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.

Contents

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.

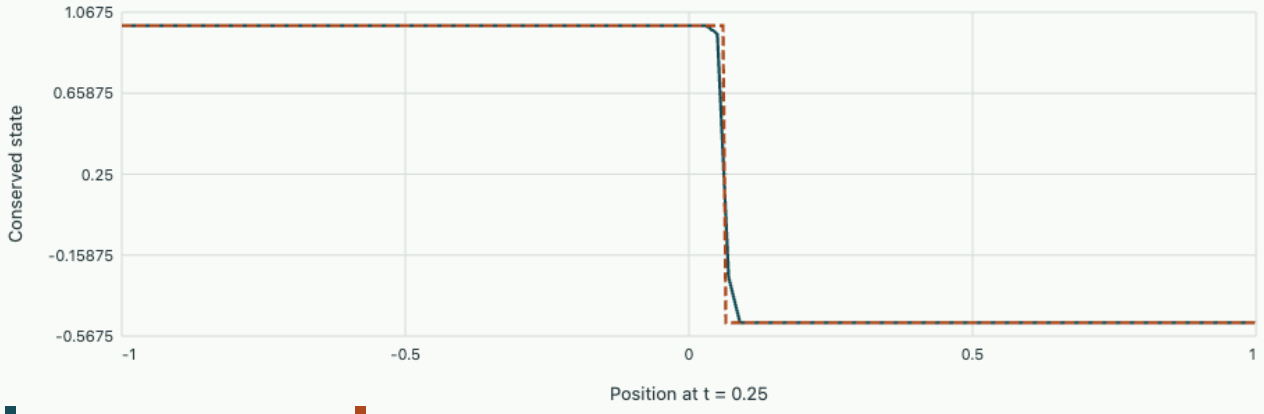
Open this experiment online

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.

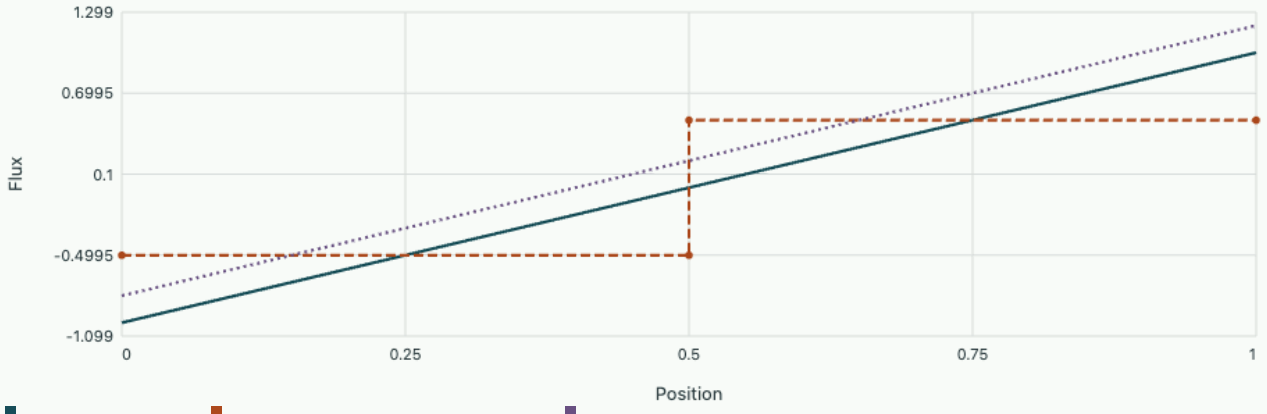
Increase the number of cells with the offset held fixed. The raw staircase becomes more accurate, but the reconstruction bound eventually plateaus near the absolute offset because the reconstructed flux still contains its deliberate uniform error. Set the offset to zero to recover the sharp certificate 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 $|k| \leq K$, 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.

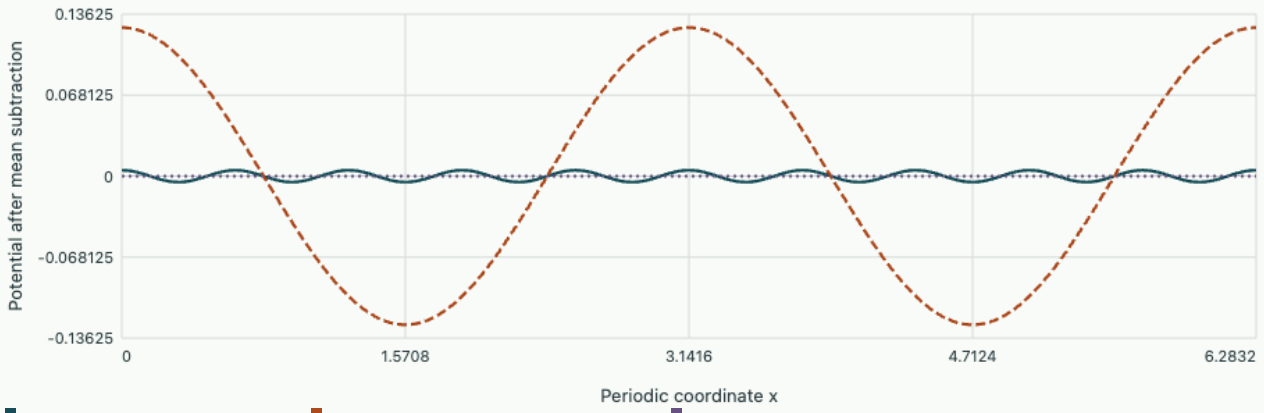
Increase N until the true product is representable. The three answers then agree, including the phase of the spatial variation. The control limits the actual input frequency to a representable non-Nyquist input and reports the mode used, so changing the grid cannot silently inject an impossible input. 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.

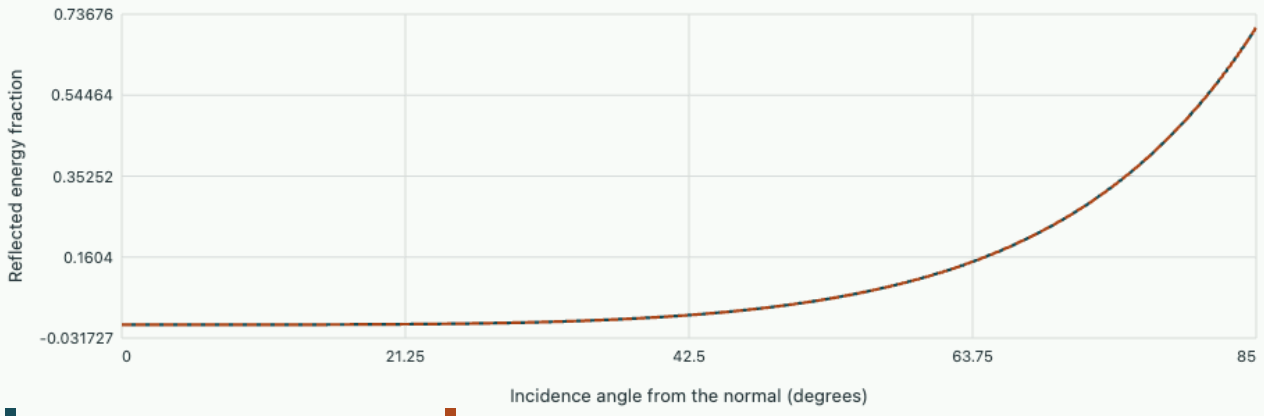
Choose a boundary speed larger than the interior speed. There is then an oblique angle at which the mismatch exactly compensates the projected normal phase in this local equation. It obeys cosine of the angle equal to the reciprocal speed ratio. For a ratio 1.25 this is about 36.87 degrees. Matching 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.

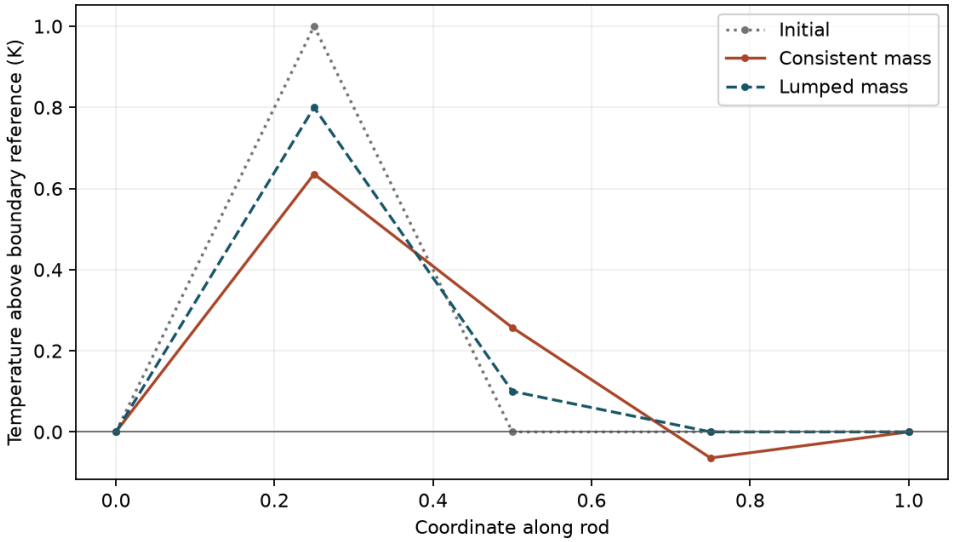
One alternative assigns half of each element's total heat capacity to each endpoint, before removing boundary degrees of freedom. Each interior mass then equals h and the explicit stencil has weights $r, 1-2r, r$, with $r = \kappa\Delta t/h^2$. These coefficients are all nonnegative for $0 \le r \le 1/2$. The first and last interior equations also couple to the fixed zero boundary temperatures.

Compute the remote temperature after one update

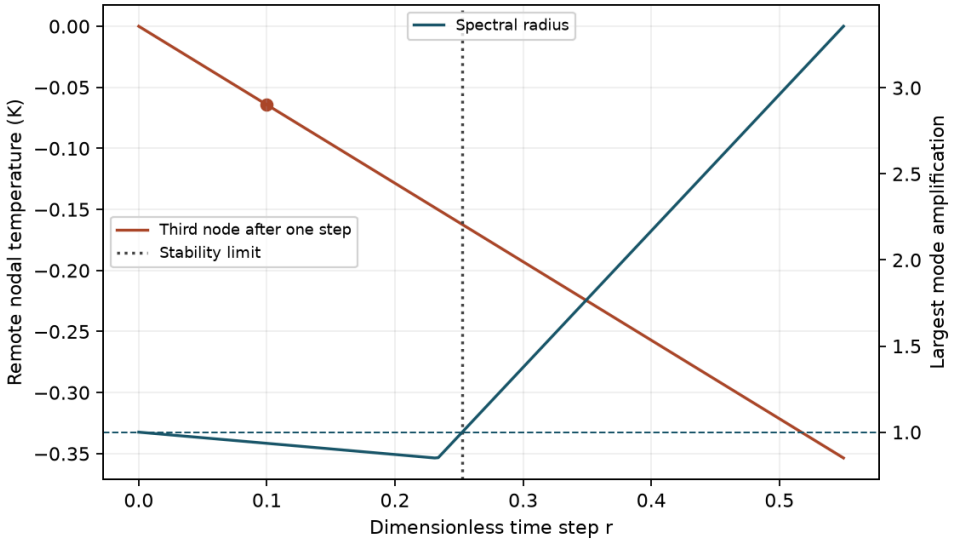
With three interior nodes, $h = 1/4$, initially set the left interior node to 1 K above the boundary reference and both others to zero. Inverting the consistent mass in this three-by-three system gives the first column of $(h^2/\kappa)M^{-1}K$ as $(51, -36, 9)/14$. At $r = 0.1$ the interior temperatures 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

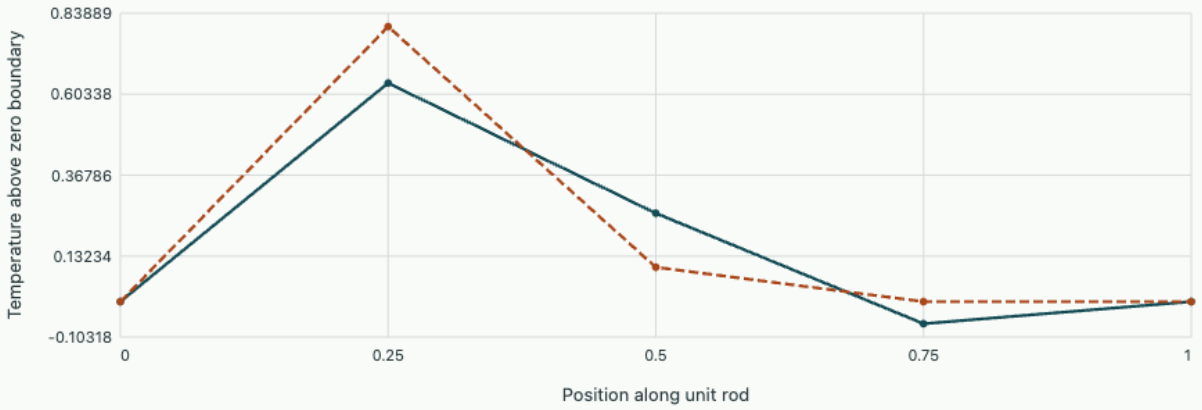
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

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^2XY 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 $s^2XY(U^{n+1} - U^n) = O(h^3)$ 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

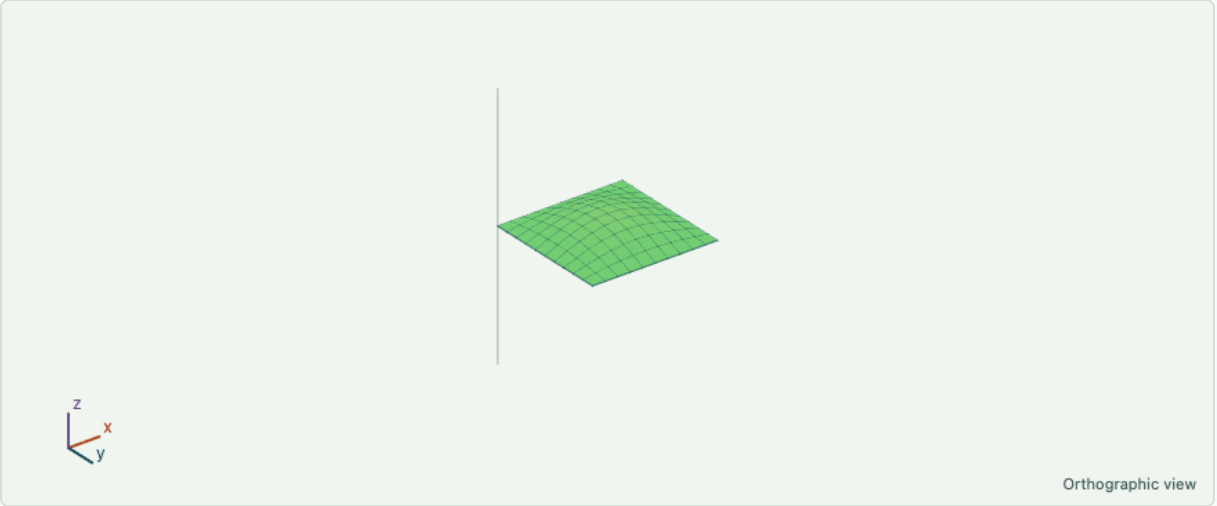
$$G = \frac{1 + h\lambda_x/2}{1 - h\lambda_x/2}\,\frac{1 + h\lambda_y/2}{1 - h\lambda_y/2}.$$

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.

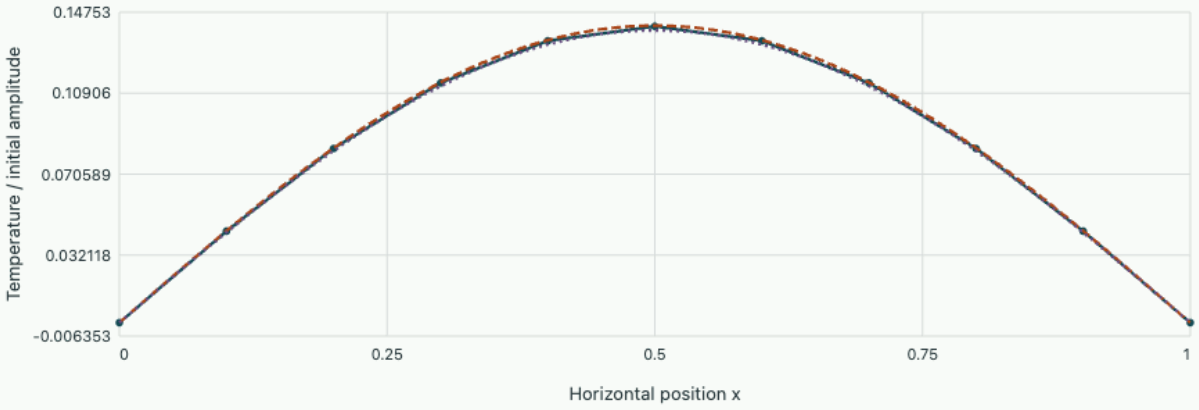
Open the live experiment

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

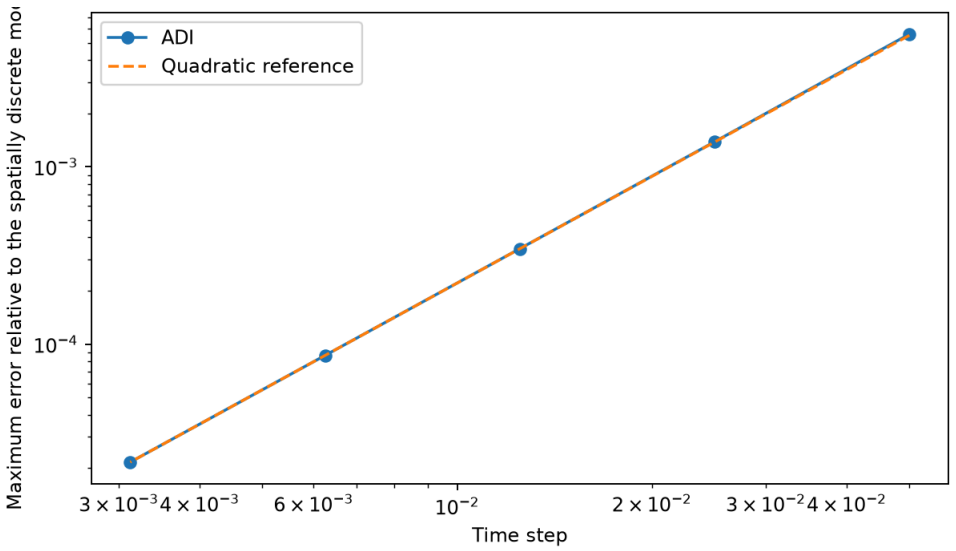
ADI amplitude = 0.14083. Temporal amplitude error = 0.00034519; spatial amplitude error = 0.0022661. The slice is at y = 1/(2q) = 0.5. Diffusivity and side length are one.

```
"""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.

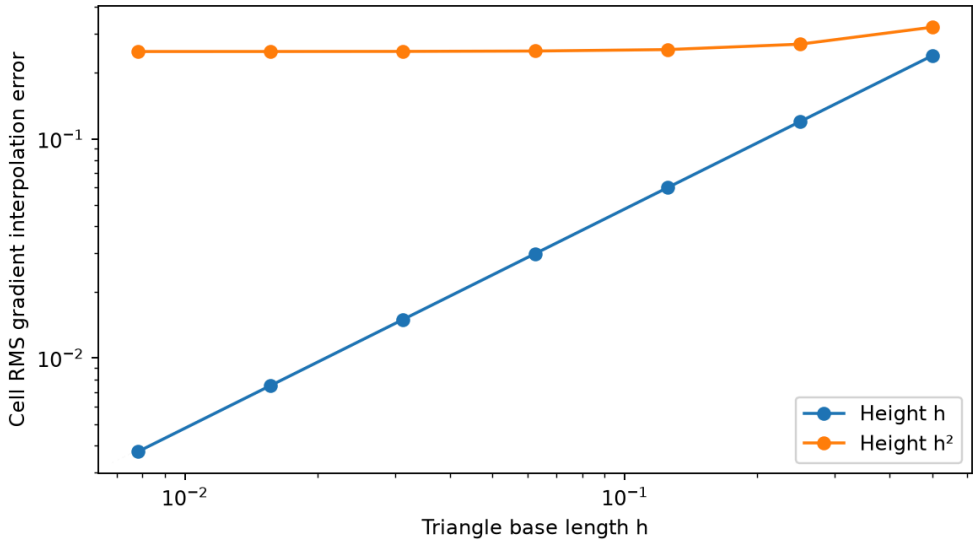
For the mean-square error solve the dual problem $a(v, z) = (v, e)$ with the same boundary conditions. If it satisfies the regularity estimate $\|z\|_{H^2} \le C_{\mathrm{reg}} \|e\|_{L^2}$, then

$$\|e\|_{L^2}^2 = a(e, z - I_h z) \le Ch \|e\|_a \|z\|_{H^2}.$$

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

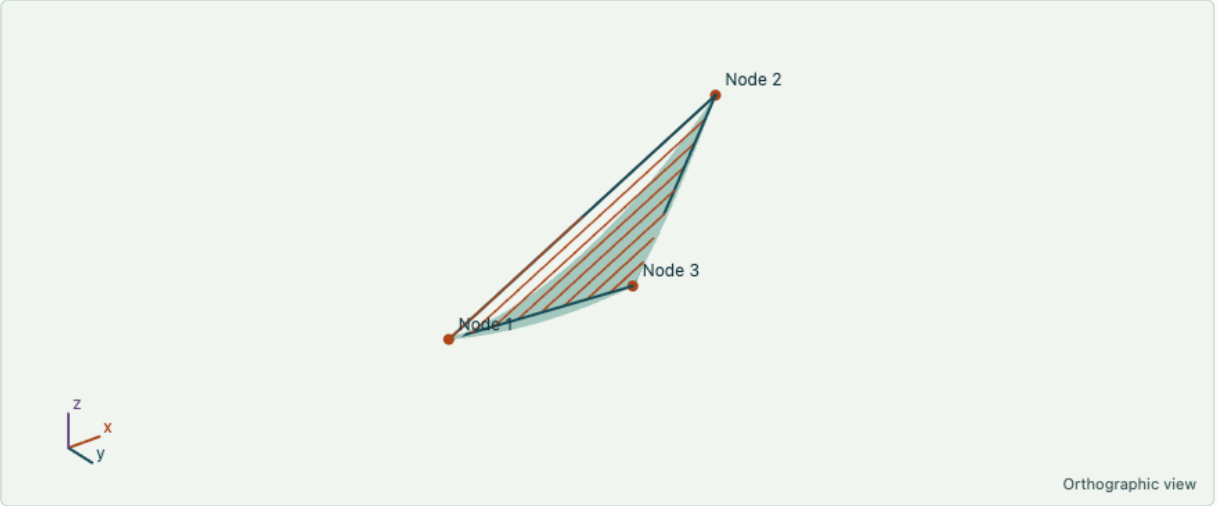
Change parameters · Following the error from one reference triangle to an entire mesh

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.

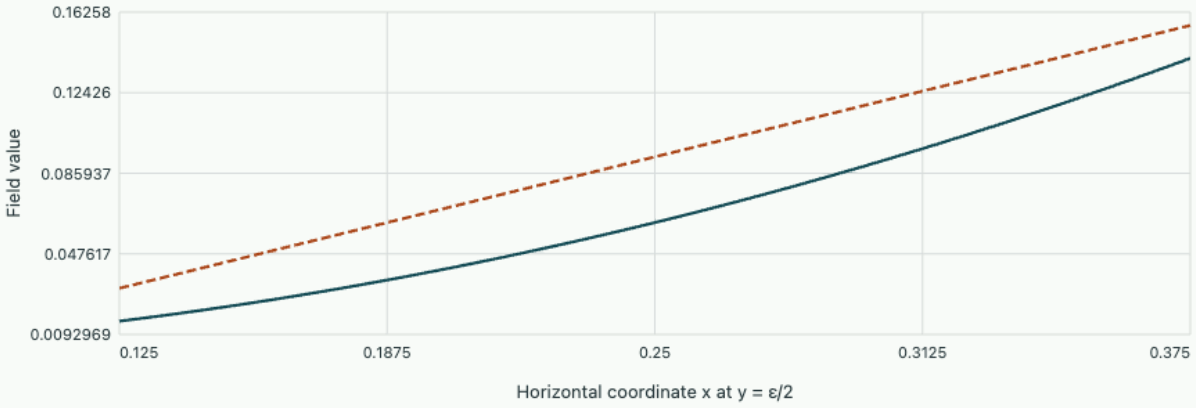
Open this calculation with adjustable inputs

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
```

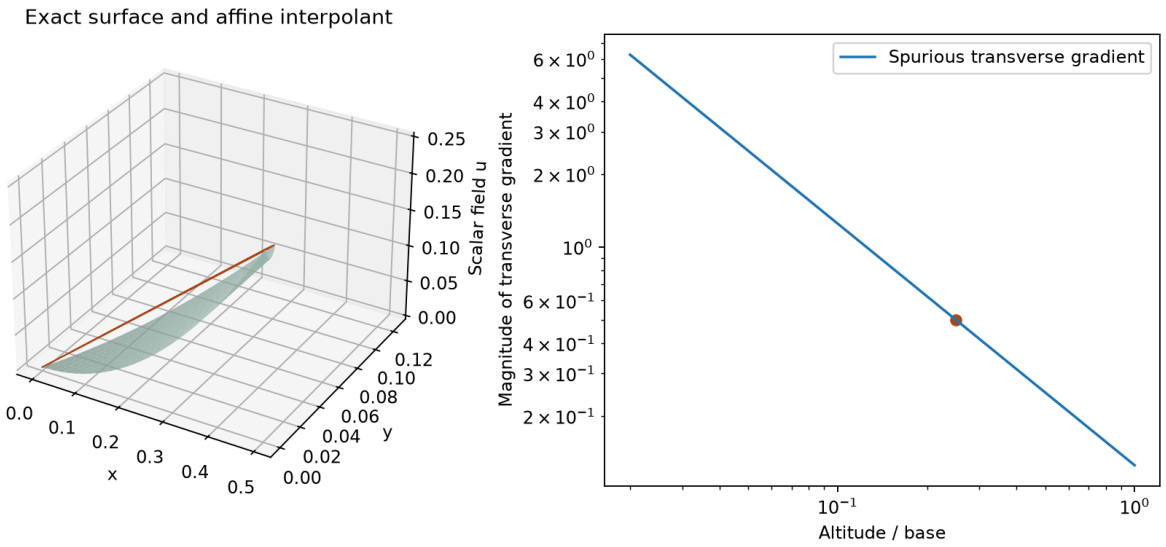
```
import numpy as np
from onesim.numerics.linear.matrices.dense_linear_algebra import solve_dense


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).

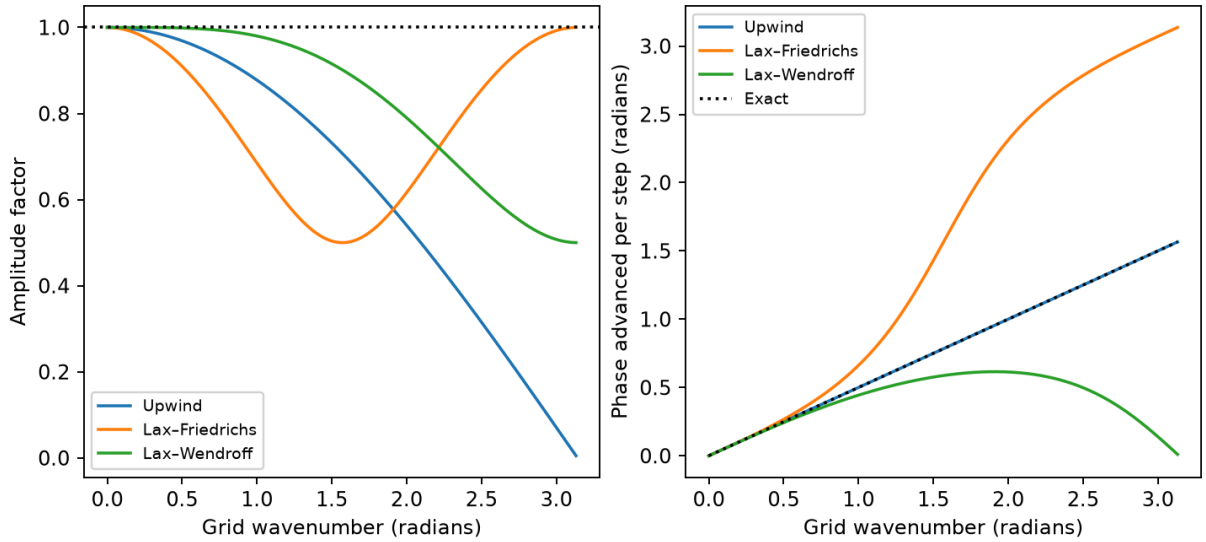
For $0 \le \nu \le 1$ these are stable in the periodic discrete L^2 norm. For the last factor direct expansion gives $|G_W|^2 = 1 - 4\nu^2(1-\nu^2)\sin^4(\theta/2)$. The exact factor $e^{-i\nu\theta}$ has modulus one. At $\nu = 1/2, \theta = \pi/2$, upwind gives $0.5 - 0.5i$, Lax–Friedrichs gives $-0.5i$ 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?

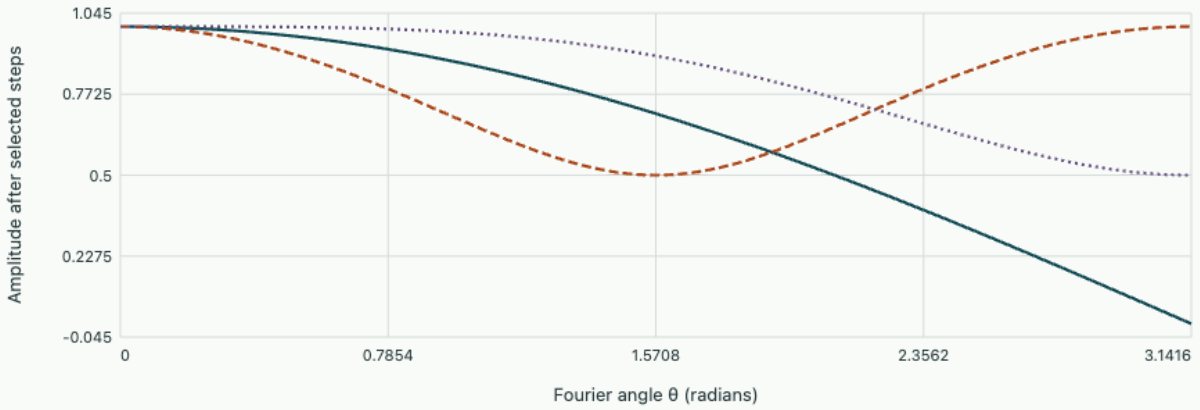
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Courant number ν **0.5**

Propagation steps **1**



Upwind Lax–Friedrichs Lax–Wendroff

Exact transport retains unit amplitude at every θ. All chosen Courant numbers lie in the stability range; grid-scale damping differs substantially between methods.