy()
    for _ in range(count):
        half = solve_dense(minus, state @ plus)
        state = solve_dense(minus, (plus @ half).T).T
    errors.append(np.max(abs(state-ref)))
orders = np.log2(np.array(errors[:-1])/errors[1:])
print(f"Finest-grid temporal order: {orders[-1]:.4f}")
assert abs(orders[-1]-2) < 0.01
fig, ax = plt.subplots(figsize=(7, 4), layout="constrained")
ax.loglog(end/steps, errors, "o-", label="ADI")
ax.loglog(end/steps, errors[-1]*(steps[-1]/steps)**2, "--", label="Quadratic reference")
ax.set(xlabel="Time step", ylabel="Maximum error relative to the spatially discrete mode")
ax.legend()
fig.savefig("numerics-adi-heat.png", dpi=170)
```

Output from the reference run

```
Finest-grid temporal order: 2.0003
```



Directional solves give second-order temporal convergence against the exact semidiscrete solution.

Measure transport, time and splitting independently

A time-dependent transport–reaction problem can have noncommuting operators even on a periodic grid. Let $Au = -cu_x,\ Bu = -k(x)u$. Direct differentiation gives $(AB - BA)u = ck'(x)u$. Expanding a product $e^{hB}e^{hA}$ through h^2 and subtracting $e^{h(A+B)}$ leaves $h^2(BA - AB)u/2$. Reversing the order reverses this leading defect. The symmetric composition $e^{hA/2}e^{hB}e^{hA/2}$ cancels it, giving second-order global convergence for sufficiently smooth data and compatible boundaries.

An independent solution follows characteristics: $u(x,t) = u_0(x - ct)\exp[-\int_0^t k(x - c(t - s))\,ds]$. For $k = k_0 + k_1 \sin x$, the integral is $k_0 t + (k_1/c)[\cos(x - ct) - \cos x]$. Compare the split method first to a tightly integrated unsplit system on the same mesh, then to this continuum formula. Halve time at fixed mesh and halve the mesh at a suitably small time step. A single combined refinement cannot identify which contribution dominates.

> **How expensive is this factorisation?**
>
> On an m-by-m grid each half-step contains m tridiagonal systems of length m. A tridiagonal factorisation and substitution lead to $O(m^2)$ work per step. The small reference notebook deliberately uses the OneSimulation dense solver for each block of right-hand sides so the matrices can be inspected directly.

Tutorial on damping and noncommutation

Calculate the amplification of a strongly anisotropic mode

▼ **Work through the solution**

With $s\lambda_x = -10$ and $s\lambda_y = -0.1$, the factors are $-9/11$ and $9/11$. Their product is $-81/121$. The exact one-step decay is $e^{-20.2}$, which is much smaller. Halve the time step until damping agrees with the resolution required.

Design a test where reversing the split should change nothing

▼ **Work through the solution**

A constant reaction rate has $k' = 0$, so advection and reaction commute for periodic boundaries. With exact component updates the two orders agree at any time step. Change k to a sinusoid and use the order difference to expose the commutator.

LeVeque · diffusion and mixed time-dependent problems

Spatial discretisation and field equations · 7. Following the error from one reference triangle to an entire mesh

Derive a first-order energy error estimate and a second-order displacement estimate for a conforming linear triangular discretisation. Track the assumption on element shape at the point where it enters the proof.

Remove everything a linear element can reproduce

Let $\widehat{K}$ be a fixed triangle and $\widehat{I}$ its affine nodal interpolant. In two dimensions point evaluation is continuous on $H^2(\widehat{K})$, so $\widehat{I}: H^2 \to H^1$ is bounded. Choose an affine p to match the means of ∇u and of u on the triangle. Apply the Poincaré inequality first to the two components of $\nabla(u-p)$, then to $u-p$. This bounds the whole H^2 norm of $u-p$ by a constant times the L^2 norm of the second derivatives. Since $(I-\widehat{I})p = 0$,

$$|\widehat{u} - \widehat{I}\widehat{u}|_{H^1(\widehat{K})} \le C|\widehat{u}|_{H^2(\widehat{K})}.$$

This argument is the affine case of polynomial approximation by a seminorm. No constant in it depends on the eventual mesh. Choosing nodal values would require extra care for rougher solutions for which point evaluation is undefined.

Transform derivatives and area separately

Let $F(\xi) = B\xi + b$ map the reference triangle onto K. Physical gradients are B^{-T} times reference gradients. Pulling back second derivatives produces two factors of B. The square-root determinant factors from integrating and undoing the map cancel in the resulting estimate:

$$|u - I_K u|_{H^1(K)} \le C\|B^{-1}\|\,\|B\|^2\,|u|_{H^2(K)}.$$

A shape-regular family has uniformly bounded $\|B\|\|B^{-1}\|$ and element diameter comparable to $\|B\|$. Thus the right side is $Ch_K|u|_{H^2(K)}$. For a triangle with legs h and h^2 the general bound gives order one for its coefficient, since $\|B^{-1}\| \simeq h^{-2}$. This shows where the guarantee is lost. Some thin triangles and specially oriented solutions still interpolate accurately.

Use Galerkin orthogonality to turn approximation into a PDE estimate

For homogeneous Dirichlet diffusion with $0 < \kappa_{\min} \le \kappa(x) \le \kappa_{\max}$, set $a(v,w) = \int \kappa \nabla v \cdot \nabla w$. The conforming Galerkin solution satisfies $a(u-u_h, v_h) = 0$. Write $e = u - u_h$. Then $\|e\|_a^2 = a(e, u - I_h u) \le \|e\|_a \|u - I_h u\|_a$. Summing squared local bounds gives $\|e\|_a \le Ch\sqrt{\kappa_{\max}}|u|_{H^2}$. Here $u \in H^2$, boundary conformity and accurate assembly are substantive assumptions.

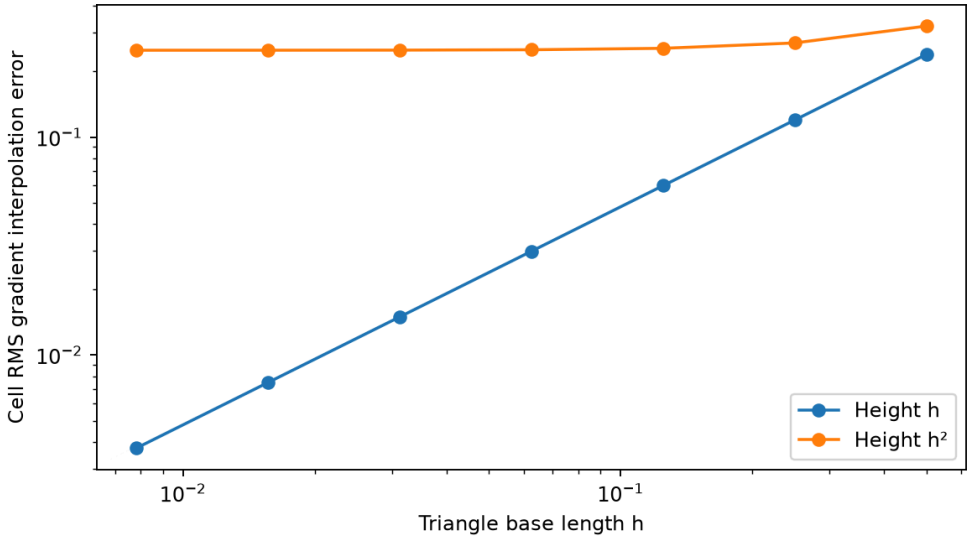
For the mean-square error solve the dual problem $a(v, z) = (v, e)$ with the same boundary conditions. If it satisfies the regularity estimate $\|z\|_{H^2} \le C_{\mathrm{reg}} \|e\|_{L^2}$, then

$$\|e\|_{L^2}^2 = a(e, z - I_h z) \le Ch \|e\|_a \|z\|_{H^2}.$$

Divide by $\|e\|_{L^2}$ to obtain the extra power of h. A reentrant corner or a coefficient interface can obstruct this regularity. Use the actual energy or output norm in the convergence study.

Why can a bad cell evade an affine patch test?

Every nonsingular triangle, including a very thin one, exactly reproduces an affine function. Such a patch tests assembly and transformation consistency. Curved test functions in several orientations are needed to expose geometry-dependent interpolation errors.



For vertices (0,0), (h,0), (h/2,ε) and u=x², the interpolant gradient is (h,−h²/(4ε)). The cell average of (2x−h)² is h²/6. At ε=h² the spurious transverse gradient persists as the cell shrinks.

Tutorial on local and global accuracy

Compute the interpolation error on a right reference triangle

▼ **Work through the solution**

For $u = x^2$ on vertices $(0,0), (h,0), (0,h)$, $I_K u = hx$. The squared energy error for unit diffusion is $\int_0^h (2x - h)^2 (h - x)\, dx = h^4/6$. On a family of similar cells the local energy error scales like h^2; the sum of squared errors over $O(h^{-2})$ cells scales like h^2.

Choose a mesh test for a collapsing angle

▼ **Work through the solution**

Compare a uniformly refined mesh to one containing triangles with B singular values h, h^2. Record the largest $\|B\|\|B^{-1}\|$, solution and energy errors, and compare to the local scaled bounds. Growth of a bound alone does not prove a particular solution diverges. Rotate the test Hessian to probe whether the poor direction couples to curvature.

Brenner and Scott · finite element approximation

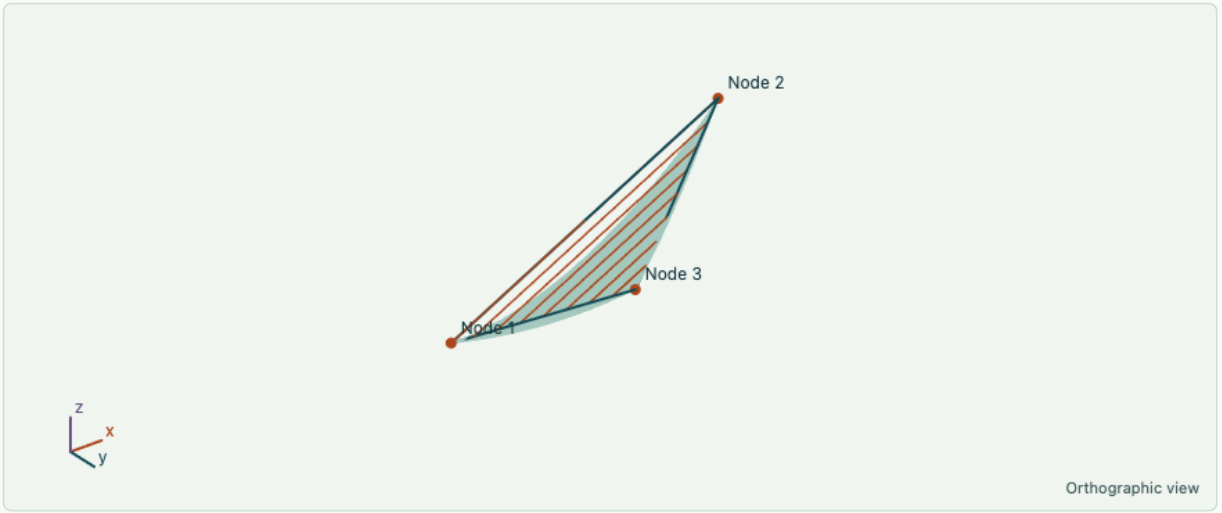
Change parameters · Following the error from one reference triangle to an entire mesh

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.

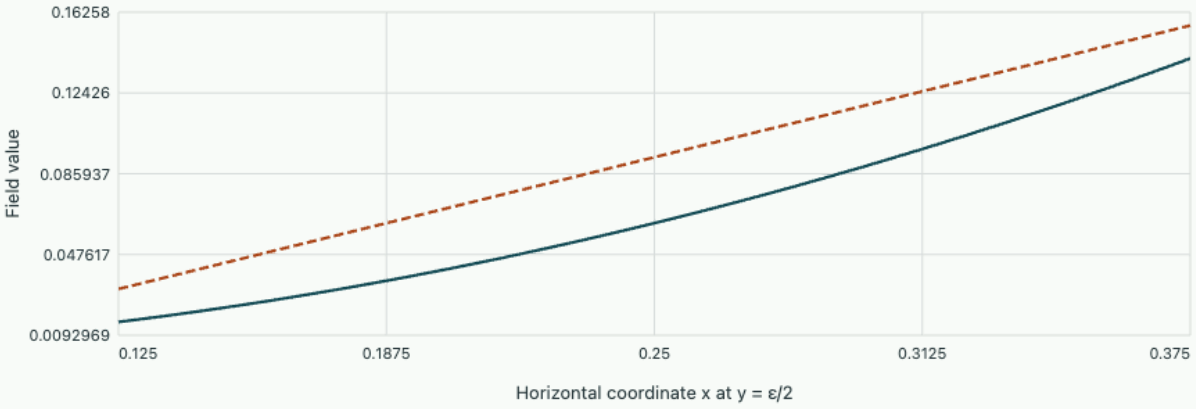
Open this calculation with adjustable inputs

Try the calculation

Base length h	**0.5**
Altitude / base	**0.25**



Height represents the scalar field, not a third spatial coordinate. Green faces: exact quadratic field on a two-dimensional triangle. Orange wire lines: its affine nodal interpolant. Thin elements amplify the transverse gradient error. Axes: x = x; y = y; z = field u.



Exact u = x² | Linear nodal interpolant at half altitude

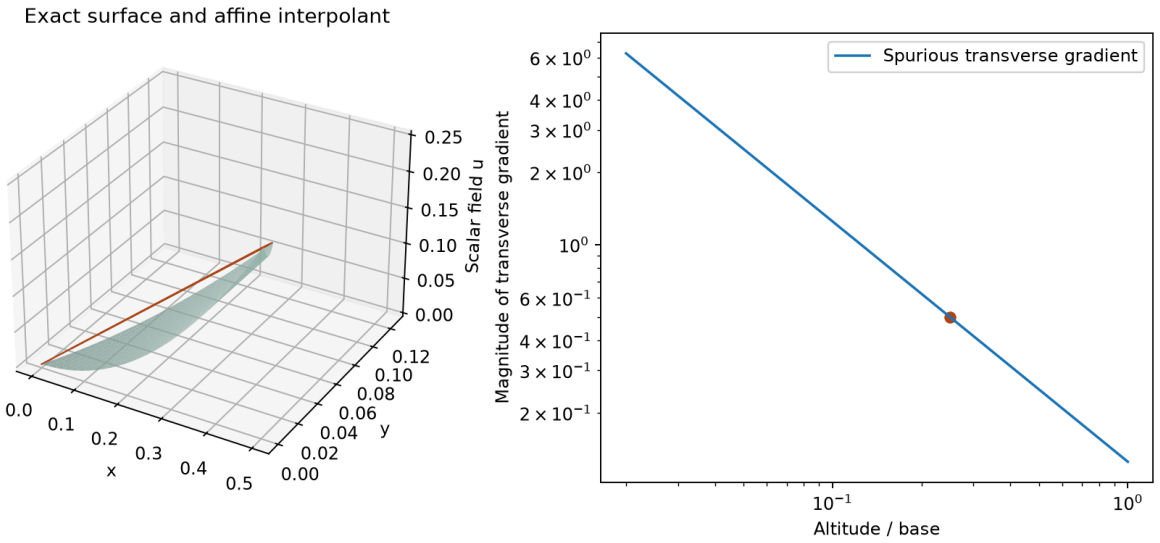
Interpolant gradient (0.5, -0.5). L² gradient-error norm 0.09547. A thin triangle creates a spurious transverse derivative even though u is independent of y.

Download the executed OneSimulation notebook · Python source

The notebook reproduces the reference case and exposes the mesh or geometry parameters. Its saved output below comes from the actual OneSimulation solver.

```
"""Inspect shape-dependent interpolation error with a OneSimulation dense solve."""
import argparse
import matplotlib.pyplot as plt
import numpy as np
```

```
from onesim.numerics.linear.matrices.dense_linear_algebra import solve_dense


def experiment(base=0.5, ratio=0.25):
    if base <= 0 or ratio <= 0:
        raise ValueError("Base and altitude/base ratio must be positive")
    altitude = base * ratio
    vertices = np.array([[0., 0.], [base, 0.], [base/2, altitude]])
    # The affine interpolation coefficients reproduce the three nodal values.
    vandermonde = np.column_stack((np.ones(3), vertices))
    coefficients = solve_dense(vandermonde, vertices[:, 0]**2)
    expected = np.array([0., base, -base**2/(4*altitude)])
    np.testing.assert_allclose(coefficients, expected, atol=2e-14)
    gradient_error = np.sqrt(base*altitude/2 * (base**2/6 + base**4/(16*altitude**2)))
    print(f"Affine coefficients: {coefficients}")
    print(f"L2 norm of gradient error: {gradient_error:.8f}")

    barycentric = np.array([(i/30, j/30) for i in range(31) for j in range(31-i)])
    points = np.column_stack((base*(barycentric[:, 0]+barycentric[:, 1]/2), altitude*barycentric[:,
1]))
    exact = points[:, 0]**2
    linear = coefficients[0] + points @ coefficients[1:]
    fig = plt.figure(figsize=(10, 4.4), layout="constrained")
    ax = fig.add_subplot(121, projection="3d")
    ax.plot_trisurf(points[:, 0], points[:, 1], exact, color="#a6c8be", alpha=.8)
    for fraction in np.linspace(0, 1, 9):
        y = altitude*fraction
        x = np.linspace(base*fraction/2, base*(1-fraction/2), 40)
        ax.plot(x, np.full_like(x, y), coefficients[0]+coefficients[1]*x+coefficients[2]*y,
                color="#ad491d", lw=1)
    ax.set(xlabel="x", ylabel="y", zlabel="Scalar field u", title="Exact surface and affine
interpolant")
    error = fig.add_subplot(122)
    ratios = np.geomspace(.02, 1, 100)
    error.loglog(ratios, base/(4*ratios), label="Spurious transverse gradient")
    error.scatter(ratio, abs(coefficients[2]), color="#ad491d")
    error.set(xlabel="Altitude / base", ylabel="Magnitude of transverse gradient")
    error.legend()
    fig.savefig("complete-numerics-triangle.png", dpi=170)
    return coefficients, gradient_error


if __name__ == "__main__":
    parser = argparse.ArgumentParser(description=__doc__)
    parser.add_argument("--base", type=float, default=.5)
    parser.add_argument("--ratio", type=float, default=.25)
    args, _ = parser.parse_known_args()
    experiment(args.base, args.ratio)
    plt.show()


Affine coefficients: [ 0.   0.5 -0.5]
L2 norm of gradient error: 0.09547033
```



The exact quadratic scalar field and its affine nodal interpolant, whose coefficients are computed with the OneSimulation dense solver.

Spatial discretisation and field equations · 8. Which waves are allowed to enter a numerical boundary?

Diagonalise linear acoustics, impose one physical condition per incoming wave and compare diffusive and dispersive transport schemes on the same spatial Fourier mode.

Separate acoustic waves by their directions

For small perturbations in a uniform fluid, $p_t + \rho c^2 v_x = 0,\ v_t + p_x/\rho = 0$ with density $\rho > 0$, speed $c > 0$. Add and subtract ρc times the second equation. The combinations $w_\pm = p \pm \rho c v$ satisfy $\partial_t w_\pm \pm c\,\partial_x w_\pm = 0$. Reconstruct $p = (w_+ + w_-)/2,\ v = (w_+ - w_-)/(2\rho c)$. At the left boundary w_+ enters and w_- leaves. These roles exchange at the right boundary.

For a rigid wall $v = 0$, set the incoming characteristic equal to the outgoing one. The reflected pressure has the same sign. At a pressure-release boundary $p = 0$, it has opposite sign. Prescribing p and v independently at the same boundary generally overspecifies a first-order system. For a diagonalizable hyperbolic matrix $A = R\Lambda R^{-1}$ the same counting follows from $R^{-1}u$, subject to independent boundary conditions on incoming components. Zero-speed or glancing modes require separate analysis.

Derive one amplitude and phase experiment for three methods

For scalar right-going $u_t + a u_x = 0$, set $\nu = a h/\Delta x$. First-order upwind uses the upstream difference. Lax–Friedrichs replaces the central value by the average of its neighbours before taking the centred first derivative. Lax–Wendroff expands time to second order and uses $u_{tt} = a^2 u_{xx}$. With Fourier angle $\theta = k\Delta x$, the three amplification factors are

$$G_U = 1 - \nu(1 - e^{-i\theta}),$$

$$G_F = \cos\theta - i\nu \sin\theta,$$

G_W=1-i\nu\sin\theta+\nu^2(\cos\theta-1).

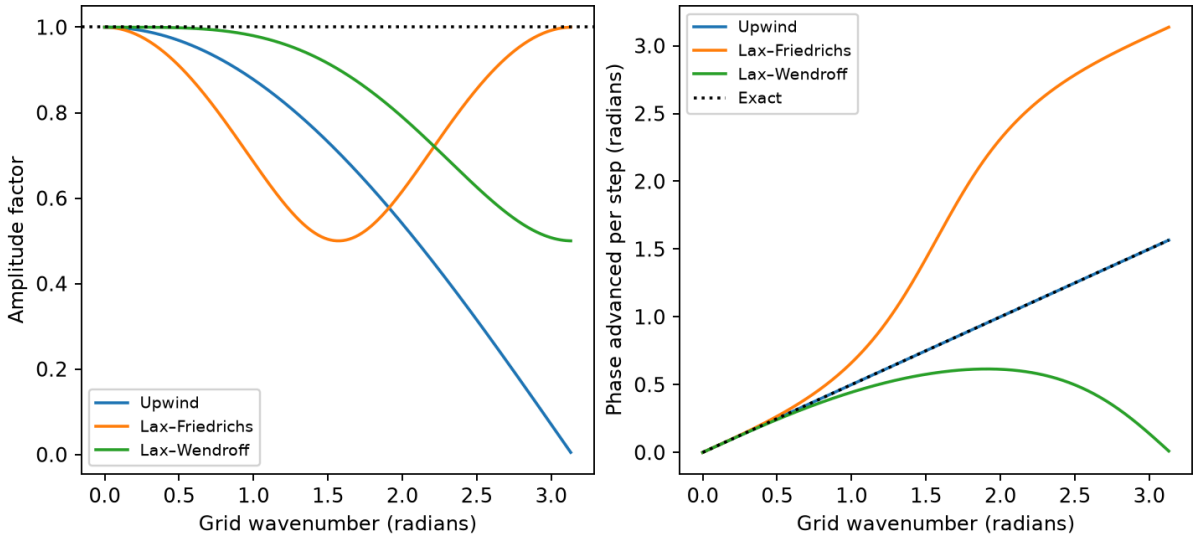
For $0 \le \nu \le 1$ these are stable in the periodic discrete L^2 norm. For the last factor direct expansion gives $|G_W|^2 = 1 - 4\nu^2(1-\nu^2)\sin^4(\theta/2)$. The exact factor $e^{-i\nu\theta}$ has modulus one. At $\nu = 1/2, \theta = \pi/2$, upwind gives $0.5 - 0.5i$, Lax–Friedrichs gives $-0.5i$ and Lax–Wendroff gives $0.75 - 0.5i$. Their magnitudes and arguments distinguish loss of contrast from arrival-time error.

Transfer the calculation to both members of the system

On a uniform grid update w_+ by a right-going scheme and w_- by its mirrored left-going counterpart. For first-order upwind the outgoing stencil at either wall uses only interior data. Supply the incoming value using the physical relation. For centred and second-order stencils, ghost values or one-sided closures must also approximate derivatives at the boundary with a compatible accuracy and stability. A periodic amplification test alone cannot certify them.

Why can a second-order scheme create negative overshoots?

In Lax–Wendroff the downstream coefficient $(\nu^2 - \nu)/2$ is negative for $0 < \nu < 1$. At a step discontinuity the update can leave the interval spanned by neighbouring data. Its Fourier stability bound controls a norm and does not imply a maximum principle.



Perfect propagation has unit amplitude and phase advance νθ. Near the shortest resolved wavelengths, amplitude loss and phase error select different methods even at identical Courant number.

Tutorial on waves in an acoustic tube

Calculate one update next to a marked step

▼ Work through the solution

Take $\nu = 1/2$ and $u_{j-1} = u_j = 0,\ u_{j+1} = 1$. Upwind gives zero, Lax–Friedrichs gives $1/4$, Lax–Wendroff gives $-1/8$. The exact right-going step has not reached this upstream cell centre.

Infer the reflected wave from a non-ideal termination

▼ Work through the solution

At the right boundary impose $p = Zv$. Substituting the characteristic pair gives $w_- = (Z - \rho c)w_+/(Z + \rho c)$. A matched termination $Z = \rho c$ transmits the wave with zero reflection. Verify the rigid and pressure-release limits before using this relation in ghost cells.

LeVeque · advection equations and hyperbolic systems

Change parameters · Which waves are allowed to enter a numerical boundary?

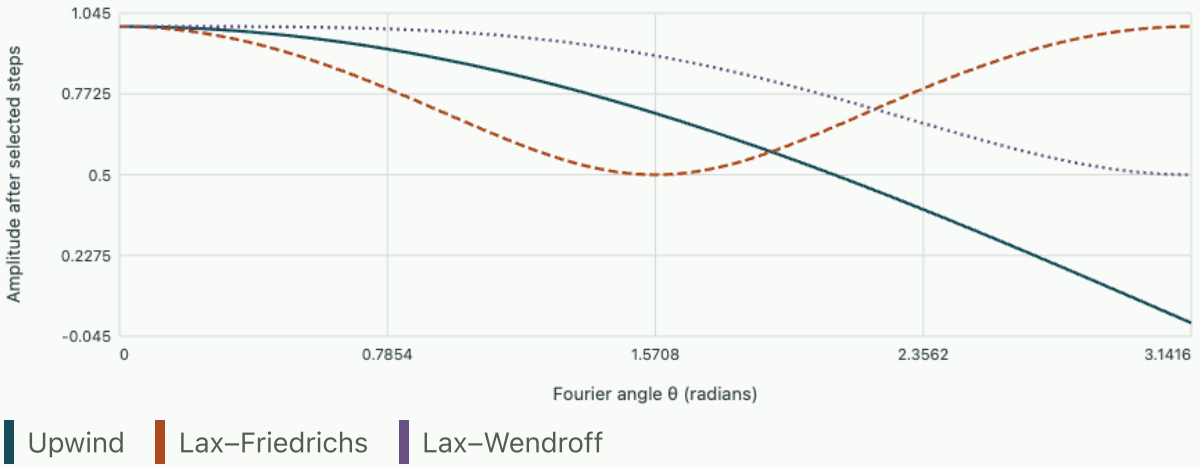
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Courant number ν **0.5**

Propagation steps **1**



Exact transport retains unit amplitude at every θ. All chosen Courant numbers lie in the stability range; grid-scale damping differs substantially between methods.

Boundary data, uniqueness and singular sources

Derive the continuous problem before assembling a discrete row. Test what happens to uniqueness, source flux and interface voltage under a change of boundary data.

1. Vector calculus through flux, circulation and singular sources
2. When do boundary measurements determine a unique electrostatic field?
3. Green functions, the Poisson kernel and what a small electrode can influence
4. Rectangular geometries, orthogonality and sources that drive individual modes
5. Cylindrical geometry, coaxial boundaries and Bessel spectra
6. A converged field solver can still give the wrong voltage
7. Measure the capacitance of a continuously graded dielectric
8. Is a computed screening layer resolved?
9. What can a mesh measure about a point charge?

Vector calculus through flux, circulation and singular sources

Electrostatic field theory and measurement

A closed-surface flux and a closed-path circulation answer different experimental questions. Starting from small physical collecting surfaces shows why the associated differential operators look different in Cartesian, cylindrical and spherical coordinates.

How a short displacement detects a derivative

A voltmeter measures a difference between endpoints. If a small displacement $d\boldsymbol{\ell}$ changes a smooth potential by dV, the gradient is the unique vector satisfying $dV = \nabla V \cdot d\boldsymbol{\ell}$ for every displacement. Its component along an arbitrary unit vector is the slope in that direction. Its largest rate of increase occurs when the displacement is parallel to the gradient. A unit step in a coordinate label does not necessarily mean a unit physical displacement. This distinction is the source of the metric factors in curvilinear coordinates.

At position (s, ϕ, z) in cylindrical coordinates, the three physical components of displacement are ds, $s\,d\phi$, dz. Equating the general dot product to $\partial_s V\,ds + \partial_\phi V\,d\phi + \partial_z V\,dz$ gives

$$\nabla V = \hat{\mathbf{s}}\partial_s V + \hat{\boldsymbol{\phi}} s^{-1}\partial_\phi V + \hat{\mathbf{z}}\partial_z V.$$

The azimuthal factor converts a voltage change per radian into a voltage change per metre around a circle. For $V = V_0(s/a)\cos\phi$ the radial and azimuthal components of $-\nabla V$ are $-V_0\cos\phi/a$ and $+V_0\sin\phi/a$. Together they equal the constant Cartesian field $-V_0\hat{\mathbf{x}}/a$, since $s\cos\phi = x$. The apparent angular variation comes only from the rotating local unit vectors. Leaving out $1/s$ would predict a nonuniform field and components with incompatible dimensions.

Count what leaves the sides of a shrinking cell

Consider a vector $\mathbf{F}$, perhaps electric displacement measured in C/m². Define its divergence at a smooth point as the limiting outward flux per cell volume. In a Cartesian box $\Delta x\Delta y\Delta z$, the opposing x faces give $[F_x(x+\Delta x) - F_x(x)]\Delta y\Delta z$. Dividing by volume and adding the other pairs gives $\partial_x F_x + \partial_y F_y + \partial_z F_z$. The signed flux measures the local excess of outflow over inflow. A field can be large and uniform while its divergence is zero.

Repeat the count in a thin cylindrical wedge between s and $s + ds$, angular width $d\phi$, height dz. The outer curved face has larger area than the inner one. Their net flux is $\partial_s(sF_s)ds\,d\phi\,dz$. The two cuts at fixed angle each have physical area $ds\,dz$ and differ by $\partial_\phi F_\phi d\phi$. The axial faces contribute $\partial_z F_z\,s\,ds\,d\phi\,dz$. Division by the wedge volume gives

$$\nabla\cdot\mathbf{F} = \frac{1}{s}\partial_s(sF_s) + \frac{1}{s}\partial_\phi F_\phi + \partial_z F_z.$$

A radial flow $F_s = A/s$ has zero divergence at positive s because the same flux crosses every cylindrical side of fixed height. A constant radial magnitude instead has divergence A/s: larger cylinders carry greater total flux and therefore enclose additional source. Specifying that a magnitude is constant says nothing about the divergence of its direction field.

In a spherical sector the radial areas grow as r^2. For a spherically symmetric radial field $F_r(r)$, the same count gives $r^{-2}d(r^2F_r)/dr$. A sphere of radius r then has flux $4\pi r^2 F_r(r)$. Choosing $F_r = Ar$ produces $3A$ everywhere instead of A; the two missing contributions in a one-dimensional derivative account for growth of transverse area. For a uniform charge density ρ, setting $3A = \rho/\epsilon_0$ immediately reconstructs the interior electrostatic field of a sphere.

Inspect a small loop before asserting that a potential exists

The component of curl normal to a surface is the limiting circulation of $\mathbf{F}$ around an oriented small loop divided by its area. In an xy rectangle the upper side is traversed oppositely to the lower one. Opposite x-directed sides leave $-\partial_y F_x$; opposite y-directed sides leave $+\partial_x F_y$. Thus

$$(\nabla \times \mathbf{F})_z = \partial_x F_y - \partial_y F_x.$$

For a planar polar loop the curved segments have different lengths, giving $(\nabla \times \mathbf{F})_z = s^{-1}[\partial_s(sF_\phi) - \partial_\phi F_s]$. Curl and divergence use different faces and signs even when their cylindrical formulas share a factor $1/s$.

Join many neighboring loops. Circulations on common edges cancel in pairs, leaving only the outer boundary; the summed area integrals of the curl give Stokes's theorem. A similar cancellation of internal face fluxes gives the divergence theorem. These are physical bookkeeping identities, and they motivate discretisations that attach oriented fluxes to shared faces and voltages to shared edges. Assigning two independent approximations to one shared face can spoil exact cancellation in a computed conservation budget.

Take $\mathbf{F} = A\hat{\boldsymbol{\phi}}/s$ outside a thin excluded cylinder. Its ordinary curl is zero there. Any circle about the cylinder has circulation $2\pi A$. A disk spanning this circle would pass through the excluded region, so a surface lying entirely inside the stated domain cannot supply the usual zero-curl conclusion. Locally $\mathbf{F} = \nabla(A\phi)$ on any small patch avoiding a branch cut, while the putative potential changes by $2\pi A$ after a full turn. This example explains the role of domain topology. A genuine static electrostatic source in ordinary free space has a globally consistent potential; a changing flux through an excluded region can instead generate precisely such an exterior circulation.

Measure curvature through neighboring voltages

Taking the divergence of a gradient defines $\nabla^2 V$. A Taylor expansion at a point gives

$$[V(x+h) - 2V(x) + V(x-h)]/h^2 = \partial_x^2 V + O(h^2).$$

Adding the three Cartesian expressions relates the six neighbouring voltage readings to the local charge density. Equivalently, averaging a potential over directions at radius h removes all terms odd in the displacement. The angular mean of $n_i n_j$ is $\delta_{ij}/3$, so for smooth V

$$\langle V(\mathbf{x}+h\mathbf{n})\rangle_\Omega = V(\mathbf{x}) + \frac{h^2}{6}\nabla^2 V(\mathbf{x}) + O(h^4).$$

In a region with positive charge, Poisson's equation makes the spherical average smaller than the centre value. An uncharged saddle can have strong curvature along each of two axes that cancels in the sum. A local maximum along a single measurement line therefore does not prove that an interior source exists. To make such an inference one needs curvature in every independent spatial direction or independent symmetry information.

Near a boundary or dielectric interface this smooth Taylor analysis may break down. Enforce the integral flux balance on a cell straddling the interface instead. In media of different permittivity the normal derivative of V can jump even where there is no free charge. The physical differential operator is then $\nabla\cdot(\epsilon\nabla V)$; simply inserting a pointwise averaged permittivity in front of a smooth Laplacian may erase the correct jump condition.

Use a test volume to give singular expressions meaning

Outside the origin the divergence of $\hat{\mathbf{r}}/r^2$ is zero, yet every sphere centred there has flux 4π. To characterize its source measure the field against a smooth test function f of compact support. Excise a ball of radius η, integrate by parts in the remaining exterior and note that its inward-facing inner boundary normal is $-\hat{\mathbf{r}}$. The zero bulk divergence eliminates the interior term, leaving $-\int(\hat{\mathbf{r}}/r^2)\cdot\nabla f\,dV \to 4\pi f(0)$ as the ball shrinks. This is the defining action of the three-dimensional delta distribution:

$$\nabla\cdot(\hat{\mathbf{r}}/r^2) = 4\pi\delta^3(\mathbf{r}),$$
$$\nabla^2(1/r) = -4\pi\delta^3(\mathbf{r}).$$

The factors and signs follow from an ordinary finite flux calculation. A delta distribution has no finite peak height to compare with a mesh value. A convergent field computation should reproduce its integral and its action on smooth functions. In spherical coordinates a radially concentrated charge can be represented with a radial delta only when the $4\pi r^2$ integration weight and the convention for the endpoint are treated consistently. It is often safer to specify the enclosed charge directly.

For a finite regularisation take $V_a = q/(4\pi\epsilon_0\sqrt{r^2+a^2})$. Differentiation gives

$$E_r = qr/[4\pi\epsilon_0(r^2+a^2)^{3/2}].$$

A spherical flux measurement encloses $Q/q = r^3/(r^2+a^2)^{3/2}$. Its derivative divided by $4\pi r^2$ gives

$$\rho_a = 3qa^2/[4\pi(r^2+a^2)^{5/2}].$$

For every $a > 0$ the density is smooth; its peak scales as a^{-3}, preserving a finite total charge. For fixed $r > 0$, $Q/q \to 1$ as $a \to 0$. Shrinking the collecting radius and the source core simultaneously at fixed r/a instead retains a nontrivial fraction. The resolution of the probe is part of the limiting experiment.

A sheet of dipoles produces a voltage jump without a net sheet charge

Replace an ideal neutral polarised layer by two parallel charge sheets, one at $z = -d/2$ with density $-\sigma$, another at $z = +d/2$ with density $+\sigma$. Assume both planes extend far beyond the detectors so their fields can be treated as infinite-plane fields. The field outside cancels. Between them it points from positive to negative,

$$E_z = -\sigma/\epsilon_0.$$

A voltmeter moved from below to above the pair rises by $+\sigma d/\epsilon_0$. The result can stay finite when d shrinks provided that the dipole moment per area $p_s = \sigma d$ is fixed. The fields away from the layer then converge to zero on both sides while the voltage difference persists. This model is useful for a surface dipole layer, for example when positive and negative charge distributions are displaced by molecular-scale distances.

It is essential not to infer the voltage jump from the net charge enclosed in a pillbox, since the net is zero for every d. Write the signed density as $\rho_d = \sigma[\delta(z - d/2) - \delta(z + d/2)]$. For any smooth test voltage $f(z)$, integration gives $\sigma[f(d/2) - f(-d/2)] \to p_s f'(0)$. The distribution with action $f \mapsto p_s f'(0)$ is $-p_s \delta'(z)$. In one dimension Poisson's equation is $-\epsilon_0 V'' = \rho$, so a solution is

$$V = (p_s/\epsilon_0) H(z) + \text{constant},$$

where H is a step. Its derivative is a delta, its second derivative a delta derivative. Reversing the layer orientation reverses both the jump and the sign of this source distribution.

A conventional monopole sheet instead has $\rho = \sigma_f \delta(z)$. Its potential is continuous and its normal derivative jumps. The two types of ideal source thus leave different signatures in electrode boundary models. For a voltage jump of 20 mV, the required $p_s = \epsilon_0 (0.020) = 1.77 \times 10^{-13}$ C/m in vacuum. Spreading it across 1.0 nm requires $\sigma = 1.77 \times 10^{-4}$ C/m² and an internal field of 20 MV/m in that thin-layer model. The large internal field is consistent with a small voltage step because it acts only over one nanometre. At fixed p_s the field energy per area grows as $p_s^2/(2\epsilon_0 d)$. Thus taking a perfect zero-thickness dipole layer gives useful exterior fields and voltage jumps but not a finite internal energy. A physical model of finite layer thickness is necessary for that observable.

Moving a numerical grid across the sheet illustrates how to recognise the two cases. If its potential samples resolve a continuous cusp, integrating the inferred charge across one symmetric layer gives finite sheet charge. If instead the voltage contains a step, two neighbouring density bins can acquire large opposite entries whose sum tends to zero. Their first moment stays finite. A useful verification uses both $\int \rho \, dz$ and $\int z\rho \, dz$ and compares potential readings on each side. A code that preserves

charge but smooths a dipole step into an incorrect zero voltage would pass the first test but fail the others.

An explicit flux audit on a nonuniform mesh

To separate geometrical signs from a particular physical equation, take a rectangular box $0 < x < L_x,\ 0 < y < L_y,\ 0 < z < L_z$ and an exactly known displacement $\mathbf{D} = Axy\,\hat{\mathbf{x}} + By^2\,\hat{\mathbf{y}} + Cz\,\hat{\mathbf{z}}$. The constants have units making each term C/m². Its divergence is $(A + 2B)y + C$. Integrating over the volume gives total free charge $L_x L_z[(A + 2B)L_y^2/2 + CL_y]$. At $x = 0, y = 0, z = 0$ the relevant normal components vanish. At $x = L_x$, integrating over the face gives $AL_xL_zL_y^2/2$. The $y = L_y$ face gives $BL_xL_zL_y^2$, and the top gives $CL_xL_yL_z$. Addition reproduces the volume calculation, including its factors of two.

Choose numerical constants by specifying $L_x = L_y = L_z = 1$ m, $A = 2$ nC/m⁴, $B = 3$ nC/m⁴ and $C = 4$ nC/m³. The three positive-face contributions are 1, 3 and 4 nC, for total 8 nC. Split the volume at $y = 0.30$ m. The bottom subvolume has volume charge $\int_0^{0.3}(8y + 4)\,dy = 1.56$ nC. Its new upper face carries $B(0.30)^2 = 0.27$ nC. The bottom section of the $x = L_x$ face carries 0.090 nC and that of the top carries 1.20 nC. Their sum is 1.56 nC. For the upper subvolume the same internal face must be counted with the opposite sign. Physical flux is single-valued on that shared face even if the cells on its sides have very different thicknesses.

Suppose a program evaluates the shared face once using a derivative from the lower cell and independently using another from the upper cell. A mismatch of 0.03 nC in those estimates generates a global charge defect when summing cells, although each side may display a locally plausible gradient. A conservative scheme stores one oriented face flux and uses the same number with opposite signs in both cells. Flux conservation can then hold to floating-point accuracy on any partition, without guaranteeing the shared flux is physically accurate. In our manufactured example a shared wrong value would move the two cell errors in opposite directions and leave the total exact. Compare the two individual subvolume charges as well as the global one.

Is this manufactured $\mathbf{D}$ an electrostatic vacuum solution? In a homogeneous scalar dielectric, $\nabla \times \mathbf{D}$ would have to vanish for a purely electrostatic field. Its z component here is $-Ax$. Thus the displacement tests the divergence theorem and a face-flux data structure perfectly, but it does not simultaneously test a potential-gradient solver with constant permittivity. For the latter choose $\mathbf{D} = -\epsilon\nabla V$ from an explicit potential before manufacturing the charge. This example separates validation of a mathematical operation from validation of a whole physical model, an important distinction when a beautifully closed flux budget is the only available diagnostic.

Reconstruct a field from its divergence and curl in a periodic measurement cell

Suppose a three-dimensional field map is tabulated on a periodic cube and you know two derived maps: its divergence d and its curl $\mathbf{w}$. Can these determine the original field? Fourier analysis gives a practical construction and reveals the missing information. For a nonzero spatial wavevector $\mathbf{k}$, use the

convention $\mathbf{F}(\mathbf{x}) = \sum_{\mathbf{k}} \mathbf{F}_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}}$. Then $d_{\mathbf{k}} = i\mathbf{k} \cdot \mathbf{F}_{\mathbf{k}}$ and $\mathbf{w}_{\mathbf{k}} = i\mathbf{k} \times \mathbf{F}_{\mathbf{k}}$. Decompose $\mathbf{F}_{\mathbf{k}}$ into its component parallel to $\mathbf{k}$ and the remaining transverse component. Elementary vector identities give

$$\mathbf{F}_{\mathbf{k}} = -\frac{i\mathbf{k}}{k^2} d_{\mathbf{k}} + \frac{i}{k^2} \mathbf{k} \times \mathbf{w}_{\mathbf{k}} \qquad (\mathbf{k} \neq 0).$$

The first term supplies the requested divergence and zero curl. The second supplies the requested curl when $\mathbf{k} \cdot \mathbf{w}_{\mathbf{k}} = 0$, a compatibility requirement because the divergence of a curl vanishes. At $\mathbf{k} = 0$ both operations annihilate the constant field. Its mean must be measured independently. In a nonperiodic region suitable boundary data enter as well. This is a constructive version of the statement that divergence and curl determine a field only after specifying its domain, boundary information and any harmonic freedom.

For a concrete field take $\mathbf{F} = (A \cos kx, B \sin kx, C)$. Its divergence is $-kA \sin kx$, its only curl component is $w_z = kB \cos kx$, and the constant offset $C\hat{\mathbf{z}}$ drops out of both maps. The first input alone reconstructs the x component and the second the y component. A purely electrostatic solver may legitimately require $B = 0$ but will not recover C from the local charge distribution alone. A uniform applied field can be imposed by suitable distant electrodes. A strictly periodic scalar potential cannot represent a nonzero mean field without separately adding a linear background.

A grid-based reconstruction needs its own consistency check. A centred difference on a periodic grid maps e^{ikx} to $i \sin(kh) e^{ikx}/h$, not exactly ike^{ikx}. A source plane wave at the highest alternating grid mode $k = \pi/h$ has zero centred derivative at every node despite alternating sign. Treating discrete divergence measurements as exact continuum coefficients near that grid scale would infer falsely that an oscillatory field has vanished. Use a consistent discrete operator, stagger quantities if appropriate, and restrict quantitative interpretation to adequately sampled modes. The continuous Helmholtz identity is exact; the data acquisition and differentiation scheme decides which components of that identity are actually observable.

Enclose different amounts of a regularised source

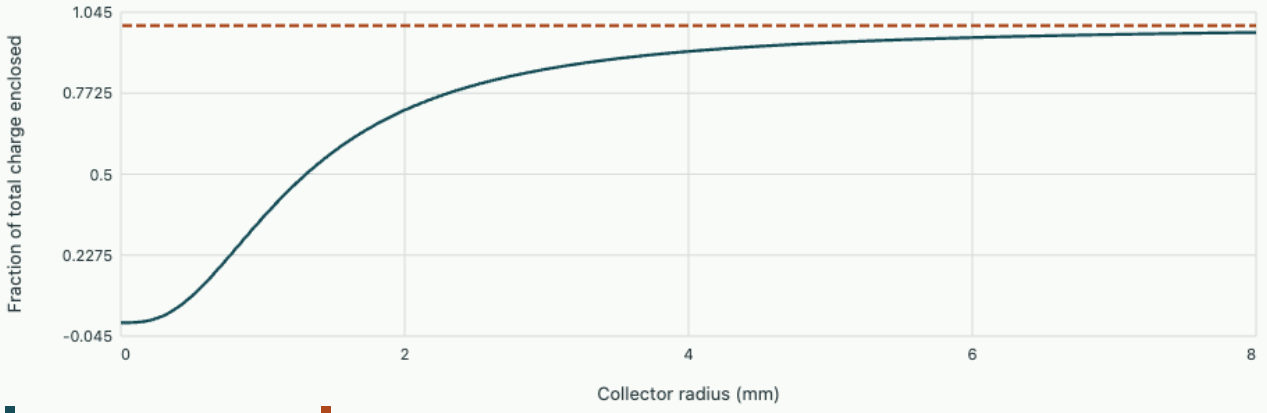
Predict how the flux of a small sphere changes as it crosses the charge cloud. Compare the enclosed fraction with the local source density.

Open this experiment online

Try the calculation

Core radius a (mm) 1

Collector radius R (mm) 2



Smooth charge core Ideal point source, R > 0

At 2 mm the collector encloses 71.6% of the charge. Doubling its radius would enclose 91.3%.

Tutorial · turn field values into reliable local measurements

Reconstruct a sheet from a voltage cusp

Two electrodes at $z = \pm L$ are grounded. Inside the gap a measured potential is $V_0(1 - |z|/L)$. Find both fields and reconstruct the source, then explain what a centred finite difference reports at the cusp.

▼ **Work through the solution**

For $z < 0$, $E_z = -V_0/L$; for $z > 0$, $E_z = +V_0/L$. The jump gives a sheet $\sigma = 2\epsilon_0 V_0/L$ at zero. A centred estimate of the first derivative there returns zero by symmetry, reflecting the arithmetic mean of the two one-sided fields. It cannot be interpreted as either physical one-sided limit. The centred second difference is $-2V_0/(Lh)$, so the inferred cell density grows as $1/h$. Multiplying by a cell width h returns the finite sheet charge. This is the correct quantity to compare when refining the grid.

With $V_0 = 2$ V and $L = 2$ mm the fields have magnitudes 1000 V/m and the sheet charge is 17.7 nC/m². Reversing V_0 reverses both field vectors and the sheet sign without moving the cusp.

Can a voltage map distinguish a saddle from a charge?

Near the origin a section shows $V(x, 0, 0) = V_c - Ax^2$, $A > 0$. Give a source-free three-dimensional extension and a second extension consistent with uniform positive charge. Propose a minimum additional measurement given no prior symmetry.

▼ **Work through the solution**

The first is $V_c - Ax^2 + Ay^2$, independent of z. Its Laplacian vanishes. The second is $V_c - A(x^2 + y^2 + z^2)$, whose Laplacian is $-6A$, giving $\rho = 6\epsilon_0 A$. Both match the entire original line. Curvatures along y, z together with the given x-curvature determine the Laplacian. A small enclosing surface measuring normal electric flux provides the same information integrated over a volume and does not require numerical second differentiation. Neither a second derivative along one line nor a visibly crowded pair of voltage contours is a universal charge detector.

Further reading

- Vary the width and collecting radius of a regularised charge
- Fließbach, chapter 3 · distributions
- Griffiths, 5th ed., chapter 1 · vector analysis and the delta function

When do boundary measurements determine a unique electrostatic field?

Electrostatic field theory and measurement

A boundary-value solver may converge while solving the wrong electrical experiment. Establishing which potentials, charges and fluxes are fixed gives a uniqueness proof and a way to turn an approximate field into an upper bound.

Specify the experiment before specifying the equation

An insulating region occupies a bounded volume Ω with smooth conducting or insulating boundaries. A known free charge density ρ_f is embedded in the region. For a positive scalar linear permittivity $\epsilon(\mathbf{x})$, electrostatics reduces to

$$-\nabla \cdot (\epsilon \nabla V) = \rho_f.$$

To turn this into an experiment one must decide how each boundary is prepared. A conductor attached to a source supplies a fixed voltage; an isolated conductor keeps a fixed total charge while becoming an equipotential; an insulating surface with known displacement flux prescribes a normal component. These preparations require different mathematical data. The same electrode shape can participate in any of them.

At an interface between two ordinary dielectrics the potential is continuous in the absence of a prescribed dipole sheet. Integrating Gauss's law across the interface gives

$$(\mathbf{D}_2 - \mathbf{D}_1) \cdot \mathbf{n} = \sigma_f,$$

where $\mathbf{n}$ points from medium 1 to 2. Thus the appropriate field problem can be posed either in a weak integral form across both materials or as two equations joined by voltage and flux conditions. The latter view is especially helpful when a measured charge appears on just one internal sheet. A jump in permittivity alone does not permit an arbitrary jump in voltage.

Prove uniqueness by testing the difference, not the solution

Suppose two solutions V_1, V_2 fit identical prescribed volume sources, material data and boundary voltages. Define $w = V_1 - V_2$. It satisfies $\nabla \cdot (\epsilon \nabla w) = 0$ and vanishes at the boundary. Multiply by w and apply the product rule to $\nabla \cdot (w \epsilon \nabla w)$. Integration yields

$$\int_\Omega \epsilon |\nabla w|^2 \, dV = \oint_{\partial\Omega} w \epsilon \, \partial_n w \, dA.$$

The boundary term is zero. The integrand on the left is nonnegative and $\epsilon > 0$, so the gradient vanishes wherever the material is regular. The function must be constant on each connected component; its fixed boundary value makes it zero. This proves that at most one field solves the stated problem. It does not

itself construct a solution. It converts a difficult task of checking every possible competitor into checking the PDE and boundary data for one candidate.

The same reasoning works for a symmetric positive-definite permittivity tensor: replace $\epsilon|\nabla w|^2$ by $(\nabla w)^T \boldsymbol{\epsilon} \nabla w$. It also applies with internal material interfaces. Integrate separately on either side. Their face terms cancel because w is continuous and the prescribed free sheet makes the jump in the difference flux zero. An unaccounted interface charge would supply precisely the term that invalidates this cancellation.

If only the normal flux is fixed everywhere, the boundary difference flux vanishes and the left integral again gives a constant w. Constants are now indistinguishable by the boundary data; the electric field remains unique. Gauss's law also imposes compatibility $\oint_{\partial\Omega} \mathbf{D} \cdot \mathbf{n}\, dA = \int_\Omega \rho_f \, dV$. Choosing flux values whose sum disagrees with the enclosed charge produces no solution, even after choosing a voltage reference. Numerically this can appear as a residual that never falls below a threshold, or as a solver modifying the source silently to make the equations compatible.

Prove uniqueness when only total conductor charge is known

Take a region surrounding several isolated conductors. The value on each metal is constant but initially unknown. For the difference of two candidate solutions denote these constants by w_i. Then each conductor's contribution to the right-hand side of the energy identity factors as $w_i \int_{S_i} \epsilon \partial_n w \, dA$. With outward normal from the dielectric into the metal this integral is the difference in the conductor's free charge. Prescribing total charge for each metal makes every such term vanish individually. Provided a reference conductor or condition at infinity removes an overall constant, the potential is unique.

Knowing only the sum of charges over several electrically disconnected metals can be insufficient. For two isolated spheres, allow one to take charge $+\delta q$ and the other $-\delta q$ while fixing their total at zero. Their equipotential values can change relative to each other, and an external field may appear. If instead they are joined by an ideal wire they must share one voltage, so the common boundary contribution factors against their total charge. The proof itself identifies the physical distinction between strings of electrically connected metal and separate conductors.

A functional whose variation is Gauss’s law

For fixed boundary voltage consider

$$\mathcal{I}[V] = \int_\Omega \left[\frac{1}{2} \nabla V \cdot \epsilon \nabla V - \rho_f V \right] dV.$$

The two meanings of V in this notation are distinguishable from context: $V(\mathbf{x})$ is potential and dV is a volume element. Add a trial perturbation η that vanishes on the fixed-voltage boundary. The first-order change is $\int [\nabla \eta \cdot \epsilon \nabla V - \rho_f \eta] dV$. Integrate the first term by parts. Requiring zero for every admissible η gives

$$-\nabla \cdot (\epsilon \nabla V) = \rho_f.$$

On a boundary portion where flux rather than voltage is prescribed, an additional linear boundary contribution accounts for the specified normal derivative. It is not generally correct to hold every trial voltage zero there.

For an exact fixed-voltage solution V_* expand $\mathcal{I}[V_* + \eta]$. The first-order terms vanish, leaving

$$\mathcal{I}[V_* + \eta] - \mathcal{I}[V_*] = \frac{1}{2}\int_\Omega \nabla\eta \cdot \epsilon \nabla\eta \, dV.$$

The exact solution minimises the functional, and the gap measures the field error in an energy-weighted norm. If $\rho_f = 0$ the functional is the stored field energy. With embedded sources the linear term matters; the value of the functional is generally not just field energy. Restrict the trial fields to a finite space and minimize there. This is the variational origin of a potential-based finite-element method. A small change in the energy between two grids can coexist with a noticeably changing local peak field because the integral weights errors by volume.

A complete numerical comparison without a black box

Take a unit area filled by two dielectric slabs each of thickness $d = 1$ mm and relative permittivities 2 and 8. Set one electrode at 10 V, the other at zero, with no free charge inside. A trial straight line through the whole gap obeys both electrode voltages. Its energy per area is $\epsilon_0(2+8)d(5000)^2/2 = 1.107\ \mu\mathrm{J/m^2}$ after unit conversion. The exact interface problem requires constant normal displacement. It gives voltage drops of 8 and 2 V, fields of 8000 and 2000 V/m and energy per area $0.708\ \mu\mathrm{J/m^2}$. A single continuous slope produced too much energy by about 56%. The trial is admissible despite its incorrect interface flux, because stationarity, rather than a kinematic constraint, enforces that flux.

Recover the exact answer as a one-variable minimisation. Let u be the voltage at the interface measured relative to the grounded end, with the first material next to the 10 V electrode. The energy per area is $\epsilon_0[2(10-u)^2 + 8u^2]/(2d)$. Differentiation gives $-2(10-u) + 8u = 0$, hence $u = 2$ V. The derivative with respect to an interior trial voltage measures the flux imbalance between neighbouring elements. This makes the algebraic finite-element equation a local charge balance. Conversely a trial that fails to equal an electrode's prescribed voltage is inadmissible, even if its energy appears lower than the computed answer.

A maximum principle as a diagnostic independent of energy

In a charge-free region with uniform permittivity the potential at the centre of a sphere equals its mean over the sphere. To see this, differentiate the angular average with respect to its radius. The result is proportional to the total normal derivative, which is zero by Gauss's law for an empty sphere. Shrink the sphere to its centre to identify the constant average. Therefore a nonconstant harmonic function cannot have a strict interior maximum or minimum. If the centre assumes the maximum value, all sufficiently nearby directions must assume that same value to keep the average equal; continuing through the connected region forces constancy.

For source-free Poisson solvers an interior voltage above all fixed outer values indicates an issue with boundary data, source terms or discretisation. But the same observation is entirely possible with positive internal charge, as the one-dimensional grounded-plane parabola demonstrates. The maximum principle concerns potentials; it does not say that a field magnitude, an individual component in an inhomogeneous medium, or the potential restricted to a line must remain between the corresponding boundary values on that line.

A thin dielectric coating turns a boundary voltage into a Robin condition

Suppose a grounded metal enclosure is lined with an insulating film of thickness d_f and permittivity ϵ_f. We want the electrostatic potential in an interior region large compared with the film thickness. Simply declaring the film's exposed face grounded neglects its voltage drop. If the tangential scale of the interior field and the radius of curvature of the surface are both much larger than d_f, the film behaves locally like a plane capacitor with capacitance per area $c_f = \epsilon_f / d_f$. Let $\mathbf{n}$ point from the interior into the film, let the backing metal be at V_b, and denote the interior-side surface potential by v. Without free charge at the contact, the outward displacement from the interior equals $c_f(v - V_b)$. Since $\mathbf{D} = -\epsilon \nabla V$ inside, the effective boundary condition is

$$\epsilon \, \partial_n V + c_f (V - V_b) = 0.$$

This is a Robin boundary condition: both the potential and its derivative enter. The coefficient c_f has units F/m², so both terms are surface charges. For $c_f \to \infty$ a bounded normal field forces V to the backing voltage. For $c_f \to 0$ a finite voltage difference can sustain almost no normal flux. The two familiar boundary ideals are limiting cases of one measurable material preparation. The derivation requires the film to contain no embedded free charge and to respond linearly; measured leakage over long times adds a current-balance problem.

The variational proof above extends naturally. Add $\frac{1}{2} \int_{S_f} c_f (V - V_b)^2 dA$ to the interior functional. Its first variation at the exposed face gives $\epsilon \partial_n V + c_f (V - V_b)$ times the arbitrary boundary variation. The added quadratic term is exactly the electrostatic energy stored in the eliminated thin layer under the one-dimensional approximation. For the difference of two solutions the energy identity becomes

$$\int \epsilon |\nabla w|^2 dV + \int c_f w^2 dA = 0.$$

Either a prescribed reference voltage or a positive-area Robin patch with $c_f > 0$ removes the constant null mode. This follows without pretending that the exposed potential itself was prescribed.

Consider a planar interior region of thickness 2.0 mm filled with $\epsilon_i = 4\epsilon_0$, separated from its 0 V backing by a 50 µm film with $\epsilon_f = 2\epsilon_0$. Set the opposite electrode to 10 V. The exact one-dimensional displacement is

$$D = 10 / [L/\epsilon_i + d_f/\epsilon_f].$$

The effective electric length in vacuum units is $0.50 + 0.025 = 0.525$ mm, hence $D = 169$ nC/m² and the film takes $10(0.025/0.525) = 0.476$ V. Imposing a direct 0 V boundary at the exposed film face would overpredict displacement by 5%. Solving with the Robin condition instead reproduces the series result exactly in the flat one-dimensional model. For a patterned electrode with wavelength comparable to film thickness, tangential field in the film cannot be ignored and the effective boundary response depends on spatial frequency.

What a voltage error bound does and does not control

If two source-free solutions have slightly different Dirichlet data, their difference is harmonic, not necessarily zero. Let their boundary discrepancy have maximum magnitude δ. The maximum principle applied to the difference and its negative guarantees $|V_1 - V_2| \le \delta$ throughout the region. This gives a stability estimate for voltage even where the exact solution is unknown. It requires an error bound on the whole boundary and identical sources inside. Collocation residuals evaluated only at a few points do not establish the premise between those points.

Does the same number bound the error in local electric field? Consider the upper half-space with boundary error $\delta\cos(kx)$. The harmonic difference at height h is $w = \delta e^{-kh}\cos(kx)$. Immediately at the boundary the field-error amplitude is $k\delta$, which can be arbitrarily large as the spatial wavelength shrinks even though the voltage error stays below δ. At nonzero h the field-error amplitude for this single mode is $k\delta e^{-kh}$. For fixed h it is maximal at $k = 1/h$, with value $\delta/(eh)$. Depth regularises the sensitivity to ever finer ripples. A global bound on arbitrary superpositions requires accounting for how their amplitudes combine; maximising one sinusoid does not prove that every boundary function attains the same bound.

Suppose a nominally grounded instrument panel has a sinusoidal ripple of peak voltage 5 mV and wavelength 1.0 mm. At the surface it creates an electric-field amplitude of 31.4 V/m. Just 1.0 mm away its amplitude is multiplied by $e^{-2\pi} = 0.00187$, leaving 0.0587 V/m. A sensor directly at the surface and one millimetre inside may therefore make very different accuracy demands on the boundary model, despite the same nominal voltage tolerance. A field solver claiming agreement of all electrode voltages to 5 mV still needs a spatial smoothness or resolution statement before claiming local field accuracy.

A useful related identity controls energy error when an approximate potential has an interior residual. Let V_h satisfy the exact boundary values, and define $r = \rho + \nabla\cdot(\epsilon\nabla V_h)$. With $e = V_* - V_h$ zero on the boundary,

$$-\nabla\cdot(\epsilon\nabla e) = r.$$

Multiplying by e gives $\int \epsilon|\nabla e|^2 dV = \int re\, dV$ for a smooth admissible approximation. If a piecewise field has jumps at internal element faces their residual contributions must be included. A tiny volume residual computed inside each piece can coexist with incorrect flux matching at every face. In the two-layer example, a straight-line trial has zero second derivative in both layers, yet its displacement mismatch at the interface is exactly what forces a different physical solution.

Why a conductor becomes equipotential when charge can move

So far conductor voltages or charges entered as boundary conditions. A related variational argument explains the equipotential condition itself. Imagine charges on one connected conductor that may redistribute while their total stays fixed. Treat the final material state as ideal and static, with no ongoing currents or dissipative constraints. The change in electrostatic field energy for an infinitesimal change in free surface charge is $\delta U = \int_S V(\mathbf{x})\delta\sigma(\mathbf{x})\, dA$. This follows either by slowly bringing in each incremental charge at its current potential or by varying the symmetric double integral $\frac{1}{2}\int\int \rho(\mathbf{x})G(\mathbf{x},\mathbf{y})\rho(\mathbf{y})d^3x\, d^3y$. The half in the energy disappears upon variation because either source factor may vary.

Move a small positive charge dq from a patch at potential V_A to another at V_B, preserving total conductor charge. The change is $(V_B - V_A)dq$. If the two voltages differ, one direction of transfer lowers the energy. At equilibrium every such redistribution must have zero first-order energy change, hence the potential is constant over the connected conductor. For disconnected conductors whose individual totals are fixed the same reasoning gives a different allowed constant on each. If an ideal wire joins them, transfers across the wire become admissible and their two constants equalise. This is a statement about electrical connectivity and constraints, not merely the smoothness or shape of the conducting material.

The second-order change for a neutral redistribution is its own nonnegative field energy. For a fixed-charge isolated conductor it follows that the electrostatic equilibrium distribution minimises energy over all allowed placements of the same surface charge. This is a different trial principle from prescribing electrode voltages and varying a potential throughout the space. A trial charge that obeys a total-charge constraint need not produce an exactly equipotential conductor; its excess energy measures a failure of equilibrium. Confusing these two variational principles can switch the direction of an inferred capacitance bound.

As a measurable limit, take two small spheres far apart, with approximate capacitances $C_1 = 4\pi\epsilon_0 a_1$ and $C_2 = 4\pi\epsilon_0 a_2$ to infinity. Temporarily neglect mutual coupling. With total charge Q, write $U(q) = q^2/(2C_1) + (Q-q)^2/(2C_2)$. Differentiating gives $q/C_1 = (Q-q)/C_2$, the equipotential solution after wiring. For radii 10 and 20 mm and $Q = 30$ pC this gives shares 10 and 20 pC at a common voltage about 9.0 V. Putting all the charge initially on the smaller sphere would give energy $Q^2/(2C_1)$. After connection the energy is $Q^2/[2(C_1 + C_2)]$, one third as large. The missing field energy leaves during transient currents through radiation or dissipation; static uniqueness alone describes neither time history nor dissipation partition. As sphere spacing is reduced, mutual capacitance must be included and the simple radius ratio becomes an approximation to test rather than an exact rule.

Check a potential independently of its residual

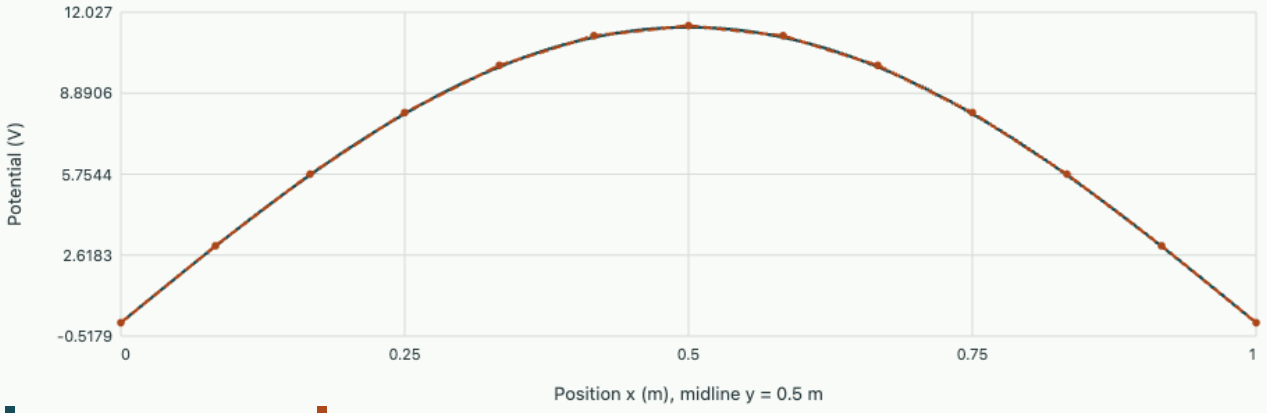
Refine a grounded square with known boundary data. Compare centre voltage with the series solution as well as the discrete residual. The accompanying OneSimulation model solves the same field problem.

Open this experiment online

Try the calculation

Cells per side **12**

Charge amplitude ρ_0 (nC/m³) **2**



Continuous solution Five-point grid solution

Centre potential 11.50889 V on the grid, 11.44331 V in the continuum. Nodal error 0.06558 V.

Download the OneSimulation model

In your OneSimulation environment, run the model with these inputs.

```
python grounded-square.py --cells 12 --charge 2
```

Tutorial · can the proposed computation represent this experiment?

Find a compatible flux and determine the remaining freedom

A slab $0 < x < L$ contains uniform positive free charge ρ. At $x = 0$ the outward displacement is measured as $D_n = -D_0$. Determine what flux must be prescribed at L, the electric field throughout, and one piece of information needed for the absolute potential.

▼ Work through the solution

Positive x displacement at the left is D_0 because that face's outward normal points towards negative x. Integrating $dD_x/dx = \rho$ gives

$$D_x = D_0 + \rho x.$$

Therefore the outward flux at L is $D_0 + \rho L$. Their sum per unit area equals ρL. In constant permittivity $E_x = (D_0 + \rho x)/\epsilon$ and $V(x) = V(0) - (D_0 x + \rho x^2/2)/\epsilon$. A reference voltage fixes $V(0)$; none can

fix an incompatible pair of flux measurements. A negative right-hand outward reading with very large positive enclosed charge would contradict Gauss's law before one even solves for V.

Bound a capacitance with an intentionally poor trial

In a charge-free parallel-plate vacuum gap of thickness L, use $V(x) = V_0(x/L)^2$ to bound the capacitance per area. Compare with the exact linear voltage. How does improving the potential shape turn into a monotonic bound?

▼ **Work through the solution**

The trial satisfies the endpoints. It gives energy per area $\epsilon_0 \int_0^L (2V_0 x/L^2)^2 dx/2 = 2\epsilon_0 V_0^2/(3L)$, hence a capacitance estimate $4\epsilon_0/(3L)$. The exact value is ϵ_0/L, lower by the predicted variational inequality. Expanding a trial space to include both the old functions and new ones cannot raise its optimal energy because the old minimiser remains available. This monotonic property requires nested trial spaces and consistent integration; changing quadrature or replacing the physical geometry changes the comparison.

Further reading

- Perturb a potential and measure the quadratic energy penalty
- Recover an interface charge using independent field measurements
- Griffiths, 5th ed., §§3.1.5–3.1.6 · uniqueness theorems

Green functions, the Poisson kernel and what a small electrode can influence

Electrostatic field theory and measurement

A Green function is the response of an entire boundary-value problem to a unit source. Once that response is known, a continuous charge distribution and a patterned boundary voltage can both be treated by integration. The two integrations require different kernels.

Define a point response with its boundary conditions

Let Ω be a vacuum domain with boundary on which the voltage is given. Use the convention $-\nabla_{\mathbf{y}}^2 G(\mathbf{x},\mathbf{y}) = \delta^3(\mathbf{y}-\mathbf{x})$ and $G = 0$ on the boundary as a function of $\mathbf{y}$. In free space $G = 1/(4\pi|\mathbf{x}-\mathbf{y}|)$ if decay is required at infinity. This definition places no permittivity in G; physical charge enters as ρ/ϵ_0. A different frequent convention puts 4π in the defining equation, changing all subsequent prefactors. Comparing the near-source flux to unity is a straightforward way to fix which convention is being used.

Begin with two differentiable functions u, v. The difference $\nabla\cdot(u\nabla v - v\nabla u) = u\nabla^2 v - v\nabla^2 u$ has no mixed-gradient term. Integrate this identity with $u = V(\mathbf{y})$,

$$v = G(\mathbf{x},\mathbf{y}).$$

The source equation for G samples V at $\mathbf{x}$, while Poisson's equation for V supplies a volume density. With the outward normal taken from Ω,

$$V(\mathbf{x}) = \int_\Omega G(\mathbf{x},\mathbf{y})\frac{\rho(\mathbf{y})}{\epsilon_0}\,d^3y - \oint_{\partial\Omega} V(\mathbf{y})\partial_{n_{\mathbf{y}}} G(\mathbf{x},\mathbf{y})\,dA_{\mathbf{y}}.$$

The $G\partial_n V$ boundary term vanishes because $G = 0$ there. If the singularity causes concern, exclude a small sphere about $\mathbf{x}$, apply the ordinary identity, and shrink the sphere. The inward normal on the excised sphere produces the same $V(\mathbf{x})$ term. The representation separates charges in the volume from voltages maintained by electrode sources; it does not combine them into one fictitious material distribution.

Solve the half-space and differentiate with respect to source position

Consider $\mathbf{x} = (\mathbf{X}, h)$ with $h > 0$ above the plane $z = 0$. For an interior source $\mathbf{y} = (\mathbf{Y}, z)$, denote its reflection by $\mathbf{y}^* = (\mathbf{Y}, -z)$. The grounded half-space Green function is

$$G = \frac{1}{4\pi}\left[\frac{1}{\sqrt{|\mathbf{X}-\mathbf{Y}|^2 + (h-z)^2}} - \frac{1}{\sqrt{|\mathbf{X}-\mathbf{Y}|^2 + (h+z)^2}}\right].$$

It vanishes as the source coordinate approaches the plane. The outward domain normal there is $-\hat{\mathbf{z}}$; therefore the coefficient of a prescribed boundary voltage is

$$P_h = -\partial_n G = \partial_z G|_{z=0}.$$

Differentiating the two terms gives equal positive contributions:

$$P_h(\mathbf{X}-\mathbf{Y}) = \frac{h}{2\pi[|\mathbf{X}-\mathbf{Y}|^2+h^2]^{3/2}}.$$

Its units are inverse area, while G has inverse length in this three-dimensional convention. For no interior charge $V(\mathbf{X},h) = \int P_h(\mathbf{X}-\mathbf{Y}) f(\mathbf{Y})\, d^2Y$. The kernel is positive. Integrate over the whole plane using $2\pi s\, ds$; the result is one. Hence a constant boundary voltage is reproduced, and an arbitrary voltage bounded between two numbers produces values between the same bounds inside. The integral solves the bounded half-space problem with the far-field condition appropriate to the data. Demanding decay for a nonzero constant boundary value would be inconsistent.

A detector at height h takes a weighted average over a neighbourhood of width comparable with h. The kernel has algebraic tails, so its support is not a sharp circular footprint. The contribution from radii smaller than b is $1 - h/\sqrt{b^2+h^2}$. Half the total weight lies within $b = \sqrt{3}h$. Ninety per cent requires

$$b = \sqrt{99}h \simeq 9.95h.$$

These quite different radii show why a full-width-at-half-maximum measurement of the kernel cannot by itself define the region that contains most of its influence.

A finite driven electrode and a reciprocity experiment

Fix a circular disk of radius b in the plane at voltage V_d while grounding the rest of the plane, with ideal thin insulating seams. Directly over its centre,

$$V(0,h) = V_d\left(1 - \frac{h}{\sqrt{h^2+b^2}}\right),$$

$$E_z(0,h) = V_d \frac{b^2}{(h^2+b^2)^{3/2}}.$$

At $h \to 0^+$ the electrode voltage is recovered at its centre. For $h \gg b$ the voltage scales as $V_d b^2/(2h^2)$. A small electrode surrounded by ground does not behave at long distance as an isolated charged electrode with a $1/h$ potential. Its surrounding return charge makes a combined disturbance whose far field has a different order. At the sharp ideal edge the boundary voltage jumps, and local fields there require an actual seam geometry if one wants a bounded peak or a breakdown prediction.

For $b = 5.0$ mm, $V_d = 4.0$ V and $h = 5.0$ mm the axial voltage is 1.17 V and field is 283 V/m. A separate experiment grounds every electrode and places a point charge q at that observation point. Green reciprocity relates the net charge on the disk to the previous weighting potential $w = V/V_d$:

$Q_d = -q\,w$. With $q = 20$ pC the disk collects -5.86 pC. Direct integration of the image-plane surface charge $\sigma(s) = -qh/[2\pi(s^2 + h^2)^{3/2}]$ gives the same result. It is precisely $-qP_h$. Thus the boundary kernel has both a voltage-producing interpretation and a charge-collection interpretation measurable with different instrumentation.

Recover the kernel of an electrode strip

A boundary pattern invariant in y should not require a two-dimensional source integral. Integrate P_h over the unused coordinate. With lateral separation $s = X - X'$, the result is

$$p_h(s) = h/[\pi(s^2 + h^2)].$$

It integrates to one over the real line. A strip between $-b, b$ at V_d gives on its central normal $V(0, h) = 2V_d \arctan(b/h)/\pi$. At $h = b$ the value is half the applied voltage, distinct from the circular electrode's $1 - 1/\sqrt{2}$. The strip extends in one more direction and occupies a larger angular range of boundary influence.

For sinusoidal boundary data $f(X) = \cos kX$ there is an independent way to determine the result without evaluating a difficult oscillatory integral. Translational symmetry and linearity preserve the sinusoid. Solving Laplace's equation with vertical factor $Z(h)$ gives

$$Z'' = k^2 Z.$$

The bounded decaying solution with unit boundary amplitude is $e^{-|k|h}$. Thus the Fourier transform of the strip kernel is $e^{-|k|h}$. Convolving twice with heights h_1, h_2 is equivalent to convolution once with height $h_1 + h_2$. This composition property checks a numerical propagation code without requiring separate analytic solutions for many different boundary patterns.

Which symmetries does an influence table possess?

Exchange source and detector in the defining equations for $G(\mathbf{x}, \cdot)$ and $G(\mathbf{z}, \cdot)$. Their difference identity gives $G(\mathbf{x}, \mathbf{z}) - G(\mathbf{z}, \mathbf{x}) = 0$, since both vanish on the same boundary. Symmetry extends to scalar or symmetric-tensor positive permittivity when the operator is defined as $-\nabla \cdot \epsilon \nabla$. It refers to exactly the same material, boundary preparation and domain in both experiments. Replacing one experiment's grounded conductor by a floating conductor changes the Green function.

On a nonuniform mesh it matters whether influence entries correspond to equal point charges or equal charge densities in unequal cells. If a finite-volume solver stores sources as densities, a column's total injected charge scales with its cell volume. An unweighted table of sampled voltages can look nonsymmetric even when the discrete physics is reciprocal. Form the matrix with unit total source and consistent detector averaging, or retain the correct volume-weighted inner product. Test positivity for positive sources as a second check. Symmetric influence entries by themselves cannot expose a common incorrect normalisation or a misplaced outer boundary.

A kernel that reproduces every harmonic around a circle

In a two-dimensional circular domain of radius a, prescribe a voltage $f(\phi')$ around the wall. For a detector at (r, ϕ) let $t = r/a < 1$. The angular Fourier component with order m is multiplied by $t^{|m|}$ on continuation into the disk. This follows directly from the regular radial power solution of Laplace's equation. Substitute the Fourier coefficient integral into the series and interchange the integral and the absolutely convergent series at fixed $t < 1$. The boundary response per unit angle is then

$$P_t(\delta) = \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} t^{|m|} e^{im\delta} = \frac{1 - t^2}{2\pi(1 - 2t\cos\delta + t^2)},$$

$$\delta = \phi - \phi'.$$

To obtain the rational form, separate the $m = 0$ term and sum the two conjugate geometric series $te^{\pm i\delta}/(1 - te^{\pm i\delta})$. The result is positive for every interior point. Integrating it over angle leaves only the constant Fourier mode, so it is normalised to unity. At the centre the kernel is simply $1/(2\pi)$, recovering the exact mean-value theorem for harmonic functions. Near the wall it becomes very sharply concentrated around the detector's azimuth while retaining a small nonlocal tail. Using prescribed boundary data as a flat average at every interior point would reproduce only the centre reading.

For a concrete experiment put a short arc of half-angle α centred on $\phi = 0$ at 1 V and ground the remaining circular boundary. Along $\phi = 0$ the potential equals the integral of P_t between $-\alpha, \alpha$. With $u = \tan(\phi'/2)$ one has $\cos\phi' = (1 - u^2)/(1 + u^2)$, $d\phi' = 2du/(1 + u^2)$. Substitution gives the weight

$$w(t) = \frac{2}{\pi} \arctan\left[\frac{1 + t}{1 - t} \tan(\alpha/2)\right].$$

At $t = 0$ it gives α/π, exactly the fraction of the circumference occupied by the arc. For a 60°-wide arc $\alpha = \pi/6$, hence the centre reads $1/6$ V. At $t = 0.6$ on the centreline it reads about 0.522 V. Placing a small positive line charge there while grounding the whole cylindrical wall induces -0.522 times the line charge on that arc by the same reciprocity argument used on the flat disk. The angular share depends on position inside the circle even though every point on the conductor has the same actual grounded voltage.

An independent check takes the limit of a very large circle with the probe only a distance $h = a - r$ from the nearest boundary point. For small angular displacement $\delta = s/a$, $1 - 2t\cos\delta + t^2 \simeq (h/a)^2 + (s/a)^2$ and $1 - t^2 \simeq 2h/a$. The angular probability element $P_t(\delta)d\delta$ therefore tends to $h\,ds/[\pi(h^2 + s^2)]$, precisely the translation-invariant strip kernel obtained by integrating the half-space formula over one coordinate. Curvature corrections can be estimated by increasing a/h at fixed physical arc width and comparing both predictions.

Why an insulated Green function needs a zero-mode subtraction

Consider the simpler interval $0 < x < L$ with prescribed zero flux at both ends. We would like a point response satisfying $-G'' = \delta(x - s)$, $G'(0) = G'(L) = 0$. Integrate this proposed equation over the

interval: the left-hand side is $-G'(L) + G'(0) = 0$, whereas the right-hand side is 1. The attempted kernel violates compatibility. Physically a sealed insulating region cannot take a net free test charge without changing its prescribed outer flux. Adding an arbitrary constant to G cannot fix this contradiction.

A useful alternative removes the constant source mode and fixes the remaining voltage reference. Define $-G_N'' = \delta(x - s) - 1/L$, $G_N'(0) = G_N'(L) = 0$, and $\int_0^L G_N(x, s)\, dx = 0$. Away from the point the curvature is $+1/L$. The left derivative must be x/L to vanish at zero, the right derivative $x/L - 1$ to vanish at L. They jump by -1 at $x = s$. Integrating, imposing continuity and subtracting the mean gives the compact symmetric expression

$$G_N(x, s) = \frac{x^2 + s^2}{2L} - \max(x, s) + \frac{L}{3}.$$

Differentiating verifies both end conditions and the jump. Integrating the quadratic part and the piecewise maximum over x verifies the mean condition. Its sign can be negative because the response includes an opposite uniform return source and uses a zero-mean reference. Unlike the positive grounded Dirichlet kernel, it is not a positive voltage weight to use for a maximum-principle test.

Apply a neutral pair of sheets $+\sigma$ at $s_1 = 0.2L$ and $-\sigma$ at $s_2 = 0.7L$. Their voltage is $\sigma[G_N(x, s_1) - G_N(x, s_2)]/\epsilon$. The two artificial uniform return sources cancel exactly. To the left of both sheets the field vanishes, between them it equals $+\sigma/\epsilon$, and to the right it vanishes again. The potential drops by $0.5L\sigma/\epsilon$ across the two sheets. This matches direct Gauss-law jumps, and the zero-mean potential fixes the constant that Gauss's law could not determine. For more complex closed domains numerical solvers implement the same idea by projecting out the constant eigenvector and checking the compatible total charge before inverting. An arbitrary tiny diagonal regularisation of a singular matrix can hide a physically impossible net-source input.

Construct a point response when the medium itself screens charge

A linear ionic or phenomenological screening model between two grounded parallel electrodes leads in one dimension to $[-\partial_x^2 + \kappa^2]V = \rho_{\text{imposed}}/\epsilon$, $0 < x < L$. Here ρ_{imposed} excludes the mobile linear response already represented by $\kappa^2 V$. We seek $G_\kappa(x, s)$ with zero voltage at both ends. Away from its unit source the equation has hyperbolic solutions. Use $A\sinh(\kappa x)$ on the left to vanish at zero and $B\sinh[\kappa(L - x)]$ on the right to vanish at L. Continuity at $x = s$ removes one constant. Integrating across the source gives $G_\kappa'(s^+) - G_\kappa'(s^-) = -1$; the finite mass term integrates to zero in a shrinking source interval.

To solve transparently use the functions $u(x) = \sinh(\kappa x)$,

$$v(x) = \sinh[\kappa(L - x)].$$

Their Wronskian $u'v - uv'$ is $\kappa\sinh(\kappa L)$, independent of x. Then continuity, the derivative jump, and symmetry under $s \leftrightarrow x$ are manifest in

$$G_\kappa(x,s) = \frac{\sinh(\kappa x_<)\sinh[\kappa(L-x_>)]}{\kappa\sinh(\kappa L)},$$
$$x_< = \min(x,s),$$
$$x_> = \max(x,s).$$

For $\kappa \to 0$, expand each hyperbolic sine to first order. The result is $x_<(L-x_>)/L$, the grounded unscreened response. When both observation and source lie many screening lengths from every boundary, the kernel tends to $e^{-\kappa|x-s|}/(2\kappa)$, the infinite-line screened solution. Thus one formula contains both geometric screening by grounded conductors and material screening by mobile charges.

At the midpoint source and detector coincide. The finite kernel reduces to $\tanh(\kappa L/2)/(2\kappa)$. Compared with the unscreened midpoint value $L/4$, its ratio is $2\tanh(\kappa L/2)/(\kappa L)$. At $\kappa L = 2$ a fixed inserted sheet therefore creates only 76.2% of its unscreened midpoint voltage. A naïve replacement of plate separation by the screening length would give a constant ratio and miss the smooth small-κ crossover. For long gaps the measured voltage saturates at fixed sheet charge, because the countercharge response near the sheet now sets the local scale.

Check the convolution using a uniform positive imposed charge density ρ_0. Without integrating the kernel directly, seek a constant particular solution plus a symmetric homogeneous correction. The result is

$$V(x) = [\rho_0/(\epsilon\kappa^2)]\{1 - \cosh[\kappa(x-L/2)]/\cosh(\kappa L/2)\}.$$

Twice differentiation reconstructs the source and zero endpoints. In the small-κL limit it tends to $\rho_0 x(L-x)/(2\epsilon)$. A numerical integration of the point response against a uniform source should match this independent closed solution throughout the interval, including the near-electrode regions. It is a stronger test than checking the point kernel at the centre alone.

Integrate the original equation across the gap. The sum of outward displacement fluxes equals imposed charge minus $\epsilon\kappa^2\int V\,dx$, the magnitude of the compensating mobile charge. For the uniform source the fraction of imposed charge balanced by electrode rather than mobile charge is $2\tanh(\kappa L/2)/(\kappa L)$. For a long gap this fraction falls like twice the screening length divided by gap width, because only the two boundary regions communicate appreciably with the electrodes. A current meter attached to the electrodes and a chemical measurement of accumulated mobile ions can therefore close the same source budget independently.

Inspect the response to a localised source

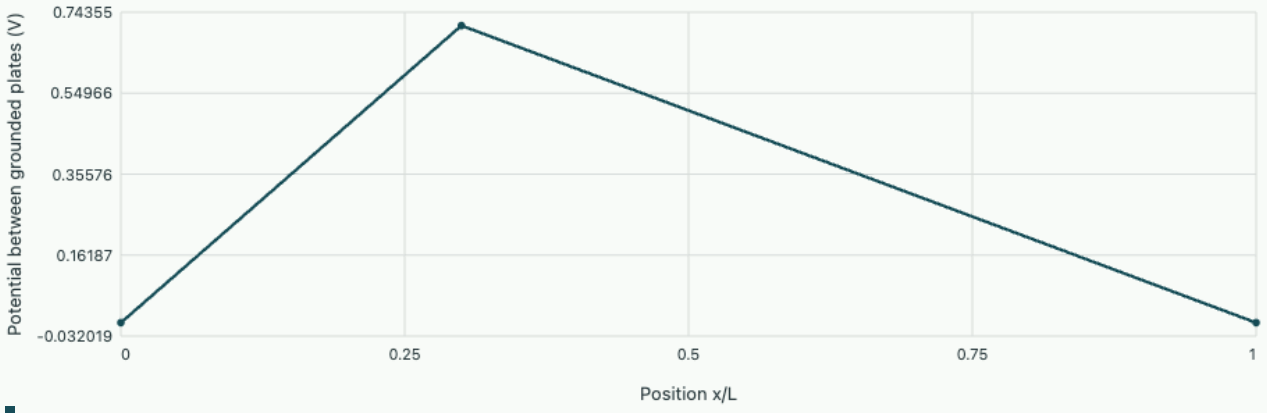
Move the source while holding the boundaries fixed. Test the symmetry between source and observation point predicted by reciprocity.

Open this experiment online

Try the calculation

Sheet position s/L **0.3**

Sheet charge density (nC/m²) **3**



Grounded 10 mm gap

Induced densities are -2.1 nC/m² on the left and -0.9 nC/m² on the right. Potential at the sheet is 0.712 V.

Design a differential collecting electrode from its weighting functions

Two concentric electrodes supply one signed position signal

A grounded plane is divided into a central disk of radius a, a surrounding annulus from a to $2a$, and the remainder. A small charge moves along the central normal. Two charge amplifiers measure the disk and annulus. Construct an output that vanishes at $h = a$, then find its local height sensitivity. Why can that output alone not distinguish uncertain charge from height away from the null?

▼ **Derive and check the result**

Let $w_a = 1 - h/\sqrt{h^2 + a^2}$ and $w_{2a} = 1 - h/\sqrt{h^2 + 4a^2}$. The disk reads $Q_d = -qw_a$; the annulus reads $Q_r = -q(w_{2a} - w_a)$. An output $S = Q_d - \gamma Q_r$ vanishes at $h = a$ if $\gamma = (1 - 1/\sqrt{2})/(1/\sqrt{2} - 1/\sqrt{5}) = 1.127$ approximately. Both readings are negative for positive q, so it is their difference, not their sum, that rejects the reference position. The ideal zero does not depend on the magnitude of any nonzero q.

Differentiate with respect to height using $w_b' = -b^2/(h^2 + b^2)^{3/2}$. At the null $a\,w_a' = -1/(2\sqrt{2})$ and $a\,w_{2a}' = -4/(5\sqrt{5})$. Hence

$$a\,d(S/q)/dh = -(1 + \gamma)aw_a' + \gamma a w_{2a}' \simeq 0.349.$$

For $q = 10$ pC, $a = 2.0$ mm, a small upward displacement of 0.10 mm produces approximately $+0.175$ pC. Finite differences of the full two-disk formula at heights on both sides of a check the linear slope. For displacements that are no longer small compared with a, use the full expression rather than extrapolating that single sensitivity.

An additional sum channel $Q_d + Q_r = -qw_{2a}$ permits a dimensionless output $S/(Q_d + Q_r)$ that cancels the magnitude and sign of q while retaining height dependence. This ratio becomes noise sensitive far from the electrodes, where collected charge is small. A null by itself can also mean that there is no charge at all, so a usable position detector checks the sum channel at the same time. The exercise combines electrostatic reciprocity, electronic weighting and a finite signal-to-noise constraint while requiring only two precomputed boundary responses.

Tutorial · predict before numerically integrating the kernel

How much charge belongs to an annular collector?

An annular electrode with inner radius h and outer radius $3h$ lies in a grounded plane. A charge q sits at height h over its centre. Derive its induced charge. If a numerical integration omits all radii beyond $3h$, where does the missing charge go?

▼ **Work through the solution**

Subtract two circular weights. The annulus receives $Q/q = -[1/\sqrt{2} - 1/\sqrt{10}] = -0.3909$. The central disk receives $-(1 - 1/\sqrt{2})q = -0.2929q$. The whole plane should sum to $-q$, so the part beyond the outer edge supplies

$$-q/\sqrt{10} = -0.3162q.$$

Nothing has been lost by the physical system; a finite integration patch excludes a considerable piece of the long-tailed kernel. Radial extrapolation with the analytic far tail or an expanding integration domain distinguishes missing area from poor quadrature within the patch.

Use resistance to a modelling mistake

A student proposes G itself as the weight for setting electrode voltages on the plane. Another replaces it by $2P_h$, arguing that an image doubles the response. Predict two different tests that diagnose these errors.

▼ **Work through the solution**

The first proposal gives zero contribution since the Dirichlet Green function is zero at the boundary. It confuses volume source response with a boundary normal derivative. The second produces twice the prescribed constant voltage because the kernel would integrate to two. Independent electrostatic

reciprocity would assign $-2q$ to the whole grounded plane around a charge q, violating the charge sum derived from the real and image fields. For a numerical test, set successively a constant, a sinusoid, and a bounded disk on the same plane. Those cases examine normalisation, spectral attenuation, and spatial tails separately.

Further reading

- Move a source between two grounded electrodes
- Compare vertical attenuation for two boundary harmonics
- Fließbach, chapter 7 · electrostatic boundary problems

Rectangular geometries, orthogonality and sources that drive individual modes

Electrostatic field theory and measurement

Fourier modes are solutions selected by their electrode walls. On a rectangle they turn a boundary-voltage problem into independent vertical equations. When a charge fills the rectangle the same eigenfunctions give a different denominator and a different physical response.

Determine which sign of a separation constant can fit the walls

Take a vacuum rectangle $0 < x < a,\ 0 < y < b$, invariant along z. Ground the left, right and bottom edges and impose $V(x, b) = f(x)$ at the top. If the boundary data disagree at a corner, the continuum problem can have large edge fields; we first choose smooth f vanishing at both endpoints. In the charge-free interior, a product solution $X(x)Y(y)$ gives $X''/X = -Y''/Y$, equal to a constant wherever the product is nonzero. If $X'' = k^2 X$, the only combination of hyperbolic functions zero at both $x = 0, a$ is zero. The same is true for the linear case $k = 0$. Thus $X'' = -k^2 X$. The left boundary picks a sine and the right boundary enforces $k = n\pi/a,\ n = 1, 2, \ldots$.

Then $Y'' = k^2 Y$, with the bottom condition choosing $\sinh ky$. Choosing its top value as unity gives the transfer function $Y_n = \sinh(k_n y)/\sinh(k_n b)$. Summing products,

$$V(x, y) = \sum_{n \geq 1} f_n \sin(k_n x) \frac{\sinh(k_n y)}{\sinh(k_n b)},$$

$$f_n = \frac{2}{a} \int_0^a f(x) \sin(k_n x)\, dx.$$

These coefficients follow by multiplying the top boundary by $\sin(k_m x)$ and integrating. For $n \neq m$ the product reduces to cosines with integer half-wave multiples and integral zero; for equal indices its integral is $a/2$. This orthogonality incorporates the edge boundary values. Periodically repeating a voltage function without regard to its endpoint conditions can give cosine or constant terms that violate the grounded side walls.

Track how geometry filters electrode detail

A mode's vertical derivative is $k_n \cosh(k_n y)/\sinh(k_n b)$, so the field normal to the electrode weights high modes differently from potential. At the top the magnitude multiplier is $k_n \coth(k_n b)$. For a half-space, high-frequency field content can be large very near a patterned electrode even if it disappears rapidly a short distance inside. For large $k_n b$, direct numerical evaluation of hyperbolic sines may overflow. The algebraically equivalent ratio $e^{-k_n(b-y)}[1 - e^{-2k_n y}]/[1 - e^{-2k_n b}]$ stays bounded; for very small arguments evaluate $1 - e^{-t}$ with a numerically stable exponential-difference routine.

Suppose the rectangle is square and the electrode has $f = 4\sin(\pi x/a) - 2\sin(3\pi x/a)$ V. At mid-height $y = a/2$, the fundamental transfer factor is $1/[2\cosh(\pi/2)] = 0.1993$, while the third is

$$1/[2\cosh(3\pi/2)] = 0.00898.$$

At $x = a/2$ the first sine is 1 and the third is -1, so the potential is 0.815 V approximately, compared with the 6 V drive at the same x on the electrode. Taking the simple arithmetic average of the top and bottom voltages at that x would predict 3 V and miss leakage to the grounded side walls.

If a second boundary carries a pattern, superpose the source-free solution produced by it with the other boundaries grounded. For example a nonzero bottom voltage pattern uses $\sinh[k_n(b-y)]/\sinh(k_n b)$. This works because the field equation and boundary conditions in fixed-voltage form are linear. A floating conductor imposes a different constraint coupling many modal coefficients, not the prescription of one extra arbitrary voltage profile.

Integrate a discontinuous electrode and distinguish two kinds of convergence

Let a short strip centred on the top boundary span $a/2 - w/2 < x < a/2 + w/2$ with voltage V_d, the remaining top grounded. Integrating the sine over that interval gives

$$f_n = 4V_d \sin(n\pi/2)\sin(n\pi w/(2a))/(n\pi).$$

Even modes vanish by mirror symmetry. Individual allowed coefficients decrease as $1/n$ at most because the boundary has jumps. On the boundary partial sums overshoot close to a jump. Adding terms narrows that region rather than removing its limiting fractional overshoot. At fixed interior distance the exponential transfer renders the series much better behaved.

For $w = a/3$, the first and third coefficients are $2V_d/\pi$ and $-4V_d/(3\pi)$. Along the centreline both contribute positively because their sine factors have opposite signs. A finite-width electrode cannot be captured exactly with a single spatial mode even when its interior signal far away is mainly fundamental. Recovering the electrode width from a distant scan by dividing attenuated high modes by their transfer factors will amplify measurement errors. Forward propagation and inverse reconstruction need different treatments of frequency content.

Let the volume rather than the electrode excite the modes

Ground all four edges and embed a known charge density $\rho(x, y)$. Choose double sines $\sin(m\pi x/a)\sin(n\pi y/b)$. They already vanish on all edges and are eigenfunctions of the negative Laplacian with eigenvalues $\lambda_{mn} = (m\pi/a)^2 + (n\pi/b)^2$. Project the charge using $\rho_{mn} = 4(ab)^{-1}\int_0^a \int_0^b \rho \sin(m\pi x/a)\sin(n\pi y/b)\,dy\,dx$. Applying Poisson's equation coefficient by coefficient yields

$$V_{mn} = \frac{\rho_{mn}}{\epsilon_0 \lambda_{mn}}.$$

High spatial frequencies of a volume charge produce weaker potentials by an inverse-square spectral factor. The corresponding normal field on a boundary involves an additional derivative. This polynomial

filtering differs from the exponential decay of a boundary pattern measured away from its source. Both follow from the same PDE but answer different experiments.

For a manufactured test choose a square of side 10 mm with $\rho = \rho_0 \sin(\pi x/a)\sin(2\pi y/a)$. The voltage magnitude peaks at $\rho_0 a^2/(5\pi^2\epsilon_0)$. With $\rho_0 = 1.0\ \mu\mathrm{C/m^3}$ it is 0.229 V. The total charge integrates to zero, since the second sine is odd about the mid-height. The field and stored energy are nonzero. At $x = a/2, y = a/4$ the voltage is a positive maximum, consistent with a positive local source. At $x = a/2, y = 3a/4$ the same amplitude is negative.

Use an exact discrete eigenvalue for a rigorous code check

On a uniform square grid of spacing $h = a/N$, replace $-\partial_x^2$ by its centred three-point formula. Acting on the sampled sine gives eigenvalue $4h^{-2}\sin^2(m\pi h/(2a))$. The two-dimensional sum of the corresponding factors is therefore the exact discrete eigenvalue for this interior mode with zero edges. A computed solution can be compared first to $\rho_{mn}/(\epsilon_0\lambda_{mn,h})$, which tests the solver algebra, and then to the continuous $\rho_{mn}/(\epsilon_0\lambda_{mn})$, which tests spatial discretisation. Reducing the iterative residual alone cannot remove the difference between these two physical amplitudes.

At $N = 10$, modes $(1, 2)$ have $a^2\lambda_h = 4N^2[\sin^2(\pi/20) + \sin^2(\pi/10)] = 47.99$, versus $5\pi^2 = 49.35$. The discrete amplitude is about 2.8% high despite solving its linear equations exactly. For fixed mode under grid halving, expansion of the sine gives an eigenvalue error proportional to h^2. At wavelengths only a few cells wide this expansion is unreliable. Try a sequence of grids with the same physical mode and compare actual amplitude ratios. Holding instead a fixed number of cells per wavelength while changing the physical wavelength tests a different convergence question.

Two transverse dimensions select a different attenuation length

Extend the rectangle into a rectangular conducting shaft $0 < x < a,\ 0 < z < c$, running to $y > 0$. Ground all four side walls, require bounded potential far down the shaft and prescribe a voltage on its mouth at $y = 0$. Separation in the cross section now gives products $\sin(m\pi x/a)\sin(n\pi z/c)$ with both $m, n \geq 1$. The axial equation is

$$Y'' = [(m\pi/a)^2 + (n\pi/c)^2]Y.$$

Each boundary coefficient therefore propagates as $e^{-\gamma_{mn}y}$,

$$\gamma_{mn} = \pi\sqrt{m^2/a^2 + n^2/c^2}.$$

A narrow second dimension suppresses penetration even when the first dimension is very broad. Approximating the shaft by a two-dimensional strip removes an entire contribution to the decay constant.

If the aperture is driven uniformly at V_0, projecting gives $f_{mn} = 16V_0/(\pi^2 mn)$ for odd m, n, zero otherwise. The ideal mouth voltage jumps at the grounded rim. Deep inside, only the lowest odd pair

survives and its amplitude exceeds V_0 before the exponential is applied; the higher terms restore the actual boundary value at $y = 0$. With $a = 20$ mm and $c = 10$ mm the slowest decay length is

$$1/\gamma_{11} = 2.85\,\mathrm{mm}.$$

A two-dimensional strip ignoring the 10 mm side would instead give

$$a/\pi = 6.37\,\mathrm{mm}.$$

After 10 mm penetration the first-mode attenuation factors are 0.0298 and 0.208, almost a factor of seven apart.

A simple numerical test samples the centreline voltage at several heights and calculates adjacent log slopes. Close to the mouth many odd modes matter, while deep inside the ratio between two positions Δy apart approaches $e^{-\gamma_{11}\Delta y}$. Compare with a second geometry that swaps a, c. The two lowest-order decay lengths must agree because the eigenvalue is symmetric in the two dimensions; an implementation that uses only the nominal 'width' may fail this check. A second observation path away from the centre can sense different mode weights, even though every mode's exponential is unchanged.

Suppose the mouth instead has opposite voltages on the left and right halves, antisymmetric about $x = a/2$. Its coefficients vanish for odd m, so the $(1, 1)$ mode is absent by symmetry. The leading allowed mode can be $(2, 1)$, shortening the decay length. A centreline probe would then read zero everywhere by antisymmetry and could not fit any length at all. Move the probe to $x = a/4$ while keeping it at mid-height of the cross section. This illustrates how an excited spectrum, a geometry's complete set of eigenvalues, and an instrument's observed spectrum may differ.

Why the energy of an ideal electrode voltage step depends on edge resolution

Return to a two-dimensional strip or half-plane. To focus on a single boundary Fourier mode take $V = A\sin(kx)e^{-ky}$ in $0 < x < a,\ y > 0$, with $ka = n\pi$. Its gradients in x and y have equal prefactor kAe^{-ky}. The sum of their squared sine and cosine factors is one, so the energy per perpendicular length is

$$U_n' = \epsilon a k A^2/4.$$

For many orthogonal sine modes cross terms vanish after integration. Thus the stored field energy involves the weighted sum $\sum_n n|f_n|^2$, while the mean-square boundary voltage itself involves simply $\sum_n |f_n|^2$. The extra factor of n comes from the steep near-electrode field of short-wavelength structure.

For a finite electrode patch abruptly stepping from 0 to V_d, the allowed Fourier amplitudes typically decay as $1/n$ along nonzero subsequences. The voltage series converges in mean square, but its energy sum can contain a logarithmically diverging $\sum 1/n$. This is the local singularity of the zero-

width voltage discontinuity. It does not prevent a well-defined voltage at any interior distance, but it prevents assigning a finite total energy or a cut-off-independent peak electric field to that exact sharp-edge idealisation. Mesh refinement can produce progressively higher energy estimates even as interior voltage curves appear perfectly stable.

There is a local way to see the same conclusion without a Fourier series. Consider a small half-disk of vacuum around a junction where an otherwise flat boundary jumps between voltages zero and V_d. In polar coordinates about the junction, the leading harmonic solution depends only on angle: $V = V_d\phi/\pi$ with angle origin chosen on the grounded side. Its field magnitude is $V_d/(\pi r)$. Integrating $\epsilon E^2/2$ over a half-annulus of inner radius δ and outer radius R gives

$$U' = \epsilon V_d^2 \ln(R/\delta)/(2\pi).$$

Halving the microscopic length changes this single-edge contribution by $\epsilon V_d^2 \ln 2/(2\pi)$. The independence of R in that change is a practical test of the logarithmic asymptotic regime.

A finite insulation seam between real electrodes supplies a different boundary problem. For a deliberately smoothed transition of width δ, the spectrum falls faster above $k \sim 1/\delta$; the local singularity is cut off. One can compare a family of specified physical seam widths while separately refining each numerical grid until it resolves its own seam. Shrinking seam width along with mesh spacing instead changes the physical geometry at each refinement. Report the interior voltage, near-edge energy within a named observation region and the physical seam model separately. Each can converge at a different rate even in a technically correct Laplace solver.

A discrete residual can fall quickly while the wrong spatial scale remains

The separated sine basis makes it possible to analyse the convergence of a concrete numerical Laplace solver as well as the continuum solution. On a square grid with equal spacing, the five-point discretisation in a charge-free cell says its voltage equals the mean of its four neighbours. A Jacobi sweep sets all new values to this mean of the old grid, preserving fixed boundary values meanwhile. Subtract the exact discrete solution from the current approximation: the error evolves by the same averaging step with zero boundary values, independently of the inhomogeneous source that generated the original solution.

Insert an error pattern $e_{ij} = \sin(m\pi i/N)\sin(n\pi j/N)$. Elementary sine addition on the two x-neighbours produces $2\cos(m\pi/N)e_{ij}$; the y-neighbours produce the analogous factor. Hence one sweep multiplies this error mode by $\mu_{mn} = [\cos(m\pi/N) + \cos(n\pi/N)]/2$. For the longest $(1,1)$ mode, $\mu \simeq 1 - \pi^2/(2N^2)$. Its amplitude takes roughly $2N^2/\pi^2$ sweeps to drop by $1/e$. Doubling resolution therefore makes this relaxation about four times slower per physical longest mode even before counting the four times as many cells per sweep.

Choose $N = 64$. Then $\mu_{11} = 0.99880$. After 100 sweeps about 88.6% of that initial smooth error still remains, although the update per cell has become small. At the opposite corner of the spectrum $(N-1, N-1)$ gives nearly -1, so a checkerboard mode also decays slowly while changing sign on

successive sweeps. Weighted Jacobi replaces an old value by $(1-\omega)$ times itself plus ω times the neighbour mean. At $\omega = 2/3$ that checkerboard amplification becomes approximately $-1/3$, removing short-wavelength oscillations quickly. The long smooth mode still has amplification near one. An attractive smooth colour map after a few sweeps can thus be misleading.

A multigrid correction exploits these distinct errors. First relax high-frequency errors on the fine grid. Compute the residual of the original discrete equation and restrict it to a coarser grid. A long fine-grid mode becomes easier to change there. Solve the coarse error equation, prolong its correction back to the fine grid, and smooth again. The correction must solve for error with the appropriate homogeneous boundary constraints, not repeat the original boundary excitation a second time. The residual and prolongation should respect the same physical units and grid volumes as the original source. Otherwise a visually plausible coarse correction can inject net charge at each level.

For a manufactured check initialise the fine-grid error to a known sine mode on top of an exact discrete solution. After one weighted sweep measure the inner product of the new error with the original mode and divide by its squared norm. The observed ratio must equal $1 - \omega + \omega\mu_{mn}$ before roundoff or boundary effects. Repeat with $(1, 1)$, a mixed $(1, N-1)$ mode, and the checkerboard-like $(N-1, N-1)$ mode. This directly tests the implementation's indexing, old-versus-new updates and boundary handling; it does not require the continuum equation to be discretisation exact.

Finally compare iterative and truncation errors in measurable volts. For a discrete mode, if the algebraic residual coefficient is r_{mn} in the form $-\nabla_h^2 e = r$, its error coefficient is $r_{mn}/\lambda_{mn,h}$. The same residual amplitude in a long-wavelength mode produces a larger voltage error than in a high-frequency mode. An absolute solver tolerance with no physical scaling is therefore insufficient to certify a voltage reading. Stop iterating when the solver error in the actual observable is small compared with discretisation and experimental uncertainty, then refine the spatial representation separately.

Resolve a short-wavelength electrode pattern

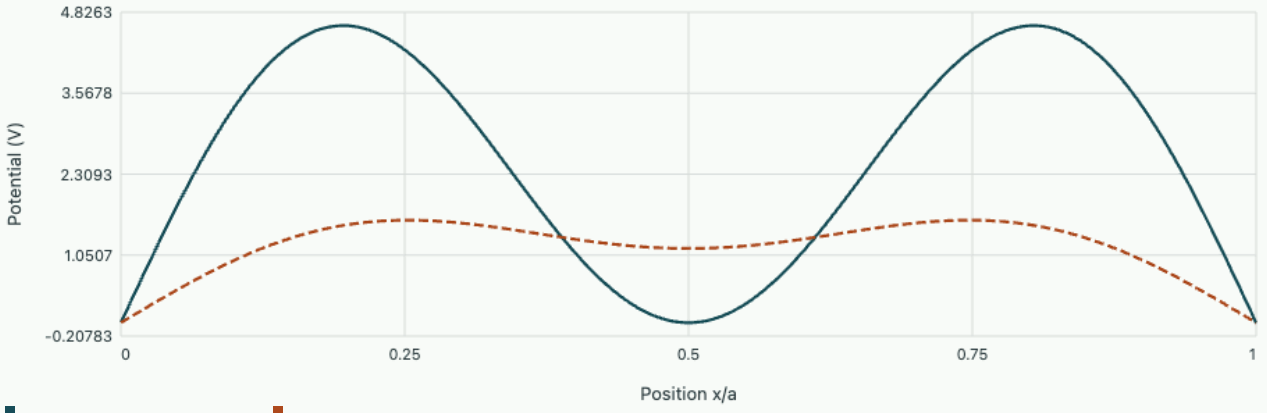
Increase the transverse mode number. Predict how far the pattern reaches into the strip before inspecting the voltage profile.

Open this experiment online

Try the calculation

Height y/a **0.18**

Third harmonic at the boundary (V) **3**



At the middle of the strip the potential is 1.15 V. The third harmonic retains 18.3% of its boundary amplitude.

Build a resistor-network analogue of the rectangular boundary problem

Use Kirchhoff measurements to check a field solver

A square lattice of nodes is connected by identical resistors R between nearest neighbours. Outer nodes are held at prescribed voltages. Derive the interior equation from charge conservation at a node. Which measured electrical quantity can test a normal field at the boundary? What changes on an anisotropic grid?

▼ **Derive and check the result**

For a steady interior node with no supplied current, the sum of currents $(V_{ij} - V_{\text{neighbour}})/R$ is zero. Rearrangement gives the same mean-of-four equation as the square-grid discrete Laplacian. Thus the resistor array reproduces the *discrete* Dirichlet solution for the same pinned boundary voltages. Its Joule dissipation $\sum_{\text{edges}} (V_i - V_j)^2/R$ is stationary against variations of the unfixed voltages, giving the same equations a second way. The numerical voltage maps can be compared node by node; the physical power and the electrostatic field energy are different dimensional quantities even though their quadratic forms are analogous.

To estimate the normal field near a flat boundary, measure the signed current from a boundary node along the edge connecting it to its nearest interior neighbour. This equals the observed voltage difference divided by R; multiplying by R/h converts it to the corresponding one-sided field magnitude

in volts per metre. At sharp corners the one-sided difference has large discretisation error and the continuum physical field may be singular. Summing all currents supplied to a terminal edge gives a more robust integrated flux analogue. Account for any links between nodes at different prescribed boundary voltages if those links contribute separate current to the laboratory power supplies.

For horizontal spacing h_x and vertical spacing h_y, replacing every resistor by the same value no longer matches the isotropic Cartesian Laplacian. A finite-volume cell requires conductance proportional to face length divided by edge length: horizontal links have conductance proportional to h_y/h_x, vertical links to h_x/h_y. Their ratio is h_y^2/h_x^2. This restores a discrete conservation law with the correct physical area factors. A measured profile agreeing exactly with an equal-resistor lattice on a stretched geometry could therefore validate the wrong continuum material or geometry unless those weights were checked first.

Tutorial · design a boundary measurement and a solver benchmark

Identify a forbidden mode

A symmetric positive top electrode has its centre at $x = a/2$. A numerical sine transform reports a large $n = 2$ coefficient. Is it evidence of physical high-frequency content? Which controlled change could verify the sign of the defect?

▼ **Work through the solution**

Mirror substitution $x \mapsto a - x$ preserves the boundary voltage but changes $\sin(2\pi x/a)$ to its negative. Its exact projection is zero. Sharp electrode edges may require many odd modes; they cannot explain an even coefficient under exact symmetry. Check coordinate origin, edge alignment on the grid and measurement offsets. Shift the actual patterned patch by a small amount to one side, then the same amount to the opposite side. A genuine even coefficient changes sign under these mirrored displacements, while odd-mode amplitudes agree. The response at interior height has the same parity because the transfer factor depends on mode number alone.

Measure where electrode charge resides

For a single sinusoidal top voltage $V_d \sin(\pi x/a)$, derive the total free charge per perpendicular length on the upper electrode face. Why does it depend on the bottom distance even though the top voltage profile is the same?

▼ **Work through the solution**

At the upper metal the normal from metal into the dielectric is $-\hat{\mathbf{y}}$. With $E_y = -\partial_y V$, the metal charge density is

$$+\epsilon_0 \partial_y V = \epsilon_0 k V_d \coth(kb) \sin(kx),$$

$$k = \pi / a.$$

Integrating along $0 < x < a$ yields $2\epsilon_0 V_d \coth(\pi b / a)$ per length. Bringing the bottom closer increases $\coth(kb)$. For $kb \ll 1$ the expression tends to $\epsilon_0 b^{-1} \int_0^a f(x) dx$, the local parallel-plate limit. A segmented electrode supplies a nonuniform fixed profile; the result counts its integrated face charge and assumes the ideal voltage boundary remains prescribed in each segment.

Further reading

- Move the measurement plane above a patterned electrode
- Evaluate a grounded-enclosure boundary problem
- Griffiths, 5th ed., §3.3.1 · Cartesian separation

Cylindrical geometry, coaxial boundaries and Bessel spectra

Electrostatic field theory and measurement

The radial coordinate of a cylinder leads to logarithms in one experiment and to Bessel functions in another. The difference is set by the variation along the axis and around the circumference. Recovering the radial weight from flux clarifies both.

A conducting annulus selects a logarithmic radial potential

Put a long cylindrical metal electrode of radius a inside a coaxial grounded tube of radius b, with homogeneous dielectric of permittivity ϵ in the gap. Long means that probes are far enough from both ends for axial variation to be negligible. With no free gap charge and no angular variation, Laplace's equation reduces to $\partial_r(r\partial_r V) = 0$. One integration gives $rV' = C$, a statement of conserved radial flux through cylinders; the second gives a logarithm. With inner electrode at V_0,

$$V(r) = V_0 \frac{\ln(b/r)}{\ln(b/a)},$$

$$E_r = \frac{V_0}{r \ln(b/a)},$$

$$C' = \frac{2\pi\epsilon}{\ln(b/a)}.$$

Here C' is capacitance per length. To obtain it, integrate D_r over the inner metal face per unit axial length and divide by V_0. Another route integrates the energy density $\epsilon E_r^2/2$ times $2\pi r\,dr$, giving $C'V_0^2/2$. For $a = 2$ mm, $b = 8$ mm and $\epsilon = 2\epsilon_0$, $C' = 80.3$ pF/m. At $r = \sqrt{ab} = 4$ mm the voltage is $V_0/2$; the half-voltage radius is the geometric mean. An equally spaced radial ruler does not sample equal voltage steps.

The logarithm is not allowed at the axis of a solid charge-free cylinder because it diverges and carries line-source flux. In the annular domain the singular axis is outside the field region. The same radial function can therefore be admissible or inadmissible in two problems with very similar geometry. Before discarding a mathematical solution, check which radii actually lie in the domain and whether singular sources were explicitly prescribed.

Nonuniform wall voltages in a two-dimensional disk

Next take a full disk of radius a with prescribed side voltage $f(\phi)$, invariant in z. Seek $R(r)e^{im\phi}$. Single-valuedness after $\phi \to \phi + 2\pi$ forces integer m. The radial equation $r^2R'' + rR' - m^2R = 0$ is solved by $r^{|m|}$ and $r^{-|m|}$ for $m \neq 0$; regularity at the centre chooses the growing power. The zero mode is constant, discarding its logarithmic partner. Thus an angular Fourier component on the wall penetrates as $(r/a)^{|m|}$. At $r = a/2$, a third harmonic retains one eighth while a sixth harmonic retains only one sixty-fourth. Both can coexist on the wall without being equally visible in the interior.

For $f(\phi) = 2 + 3\cos(2\phi)$ V, the interior is

$$2 + 3(r/a)^2 \cos(2\phi) = 2 + 3(x^2 - y^2)/a^2.$$

The potential has a saddle at the centre, with zero field there and opposite Cartesian curvatures. A scan along a single diameter could show a conspicuous maximum without detecting any charge. The two-dimensional Laplacian is zero because the curvatures cancel. This is also a useful finite-difference benchmark: a centred Cartesian Laplacian differentiates this quadratic exactly in the interior when the circular boundary has been represented correctly.

End effects produce a Sturm–Liouville problem

Now ground the side of a circular tube, excite its end at $z = 0$ and look for fields decaying into $z > 0$. For an axisymmetric source-free potential $V = R(r)Z(z)$, choose $Z = e^{-kz}$ for decay. Laplace's equation becomes

$$(rR')' + k^2 rR = 0.$$

The solutions are $J_0(kr)$ and $Y_0(kr)$. The former is regular with value one at zero; the latter diverges logarithmically. Grounding at $r = a$ requires $J_0(ka) = 0$, so $k_n = j_n/a$ with positive roots $j_1 = 2.40483,\ j_2 = 5.52008, \ldots$. Unlike sine wavenumbers they are not evenly spaced at low index.

Multiplying two such eigen-equations by the opposite radial function, subtracting and integrating gives $(k_m^2 - k_n^2) \int r R_m R_n \, dr$ equal to the endpoint term $r(R_n' R_m - R_m' R_n)$. It vanishes at a grounded wall and at a smooth axis. Hence modes at distinct eigenvalues are orthogonal with weight r. This weight arises physically from the circumference of a radial ring. Normalize dimensionless radius $x = r/a$. For a zero j_n, the squared modal norm is $J_1(j_n)^2/2$. An end profile $f(x)$ therefore gives

$$V(r, z) = \sum_n c_n J_0(j_n r/a) e^{-j_n z/a},$$

$$c_n = \frac{2}{J_1(j_n)^2} \int_0^1 x f(x) J_0(j_n x) \, dx.$$

For $f = V_0$ across the cap, $\int_0^1 x J_0(jx) dx = J_1(j)/j$ by differentiating $xJ_1(jx)$. Then $c_n = 2V_0/[j_n J_1(j_n)]$, with alternating signs. The cap voltage jumps at the wall corner, making its near-rim reconstruction slow. A smooth end pattern that already vanishes at a yields faster coefficient decay. In either case a narrow radial strip at the axis should not be given as much integration weight as an equal width near the circumference.

Two ways to see which axial scale survives

Far enough from the cap the first excited nonzero coefficient dominates. Its length is

$$a/j_1 = 0.416a.$$

Ratios of distant on-axis voltages can then identify tube radius without absolute calibration of the cap voltage. For $a = 15.0$ mm, an added 5.0 mm of distance should give factor $e^{-2.40483/3} = 0.449$. Repeat this estimate at several starting positions. The ratio only approaches that value once competing modes have decayed. A datum deep enough to suppress all higher modes may also be close to the instrument noise level, which biases a fitted decay length.

Instead impose an axial sinusoid e^{ikz} on a very long cylinder's side wall. The radial equation now has the opposite sign, $(rR')' - k^2 rR = 0$, whose regular solution is the modified Bessel function $I_0(kr)$. The potential is $V_0 I_0(kr)\cos(kz)/I_0(ka)$. The experiment selects a real oscillation along z, not an exponential attenuation away from an end. Near the axis $I_0(\xi) = 1 + \xi^2/4 + \cdots$; the field generated by a long-wavelength side modulation is nearly uniform across a narrow cross section. At large ka, $I_0(ka)$ grows exponentially and the axis scarcely senses a short-wavelength wall pattern. Calling every radial solution simply a Bessel mode without tracking the separation sign can exchange these two physical predictions.

Construct a grid that treats the axis as a regular point

A conservative radial discretisation starts from $\int_{r_-}^{r_+} r\,\nabla_r^2 V\,dr = [rV']_{r_-}^{r_+}$. Use annular cells and fluxes at shared ring faces. Near $r = 0$ the left face has zero circumference, requiring no unknown point beyond the axis. For a nodal grid with Taylor expansion $V(r) = V(0) + Ar^2 + O(r^4)$, the radial operator at zero is $4A$, so $\nabla_r^2 V(0) \simeq 4[V(h) - V(0)]/h^2$. Simply setting $1/r = 0$ and using the Cartesian one-dimensional second derivative misses a factor of two.

Compare a numerical response to the analytic J_0 mode at a fixed z/a. Test the wall value, axis derivative, radial energy norm and distant decay rate. These are independent observables. A discretisation can reproduce the slowest length while misrepresenting an edge profile composed of many faster modes. Conversely, matching a discontinuous cap perfectly at mesh nodes does not demonstrate that high-order interior amplitudes are accurate.

Can a coaxial dielectric be graded to produce a uniform field?

The homogeneous coaxial field grows as $1/r$ and reaches its highest magnitude at the inner wire. A designer proposes grading the dielectric radially so the full annular gap experiences one common radial field. Assume both metal faces are equipotentials, the scalar permittivity $\epsilon(r)$ may vary continuously, and there is no free volume charge. Cylindrical symmetry still gives $D_r = \lambda/(2\pi r)$, independent of how the material is graded. A constant electric-field magnitude E_g requires $\epsilon(r) = \lambda/(2\pi E_g r)$, that is, permittivity proportional to $1/r$. This construction replaces a uniform material by a known spatial distribution; applying $\epsilon(r)\nabla^2 V = 0$ would miss the material gradient and give the wrong field.

Write $\epsilon(r) = \epsilon_a a/r$. At applied voltage V_0, the constant field is $E_g = V_0/(b - a)$, and charge per length is

$$\lambda = 2\pi\epsilon_a a V_0/(b - a).$$

Thus $C' = 2\pi\epsilon_a a/(b-a)$. For $a = 2.0$ mm, $b = 6.0$ mm, $\epsilon_a = 6\epsilon_0$, the outer relative permittivity is 2, and $C' = 167$ pF/m. At 12 V every radius experiences 3000 V/m. For a homogeneous dielectric at the same voltage the inner surface field would be $12/[0.002 \ln 3] = 5461$ V/m regardless of its constant permittivity. The gradient reduces peak field at fixed voltage. Whether a real material benefits also depends on its local field limit, losses and interfaces.

Where is the polarisation charge that makes this possible? In the dielectric the true total charge satisfies

$$\rho_{\text{tot}} = \epsilon_0 \nabla \cdot \mathbf{E} = \epsilon_0 E_g / r.$$

All of it is bound since $\rho_f = 0$. Direct evaluation from $\mathbf{P} = (\epsilon(r) - \epsilon_0) E_g \hat{\mathbf{r}}$ gives $-r^{-1}\partial_r(rP_r) = +\epsilon_0 E_g / r$, the same result. Material surfaces add bound charges $P_r(b)$ on the outer radial face and $-P_r(a)$ on the inner face. Integrating their surface contributions and the volume charge over a full annulus returns zero net bound charge, as required for a polarised neutral body. A constant radial electric magnitude can therefore coexist with zero free charge and nonzero total charge.

A manufactured graded material is often approximated by concentric homogeneous shells. For shell j from r_j to r_{j+1}, its voltage drop at fixed λ is $\lambda \ln(r_{j+1}/r_j)/(2\pi\epsilon_j)$. To reproduce the exact voltage drop of the continuously graded design over each shell, choose $\epsilon_j = \epsilon_a a \ln(r_{j+1}/r_j)/(r_{j+1} - r_j)$. This matches electrode capacitance even with relatively few shells. Yet the field inside each discrete shell still falls as $1/r$ and jumps at its interfaces. Matching terminal capacitance alone does not verify the target uniform-field profile. Compute the field just inside each shell edge and refine the radial layering against a local peak-field tolerance.

What a small displacement of the end electrode excites

An axisymmetric tube calculation retains only angular order $m = 0$. Now impose at the tube entrance a voltage profile proportional to $r\cos\phi$ near the axis and taper it to zero at the wall. This is a dipolar lateral pattern with positive voltage on one side and negative on the other. Separating $R(r)\cos(m\phi)e^{-kz}$ into Laplace's equation yields

$$(rR')' + \left(k^2 r - \frac{m^2}{r}\right) R = 0.$$

The regular solutions are $J_m(kr)$. Near the axis their leading powers are $r^{|m|}$, so an $m = 1$ potential vanishes on the centreline while its transverse gradient generally does not. The grounded wall selects zeros of J_m. The first zero of J_1 is 3.83171, larger than the first axisymmetric zero, so the slowest lateral disturbance decays faster than the slowest axisymmetric voltage. Probe design therefore changes which decay laws are experimentally accessible.

For a cap shaped exactly as $f = V_1 J_1(jr/a)\cos\phi$, $j = 3.83171$, the whole solution is a single mode. Near the axis $J_1(jr/a) \simeq jr/(2a)$; using $r\cos\phi = x$,

$$V \simeq V_1 j x e^{-jz/a}/(2a).$$

Hence $E_x(0,z) = -V_1 j e^{-jz/a}/(2a)$. With $V_1 = 2.0$ V and $a = 10$ mm, the lateral field starts at 383 V/m and falls to 8.30 V/m after an axial distance equal to one radius. A centreline voltmeter sees zero at both positions and gives no hint of the finite transverse force. A pair of voltmeters at small opposite x or a field probe oriented along x measures the relevant derivative.

For a realistic shifted electrode the input can contain both $m = 0$ and $m = 1$. Place detectors at $(r, 0, z)$ and (r, π, z). Their average selects even azimuthal orders; their difference selects odd ones. The ratio of the difference at two heights tests $m = 1$ decay after higher odd modes subside. The average also includes $m = 2, 4, \ldots$, so attributing it to $m = 0$ requires symmetry information or more angular sensors. Reversing the applied electrode displacement changes the sign of an odd response but leaves even corrections unchanged to leading order. This experiment links angular Fourier selection and radial eigenvalues to a measurable alignment tolerance without requiring an entire three-dimensional voltage map.

Project a ring source without calling an axis cell a point detector

Deposit a thin circular ring of total free charge q at radius s in the plane $z = z_0$ inside a long grounded cylindrical tube. Its volume density in an axisymmetric model is $q\,\delta(r - s)\delta(z - z_0)/(2\pi s)$, for $s > 0$. The $2\pi s$ removes the circumference introduced by the volume element. A radial code that deposits q directly as a density in one mesh bin, without dividing by annular volume, changes the total charge each time the radius or mesh spacing changes. In this calculation one can track that normalisation all the way to a voltage coefficient.

Take the tube infinite in both axial directions for simplicity. Expand $V = \sum_n v_n(z) J_0(j_n r/a)$ in the grounded radial eigenfunctions. Project Poisson's equation using weight $r\,dr$ and the norm $a^2 J_1(j_n)^2/2$. The ring produces

$$[-d_z^2 + (j_n/a)^2] v_n = \frac{q J_0(j_n s/a)}{\pi \epsilon a^2 J_1(j_n)^2}\,\delta(z - z_0).$$

The infinite-line Green response for $-d_z^2 + k^2$ is $e^{-k|z-z_0|}/(2k)$, as can be checked by integrating its derivative across z_0. Thus each modal voltage has a finite amplitude and two decaying tails, and the complete field is obtained by summing radial modes. At a detector separated axially from the ring the exponential factors ensure rapid convergence. Exactly on the mathematical ring the local potential diverges logarithmically like that of a thin filament; comparing finite-element peak voltages there requires a finite wire cross section or a specified detector average.

An unexpected and testable consequence concerns the source radius. The coupling to mode n is proportional to $J_0(j_n s/a)$. It becomes exactly zero when the source sits on a radial node of that mode. The second mode has an interior node at $s/a = j_1/j_2 \simeq 0.4357$. A ring there still carries charge and excites many other modes but contributes no second modal coefficient. As an experimental analogy, a weighting potential that changes sign across part of a sensor can receive cancelling contributions from an extended excitation. Here the cancellation is encoded exactly in orthogonality and the pointwise coupling value.

A check independent of the ring-charge normalisation uses reciprocity. Imagine two narrow coaxial source/detector rings of radii s, r, each with the same definition of total charge and voltage measured around its circumference. The modal response contains the symmetric product $J_0(j_n s/a)J_0(j_n r/a)$ multiplied by a factor that depends only on axial separation. Swapping radius and axial position therefore leaves voltage per total charge unchanged. If a numerical table instead looks symmetric per charge density but not per total ring charge, check whether the radius-dependent integration weights were applied on both deposition and sampling.

Another useful test fixes total deposited charge but reduces a resolved annular thickness around s. First compare with an integral of the analytic ring solution across the same finite distribution. Then take the thin-ring limit at a detector away from the deposition region. Taking grid thickness and source thickness to zero together while comparing a self-voltage at the source conflates numerical refinement with a mathematically singular observable. Stating detector location, normalisation and order of limits turns a superficially straightforward ring source into a robust benchmark for cylindrical Poisson solvers.

Measure a radial cylindrical mode

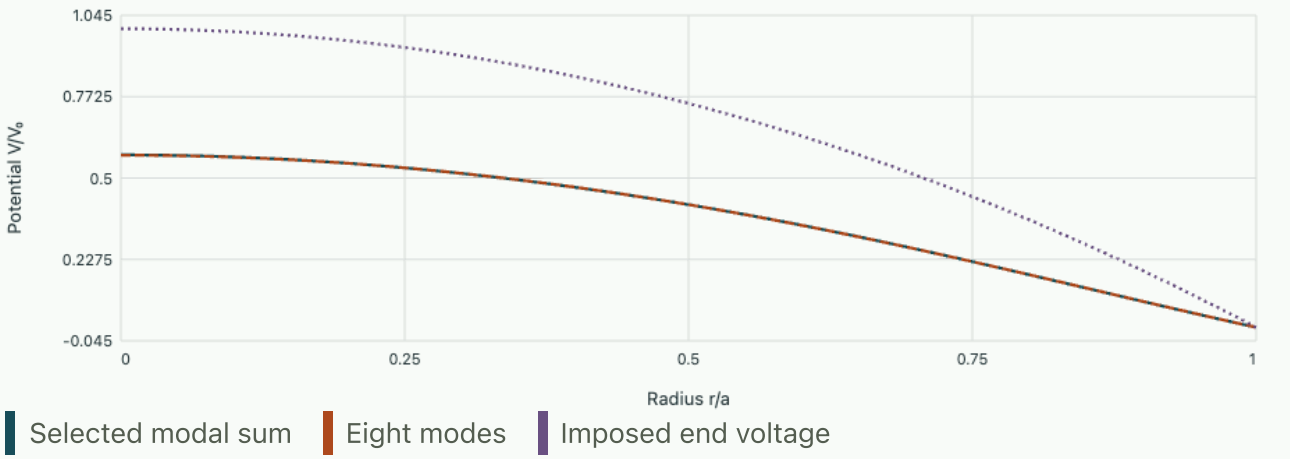
Change the electrode mode and compare radial nodes with axial decay. A cylindrical boundary chooses a different eigenbasis from a rectangular one.

Open this experiment online

Try the calculation

Height z/a **0.25**

Retained modes **3**



The on-axis potential is 0.5774 V_0 with the selected modes, versus 0.5765 V_0 using eight. The slowest mode has e-folding length 0.4158 a.

A cylindrical interface can carry induced bound charge with an entirely radial field

Separate the two charges detected at a coaxial material boundary

A long coaxial gap contains two homogeneous radial layers with interface at s, no free sheet there and fixed inner line charge $+\lambda$. Find the bound interface charge per length, verify it from the field jump, and determine the interface radius at which the layers take equal voltages. Evaluate the result for $a = 2$ mm, $b = 18$ mm and relative permittivities 2 and 4.

▼ **Derive and check the result**

For both regions $D_r = \lambda/(2\pi r)$. In material i the polarisation is $P_i = D_r(1 - \epsilon_0/\epsilon_i)$ along the radial direction. At the interface the outer face of material 1 has bound density $+P_1$, the inner face of material 2 has $-P_2$. Their sum times the circumference gives bound line charge $\lambda_b = \lambda(\epsilon_0/\epsilon_2 - \epsilon_0/\epsilon_1)$. It equals $-\lambda/4$ for the given ordering; the value per length is independent of s because the $1/s$ field cancels the interface circumference. A local density meter would see a different value at a different radius even though the integrated result stays the same.

Direct use of Gauss's law for the true electric field gives $\epsilon_0\, 2\pi s(E_{2r} - E_{1r}) = \lambda[\epsilon_0/\epsilon_2 - \epsilon_0/\epsilon_1]$, confirming the sign and magnitude. A check with displacement alone gives zero jump and thus correctly reports no *free* charge at that same surface. The two field laws answer different questions. Extending the shell calculation to include both exterior material faces returns zero total bound charge per material, so the internal negative interface charge never represents a net creation of charge.

The voltage drops are proportional to $\ln(s/a)/\epsilon_1$ and $\ln(b/s)/\epsilon_2$. Equality gives $s^{\epsilon_1+\epsilon_2} = a^{\epsilon_2} b^{\epsilon_1}$, or $s = a(b/a)^{\epsilon_1/(\epsilon_1+\epsilon_2)}$, a dimensionally transparent form. Numerically $s = 2(9)^{1/3} = 4.16$ mm. A half-voltage location at 4.16 mm is far from the arithmetic centre of this gap. It is also distinct from the 6.0 mm geometric mean for a homogeneous annulus, because the inner lower-permittivity layer needs less geometric thickness to take the same voltage.

Tutorial · radial modes with different physical boundaries

Replace the grounded wall by a reflecting boundary

A long tube has $\partial_r V = 0$ on its side instead of $V = 0$, and its end voltage profile is nonuniform. What happens to the radial eigenvalues and to the assertion that every end disturbance dies away?

▼ **Work through the solution**

The wall now requires

$$J_0'(ka) = -J_1(ka) = 0.$$

In addition a constant radial mode with $k = 0$ satisfies both side and axis conditions. Its axial equation is $Z'' = 0$, so boundedness allows a constant component. The end profile's weighted cross-sectional mean supplies that persistent offset. Only the components orthogonal to the constant decay. Requiring the potential to tend to zero at infinity while prescribing a nonzero mean cap voltage would be incompatible for this insulated side model unless other boundaries or sources are introduced.

Infer a radial charge density from a coaxial reading

Suppose the measured potential in a coaxial gap is $V = A\ln(b/r) + B(b^2 - r^2)$ with constant permittivity. Determine the volume charge and the signs of its contribution to the two face fluxes. Which part can be changed by externally altering the applied voltage at fixed charge density?

▼ Work through the solution

The logarithmic term is harmonic away from the excluded axis. For the quadratic term $\nabla_r^2[-Br^2] = -4B$, giving uniform $\rho = 4\epsilon B$. The radial field is $A/r + 2Br$. The net outward displacement flux through the gap per length is $2\pi\epsilon[bE(b) - aE(a)] = 4\pi\epsilon B(b^2 - a^2)$, exactly $\rho\pi(b^2 - a^2)$. At positive B the quadratic term contributes positive outer flux and negative inward-face flux, with the larger outer contribution from both area and field. Changing only the source-free boundary excitation adds a logarithmic term and, if the reference permits it, a constant. It alters both field readings without altering the inferred uniform volume density.

Further reading

- Reconstruct a radial electrode profile with one to eight Bessel modes
- Fließbach, chapter 10 · cylindrical problems

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 requires its own 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.

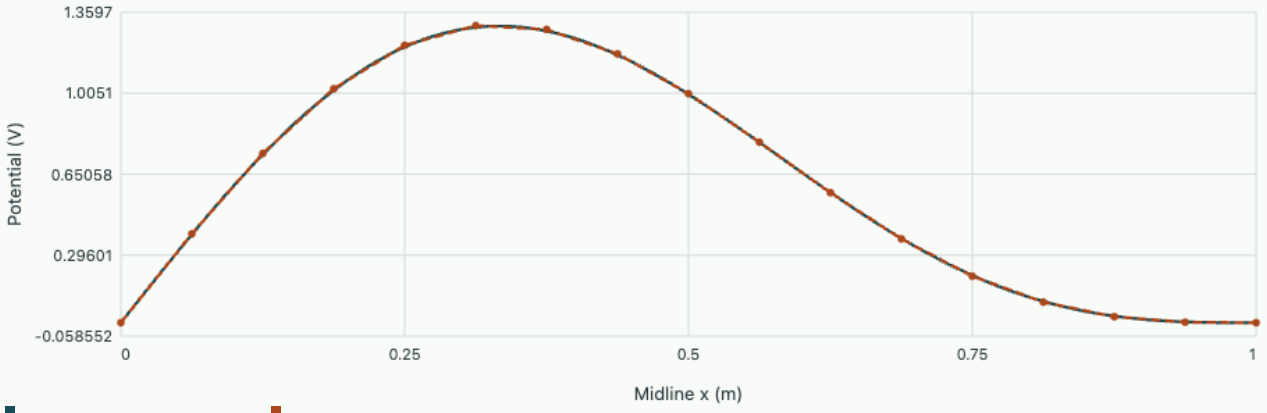
Run the full field and compare genuinely different observations

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.

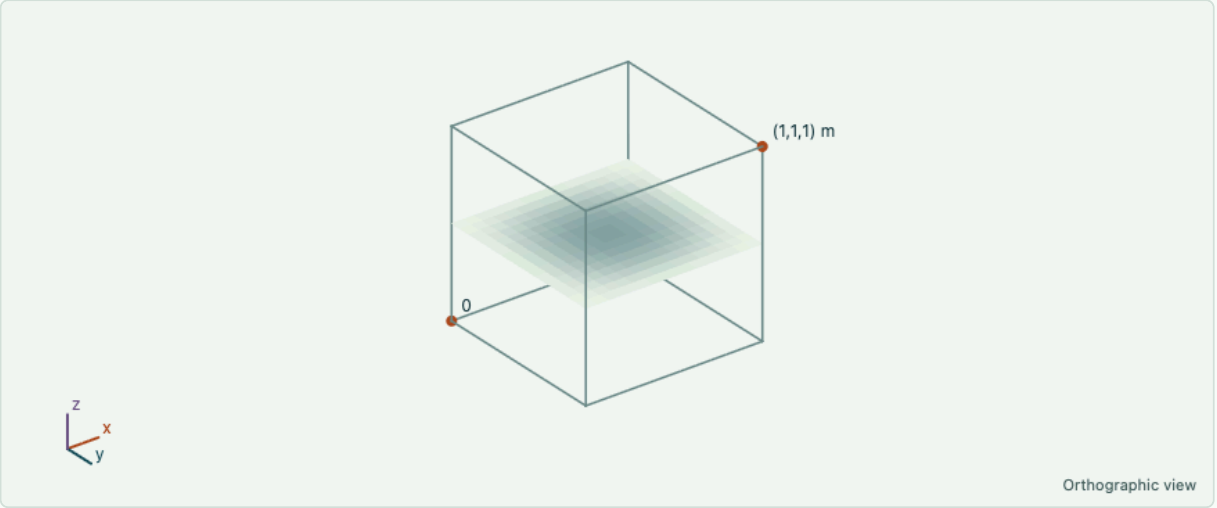
Try it · Rotate a 3D field, move its slice and refine its mesh

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.

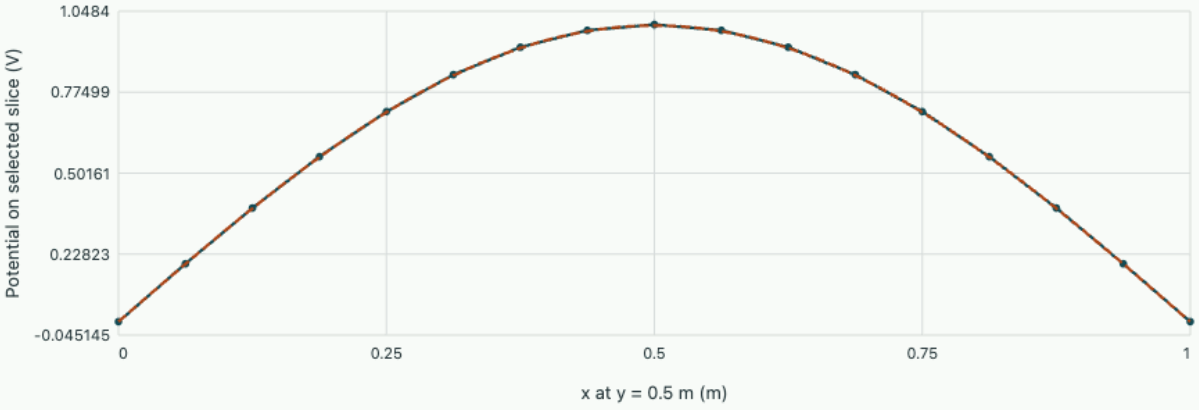
Open the live 3D experiment

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

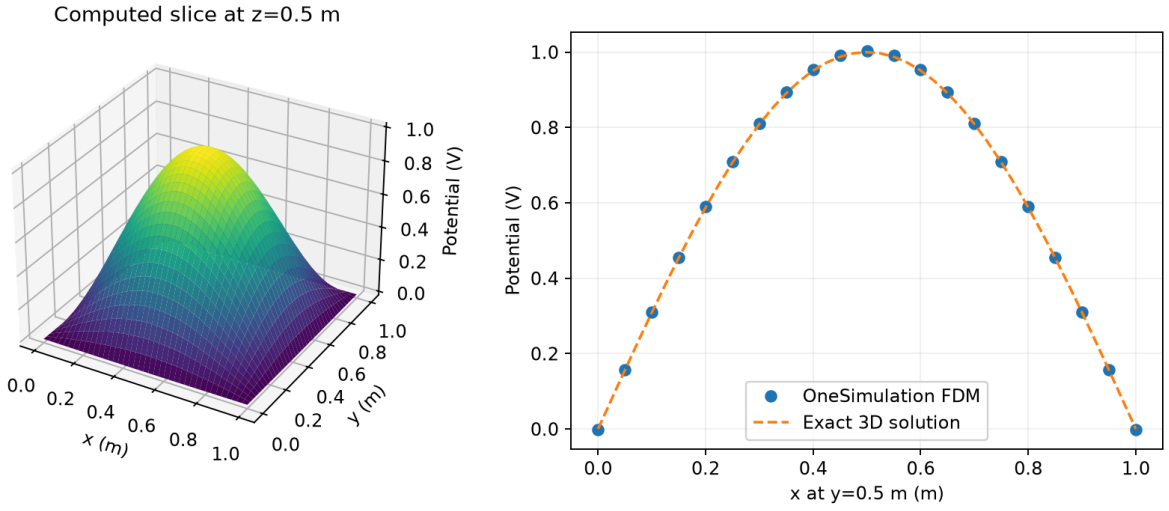
The notebook assembles and solves the full 3D OneSimulation finite-difference model. It reports an RMS error at cell centres; the browser reports a maximum error at nodes for its stated seven-point operator. These are different diagnostics. Compare each against its own continuous reference and repeat 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 the capacitance of a continuously graded dielectric

Numerical field laboratory

Find the exact solution of a variable-coefficient field equation, then estimate the same capacitance from charge, voltage drop and electrostatic energy.

Build the problem from what is continuous

Two very large parallel plates are separated by $L = 1$ m. Set their potentials to 0 and 1 V. Edge effects are negligible in the observation region. The relative permittivity of the material varies as $\epsilon_r(x) = 1 + gx/L$, where $g \geq 0$. There is no free volume charge. A common first guess is a linear potential between the plates, because the boundary voltages impose an average field. Test that guess using the local field law.

For a thin pillbox whose faces are perpendicular to x, zero enclosed free charge means equal normal displacement at both faces. Let $D_* > 0$ denote the magnitude of the constant displacement. Since potential increases with x, the electric field points towards negative x and $D_x = -D_*$. Hence $V' = D_*/[\epsilon_0(1 + gx/L)]$. Regions of low permittivity must sustain a larger electric field and account for a larger voltage drop per metre. We have a local constraint before solving a single matrix equation.

Obtain the voltage and both expressions for capacitance

Integrate V' from left to right and impose $V(L) - V(0) = \Delta V$. For $g > 0$ this gives

$$\frac{C}{A} = \frac{D_*}{\Delta V} = \frac{\epsilon_0}{L}\frac{g}{\log(1+g)},$$

$$\frac{V(x)}{\Delta V} = \frac{\log(1+gx/L)}{\log(1+g)}.$$

As g approaches zero, both expressions have removable singularities. Since $\log(1+g) = g - g^2/2 + O(g^3)$, the effective relative permittivity approaches $1 + g/2 + O(g^2)$. An arithmetic average can therefore look convincing at very small gradients despite having the wrong general derivation. For computation near zero, use a logarithm specifically evaluated as $\log(1+g)$ with a small-argument routine such as `log1p`.

Check the result through stored energy. Each volume element contributes $\epsilon E^2/2$, so

$$U/A = (D_*^2/2)\int_0^L dx/\epsilon(x).$$

Substituting the voltage relation gives

$$U/A = D_*\Delta V/2.$$

The same capacitance follows from $C/A = 2U/(A\Delta V^2)$. This equality requires a linear, lossless dielectric and a reversible charging path. A material whose permittivity changes with field needs the corresponding integral over the constitutive curve.

Work one case numerically before assembling the grid

At $g = 3$, the material goes from relative permittivity 1 to 4. The effective value is $3/\log 4 \simeq 2.1640$, smaller than the arithmetic mean 2.5. The midplane voltage is $\log 2.5/\log 4 \simeq 0.6610$ V, not 0.5 V. The electric field at the left is four times as large as at the right. A measurement of the midpoint voltage alone is enough to reject a uniform-field model if the plates and material are well characterised.

As a dimensional check, $C/A \simeq 19.16$ pF/m² for $L = 1$ m. For a geometrically similar layer one millimetre thick the same effective permittivity gives 19.16 nF/m². The voltage profile versus x/L stays fixed when g and the plate voltages stay fixed, while the field and free surface charge increase in inverse proportion to separation.

Assemble by balancing fluxes before dividing by a coefficient

Divide the gap into n cells. Across cell i approximate the displacement magnitude by $D_i = \epsilon_{i+1/2}(V_{i+1} - V_i)/h$. Enforcing $D_i = D_{i-1}$ at each internal node gives a tridiagonal linear system. An independent route views each cell as a capacitor with inverse capacitance per area $h/\epsilon_{i+1/2}$. These series contributions add, giving $D_*^{(h)} = \Delta V/\sum h/\epsilon_{i+1/2}$. Cumulative sums of the same drops then recover the entire grid potential. Both routes implement identical physics and the same discrete constitutive assumption.

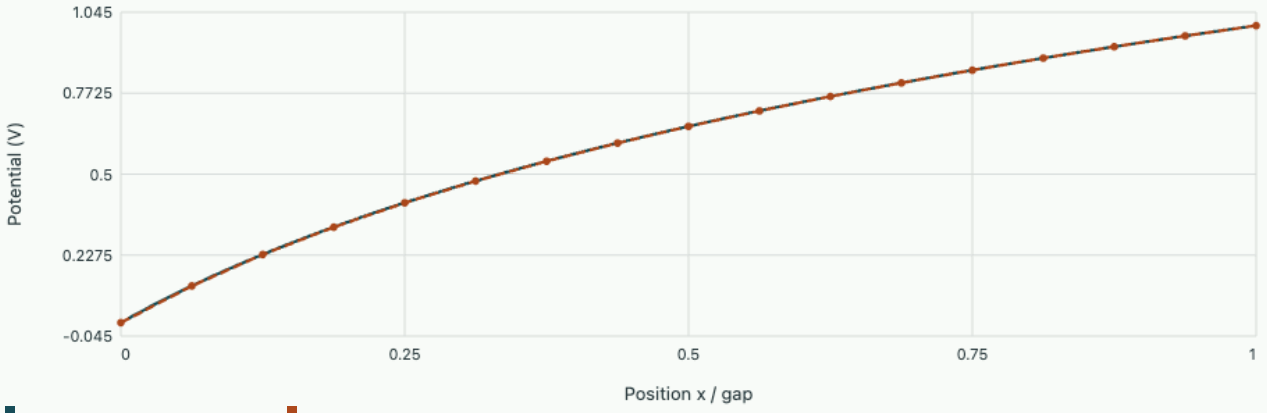
The browser uses this cumulative series construction. The OneSimulation input instead supplies the spatial coefficient $1 + gx/L$, zero source and the two exterior potential values to the coefficient-form PDE solver. Their agreement checks assembly against a distinct algorithm. In this example the integral of inverse permittivity also tells us what the midpoint discretisation approximates. Since $1/(1 + gx/L)$ is convex, midpoint quadrature underestimates its integral and therefore overestimates capacitance. This predicts the bias before looking at a convergence table.

Open the live experiment

Try the calculation

Cells across gap 16

Permittivity increase g 3



Analytical voltage Flux balance

Effective relative permittivity 2.1662 on the grid and 2.164 analytically. Midpoint voltage 0.66073 V.

Download the OneSimulation model

In your OneSimulation environment, run the model with these inputs.

```
python field-laboratory.py capacitor --cells 16 --contrast 3
```

Compare a local field measurement with an integral quantity

In the experiment, plot V across the dielectric and record the effective relative permittivity. At $g = 2$, $n = 16$, the computed potential at $x = 0.1875$ m is about 0.28974 V; the analytical reference is about 0.28987 V. The potential there is low while the computed capacitance is high. A voltage error at an arbitrary coordinate need not share the sign of the error in an integral quantity.

Now estimate the left electric field from the first-cell drop and from the analytical derivative at $x = 0$. The first-cell drop divided by h is an average across that cell, representing its midpoint to second order for a smooth field. Interpreting it as a value exactly at the wall introduces an $O(h)$ offset. Label sample positions when comparing a derivative against an analytic curve. The grid solution can converge quadratically in voltage while a carelessly placed flux diagnostic appears first order.

Tutorial

The gradient is reversed so $\epsilon_r = 1 + g(1 - x/L)$, but the plates remain at the same voltages. Which earlier observations change?

▼ **Work through the solution**

The integral of inverse permittivity is invariant under $x \mapsto L - x$, so the capacitance and stored energy stay the same. The new potential obeys $V_{\mathrm{rev}}(x) = \Delta V - V(L - x)$. The largest field is now at the right wall, and the midpoint voltage at $g = 3$ falls from about 0.661 to 0.339 V.

A sensor measures an apparently linear voltage profile and C/A greater than the vacuum value. Is the inferred dielectric necessarily uniform?

▼ **Work through the solution**

Under the one-dimensional, source-free and linear assumptions, exactly constant nonzero field and constant normal displacement require constant permittivity. A sparse set of voltage probes may simply undersample a graded profile. Lateral variation, free volume charge or uncertain boundary voltages would break the inference.

Derive bounds on the effective permittivity if only its material range $\epsilon_{\min}$ to $\epsilon_{\max}$ is trusted.

▼ **Work through the solution**

Integrate $1/\epsilon_{\max} \le 1/\epsilon(x) \le 1/\epsilon_{\min}$, multiply by $1/L$, and invert the positive quantities. The effective value lies between these two limits. A simulation reporting an effective permittivity outside that interval in this passive one-dimensional model fails this check regardless of mesh resolution.

Is a computed screening layer resolved?

Numerical field laboratory

Use a measured decay length to select a mesh, derive the exact discrete penetration constant and decide whether a discrepancy belongs to the physical screening law or the numerical grid.

Derive the screened field and state its range

A dilute sample adjacent to an electrode can often be approximated, for small applied potentials, by a relation $\rho(x) = -\epsilon V(x)/\lambda^2$. It says that an excess positive potential attracts net negative charge. The parameter λ is independently calibrated for the ambient material. Combining this relation with Poisson's equation yields

$$V'' = V/\lambda^2.$$

A plus sign is essential: the physical disturbance decays into the sample. Changing the sign would turn a decaying layer into an oscillatory boundary-value problem.

Set $V(0) = V_s$ and $V(L) = 0$. Scale the voltage by V_s, position by L, and write $g = L/\lambda$. The dimensionless problem is

$$-u'' + g^2 u = 0,\ u(0) = 1,\ u(1) = 0.$$

The program solves for unit boundary voltage as a normalisation. For a dilute monovalent electrolyte at room temperature the linearised microscopic model requires the physical $|V_s|$ to be small compared with $k_B T/e$, roughly 26 mV. A computed unit-amplitude solution can represent a small physical voltage by multiplication. Applying 1 physical volt to the linearised electrolyte would generally exceed this assumption.

Solve the finite slab before taking the half-space limit

The general solution contains e^{gx} and e^{-gx}. Impose the right voltage first, using $\sinh[g(1-x)]$, then divide by $\sinh g$ to set the left voltage. Thus $u = \sinh[g(1-x)]/\sinh g$. For $g = 0$ take the limit to get $u = 1 - x$. For a thick slab and fixed physical distance from the left, use $e^{-gx}(1 - e^{-2g(1-x)})/(1 - e^{-2g})$. Close to the right wall the subtracted exponential is needed to satisfy grounding. In the interior many screening lengths away, even a big relative error can correspond to a negligible physical voltage.

At $g = 4$, the midpoint voltage fraction is $1/(2\cosh 2) \simeq 0.1329$. The half-space estimate is

$$e^{-2} = 0.1353.$$

A sample only four screening lengths across is already close to the half-space answer at its middle, though its right edge still has exactly zero voltage rather than e^{-4}. Determine where your instrument can distinguish the profiles before treating one as a better physical model.

Obtain the grid behaviour without running a solver

With grid spacing $h = 1/n$, the interior stencil is

$$-u_{i-1} + (2 + g^2 h^2) u_i - u_{i+1} = 0.$$

Try a geometric progression $u_i = e^{-\gamma_h i h}$. Divide through by the trial value. The sum of its neighbours gives

$$2\cosh(\gamma_h h) = 2 + (g h)^2.$$

Since $\cosh(2z) = 1 + 2\sinh^2 z$, the effective decay constant is

$$\gamma_h = \frac{2}{h}\operatorname{asinh}(g h/2) = g - \frac{g^3 h^2}{24} + O(g^5 h^4).$$

Thus the numerical layer is slightly too broad for a coarse mesh. The relevant resolution is $g h = h/\lambda$ in consistent units, not the cell count alone. Keeping n fixed while increasing g makes a visually unchanged mesh physically coarser. The finite-slab grid reference follows by replacing g with γ_h in the hyperbolic-sine ratio at the nodes.

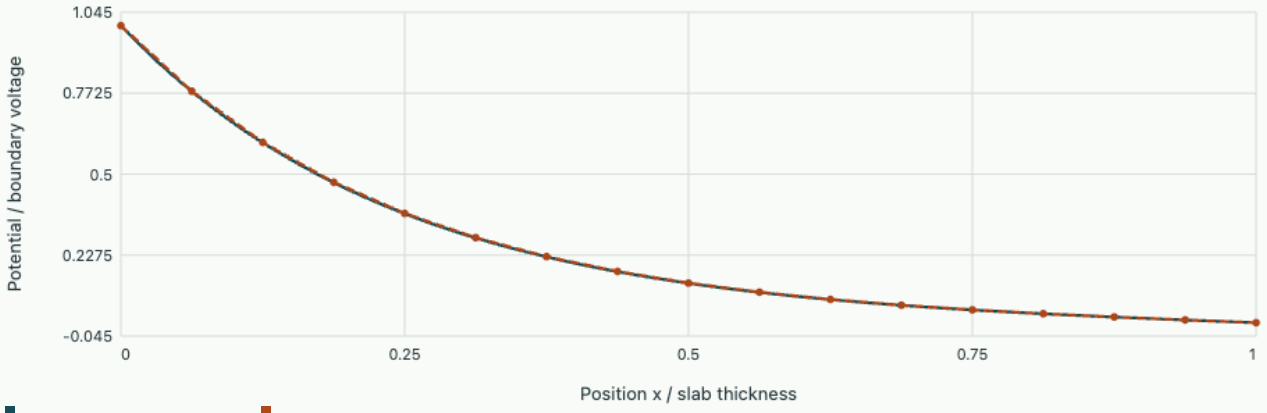
Separate the physical and computational experiments

Open the live experiment

Try the calculation

Cells across slab **16**

Gap / screening length **4**



Continuum slab Solved grid values

Midpoint 0.13357, exact 0.1329. Discrete decay constant × thickness = 3.9897.

Download the OneSimulation model

In your OneSimulation environment, run the model with these inputs.

```
python field-laboratory.py screen --cells 16 --contrast 4
```

The browser solves the tridiagonal boundary equations by forward elimination and backward substitution. The downloadable OneSimulation model supplies an absorption term g^2u and the same two Dirichlet values. For $g = 2,\ n = 16$, the potential fraction at $x = 0.1875$ is about 0.67311 on the grid and 0.67297 in the continuum. Reverse the grid spacing in the analytical expression for γ_h and independently predict whether doubling the number of cells raises or lowers the estimate there.

A typical measurement infers a length from the logarithmic slope,

$$1/\lambda_{\mathrm{fit}} = -d\log V/dx.$$

For two nearby positive voltages use the difference of their logarithms divided by probe separation. As V approaches the instrument floor, small absolute noise has a large effect on this ratio. At the grounded right wall the logarithm does not exist. Exclude that point and justify the usable region using both boundary physics and noise. Refining the field mesh cannot improve a voltage measurement below its resolution.

Tutorial

A grid has 16 cells and $g = 16$. Predict the fractional error in the effective decay rate, using the exact discrete expression rather than only the expansion.

▼ **Work through the solution**

Here $gh = 1$, so

$$\gamma_h / g = 2\,\mathrm{asinh}(0.5) \simeq 0.9624.$$

The numerical decay is slow by about 3.76%. The quadratic estimate gives 4.17%, reasonable but visibly imperfect at one cell per screening length.

For a desired relative decay-rate error below 0.1%, estimate the largest acceptable gh in the small-step regime.

▼ **Work through the solution**

The leading relative error is $(gh)^2/24$. Requiring it below 0.001 gives $gh < 0.155$, approximately 6.5 cells per screening length. This condition concerns the exponential rate. The derivative at the electrode may need a separate boundary stencil and its own check.

A calibration reports a longer decay length as electrode voltage increases, even after mesh refinement. Can the numerical dispersion formula alone explain this trend?

▼ **Work through the solution**

In the linear model the normalised profile is independent of the voltage amplitude. The discretisation error depends on g and the mesh. At fixed g, n it cannot create an amplitude-dependent normalised length. Recheck the linear-response regime, electrode boundary condition and instrument response before fitting a new λ.

What can a mesh measure about a point charge?

Numerical field laboratory

Distinguish conserved charge and measurable exterior fields from a numerical peak that diverges as the source becomes more localised.

Ask three detectors different questions

Imagine three instruments looking at a small charged body in vacuum. One measures the voltage on a spherical probe outside it, one measures the integrated electric flux through a closed surface, and one estimates the largest electric field somewhere near the source. If the body is idealised as a point, which readings have finite limits? The distinction matters when a calculation displays a dramatic change in maximum field after a perfectly reasonable mesh refinement.

For a positive point charge q, the exterior potential $V = kq/r$ tends to infinity as the observation point approaches the source. Its radial field magnitude kq/r^2 also diverges. Gauss's law gives a different answer for the flux through any sphere of positive radius:

$$4\pi r^2 E = q/\epsilon_0.$$

A finite integral and an unbounded maximum are compatible. Knowing what the instrument measures determines which numerical output should be expected to converge.

Use a smooth source with a known answer

Replace the potential by $V_a = kq/\sqrt{r^2 + a^2}$, with $a > 0$ a distance. This is a particular physical regularisation. Differentiating before sampling gives

$$E_a = kqr/(r^2 + a^2)^{3/2}.$$

Divide the flux by q/ϵ_0. The putative enclosed-charge fraction becomes

$$F_a(r) = \frac{r^3}{(r^2 + a^2)^{3/2}}.$$

To recover the corresponding density, differentiate $qF_a(r)$ with respect to radius and divide by shell area. The terms involving r^4 cancel, leaving $\rho_a(r) = 3qa^2/[4\pi(r^2 + a^2)^{5/2}]$. This density is smooth, nonnegative and integrates over all space to q, since $F_a(0) = 0$ and $F_a(\infty) = 1$. It has an extended tail. The parameter a is a smoothing scale rather than an outer radius where charge abruptly stops.

At $r = a$ only $1/(2\sqrt{2})$, about 35.36%, of the charge is enclosed. At $r = 2a$ the fraction is about 71.55%. A common intuition that two smoothing distances already contain essentially everything fails for this distribution. Measure the probe radius in units of the source size before blaming the flux calculation.

Hold the observable fixed while refining

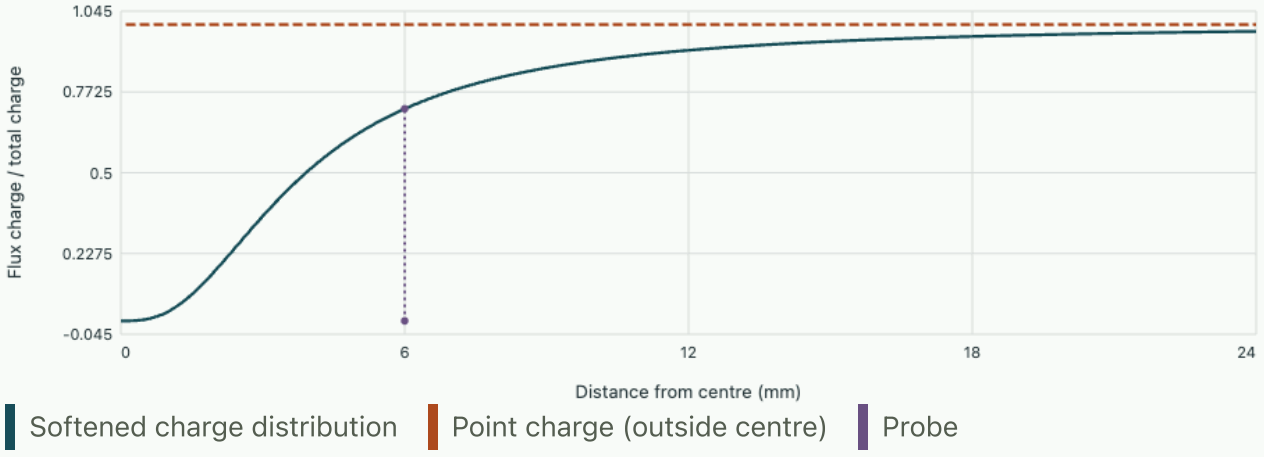
The exterior-field ratio relative to the true point charge is $[1 + (a/r)^2]^{-3/2}$. At fixed physical r, expanding the bracket gives $1 - \frac{3}{2}(a/r)^2 + O((a/r)^4)$. Thus a probe far enough away sees a second-order regularisation error. Locating the maximum of E_a, however, requires setting $a^2 - 2r^2 = 0$, giving $r = a/\sqrt{2}$ and a peak proportional to $1/a^2$. Refine a and move the observer with the maximum: the reported peak grows even as a fixed exterior probe reading converges.

Open the live experiment

Try the calculation

Charge-cloud radius a (mm) **3**

Probe radius r (mm) **6**



The probe encloses 71.554% of the charge. Reducing a by half gives 91.308% at the same physical probe.

Keep the probe at 6 mm and reduce a from 3 to 1.5 mm. The flux rises from about 71.55% to 91.31% of the point-charge value. Now instead halve both the probe radius and a. The measured fraction stays the same. The two procedures have distinct limiting observations; writing “halve the mesh” without recording the source representation and probe definition would make the experiment irreproducible.

Use continuity as a numerical diagnostic

Suppose a code deposits a prescribed charge into nearby cells and then solves for a nodal potential. Summing deposited cell charges tests normalisation of the source. Summing outward face fluxes over a larger box tests a separate constraint on the computed field. A correct finite-volume implementation cancels interior face pairs. Interpolating a nodal field to a smooth plotting surface with an unrelated quadrature need not conserve the same discrete quantity. A box-flux test using the solver's own numerical face fluxes and a smooth-surface test using the reconstructed field answer complementary questions.

A second check concerns energy. For a true point charge the field-energy integral near the origin scales as $\int_0 r^2 (r^{-2})^2 dr$, which diverges. For a body of radius a this short-distance contribution typically scales like $1/a$. Comparing two source regularisations through their total self-energy without specifying a physical charge radius is therefore an ill-defined validation exercise. Interaction energy with a smooth distant field, in contrast, can have a well-defined point-particle limit.

Tutorial

How far from this smooth charge should a flux probe be placed to enclose 99%?

▼ Work through the solution

Solve $[1 + (a/r)^2]^{-3/2} = 0.99$. This gives

$$r/a = [0.99^{-2/3} - 1]^{-1/2} \simeq 12.2.$$

The answer depends on the chosen source shape. A uniform finite ball would have a different exterior behaviour.

A code uses $a = 2h$ and reports the field one cell away, $r = h$, on every refined mesh. What do you expect for its flux fraction and pointwise field?

▼ Work through the solution

The fixed ratio $a/r = 2$ gives $F = 5^{-3/2} \simeq 0.0894$ at every resolution. The field still scales as r^{-2} times that factor and grows like h^{-2}. The observer is moving towards the singular source.

If total deposited charge differs from q by 2% independently of mesh size, can adjusting a cure the exterior result?

▼ Work through the solution

At large radius the flux follows the total charge that was actually deposited. Changing width cannot repair an incorrect integral. Fix the source normalisation and cell-volume weights before interpreting a convergence slope.

Heat, momentum and wave transport

Test the same spatial method on diffusion, a propagating pulse and an elastic mode. Compare dissipation, dispersion and boundary residuals on a common footing.

1. How a moving wall sends momentum into a liquid
2. Critical depth and the hydraulic jump
3. Thermal diffusion and transient surface cooling
4. A moving boundary in a freezing layer
5. How a longitudinal pulse reflects from a change in rod section
6. Why the frequencies of a bending beam are not equally spaced
7. How a magnetic field relaxes into a finite conductor
8. Posing an electromagnetic evolution problem
9. Measure the phase speed of a numerical electromagnetic wave
10. Action density, stress waves and the boundary that reflects them
11. From binary collisions to the Boltzmann equation
12. Transport coefficients from equilibrium fluctuations

How a moving wall sends momentum into a liquid

Fluid mechanics

A plate oscillating horizontally in a deep bath has no average velocity. It nevertheless dissipates energy continuously. A velocity measurement at two heights can reveal the fluid's viscosity.

Keep the unsteady term in a parallel flow

Let y increase upward from a very broad, horizontal plate. The liquid is Newtonian and incompressible with constant kinematic viscosity ν; the plate moves parallel to itself with speed $U\cos\omega t$. Ignore edges, any pressure gradient along the plate and transients left from starting the motion. By symmetry the velocity has a single component $u(y,t)$. A parcel has no vertical speed and meets no streamwise velocity variation, so convective acceleration vanishes even though the motion is time dependent. Horizontal momentum balance on a slice of thickness dy compares its acceleration with the shear-stress difference on its two faces.

$$\rho\,dy\,\partial_t u = \tau(y+dy) - \tau(y),$$

$$\tau = \mu\,\partial_y u,$$

$$\partial_t u = \nu \partial_y^2 u.$$

Velocity is prescribed at the plate by no slip and tends to zero far away. The same equation describes thermal diffusion with temperature in place of velocity. Here the quantity transferred is horizontal momentum, and a velocity gradient sets its transfer rate.

Read decay and phase from one complex root

Represent periodic velocity by the real part of $\widehat{u}(y)e^{i\omega t}$. Inserting it into the equation gives

$$\widehat{u}'' = (i\omega/\nu)\widehat{u}.$$

Since $(1+i)^2 = 2i$, the decaying branch of the solution has an exponent with equal real and imaginary parts.

$$\delta = \sqrt{2\nu/\omega},$$

$$u(y,t) = U e^{-y/\delta}\cos(\omega t - y/\delta).$$

Every increase in height by δ divides the amplitude by e and delays the phase by one radian. This δ is a scale rather than a physical edge. The unbounded solution requires the top and sidewalls to be far enough away for the amplitude there to be negligible. At five penetration depths it is less than one percent of its value at the plate.

Worked example · interpret a submerged velocity probe

In an oil with $\nu = 5.0 \times 10^{-5}$ m²/s, set $f = 2.00$ Hz and $U = 0.080$ m/s. Using $\omega = 2\pi f$ gives $\delta = 2.82$ mm. At 3.00 mm the predicted amplitude is

$$0.080 e^{-3/2.82} = 0.0276\,\mathrm{m/s}.$$

The phase lag is 1.06 rad or 0.0846 s. A synchronized velocity snapshot can therefore show nearly stationary liquid while the wall speed is large. A single instantaneous ratio generally differs from the ratio of fitted amplitudes.

Worked example · power delivered to the wall layer

The wall exerts on the fluid a tangential traction $-\mu u_y(0,t)$. Its complex amplitude is $\mu U(1+i)/\delta$. Only the in-phase part contributes to power averaged over a cycle, since averaging a product of two cosines gives one half of the cosine of their phase difference. Thus

$$\overline{P}/A = \mu U^2/(2\delta).$$

For the same oil at density 900 kg/m³, $\mu = 0.045$ Pa s and the mean power is 0.0510 W/m². The positive value is consistent with viscous dissipation even though mean shear and mean speed separately vanish.

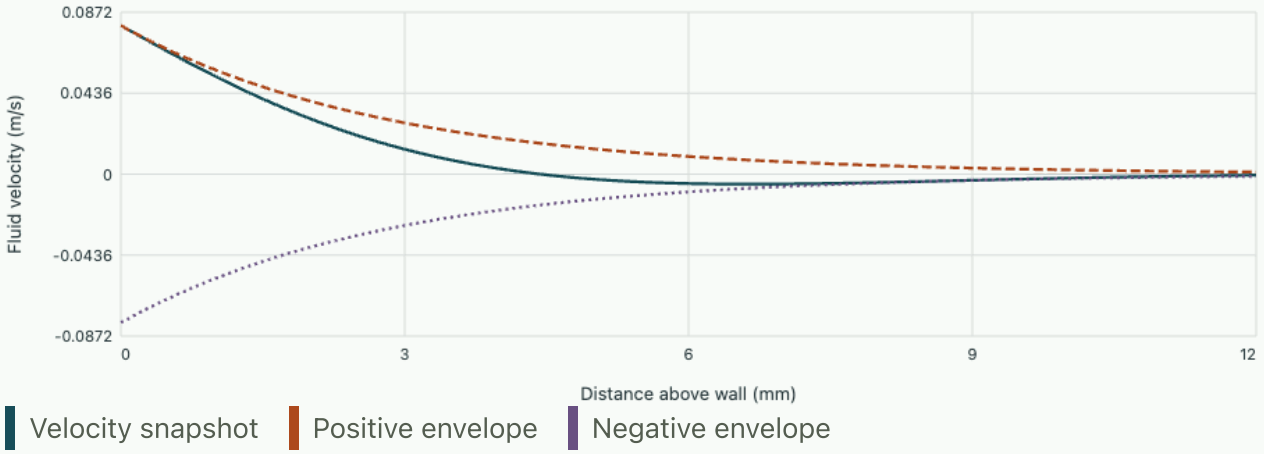
Freeze the phase at a quarter cycle and find the depth at which the velocity first changes sign away from the wall. Sweep the phase to distinguish that moving zero from the stationary amplitude envelope, then double the frequency and predict the new penetration depth.

Open this experiment online

Try the calculation

Wall frequency (Hz) 2

Cycle phase (°) 0



The penetration depth is 2.82 mm. The 3 mm probe has a speed amplitude of 0.0276 m/s and lags by 0.085 s.

A related OneSimulation calculation

The course flat-plate calculation provides a second momentum-diffusion example. Predict what doubling viscosity does to the measured layer thickness at a fixed downstream station, then compare the square-root scaling across the two geometries.

Measure boundary-layer thickness in OneSimulation

Tutorial · estimate a material property without a force sensor

At 1.00 Hz two probes spaced 2.00 mm vertically report fitted amplitudes in the ratio upper/lower = 0.500. Estimate δ, ν, and the phase difference. Identify a measurement that would falsify this model even if the amplitude ratio agreed.

▼ **Use both the magnitude and phase predictions**

The spacing divided by δ is $\ln 2$, giving $\delta = 2.89$ mm. Then $\nu = \omega\delta^2/2 = 2.62 \times 10^{-5}$ m²/s. The upper signal must lag by 0.693 rad, equivalent to 0.110 s. A substantial phase discrepancy suggests confinement, an unremoved transient or motion beyond a one-dimensional shear layer. Fitting both observables is a stronger test than assigning a viscosity to every amplitude pair.

Further reading · MIT 2.25, Stokes' second problem (PDF)

Critical depth and the hydraulic jump

Fluid mechanics

In a wide shallow channel the same discharge can pass as a shallow fast stream or as a deeper slow stream. A gradual hump and an abrupt jump connect depths through different conservation laws.

Minimise the energy needed to carry a discharge

Use a broad rectangular channel and let $q = Q/b$ be its volume flow per unit width. At a section with approximately uniform speed u and water depth h, mass conservation gives $u = q/h$. Take pressure to be hydrostatic over the depth and neglect losses between adjacent smooth sections. Measured relative to the local channel bottom, specific energy in metres of head is

$$E(h) = h + \frac{q^2}{2gh^2},$$

$$\frac{dE}{dh} = 1 - \frac{q^2}{gh^3} = 1 - \mathrm{Fr}^2.$$

The derivative vanishes at $h_c = (q^2/g)^{1/3}$, where $\mathrm{Fr} = u/\sqrt{gh} = 1$. The second derivative $3q^2/(gh^4)$ is positive. Critical depth is the unique minimum, with $E_c = 3h_c/2$. The speed $\sqrt{gh}$ is the speed of small long gravity waves relative to shallow water. When u exceeds it, even a disturbance travelling upstream relative to the liquid is carried downstream relative to the bed.

Background · Froude number

$$\mathrm{Fr} = U/\sqrt{gh}$$

For shallow rectangular flow with gravity waves this compares material speed to wave speed. Critical depth occurs when the two are equal.

Worked example · decide if a smooth rise can pass the flow

A channel carries $q = 0.40$ m²/s at depth 0.50 m. Speed is 0.80 m/s and specific energy is

$$0.50 + 0.80^2/(2g) = 0.5326\,\mathrm{m}.$$

Critical depth is 0.2536 m and the minimum specific energy is 0.3804 m. Since total head includes bed elevation, a local rise of up to about 0.152 m can be accommodated in the ideal fixed-upstream-state model. A higher rise requires the upstream water level or discharge to adjust. For a small rise on the subcritical branch the available specific energy decreases and depth decreases toward the critical value.

Keep momentum across a short rapid transition

A hydraulic jump is turbulent and strongly dissipative. Place control surfaces far enough upstream and downstream for pressure to be hydrostatic again. Across a short horizontal reach, neglect net bed shear compared with the momentum and end-pressure forces. The hydrostatic force on an end section is $\rho g h^2/2$ per unit width, obtained by integrating a triangular pressure profile. Momentum flux is $\rho q^2/h$. Requiring equal momentum functions gives

$$\frac{h_1^2}{2}+\frac{q^2}{g h_1}=\frac{h_2^2}{2}+\frac{q^2}{g h_2}.$$

Factor out h_2-h_1 and take the distinct-depth root. With $r=h_2/h_1$, the remaining relation is

$$r(r+1)=2\mathrm{Fr}_1^2.$$

The positive solution and associated energy drop are

$$r=\frac{\sqrt{1+8\mathrm{Fr}_1^2}-1}{2},$$

$$E_1-E_2=\frac{(h_2-h_1)^3}{4h_1h_2}.$$

The loss is positive for a jump from shallow to deep flow. Reversing the depths would require a gain in mechanical energy without a source, which excludes that continuation of the same algebraic equation.

Worked example · locate the conjugate depth

At a flume section let $h_1=0.100$ m and $u_1=2.00$ m/s. Then $q=0.200$ m²/s and $\mathrm{Fr}_1=2.02$. The depth ratio is 2.40, giving $h_2\simeq 0.240$ m and $u_2\simeq 0.835$ m/s. Upstream energy is 0.304 m and downstream energy about 0.275 m. Momentum balance predicts a head drop near 0.0285 m. Its location along the channel also depends on the imposed tailwater level; the two-section balance itself contains no position variable.

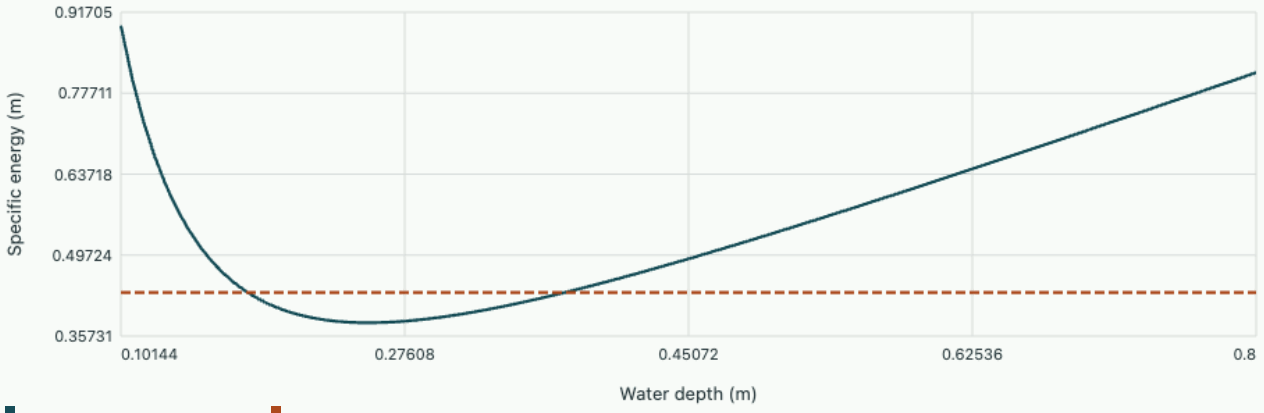
Follow the head over the bump until its horizontal line just touches the energy curve. Estimate the maximum admissible bed rise from the two vertical readings. Predict how changing discharge moves that tangency before changing the input.

Open this experiment online

Try the calculation

Discharge per width (m²/s) **0.4**

Bed rise (m) **0.1**



Critical depth is 0.254 m. The available head over the bump is 0.433 m versus a minimum of 0.380 m. The specified upstream state can pass this discharge.

Tutorial · infer discharge from two depth gauges

Two gauges straddle a stationary jump and report 0.12 and 0.30 m. Use the momentum relation to recover the discharge per unit width, determine which side is upstream and calculate the specific mechanical-energy loss.

▼ **Recover the flow and check its direction**

Factoring the momentum equality gives $q^2 = g h_1 h_2 (h_1 + h_2)/2$, hence $q = 0.2724$ m²/s. The shallow side has speed 2.27 m/s and Froude number 2.09. The deeper side has speed 0.908 m/s and Froude number 0.529. Taking the shallow side upstream yields energy loss $0.18^3/[4(0.12)(0.30)] = 0.0405$ m. Both discharge and the sign of dissipation support this ordering.

Further reading · W. Smyth, All Things Flow, §9.3, hydraulic jumps and bores

Thermal diffusion and transient surface cooling

Engineering thermodynamics and heat transfer

Use diffusion time, Biot number and penetration depth to choose between a uniform-body model and a spatial temperature field.

FROM FIRST PRINCIPLES

Derive thermal diffusion from stored energy and conductive flux

Let a stationary homogeneous solid have density ρ, specific heat c and thermal conductivity k. Assume no heat generation and consider one-dimensional temperature variation. A slice of area A and thickness dx stores energy at rate $\rho c A\,dx\,\partial_t T$. The difference between the conductive flux entering and leaving gives

$$\rho c A\,dx\,\partial_t T = A[q''(x) - q''(x+dx)] \quad \Longrightarrow \quad \rho c\,\partial_t T = -\partial_x q''.$$

Fourier's law $q'' = -k\partial_x T$ closes this balance. For constant properties the diffusion equation and its characteristic time over distance L follow:

$$\partial_t T = \alpha\,\partial_{xx} T,$$
$$\alpha = \frac{k}{\rho c}, \qquad t_{\mathrm{diff}} \sim L^2/\alpha.$$

Initial temperatures and boundary heat-flux or temperature conditions are needed to select a specific solution. If conduction inside the solid is sufficiently rapid relative to transfer at its surface, one can integrate the balance using an almost uniform interior temperature. Comparing the internal and external resistance scales yields the Biot number hL_c/k used for that simplification.

Background · Biot number

$$\mathrm{Bi} = hL_c/k$$

Here h is a surface heat-transfer coefficient and k is conductivity. A small Biot number supports a nearly uniform specimen temperature for the specified boundary condition.

DERIVATION AND APPLICATIONS

Predict how a thermal probe follows a changing fluid temperature

A sensor calibrated by an abrupt temperature step can lag behind a fluid that is itself heating steadily. Take a body of mass m, heat capacity c, exposed area A and film coefficient h with negligible internal temperature gradients. This lumped description requires internal conduction to be fast compared with surface heat transfer. Treat properties and h as constant and neglect self-heating, radiation and lead conduction. Denote the body temperature by T, fluid temperature by T_f, and set $\tau = mc/(hA)$.

An instantaneous energy balance gives a first-order differential equation. For a linearly increasing fluid temperature $T_f = T_0 + \beta t$, define the lag $D = T_f - T$:

$$mc\dot{T} = hA(T_f - T),$$

$$\dot{D} + D/\tau = \beta.$$

Suppose the probe and fluid start in equilibrium at T_0. Multiply the lag equation by $e^{t/\tau}$ and integrate from zero. The integrating factor converts the left side to a single derivative:

$$e^{t/\tau} D(t) = \beta \int_0^t e^{s/\tau}\, ds = \beta\tau(e^{t/\tau} - 1).$$

Subtract the resulting lag from the prescribed fluid history to obtain the reading:

$$T(t) = T_0 + \beta[t - \tau(1 - e^{-t/\tau})].$$

For $t \ll \tau$ the rise begins quadratically, since the initial film temperature difference was zero. For $t \gg \tau$ the probe warms at the same rate as the fluid but reads low by $\beta\tau$. Cooling ramps reverse the sign. A measured constant lag is therefore evidence for this model only after the startup transient has decayed.

WORKED EXAMPLE

Evaluate the reading during a ramp calibration

Change the inputs and test this calculation.

A probe has $mc = 0.800$ J/K and $hA = 0.100$ W/K, hence $\tau = 8.00$ s. Start at 20.0°C and heat the fluid by 0.500 K/s. At $t = 16.0$ s the fluid is at 28.0°C and the lag is $4.00(1 - e^{-2}) = 3.46$ K, so the probe reads 24.5°C. Simply subtracting the eventual 4.00 K would overcorrect during startup.

SECOND APPLICATION

Estimate a time constant without an abrupt temperature step

Change the inputs and test this calculation.

During a long calibration ramp of 0.200 K/s, a second probe remains 1.20 K below an independent fluid reference. Its inferred time constant is

$$\tau = 1.20/0.200 = 6.00\,\mathrm{s}.$$

If the fluid is then held fixed, the remaining lag decays exponentially. Waiting $6.00\ln(1.20/0.100) = 14.9$ s brings the error below 0.100 K. The hold and the ramp test two different responses using the same fitted parameter.

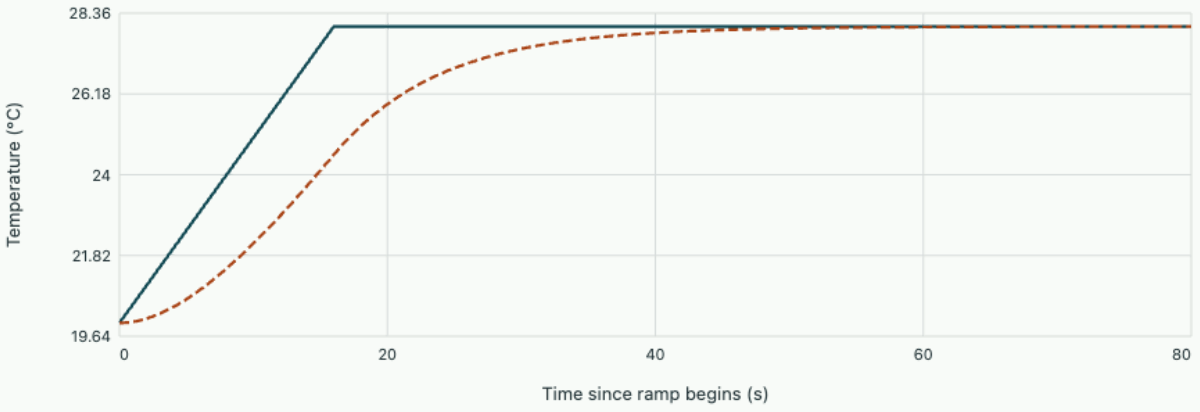
Try it · Probe ramp

Set the inputs to the values in the worked example, then vary one quantity while holding the others fixed. Compare the curve and numerical balance with the second application. The calculation uses the assumptions derived above; its readout identifies bounds such as choking, phase-region changes and inconsistent measurements.

Open this live calculation

Try the calculation

Probe time constant (s)	8
Fluid ramp rate (K/s)	0.5
Ramp duration before hold (s)	16



Fluid ramp then hold Lumped probe reading

At end of ramp fluid 28 °C, probe 24.5413 °C, signed lag 3.45866 K. Long-ramp lag βτ=4 K; the hold tests the same time constant independently. Internal temperature gradients are neglected.

Allow energy to accumulate in the material

A surface thermocouple responds rapidly to a cold wash, while the interior of a thick part can remain hot. To predict the time needed to cool the centre, steady thermal resistance alone is insufficient. Enclose a differential slice of material. Incoming conductive heat minus outgoing heat plus local generation equals the rate of change of its stored energy. Substituting Fourier’s law in a homogeneous stationary solid gives

$$\rho c_p \frac{\partial T}{\partial t} = k \nabla^2 T + \dot{q}_v,$$

$$\alpha = \frac{k}{\rho c_p}.$$

Thermal diffusivity α in m²/s compares a material’s ability to spread heat with its ability to store it. A large conductivity does not alone guarantee fast adjustment if volumetric heat capacity is also large. Without heat generation, a length L gives a characteristic diffusion time $t_d = L^2/\alpha$. The dimensionless Fourier number $Fo = \alpha t / L^2$ measures how long diffusion has acted relative to that scale.

Decide whether the interior stays nearly uniform

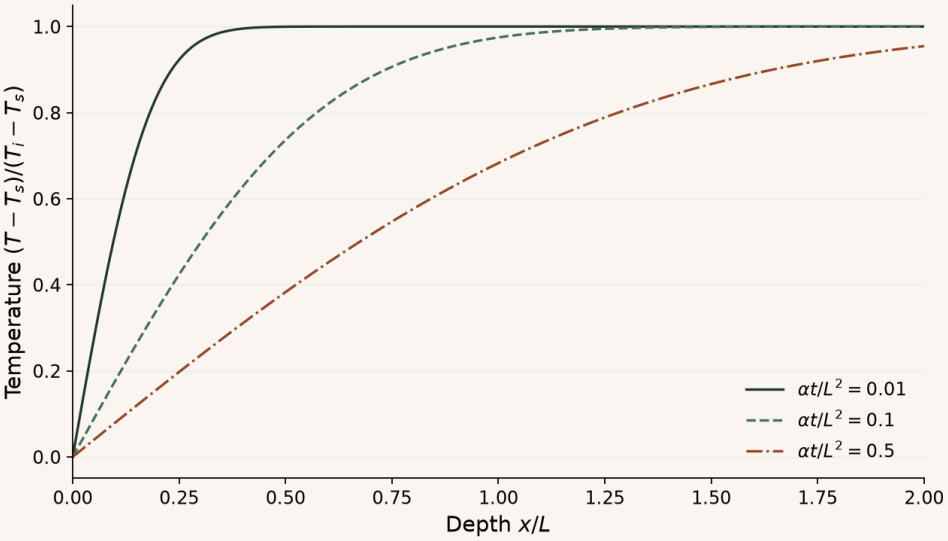
At a convecting boundary $-k \partial T/\partial n = h(T_s - T_\infty)$. Comparing the resistance of the interior with a fluid film leads to $Bi = hL_c/k$, often using $L_c = V/A_s$ for a lumped model. If this ratio is small, commonly below about 0.1 as a first engineering check, the internal temperature difference is small compared with that across the film. Integrating energy over a uniform body then gives

$$\rho c_p V \, dT/dt = -h A_s (T - T_\infty),$$

$$(T - T_\infty)/(T_0 - T_\infty) = e^{-t/\tau},$$

with $\tau = \rho c_p V/(h A_s)$.

The Biot and Fourier numbers answer different questions. A small Fourier number says a disturbance has penetrated only a short distance compared with L. A small Biot number says the boundary transfers heat slowly enough for internal gradients to remain small. Choose the length for each comparison explicitly, especially for a slab cooled on one side rather than both.



Semi-infinite cooling profiles plotted against a fixed reference length L. Increasing time by a factor of ten increases the depth of a given temperature ratio by the square root of ten.

Solve a sudden prescribed-temperature change

Consider initially uniform material occupying $x > 0$ whose surface is suddenly held at T_s, with its far interior still at T_i. Neglect internal generation and use constant properties. The initial and boundary conditions have no geometric length scale. Try the similarity coordinate $\eta = x/(2\sqrt{\alpha t})$ and set $\Theta = (T - T_s)/(T_i - T_s)$. Differentiating and substituting into the diffusion equation yields $\Theta'' + 2\eta\Theta' = 0$, with derivatives in η. Integrating once gives

$$\Theta' = Ce^{-\eta^2}.$$

The limits $\Theta(0) = 0,\ \Theta(\infty) = 1$ give

$$\Theta = \operatorname{erf}(\eta).$$

At $x = 0$ the imposed surface temperature holds exactly. At a fixed depth and very early time η is large, so the temperature tends to its initial value. The distance of any fixed fractional temperature change grows as $\sqrt{t}$. Quadrupling exposure time doubles that distance, a scaling quite different from a body travelling at a constant speed.

Predict depth and validate the time window

A thick solid initially at 373 K is held at 293 K at its surface. With $\alpha = 1.0 \times 10^{-5}$ m²/s and $t = 10$ s, $2\sqrt{\alpha t} = 0.020$ m. At depth 10 mm, $\eta = 0.50$ and $\operatorname{erf}(0.50) = 0.5205$, giving $T = 293 + 80(0.5205) = 334.6$ K. At 20 mm, $\operatorname{erf}(1) = 0.8427$ gives 360.4 K. The surface has cooled by 80 K while the deeper sensor has cooled by only 12.6 K.

At 40 s the 20 mm sensor reaches the same fraction previously measured at 10 mm. This prediction assumes the remote boundary has not begun to influence the measurement and that the surface temperature is actually prescribed. A finite slab or a convective wash may require a finite-domain diffusion solution instead. Record surface and interior histories to distinguish uncertain boundary conditions from uncertain diffusivity.

Interpret a penetration depth

If two embedded sensors are at 2 and 6 mm in the same semi-infinite specimen, the deeper sensor reaches the same normalised temperature at nine times the shallow sensor's time because $x/\sqrt{\alpha t}$ must match. An observed factor much smaller than nine can indicate a heat leak along a probe or a far boundary that has begun to influence the temperature field.

Resolve a slab with a finite surface film

A prescribed surface temperature and a uniformly cooling lump are two useful limits. A finite slab exchanging heat through a finite film generally satisfies neither. Let L be half-thickness, $x = 0$ the midplane, and T_i the initial uniform temperature. Both sides suddenly see fluid at T_∞ with constant h. Using $X = x/L,\ Fo = \alpha t/L^2$ and $\Theta = (T - T_\infty)/(T_i - T_\infty)$, the diffusion equation becomes

$$\Theta_{Fo} = \Theta_{XX},$$

with symmetry $\Theta_X(0) = 0$ and film condition $-\Theta_X(1) = Bi\,\Theta(1)$, where $Bi = hL/k$.

Seek a separated mode $\Theta = \cos(\lambda X)e^{-\lambda^2 Fo}$. Its derivative vanishes at the midplane. Imposing the convective surface condition gives

$$\lambda \tan \lambda = Bi.$$

The uniform initial condition fixes modal amplitudes by cosine orthogonality on $0 < X < 1$:

$$\Theta = \sum_n A_n \cos(\lambda_n X) e^{-\lambda_n^2 Fo},$$

$$A_n = \frac{4 \sin \lambda_n}{2\lambda_n + \sin(2\lambda_n)}.$$

For $Bi \ll 1$, the smallest root has $\lambda_1^2 \simeq Bi$, its shape approaches a constant, and $A_1 \simeq 1$. The decay becomes $e^{-Bi\,Fo}$, exactly the lumped time scale because $Bi\,Fo = ht/(\rho c L)$. For a large film coefficient the first root tends towards $\pi/2$, corresponding to a prescribed surface temperature. The same boundary-value model thus recovers both limiting behaviours.

Predict when the core and surface meet different targets

A 40 mm thick flat slab with $k = 2.0\,\mathrm{W/(m\,K)}$ and $\alpha = 1.0 \times 10^{-6}\,\mathrm{m^2/s}$ is cooled on both sides by $h = 50\,\mathrm{W/(m^2K)}$. It starts at 100 °C in fluid at 20 °C. Here $L = 0.020$ m and $Bi = 0.50$. The first root of $\lambda \tan \lambda = 0.5$ is 0.6533 and $A_1 = 1.070$. After 600 s, $Fo = 1.50$. The first term gives centre excess ratio $1.070 \exp[-(0.6533)^2(1.50)] = 0.564$, hence $T_c = 65.1^\circ\mathrm{C}$. At the surface multiply by $\cos(0.6533) = 0.794$, giving about 55.9 °C. Higher roots are strongly attenuated at this Fourier number.

A process requiring the centre to fall below 50 °C needs centre ratio $(50 - 20)/(100 - 20) = 0.375$. Solve $Fo = \ln(1.070/0.375)/0.6533^2 = 2.46$, or approximately 984 s. By contrast a lumped estimate gives $Fo = -\ln(0.375)/0.5 = 1.96$, about 785 s. It underestimates the core cooling time by over three minutes because the film and internal gradients both matter. An early-time surface prediction requires more modes than this late-time centre criterion. In particular the first term alone gives a centre ratio above unity at time zero, signalling that the remaining terms are necessary to reconstruct the original uniform state.

If product quality depends on average energy content rather than the hottest spot, average the first-mode shape over half the thickness. Its factor $\int_0^1 \cos(\lambda_1 X)dX = \sin\lambda_1/\lambda_1 = 0.930$ gives mean excess ratio $0.564(0.930) = 0.525$ after 600 s. The predicted bulk temperature is therefore about 62.0 °C, between centre and surface as expected. A calorimetric measurement made after removing and mixing the material is closer to the mean than to either point thermometer. A decision to stop cooling based on such a reading can therefore satisfy an average enthalpy requirement while leaving the centre several kelvins above a local temperature limit.

The surface flux can be checked against the rate of thermal energy loss. At 600 seconds each face transfers $50(55.9 - 20) = 1795\,\mathrm{W/m^2}$ to the air. Per square metre of plan area the slab therefore loses approximately 3590 W. The volumetric heat capacity obtained independently from k/α is $2.0 \times 10^6\,\mathrm{J/(m^3K)}$. Multiplying by total 0.040 m thickness gives an areal capacity of 80000 J/(m² K), and a predicted instantaneous mean cooling rate $3590/80000 = 0.0449\,\mathrm{K/s}$. Differentiating the one-term mean temperature gives the same decay rate to within rounding.

This independent balance is especially useful if measurements suggest two different film coefficients on opposite sides. An unequal pair of measured fluxes cannot be fitted by a strictly symmetric half-slab model even if it reproduces one centre temperature. The full-thickness conduction problem then requires both surface boundary conditions; simply replacing them by their arithmetic mean may match total heat over one interval while shifting the hottest location away from the midplane.

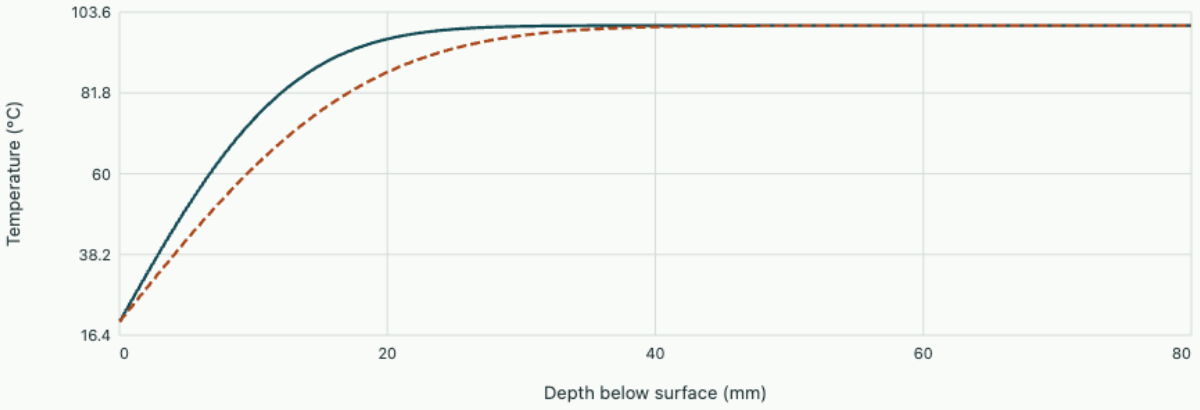
Interactive calculation · Change the inputs of this example

Change the parameters of this reference plot and compare its limiting cases. The fixed illustration remains beside the derivation; the adjustable calculation exposes the same dependence on geometry, material properties or loading.

Open this live calculation

Try the calculation

Thermal diffusivity (mm²/s)	5
Elapsed time (s)	10
Initial temperature (°C)	100



The diffusion scale2√(αt) is 14.142 mm. Semi-infinite solid, initially uniform, surface suddenly held at20°C.

A moving boundary in a freezing layer

Engineering thermodynamics and heat transfer

A flat cold surface grows a layer of ice from water initially at its melting temperature. Estimate when it will reach a specified thickness, then discover how a finite heat-transfer resistance changes its early growth.

Write a balance at a boundary whose position is unknown

Let the solid occupy $0 < x < s(t)$. The boundary at x=0 is held at $T_w < T_m$, the liquid remains at T_m, and the solid-liquid interface is at its equilibrium melting temperature. Treat density, latent heat of fusion L_f and solid conductivity k as constants. Ignore volume change and fluid convection. When the front moves by ds, each unit area releases latent energy $\rho L_f \, ds$. In this one-phase problem conduction into the cold side removes that energy. Taking the magnitude of heat extracted, the interface condition reads

$$\rho L_f \, \frac{ds}{dt} = k \, \frac{\partial T}{\partial x}\bigg|_{s^-}.$$

The derivative is positive because temperature increases towards the liquid. Solving the full problem requires time-dependent conduction in a domain with a moving outer end. A useful approximation emerges when the sensible heat needed to cool the new solid is small beside its latent heat. Measure that ratio with the Stefan number $\mathrm{Ste} = c_p(T_m - T_w)/L_f$.

Derive the square-root law and check the approximation

For small Ste the temperature profile in the solid adjusts quickly compared with front growth. Approximate the gradient by $(T_m - T_w)/s$. Multiplying the interface balance by s separates the variables. Integrating from an initial layer s_0 gives

$$s \, ds = \frac{k(T_m - T_w)}{\rho L_f} \, dt,$$

$$s^2 = s_0^2 + \frac{2k(T_m - T_w)}{\rho L_f} t.$$

With zero initial thickness the predicted speed starts unbounded because the model inserts a finite temperature difference across zero resistance. A finite surface heat-transfer coefficient removes this singularity. If coolant at T_c reaches the wall through a film $1/h$, the heat flux is $(T_m - T_c)/(1/h + s/k)$. For a layer growing from zero, integration gives

$$t = \frac{\rho L_f}{T_m - T_c} \left(\frac{s}{h} + \frac{s^2}{2k} \right).$$

The linear term dominates while the film presents the larger resistance. Once the solid is thick, the quadratic term recovers diffusion-controlled growth. These two regimes can be distinguished by

recording thickness at equal time intervals.

An ideal wall and a finite film

For water freezing at 0 °C take $k = 2.2$ W/(m K), $\rho = 917$ kg/m³, $L_f = 334$ kJ/kg and $c_p = 2.1$ kJ/(kg K). With a wall at −5 °C the Stefan number is 0.0314. The ideal-wall time to form 10 mm is $917(334000)(0.01)^2/[2(2.2)(5)] = 1392$ s, about 23 minutes. Reaching 20 mm requires four times that duration. The increasing conduction distance accounts for the slowing front.

With coolant at −5 °C and $h = 200$ W/(m² K), the added time for 10 mm is 3063 s. The resulting total is about 74.2 minutes. At that point the film resistance is 0.005 m² K/W and the ice contributes 0.00455 m² K/W. Both terms are of the same order. Vary h and find when a second thickness measurement would provide more information about ice conductivity than a first one.

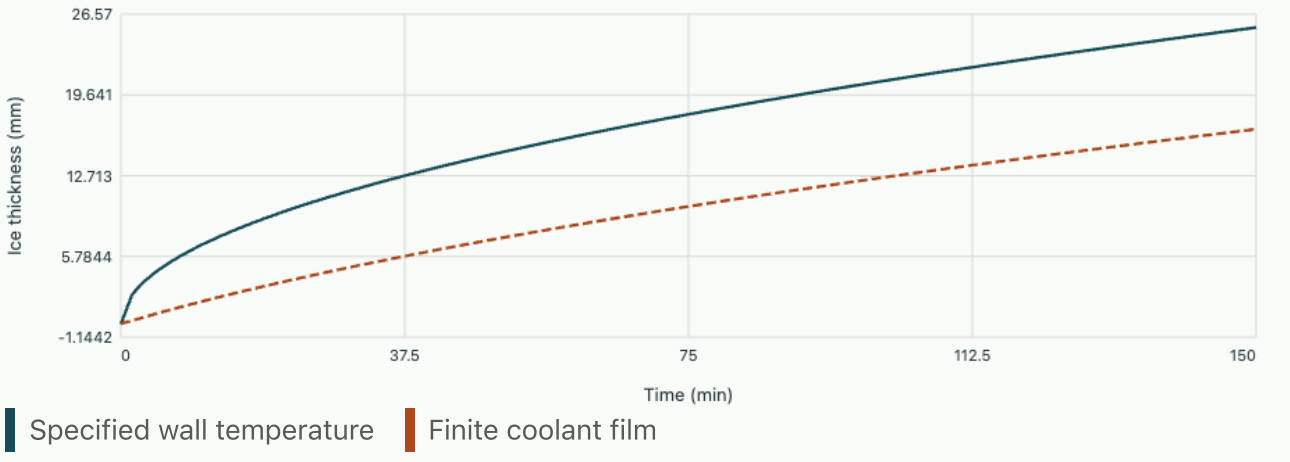
At a fixed coolant temperature compare the early and late parts of the thickness histories. Increase the film coefficient strongly. Predict where the finite-film curve should approach the fixed-wall result first in relative terms, then read the completion time.

Open this experiment online

Try the calculation

Coolant depression below freezing (K) **5**

Film coefficient (W/m² K) **200**



At the selected film coefficient 10 mm requires 74.2 minutes. The fixed-wall prediction is 23.2 minutes.

Tutorial · recover the coolant coefficient from one observation

A 15 mm layer is observed after 8000 s with the same properties and coolant temperature. Estimate h from the integrated equation. Would a time below the ideal-wall value be compatible with a positive h in

this model?

▼ **Show the resistance fit**

The ideal-wall contribution is about 3132 s. The remaining 4868 s equals $\rho L_f s/[h(T_m - T_c)]$, giving $h \simeq 189$ W/(m² K). Both resistances are positive. A shorter time than the ideal-wall prediction would require checking the temperature, initial ice thickness, geometry or material data before assigning a heat-transfer coefficient.

Further reading · RWTH Aachen · Computational Modeling of Multiphysics, chapter 12 on the Stefan problem

How a longitudinal pulse reflects from a change in rod section

Machine dynamics

A tap on one end of a rod reaches a receiver after a finite delay. Two rods joined end to end can even return a pulse before the receiver sees a far-end reflection. Derive propagation speed and distinguish the displacement amplitude from the energy carried by a pulse.

From the cut force to a travelling disturbance

Measure the small axial displacement $u(x,t)$ from a stress-free reference. In a slender homogeneous rod of area A, tensile force is

$$N = EAu_x.$$

A slice of length dx and mass $\rho A\,dx$ accelerates because its right and left forces differ by $N_x dx$. For constant properties, cancelling A gives

$$\rho A u_{tt} = (EAu_x)_x,$$

$$u_{tt} = c^2 u_{xx},$$

$$c = \sqrt{E/\rho}.$$

Either $u = f(x - ct)$ or $u = g(x + ct)$ satisfies this equation by differentiation. Curves of constant phase move right or left with speed c. Long-wavelength assumptions suppress lateral inertia and radial modes. Rod area cancels from wave speed but remains in the force carried at given particle velocity. For a right-moving pulse $v = u_t = -cu_x$, so tensile force $N = -Zv$, with mechanical impedance $Z = \rho Ac = EA/c$. The sign reverses for a left-moving wave.

Match velocity and force at a bonded junction

Let amplitudes I, R and T denote incident, reflected and transmitted particle velocity at the junction. A massless, perfectly bonded interface requires equal velocities and transmitted force. Apply the preceding sign convention separately to each travelling direction. The two equations determine two unknowns.

$$I + R = T,$$

$$-Z_1 I + Z_1 R = -Z_2 T,$$

$$\frac{R}{I} = \frac{Z_1 - Z_2}{Z_1 + Z_2},$$

$$\frac{T}{I} = \frac{2Z_1}{Z_1 + Z_2}.$$

The instantaneous power transported to the right is $-Nv = Zv^2$ for a single right-moving wave. For the same pulse shape the reflected energy fraction is $(R/I)^2$ and the transmitted fraction is $(Z_2/Z_1)(T/I)^2$. They sum to one. A transmitted velocity amplitude exceeding the incident amplitude can thus be compatible with conservation of energy.

A stepped rod and a pulse-echo observation

Join two aluminium rods of areas 80 and 160 mm², $E = 70$ GPa and $\rho = 2700$ kg/m³. Both have $c = 5092$ m/s, but the second impedance is twice the first. An incident velocity of 0.30 m/s produces a reflected −0.10 m/s and a transmitted 0.20 m/s. The reflected pulse carries one ninth of the energy. Reversing the experiment makes the reflection positive and the transmitted velocity 0.40 m/s for the same incident amplitude. The two energy fractions are unchanged on reversal.

A receiver at the tap detects a reflected pulse 0.40 ms after launch, before any far-end reflection contaminates it. The return path covers the distance twice, so the inferred junction is $c(0.00040)/2 = 1.02$ m away. Measure pulse velocity rather than strain when comparing the signs to the plot. Incident tensile force has sign opposite incident velocity with the convention used here.

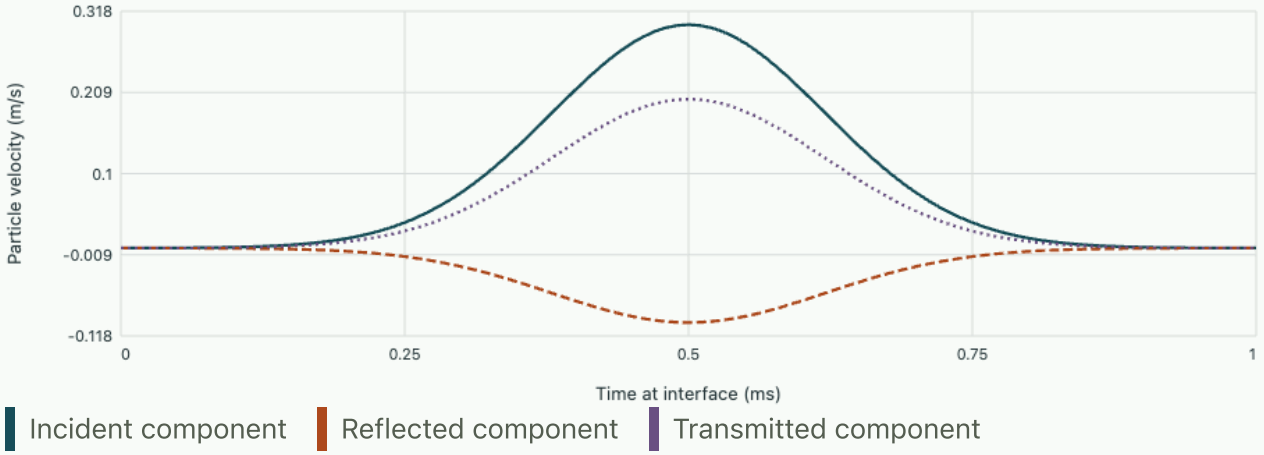
Keep the incoming pulse fixed and change the area ratio from 2 to 0.5. Determine which velocity component reverses sign and compare reflected energy fractions. Explain the difference by writing power balance for the two experiments.

Open this experiment online

Try the calculation

Second / first section area **2**

Incident peak velocity (m/s) **0.3**



Peak reflected velocity is -0.100 m/s and transmission is 0.200 m/s. The reflection carries 11.1% of the incident energy.

Tutorial · infer a second cross-section

On a uniform material an isolated junction returns 25% of the incident velocity amplitude with the same sign. Infer the area ratio and transmitted energy fraction. How could an energy-only measurement leave the geometry ambiguous?

▼ **Show the inverse relation**

Let $z = Z_2/Z_1 = A_2/A_1$. Then $0.25 = (1 - z)/(1 + z)$, so $z = 0.60$. The transmitted velocity is $1.25I$. Its energy fraction is

$$0.60(1.25)^2 = 0.9375.$$

Measuring only the reflected energy gives 0.0625 and loses the sign of the amplitude. A thickening by a factor $1/0.60$ would give the same energy reflection and the opposite sign in the velocity trace.

Further reading · Bower, Applied Mechanics of Solids · solutions to problems involving vibrations and wave propagation

Why the frequencies of a bending beam are not equally spaced

Machine dynamics

A distributed beam can bend into infinitely many shapes. Derive the relation between wavelength and oscillation frequency, then compare the measured response of a beam with the mode family predicted by its supports.

Combine transverse force, moment and curvature

Take a straight uniform beam with small transverse deflection $w(x,t)$, density ρ, area A and second moment of area I about the bending axis. Euler–Bernoulli kinematics keeps cross-sections perpendicular to the centreline. The bending resistance is proportional to curvature EIw_{xx}. For a load-free part of the beam the gradient of moment supplies shear and the gradient of shear accelerates the mass. With restoring forces opposing deflection, the resulting equation is

$$\rho A\, w_{tt} + EI\, w_{xxxx} = 0.$$

The fourth spatial derivative calls for four boundary conditions. At a simple support the displacement vanishes and the bending moment is zero, hence

$$w = w_{xx} = 0.$$

A clamp instead fixes displacement and slope. Interchanging a pin and a clamp changes the allowed shape family even if the span and material remain fixed.

Obtain the simply supported series

Try $w = W\sin(kx)\cos(\omega t)$. Two spatial derivatives bring down $-k^2$, and four give $+k^4$. Substitution balances the inertia and stiffness terms. The conditions at x=0 hold for every k in this sine family. At x=L the sine must vanish, yielding

$$\omega^2 = \frac{EI}{\rho A}k^4,$$

$$k_n = \frac{n\pi}{L},$$

$$f_n = \frac{n^2\pi}{2L^2}\sqrt{\frac{EI}{\rho A}}.$$

A long travelling sinusoid has phase velocity $\omega/k = k\sqrt{EI/(\rho A)}$. The group velocity $d\omega/dk$ is twice that value in this model. A localized pulse contains many wavelengths, which can therefore separate as it propagates. Predictions at wavelengths near the thickness require accounting for rotary inertia and shear deformation.

Choose sensor positions as carefully as the frequency range

Use a 1 m steel strip 20 mm wide and 4 mm thick with $E = 200$ GPa and $\rho = 7800$ kg/m³. For bending about the thin direction $I = bh^3/12$ and $I/A = h^2/12$. The first three simply supported frequencies are roughly 9.18, 36.7 and 82.7 Hz. Doubling the width alone doubles stiffness and mass together and leaves these frequencies unchanged. Doubling thickness doubles each frequency. Shortening the span by 20% increases all three by $1/0.8^2 = 1.56$.

At the centre a displacement pickup sees $\sin(n\pi/2)$. It sees the first mode at its antinode and misses the second mode entirely. At the quarter point it sees that second mode strongly. An absent spectral peak can therefore result from excitation or sensing at a node. Move the measuring point before adjusting a fitted damping value to explain its absence.

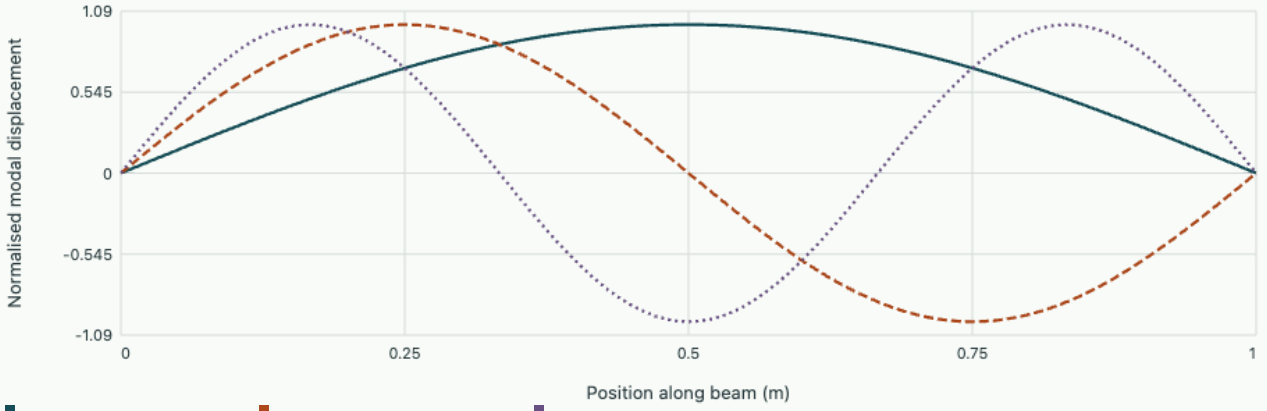
Find a common node of the first and third modes and a location that suppresses the second mode alone. Change the thickness and then the span. Use the resulting frequencies to test whether each scaling is linear or quadratic.

Open this experiment online

Try the calculation

Span (m) 1

Thickness (mm) 4



At midspan the second mode has a displacement node. The three frequencies are 9.18, 36.74, 82.66 Hz.

A related OneSimulation calculation

For a frame whose geometry and end payload preclude the simply supported sine family, compare the first three computed modes after changing the tip mass. Test whether all frequencies scale by the same factor.

Compare frame modes with OneSimulation

Tutorial · distinguish distributed bending from a tensioned string

A simply supported specimen has prominent peaks at 12 Hz and 108 Hz with the receiver at midspan. Infer which bending modes these could be and predict the frequency and receiver location for the missing mode. What would the third-to-first frequency ratio be for an ideal string?

▼ Show the mode test

The ratio $108/12 = 9$ agrees with n^2 for $n = 3$. The second bending mode should lie near $4(12) = 48$ Hz. Measure at one quarter span and excite away from its nodes. An ideal uniform string has $\omega = ck$ and a third-to-first ratio of 3. Checking both the midspan node and the shifted pickup spectrum distinguishes these models more effectively than a single peak.

Further reading · MIT 2.003SC · lecture 27 on vibration of continuous structures

How a magnetic field relaxes into a finite conductor

Magnetism, induction and field dynamics

A conducting sheet initially shields its interior from an applied magnetic step. Given time, a normal resistive metal admits a static tangential magnetic field. The same equations give a finite slab a discrete relaxation spectrum and a periodic drive a complex penetration depth. These are different initial and boundary conditions for one diffusion operator.

Identify the time derivative lost in the approximation

Let the conductor be stationary, Ohmic, uniform and isotropic with conductivity σ, permeability μ and permittivity ϵ. In the full Maxwell equations both $\sigma \boldsymbol{E}$ and $\epsilon \partial_t \boldsymbol{E}$ contribute to the curl of $\boldsymbol{H}$. At a characteristic angular frequency ω, the ratio of the latter to the former is $\omega\epsilon/\sigma$. For a good conductor at low frequencies this is very small. Neglect displacement current only in that constitutive region; a vacuum gap or a capacitor elsewhere can still carry significant displacement current. Ampère then gives

$$\boldsymbol{E} = (\nabla \times \boldsymbol{B})/(\mu\sigma).$$

Substitution into Faraday and use of $\nabla \cdot \boldsymbol{B} = 0$ yield

$$\partial_t \boldsymbol{B} = D\nabla^2 \boldsymbol{B},$$
$$D = 1/(\mu\sigma).$$

The diffusion time for scale a is of order a^2/D. Increasing conductivity reduces D, because smaller electric fields can sustain stronger opposing currents. In an ideal mathematical conductor with $\sigma \to \infty$, an initial flux can be frozen in this model. That is a different preparation from a superconductor expelling a field in equilibrium. Magnetic diffusion alone contains neither a thermodynamic superconducting transition nor London screening. At sufficiently short times the dropped displacement current also matters: an ideal step is an approximation to a finite-risetime drive, and the diffusion equation's nonzero Gaussian tails at arbitrarily large distances should not be used to infer a superluminal signal.

Obtain a transient similarity profile

Take a half-space $x > 0$ with field $\boldsymbol{B} = B_z(x,t)\hat{\boldsymbol{z}}$; the surface is driven to B_s at $t = 0^+$, while initially the interior has zero field. Scaling x by $\sqrt{Dt}$ suggests $B_z/B_s = F(\eta)$, $\eta = x/(2\sqrt{Dt})$. Compute $\partial_t F = -\eta F'/(2t)$ and $D\partial_x^2 F = F''/(4t)$; hence

$$F'' + 2\eta F' = 0.$$

Its first integral is

$$F' = Ce^{-\eta^2}.$$

Set $F(0) = 1, F(\infty) = 0$ and use the Gaussian integral to obtain

$$F(\eta) = \operatorname{erfc}\eta.$$

The associated current $J_y = -(1/\mu)\partial_x B_z$ is positive for this sign of surface field and has a Gaussian depth profile.

$$J_y(x,t) = \frac{B_s}{\mu\sqrt{\pi D t}} e^{-x^2/(4Dt)},$$

$$\int_0^\infty J_y\,dx = \frac{B_s}{\mu}.$$

The current integral remains fixed as the profile spreads, even though the surface current density falls as $t^{-1/2}$. A finite conductivity does not imply uniform current at every instant. For $\sigma = 5.8 \times 10^7\,\mathrm{S/m}$ and $\mu = \mu_0$, $D = 0.01372\,\mathrm{m^2/s}$. The point reaching half the surface field is

$$x = 2\operatorname{erfc}^{-1}(1/2)\sqrt{Dt} = 0.9539\sqrt{Dt}.$$

At $0.10\,\mathrm{ms}$ it is $1.12\,\mathrm{mm}$ deep. At $0.40\,\mathrm{ms}$, four times the time, it is twice as deep. Measuring several half-amplitude positions against time lets one fit D without calibrating the absolute magnetic field strength.

Satisfy both faces of a finite slab

Let the conductor occupy $-a < x < a$. Impose the same sudden tangential field B_s at both faces. At long time the unique steady solution with these equal boundaries is uniform B_s. Subtract this final value by defining the deficit $u = 1 - B_z/B_s$. It satisfies homogeneous boundary values $u(\pm a, t) = 0$ and begins at one inside. Even symmetry permits cosine modes $\cos(k_m x)$ with $k_m a = (m + 1/2)\pi$. Separating time gives the multiplier $e^{-D k_m^2 t}$. Orthogonality on $(-a, a)$ gives squared norm a for each cosine and projection of the constant equal to $2\sin(k_m a)/(a k_m)$. Thus

$$\frac{B_z(x,t)}{B_s} = 1 - \frac{4}{\pi}\sum_{m=0}^{\infty} \frac{(-1)^m}{2m+1} \cos\frac{(2m+1)\pi x}{2a}\,\exp\!\left[-(2m+1)^2 t/\tau_0\right],$$

$$\tau_0 = \frac{4a^2}{\pi^2 D}.$$

The initial boundary jump is real in the idealisation; the series reproduces the interior initial value in the usual Fourier sense, while the imposed boundary value applies for $t > 0$. Later times damp the higher spatial harmonics most rapidly. At the centre and $t = \tau_0$, the first term predicts

$$B/B_s = 1 - (4/\pi)e^{-1} = 0.5316.$$

The next correction is $4e^{-9}/(3\pi) = 0.0000524$, so this is already accurate to within 6×10^{-5}. At very early times many terms are needed because their cancellations encode a nearly field-free interior. A half-space similarity expression near each face can be more efficient there.

For a $4.0\,\mathrm{mm}$-thick copper slab $a = 2.0\,\mathrm{mm}$, giving $\tau_0 = 0.118\,\mathrm{ms}$. At late times the centre deficit $1 - B/B_s$ approaches a single exponential with log slope $-1/\tau_0$. Reaching 90 percent at the centre requires

$$t \simeq \tau_0 \ln(4/(0.1\pi)) = 2.54\tau_0 = 0.300\,\mathrm{ms}.$$

The initial square-root penetration law cannot simply be extrapolated into this final regime, since the penetration fronts from both sides interact. Doubling full thickness at constant material quadruples the long relaxation time.

Compare a periodic solution and a surface impedance

Instead drive the half-space sinusoidally and adopt $e^{i\omega t}$ peak phasors. Solving $i\omega\widetilde{B} = D\,d^2\widetilde{B}/dx^2$ gives $\widetilde{B} = B_s e^{-(1+i)x/\delta}$ with the root $\delta = \sqrt{2D/\omega}$ that decays towards infinity. The envelope falls as $e^{-x/\delta}$, while phase lags the boundary by x/δ. The electrical surface field follows by differentiating, yielding $\widetilde{E}_y = (1+i)\widetilde{H}_z/(\sigma\delta)$. The surface impedance in these orientations is $Z_s = (1+i)/(\sigma\delta)$. Its positive real part measures dissipation, and the equal quadrature term records energy storage with this convention.

At $f = 1.0\,\mathrm{kHz}$ in the copper model, $\delta = 2.09\,\mathrm{mm}$ and $\mathrm{Re}\,Z_s = 8.25\,\mu\Omega$. A tangential surface peak field $H_s = 1000\,\mathrm{A/m}$ gives mean incoming power density $\mathrm{Re}\,Z_s|H_s|^2/2 = 4.13\,\mathrm{W/m^2}$. The same number follows from volume Joule heating. Integrate $\sigma|\widetilde{E}_y|^2/2$ over depth; the squared envelope $e^{-2x/\delta}$ integrates to $\delta/2$. An experiment monitoring just the in-phase field can find a zero at $x = \pi\delta/2$ while the full harmonic magnitude is still $e^{-\pi/2} = 0.208$ of the surface amplitude.

Compare a half-space step and a sinusoidal depth profile using the same material

Use the solution to design a convergence test

A mesh across the finite slab can step $B_j^{n+1} = B_j^n + r(B_{j+1}^n - 2B_j^n + B_{j-1}^n)$, with $r = D\Delta t/h^2$. A Fourier mode e^{ikjh} is multiplied by $1 - 4r\sin^2(kh/2)$, so forward Euler is stable for $0 \le r \le 1/2$. Above that bound a high-frequency grid mode can grow despite the physical diffusion being dissipative. For $h = 0.20\,\mathrm{mm}$ in the copper model, $\Delta t \le 1.46\,\mu\mathrm{s}$. Halving the mesh spacing demands quartering this explicit time step. An implicit scheme can overcome the same stability restriction but still needs small enough steps to resolve the measured transient.

Set both face values, initialise the interior once and record the centre at $t = \tau_0$. Compare first against the series, then against a second mesh and time step. A mistaken use of full thickness in τ_0 would shift the inferred time by a factor four even if the code has converged perfectly. Independent checks include reflection symmetry about the slab centre, zero gradient at that centre, a bounded response between zero and B_s for the continuous problem, and ohmic heating that is nonnegative at every point. The latter can catch sign mistakes in an energy budget even though squaring J hides the sign of a local measured current.

Compare surface and centre in a periodically driven finite slab

At a fixed angular frequency a finite slab with equal tangential field on both faces has a different profile from a half-space. Write the phasor equation $d^2\widetilde{B}/dx^2 = k^2\widetilde{B}$, $k = (1+i)/\delta$ for temporal dependence $e^{i\omega t}$, and impose $\widetilde{B}(\pm a) = B_s$. Even symmetry eliminates $\sinh kx$, giving $\widetilde{B}(x) = B_s \cosh(kx)/\cosh(ka)$. The centre amplitude is $1/|\cosh[(1+i)a/\delta]|$ of the applied field, and its phase lag is the argument of that denominator. Both incoming penetration profiles must be combined as complex amplitudes satisfying boundary conditions; adding two decaying real envelopes is not a valid phasor solution.

The low-frequency limit has $ka \ll 1$, yielding $B(0)/B_s \simeq 1 - (ka)^2/2 = 1 - i(a/\delta)^2$. The leading correction is a phase lag, while the magnitude change starts at higher order. Thus a sensitive phase detector can pick up eddy-current effects before a large change of amplitude is visible. At $a/\delta = 1$, $\cosh(1+i) = \cosh 1 \cos 1 + i \sinh 1 \sin 1 \simeq 0.834 + 0.989i$, giving centre amplitude 0.773 and lag 0.870 radians. An isolated half-space viewed one skin depth in from a single face would have amplitude 0.368 and lag 1.00 radian. These are genuinely different field boundary conditions, not two definitions of skin depth.

The phasor of the current varies as $\sinh(kx)$ and is odd. Current on one half of the slab runs opposite to that on the other; the net integrated current across both halves vanishes for equal exterior fields. That cancellation does not suppress losses, because local Joule heating is proportional to $|J|^2$. In the thin-slab limit expand Faraday directly. If $B_z(t)$ is nearly uniform,

$$\partial_x E_y = -\dot{B}_z.$$

Symmetry about the centre gives

$$E_y \simeq -x\dot{B}_z.$$

For a sinusoidal peak field B_s, the mean volumetric loss averaged across thickness is $\sigma\omega^2 B_s^2 a^2/6$. The a^2 scaling at fixed total magnetic volume explains why dividing a core into electrically insulated thin laminations can reduce classical eddy losses when the applied field lies along the lamination faces. The thin-slab approximation, isolation between laminae and field orientation are essential parts of that conclusion.

For a material with $\sigma = 2.0 \times 10^6\,\mathrm{S/m}$, half-thickness $0.10\,\mathrm{mm}$, sinusoidal peak field 0.10 T and frequency 50 Hz, the formula gives $3.29\,\mathrm{W/m^3}$. If effective permeability were $100\mu_0$, the skin depth at that frequency would be about $5.03\,\mathrm{mm}$, so the thin-slab ordering is internally consistent for these model numbers. Hysteresis or anomalous dynamic losses can still dominate a measured sample. Halving full lamination thickness at fixed material volume should quarter only this classical contribution, not every loss mechanism in the test fixture.

A thin conducting cylindrical shell gives a single time constant

A different finite system exposes how geometry changes a diffusion transient. Consider a long cylindrical shell of radius a and wall thickness $w \ll a$ surrounding vacuum. Assume the skin depth in the wall exceeds w on the important time scales, so current is approximately uniform across the wall, and the solenoid producing an approximately uniform external axial field is sufficiently long. An azimuthal sheet current K_φ produces an additional axial field inside and negligible external field in the long-cylinder approximation. The boundary condition is

$$B_{\mathrm{in}} = B_{\mathrm{ext}} + \mu_0 K_\varphi.$$

Faraday's law around the shell gives

$$2\pi a E_\varphi = -\pi a^2 \dot{B}_{\mathrm{in}}.$$

The sheet obeys $K_\varphi = \sigma w E_\varphi$. Combining the three relations gives

$$B_{\mathrm{in}} + \tau_s \dot{B}_{\mathrm{in}} = B_{\mathrm{ext}},$$

$$\tau_s = \mu_0 \sigma w a / 2.$$

For a step from zero to B_s the field inside grows as $B_s(1 - e^{-t/\tau_s})$. For a sinusoid the interior phasor transfer is $1/(1 + i\omega\tau_s)$. This model has a single time constant instead of a slab's infinite relaxation spectrum because the wall thickness has been collapsed into one azimuthal sheet-current degree of freedom; magnetic energy is stored mainly throughout the bore. Take copper with $a = 0.030\,\mathrm{m}, w = 0.10\,\mathrm{mm}$ to obtain $\tau_s = 0.109\,\mathrm{ms}$. At its half-power frequency $f = 1/(2\pi\tau_s) \simeq 1.46\,\mathrm{kHz}$, the copper skin depth is about 1.73 mm, comfortably larger than the specified wall. At 100 times this frequency the uniform-wall assumption is much weaker.

This result also suggests a discriminating experiment. Compare shells made of the same material with fixed wall thickness and different bore radius. The leading screening time grows linearly with radius. Compare solid sheets of different normal penetration thickness in the earlier slab experiment and their time grows with thickness squared. Either system can display a late exponential, so an observed straight tail in a semilog trace does not alone reveal which diffusion path limits the response. The thickness and radius dependences, together with the phase response and the known current paths, provide independent evidence. A longitudinal insulating slit in the cylindrical shell interrupts its closed eddy-current path, substantially altering the ideal shielding despite leaving most of the metal present.

Distinguish propagation from diffusion

Change conductivity or drive frequency and inspect how the field attenuates with depth. Compare its amplitude decay and phase lag.

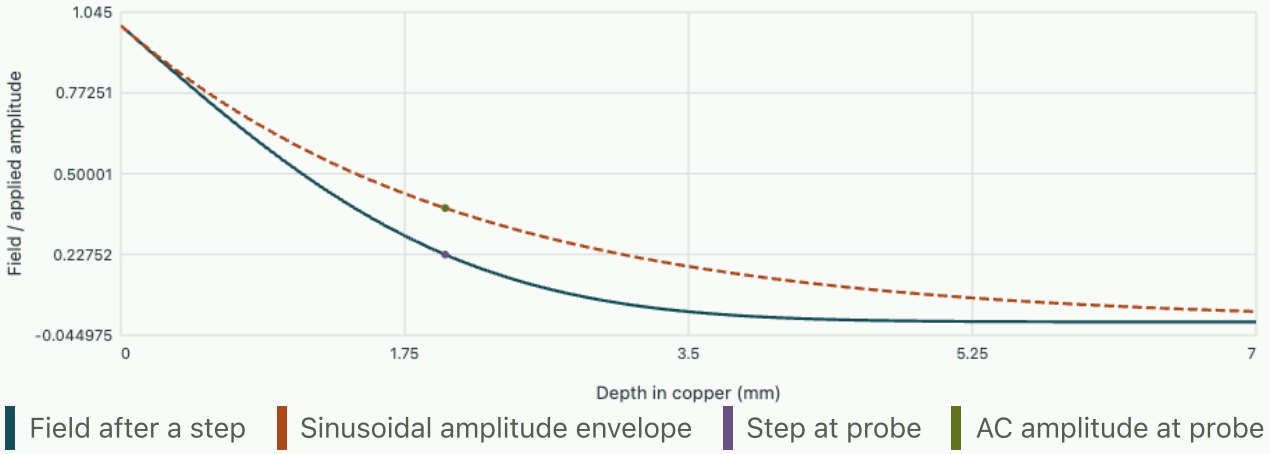
Open this experiment online

Try the calculation

Time after field step (ms) **0.1**

Sinusoidal drive frequency (kHz) **1**

Probe depth (mm) **2**



Half-step depth 1.117 mm. AC skin depth 2.09 mm. At the probe the step fraction is 0.2273, AC amplitude fraction 0.384 and AC phase lag 0.957 rad.

Account for every joule admitted by a magnetic step

A conducting half-space initially has zero internal field. At $t = 0$ a source clamps the surface induction to $B_s\hat{z}$. Let $x > 0$ measure inward distance and $D = 1/(\mu\sigma)$. The previously derived field profile is $B_z = B_s \operatorname{erfc}[x/(2\sqrt{Dt})]$. An observer sees an ever-deepening field front even though the surface induction subsequently remains constant. Someone must provide both the magnetic field energy and the Joule heat generated by the induced current. The appropriate bookkeeping is Poynting's theorem, with positive surface power defined into the conductor.

In this geometry Ampère's law gives

$$J_y = -(1/\mu)\partial_x B_z.$$

Differentiating the error function produces a Gaussian current density $J_y = B_s[\mu\sqrt{\pi D t}]^{-1} \exp[-x^2/(4Dt)]$. Ohm's law sets $E_y = J_y/\sigma$. At the surface the Poynting vector $E_y H_z \hat{x}$ points into the material and has magnitude

$$\frac{P_{\text{in}}}{A} = \frac{B_s^2}{\mu}\sqrt{\frac{D}{\pi t}}.$$

The initial mathematical step creates an integrable square-root singularity in the surface power. Holding B_s fixed still requires the source to deliver power for $t > 0$. Faraday's law relates the surface electric

field to the changing flux below the surface. Its value falls with time because the depth derivative of the induction spreads and weakens. A physical source with a finite rise time smooths the initial singularity and sets the earliest time for which the ideal step is a useful approximation.

To calculate heat production, square the full depth-dependent current and integrate J_y^2/σ from zero to infinity. The squared Gaussian has exponent $-x^2/(2Dt)$, and its integral on the half-line is $\sqrt{\pi Dt/2}$. The result is

$$P_{\mathrm{heat}}/A = (B_s^2/\mu)\sqrt{D/(2\pi t)}.$$

Consequently the heat fraction of instantaneous input is $1/\sqrt{2}$, about 70.7%. The rest, $1 - 1/\sqrt{2}$, supplies an increasing store of magnetic field energy. Both powers decay as $t^{-1/2}$; their integrals up to a finite time increase as $\sqrt{t}$. These fractions belong to the step-driven semi-infinite problem. They do not describe steady sinusoidal loss or a finite specimen after the diffusion front traverses it.

With $B_s = 10\,\mathrm{mT}$, copper conductivity $5.8 \times 10^7\,\mathrm{S/m}$, and permeability close to μ_0, the diffusion coefficient is $0.0137\,\mathrm{m^2/s}$. At $100\,\mu\mathrm{s}$, the incoming power is about $526\,\mathrm{W/m^2}$. Joule heating accounts for $372\,\mathrm{W/m^2}$, and field energy is increasing at about $154\,\mathrm{W/m^2}$. Integration from the ideal step to this time gives $2tP_{\mathrm{in}}(t)/A = 0.105\,\mathrm{J/m^2}$, split into approximately $0.0744\,\mathrm{J/m^2}$ of heat and $0.0308\,\mathrm{J/m^2}$ of stored field energy. At that time $2\sqrt{Dt}$ is about $2.34\,\mathrm{mm}$. Applying these numbers to a foil much thinner than a millimetre would use the half-space solution beyond its range.

A finite slab of thickness d is driven from both faces by the same step. How much magnetic energy is eventually stored, and why must a decaying input replace the half-space power law?

▼ **Work through the solution**

At long times, Ampère's law allows a uniform induction B_s with zero current and the prescribed values on both faces. Its stored magnetic energy is $dB_s^2/(2\mu)$ per unit footprint, independent of conductivity. With $d = 5.0\,\mathrm{mm}$ and the parameters above this is about $0.199\,\mathrm{J/m^2}$. At early times each face behaves approximately as an independent half-space. Once their diffusion profiles overlap, extended use of that approximation would predict unlimited energy accumulation in a finite volume with bounded field magnitude. The finite-slab modes replace the late algebraic law by exponential decay of the current and the surface electric field.

What is the total heat in a finite linear slab after full penetration if both surface inductions are clamped to the same ideal step?

▼ **Work through the solution**

Integrate the inward surface Poynting power over all time. With symmetry about the midplane, Faraday's law on the upper half relates the magnitude of the surface electric field to $d[\int_0^{d/2} B\,dx]/dt$. The surface magnetic field after the step is the constant B_s/μ. Including both faces gives total input energy per area $B_s^2 d/\mu$. Half remains stored, so half has been dissipated as heat, or about $0.199\,\mathrm{J/m^2}$ in the example. The integrated fraction one half over the complete finite transient differs from the early-time

half-space fraction. This energy identity provides a stringent check of a discretised slab calculation because it tests the surface flux, the volume heating and their time integration together.

Hermann Haus and James Melcher, Magnetoquasistatic Relaxation and Diffusion, including experimental demonstrations.

Tutorial on extracting material parameters

The late centre deficit of a slab of full thickness 6.0 mm falls by a factor e in $0.60\,\mathrm{ms}$. Find D, and estimate σ if $\mu = 2\mu_0$.

▼ **Work through the solution**

Using $a = 3.0\,\mathrm{mm}$ gives

$$D = 4a^2/(\pi^2\tau_0) = 0.00608\,\mathrm{m}^2/\mathrm{s}.$$

Then

$$\sigma = 1/(\mu D) = 6.55\times10^7\,\mathrm{S/m}.$$

A transient alone determines $\mu\sigma$; the independent permeability assumption is essential.

At fixed frequency the signal amplitude at one depth is 10 percent of the surface amplitude in a periodic half-space experiment. Find the phase lag there and predict the new amplitude when frequency drops to one quarter.

▼ **Work through the solution**

The depth is $x = \delta\ln 10$, so the lag is 2.303 radians. Quartering frequency doubles the skin depth; the ratio becomes

$$e^{-(\ln 10)/2} = 0.3162.$$

Treating the original 10 percent as an intensity ratio instead of a field ratio would produce a different inferred depth.

A step field applied at $t = 0$ is removed again at $t = T$. Obtain the field after removal and explain why it need not vanish immediately in the interior.

▼ **Work through the solution**

By linearity superpose a positive step and a delayed negative step. For $t > T$ the normalised half-space profile is $\mathrm{erfc}[x/(2\sqrt{Dt})] - \mathrm{erfc}[x/(2\sqrt{D(t-T)})]$. It vanishes at the driven boundary but remains

positive at interior points where the later disturbance has not substantially diffused. The induced currents subsequently relax and dissipate energy.

Posing an electromagnetic evolution problem

Magnetism, induction and field dynamics

A solver cannot evolve 'a pulse in a box' from geometry and a time step alone. It needs material laws, compatible initial fields, a charge-conserving source, and enough boundary data to specify what enters the domain. The same bookkeeping is required when interpreting a real transient measurement.

Separate the evolution equations from the constraints

In vacuum with prescribed sources the evolution system can be written $\partial_t \boldsymbol{B} = -\nabla \times \boldsymbol{E}$ and $\partial_t \boldsymbol{E} = c^2 \nabla \times \boldsymbol{B} - \boldsymbol{J}/\epsilon_0$. The divergence conditions $\nabla \cdot \boldsymbol{B} = 0$ and $\nabla \cdot \boldsymbol{E} = \rho/\epsilon_0$ constrain the initial data. Differentiating the first in time gives zero from the divergence of a curl. Differentiating the second constraint residual gives $-[\nabla \cdot \boldsymbol{J} + \partial_t \rho]/\epsilon_0$. Compatible charge evolution therefore preserves an initially satisfied Gauss constraint. An arbitrarily chosen charge movie and unrelated current movie can violate it before a single boundary reflection occurs.

A numerical solver based only on the two curl equations does not automatically repair bad initial data. In an initial periodic vacuum box a uniform charge density of one sign is inadmissible because the integral of a periodic divergence is zero. One can prescribe a neutralising background as part of a separately justified model, choose net-neutral initial sources, or choose open boundaries that permit net electric flux. Silently discarding the zero Fourier mode of the charge would solve a different problem. Similarly, preserving a constant divergence error exactly during time integration shows consistency of the discrete curls with the chosen derivative, not correctness of the initial physical state.

Construct a right-running pulse from two initial fields

In one-dimensional vacuum choose $\boldsymbol{E} = E_y(x,t)\hat{\boldsymbol{y}},\ \boldsymbol{B} = B_z(x,t)\hat{\boldsymbol{z}}$. Then

$$\partial_t E_y = -c^2 \partial_x B_z,\ \partial_t B_z = -\partial_x E_y.$$

Define $w_+ = E_y + cB_z,\ w_- = E_y - cB_z$. Addition and subtraction give

$$\partial_t w_+ + c\partial_x w_+ = 0,$$
$$\partial_t w_- - c\partial_x w_- = 0.$$

Thus w_+ travels towards $+x$ and w_- towards $-x$. The initial conditions $E_y = E_0 e^{-(x-x_0)^2/(2s^2)}$, $B_z = E_y/c$ specify a right-running packet. Prescribing the same electric shape but $B_z = 0$ launches two oppositely directed electric pulses each of half the original height. Measuring only the initial electric waveform would fail to distinguish the two preparations. At $E_0 = 60\,\mathrm{V/m}$, the peak magnetic field for the first preparation is $0.200\,\mu\mathrm{T}$. An observer 0.90 m farther right sees its peak 3.00 ns later in vacuum.

The component energies provide another independent comparison. Initialising only E_y contains half as much total energy as initialising E_y and its matched B_z. At late times when the two pulses no longer overlap, each half-amplitude pulse has one quarter of the energy of the full matched pulse; the sum is half, as required. In the overlap region it is better to use the characteristic combinations themselves. In

these units $u = \epsilon_0(w_+^2 + w_-^2)/4$ and $S_x = \epsilon_0 c(w_+^2 - w_-^2)/4$ for the transverse fields. Equal oppositely running amplitudes can store energy with zero net flux.

Specify only incoming characteristic information at an open end

In $0 < x < L$, w_+ enters at the left end and w_- at the right. Values travelling out are determined by the interior solution. For a pure outgoing plane wave at the right boundary prescribe $w_- = 0$, equivalently $E_y - cB_z = 0$, while allowing w_+ to exit. At a perfectly conducting fixed wall the tangential electric field vanishes, so there $w_- = -w_+$, reflecting the electric amplitude with a sign reversal. A boundary with a real voltage-like reflection coefficient r gives $w_- = r w_+$ in this one-dimensional equal-medium case. The time-integrated returned energy fraction is r^2. Specifying both E_y and B_z independently at all times generally overdetermines a wave boundary and can amount to fixing both incident and reflected waves inconsistently.

Suppose $L = 1.20\,\mathrm{m},\ x_0 = 0.30\,\mathrm{m}$. The centre of a right-running pulse reaches the conducting end at $3.00\,\mathrm{ns}$. A detector at $x = 0.60\,\mathrm{m}$ first sees the incident centre after $1.00\,\mathrm{ns}$, then the reflected centre after $(0.90 + 0.60)\mathrm{m}/c = 5.00\,\mathrm{ns}$. The two electric traces have opposite signs, while their magnetic contributions at the wall add. Swapping the conducting end for a matched boundary removes the reflected peak within this exact one-dimensional model. A local planar absorbing condition on a curved three-dimensional boundary may still reflect oblique waves; the exact statement depended on one normal mode and its wave impedance.

Derive uniqueness from energy

Take two solutions with the same sources, initial data and perfectly conducting closed boundaries in vacuum. Their difference satisfies homogeneous Maxwell equations with zero initial fields and zero tangential electric field at the wall. Poynting's theorem for the difference says $dU_\Delta/dt = -\oint (\boldsymbol{E}_\Delta \times \boldsymbol{H}_\Delta)\cdot \boldsymbol{n}\, dA = 0$ because the flux normal to a conducting wall vanishes. Since U_Δ is a sum of nonnegative squared fields and initially zero, it stays zero. Thus the interior fields agree. A passive outgoing boundary can only remove this difference energy, so the same conclusion holds given no incoming difference signal. This proof shows why specifying incoming boundary data and initial fields correctly is more fundamental than choosing a particular potential gauge.

The argument also supplies a stability target for computation. With no drive and a perfectly reflecting cavity the discrete total energy should remain bounded, and with an absorbing passive end the loss of field energy should match outgoing flux. An unphysical sign in the wall update can inject energy, producing a growing solution even when the interior scheme is stable. Fixing the numerical grid spacing alone cannot cure an active boundary condition. At a material interface Maxwell jump conditions must be applied to the appropriate tangential or normal fields, since setting all Cartesian components continuous would generally violate flux or polarisation physics.

Discretise space and time without destroying wave kinematics

For the one-dimensional system use E_j^n at integer time and integer position, with $B_{j+1/2}^{n+1/2}$ at half steps in both. Integrating Faraday on a small rectangular loop gives

$$B_{j+1/2}^{n+1/2} = B_{j+1/2}^{n-1/2} - (\Delta t/h)(E_{j+1}^n - E_j^n).$$

Ampère then gives $E_j^{n+1} = E_j^n - c^2(\Delta t/h)(B_{j+1/2}^{n+1/2} - B_{j-1/2}^{n+1/2})$ without current. For material sources include the appropriate time-centred current. The staggered layout places discrete curls where the updated component lives. In several dimensions a compatible divergence of the curl vanishes by cancellation of oriented face-edge contributions.

Substitution of waves $e^{i(kjh-\omega n\Delta t)}$ with the matching half-step phases yields $\sin(\omega\Delta t/2) = (c\Delta t/h)\sin(kh/2)$ on the positive-frequency branch. Stability on a one-dimensional uniform grid requires $c\Delta t/h \le 1$. For $h = 5.0\,\mathrm{mm}$, the upper vacuum time step is $16.7\,\mathrm{ps}$. A cubic three-dimensional mesh imposes the more restrictive $c\Delta t/h \le 1/\sqrt{3}$, about $9.63\,\mathrm{ps}$. In one dimension the equality case transports compatible resolved plane waves at the correct speed on this grid. Numerical dispersion generally returns at smaller Courant number or for diagonal propagation in multiple dimensions. Sample a pulse across enough sites that phase and dispersion errors lie below the intended time-delay uncertainty.

A careful initialisation also respects the half time step. Assign $B^{1/2}$ from the analytic field at that shifted time or use a consistent half-step update. Assigning B^0 directly where $B^{1/2}$ is required can create a spurious left-running component in what should have been a one-way pulse. Examine w_- reconstructed at common space and time before attributing a tiny backward peak to the material. For a real calculation, refine both space and time, compare detector arrival times as well as peak amplitudes, and integrate energy flux over complete transit windows. Solver convergence, charge compatibility, and a pulse moving in the intended direction are independent tests.

Compare a matched return and a reflected pulse on a line

Match a travelling pulse at a material boundary

A single homogeneous characteristic impedance was built into the one-dimensional end conditions. Replace the boundary by an interface between lossless homogeneous nondispersive media with positive real impedances Z_1, Z_2. At normal incidence from the left, write electric amplitudes E_i, E_r, E_t and choose the magnetic sign so a right-running wave has $H = E/Z$ and a left-running one $H = -E/Z$. With no free sheet current and no singular electric field at the interface, tangential continuity gives

$$E_i + E_r = E_t,$$
$$(E_i - E_r)/Z_1 = E_t/Z_2.$$

Solving these two equations yields

$$r = \frac{Z_2 - Z_1}{Z_2 + Z_1},$$

$$t = \frac{2Z_2}{Z_2 + Z_1},$$

$$\mathcal{R} = |r|^2,$$

$$\mathcal{T} = \frac{Z_1}{Z_2}|t|^2.$$

The impedance factor in transmitted power is required even when electric continuity makes $t = 1 + r$. With $Z_1 = 300\,\Omega$, $Z_2 = 100\,\Omega$, the electric reflection is $-1/2$ and transmission $+1/2$. The reflected energy is one quarter, while transmitted energy is three quarters. Mistaking both energy fractions for one quarter merely because the two electric amplitudes have equal magnitudes leaves half the input pulse energy unexplained. At $Z_2 \to 0$ the coefficients approach an electric short: the tangential electric field vanishes, even though the magnetic field doubles in the incident medium. Matching the impedances gives zero reflection even if material wave speeds differ, since those speeds depend on the product $\mu\epsilon$ whereas impedance depends on their ratio.

Inject a short pulse with known area and width a sufficient distance to the left that it initially lies entirely in medium one. If the media are nondispersive and impedance independent of frequency across the pulse band, each electric waveform copies the incident temporal shape multiplied by its amplitude coefficient. A transmitted wave travelling at a different speed has a different spatial width for the same temporal width, so comparing stored energy by squaring peak fields and comparing equal geometrical intervals can easily mislead. A fixed detector measuring $\int (EH)\,dt$ across a complete pulse directly measures transmitted energy per area. In a simulation integrate forward and backward flux at separate planes, include any energy still between the planes, and check that the field boundary gives the same answer when the detector positions are changed.

Manufacture a source that obeys charge continuity exactly

A verification case with prescribed charge and current can test different components from a source-free travelling wave. Consider a one-dimensional periodic interval with period $2\pi/k$. Choose a longitudinal electric field $E_x(x,t) = E_a \sin(kx)\cos(\omega t)$ and zero magnetic field. Gauss's law fixes $\rho = \epsilon_0 k E_a \cos(kx)\cos(\omega t)$. Ampère–Maxwell requires

$$J_x = -\epsilon_0 \partial_t E_x = \epsilon_0 \omega E_a \sin(kx)\sin(\omega t).$$

Faraday gives no magnetic-field change because a purely x-directed function of x has zero curl. Differentiating the two source expressions verifies $\partial_t \rho + \partial_x J_x = 0$, and the periodic spatial mean of the charge vanishes. This is a prescribed-source test of the equations, not a claim that an unspecified material spontaneously supplies those currents.

For period $0.20\,\mathrm{m}$, $E_a = 100\,\mathrm{V/m}$, and $f = 2.0\,\mathrm{MHz}$,

$$k = 31.42\,\mathrm{m}^{-1}.$$

The charge-density amplitude is $27.8\,\mathrm{nC/m^3}$ and the free-current-density amplitude $0.0111\,\mathrm{A/m^2}$. At $t = 0$ the longitudinal electric energy is maximal and current vanishes. One quarter period later the electric field and charge are zero while the source current is maximal at the spatial electric-field antinodes. At every point where E, J are nonzero their product is $-\partial_t(\epsilon_0 E^2/2)$; the prescribed external charges exchange energy with the field without a magnetic Poynting flux. A discrete divergence residual and an independently measured field-energy change can therefore check the signs and amplitudes of the source implementation.

Prescribing those continuous formulas sampled separately on a staggered mesh does not necessarily produce exact *discrete* continuity. If the numerical current is sampled at half time and half space, integrate the cell continuity equation to update ρ from numerical boundary currents. Alternatively use the corresponding discrete sinusoidal differentiation factors $2\sin(kh/2)/h$ and $2\sin(\omega\Delta t/2)/\Delta t$ when manufacturing exact discrete data. Comparing the continuous and discrete constructions separates a model truncation error that converges with mesh refinement from a genuinely nonconserving source injection. A code can look excellent on transverse vacuum pulses and still fail this test because those pulses never challenge its charge handling.

Decide what a late echo means

A narrow pulse incident on a material slab can return signals from both interfaces. Separate the two in time when possible. If the slab thickness is d and its propagation speed v_2, the round-trip delay inside it is $2d/v_2$. Later echoes spaced by that delay suggest repeated internal reflections. Changing the outer computational boundary location while keeping the material and detector fixed provides a different test: an artificial wall echo moves in time by twice the boundary displacement divided by exterior speed. A genuine material echo keeps its material-defined delay. A calibrated single detector trace and these two perturbations can thus identify which part of the model causes a late pulse.

For $d = 0.30\,\mathrm{m}$ and $v_2 = 1.5 \times 10^8\,\mathrm{m/s}$ the material echo spacing is $4.0\,\mathrm{ns}$. Moving an external vacuum boundary outward by $0.45\,\mathrm{m}$ would delay a returning boundary echo by $3.0\,\mathrm{ns}$. If the incident pulse lasts 8 ns the echoes overlap strongly and individual peaks are difficult to assign. Broadening the bandwidth with a shorter source or fitting the full complex frequency response is then more informative than selecting the largest local maximum. Any comparison must include the source bandwidth and detector response because a perfectly converged field solution can still predict a different voltage from a finite-bandwidth probe.

Check a discretised field evolution against its phase

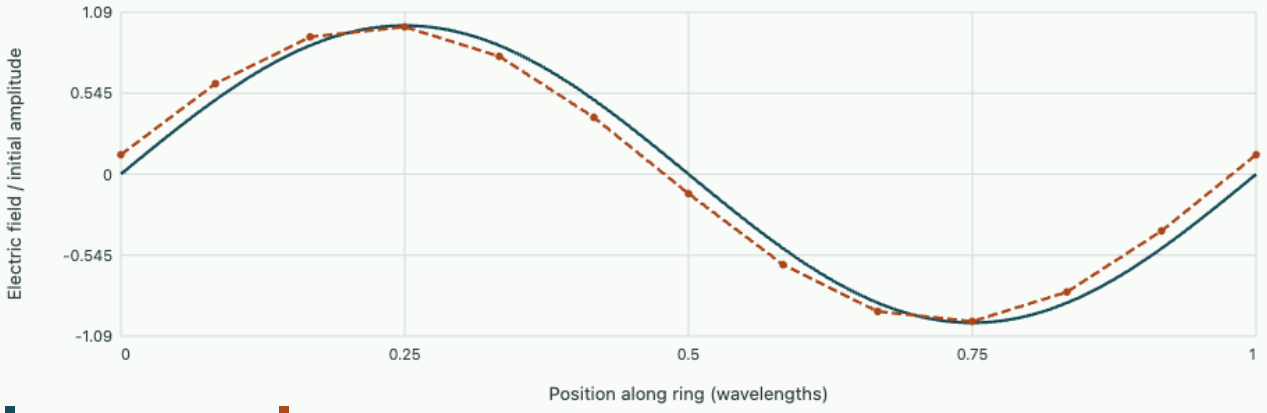
Keep the equations and initial mode fixed while refining the grid. Examine numerical phase delay alongside the physical propagation law.

Open this experiment online

Try the calculation

Points per wavelength 12

Courant number 0.8



75 steps, 5 physical periods. Phase speed ratio 0.99582. Accumulated phase lag 7.5244°.

Set a cavity ringing from specified initial fields

A one-dimensional lossless cavity consists of perfectly conducting parallel planes at $x = 0$ and $x = L$. Use a polarisation E_y, B_z and ignore lateral edges. With no sources inside, the incoming characteristic at each wall is set by the condition $E_y = 0$. Instead of injecting a pulse during the calculation, prescribe an initial field $E_y(x, 0) = E_0 \sin(\pi x / L)$ and $B_z(x, 0) = 0$. These data obey the tangential electric wall condition. They also obey the divergence constraints because the spatial variation is along x while the fields are transverse. Initial data that satisfy a wave equation but violate either of these checks would model a different physical preparation.

Let $k = \pi / L$ and try $E_y(x, t) = E_0 \sin(kx) F(t)$. The wave equation gives

$$F'' + c^2 k^2 F = 0.$$

The initial magnetic field has no spatial gradient and the source-free Maxwell–Ampère law therefore sets $F'(0) = 0$. It follows that $F = \cos \omega t$, with $\omega = ck$. Use Faraday's first-order equation to find the sign and phase of B_z, obtaining

$$E_y = E_0 \sin(kx) \cos \omega t,$$

$$B_z = -\frac{E_0}{c} \cos(kx) \sin \omega t.$$

For a first-order time solver this pair gives a stronger check than the electric standing-wave frequency alone. A code with an incorrect sign in one curl can reproduce an electric snapshot but will transfer energy incorrectly on the next advance. The signs of the magnetic fields at the two ends are opposite.

Their amplitudes peak one quarter period after the initial electric maximum. Perfectly conducting walls carry the necessary surface currents. The standing solution has no Poynting flux through them because the tangential electric field at each wall vanishes.

Per unit transverse area the total initial energy is $\epsilon_0 E_0^2 L/4$, since the integral of $\sin^2(kx)$ over the cavity is $L/2$. At a general time the electric share is that total multiplied by $\cos^2 \omega t$, and the magnetic share is the same total multiplied by $\sin^2 \omega t$. The integral of their sum is constant, although energy continually flows within the cavity. For $L = 0.30\,\mathrm{m}$ the fundamental frequency is $c/(2L) = 499.65\,\mathrm{MHz}$ and its quarter period is about $0.500\,\mathrm{ns}$. If $E_0 = 100\,\mathrm{V/m}$, the stored energy is $6.64\,\mathrm{nJ/m^2}$ and the peak induction is $0.334\,\mu\mathrm{T}$.

The experiment can also reveal the discrete wave speed. A staggered one-dimensional mesh with spacing $h = L/10$ and Courant number $C = c\Delta t/h = 0.5$ obeys $\sin(\omega_{\mathrm{num}}\Delta t/2) = C\sin(kh/2)$. The predicted frequency ratio is $2\arcsin[C\sin(kh/2)]/(Ckh)$, approximately 0.9969 here. Measuring too short a time interval to distinguish this phase lag from sampling uncertainty would give little information about spatial accuracy. Track many periods, use consistent time-centred field estimates, and halve h at fixed C. At these small kh, the leading phase error should decrease by roughly a factor of four.

A student prepares a mixture $E_y(x,0) = E_0[\sin(\pi x/L) + 0.30\sin(3\pi x/L)]$, still with initial $B_z = 0$. What should a midplane electric probe observe, and what fraction of the initial total energy is in the third mode?

▼ **Work through the solution**

At the midplane the fundamental sine equals $+1$, while the third sine equals -1. The probe therefore reads $E_0[\cos(\omega t) - 0.30\cos(3\omega t)]$. Integrated over the full cavity, cross terms vanish because sine modes of different integer index are orthogonal. The third mode's energy fraction is

$$0.30^2/(1 + 0.30^2) = 0.0826.$$

A single spatial probe samples mode amplitudes times their local shape and cannot identify an energy fraction directly. At the midplane all even modes have a node and would escape detection even if they contained substantial energy.

A solver allows an arbitrary spatially uniform constant to be added to B_z while leaving the computed oscillating electric field unchanged. Is this necessarily a bug?

▼ **Work through the solution**

A constant transverse magnetic induction has zero curl and divergence in the ideal interior. It can be added as a static bias if the physical initial magnetic field and the exterior wall-current preparation permit it. In the test posed above the initial magnetic field was explicitly specified as zero, so a later spontaneously generated offset would be a numerical error. Monitoring the spatial mean of B_z provides a direct test: integrating Faraday's law over the interval gives its rate of change proportional to the

difference of the wall electric fields, which is zero. This check tracks an initial-data constraint invisible to a comparison of electric oscillation frequencies alone.

Hermann Haus and James Melcher, Electrodynamic Fields, on wave propagation and standing waves.

Tutorial on initial and boundary compatibility

A vacuum pulse has at its centre $E_y = 100\,\mathrm{V/m}$ and $cB_z = 60\,\mathrm{V/m}$. If both fields have the same spatial shape, infer the two travelling electric amplitudes and the ratio of the energies associated with the two directions.

▼ Work through the solution

The characteristics are $w_+ = 160$ and $w_- = 40\,\mathrm{V/m}$. Since $E = (w_+ + w_-)/2$, the component electric amplitudes are 80 and 20 V/m. Their separate directional energy ratio is $160^2/40^2 = 16$, assuming identical envelopes. The total local flux initially points right although neither characteristic is zero.

A numerical field in a periodic box starts with $\nabla \cdot \boldsymbol{B} = 10^{-4}$ in some cells. The code uses a discrete divergence of its discrete curl that is identically zero and carries that pattern unchanged. Does reaching a small time-step tolerance guarantee physically acceptable initial conditions?

▼ Work through the solution

No. The curl evolution preserves its initially supplied divergence. A compatible projection or constrained initialisation must remove the error in the correct discrete divergence norm, subject to boundary conditions and any physical mean field. Refining the time integration of inconsistent initial data leaves inconsistency.

A detector sees an amplitude echo 0.15 of an incident Gaussian pulse after reflection at a passive boundary. What is the reflected energy fraction and how much of the input energy remains unaccounted for if measured transmitted energy is 92 percent?

▼ Work through the solution

Same-medium, same-shape electric amplitudes give $0.15^2 = 0.0225$ reflected energy. With 0.92 transmitted, the remainder is $1 - 0.0225 - 0.92 = 0.0575$ or 5.75 percent. It may include absorption, energy still in the structure, lateral flux or measurement error. A completed energy audit distinguishes these possibilities.

Measure the phase speed of a numerical electromagnetic wave

Numerical field laboratory

Start with a lossless Maxwell wave, build a second-order time update and compare its actual travel time with what the continuum wave equation predicts.

A pulse can arrive late without losing energy

In a uniform lossless one-dimensional material take electric field along y, magnetic field along z and propagation along x. Maxwell's curl laws reduce to

$$\epsilon E_t = -H_x,$$
$$\mu H_t = -E_x.$$

Differentiate the first in time and substitute the second to obtain

$$E_{tt} = c^2 E_{xx},$$
$$c = 1/\sqrt{\mu\epsilon}.$$

A real laboratory signal is a combination of spatial frequencies. Different errors in travel time for different components can distort its shape even when the scheme does not attenuate any resolved Fourier mode.

Use periodic boundaries so that reflection from an end cannot be confused with this effect. Scale distance by one wavelength and time by one exact period. The periodic field $E(x,t) = \sin[2\pi(x-t)]$ has unit amplitude and returns to its initial shape after every integer period. A probe comparing just its envelope would miss a discrepancy in phase.

Derive the discrete dispersion relation

Take points separated by h and times separated by Δt. Centred second differences give

$$u_j^{n+1} = 2u_j^n - u_j^{n-1} + C^2(u_{j+1}^n - 2u_j^n + u_{j-1}^n),$$

with $C = c\Delta t/h$. Insert the complex trial field $e^{i(j\theta - n\Omega)}$. The time pair contributes $2\cos\Omega - 2$, the spatial pair $2\cos\theta - 2$. The physical positive-frequency branch obeys

$$\sin(\Omega/2) = C\sin(\theta/2),$$
$$\frac{c_{\text{phase,grid}}}{c} = \frac{\Omega}{C\theta}.$$

If $0 < C \le 1$ a real Ω exists for every grid phase, and each Fourier mode has constant discrete amplitude in exact arithmetic when started on one travelling branch. For $C > 1$ sufficiently rapid spatial

oscillations have no real frequency. One time factor then grows. Testing stability on a single very smooth wave can miss this unstable part of the supported spectrum.

For long waves expand $\sin$ and $\arcsin$:

$$\Omega = C\theta + C(C^2 - 1)\theta^3/24 + O(\theta^5).$$

The relative phase-speed error is $(C^2 - 1)\theta^2/24$ at leading order. Here increasing C towards one improves wave propagation. This special cancellation depends on a uniform one-dimensional grid, constant wave speed and the selected difference operators.

Initialise a clean numerical experiment

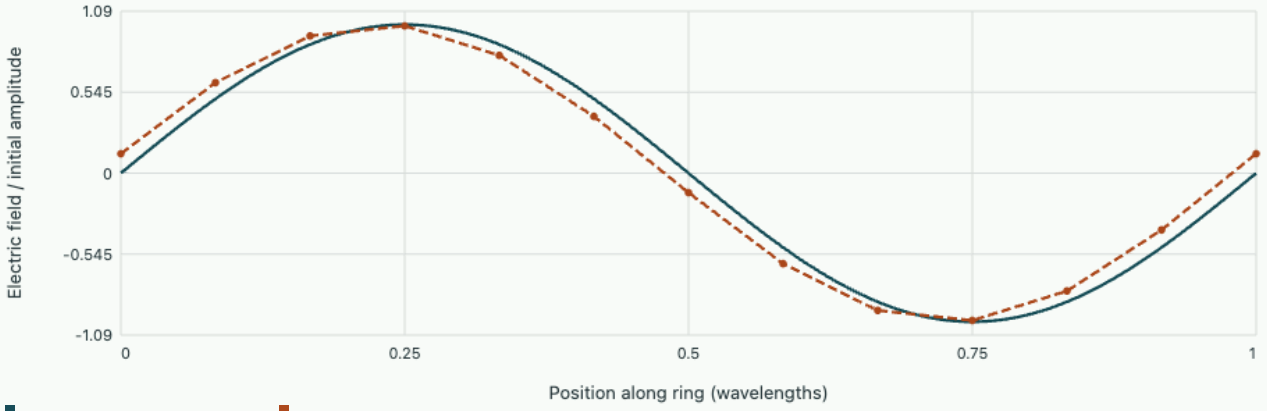
A second-order recurrence needs two time levels. Taking $u_j^0 = \sin(j\theta)$ and $u_j^{-1} = \sin(j\theta + \Omega)$ excites a single right-going numerical mode. Using the exact continuum phase $C\theta$ at the second level instead introduces a small mixture of right- and left-travelling grid modes. That is a valid test of initialisation, but it mixes errors if the quantity of interest is solely the numerical dispersion relation. State in your notes how the second level was produced.

Open the live experiment

Try the calculation

Points per wavelength **12**

Courant number **0.8**



75 steps, 5 physical periods. Phase speed ratio 0.99582. Accumulated phase lag 7.5244°.

The browser actually advances the nodal difference equations. It then reports the dispersion prediction as an independent check. The time is chosen as an integer number of numerical steps close to five exact periods and is stated with the result. At twelve points per wavelength and $C = 0.8$ there are

fifteen steps per period. Both exact and discrete solutions have the same amplitude to roundoff. At the end of five periods examine how far their zero crossings have separated.

From a phase test to a pulse test

For a narrow packet, envelope motion is set by group velocity $d\omega/dk$. Differentiate the discrete relation at fixed $h, \Delta t$. The result is $v_g/c = \cos(\theta/2)/\sqrt{1 - C^2 \sin^2(\theta/2)}$ on the positive branch for resolved nonendpoint modes. At $C = 0.8$ a mode close to the grid Nyquist limit has small group speed. A wave carrier and its envelope can show different speed errors. In a real dielectric wavelength dependence may also come from the material itself; a grid-refinement test with unchanged material law separates these contributions.

As an independent benchmark, set $C = 1$. Then $\Omega = \theta$ for $0 \le \theta \le \pi$, and the right-going recurrence translates the samples exactly by one grid cell at each time step. The lab should align both curves even on a surprisingly coarse grid for this particular smooth periodic mode. This checks the sign and indexing of propagation without implying exact accuracy for arbitrary geometry and material interfaces.

Tutorial

A signal must accumulate less than 1 degree of phase error over ten periods. Estimate a required points-per-wavelength count for $C = 0.5$ using the leading error.

▼ **Work through the solution**

The total continuum phase advance is 20π rad. Requiring $20\pi(1 - C^2)(2\pi/N)^2/24 < \pi/180$ gives $N^2 > 450\pi^2$, or $N \gtrsim 67$. Evaluate the exact dispersion relation to check the chosen integer.

Why is measuring only the field at its crest a poor detector of small phase lag?

▼ **Work through the solution**

The derivative of a sinusoid vanishes at its crest. A small phase shift changes that value only quadratically; at a zero crossing the change is linear. A phase fit using the entire ring is more robust than either single sample.

The numerical envelope changes after a wall is introduced. What additional tests would separate dispersion from reflection errors?

▼ **Work through the solution**

First validate travel time on a uniform periodic domain. Next use a single known interface with an analytical reflection coefficient. Move the observation plane and refine the mesh separately. Compare

reflected phase and amplitude with an energy-flux balance so a boundary error cannot masquerade as bulk dispersion.

Action density, stress waves and the boundary that reflects them

Analytical mechanics from constraints to invariants

A chain of coupled masses becomes an elastic field when the wavelength is long compared with the spacing. Derive the field equation and its boundary traction from the same variation, follow energy and stress across a junction, and decide whether a measured pulse is limited by material waves or by numerical dispersion.

Pass from a discrete rod to an action density

Let equal masses of spacing a move longitudinally by $u_j(t)$, with nearest-neighbour spring constant k_s and mass m_s per site. Their Lagrangian is $\sum_j [m_s \dot{u}_j^2/2 - k_s(u_{j+1} - u_j)^2/2]$. For variations smooth on the spacing, $(u_{j+1} - u_j)/a \to u_x$ and $\sum_j a\, f_j \to \int f(x)\, dx$. Hold the line density $\rho A = m_s/a$ and axial rigidity $EA = k_s a$ fixed. The continuum Lagrangian density per unit reference length is

$$\mathcal{L} = \rho A\, u_t^2/2 - EA\, u_x^2/2,$$

with A area, ρ volume density and E Young modulus. Both terms have units of joules per metre. This is a slender-rod model with small strain, primarily uniaxial stress and wavelengths long compared with the cross-sectional dimensions.

Vary the field in the action over $[0, L] \times [t_1, t_2]$. Integration of $\rho A\, u_t \delta u_t$ by parts in time, with fixed endpoint fields, produces $-\rho A\, u_{tt} \delta u$. Integration of $-EA u_x \delta u_x$ by parts in space produces $+\partial_x(EA u_x)\delta u$ and a remaining term $-[EA u_x \delta u]_0^L$. Thus the interior equation is $\rho A u_{tt} = \partial_x(EA u_x)$, and the end force conjugate to displacement is axial force $N = EA u_x$. Adding external work $F_L(t)\delta u(L)$ at the right face yields

$$N(L) = F_L.$$

A prescribed displacement instead imposes $\delta u(L) = 0$, leaving its reaction as an output. Keeping this boundary term prevents prescribing displacement and independent traction simultaneously at the same end of a second-order rod.

Determine the energy flux and characteristic impedance

Assume time-independent coefficients. Multiply the field equation by u_t and rearrange to obtain

$$\partial_t[\rho A u_t^2/2 + EA u_x^2/2] + \partial_x[-EA u_x u_t] = 0.$$

The energy flux is $-N u_t$, with positive sign denoting energy moving towards increasing x. The minus sign follows the convention of positive tensile N: positive mechanical power entering through a right face is $N u_t$, which is inward with respect to the rod on its left. This identity controls energy transport across both smooth material changes and sharp joints.

For a homogeneous right-moving wave $u = F(x - ct)$, $c = \sqrt{E/\rho}$, particle velocity $v = u_t = -cF'$ and axial force $N = EAF' = -Zv$, where $Z = \rho Ac = EA/c$. A left-moving wave has $N = +Zv$. The impedance is force divided by particle velocity and has units $\mathrm{kg/s}$. If one instead uses stress, impedance is ρc per unit area. Doubling the area of the same material keeps wave speed unchanged but doubles the force–velocity impedance. A transit-time measurement and a reflection measurement therefore provide independent material or geometry information.

For a rod with $E = 70\,\mathrm{GPa}$, $\rho = 2700\,\mathrm{kg/m^3}$ and $A = 1.0 \times 10^{-4}\,\mathrm{m^2}$, $c \approx 5092\,\mathrm{m/s}$ and $Z \approx 1375\,\mathrm{kg/s}$. A right-going compressive pulse with particle speed $0.020\,\mathrm{m/s}$ carries axial force $-27.5\,\mathrm{N}$, compressive stress $0.275\,\mathrm{MPa}$, and instantaneous power $Zv^2 = 0.550\,\mathrm{W}$ across a section. A rectangular approximation lasting $100\,\mu\mathrm{s}$ carries about $55\,\mu\mathrm{J}$ and occupies $0.509\,\mathrm{m}$. The pulse speed is kilometres per second; the material particles move only millimetres per second. Treating these as the same velocity would spoil force, impulse and energy estimates.

Match a bonded interface and distinguish amplitude from power

At a perfect massless bond displacement and axial force are continuous. For an incident right-going velocity amplitude v_i in rod 1, let v_r be reflected left-going velocity and v_t transmitted right-going velocity. Velocity continuity gives

$$v_i + v_r = v_t.$$

Force continuity gives

$$-Z_1 v_i + Z_1 v_r = -Z_2 v_t.$$

Thus

$$r_v = \frac{Z_1 - Z_2}{Z_1 + Z_2},$$

$$t_v = \frac{2Z_1}{Z_1 + Z_2},$$

$$\mathcal{R} = r_v^2,$$

$$\mathcal{T} = \frac{Z_2}{Z_1} t_v^2 = \frac{4Z_1 Z_2}{(Z_1 + Z_2)^2}.$$

The power fractions add to one for real positive impedances. The reflected force coefficient has opposite sign to the reflected velocity coefficient because the wave direction reverses. At a free end $Z_2 \to 0$, reflected velocity equals incident velocity, so end velocity doubles while total force vanishes. At a rigid end $Z_2 \to \infty$, velocity reverses and force doubles. The sign a strain gauge reports can therefore differ from that of a velocity-sensitive optical measurement at the same reflection.

A step from the worked rod into one of the same material with twice the cross-sectional area gives $Z_2 = 2Z_1$, $r_v = -1/3$, $t_v = 2/3$, and power fractions $1/9$ and $8/9$. A reduced transmitted velocity amplitude does not imply that most energy is reflected. If the interface includes a concentrated mass

M_b, force jumps instead obey $N_2 - N_1 = M_b \dot{v}_b$ with signs determined by forces exerted on that mass from either side. At high frequency the reaction of that mass can mimic a much stiffer termination. Frequency-dependent reflection is then an expected boundary effect, rather than evidence that the bulk modulus varies with frequency.

Connect continuum modes to their lattice predecessor

For a periodic chain the exact discrete equation is

$$m_s \ddot{u}_j = k_s(u_{j+1} - 2u_j + u_{j-1}).$$

Substitute $u_j = U e^{i(kja - \omega t)}$ to obtain

$$\omega = 2\sqrt{k_s/m_s}\,|\sin(ka/2)|.$$

In the first positive Brillouin zone, $\omega/(ck) = \sin(ka/2)/(ka/2)$ and group velocity $d\omega/dk = c\cos(ka/2)$, where $c = a\sqrt{k_s/m_s}$. The continuum rod retains only the leading low-ka term. A wave with eight sites per wavelength has phase speed about $0.9745c$ and group speed $0.9239c$. A narrowband packet's peak can arrive later than the continuum prediction even when a single crest is timed fairly accurately.

A finite-difference code approximating the continuum rod can produce the same spatial dispersion with its mesh spacing in place of a physical lattice spacing. Distinguish them experimentally or numerically by changing the computational mesh while holding the physical structure fixed. Lattice effects persist as the numerical grid is refined; artificial spatial dispersion diminishes. A separate time-step study is needed if a time integrator also disperses the wave. Comparing both peak arrival and phase at two frequencies tests more of the dynamics than fitting a single transit time.

Derive transverse beam waves from curvature energy

A slender beam bending laterally is governed by a different elastic energy from the axial rod. Let $w(x,t)$ denote small transverse deflection of its neutral axis. Under the Euler–Bernoulli assumption that cross sections remain plane and normal to that axis, the axial strain a distance z from the neutral axis is approximately $-z w_{xx}$. Integrating the local elastic energy $E\epsilon_{xx}^2/2$ across the section gives $EI_A w_{xx}^2/2$ per unit length, where $I_A = \int z^2\,dA$ is a geometrical second moment of area, with units m^4. It is distinct from a mass moment of inertia. The kinetic contribution uses line density ρA as before.

Varying $\mathcal{L} = \rho A w_t^2/2 - EI_A w_{xx}^2/2$ requires integrating twice by parts in space. With constant coefficients the interior variation gives

$$\rho A w_{tt} + EI_A w_{xxxx} = 0.$$

The boundary terms are $-[EI_A w_{xx}\delta w_x] + [(EI_A w_{xx})_x \delta w]$. Thus slope variations pair with bending moments and displacement variations with shear forces, with signs set by the chosen face orientation. A clamped end prescribes both deflection and slope; a free end sets moment and shear to zero. A simply supported end prescribes deflection and zero bending moment. Using only one boundary condition per end, as for the axial rod, leaves the fourth-order beam problem underdetermined.

A sinusoid $e^{i(kx-\omega t)}$ gives

$$\omega = \sqrt{EI_A/(\rho A)}\, k^2.$$

Phase speed grows proportionally to wavenumber and group speed is twice phase speed in this model. A narrow transverse pulse in a long beam consequently spreads even in a perfectly homogeneous non-dissipative continuum. By contrast the axial rod was nondispersive in its long-wavelength continuum approximation. At high enough frequency the simple bending model omits rotary inertia, shear deformation and eventually three-dimensional elastic behaviour. Extrapolating its unbounded speed to arbitrary k would use the slender, low-frequency approximation far beyond its derivation.

For a simply supported beam with length L, the sine modes satisfy both end conditions and give

$$\omega_n = (n\pi/L)^2 \sqrt{EI_A/(\rho A)}.$$

For a rectangular section of width w_b and thickness d in the bending direction, $I_A = w_b d^3/12$ and $A = w_b d$. Doubling thickness doubles all the natural angular frequencies at fixed L, E, ρ, whereas doubling width alone leaves the ideal frequencies unchanged. The stiffness increases and so does the inertia. A beam with $L = 0.50\,\mathrm{m}$, $d = 5.0\,\mathrm{mm}$ and the earlier aluminium properties has $\sqrt{EI_A/(\rho A)} \approx 7.35\,\mathrm{m^2/s}$, giving fundamental $46.2\,\mathrm{Hz}$. The second bending mode is near four times that frequency, a clear difference from the approximately harmonic axial mode sequence.

Match a force pulse to a support spring and dashpot

A right termination exerts force on the rod through a dashpot with mechanical impedance Z_L. For a real positive dashpot impedance the reflected velocity fraction is $r = (Z - Z_L)/(Z + Z_L)$. Taking $Z_L = Z$ gives zero reflection at the termination even though energy leaves a finite rod: the dashpot dissipates the arriving power. With $Z_L = 3Z$ one obtains a reflected velocity of $-1/2$ of the incident value, a reflected energy fraction one quarter and absorbed power fraction three quarters. A free or rigid end also reflects energy, but it absorbs no net power in the ideal limit. A measured weak reflection alone is therefore not proof of weak damping inside the rod.

For a pure end spring of stiffness k_L attached to a rigid support, the velocity impedance is $Z_L = k_L/(i\omega)$ under $e^{i\omega t}$. It is imaginary, so the reflection coefficient has unit magnitude at every frequency and a frequency-dependent phase. At low frequency the spring impedance is large and the end looks nearly fixed. At high frequency its impedance is small and the end looks nearly free within the lumped approximation. An impulsive signal contains many frequencies and can return with a changed

waveform even without any transmitted or dissipated energy. A support combining a resistor-like dashpot and a spring gives both amplitude loss and phase variation.

A measurement with strain gauges at two positions can separate forward and backward waves if their path delays are resolved. In a narrow frequency band the two records are linear combinations of advancing and returning phase factors. If the gauge spacing corresponds to an integer number of half wavelengths, these combinations become linearly dependent and the separation is ill-conditioned. Change the spacing, use a broadband pulse, or combine a velocity measurement with a strain measurement at one station so that the two characteristic directions enter with opposite signs. The relation $v = (v_+ + v_-)$, $N = Z(-v_+ + v_-)$ then gives a direct local decomposition.

Derive sound propagation from mass and momentum conservation

The displacement-action argument has a fluid analogue. Take an initially uniform fluid of density ρ_0 and pressure p_0, moving little from equilibrium in one spatial dimension. Write small perturbations ρ', p', v. Linearised mass conservation is

$$\partial_t \rho' + \rho_0 \partial_x v = 0.$$

Momentum balance for a short fluid parcel is $\rho_0 \partial_t v = -\partial_x p'$. If compression over an acoustic cycle is approximately isentropic, $p' = c_s^2 \rho'$ with $c_s^2 = (\partial p / \partial \rho)_s$. Differentiate the first equation in time, eliminate the velocity derivative with momentum balance, and obtain $\partial_t^2 p' = c_s^2 \partial_x^2 p'$.

The thermodynamic derivative matters because pressure can respond differently when compression is slow enough for heat exchange. For a perfect gas with ratio of specific heats γ, isentropic perturbations give

$$c_s^2 = \gamma p_0 / \rho_0.$$

Taking an isothermal derivative would give p_0/ρ_0 instead. For $p_0 = 101\,\mathrm{kPa}$, $\rho_0 = 1.20\,\mathrm{kg/m^3}$, and $\gamma = 1.40$, the isentropic speed is about $343\,\mathrm{m/s}$, while the isothermal calculation gives $290\,\mathrm{m/s}$. A measured transit time can discriminate between these constitutive conditions. Wall heat exchange in a very narrow tube can alter which approximation is appropriate over a chosen frequency band.

For a right-moving sound wave $p' = \rho_0 c_s v$, with the pressure positive in compression. This differs in sign from the tensile-positive axial force used for the rod, so match physical force conventions before borrowing formulas across the two examples. The acoustic power per area is $p' v$. For sinusoidal pressure with rms amplitude $1.0\,\mathrm{Pa}$ in the stated gas, the mean intensity is

$$p_{\mathrm{rms}}^2 / (\rho_0 c_s) = 2.43 \times 10^{-3}\,\mathrm{W/m^2}.$$

A peak amplitude of $1.0\,\mathrm{Pa}$ would yield half this number. Writing rms or peak explicitly prevents a routine factor-of-two error when comparing a microphone measurement with an energy flux.

Determine whether a measured transit reveals a modulus or a geometry

Suppose two axially loaded slender rods have measured wave speeds agreeing within their uncertainties. That alone determines only the ratio E/ρ under the slender-rod assumptions. Measuring force–velocity impedance $Z = \rho A c$ and cross section as well gives $\rho = Z/(Ac)$ and $E = Zc/A$. With $Z = 1375\,\mathrm{kg/s}$, $c = 5092\,\mathrm{m/s}$ and $A = 1.0 \times 10^{-4}\,\mathrm{m^2}$, the inversion gives

$$\rho = 2700\,\mathrm{kg/m^3},$$

$$E = 70.0\,\mathrm{GPa}.$$

If area were overestimated by five percent, both inferred quantities would be underestimated by almost five percent while their ratio remained correct.

Now examine a thick elastic solid under a narrow compressional disturbance with wavelengths short compared with its transverse dimensions. Lateral deformation can be constrained by surrounding material. In an isotropic three-dimensional continuum the longitudinal plane-wave modulus is $\lambda + 2\mu_s$, where λ, μ_s are Lamé coefficients, rather than the uniaxial-stress Young modulus of a slender rod. The shear-wave speed satisfies $c_T^2 = \mu_s/\rho$ and longitudinal speed $c_L^2 = (\lambda + 2\mu_s)/\rho$. Measuring both, together with density, can recover the two independent elastic moduli. Using only one longitudinal speed to infer Young modulus without specifying the deformation geometry can give a physically incorrect value.

For $\rho = 2500\,\mathrm{kg/m^3}$, $c_L = 5000\,\mathrm{m/s}$ and $c_T = 3000\,\mathrm{m/s}$, these relations give

$$\mu_s = 22.5\,\mathrm{GPa},$$

$$\lambda = 17.5\,\mathrm{GPa}.$$

The implied Young modulus $E = \mu_s(3\lambda + 2\mu_s)/(\lambda + \mu_s) = 54.8\,\mathrm{GPa}$, while $\rho c_L^2 = 62.5\,\mathrm{GPa}$. Both numbers represent material stiffness, but they enter different mechanical boundary conditions. A long thin specimen of the same material should have an axial speed closer to $\sqrt{E/\rho} = 4680\,\mathrm{m/s}$. Comparing the bulk and rod experiments tests the constitutive and geometrical assumptions behind the measured speed.

Derive orthogonality when a rod changes material along its length

For a rod with positive varying density $\rho(x)$, area $A(x)$ and modulus $E(x)$, a separated free oscillation $u(x,t) = \phi(x)\cos\omega t$ satisfies

$$-[EA\phi']' = \omega^2 \rho A \phi.$$

Suppose two modes ϕ_n, ϕ_m satisfy the same homogeneous fixed or free boundary conditions and the correct continuity of displacement and axial force across every perfectly bonded interface. Multiply the

first equation by ϕ_m, the second by ϕ_n, subtract and integrate. Interior terms produced by integration by parts cancel pairwise, including the interface flux terms, and the endpoint terms vanish. Thus

$$(\omega_n^2 - \omega_m^2) \int \rho A \phi_n \phi_m \, dx = 0.$$

Mode shapes with distinct frequencies are orthogonal under the mass-weighted inner product, not generally under a simple average of displacements measured at uniformly spaced stations. An initial velocity distribution $v_0(x)$ projects onto mode n with coefficient $\int \rho A v_0 \phi_n \, dx / \int \rho A \phi_n^2 \, dx$. In a rod with a heavy section, the same imposed velocity there stores more energy and carries more modal weight. An experiment interpreting raw camera amplitudes without line-density weighting can misassign energy among modes. Normalising each measured mode shape to unit maximum does not fix the mass weighting needed for energies.

Multiplying one eigenvalue equation by its own shape gives

$$\omega_n^2 = \int EA(\phi_n')^2 dx / \int \rho A \phi_n^2 dx.$$

For an admissible trial function satisfying prescribed displacement conditions, the same quotient bounds the fundamental squared frequency from above when rigid translations are excluded. Consider a uniform fixed–free rod and trial $\phi = x/L$. The numerator is EA/L and denominator $\rho AL/3$, giving $\omega_{\mathrm{trial}} = \sqrt{3}\, c/L$. The exact fundamental is $\pi c/(2L)$, so this one-parameter shape overestimates frequency by about 10.3%. It need not satisfy the free-end traction condition itself because that is a natural condition of the variational minimiser; it does need to satisfy the essential fixed-end displacement.

For a free–free rod the constant displacement mode has zero frequency. Using an arbitrary positive trial function without removing its mass-weighted mean can drive a quotient towards zero by adding a large constant. To estimate the first deforming mode impose orthogonality to the rigid translation, $\int \rho A \phi \, dx = 0$. This has an immediate mechanical interpretation: the deformation mode should not translate the centre of mass of an isolated rod. The null mode and its mass-weighted constraint reappear in numerical eigenproblems built from discretised stiffness and mass matrices.

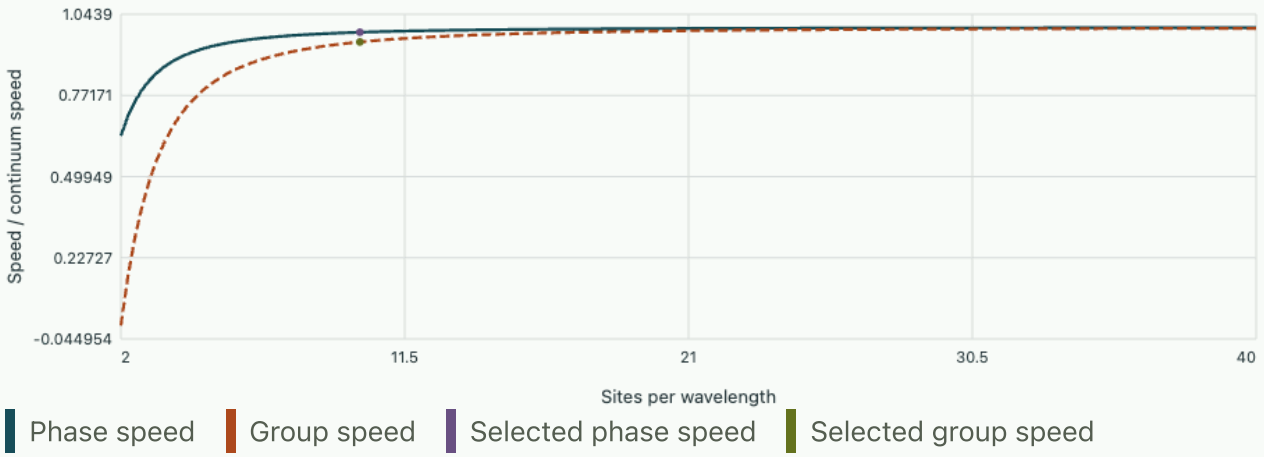
Find where a discrete elastic chain ceases to look continuous

Shorten the wavelength relative to the lattice spacing. Compare the phase and group velocities with those inferred from a continuum action.

Open this experiment online

Try the calculation

Sites per wavelength 10



Frequency and phase speed are 98.36% of the continuum prediction. Group speed is 95.11%.

Track work and impulse in a one-dimensional pulse experiment

A finite displacement pulse provides a useful warning about integrating strain and velocity records. A right-going strain pulse of one sign can carry net momentum $\int \rho A v \, dx$. Because $v = -c u_x$, its integrated material momentum is $-\rho A c[u(+\infty) - u(-\infty)]$ at a fixed time. A localised displacement bump that returns to the same displacement on both sides has zero net ordinary material momentum, even though its energy is positive and travels to the right. Its strain and particle-velocity records contain positive and negative parts. A unipolar velocity pulse leaves behind a displacement step, produced by a source that has supplied net impulse.

For a right-moving rectangular velocity pulse of amplitude $0.020\,\mathrm{m/s}$, duration $100\,\mu\mathrm{s}$ in the worked rod, the local displacement jump measured in time is $2.0\,\mu\mathrm{m}$. Its impulse magnitude across a section is $Zv\Delta t = 2.75 \times 10^{-3}\,\mathrm{N\,s}$, while its transmitted energy is

$$Zv^2\Delta t = 55.0\,\mu\mathrm{J}.$$

A second equal pulse of opposite velocity sign launched later cancels the net displacement and material momentum but doubles the transmitted energy. Thus integrating a signed velocity signal and integrating its square distinguish different properties of the same experimental record.

A displacement sensor has a small linear drift during a burst. Which inferred observables are particularly sensitive?

▼ **Work through the solution**

Differentiating adds a constant offset to particle velocity. Integrating that apparent velocity over a long window produces a growing spurious displacement jump and momentum estimate. Squaring it adds

both a cross term with the physical velocity and a positive baseline energy. Measure a pre-pulse baseline, choose an integration interval covering the physical pulse, and test that the result is stable when the window is moderately extended. An independent strain gauge gives another characteristic combination; comparison with $N = -Zv$ can expose drift or an overlapping reflection.

Can a finite-element rod model with a plausible first eigenfrequency still fail an impulse test?

▼ Work through the solution

The first frequency mostly probes a long-wavelength weighted stiffness-to-mass ratio. A brief pulse contains higher modes and is sensitive to temporal resolution, dispersion, boundary tractions and the way discrete source work enters. Verify modal frequencies at several wavelengths, check the pulse's positive energy and signed impulse separately, and compare the measured incoming and reflected characteristic fluxes. A single standing-wave eigenvalue cannot validate all of those physically independent balances.

Recover a jump condition at an interface with a point mass

For a concentrated mass M_b joining two rods include $M_b u_t(x_b)^2/2$ as a separate contribution to the Lagrangian. Its time variation contributes $-M_b u_{tt}\delta u$ at the joint. The left and right rod boundary terms contribute respectively $-N_1\delta u$ and $+N_2\delta u$. Requiring stationarity for an independently varying shared displacement gives exactly $N_2 - N_1 = M_b u_{tt}$. This derivation identifies both the sign and the constitutive assumption of a perfectly bonded, nonslipping joint.

For identical rods and a sinusoidal drive at angular frequency ω, the reflected velocity coefficient becomes $r = -i\omega M_b/(2Z + i\omega M_b)$ under the stated phasor convention. At $\omega M_b = 2Z$ it reflects half the power and transmits half. The mass itself absorbs no cycle-averaged energy, though it stores kinetic energy during a cycle. For $Z = 100\,\mathrm{kg/s}$ and $M_b = 0.10\,\mathrm{kg}$, the equal-power frequency is about $318\,\mathrm{Hz}$. A real joint that loses appreciable power at that frequency needs a damping or radiation term in addition to its measured mass.

Tutorial on stress waves

A gauge $0.80\,\mathrm{m}$ from a free end records the echo of a sharp disturbance in the worked rod. What echo delay and strain sign should it expect?

▼ Work through the solution

The extra path from gauge to end and back is $1.60\,\mathrm{m}$, giving $1.60/5092 = 0.314\,\mathrm{ms}$. A free end has zero net axial force, so the reflected force and strain have opposite signs to the incident ones. The velocity reflection keeps its sign. Finite pulse width and any change of material between gauge and end can obscure that simple result.

A bonded step has measured reflected velocity coefficient -0.20 and no detectable dissipation. Infer its impedance ratio and transmitted energy fraction.

▼ **Work through the solution**

Rearranging $r_v = (1 - r)/(1 + r)$ gives

$$r = Z_2/Z_1 = (1 - r_v)/(1 + r_v) = 1.50.$$

The energy reflection is 0.04 and transmission 0.96. The velocity transmission is 0.80, whose square alone is only 0.64. Multiplying by the transmitted-to-incident impedance ratio recovers the correct energy fraction.

A rod model with a prescribed displacement at its right end reports nonzero work there, even though the variational boundary term vanished in deriving its equation. Is there a contradiction?

▼ **Work through the solution**

Vanishing of $N\delta u$ follows because the allowed virtual displacement at a prescribed boundary is zero. Actual boundary velocity can be nonzero when the prescribed displacement varies in time. The actuator supplies power Nu_t at the right boundary. Compare the time-integrated input with the change of stored energy plus any losses or outgoing energy rather than identifying an admissible variation with actual displacement.

Compare a finite chain with its long-wavelength continuum.

Fließbach includes string vibrations, beam bending and sound in his continuum mechanics treatment.

From binary collisions to the Boltzmann equation

Statistical mechanics from microscopic counting to material response

An initially anisotropic velocity distribution can relax even though every individual collision conserves energy and momentum. Describe a dilute gas with a phase-space density, derive the structure of its collision term and prove which distribution can stop producing entropy. Then examine the assumptions behind that proof and the measurements most likely to reveal their failure.

What the one-particle function counts

Let $f(\mathbf{r}, \mathbf{v}, t) d^3r\, d^3v$ be the expected number of particles in a small phase-space cell. Integrating over velocities gives number density n, and the mean velocity is $\mathbf{u} = n^{-1} \int \mathbf{v} f\, d^3v$. Define the peculiar velocity $\mathbf{c} = \mathbf{v} - \mathbf{u}$; its second moment sets local kinetic energy after removing the organised flow. A gas can have large bulk speed but low temperature.

Follow a particle ballistically over a short interval under acceleration $\mathbf{a} = \mathbf{F}/m$. With no collisions, phase-space probability is advected along its trajectories. For a standard position-dependent force, phase-space flow has zero divergence, giving the streaming operator $\partial_t f + \mathbf{v} \cdot \nabla_r f + \mathbf{a} \cdot \nabla_v f$. Magnetic Lorentz acceleration also has zero velocity-space divergence. A dissipative velocity-dependent drag generally requires the divergence form $\nabla_v \cdot (\mathbf{a} f)$ instead. It should not silently be inserted into the Hamiltonian streaming expression.

We now assume a dilute classical gas of identical particles with short-ranged binary elastic collisions, approximated as local on the spatial scale resolved by f. Different particles approaching collision are assumed uncorrelated apart from their individual one-particle distributions at that position. This molecular-chaos approximation is an assumption about incoming pairs, not a claim that the outgoing particles remain uncorrelated. It is the step that allows the evolution to close on a one-particle function.

Count pairs crossing a scattering area

Consider a tagged incident velocity $\mathbf{v}$ and a possible partner $\mathbf{v}_1$. Their relative speed $g = |\mathbf{v} - \mathbf{v}_1|$ converts a differential scattering cross section $d\sigma/d\Omega$ into the swept interaction volume per unit time. A binary collision rotates the relative velocity in the centre-of-mass frame while preserving its magnitude. With equal masses, set $\mathbf{V} = (\mathbf{v} + \mathbf{v}_1)/2$ and $\mathbf{g} = \mathbf{v} - \mathbf{v}_1$. The partner and tagged postcollision velocities are $\mathbf{V} \pm \mathbf{g}'/2$, where $|\mathbf{g}'| = |\mathbf{g}|$. This parametrisation enforces momentum and kinetic energy conservation before choosing a scattering direction.

The loss term counts incoming pairs with velocities $\mathbf{v}, \mathbf{v}_1$. The gain counts inverse collisions that produce the specified velocity. Microreversibility of elastic scattering and invariance of the integration measure allow the compact expression

$$C[f](\mathbf{v}) = \int d^3v_1\, d\Omega\, g \frac{d\sigma}{d\Omega} \left[f(\mathbf{v}') f(\mathbf{v}_1') - f(\mathbf{v}) f(\mathbf{v}_1) \right].$$

Both products refer to the same position and time. A change in which pair is called primed merely swaps gain and loss if one reverses the parametrisation consistently. This formula uses a consistent

distinct-partner convention for the full scattering measure. When counting the total number of collision events per volume rather than the loss of tagged particles, divide the sum over all ordered pairs by two. Confusing those two observables gives an erroneous collision frequency.

For hard spheres of molecular diameter d, the total geometric cross section is πd^2. In an equilibrium Maxwellian the mean relative speed is $\sqrt{2}$ times the mean single speed for identical masses. The mean free path under the usual hard-sphere convention is $\ell = 1/(\sqrt{2} n \pi d^2)$. For $d = 0.36$ nm, $n = 2.4 \times 10^{25}$ per cubic metre and molecular mass 28 unified atomic units at 300 K, $\ell = 72.4$ nm and $\bar{v} = 476$ m/s. Their ratio gives a characteristic flight time 0.152 ns.

Take a weak moment of the collision operator

For a test function $\psi(\mathbf{v})$, multiply the collision term by ψ and integrate over tagged velocity. Exchange the identities of both particles and then exchange incoming and outgoing velocities in the gain integral. The result expresses the change of the pair's total test-function value:

$$\int \psi C[f] \, d^3v = \frac{1}{2} \int d^3v \, d^3v_1 \, d\Omega \, g \frac{d\sigma}{d\Omega} f f_1 [\psi' + \psi_1' - \psi - \psi_1].$$

For $\psi = 1$, $m\mathbf{v}$ and $mv^2/2$ the square bracket vanishes collision by collision. Integrating the kinetic equation gives exact local conservation of mass, momentum and energy with appropriate streaming fluxes, regardless of how close the distribution is to Maxwellian. Conservation alone leaves a hierarchy: the momentum flux depends on the pressure tensor and the energy flux on a heat flux, both higher velocity moments.

A numerical collision solver should check all these invariants separately. Equal masses allow a direct test with velocities $(300, 100, 0)$ and $(-100, 100, 0)$ m/s. Their centre velocity is $(100, 100, 0)$ and relative velocity has magnitude 400 m/s. Rotating it from x to y gives outgoing velocities $(100, 300, 0)$ and $(100, -100, 0)$. Total velocity and total squared speed agree before and after, but the directional kinetic energy has changed. A field of such collisions can isotropise the pressure tensor while leaving total energy fixed.

Derive a sign-definite entropy production

Define the local kinetic H density as $\int f \log(f/f_\star) d^3v$ with a constant reference of matching units. The reference only changes a conserved number contribution. For a homogeneous isolated gas the collision contribution to its time derivative follows by applying the weak form with $\psi = \log f$. Average the result with its incoming–outgoing interchange. Setting $A = f f_1$, $B = f' f_1'$, one obtains

$$\frac{dH}{dt} = -\frac{1}{4} \int d^3v \, d^3v_1 \, d\Omega \, g \frac{d\sigma}{d\Omega} (A - B) \log(A/B) \le 0.$$

For positive A, B, logarithm and difference have the same sign; appropriate limits handle zero occupation. Thus the Boltzmann entropy $-k_B H$ does not decrease from these local binary collisions. In

an inhomogeneous region, advection additionally transports entropy through the boundaries. A local cell's kinetic entropy can drop because entropy leaves the cell, while collision production remains nonnegative.

Equality requires $ff_1 = f'f_1'$ for collisions connected by the scattering kernel. Equivalently $\log f + \log f_1$ is a collision invariant. For sufficiently rich elastic scattering in three dimensions the smooth additive collision invariants are linear combinations of particle number, momentum and energy. Therefore $\log f = a + \mathbf{b}\cdot\mathbf{v} + cv^2$, with $c < 0$ for normalisation. Completing the square and matching moments yields the local Maxwellian:

$$f_M = n\left(\frac{m}{2\pi k_B T}\right)^{3/2} \exp[-m|\mathbf{v}-\mathbf{u}|^2/(2k_B T)].$$

A Maxwellian at each location makes the collision term zero locally, but spatial gradients usually leave a nonzero streaming term. Such a distribution can evolve because neighbouring cells exchange particles. Conversely, a truly global equilibrium in an external conservative potential includes the appropriate spatial Boltzmann factor. In one-dimensional equal-mass elastic collisions particles merely exchange velocities, so the assumed richness of three-dimensional scattering fails; an arbitrary velocity histogram need not thermalise through those collisions.

Assess the kinetic and hydrodynamic descriptions separately

Define Knudsen number $Kn = \ell/L$ using the length over which the distribution or boundary condition changes. In the earlier gas $Kn = 7.2 \times 10^{-5}$ for a 1 mm chamber feature but 0.072 for a 1 micrometre channel gap. The same binary collision model can remain dilute in both, yet a no-slip bulk hydrodynamic closure may be much less accurate in the microchannel. A mean free path small relative to the overall device length is insufficient if a sharp boundary layer develops on a much shorter scale.

A useful relaxation-time collision approximation is $C_{\mathrm{BGK}} = -(f - f_M)/\tau$, with f_M computed from the local conserved moments of f. This enforces their conservation and models return to local Maxwellian without evaluating an angular integral. If an externally fixed Maxwellian is used irrespective of actual local density, momentum and energy, the approximation can remove these quantities spuriously. A single fitted relaxation time also need not reproduce both stress and heat transport for a real gas.

A particle simulation resolving binary encounters must choose spatial cells small enough to represent local variation, and event or collision timesteps short against relevant flight or collision times. Statistical noise decreases with independent sampled pairs; refining cells while keeping total simulated particle count fixed can make local collision estimates noisier. Compare conservation, relaxation of anisotropic moments and a transport observable under separate refinements of geometry and sample count. A correct mean pressure alone is a weak test of a collision operator.

Compare the flux-weighted velocities of escaping particles with the bulk velocity distribution

Take the next moment and exhibit the source of viscosity

Write momentum density as $\rho u_i = m \int v_i f \, d^3v$. The flux of i-momentum through a surface with normal in direction j is $m \int v_i v_j f \, d^3v = \rho u_i u_j + P_{ij}$, where $P_{ij} = m \int c_i c_j f \, d^3v$. The scalar pressure is one third of its trace in three dimensions. Multiplying the kinetic equation by $m v_i$ and using collision invariance yields the momentum balance with $\partial_j P_{ij}$ explicitly. No constitutive assumption has yet set the off-diagonal components to zero or made them proportional to strain rate.

For a local Maxwellian, the odd moments of peculiar velocity vanish and $P_{ij} = n k_B T \delta_{ij}$. Now consider slow shear with mean velocity $u_x(y)$, and approximate collisions by a relaxation time τ. In a local coordinate frame moving with the fluid write $f \simeq f_M - \tau v_y \partial_y f_M$ to first order in gradient. At constant density and temperature,

$$\partial_y f_M = (m c_x / k_B T) f_M \partial_y u_x.$$

The off-diagonal momentum flux correction is therefore a fourth Gaussian velocity moment:

$$P_{xy} = -\tau \frac{m^2}{k_B T} \partial_y u_x \int c_x^2 c_y^2 f_M d^3v = -n k_B T \tau \, \partial_y u_x.$$

The integral factorises into variances $k_B T / m$ in the two directions. Under the kinetic pressure-tensor convention the result has a minus sign, because random motion transports high x-momentum toward lower mean velocity. The viscous part of the Cauchy stress is commonly written with the opposite sign convention. State which tensor enters the balance before declaring a predicted shear stress negative. The dynamic viscosity of this simple model is $\eta = p\tau$.

If $p = 1.0 \times 10^5$ Pa and the fitted stress relaxation time is 1.8×10^{-10} s, this gives 1.8×10^{-5} Pa s. At a shear rate of 10^3 per second the dissipative stress magnitude is 0.018 Pa, tiny compared with equilibrium pressure and with dimensionless shear rate $\tau |\partial_y u_x| = 1.8 \times 10^{-7}$. A stress measurement must resolve a small anisotropic contribution on top of a large isotropic pressure. A pressure gauge alone cannot test this kinetic viscosity relation.

In a dilute hard-sphere gas the mean collision time scales inversely with density. The simultaneous pressure factor then makes dynamic viscosity approximately independent of density at fixed temperature in the dilute continuum regime. The kinematic viscosity, η / ρ, still changes strongly. A gas made sufficiently rarefied no longer supports ordinary no-slip continuum flow through a fixed channel, so an apparent pressure-dependent flow resistance can coexist with an essentially density-independent bulk dynamic viscosity.

Infer accommodation from a wall momentum measurement

The interior collision operator does not prescribe the boundary condition at a solid wall. Specular reflection reverses normal velocity while preserving tangential velocity in the wall frame. Diffuse thermal re-emission from a fully accommodating stationary wall instead draws outgoing velocities from a wall-temperature distribution appropriate to the outward flux and erases incident tangential drift. A real clean

or rough surface may lie between these limits. The boundary scattering model matters particularly when particles encounter the wall as often as other particles.

For a simple moving plate in a dilute gas at the same temperature, suppose a fraction α of incident particles accommodates tangentially to plate speed U, and the rest reflects specularly. To leading order in small tangential drift, accommodated particles carry away an average tangential momentum increment mU if incident gas has zero mean tangential speed. The incident number flux on one plate face is $n\bar{v}/4$. Thus the leading tangential drag magnitude on the plate is $\alpha m U n \bar{v}/4$ per area in this ideal free-molecular geometry.

Take $n = 10^{21}$ per cubic metre, molecular mass 4.65×10^{-26} kg, $\bar{v} = 480$ m/s, $U = 2.0$ m/s and $\alpha = 0.80$. The predicted shear is about 8.93×10^{-3} Pa per exposed face. The effect scales linearly with density here, unlike the dilute continuum viscosity at a fixed macroscopic velocity gradient. Both results derive from momentum transport, but the first describes particles crossing a wall-bounded collisionless region and the second local shear sustained by many intermolecular collisions. Compare them only after specifying the Knudsen number and wall geometry.

To test the accommodation model, reverse the plate motion to remove static force offsets, vary density across a regime that remains free molecular and change the exposed surface while keeping area fixed. A single measurement gives the combination of accommodation fraction, incident number flux and velocity difference. It cannot separate a changing incident gas drift from a changing material accommodation coefficient. Measuring the upstream velocity distribution or using opposed plates with known relative motion supplies additional constraints.

A discrete collision quartet demonstrates the H theorem numerically

A single allowed reversible collision connects two incoming velocity cells to two outgoing cells. Suppose cell-volume and scattering factors have been normalised so that its net flux is $K(A - B)$, with $A = f_1 f_2$, $B = f_3 f_4$. It removes particles from cells one and two and adds them to three and four when $A > B$. Conservation requires the representative cell velocities to satisfy the same pair momentum and energy constraints used for a physical collision. Assigning arbitrary four velocities may conserve the number of particles but will generally create energy.

Choose the velocities from the earlier rotated-pair example and occupancies $f_1 = 4, f_2 = 1, f_3 = 2, f_4 = 1$ in a common dimensionless reference. Then $A = 4$, $B = 2$ and the forward flux is $2K$. Its contribution to dH/dt is $-2K \log 2$, negative. The exact continuous rate tends to zero as the products equalise. A forward Euler step larger than $f_2/(2K)$ would make the second occupancy negative and destroy the logarithm's meaning. Nonnegative entropy production for the continuum collision model does not guarantee it for an arbitrary numerical time step.

An implicit or positivity-preserving update can avoid this failure, but still requires independent tests of conserved moments and of the correct equilibrium. For this isolated quartet the condition $f_1 f_2 = f_3 f_4$ alone leaves many possible stationary cell distributions. A network of collision quartets sufficiently connecting the velocity space is needed to enforce a Maxwellian-like equilibrium. On a coarse discrete velocity lattice too few permitted momentum- and energy-conserving quartets may leave artificial

invariants. A code can display exactly conserved energy and monotonically decreasing H while relaxing to the wrong distribution because its collision graph is incomplete.

A velocity-selective aperture tests collision equilibration

A thin aperture feeding a vacuum produces a flux-weighted distribution from a local Maxwellian reservoir. A detector placed downstream generally selects a finite angular cone and may respond differently to fast and slow particles. Consider two narrow speed bins with the same width in speed and the same angular acceptance. For an effusive Maxwellian their flux ratio contains a cube of speed multiplied by a Boltzmann exponential. A bulk sampler instead has a square of speed. Confusing these prefactors shifts a temperature inferred from ratios even if the gas truly is locally equilibrated.

For bins at speeds v and $2v$, and choosing $mv^2/(2k_BT) = 1$, the effusive high-to-low ratio is

$$8e^{-3} = 0.398.$$

The bulk number-density ratio at those speeds would be $4e^{-3} = 0.199$. A downstream detector measuring particles present in a fixed optical observation volume may weight by travel time, introducing an additional approximate inverse speed that moves the response closer to density weighting. Distinguish arrivals counted per time from particles counted during an exposure.

Changing aperture diameter is an independent control. A broad opening or high upstream density can produce collisional expansion and a directed jet, altering both angular and speed distributions. Measure pressure and beam shape across aperture sizes and compare the small-aperture limit before attributing a non-Maxwellian downstream spectrum to a failure of local thermal equilibrium throughout the reservoir.

A flight-time measurement tests free paths and speed selection separately

Imagine a dilute molecular tracer introduced into a uniform bath. For a deliberately simplified constant collision-rate model with rate r, probability of no collision by time t is e^{-rt}. For tracers conditioned to have one speed v, the distribution of collision-free distance is then exponential with mean v/r. If the collision rate itself depends on speed through the average relative velocity against the bath, the conditional mean free path is $v/r(v)$. Mixing all initial speeds need not produce a single exponential in distance, even if every conditional waiting time is exponential.

Place a detector a distance L downstream that accepts only trajectories maintaining their initial direction. For a narrow injected speed class, the unscattered fraction is approximately $e^{-r(v)L/v}$ if any collision removes it from the acceptance. For a class at 500 m/s and mean free path 0.20 m, a detector at 0.10 m receives an unscattered fraction 0.607; at 0.30 m it receives 0.223. A ratio of the two readings removes the unknown injection strength and recovers the attenuation coefficient if detector acceptances and geometric factors match.

If forward-scattered trajectories remain inside the acceptance, the apparent attenuation is smaller than the total collision rate. Angularly narrowing the receiver should approach the total-removal model for suitable scattering. A finite source pulse and finite time-of-flight bin also select velocities even if the source initially emitted a Maxwellian. Fit survival separately in narrow speed bins first. A nonexponential whole-beam attenuation then need not imply that the microscopic collision process has memory; it can arise purely from an unresolved distribution of rates.

A simulation can distinguish these alternatives by logging every collision event and comparing the full collision count with the detector-defined removed count. First test conditional survival for a prepared monochromatic beam, then add a known distribution of incident speeds and calculate the mixture prediction explicitly. Finally include a finite angular acceptance and integrate the differential cross section over the appropriate retained directions. These steps separate the kinetic collision operator, the initial distribution and the observation rule.

The same logic applies to a relaxation experiment reporting a single characteristic mean free time. The decay time of a directional flux depends on how much momentum is deflected per collision, not simply on how often particles meet. Strongly forward-peaked scattering can have a large total cross section and a much smaller momentum-transfer cross section weighted by $1 - \cos\theta$. A measured attenuation and a measured viscosity may consequently constrain different moments of precisely the same differential scattering process.

For example, if every collision deflects a test particle through a small fixed angle of ten degrees in the relevant frame, the momentum-transfer weighting is

$$1 - \cos 10^\circ = 0.0152.$$

The corresponding transport cross section can then be far smaller than the total cross section even when collisions are frequent. A velocity-randomising approximation that uses the total collision time as its stress-relaxation time would predict much faster momentum relaxation than this angular picture supports. Actual stress and energy transport involve additional angular weights and velocity averages, so this numerical comparison is a scale diagnostic rather than an exact viscosity formula. Measure several angular moments of the differential scattering kernel or compare at least two independently sensitive transport observables before assigning a universal relaxation time to a real collisional gas.

Follow an H functional during a reversible elementary collision

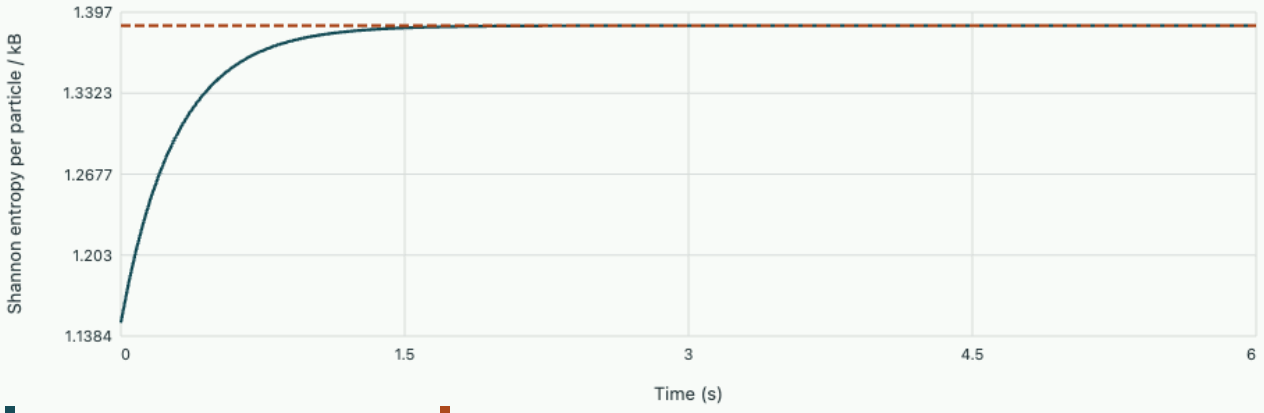
Prepare unequal populations in two velocity pairs. Count forward and reverse collision rates and compare the sign of entropy change with the population imbalance.

Open this experiment online

Try the calculation

Initial fraction in velocity pair A **0.83**

Collision rate coefficient (s⁻¹) **0.8**



Initial clockwise-minus-reverse pair imbalance x²−(1−x)² = 0.66. Initial d(S/kB)/dt = 0.83721 s⁻¹.

Tutorial · Find the equilibrium temperature of an anisotropic preparation

A homogeneous gas has zero mean flow and component kinetic temperatures 600, 300 and 150 K, defined by $m\langle v_i^2\rangle = k_B T_i$. In an isolated elastic gas, predict its equilibrium temperature and pressure tensor change. Could a thermostat holding the velocity variance in one direction at its original value permit the same inference?

▼ **Retain all three conserved energy contributions**

Conservation of total kinetic energy implies $T_f = (600 + 300 + 150)/3 = 350$ K. Initially the diagonal kinetic pressure components are nk_BT_i; ultimately each becomes $nk_B(350\,\mathrm{K})$. Their sum is unchanged. A component singled out by a thermostat can exchange work or heat with its environment. Its final common temperature, if one exists, must be obtained from the full energy balance including that exchange. Isotropisation cannot simply be assumed to retain the initial isolated mean.

Tutorial · Estimate the error in a collision-frequency inference

A detector reports a mean speed of 480 m/s and a geometric collision cross section 4.0×10^{-19} square metres at $n = 2.0 \times 10^{25}$ per cubic metre. Estimate the collision rate per tagged particle and total events per unit volume per second for an equilibrium identical-particle gas. What changes in a molecular beam with almost parallel velocities?

▼ **Use the relative rather than the single-particle speed**

The equilibrium mean relative speed is $\sqrt{2}(480) = 679$ m/s. The rate per tagged particle is $ng\sigma \simeq 5.43 \times 10^{9}$ per second. Multiplying by $n/2$ gives event rate 5.43×10^{34} per cubic metre per second. In a nearly monoenergetic collimated beam the relative speed can be much less than the mean laboratory speed. Evaluate its pair-distribution average explicitly; translating every velocity by a common vector leaves all collision rates unchanged at fixed distribution of relative velocities.

Tutorial · Does reversing all velocities contradict the H theorem?

In a dilute-gas simulation, reverse all particle velocities at a chosen instant with exact precision. An ideal reversible microscopic integrator would retrace the prior trajectory. Explain how this outcome can coexist with a kinetic equation predicting nondecreasing entropy, and propose a comparison that isolates the missing information.

▼ **State which distributions are controlled by the reversal**

The reversed incoming pairs carry precise correlations inherited from earlier outgoing collisions. Their two-particle distribution generally does not factor into the product of their one-particle distributions just before future collisions. The molecular-chaos step used to derive the collision operator therefore fails for this deliberately prepared history. Microscopic Hamiltonian evolution and the coarse one-particle closure make different state specifications.

Prepare a second gas with the same reversed one-particle velocity histogram but with spatial and partner correlations randomised while conserving total momentum and energy. Compare the two subsequent anisotropy histories. The kinetic equation has the same initial one-particle input for both, while a microscopic simulation can distinguish them through higher-order correlations. Imperfect numerical reversal introduces a separate error; test exactness and time-step dependence before interpreting a failure to retrace as a fundamental thermodynamic observation.

Cambridge Kinetic Theory · the Boltzmann equation, collision invariants and hydrodynamics

Transport coefficients from equilibrium fluctuations

Statistical mechanics from microscopic counting to material response

How can a fluid in equilibrium reveal the diffusion constant that would govern a spreading drop, or the viscosity that would resist an applied shear? Derive an exact connection between displacements and velocity correlations, generalise it to stress and heat currents and test how a finite record can bias each number. The assumptions behind convergence matter as much as the prefactor.

Integrate the velocity correlation before taking a long-time limit

Observe one Cartesian coordinate of a tagged particle in a homogeneous equilibrium fluid with zero mean drift. The displacement over time t is the integral of its velocity. Square that integral, average over the equilibrium ensemble and use stationarity to replace the two-time expectation by $C_v(t_1 - t_2)$. For a scalar component in a time-reversal-symmetric equilibrium state this correlation is even. Integrate over the two triangles of the time square:

$$\langle [x(t) - x(0)]^2 \rangle = 2 \int_0^t (t - s) C_v(s)\, ds,$$

$$\frac{1}{2} \frac{d}{dt} \langle \Delta x^2(t) \rangle = \int_0^t C_v(s)\, ds.$$

If the correlation is integrable and boundary effects remain absent, the long-time slope yields a diffusion coefficient $D = \int_0^\infty C_v(s) ds$. In d isotropic dimensions one may use $D = (1/d) \int \langle \mathbf{v}(s) \cdot \mathbf{v}(0) \rangle ds$ or the mean-square-displacement slope divided by $2d$. Using the full vector dot product without division by dimension gives an overestimate. Directional diffusion in an anisotropic material requires tensor components rather than a forced scalar average.

The derivation is kinematic and does not assume a Markovian motion. A fluid with velocity memory or an oscillating solid still obeys the finite-time relation. Those systems need not exhibit the same long-time proportionality to time. A bound oscillator can have persistent or alternating velocity correlations and a bounded displacement. An observation window in which the mean square grows over a fraction of an oscillation is insufficient to claim diffusion.

A solvable benchmark from ballistic to diffusive motion

For one coordinate take the underdamped Langevin equation $m\dot{v} = -\zeta v + \xi$, with zero-mean white noise satisfying $\langle \xi(t) \xi(t') \rangle = 2 k_B T \zeta\, \delta(t - t')$. In equilibrium $\langle v^2 \rangle = k_B T / m$. Condition on $v(0)$ and integrate for positive time. The later random force has no correlation with the initial velocity, giving $C_v = (k_B T / m) e^{-t/\tau_v}$, where $\tau_v = m/\zeta$. Its area and the resulting displacement are

$$D = \frac{k_B T}{\zeta},$$

$$\langle \Delta x^2(t) \rangle = 2D[t - \tau_v (1 - e^{-t/\tau_v})].$$

At $t \ll \tau_v$, the constant and linear terms inside the brackets cancel and displacement is ballistic:

$$\langle \Delta x^2 \rangle \simeq (k_B T/m) t^2.$$

At $t \gg \tau_v$ the slope approaches $2D$. Fitting a line through the origin at times only a few relaxation times overestimates the influence of the negative intercept in this long-time expression and underestimates D. Comparisons at multiple lags distinguish crossover curvature from a true anomalous power law.

For an inert sphere of radius 0.50 micrometres and density 2000 kg per cubic metre in water with $\eta = 1.00$ mPa s at 298 K, its bare mass is 1.05×10^{-15} kg and Stokes drag 9.42×10^{-9} kg/s. The estimate $\tau_v = 0.111$ microseconds is far faster than ordinary video tracking. $D = 0.437$ square micrometres per second and the one-coordinate rms displacement over 10 ms is 0.0935 micrometres. Hydrodynamic added mass and memory affect very short-time behaviour; the simple exponential is an analytical reference, not a claim that a water suspension remains Markovian down to arbitrary times.

Relate a weak imposed force to equilibrium motion

In the same soluble model a small constant force F turns the mean velocity into $F(1 - e^{-t/\tau_v})/\zeta$, so the mobility $b = \lim \langle v \rangle / F = 1/\zeta$. Together with equilibrium diffusion this gives the Einstein relation $D = k_B T b$. A broader equilibrium argument considers a dilute suspension in a weak conservative potential $U(x)$. Its current is

$$j = -D\,\partial_x n + bFn,$$

with $F = -\partial_x U$. The Boltzmann distribution $n \propto e^{-U/k_B T}$ must carry zero current; substitution gives the same relation without assuming exponential velocity relaxation.

Linear response connects the full weak-force impulse response to an equilibrium correlation. For a perturbation coupled as $-F(t)x$, classical equilibrium response of velocity has causal kernel $C_v(t)/(k_B T)$ for $t > 0$ under the usual microreversibility assumptions. Integrating the response kernel for a constant weak force gives

$$b = (1/k_B T) \int_0^\infty C_v(t) dt,$$

which recovers the kinematic diffusion relation. Strong forcing, an active bath or a nonequilibrium steady state need not obey this simple equilibrium identity.

A direct check compares independently measured weak sedimentation or calibrated drift with freely diffusing trajectories. For the previous bead a force 10 fN would produce a mean drift about 1.06 micrometres per second in the ideal Stokes regime. During a 10 ms frame its mean drift is about 0.011 micrometres, much smaller than rms Brownian motion, so many frames or many particles are needed. A single short jagged record can appear to move against the force. Use equal positive and negative applied forces to cancel constant stage drift in the estimated mobility.

Work through diffusion and mobility in the random-walk chapter

Stress and heat current carry their own normalisations

In a homogeneous isotropic laboratory fluid, define P_{xy} as a spatially averaged off-diagonal microscopic stress, with units Pa. Its equilibrium mean is zero in an isotropic state without imposed shear. A weak velocity gradient couples to momentum transport. The Green–Kubo expression for dynamic shear viscosity is

$$\eta = \frac{V}{k_B T} \int_0^\infty \langle P_{xy}(t) P_{xy}(0) \rangle \, dt.$$

If a program outputs an extensive stress moment $J_{xy} = V P_{xy}$, the prefactor is $1/(V k_B T)$. Multiplying an already extensive current by volume again is a common orders-of-magnitude error. Check dimensions: pressure squared times time times volume divided by energy gives pascal seconds. Kinetic and interaction contributions to microscopic stress must both be included according to the model; a calculation with ideal kinetic flux alone need not reproduce the viscosity of a dense fluid.

Similarly for a component of total heat current $J_{Q,x}$ with units energy times length per time, the thermal conductivity in an isotropic equilibrium single-component system is

$$\kappa = \frac{1}{V k_B T^2} \int_0^\infty \langle J_{Q,x}(t) J_{Q,x}(0) \rangle \, dt.$$

Here $k_B T^2$ means k_B times T^2. If a heat-flux density is used instead, the volume prefactor moves as for stress. Heat current must exclude energy carried by a net mass flow as appropriate to the chosen thermodynamic constraint. In mixtures several conserved species currents can couple, and the measured conductivity at zero mass flux can differ from one unprojected current coefficient. The meaning of the measured flux and the imposed condition must be specified before comparing coefficients.

For an illustrative volume 10^{-24} cubic metres at 300 K, assume one stress autocorrelation $C_P(t) = (2.0 \times 10^6 \ \mathrm{Pa})^2 e^{-t/(2.0 \ \mathrm{ps})}$. Its area is 8.0 square pascal seconds. Multiplication by $V/(k_B T)$ gives viscosity 1.93×10^{-3} Pa s. Doubling the simulated volume of an otherwise similar homogeneous fluid generally reduces the variance of its volume-averaged stress, keeping the physical transport coefficient intensive. Holding the same numerical stress variance while merely changing the prefactor would describe a different equilibrium fluctuation model.

Find a correlation integral's useful time window

An estimate at lag s averages over pairs of samples separated by that lag. In a record of duration T_{rec}, the number of available pairs decreases with lag. Overlapping time origins strongly correlate the estimates, so plotting one error bar per raw pair exaggerates precision. Integrating an estimated correlation to the maximum possible lag frequently accumulates late random noise rather than useful signal. Conversely stopping too early omits a physical tail.

For a positive single-exponential reference, integrating to three correlation times retains $1 - e^{-3} = 95.0$ percent of the area. Five times retain 99.3 percent. A correlation containing a small,

much slower component can look essentially flat on the main decay scale while contributing appreciably to the area. Divide a long run into blocks, compare running integrals and check several window lengths. Report the variation between independent blocks and any explicit tail model used beyond the measured interval.

At short lags, finite exposure averages correlated positions or velocities. Uncorrelated localisation errors in position increase nonzero-lag displacement variance by a constant $2s_x^2$ per coordinate. Numerically differentiating those same noisy positions creates large artificial velocity variance and negative correlations between adjacent increments. An apparent negative velocity lobe can therefore be physical backscattering or a measurement artefact. Comparing a raw displacement fit with the velocity integral after independently characterising detector noise is more informative than differentiating first and interpreting every feature as fluid memory.

A forced shear wave gives a second viscosity measurement

An equilibrium stress correlation estimates viscosity without imposing any flow. A separate driven measurement can test that coefficient against a directly observed penetration depth. Take a Newtonian fluid filling $y > 0$, driven by a laterally oscillating plate at $y = 0$ with speed $U_0 \cos \omega t$. Neglect boundaries at large y, compressibility and nonlinear convective effects in this spatially uniform tangential shear. The momentum equation reduces to

$$\partial_t u = \nu \partial_y^2 u,$$

with kinematic viscosity $\nu = \eta / \rho$.

Seek $u = \mathrm{Re}[U_0 e^{-\lambda y - i\omega t}]$ and choose the square root with positive real part. Then $\lambda = (1 - i)/\delta$,

$$\delta = \sqrt{2\nu/\omega}.$$

Both amplitude decay and phase lag carry information: at distance one penetration depth the amplitude is e^{-1} of the wall speed and the lag magnitude is one radian. A single complex velocity map can determine viscosity from either slope and use their agreement to test the simple model.

For a fluid of density 1000 kg per cubic metre and measured penetration depth 0.126 mm at 20 Hz, $\nu = \omega \delta^2 / 2 \simeq 1.00 \times 10^{-6}$ square metres per second and $\eta \simeq 1.00$ mPa s. A stress-fluctuation integral yielding a value twice as large would predict a penetration depth larger by $\sqrt{2}$, approximately 0.178 mm. Such a comparison can discriminate a missing component or volume factor in the equilibrium calculation if density, temperature and fluid composition match.

At high enough frequency the fluid may have a complex frequency-dependent viscosity rather than the zero-frequency coefficient from a Green–Kubo area. In linear response that generalisation is the one-sided Fourier transform of the equilibrium stress correlation with appropriate prefactors. A single-exponential stress memory gives a frequency-dependent response with a characteristic roll-off and phase lag. Comparing a high-frequency penetration measurement directly to a zero-frequency area without checking the timescale can therefore produce a genuine difference, not a statistical failure.

Remove a centre-of-mass mode only when the observable requires it

In a finite periodic particle simulation total momentum may be exactly conserved. If the initial centre of mass has a random velocity and no thermostat removes it, every particle's measured laboratory coordinate includes the common drift $V_{\mathrm{CM}}t$. Its mean-squared displacement gains a ballistic term $\langle V_{\mathrm{CM}}^2\rangle t^2$. In a canonical ensemble the centre-of-mass velocity variance in one direction is k_BT/M_{tot}. Though it falls with system size, at sufficiently long times its quadratic contribution exceeds the linear diffusive part.

For 1000 equal particles of mass m, an equilibrium centre-of-mass mode contributes one thousandth of the single-particle instantaneous velocity variance. In a long-time autocorrelation this small constant never integrates to a finite diffusion coefficient. Subtract centre-of-mass displacement or work in a zero-total-momentum ensemble before taking the long-time integral of an internal tracer's motion. In a real sample with externally imposed drift, decide whether the physical quantity sought is diffusion about that drift or absolute laboratory displacement; the subtraction has a physical meaning rather than merely a cosmetic improvement.

Conversely, indiscriminately subtracting the mean velocity of a tiny observation subregion separately at every time can remove genuine hydrodynamic long-wavelength motion. A strongly applied single-particle thermostat can alter current correlations and suppress transport even while maintaining the correct static temperature histogram. Check the thermostat strength and system size; compare an equilibrium-prepared weakly thermostatted trajectory with an energy-conserving production segment where appropriate. Maintaining the correct one-time distribution is not sufficient to preserve all dynamical correlation functions.

The distinction also matters when removing an estimated mean stress before integrating its fluctuations. The true shear mean may vanish by symmetry, but a short noisy record has a nonzero sample mean. Subtracting a slowly varying trend on the same timescale as a genuine stress tail can erase physical viscosity. Use independent samples, separately measured drifts and explicit sensitivity to the subtraction window. A persistent nonzero off-diagonal stress under nominal isotropic conditions may indicate incomplete equilibration, a deformed box or an unrelaxed solid rather than a fluid transport coefficient.

Heat and particle currents can share fluctuations

Consider two thermodynamic fluxes J_1, J_2 driven weakly by conjugate forces X_1, X_2, chosen so entropy production density is $J_1X_1 + J_2X_2$. Linear response writes $J_i = \sum_j L_{ij}X_j$. Near time-reversal-symmetric equilibrium, with flux parities chosen appropriately and no magnetic-field reversal complications, reciprocal relations give

$$L_{12} = L_{21}.$$

The symmetric part must be positive semidefinite so that entropy production is nonnegative for every small force. Thus $L_{11}, L_{22} \geq 0$ and $L_{12}^2 \leq L_{11}L_{22}$.

Suppose a thermal measurement enforces $J_1 = 0$ by allowing a compensating force X_1 to develop, as when a concentration or electrochemical gradient cancels a coupled particle flux. Solving the first row gives

$$X_1 = -L_{12}X_2/L_{11}.$$

Substitution into the second yields response coefficient $L_{22} - L_{21}L_{12}/L_{11}$. The measured constrained heat response can therefore be smaller than the coefficient at fixed $X_1 = 0$. The difference has a simple operational origin in the boundary condition enforced during the experiment.

Take consistent normalised coefficients $L_{11} = 4, L_{22} = 3, L_{12} = L_{21} = 2$. At zero first force the heat flux slope is three. At zero first flux it is $3 - 4/4 = 2$. Both are positive and consistent with the same equilibrium correlation matrix. Comparing a simulated heat-current autocorrelation corresponding to the unconstrained coefficient against a zero-particle-flux experiment requires the mixed-current correlation too. Apparent disagreement by a factor of one and a half can arise entirely from this distinction.

Green–Kubo cross-correlations must use properly defined conjugate fluxes and units. Redefining heat current by adding a constant multiple of particle current changes the matrix entries but must leave a physically defined zero-particle-flux heat measurement invariant when the forces are transformed consistently. This reference invariance is analogous to subtracting a per-particle energy reference in grand-canonical covariance. It provides a powerful test of a multicomponent transport implementation without relying on any empirical transport value.

A lattice trajectory provides an exactly known correlation integral

For a one-dimensional nearest-neighbour walk with steps of length a every interval Δt, suppose successive step directions have correlation $\langle s_{j+k}s_j\rangle = r^k$ with $|r| < 1$, and zero mean. Squaring the sum of n steps gives

$$a^2[n + 2\sum_{k=1}^{n-1}(n-k)r^k].$$

In the large-n limit divide by $2n\Delta t$ to obtain

$$D = a^2(1+r)/[2\Delta t(1-r)].$$

The uncorrelated random-walk value is recovered at $r = 0$.

With $a = 1$ micrometre and $\Delta t = 0.10$ s, uncorrelated diffusion is five square micrometres per second. Persistence $r = 0.5$ raises it to fifteen, while anticorrelation $r = -0.5$ lowers it to $5/3$. All three walks have the same step-size variance, so an estimate based only on one-frame displacement misses the entire correlation effect. The finite-lag autocorrelation sum and a long-time displacement fit should recover the same value on a long simulated record.

The discrete formula also clarifies the special zero-lag weight in numerical integration. An integral approximated from sampled symmetric correlations gives half weight to the zero lag on a one-sided trapezoidal rule. A careless double count of the instantaneous self term predicts twice the uncorrelated diffusion. Benchmark a numerical estimator with this exact walk before applying it to molecular data with unknown late tails and noisy velocity derivatives.

Measure an off-diagonal diffusion coefficient after rotating the sample

In an anisotropic sheet a tracer's two-dimensional displacement covariance need not align with the camera axes. Let its long-time principal diffusion coefficients be D_1, D_2 and the first principal axis make angle θ with laboratory x. In matrix form the laboratory tensor is $R\,\mathrm{diag}(D_1, D_2)R^{\mathsf{T}}$, with R the planar rotation. At sufficiently large lag in an effectively unbounded drift-free region, the displacement covariance matrix tends to $2Dt$. An off-diagonal covariance is a measurable correlation between displacement components even if the noise is independent in the material principal directions.

Take $D_1 = 2.0$ and $D_2 = 0.50$ square micrometres per second with $\theta = 30^\circ$. Laboratory components are $D_{xx} = 1.625$, $D_{yy} = 0.875$ and $D_{xy} = 0.650$, in the same units. At 0.10 s the cross covariance is 0.130 square micrometres. Rotating the physical sheet by ninety degrees interchanges the two diagonal values and reverses the sign of the off-diagonal one. An instrument-fixed camera readout artefact would not generally follow this material rotation.

To infer the principal axes from data, fit each covariance component versus lag using a common drift subtraction and independent position calibration. Diagonalise the symmetric fitted diffusion matrix. Positive semidefiniteness requires

$$D_{xy}^2 \le D_{xx} D_{yy}.$$

A strong violation indicates incompatible estimates, unit conversion errors or statistical uncertainty, not a physically permissible strongly anisotropic diffusion tensor. Near-degenerate principal values make the orientation poorly determined even when the mean diffusion coefficient is precise. Report uncertainties on eigenvalues and orientation separately.

A finite field of view introduces a further bias. Fast particles along the high-diffusion direction leave the image sooner, and a tracking pipeline retaining only long trajectories can favour slow or fortuitously confined paths. Compare different fields of view and minimum track durations, or fit a measurement model incorporating censoring. A physical material anisotropy should persist after correcting these effects; apparent anisotropy aligned with the short direction of a rectangular observation window demands particular scrutiny.

A weak calibrated force along different axes supplies an independent check through the tensor Einstein relation under equilibrium conditions. If the diffusion tensor predicts off-axis mean drift, the mobility experiment should observe its sign and magnitude after reversing the force. A discrepancy could reveal active forcing, nonstationarity, an incorrect temperature or a mismatch between the mobility and displacement observables. It cannot be diagnosed from a scalar mean square alone.

Another rotational invariant is the trace of the diffusion tensor. In the example its value stays 2.50 square micrometres per second for every sample orientation, while the determinant stays one square micrometre to the fourth per second squared. A comparison of two camera orientations therefore provides numerical checks independent of the fitted angle. A drift mistakenly left in the displacement covariance can violate the expected linear dependence on lag, making the inferred effective eigenvalues grow with the selected fitting window. Subtract the ensemble mean displacement at each lag and record that mean separately as a flow measurement. Conversely do not subtract an individual best-fit line from every short trajectory without modelling the resulting loss of long-time fluctuations; that procedure can bias both diffusion coefficients downward.

Choose an integration window for a transport coefficient

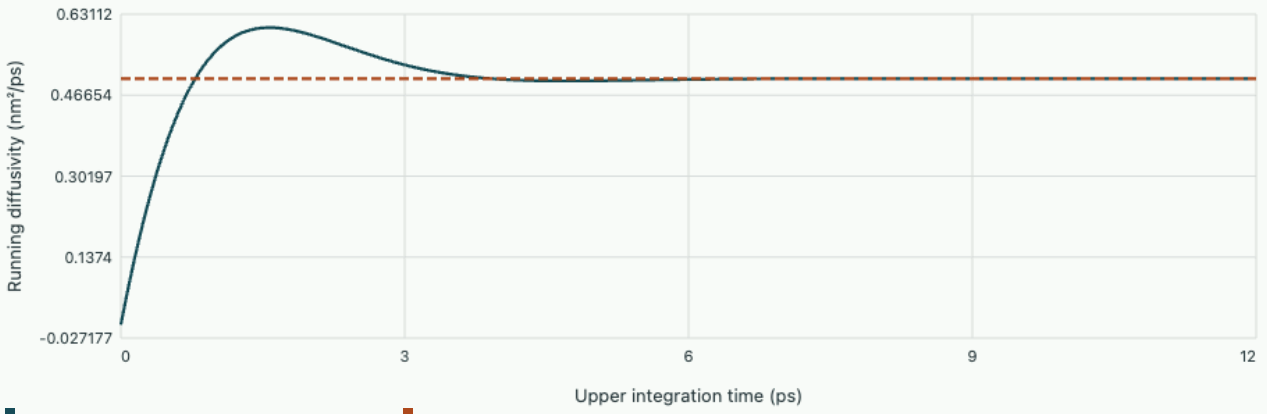
Integrate an oscillatory velocity correlation over a finite record. Compare a short apparent plateau with the independently computed infinite-time result.

Open this experiment online

Try the calculation

Correlation decay time τ (ps) 1

Oscillation frequency Ω (rad/ps) 1



With C(0) = 1 nm²/ps² the exact 1D diffusion coefficient is 0.5 nm²/ps. At 2τ the integral gives 0.58969 nm²/ps.

Tutorial · Invert diffusion measurements at two lags

In a one-dimensional diffusive window a measured displacement variance is 0.024 square micrometres at 10 ms and 0.064 at 50 ms. Assume long-time Brownian displacements plus independent frame-localisation variance per position measurement. Estimate D and localisation precision.

▼ **Fit slope and intercept separately**

The variance slope is $(0.064 - 0.024)/(0.050 - 0.010) = 1.00$ square micrometre per second, so $D = 0.500$ square micrometres per second. The intercept is 0.014 square micrometres, equal to twice the single-frame localisation variance. The implied rms localisation error is $\sqrt{0.007} = 0.0837$ micrometres. Forcing the fit through the origin at the shortest lag would give $D = 1.20$, more than twice the inferred value.

Acquire more lags to test the linear model. A static drift contributes a term growing as t^2 if its mean displacement is not removed, while finite exposure can lower the physically measured short-lag variance. Two points always define a line and cannot independently validate either noise model.

Tutorial · Estimate a hidden slow tail in viscosity

A stress correlation is modelled as $A[e^{-t/\tau} + 0.02e^{-t/(40\tau)}]$. What fraction of its true area comes from the low-amplitude tail? How much does a cutoff at 5τ capture?

▼ **Weight a correlation by its duration**

The fast and slow areas are $A\tau$ and $0.80A\tau$. The slow contribution is 44.4 percent of total. Integration to five fast times captures $A\tau[(1 - e^{-5}) + 0.80(1 - e^{-5/40})] \simeq 1.087A\tau$, or about 60.4 percent. Visual disappearance of the dominant peak therefore does not establish convergence of the transport coefficient.

The slow contribution must be supported by independent long blocks or a physically motivated model. A fitted tiny constant baseline can generate arbitrarily large apparent area if extrapolated indefinitely. Subtracting a sample mean estimated from the same short record can conversely suppress a genuine slow mode. Compare records much longer than the proposed tail time before trusting the extrapolation.

Tutorial · Is an algebraic tail compatible with ordinary diffusion?

Suppose a measured long-time velocity correlation remains positive and scales as $t^{-\alpha}$. Determine when its Green–Kubo area exists and infer the displacement growth for $0 < \alpha < 1$ and for $\alpha = 1$. Why would a finite periodic simulation require care?

▼ **Integrate the time-dependent slope before naming a diffusion coefficient**

For $\alpha > 1$ the tail integral converges. For $0 < \alpha < 1$, the running area grows as $t^{1-\alpha}$, so displacement variance grows as $t^{2-\alpha}$ under the stated persistent positive tail. At $\alpha = 1$, the running area grows logarithmically and the displacement variance is of order $t \log t$. Sign oscillations, finite-size termination or another late-time crossover can alter these conclusions.

A periodic box admits only discrete long-wavelength modes and may cut off a hydrodynamic tail. Repeating at larger sizes and comparing the range before signals traverse the box helps identify a bulk regime. Unwrap particle coordinates crossing periodic boundaries before calculating a displacement. A wrapped coordinate can falsely saturate and imitate a confined particle even in a freely diffusing periodic fluid.

Cambridge Kinetic Theory · correlations, response and transport

Applied study

Discretise a boundary-value problem and separate mesh error from solver and sampling errors.

Replace a continuum balance by nodal equations

Suppose a scalar diffuses along a unit interval, with a fixed value at each end. The steady balance is $-u'' = f$ for constant unit diffusivity. At equally spaced interior nodes $x_i = ih$, approximate the derivative at each face by the difference between its neighbouring values divided by h. Subtracting the outgoing and incoming diffusive fluxes and dividing by cell width gives

$$\frac{2u_i - u_{i-1} - u_{i+1}}{h^2} = f(x_i).$$

The endpoints are known, leaving $N - 1$ interior unknowns for N intervals. Each interior flux appears once with each sign in neighbouring balances. Summing balances therefore leaves the net boundary flux and the total source. With variable diffusivity D, multiply each face gradient by its *face* value of D before taking the difference.

Find the error made by a single stencil

Insert the exact values $u(x_i - h)$ and $u(x_i + h)$, expand both in Taylor series and add them. Odd terms cancel. The central second difference satisfies $\delta_{xx} u = u'' + h^2 u^{(4)}/12 + O(h^4)$ for sufficiently smooth u. This local consistency calculation predicts the scale of the defect when the exact solution is inserted in the discrete equations.

Write the discrete problem $A_h u_h = f_h$ and define $\tau_h = A_h u_\star - f_h$ using exact nodal values $u_\star$. Then $A_h(u_\star - u_h) = \tau_h$. A stability bound on A_h^{-1}, uniform in mesh size in the chosen norm, turns second-order consistency into second-order solution error. A tight iterative solve reduces the residual of $A_h u_h = f_h$, but leaves τ_h in place.

Use an exact sine mode to predict the discrete amplitude

Prescribe $u(0) = u(1) = 0$ and $f(x) = \pi^2 \sin(\pi x)$. The exact solution is

$$u_\star = \sin(\pi x).$$

For its grid samples $s_i = \sin(\pi i h)$, the sine addition formula yields

$$A_h s = \lambda_h s,$$

with $\lambda_h = 4 \sin^2(\pi h/2)/h^2$. The right-hand side is $\pi^2 s$, so the discrete solution has the same nodal shape and amplitude π^2/λ_h.

With four intervals, $\lambda_h = 64 \sin^2(\pi/8) = 9.3726$ and the numerical midspan value is 1.0530. The exact value is 1. The sign makes sense: the discrete stiffness of this mode is below π^2, so the same load

produces a larger amplitude. Expansion gives $\pi^2/\lambda_h = 1 + \pi^2 h^2/12 + O(h^4)$, predicting a fourfold error reduction on refinement.

Problem class

Test a polynomial before trusting the mesh

Keep the homogeneous end values and choose $u_\star = x(1-x)$. Derive f, then predict the nodal error for any uniform mesh in exact arithmetic. Would this make a good test for measuring convergence order?

▼ **Show the calculation**

Since $u_\star'' = -2$, use $f = 2$. The centred second difference of any quadratic equals its exact second derivative. The nodal vector therefore solves the discrete equations exactly. It is a useful stencil consistency check, but its zero leading error prevents a meaningful order measurement from a ratio of errors.

Manufacture a source for variable diffusion

On the unit square let $D(x) = 1 + x$, $u_\star = \sin(ax)\sin(ay)$, and $a = m\pi$ for positive integer m, using metre coordinates. What forcing makes $-\nabla \cdot (D\nabla u) + 2u = f$ hold with zero boundary values?

▼ **Show the calculation**

Apply the product rule,

$$-\nabla \cdot (D\nabla u) = -D\Delta u - D'(x)u_x.$$

Here $\Delta u_\star = -2a^2 u_\star$ and $D' = 1$. Thus $f = [2(1+x)a^2 + 2]u_\star - a\cos(ax)\sin(ay)$. Each boundary has a zero sine factor. Omitting $D'u_x$ would produce a continuous modelling error that persists even when the mesh is refined.

ONESIMULATION LABORATORY

Measure spatial convergence on a known field

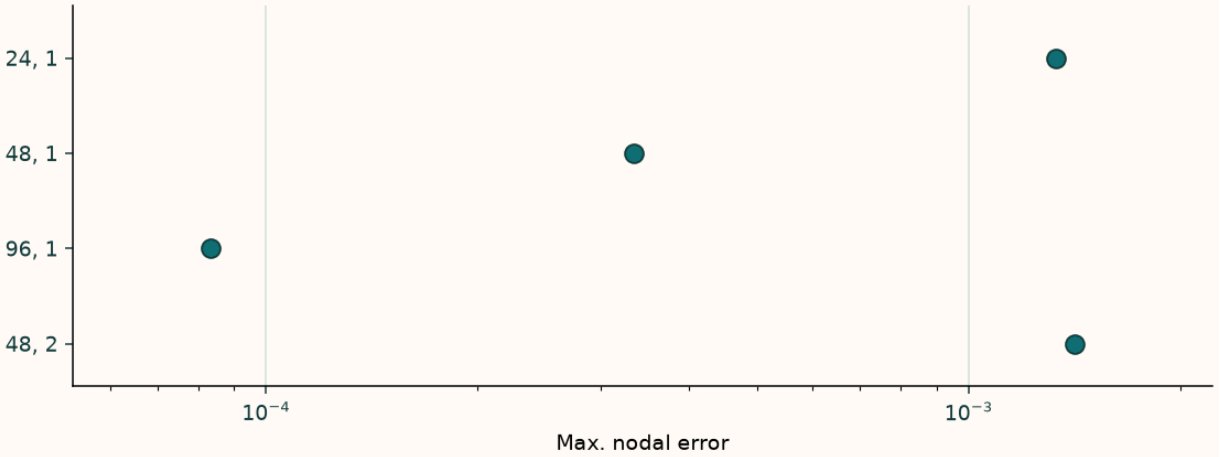
Compare recorded OneSimulation runs on this page. To test your own inputs, download the Python model and run the model in an activated OneSimulation environment.

The script solves the steady equation $-\nabla\cdot\left[(1+x)\nabla u\right]+2u=f$ on the unit square by conservative flux differences and Jacobi-preconditioned CG. Set $u=\sin(m\pi x)\sin(m\pi y)$ and obtain the source by substitution. The boundary values are zero on all four sides. Compare the maximum error at grid nodes with the solver's relative residual. Refine both axes together, then raise the sine mode on the same 48-cell mesh to see how wavelength affects the error. Download the Python experiment · OneSimulation reaction–diffusion refinement study.

```
python partial-differential-equations.py
python partial-differential-equations.py --cells 48
python partial-differential-equations.py --cells 96
python partial-differential-equations.py --cells 48 --mode 2
```

Independent solves for the dimensionless scalar field on a unit square.

Cells per axis, mode	Interior unknowns	Max. nodal error	Relative residual
24, 1	529	1.33345×10^{-3}	9.79×10^{-11}
48, 1	2,209	3.33763×10^{-4}	7.14×10^{-11}
96, 1	9,025	8.34283×10^{-5}	4.21×10^{-11}
48, 2	2,209	1.41546×10^{-3}	8.89×10^{-11}



The first three runs approximately quarter the nodal error with each grid refinement. Changing the wavelength on the same 48-cell grid increases the error even at a small algebraic residual.

At mode one the error falls by factors close to four while the number of interior unknowns roughly quadruples. On the same 48-cell mesh, mode two has half as many intervals per wavelength and a

larger error. Both modes have small algebraic residuals. Mesh convergence is a property of the resolved field and its discretisation.

Change parameters · Change the laboratory inputs

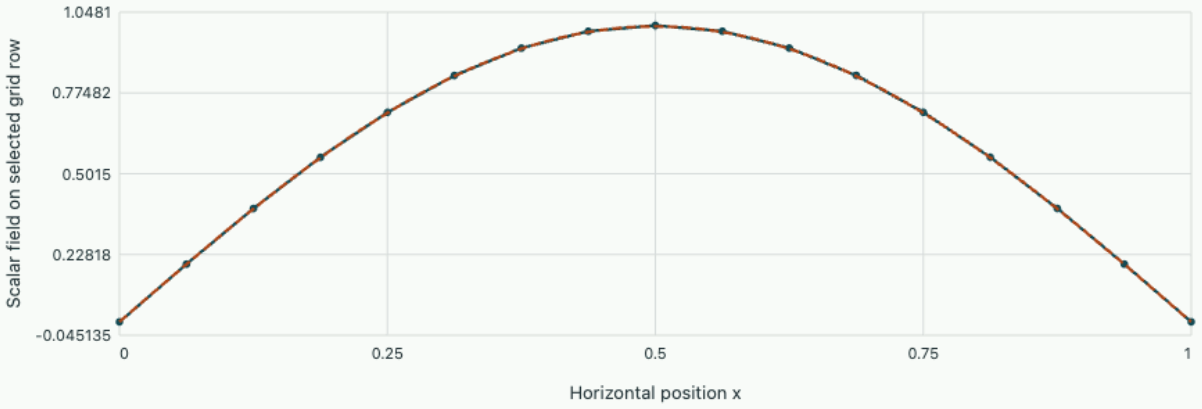
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Mesh intervals per axis **16**

Sine half-waves per direction **1**



Variable-coefficient five-point solve | Manufactured continuum field

For −∇·[(1+x)∇u] + 2u = f, maximum nodal error 0.0030008, relative algebraic residual 4.2371e-11. Slice y = 0.5; 17 Jacobi-preconditioned CG iterations. Refinement measures discretisation after the algebraic solve has converged.

Problem sheet 3 · Predict the error before making a finer grid

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.

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.

Derivation · Pure flux boundaries, compatibility and the missing constant

Conservative face flux

$$|C_i| \dot{u}_i + \sum_{f \subset \partial C_i} F_{if} = |C_i| s_i$$

Use the same face flux with opposite orientations in neighbouring cells. Interior contributions then cancel in the global balance.

Derivation · Flux differences for an elliptic balance

Courant number

$$C = |a| \Delta t / \Delta x$$

For scalar advection this compares a transported distance with a cell width. An admissible range depends on the spatial method and time integrator jointly.

Derivation · Finite volumes and moving fronts

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.

Derivation · Why two finite-element wave solvers propagate the same mode at different speeds

Observed convergence order

$$p_{\mathrm{obs}} = \log(e_h / e_{h/2}) / \log 2$$

Errors must refer to the same observable and norm. Cancellation, singular boundaries and a reference solution of insufficient accuracy can distort the result.

Derivation · Verification, refinement and local error

Weak form

$$\int k \nabla u \cdot \nabla v \, dV = \int f v \, dV + \text{boundary terms}$$

Integration by parts lowers the derivative requirement. Essential and natural boundary data enter the discrete system differently.

Derivation · Weak forms and finite elements

Further reading

- Compare stability and consistency for diffusion and advection using the same mesh. Use the accompanying examples to identify which discretisations preserve qualitative features of a solution. Randall J. LeVeque · Finite Difference Methods for Ordinary and Partial Differential Equations
- Chapters 11–13 develop diffusion, advection and two-dimensional problems. Compare the largest operator eigenvalues before choosing an explicit time step. Driscoll and Braun · Fundamentals of Numerical Computation

Compare specific topics with the university sources