Numerical ordinary differential equations

Derive and compare time integrators, preserve physical invariants, and use sensitivities and shooting to solve for trajectories and boundary data.

Read the subject in depth

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

- **Dynamics and numerical time integration**
  Turn a differential model into a time step. Track accuracy, stability and structural properties separately in a computable physical or mathematical example.

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. **Conserving motion in a physical system**
   Track an invariant in phase space. Compare a rotating charge, parametrically driven oscillator and magnetic orbit. Use phase and energy as independent error measurements.
2. **Multiple scales and identifiable dynamics**
   Derive a system with competing rates and ask which measurements can separate them. Resolve a decaying field and a coupled mechanical instrument before interpreting the numerical trajectory.

How the course fits together

1. **Chapters 1–4**
   Integrate a rate over one step. Local defects accumulate into global error, additional Runge–Kutta stages cancel Taylor terms, and stiffness can impose a stability limit even on a slowly varying solution. Multistep formulas reuse earlier information.
2. **Chapters 5–8**
   Estimate error to select the step size, then ask what structure a time update preserves. Exponential integration evolves linear modes exactly and splitting separates processes that can each be evolved simply.
3. **Chapters 9–12**
   Construct a reversible Verlet step, differentiate a trajectory with respect to a parameter, solve a boundary condition by shooting, and enforce algebraic constraints that restrict permissible states.

Contents

1. Chapter 1 · Initial values and propagated error
2. Chapter 2 · Runge–Kutta stages and order
3. Problem sheet 1 · From one update to a trajectory
4. Chapter 3 · Absolute stability and stiffness
5. Chapter 4 · Multistep formulas and parasitic modes
6. Problem sheet 2 · Damping and multistep memory
7. Chapter 5 · Adaptive steps and event locations
8. Chapter 6 · Energy and geometric integration
9. Problem sheet 3 · Events and long trajectories
10. Chapter 7 · Variation of constants and exponential steps
11. Chapter 8 · Operator splitting and noncommuting flows
12. Problem sheet 4 · Linear response and split flows
13. Chapter 9 · Störmer–Verlet and a discrete invariant
14. Chapter 10 · Tangent equations and adjoint outputs
15. Problem sheet 5 · Long trajectories and sensitivities
16. Chapter 11 · Shooting for two-point boundary values
17. Chapter 12 · Differential–algebraic equations and consistency
18. Problem sheet 6 · Boundary and algebraic conditions
19. Chapter 13 · Implicit midpoint, reversibility and the phase of an oscillator
20. Chapter 14 · Splitting a stiff relaxation from a slowly varying drive
21. Chapter 15 · Why a boundary value problem can be easier than its shooting equation
22. Chapter 16 · How two decaying modes produce a growing signal
23. Chapter 17 · Locating a grazing event between time steps
24. Chapter 18 · A constrained integrator can get the circle right and the clock wrong
25. Problem sheet 7 · Project · Choose a time integrator from measurements
26. Problem sheet 8 · Differentiate a bouncing trajectory through an impact
27. Dynamics and numerical time integration
28. Conserving motion in a physical system
29. Multiple scales and identifiable dynamics
30. Applied study
31. Notation and assumptions
32. Further reading

Chapter 1 · Initial values and propagated error

Derive an error bound for forward Euler and distinguish a poor approximation of a well-posed trajectory from uncertainty in the starting state.

FROM FIRST PRINCIPLES

Follow one local error through the remaining steps

An initial-value problem specifies a rate $y' = f(t, y)$ and a starting value. Integrating the rate over a short interval gives the exact change. Approximating the integrand by the rate at the left endpoint is forward Euler. Suppose the solution is twice differentiable with $\|y''\| \le M$. Taylor's theorem bounds the defect made by a single step starting at the exact value:

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + d_n, \qquad \|d_n\| \le Mh^2/2.$$

The next step generally starts at a computed value y_n. If changing the input state changes the rate by at most L times that change in norm, then for $e_n = y(t_n) - y_n$ subtraction gives

$$\|e_{n+1}\| \le (1 + hL)\|e_n\| + Mh^2/2.$$

Each earlier defect can therefore be amplified by every later step. With exact initial data, sum the resulting geometric series up to fixed time $T = nh$, for $L > 0$:

$$\begin{aligned} \|e_n\| &\le \frac{Mh^2}{2} \sum_{j=0}^{n-1} (1 + hL)^j \\ &= \frac{Mh}{2L} \left[(1 + hL)^n - 1 \right] \\ &\le \frac{Mh}{2L} \left(e^{LT} - 1 \right). \end{aligned}$$

The local h^2 becomes global h because the number of terms grows on refinement. When $L = 0$, their sum is $MTh/2$. An uncertain initial value contributes a separate amplified initial-error term.

Background · Local truncation error

$$d_{n+1} = y(t_{n+1}) - \Phi_h(y(t_n))$$

This is the one-step defect from exact starting data. Some conventions divide it by h. State the convention before comparing local and global orders.

DERIVATION AND APPLICATIONS

Estimate an endpoint error from two grids that share the same endpoint

A local defect need not equal the error in a reported final state. A practical test for a smooth nonstiff initial-value problem is to integrate to the identical physical time on two or more nested time grids. Suppose the method has global order p and its leading error varies smoothly with step size. Write Y_h for its value at the final time and y for the exact value there. The asymptotic expansion and its version on the refined grid are

$$Y_h = y + Ch^p + O(h^{p+1}),$$
$$Y_{h/2} = y + 2^{-p}Ch^p + O(h^{p+1}).$$

Here C depends on the problem and final time but is common to these two runs. Subtract to estimate the error of the fine result. The sign convention below is correction to the fine result, not numerical value minus exact value:

$$y - Y_{h/2} \simeq \frac{Y_{h/2} - Y_h}{2^p - 1},$$
$$Y_{\text{ext}} = Y_{h/2} + \frac{Y_{h/2} - Y_h}{2^p - 1}.$$

A third grid can test whether the nominal p applies. The unknown coefficient cancels in the ratio of successive differences:

$$p_{\text{obs}} = \log_2 \left| \frac{Y_h - Y_{h/2}}{Y_{h/2} - Y_{h/4}} \right|.$$

If two differences have incompatible signs or are close to roundoff, this diagnostic becomes unreliable. A nonsmooth switch, an approximate event location or comparing different physical final times can also destroy the shared expansion. The estimate concerns the specific endpoint output; a zero leading error there need not indicate higher order throughout the trajectory.

WORKED EXAMPLE

Extrapolate a smooth decay calculation

For $y' = -y$, $y(0) = 1$, integrate to $t = 1$ with Euler steps $h = 1/4$ and $1/8$. The endpoint values are

$$(3/4)^4 = 0.316406,$$
$$(7/8)^8 = 0.343609.$$

For $p = 1$ the fine estimate needs a positive correction of 0.027203, giving $Y_{\mathrm{ext}} = 0.370812$. The exact value $e^{-1} = 0.367879$ lies much closer to the extrapolation than either raw value. Extrapolation removes the leading bias but does not make an exact solution.

SECOND APPLICATION

Check the order before trusting an error bar

A different endpoint measurement returns 1.2480, 1.2120 and 1.2030 on successive halvings. Their differences are 0.0360 and 0.0090, yielding $p_{\mathrm{obs}} = 2.00$. The finest correction is $-0.0090/(4 - 1) = -0.0030$, so the extrapolated endpoint is 1.2000. Using a first-order denominator would triple the estimated correction.

> **Start from this chapter.** For a scalar exponential decay, the exact one-step amplification is known. Accumulate the numerical factor to predict the final concentration, then check a separate conserved sum.
>
> OneSimulation laboratory · Compare time-step error with conservation

Specify what the trajectory represents

In an initial-value problem $y' = f(t, y)$, $y(t_0) = y_0$, the time history follows from both the law and the initial state. A measurement made up to an absolute uncertainty gives a family of possible trajectories even before selecting a method. A numerical calculation adds further deviations each time it replaces the continuous flow over a short interval by an algebraic update. To estimate the final error, track how deviations arising early in a run are transported through later steps.

If f is continuous in time and locally Lipschitz in the state, the initial-value problem has a unique local solution. Lipschitz continuity is more than a technical detail. For $y' = 2\sqrt{|y|}$ with $y(0) = 0$, solutions may remain zero up to any chosen nonnegative time c and then follow $(t - c)^2$. Continuity alone gives no unique departure time. A code that keeps returning zero is computing one admissible trajectory, but there is no single exact departure trajectory against which to grade it without further modelling information.

Accumulate one-step errors

Write the exact integral from t_n to $t_n + h$. Approximating the integrand by its value at the start gives

$$y_{n+1} = y_n + h f(t_n, y_n).$$

Inserting an exact smooth solution in this update produces a one-step defect bounded by $Mh^2/2$, where M bounds $|y''|$ over the interval. This defect starts a step from the exact state. Actual steps generally start from an already perturbed state.

Let $e_n = |y(t_n) - y_n|$, with state Lipschitz bound L. Adding and subtracting a step started at the exact value gives

$$e_{n+1} \le (1 + hL)e_n + Mh^2/2.$$

Iterate this inequality as a geometric series. For $L > 0$ and elapsed time $T = nh$, use $(1 + hL)^n \le e^{LT}$ to conclude

$$e_n \le e^{LT} e_0 + \frac{Mh}{2L}(e^{LT} - 1).$$

For $L = 0$ the second term becomes $MTh/2$. A local error of order two gives order one over fixed time, while the initial uncertainty contributes separately. The bound treats every perturbation as if amplified. A dissipative system may damp them strongly, so a more problem-specific calculation can be much sharper.

Background · Which direction does an earlier local error take?

For the two-component system $y' = Ay$, an additive state error introduced at time t_j contributes $e^{A(T-t_j)}d_j$ to the endpoint. A norm estimate counts the magnitudes of those contributions. Their directions can reinforce or cancel. For undamped rotation at angular speed ω, equal displacement errors introduced half a period apart reach the endpoint in opposite directions and cancel exactly.

This is why error bounds add absolute contributions while an experiment can show a much smaller final error. A single endpoint near cancellation gives a poor estimate of an integration method's order. Check errors at more than one observation time, or use a maximum-in-time norm, when the quantity of interest covers the entire path.

Test the prediction against exact exponential growth

Take $y' = y,\ y(0) = 1$ over one time unit. Forward Euler with two steps gives $1.5^2 = 2.25$; with four it gives

$$1.25^4 = 2.44141.$$

The exact result is $e = 2.71828$. The errors 0.46828 and 0.27688 decrease by a factor about 1.69. The limiting factor of two requires a finer asymptotic regime. Expanding $N \log(1 + 1/N) = 1 - 1/(2N) + O(N^{-2})$ predicts $e - (1 + 1/N)^N = e/(2N) + O(N^{-2})$.

If the initial value is actually 1.01, its exact trajectory at time one differs from the nominal one by $0.01e$. Mesh refinement of the nominal solver cannot decide which initial reading is correct. To evaluate an integration scheme experimentally, fix the model and initial data, compare at a common endpoint and record the norm used. Resampling a coarse trajectory at extra plot positions changes the reported curve but does not alter the integration error accumulated at its original nodes.

Use a manufactured trajectory to test a solver

Instead of relying on a familiar unforced exponential, choose an exact function and derive the source that produces it. For $y' + 2y = s(t)$, selecting $y_\star = 1 + \sin t$ requires $s = \cos t + 2(1 + \sin t)$ and $y(0) = 1$. The forcing and damping are both active. A solver that silently evaluates the source at the wrong stage time can pass constant-source tests but fail this one.

Over one Euler step of 0.2 the initial derivative is one, so $y_1 = 1.2$ versus the exact $1 + \sin(0.2) = 1.19867$. This small first-step discrepancy is unusually favourable because $y_\star''(0) = 0$. It does not imply the global method has become second order. Report an error over a fixed time interval with multiple nontrivial phases of the forcing. A single specially located sample can suppress the leading contribution and give a misleading estimate.

Measure the defect of an entire reconstructed trajectory

Suppose a time integrator supplies a differentiable curve $U(t)$ between its stored nodes. Its continuous defect is $d(t) = U'(t) - f(t, U(t))$, the forcing that would make the reconstructed curve satisfy a perturbed differential equation exactly. For a scalar linear equation $y' = a(t)y + b(t)$, the state error $e = U - y$ satisfies

$$e' = a(t)e + d(t).$$

Multiplication by an integrating factor gives the exact identity

$$e(T) = \exp\left(\int_0^T a\right) e(0) + \int_0^T \exp\left(\int_s^T a(\xi)\, d\xi\right) d(s)\, ds.$$

This equation distinguishes defects by when they occur. In a dissipative equation with $a \le -\alpha < 0$, older defects receive exponentially smaller weights. If $|d| \le D$, the resulting bound is $e^{-\alpha T}|e(0)| + D(1 - e^{-\alpha T})/\alpha$. Simply multiplying the worst defect by the duration misses this damping. For a growing mode the weighting reverses: even a modest defect introduced early can dominate the final state. A defect must be examined between nodes; sampling only locations where a collocation method enforces the equation can return zero by construction.

Check a polynomial approximation to a driven relaxation

Take $y' = -2y + t,\ y(0) = 0$, and on $[0, 1]$ use the approximate curve $U(t) = t^2/2$, obtained by integrating the input while ignoring damping. Then $U' = t$, while the actual right-hand side at U is $t - t^2$. The defect is $d(t) = t^2$, zero initially and largest at the final time. Its sign shows that the approximate curve is receiving an artificial positive input. The crude bound with $D = 1, \alpha = 2$ gives final error at most $(1 - e^{-2})/2 \approx 0.4323$, already tighter than the undamped bound of one.

Solve the true equation with an integrating factor:

$$y(t) = t/2 - 1/4 + e^{-2t}/4.$$

At the endpoint the error is

$$0.5 - (0.25 + e^{-2}/4) = (1 - e^{-2})/4 \approx 0.2162.$$

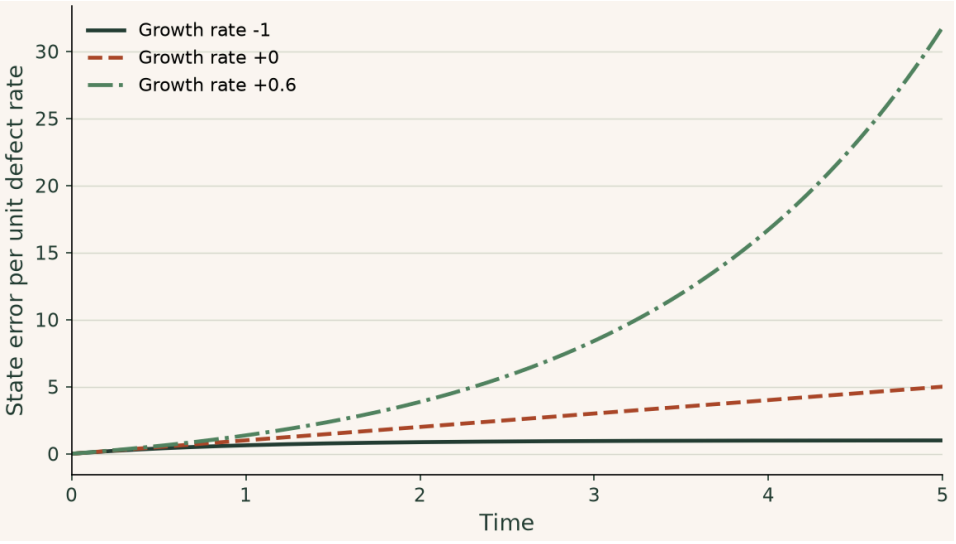
Independently integrating the defect with the correct damping weight gives $\int_0^1 e^{-2(1-s)} s^2 ds = (1 - e^{-2})/4$, exactly reproducing it. The difference between the two estimates has an interpretation: the simpler bound assumed the worst defect was present throughout, while the actual curve develops most of its error only near the end.

One can even use the defect to form a correction by solving the linear error equation and subtracting the estimated error from U. For a nonlinear problem linearising around the reconstructed path introduces a new remainder and the correction is no longer exact. Still, it targets the right mechanism: identify where the reconstructed curve fails the original equation and propagate that imbalance to the particular observation time instead of judging a trajectory only by its nodal appearance.

The same forcing identity can distinguish two curves passing through identical initial and final data. One curve may solve the differential equation nearly everywhere, while another contains positive and negative defects whose weighted integrals cancel exactly at the final time. If the experiment also observes a midpoint, that cancellation no longer guarantees agreement there. A dense-output validation should therefore compare defect over the full interval and state error at the actual observation times, especially when a numerical interpolant is used for later event finding.

Propagate a persistent local defect through an evolution

Model the accumulation of small local errors by $e' = \lambda e + 1,\ e(0) = 0$. Variation of constants integrates each earlier defect with weight $e^{\lambda(t-s)}$. Compare a contracting flow, a neutral direction and an unstable direction.



For $\lambda = -1$ the normalised error approaches one; for $\lambda = 0$ it grows linearly. At $\lambda = 0.6$ it grows exponentially. The same local accuracy target can therefore result in different long-time global errors.

▶ **Read values from the plot**

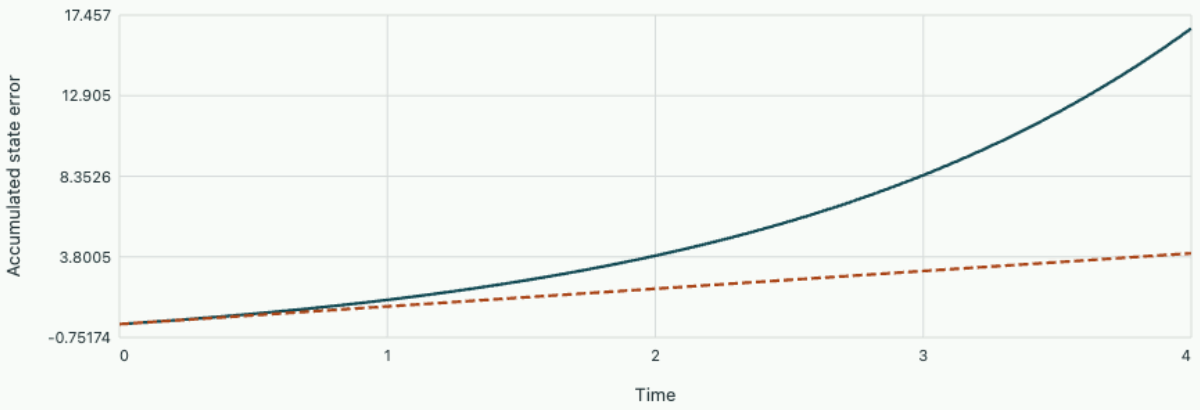
Change parameters · Propagate a persistent local defect through an evolution

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

Growth rate λ	0.6
Defect rate	1
End time	4



This solves e′ = λe + defect from e(0) = 0. Equal local defects can saturate, accumulate linearly or grow exponentially depending on the underlying evolution.

Study the full derivation

Find a crossing between accepted time steps · Dynamics and numerical time integration

An integrator can approximate a trajectory accurately yet report a poor switch time. Construct a continuous interpolant, relate a state error to an event-time error and inspect what happens when a trajectory merely touches a threshold.

Chapter 2 · Runge–Kutta stages and order

Construct a two-stage second-order integrator from Taylor conditions and calculate how different stage choices affect autonomous and time-dependent problems.

FROM FIRST PRINCIPLES

Choose a second slope so that the curvature term is correct

A rate at the start of a step gives a tangent. To see what a single tangent misses, differentiate $y' = f(t, y)$ along the solution. Time changes the slope directly and the moving state changes it through the Jacobian f_y. For a smooth right-hand side the exact increment begins

$$y(t+h) - y(t) = hf + \frac{h^2}{2}(f_t + f_y f) + O(h^3).$$

Construct a second slope at a fraction c of the interval. The exact intermediate state is unknown, so predict it with the first slope $k_1 = f(t, y)$. Expand the new evaluation:

$$k_2 = f(t + ch, y + chk_1) = f + ch(f_t + f_y f) + O(h^2).$$

Take a weighted average of both slopes,

$$y_{n+1} = y_n + h(b_1 k_1 + b_2 k_2).$$

Matching the exact coefficients of h and h^2 imposes two conditions:

$$b_1 + b_2 = 1,$$
$$b_2 c = \tfrac{1}{2}.$$

Choose $c = 1/2$ and the first slope guides the midpoint location while the second supplies the whole update. Choose $c = 1$ and the two slopes receive equal weights. Both constructions have local defect $O(h^3)$; they can disagree at the next order because slope matching has imposed no condition on cubic terms.

DERIVATION AND APPLICATIONS

Why two second-order Runge–Kutta formulas diverge on a nonlinear system

A linear decay test cannot distinguish every pair of Runge–Kutta methods. Compare a family of explicit two-stage schemes on a smooth autonomous scalar equation $y' = f(y)$. Starting from the same y,

choose a nonzero stage location a, compute $k_1 = f(y)$ and $k_2 = f(y + ahk_1)$, and update with weights b_1, b_2. Expand the internal slope at the unperturbed state:

$$k_2 = f + ahf'f + \tfrac{1}{2}a^2h^2f''f^2 + O(h^3).$$

Insert this expansion into $y_+ = y + h(b_1k_1 + b_2k_2)$. Matching the coefficients of h and h^2 with the exact Taylor series requires $b_1 + b_2 = 1$ and $ab_2 = 1/2$. The update for every such method then reads

$$y_+ = y + hf + \tfrac{1}{2}h^2f'f + \tfrac{1}{4}ah^3f''f^2 + O(h^4).$$

Differentiate the differential equation twice to obtain the exact cubic term. This isolates two distinct effects of nonlinearity and slope propagation:

$$y''' = f''f^2 + (f')^2f,$$
$$y(t+h) = y + hf + \tfrac{1}{2}h^2f'f + \tfrac{1}{6}h^3[f''f^2 + (f')^2f] + O(h^4).$$

Changing a can adjust the coefficient of $f''f^2$. No choice supplies the missing $(f')^2f$ term for arbitrary f, explaining why a third independent stage is needed for a general third-order explicit method. When f is linear, $f'' = 0$, so this two-stage family gives the same output regardless of a.

WORKED EXAMPLE

Compare one nonlinear growth step

For $y' = y^2$, $y(0) = 1$ and $h = 0.100$, explicit midpoint uses $a = 1/2$ and $y_+ = 1 + 0.1(1.05)^2 = 1.11025$. Heun uses $a = 1$, averages 1 and $(1.10)^2$, and gives 1.11050. The exact solution $1/(1-t)$ yields 1.11111. Both underpredict; Heun happens to be closer on this step. That comparison alone does not establish a generally higher order.

SECOND APPLICATION

Construct a test unable to separate the methods

With $y' = -2y$, $y(0) = 3$ and $h = 0.100$, every second-order member has amplification $1 + z + z^2/2$ at $z = -0.200$. The output is $3(0.820) = 2.46$. The exact result is

$$3e^{-0.2} = 2.45619.$$

Matching between midpoint and Heun in this test checks their shared polynomial but cannot test the implementation of their nonlinear stage terms.

Start from this chapter. Halve the step at the same final time. Use the fourth-order Runge–Kutta prediction for the error ratio and compare it with the maximum error over the whole trajectory.

OneSimulation laboratory · Compare time-step error with conservation

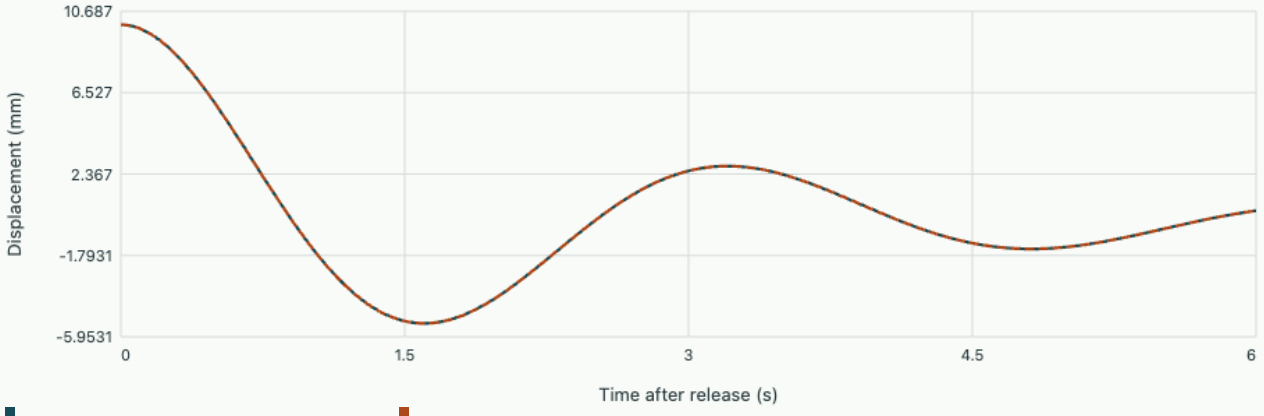
Reproduce a four-stage time integration in OneSimulation

Release a 1 kg mass from 10 mm at rest on a 4 N/m spring. Introduce damping $c = 2\zeta\sqrt{km}$. With $v = \dot{x}$ one step must advance the coupled vector (x, v). Each Runge–Kutta stage evaluates both v and $-4x - 4\zeta v$ at the same intermediate state. Using a new velocity with an old displacement changes the method. Predict whether a displacement zero crossing is possible for $\zeta = 0.5, 1, 2$.

Open this experiment online

Try the calculation

Damping / critical damping **0.2**



Fourth-order Runge–Kutta Characteristic-root solution

At 6 s the displacement is 0.49230 mm. Largest difference from the exact response on 600 steps: 2.46e-8 mm.

Download the OneSimulation model

In your OneSimulation environment, run the model with these inputs.

```
python oscillator.py --ratio 0.2
```

The plot integrates six seconds with a step of 0.01 s. Its independent reference comes from the roots of $s^2 + 4\zeta s + 4 = 0$. At the repeated root $\zeta = 1$, use $x = 10(1 + 2t)e^{-2t}$ mm. Take the same damping into the downloadable OneSimulation release model and compare individual CSV rows. If you modify its time step, compare the largest error on shared output times and check the fourth-order ratio while the error remains above roundoff.

Sample slopes without solving a new differential equation

Euler uses a single slope at the start of a step. If the direction changes significantly over the interval, a second evaluation at a predicted point can estimate that change. Consider $k_1 = f(t, y)$, $k_2 = f(t + ch, y + ahk_1)$ and $y_{new} = y + h(b_1 k_1 + b_2 k_2)$. The time and state arguments of a stage must both be specified. In a forced system, evaluating a source at the old time while using a midpoint state can remove the expected order.

Expand around the start of the step. The second stage is $f + h(cf_t + af_y f) + O(h^2)$. The exact trajectory has $y(t + h) = y + hf + h^2(f_t + f_y f)/2 + O(h^3)$. Comparing coefficients gives

$$b_1 + b_2 = 1,\ b_2 c = 1/2,\ b_2 a = 1/2.$$

For the usual internally consistent stages $a = c$. The explicit midpoint rule takes $a = c = 1/2,\ b_2 = 1$; the explicit trapezoidal or Heun rule takes $a = c = 1,\ b_1 = b_2 = 1/2$. Both have order two globally under regularity and stable error propagation.

Compute two different approximations of equal order

For $y' = y^2,\ y(0) = 1$, use $h = 0.1$. The exact solution $1/(1 - t)$ gives 1.111111 at $t = 0.1$. Explicit midpoint predicts a midpoint value 1.05, evaluates slope $1.05^2 = 1.1025$ there and returns $1 + 0.1(1.1025) = 1.11025$. Heun predicts an endpoint value 1.1, averages endpoint and initial slopes to obtain 1.105 and returns 1.11050. Both improve on Euler’s 1.1. Equal formal order allows different error constants, even for the same step size.

The exact solution develops a pole at $t = 1$. Approaching that time its derivatives grow and a step that was satisfactory at the start can perform poorly. This illustrates the dependence of all Taylor estimates on bounds over the interval of interest. A finite numerical value beyond the exact blow-up time is not automatically a valid continuation of the continuous solution. The physical or mathematical existence interval is part of any convergence claim.

Check what a higher-order stage sequence buys

Classical RK4 evaluates one initial slope, two midpoint slopes based on successive predictions and one endpoint slope. It averages them with weights $1, 2, 2, 1$ divided by six. For a constant linear scalar mode $y' = \lambda y$, substitution gives

$$R(z) = 1 + z + z^2/2 + z^3/6 + z^4/24,$$
$$z = h\lambda.$$

Its first missing exponential term has degree five. This test explains fourth order on that mode. General nonlinear fourth order requires additional stage conditions because derivatives such as $f_{yy}(f, f)$ and $f_y f_y f$ are independent.

Take $\lambda = -1,\ h = 1$ and $y_0 = 1$. The four slopes are $-1, -0.5, -0.75, -0.25$, yielding $y_1 = 0.375$ against $e^{-1} = 0.36788$. The two-stage midpoint rule gives 0.5 at the same endpoint. Cost matters: a comparison at equal steps lets RK4 use twice as many right-hand-side evaluations as a two-stage scheme. At equal evaluation budgets one can instead compare one RK4 step with two midpoint steps. Over a complex trajectory it is the measured work for an agreed error tolerance that answers which method is economical.

OneSimulation · Measure numerical damping in a lossless oscillator

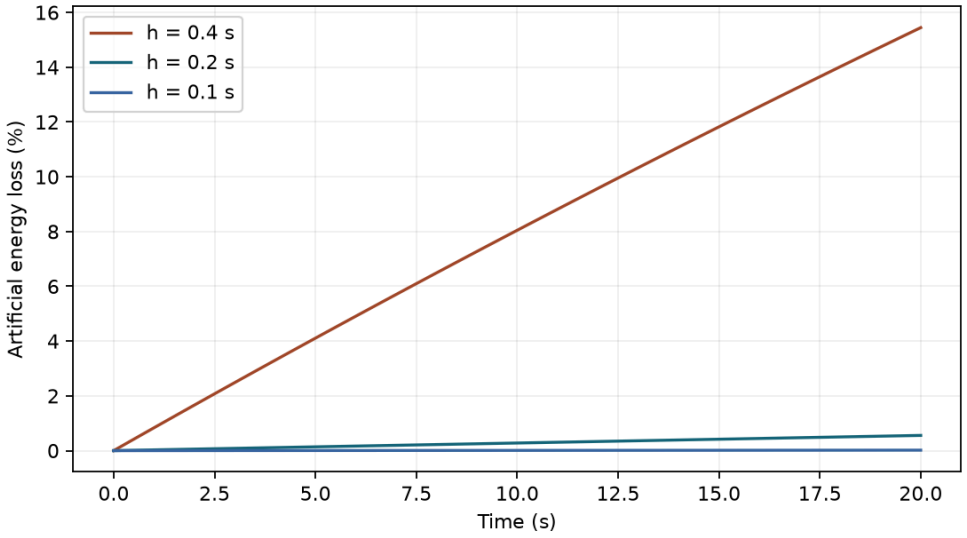
An undamped 1 kg oscillator has angular frequency 2 rad/s. The initial displacement is 1 m and the initial velocity is zero, giving constant energy 2 J. In energy-scaled coordinates the exact motion is a rotation. A classical RK4 step multiplies the two complex modes by $R(\pm iz)$, with $z = \omega h$. Consequently one step changes the energy by the exact factor

$$\frac{E_{j+1}}{E_j} = |R(iz)|^2 = 1 - \frac{z^6}{72} + \frac{z^8}{576}.$$

OneSimulation advances the original position–velocity equations. For each step size the code checks the computed energy at 20 s against the factor raised to the number of steps. The accumulated losses are 15.44%, 0.556% and 0.0177% as the step is halved twice. For fixed final time and small h, the number of steps grows like $1/h$, so the leading energy loss is proportional to h^5. This diagnostic has its own convergence rate even though the method is fourth order in the state.

```
"""Measure artificial energy loss over 20 seconds in an undamped oscillator."""
import matplotlib.pyplot as plt
import numpy as np
from onesim.numerics.time import ClassicalRungeKutta4TimeIntegrator

omega = 2.0  # rad/s
mass = 1.0  # kg
duration = 20.0  # s
method = ClassicalRungeKutta4TimeIntegrator()

def dynamics(state, time):
    position, velocity = state
    return (velocity, -omega**2 * position)

fig, ax = plt.subplots(figsize=(6.8, 3.8), constrained_layout=True)
losses = []
for n, colour in [(50, "#a4492b"), (100, "#176879"), (200, "#3868a0")]:
    h = duration / n
    state = (1.0, 0.0)  # metres, metres per second
    energies = []
    for j in range(n + 1):
        x, v = state
        energies.append(0.5 * mass * (v*v + omega**2*x*x))
        if j < n:
            state = method.advance(state, dynamics, h, time=j*h)
    fraction = np.array(energies) / energies[0]
    z = omega*h
    assert np.isclose(fraction[-1], (1 - z**6/72 + z**8/576)**n)
    losses.append(1 - fraction[-1])
```

```
    ax.plot(np.arange(n + 1)*h, 100*(1 - fraction), color=colour, label=f"h = {h:g} s")
print("Final percentage losses:", np.round(100*np.array(losses), 5))
ax.set(xlabel="Time (s)", ylabel="Artificial energy loss (%)")
ax.legend()
ax.grid(alpha=0.2)
fig.savefig("ordinary-differential-equations.png", dpi=160)
```



Energy decreases most with 0.4 s steps. The exact loss is zero at every time; the curves therefore measure damping introduced entirely by the discrete update.

Download the Python example

Check a linear invariant at the internal stages

Suppose $y = (A, B)$ and the right-hand side obeys $f_A + f_B = 0$ at every admissible state. At any explicitly constructed stage the vector of slopes still has zero component sum. In the Runge–Kutta update $y_{n+1} = y_n + h \sum_j b_j k_j$, the sum of the increments therefore vanishes for any weights. Every method of this form preserves the linear invariant $A + B$ in exact arithmetic, regardless of its order.

A protection against negative concentrations is a different requirement. With $A' = -kA$ and $kh = 3$, explicit midpoint forms the internal value $A_{\mathrm{mid}} = A_n(1 - 3/2) = -0.5A_n$ even though the final polynomial $1 - 3 + 9/2$ is positive. A nonlinear rate law involving a logarithm of concentration would be undefined at this intermediate evaluation. Checking final stored states alone misses the problem. This is why a useful stage test examines both admissibility of intermediate evaluations and the conservation of the accepted update.

Compare work at a fixed endpoint

Comparing two Runge–Kutta schemes on the same step size need not compare the same computational budget. Record the number of right-hand-side evaluations and the actual endpoint error. When evaluating the model dominates all other costs, four one-stage steps and one four-stage step cost

roughly the same, yet respond quite differently to curvature of the solution. Some implementations reuse a stage across accepted steps; count evaluations actually made.

Derive implicit stages from polynomial collocation

Let $U(\theta)$ denote a polynomial approximation to $y(t_n + h\theta)$ on one time slab, with $U(0) = y_n$. At s distinct nodes c_i, require

$$U'(c_i) = hf(t_n + hc_i, U(c_i)) = hK_i.$$

A polynomial U of degree at most s has a derivative determined by those s values. With cardinal polynomials ℓ_j through the nodes, integration gives

$$U(\theta) = y_n + h\sum_j K_j \int_0^\theta \ell_j(r)\,dr.$$

Evaluating at the nodes and at the endpoint turns a polynomial matching problem into a Runge–Kutta scheme:

$$a_{ij} = \int_0^{c_i} \ell_j(r)\,dr,$$

$$b_j = \int_0^1 \ell_j(r)\,dr,$$

$$K_i = f(t_n + hc_i, y_n + h\sum_j a_{ij}K_j).$$

The stage values generally depend on one another and must be solved simultaneously. With the single midpoint node this gives the implicit midpoint step, since $\ell_1 = 1$, $a_{11} = 1/2$ and $b_1 = 1$. With two Gaussian nodes $c_{1,2} = 1/2 \mp \sqrt{3}/6$, integrating their linear cardinal functions gives weights $1/2, 1/2$, diagonal entries $1/4$, and off-diagonal entries $1/4 - \sqrt{3}/6,\ 1/4 + \sqrt{3}/6$ in row order.

For $y' = \lambda y$, write $z = h\lambda$ and $Y = \mathbf{1}y_n + zAY$ for the stage states. Solving this two-by-two system and substituting in $y_{n+1} = y_n + z(Y_1 + Y_2)/2$ gives amplification $(1 + z/2 + z^2/12)/(1 - z/2 + z^2/12)$. For $y_0 = 1, \lambda = -3, h = 1$, the result is $1/13 \approx 0.07692$, against $e^{-3} \approx 0.04979$. The endpoint formula is fourth order as $h \to 0$. For arbitrarily large negative z this ratio tends to one. Accurate integration of smooth resolved trajectories and rapid removal of unresolved stiff modes remain distinct design goals.

Practise on sheet 1 · Measure a collocation curve between its Gaussian nodes. Further reading · ETH Zürich · Numerical Methods for Ordinary Differential Equations · §4.4.1.1 Collocation methods, pp. 53–54.

Match the two independent third-order terms

For an autonomous smooth vector field write $J = f_y$ and f'' for its second derivative. Differentiating $y' = f(y)$ twice gives $y'' = Jf$ and $y''' = f''(f,f) + J(Jf)$. A third-order step has to reproduce both distinct vector expressions. For Runge–Kutta stages $k_i = f(y + h\sum_j a_{ij}k_j)$, put $c_i = \sum_j a_{ij}$. Stage expansion gives

$$k_i = f + hc_iJf + h^2[\sum_j a_{ij}c_jJ(Jf) + (c_i^2/2)f''(f,f)] + O(h^3).$$

Multiplying by the output weights hb_i and comparing independent Taylor terms gives the conditions.

Besides $\sum b_i = 1$ and $\sum b_ic_i = 1/2$, they are $\sum b_ic_i^2 = 1/3$ and $\sum_{ij} b_ia_{ij}c_j = 1/6$. The first additional condition concerns curvature of the vector field along a displacement; the second concerns propagation of one derivative through another stage. A single test equation with linear f makes the curvature term vanish, so matching an exponential stability polynomial cannot alone certify third order on general nonlinear systems. Smoothness over the region visited by all stages underlies the Taylor argument.

Verify a three-stage method on nonlinear growth

Use the stages $k_1 = f(y)$, $k_2 = f(y + hk_1/2)$, $k_3 = f(y - hk_1 + 2hk_2)$ and output $y + h(k_1 + 4k_2 + k_3)/6$. The stage abscissae are $0, 1/2, 1$. The squared-abscissa sum is

$$(4/6)(1/4) + (1/6) = 1/3.$$

For the nested derivative only the third stage contributes:

$$b_3a_{32}c_2 = (1/6)(2)(1/2) = 1/6.$$

The two lower-order conditions are checked by summing the weights and their first moments. Together these calculations show why the intermediate stages are arranged this way.

Take $y' = y^2,\ y(0) = 1,\ h = 0.1$. The exact value is

$$1/(1-h) = 1.111111\ldots.$$

The stages are

$$k_1 = 1,\ k_2 = 1.05^2 = 1.1025,$$

$$k_3 = (1 - 0.1 + 0.2 \cdot 1.1025)^2 = 1.1205^2 = 1.25552025.$$

Their weighted combination gives

$$y_1 = 1 + 0.1(1 + 4.41 + 1.25552025)/6 \approx 1.11109200.$$

The local error is about -1.91×10^{-5}, small despite appreciable curvature in the vector field. The stages themselves are sampling states for evaluating slopes; their individual values are not third-order predictions at those times.

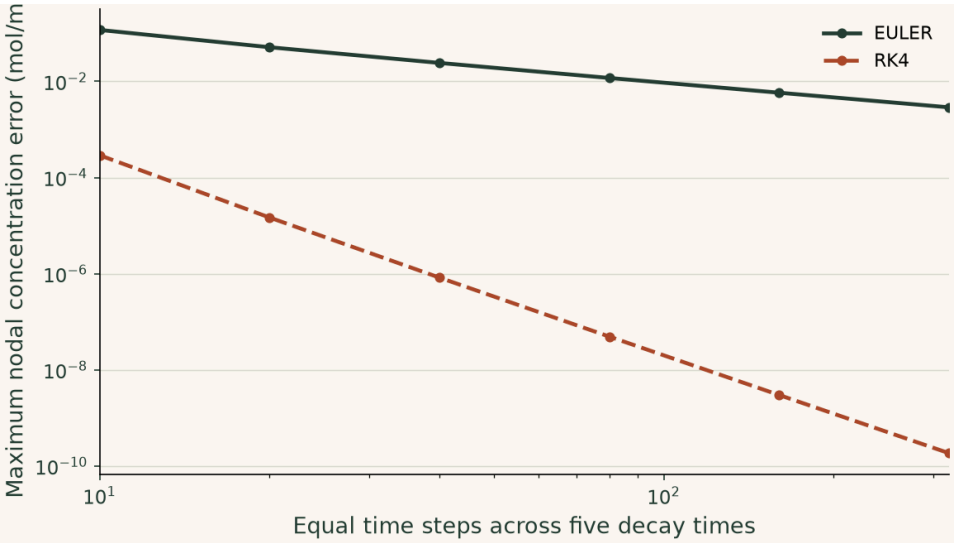
Near $t = 1$ the exact solution becomes unbounded, and the derivatives used to bound the truncation error grow sharply. A method's algebraic order is a local smoothness statement, not a promise that a fixed step will remain accurate up to a singularity. Comparing errors at successively refined fixed observation times below the singularity tests the claimed order under the assumptions actually used in the derivation.

There is a quick nonlinear comparison showing why the two third-order conditions are genuinely needed. For a constant matrix equation the term $f''(f, f)$ vanishes identically, so only the nested derivative contributes at cubic order. For a scalar quadratic field, both terms are generally nonzero. The exact-series test in the worked example can thus reveal defects that would escape every linear growth or decay test using the same code. One can isolate the curvature term more sharply by placing the initial state at a point where $J = 0$ but $f \neq 0$ and $f'' \neq 0$, for example $y' = 1 + y^2$ at $y = 0$. Its solution begins $y = t + t^3/3 + \cdots$, providing an independent check purely on the squared-abscissa moment.

In fact $y = \tan t$ in this test, and the coefficient of its cubic term can be measured directly without numerical reference integration.

Measure fourth order against a conserved decay

The OneSimulation run transfers A into B at rate 0.8 per second. Start with one mole of A per cubic metre and no B. Compare the recorded A values at every time node with the exponential solution. Both integrators conserve the total A+B, so a total-mass test alone cannot distinguish their time accuracy.



Doubling the number of steps reduces the fine-grid Euler error by about two and the RK4 error by about sixteen. Read the last two points to estimate the order yourself. Each RK4 step evaluates four slopes, so compare 40 Euler steps with 10 RK4 steps when counting right-hand-side evaluations.

▶ **Read values from the plot**

Download the OneSimulation model (Python)

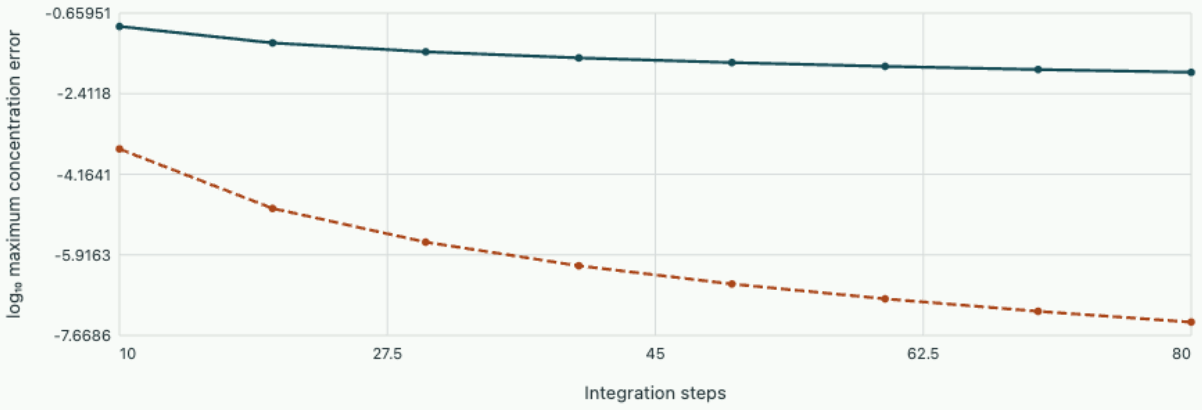
Change parameters · Measure fourth order against a conserved decay

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

Reaction rate k (1/s)	**0.8**
End time (s)	**6**
Maximum steps	**80**



The model transfers A to B with A + B = 1 and initial A = 1. Both methods conserve that sum; time accuracy differs. RK4 costs four slope evaluations per step.

Study the full derivation

Find a crossing between accepted time steps · Dynamics and numerical time integration

An integrator can approximate a trajectory accurately yet report a poor switch time. Construct a continuous interpolant, relate a state error to an event-time error and inspect what happens when a trajectory merely touches a threshold.

Problem sheet 1 · From one update to a trajectory

Set up a global error recurrence, verify Runge–Kutta order conditions and identify the role of explicitly time-dependent forcing.

Propagate a sequence of local disturbances

A one-step algorithm satisfies $e_{n+1} \le 1.05 e_n + 0.0002$ K, and initially $e_0 = 0.01$ K. Derive a bound after ten steps. How much is due to the initial uncertainty and how much is attributable to defects during the run?

▼ **Worked solution**

Iterating the recurrence gives $e_{10} \le 1.05^{10} e_0 + 0.0002(1.05^{10} - 1)/(1.05 - 1)$. Using $1.05^{10} = 1.62889$ yields

$$0.01629 + 0.00252 = 0.01881\,\mathrm{K}.$$

A more accurate numerical step reduces the second contribution directly; reducing uncertainty in the initial measurement reduces the first. The bound allows maximally aligned and amplified disturbances at every step and need not be attained on a particular trajectory.

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.

Integrate a time ramp

For $y' = 3t + 1$ with $y(0) = 0$, take one step of length 0.4. Compute the Euler, explicit midpoint and Heun values. Explain their performance and predict what happens if both Heun stages mistakenly evaluate the forcing at the old time.

▼ **Worked solution**

Euler gives 0.4. At the midpoint the slope is $1 + 3(0.2) = 1.6$, so midpoint gives 0.64. Heun averages slopes 1 and 2.2 and likewise gives 0.64. Integrating analytically gives $y = t + 1.5t^2$, hence 0.64.

A linear forcing is integrated exactly by both second-order rules. If both evaluations use the old time the two slopes coincide and Heun degenerates into the Euler estimate 0.4 here, even if its state predictor is correct.

Explore the related calculation with adjustable inputs. Change the decay rate and forcing ramp, and compare the exact convolution with a frozen-input step.

Design a second-order formula

A two-stage explicit method uses its second stage at $c = a = 2/3$. Find both output weights for order two. Apply the resulting scheme to $y' = y,\ y_0 = 2,\ h = 0.3$ and compare with the exact endpoint.

▼ **Worked solution**

The order conditions are $b_2 c = 1/2$ and $b_1 + b_2 = 1$, giving $b_2 = 3/4,\ b_1 = 1/4$. The first slope is 2 and the second is

$$2 + 0.2(2) = 2.4.$$

The update gives

$$2 + 0.3(0.25 \cdot 2 + 0.75 \cdot 2.4) = 2.69.$$

The exact value is

$$2e^{0.3} = 2.69972.$$

Every explicit two-stage second-order method has the same stability polynomial on this constant linear equation, so this test alone cannot distinguish their different nonlinear error constants.

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.

Measure a collocation curve between its Gaussian nodes

On a unit step solve $y' = t^3,\ y(0) = 0$ by the two-node Gaussian collocation method. Determine the linear interpolant of the derivative at the two nodes, integrate it, and compare the endpoint and midpoint with the exact trajectory. The two Gaussian nodes satisfy $c^2 - c + 1/6 = 0$.

▼ **Worked solution**

At either node,

$$c^3 = c(c - 1/6) = (5/6)c - 1/6.$$

Thus the interpolated derivative is $(5/6)t - 1/6$ and $U(t) = (5/12)t^2 - t/6$. Its endpoint is $1/4$, exactly the Gaussian quadrature integral of the cubic forcing.

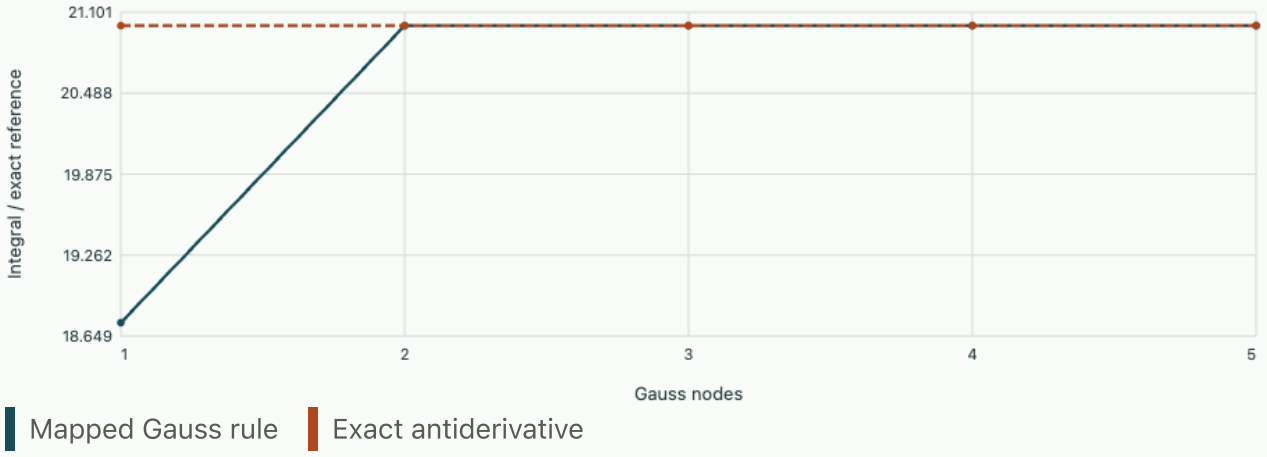
At $t = 1/2$ the collocation polynomial gives $1/48$, whereas the true value $t^4/4$ is $1/64$. The excess is $1/192$. An exact integrated endpoint is compatible with an imperfect continuous curve. Endpoint superconvergence alone does not qualify an event interpolant.

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

Open this adjustable calculation

Try the calculation

Interval left endpoint	1
Interval length	3
Monomial degree	2



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

Revisit chapter 2 · Derive implicit stages from polynomial collocation.

Predict when a conservative calculation loses all reactant

Keep the same final time $5/k$ and take five equal steps for the conversion $A \to B$ starting at $A = 1,\ B = 0$. Derive the first Euler update, the exact final A, and the five-step RK4 prediction using its fourth-degree stability polynomial. Which mass-balance diagnostic would report a mistake?

▼ **Work through the solution**

Here $kh = 1$, so Euler transfers all of A in the first step. Its final A is zero while the exact value is $e^{-5} = 0.00673795$ mol/m³. At $z = -1$ the RK4 multiplier is $1 - 1 + 1/2 - 1/6 + 1/24 = 3/8$, giving $(3/8)^5 = 0.00741577$ mol/m³. In every stage the increments to A and B are opposite, so a check on their total alone detects neither species error. Run the download first with five Euler steps and then with five RK4 steps.

Use the OneSimulation laboratory to check your answer

Chapter 3 · Absolute stability and stiffness

Determine the step sizes that damp a decaying mode and explain why fast, physically uninteresting modes can dominate the cost of explicit integration.

FROM FIRST PRINCIPLES

Find a stability limit directly from repeated multiplication

A small perturbation around an equilibrium often satisfies a linear equation $y' = \lambda y$. Begin with real $\lambda < 0$, for which the continuous value is multiplied by $e^{h\lambda}$, strictly between zero and one. Apply Euler's left- and right-endpoint approximations to the integral of this same rate. Let $z = h\lambda$:

$$y_{n+1} = y_n + z y_n,$$
$$y_{n+1} = y_n + z y_{n+1}.$$

Solve each update for its new value and iterate n steps. Persistence of roundoff or initial error is governed by the same powers:

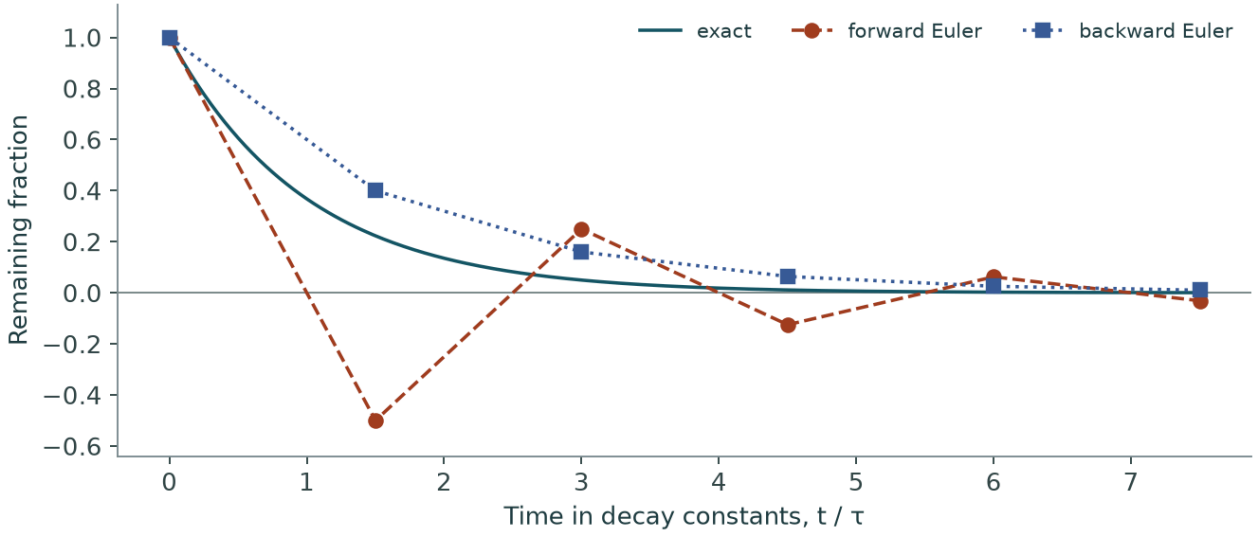
$$y_n = R(z)^n y_0,$$
$$R_{\text{forward}} = 1 + z,$$
$$R_{\text{backward}} = (1 - z)^{-1}.$$

For strict decay, $|R| < 1$ is necessary. Forward Euler on this real mode requires $-2 < z < 0$, and avoiding sign alternation further requires $z \geq -1$. Backward Euler damps every negative real z, even when the step is too large to approximate a transient accurately. With modes at -1 and -1000 s^{-1}, the fastest mode restricts a forward step to less than 0.002 s even after that transient has become tiny. This mismatch between stability and the resolution needed for the slow observable is the source of stiffness in this example.

Background · Stability function

$$y_{n+1} = R(h\lambda) y_n$$

On the scalar linear test equation, absolute stability depends on the point h times lambda. It gives only part of the story for a nonnormal matrix or nonlinear problem.



A stable update can predict an impossible concentration

Take a decay time τ and a time step h = 1.5τ. The exact one-step factor is exp(−1.5) ≈ 0.223. Forward Euler multiplies by −0.5 and backward Euler by 0.4. Both numerical magnitudes decay, yet the first update changes sign. Refinement tests accuracy even when positivity and stability hold.

DERIVATION AND APPLICATIONS

Separate unconditional stability from damping of an unresolved mode

A stiff dissipative mode may have decayed almost completely during a single time step. Merely keeping its computed amplitude bounded is then a weak requirement. Examine the family of one-step formulas that weights the old and new slopes with $1-\theta$ and θ. On the scalar test equation $y' = -\lambda y$, $\lambda > 0$, define the dimensionless decay over a step as $r = h\lambda$. Collecting the new value on the left gives

$$y_{n+1} = y_n - r[(1-\theta)y_n + \theta y_{n+1}],$$

$$G(r) = \frac{y_{n+1}}{y_n} = \frac{1-(1-\theta)r}{1+\theta r}.$$

The denominator is positive for $\theta \geq 0$. The amplification never exceeds one from above; enforcing $G \geq -1$ supplies the other half of the real-axis stability condition:

$$|G| \leq 1 \quad \Longleftrightarrow \quad 2 + (2\theta - 1)r \geq 0.$$

Thus all positive step sizes are stable along the negative real eigenvalue axis for $\theta \geq 1/2$. Now investigate what actually appears in the stored solution when the mode is unresolved:

$$\lim_{r\to\infty} G(r) = -\frac{1-\theta}{\theta} \quad (\theta > 0),$$

$$\lim_{r\to\infty} e^{-r} = 0.$$

The trapezoidal choice $\theta = 1/2$ approaches a sign-reversing factor of unit magnitude. Backward Euler $\theta = 1$ suppresses the fast mode. These properties govern the handling of transients; the order of approximation for a smooth, resolved mode is a separate question.

WORKED EXAMPLE

Check a relaxation after one coarse step

Take $\lambda = 1000$ s⁻¹, $h = 0.100$ s and an initial perturbation of 1.00 K. The exact remnant is e^{-100} K, negligible. Trapezoidal stepping leaves $(1-50)/(1+50) = -0.961$ K, while backward Euler leaves $1/101 = 0.00990$ K. After two trapezoidal steps the perturbation becomes positive again at 0.923 K. That oscillatory persistence is numerical, even though the amplification satisfies the stability condition.

SECOND APPLICATION

Choose a step for monotonically decreasing transients

Suppose a solver uses $\theta = 0.750$ and should preserve positivity for a spectrum of decay rates up to 400 s⁻¹. The numerator of G requires $1 - 0.250h\lambda \geq 0$, hence $h \leq 0.0100$ s. A 0.0200 s step would remain stable on every mode but gives $G = (1-2)/(1+6) = -1/7$ for the fastest one. A positivity requirement can impose a finite time step even when boundedness does not.

> **Start from this chapter.** Use the Euler amplification factor to distinguish stability from nonnegative concentration. Test the step that consumes the entire amount in one update.
>
> OneSimulation laboratory · Compare time-step error with conservation

Examine a single amplification factor

For the real decay problem $y' = -\lambda y$, $\lambda > 0$, the physical solution multiplies by $e^{-h\lambda}$ over a step. A linear one-step method replaces this by $R(z)$, with $z = -h\lambda$. Absolute stability at z means repeated multiplication does not amplify magnitude, conventionally $|R(z)| \leq 1$. A truly decaying mode requires a strict inequality for numerical damping. Positivity and phase fidelity impose additional requirements beyond this magnitude test.

Forward Euler has $R = 1 + z$. In the complex plane the stability region is the disk of radius one centred at -1. On the real negative axis, decay requires $0 < h\lambda < 2$; nonnegative decay requires $h\lambda \leq 1$. Backward Euler has $R = 1/(1-z)$, and the trapezoidal method has $R = (1+z/2)/(1-z/2)$. Both contain the entire left half-plane in their stability regions. As z becomes large and negative, backward Euler tends to zero while trapezoidal tends to minus one. The former strongly damps unresolved fast components; the latter can keep them as alternating numerical transients.

Follow a slow and a fast component together

Let $u' = -u,\ v' = -1000v$, with $u(0) = v(0) = 1$. At $t = 0.1$ the exact slow state is 0.90484, while the fast state is e^{-100}, effectively zero on a plot at unit scale. Suppose the slow state is the desired observable. One forward Euler step of 0.1 s gives $u = 0.9$ but $v = -99$. The method becomes unstable even though the desired slow response is well resolved. To damp the fast mode explicitly, h must stay below 0.002 s whenever that mode can be excited.

Backward Euler with $h = 0.1$ gives

$$u = 1/1.1 = 0.90909,$$

$$v = 1/101 = 0.00990.$$

Trapezoidal gives

$$u = 0.95/1.05 = 0.90476,$$

$$v = -49/51 = -0.96078.$$

The second-order rule is especially accurate on the slow mode but retains a conspicuous oscillatory fast component. Neither the maximum order nor a broad stability region alone identifies the right integrator. The observable, acceptable transient error and stiffness all enter the decision.

Trace the cost back to the model

A diffusion equation discretised on a spatial mesh produces modes with rates up to order D/h_x^2. Halving spatial spacing roughly quadruples the magnitude of the fastest eigenvalue. An explicit time step chosen for a coarse grid can therefore fail after spatial refinement, even if the initial physical signal looks smooth. Small high-frequency components from boundary changes, rounding or nonlinear interactions may excite modes absent from an idealised initial state.

Implicit stability shifts work into linear or nonlinear systems inside each time step. An oversized implicit step can miss a forcing pulse altogether, or produce a poor slow-phase estimate despite decaying every homogeneous mode. Compare the output on two time grids, distinguish the time required to resolve a source from the time required merely for stability, and inspect at least one fast component when interpreting unexpected oscillations. The distinction is visible in a reaction where exact species totals coexist with substantial errors in the individual states. Compare reaction histories under Euler and RK4.

Explain why a tiny initial fast component can still matter

Take the same two-rate model but start with $v_0 = 10^{-12}$ and apply forward Euler with $h = 0.01$. The physical fast mode should disappear almost immediately. The numerical factor is $1 - 10 = -9$, so after twelve steps its magnitude is $9^{12}10^{-12} \simeq 0.282$. An initial disturbance that was invisible in the output has become comparable with the desired slow signal in little physical time.

If the fast variable started exactly at zero in a perfectly decoupled model, exact arithmetic would keep it zero. That special trajectory would give a misleading sense of stability. Roundoff, forcing or a weak coupling can generate a nonzero fast component at later stages. A robust experiment deliberately perturbs fast directions and examines whether numerical damping agrees with the model under those perturbations. This is especially relevant when a mesh refinement adds high-frequency modes that were absent from the coarser spectrum.

Separate damping of a perturbation from tracking a moving state

A useful stiff test has an exact slow path $g(t)$ embedded in $y' = \lambda(y - g(t)) + g'(t)$, with $\lambda < 0$. Subtracting g gives

$$(y - g)' = \lambda(y - g).$$

An initial condition on the path stays on it, whereas an off-path disturbance decays exponentially. A numerical method must therefore address two different tasks: damp erroneous components perpendicular to the slow path and follow the path as it moves. A homogeneous stability test measures only the first.

Apply backward Euler and define $e_n = y_n - g(t_n)$. Subtract $g(t_{n+1})$ after the implicit update. The error satisfies

$$(1 - h\lambda)e_{n+1} = e_n + hg'(t_{n+1}) - [g(t_{n+1}) - g(t_n)].$$

The last two terms are the error in approximating the change of g using its new endpoint derivative. Even with $e_n = 0$, this forcing can create error at the next time. The factor $1/(1 - h\lambda)$ damps both old error and newly created defect. Increasing stiffness changes their balance, while changing the time step alters both the source and its propagation.

Track a quadratic path with a fast relaxation

Take $g = t^2,\ \lambda = -100$ per second, $h = 0.1$ s, $e_0 = 0$. Since $h(2t_{n+1}) - (t_{n+1}^2 - t_n^2) = h^2$, the recurrence is

$$e_{n+1} = (e_n + 0.01)/11.$$

The first error is 0.00090909, so the first computed state is 0.01090909 when the exact value is 0.01. The second error is approximately 0.00099174. Solving the scalar recurrence shows that the error tends to 0.001, the fixed point of $11e = e + 0.01$. Stability does not force it to vanish because each step creates a new tracking defect.

For comparison, trapezoidal integration of the slow-path derivative integrates the linear function $g' = 2t$ exactly over each interval. Substituting the exact nodal values $g(t_n)$ therefore satisfies its full discrete equations, so an exactly initialized trajectory stays on this particular path. However, a small off-path

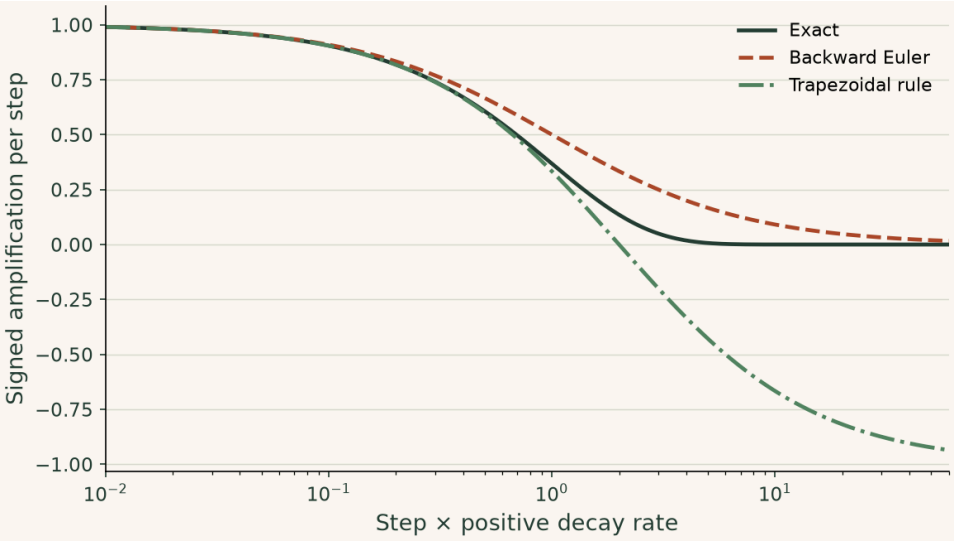
perturbation is multiplied per step by $(1-5)/(1+5) = -2/3$, whereas backward Euler multiplies it by $1/11$. Thus one method is exact on the specified moving path and the other damps unrelated fast perturbations more rapidly.

A numerical test that initialises precisely on g and uses only this quadratic path could make the trapezoidal method appear flawless. Add a known small perturbation and measure its sign and amplitude separately. Conversely a test with $g = 0$ detects damping but cannot expose a tracking defect. Both controlled initialisations are needed to tell whether an observed error arises from resolving the moving forcing or dissipating the artificially fast component.

A final scaling experiment holds the stiffness fixed and halves the step. The long-time tracking error in this quadratic case is $h/|\lambda|$, obtained by cancelling the fixed-point terms in the recurrence. With $h = 0.05$ and the same relaxation rate it becomes 0.0005. A one-step error alone, evaluated from the exact state on each grid, would not reveal the same accumulation mechanism. Report whether an observed order concerns a local defect, a settled tracking error or an initial fast transient.

Separate boundedness from physical decay in a stiff mode

For $y' = -ky$ with positive initial state, an unresolved fast mode should disappear on the reporting timescale. Compare the actual signed factors. A magnitude-only graph would conceal a spurious sign change.



At $h\lambda = 10$ the exact factor is 4.54×10^{-5}, backward Euler gives $1/11$, and trapezoidal gives $-2/3$. The trapezoidal solution remains bounded but an alternating fast component can contaminate physically nonnegative output.

▶ **Read values from the plot**

Change parameters · Separate boundedness from physical decay in a stiff mode

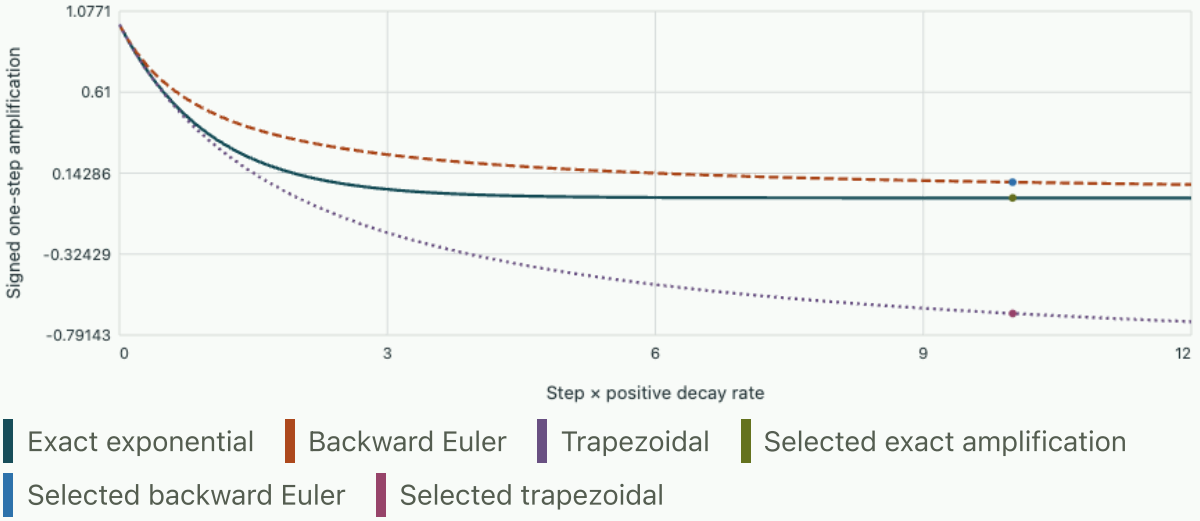
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Maximum step × rate **12**

Selected step × rate **10**



At selected hk = 10: exact 0.0000454, backward Euler 0.090909, trapezoidal -0.66667. Negative amplification creates an alternating unresolved mode.

Study the full derivation

Why stiff stability does not guarantee the expected order · Dynamics and numerical time integration

A differential equation is constructed to have a smooth, known solution and a rapidly damped error mode. Use it to distinguish stability of the homogeneous equation, tracking of a moving target and the effect of off-grid stage forcing.

Chapter 4 · Multistep formulas and parasitic modes

Derive Adams–Bashforth and BDF formulas from interpolation, test a characteristic polynomial and determine whether starting errors survive refinement.

FROM FIRST PRINCIPLES

Integrate the straight line through two past slopes

A one-step method evaluates new slopes inside its current interval. A multistep method can instead reuse the recorded rates $f_n = f(t_n, y_n)$ and f_{n-1}. On a uniform mesh, approximate the rate by the straight line through these two values. At a future offset s from t_n, extrapolation gives

$$p(s) = f_n + \frac{s}{h}(f_n - f_{n-1}), \qquad 0 \le s \le h.$$

Integrating this line gives one area hf_n from the constant part and a triangle from its change. The two-step Adams–Bashforth update follows:

$$y_{n+1} = y_n + \int_0^h p(s)\,ds = y_n + \frac{h}{2}(3f_n - f_{n-1}).$$

For smooth slopes the pointwise interpolation error is $O(h^2)$ here, so integration leaves $O(h^3)$ local defect with exact previous data. Reusing history also means more ways for an error to propagate. Test the recurrence at zero rate by inserting $y_n = \xi^n$:

$$\xi^2 - \xi = 0,$$
$$\xi = 1,\ 0.$$

One mode carries a constant initial offset and the other vanishes. More general multistep formulas require all such roots inside or on the unit circle and simple roots on the circle for zero stability. A starting method must still supply a sufficiently accurate y_1.

DERIVATION AND APPLICATIONS

Identify the physical and parasitic components of a two-step calculation

A two-step formula needs an extra starting value. Even a zero differential equation can reveal how an inaccurate start persists. Begin with a uniform grid and the second-order backward differentiation formula. Compute the unique quadratic interpolant through the most recent three values, differentiate at the newest point, and equate this derivative to the new slope. On the constant equation $y' = 0$ this gives

$$(3y_{n+1} - 4y_n + y_{n-1})/(2h) = 0.$$

To determine which patterns in the history survive, insert $y_n = r^n$ and cancel r^{n-1}. The polynomial and its two independent modes are

$$3r^2 - 4r + 1 = (r-1)(3r-1) = 0,$$
$$y_n = A + B(1/3)^n.$$

The unit root represents a constant exact solution; the second root is an unrelated decaying numerical mode. Suppose the exact initial value is Y, but the supplied second value is $Y + \delta$. At indices zero and one, solve for both constants:

$$A + B = Y,$$
$$A + B/3 = Y + \delta,$$
$$y_n = Y + \tfrac{3}{2}\delta[1 - (1/3)^n].$$

The transient numerical component dies out, but the startup error changes the persistent constant. This illustrates the root condition for zero stability: all roots must have magnitude at most one and those on the unit circle must be simple. It does not guarantee that an arbitrarily inaccurate initializer has the intended accuracy.

WORKED EXAMPLE

Track the effect of a faulty starting value

For an exact constant signal $Y = 2.000$ and a second reading 2.006, $\delta = 0.006$. The next value is

$$(4 \times 2.006 - 2.000)/3 = 2.008.$$

The limiting constant is 2.009. Reducing the time step while keeping the same fixed initializer bias leaves this limiting offset unchanged, regardless of the formal second-order truncation error.

SECOND APPLICATION

Compare an undamped parasitic root

The centred two-step formula $(y_{n+1} - y_{n-1})/(2h) = 0$ has roots $+1$ and -1. With the same initial pair, even values stay 2.000 and odd values 2.006. The sequence is bounded but alternates indefinitely. Averaging adjacent values gives 2.003 and conceals the oscillation in a plot. Checking the even and odd subsequences distinguishes a persistent computational mode from a physically varying solution.

Reuse information from earlier steps

A Runge–Kutta method constructs within-step stage values and can usually start from a single known initial state. A multistep method instead keeps a short history of states or slopes. Once the history exists, it can attain higher order with fewer new function evaluations per step. Startup, changes of step size and discontinuities need special care because the coefficients embody assumptions about where past data lie.

For a uniform step h, approximate f on $[t_n, t_{n+1}]$ by the line through f_{n-1} and f_n. In the forward interval the line is $f_n + s(f_n - f_{n-1})/h$, with $s = t - t_n$. Integrating from zero to h gives $y_{n+1} = y_n + h(3f_n - f_{n-1})/2$, the two-step Adams–Bashforth method. The negative weight results from extrapolating a straight line beyond its data. The formula integrates a linearly varying right-hand side exactly if the previous states and slopes are exact.

Differentiate a polynomial at the new point

A backward differentiation formula instead interpolates the recent state values. Differentiating the quadratic through y_{n-1}, y_n, y_{n+1} at t_{n+1} gives $(3y_{n+1} - 4y_n + y_{n-1})/(2h)$. Equating this to $f(t_{n+1}, y_{n+1})$ defines BDF2. To check the coefficients without constructing the full polynomial, substitute the constants, linear and quadratic monomials centred at t_{n+1}. Their derivatives at the evaluation point are $0, 1, 0$, respectively, which gives three linear conditions for the three weights.

For $y' = -y$ with step 0.2 and exact starts $y_0 = 1, y_1 = e^{-0.2} = 0.81873$, BDF2 gives

$$y_2 = (4y_1 - y_0)/(3 + 0.4) = 0.66910.$$

Adams–Bashforth gives

$$y_2 = y_1 + 0.1(-3y_1 + 1) = 0.67311.$$

The exact value at 0.4 is 0.67032. Exact starting data isolate the behaviour of the multistep formula. In a practical code, a single-step method of adequate order creates the first value. The starting error becomes an input disturbance for the recurrence.

Inspect zero stability before trusting the order

For $y' = 0$ with no new forcing, BDF2 obeys $3y_{n+1} - 4y_n + y_{n-1} = 0$. Try a sequence $y_n = \zeta^n$; the characteristic roots are $1, 1/3$. One root represents the constant physical solution, and the other is a decaying parasitic component. In a general constant-step linear multistep formula all characteristic roots must satisfy $|\zeta| \le 1$ and those on the unit circle must be simple for zero stability. A repeated root at one would admit a term proportional to n and hence unbounded amplification of a fixed initial mismatch as h decreases at fixed physical time.

Zero stability controls propagation over many small steps near $h = 0$. Absolute stability addresses a given finite $h\lambda$ for a dynamical test mode. A formula may satisfy the first and still require a restrictive

explicit step for the second. With consistency and a Lipschitz right-hand side, zero stability is the condition that converts small local defects and sufficiently accurate starts into convergence on fixed intervals. After a state reset or a discontinuity, reconstruct the history from the new side rather than extrapolating slopes belonging to a different evolution law.

Change the step without silently changing the order

If the preceding step had length h_p and the proposed step has length h, the slope of the line through the two old right-hand sides is $(f_n - f_{n-1})/h_p$. Integrating that line over the proposed interval gives $hf_n + h^2(f_n - f_{n-1})/(2h_p)$. The constant-step coefficients $3/2, -1/2$ are recovered only when $h = h_p$. For $h = h_p/2$, the new weights multiplying h are $5/4, -1/4$.

As a numerical check, integrate $y' = t$ starting at $t_n = 1$ with $h_p = 0.2,\ h = 0.1$. Slopes are 1 and 0.8, so the variable-step formula predicts an increment $0.1(1.25 - 0.25 \cdot 0.8) = 0.105$, exactly $\int_1^{1.1} t\,dt$. Reusing constant-step coefficients would give 0.11. Adjusting only the time interval in the integrator while leaving coefficients fixed can therefore introduce a measurable error even on linear forcing.

After an abrupt change of forcing

Suppose an input is switched from zero to one exactly at a recorded time. A two-step formula mixing a pre-switch derivative with the new value is extrapolating across a discontinuity. On the first step wholly after the switch, the true integral of the input is exactly h, whereas the old and new samples in Adams–Bashforth give $3h/2$. Restarting with a suitable one-step update reconstructs a history belonging to the new smooth segment.

An implicit formula is only as accurate as its correction solve

An implicit linear multistep method has an equation $y_{n+1} = a + h\beta f(t_{n+1}, y_{n+1})$, where a includes old states and old slopes. A predictor supplies a first iterate v_0. Functional correction uses $v_{j+1} = a + h\beta f(t_{n+1}, v_j)$. If f is Lipschitz in its state with constant L, differences contract whenever $h|\beta|L < 1$. Under that condition j corrections reduce an initial error by at least the corresponding geometric factor. This is a restriction on a particular inner solver and can be much more severe than the stability condition for the time formula itself.

For $f = \lambda y$, the fixed-point correction has error multiplier $h\beta\lambda$, even though solving the implicit time equation exactly produces a rational amplification factor. A single correction from an explicit predictor replaces that rational response by a polynomial. Accordingly, tables of stability regions for exactly solved formulas cannot be quoted unchanged when an implementation uses a fixed small number of functional corrections. Newton or a direct solve can restore the intended implicit update, provided its own arithmetic error is sufficiently controlled.

Compare a predictor, one correction and an exact implicit step

Consider $y' = -4y,\ y_n = 1,\ h = 1$. The trapezoidal implicit equation is $y_{n+1} = 1 - 2(1 + y_{n+1})$, giving $3y_{n+1} = -1$ and $y_{n+1} = -1/3$. For comparison the true value after the interval is $e^{-4} \approx 0.0183$. The exactly solved trapezoidal update remains bounded but oscillatory. Forward Euler prediction gives

$$v_0 = 1 - 4 = -3.$$

One functional trapezoidal correction yields $v_1 = -1 - 2(-3) = 5$, and another gives

$$v_2 = -1 - 10 = -11.$$

Relative to the implicit solution the correction errors are $-8/3,\ 16/3,\ -32/3$. Their absolute magnitudes double after every correction, precisely as predicted by $|h\lambda|/2 = 2$. The linear residual of the intended equation at the once-corrected value five is $3(5) + 1 = 16$. Reporting this as the result of an implicit stable solver would hide a far larger failure than the known phase error of the exactly solved trapezoidal rule. The inner iteration has failed to find the state required by the time method.

At $h = 0.1$ the correction multiplier instead becomes –0.2. The implicit update is $2/3$, while the explicit predictor is 0.6. One correction gives $0.8 - 0.2(0.6) = 0.68$, and another gives 0.664. Monitoring the actual implicit-equation residual now documents successful contraction. For nonlinear multistep implementations use the same principle: record whether each implicit equation was solved to accuracy compatible with the local time error before interpreting convergence under time refinement.

One can also solve the scalar implicit equation by Newton iteration. Here its derivative with respect to the unknown is the constant three at $h = 1$, so one Newton correction from any starting guess recovers $-1/3$ exactly. The comparison demonstrates a failure of a particular nonlinear iteration, while the mathematical time formula itself is well defined and has a unique implicit solution. More generally stiffness makes the choice of inner solver central to total runtime and reliability.

Change parameters · Explore the worked calculation

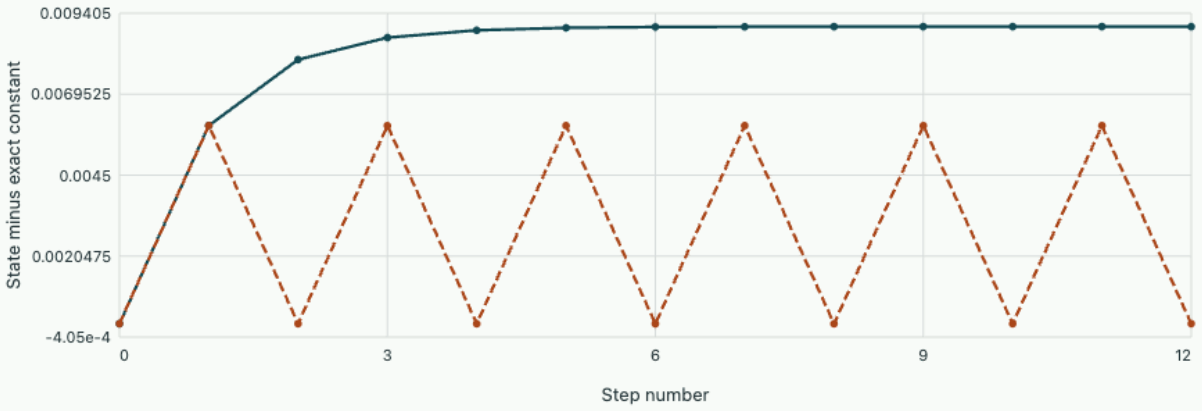
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Second-start error δ **0.006**

Updates **12**



BDF2 roots 1 and 1/3 Centred roots 1 and −1

BDF2 damps its parasitic component but retains a limiting bias 3δ/2 = 0.009. The centred method retains the even/odd oscillation.

Study the full derivation

A delayed feedback model needs a history, not one starting value · Dynamics and numerical time integration

A controller responds to the measured displacement one time delay ago. Construct its solution in successive intervals, track propagated derivative discontinuities and derive the point at which increasing delay can destabilise negative feedback.

Problem sheet 2 · Damping and multistep memory

Locate a stability region, compare two implicit formulas in the stiff limit and test whether a recurrence transports initial errors consistently.

Separate positivity from decay

A mode obeys $y' = -4y$. Starting at one, find the first two forward Euler values for steps 0.2, 0.35 and 0.55. Classify the histories by sign and decay, and find the exact positive-step threshold for each property.

▼ **Worked solution**

The three factors $1 - 4h$ are 0.2, −0.4 and −1.2. The first two values are respectively $(0.2, 0.04)$, $(-0.4, 0.16)$ and $(-1.2, 1.44)$. The first sequence is nonnegative and decreasing, the second sign-changing with decreasing magnitude, and the third sign-changing with growing magnitude.

For positive steps, nonnegative propagation requires $h \le 0.25$; decay of magnitudes requires $h < 0.5$. At 0.5 the amplitude remains one while the sign alternates.

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.

Switch off a stiff transient

At a proposed step, a fast mode has $z = -50$. Compute its factors under backward Euler and the trapezoidal method, and the magnitude after two steps. Which would you use when the desired late-time response should no longer contain the fast mode?

▼ **Worked solution**

Backward Euler gives $1/51 = 0.01961$, hence $1/2601 = 0.0003845$ after two steps. Trapezoidal gives $(1 - 25)/(1 + 25) = -12/13$, hence $144/169 = 0.8521$ in magnitude after two steps. The physical magnitude is e^{-100}.

Backward Euler suppresses this unresolved transient much more strongly. If the early transient itself were the observable, reducing the step would remain necessary even with the more strongly damping formula.

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.

Diagnose a parasitic solution

For $y' = 0$, a proposed two-step recurrence gives

$$y_{n+2} - 2y_{n+1} + y_n = 0.$$

Start at $y_0 = 1, y_1 = 1 + \epsilon$. Obtain y_n and compare with BDF2 from the same values. Which can bound the propagated starting error independently of step count?

▼ **Worked solution**

The characteristic polynomial is $(z-1)^2$. Solving for the constants in $a + bn$ gives

$$y_n = 1 + n\epsilon;$$

the starting error grows linearly with step count.

BDF2 gives

$$y_n = a + b(1/3)^n.$$

The initial conditions yield

$$b = -3\epsilon/2,$$
$$a = 1 + 3\epsilon/2.$$

Its departure from one stays bounded by $1.5|\epsilon|$ for all n. Two simple roots behave differently from a repeated unit root even when all root magnitudes are at most one.

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

Chapter 5 · Adaptive steps and event locations

Estimate local integration error with a low–high order pair, choose a dimensionally consistent tolerance and bracket a zero of an event function in time.

FROM FIRST PRINCIPLES

Obtain a step controller from the measured error of two approximations

Take a proposed step h twice from the same current state, once using an order-p formula and once using a more accurate formula whose leading defect is one order smaller. For smooth data and sufficiently short steps the two values have a common exact target y_*. Their expansion gives

$$y_{\text{low}} = y_* + Ch^{p+1} + O(h^{p+2}),$$
$$y_{\text{high}} = y_* + O(h^{p+2}),$$
$$d = y_{\text{low}} - y_{\text{high}}.$$

The difference estimates the leading local error of the lower-order formula. It has the units of the state. Scale each component by an absolute tolerance with those units plus a dimensionless relative tolerance times the current magnitude:

$$r = \max_i \frac{|d_i|}{\text{atol}_i + \text{rtol}\,\max(|y_{n,i}|, |y_{\text{high},i}|)}.$$

If $r \le 1$ the proposed step passes this error estimate. Provided its leading coefficient varies slowly, a new step should change the estimated error by the ratio of step sizes raised to $p+1$. Targeting unit scaled error gives

$$h_{\text{new}} = s\,h\,r^{-1/(p+1)}, \qquad 0 < s < 1.$$

The safety factor leaves room for variation in the neglected terms. A real implementation bounds growth and handles $r = 0$ separately. A discontinuity or an event between endpoints can violate the smooth-error assumption even when the endpoints appear accurate.

Background · Embedded error estimate

$$e = y_{n+1}^{[p]} - y_{n+1}^{[p-1]}$$

Two formulas share stages and estimate a local error. The controller also needs absolute and relative scales for each component.

DERIVATION AND APPLICATIONS

How component tolerances become a single step decision

A time step acceptable for one component can be inaccurate for another whose magnitude or units differ. Consider an embedded pair of methods sharing their internal stages and producing two endpoint estimates y_i^H, y_i^L. Let $d_i = y_i^H - y_i^L$. For a smooth right-hand side and a pair whose lower formula has order p, this difference is normally dominated by an h^{p+1} local term. Assume all components use the same step but have physically appropriate absolute scales $a_i > 0$ and relative tolerances r_i.

Convert each difference to a dimensionless ratio using the larger recent endpoint magnitude. An infinity norm requires every component to meet its own allowance:

$$s_i = a_i + r_i \max(|y_{i,n}|, |y_i^H|),$$
$$E = \max_i |d_i|/s_i.$$

With approximately constant leading error coefficient and scale over the proposed retry, the new norm is predicted from the power law. Solving for a target near one and including a safety factor $\nu < 1$ gives

$$E_{\mathrm{new}} \simeq E(h_{\mathrm{new}}/h)^{p+1},$$
$$h_{\mathrm{new}} = \nu h E^{-1/(p+1)}.$$

Accept and advance with the higher-order estimate only when $E \le 1$ under this policy. If a component is converted to another unit by a factor c, both its measured difference and absolute tolerance must scale:

$$|c d_i|/(|c|a_i + r_i|c| \max |y_i|) = |d_i|/s_i.$$

The decision is therefore invariant under consistent unit conversion. A bounded growth factor handles zero estimated differences and prevents a single unusually small defect from triggering a huge next step. This controller targets local agreement; its acceptance is not a proof of a global error bound.

WORKED EXAMPLE

Decide a two-component retry

Let an order-two embedded lower estimate at $h = 0.100$ s produce differences 0.000800 m and 0.00600 m/s. The chosen component scales are 0.00100 m and 0.00200 m/s. Their ratios are 0.800 and 3.00. Reject the step and take $\nu = 0.9$, giving $h_{\mathrm{new}} = 0.9(0.100)3^{-1/3} = 0.0624$ s. The leading-error model predicts a new norm $0.9^3 = 0.729$.

SECOND APPLICATION

Compare a componentwise limit with an averaged one

Suppose a system has 100 components, with 99 normalized errors zero and one equal to 3.00. Its root-mean-square norm is $\sqrt{9/100} = 0.300$, so an RMS policy would accept. The infinity policy rejects because one variable missed its tolerance threefold. An RMS policy can be appropriate when errors are distributed, while a strict component requirement demands the maximum or a separate check for a critical variable.

Ask different parts of a trajectory for different resolutions

A cooling transient changes rapidly at first and slowly near equilibrium. A uniform step small enough for the beginning can waste effort later. An adaptive controller proposes a step, estimates its error and accepts or rejects it. To estimate the error without the exact solution, two formulas can share function evaluations. With Euler and Heun, compute $k_1 = f(t_n, y_n)$, predict $y_E = y_n + hk_1$, then evaluate $k_2 = f(t_n + h, y_E)$ and $y_H = y_n + h(k_1 + k_2)/2$.

Their difference $y_H - y_E = h(k_2 - k_1)/2$ estimates the leading local error of the lower-order Euler step. It normally scales as h^2 in smooth regions. An embedded difference does not equal the local error of the higher-order accepted value; the higher-order value generally has a smaller error starting from accurate data. Nor does a local tolerance directly bound the error at the end of the whole run. Previous accepted steps already carry accumulated trajectory error.

Scale the estimate and choose a step

For a scalar use $w = \mathrm{atol} + \mathrm{rtol}\max(|y_n|, |y_H|)$ and the dimensionless estimate $E = |y_H - y_E|/w$. Accept when $E \le 1$. If the next difference scales as h^2, a proposed change is $h_{new} = 0.9hE^{-1/2}$. The factor 0.9 leaves a margin for changes in curvature. Practical implementations limit the ratio between successive steps and handle a zero difference by bounded growth. A rejected step is retried from the same original state and time.

For $y' = -2y,\ y_n = 1,\ h = 0.1$, Euler gives 0.8 and Heun gives 0.82, a difference 0.02. With absolute tolerance 0.005 and zero relative tolerance for this demonstration, $E = 4$. The step is rejected and a revised step is

$$0.9(0.1)/2 = 0.045.$$

Re-evaluation from the original state gives

$$y_E = 0.91,\ y_H = 0.91405,$$

with difference 0.00405 and $E = 0.81$. The accepted state corresponds to time 0.045, not 0.145. Advancing time on rejection would skip the unresolved portion.

Background · What tolerance should a solver apply near zero concentration?

Two state components can be reported in very different units. An error of 10^{-6} metres may be harmless for a position and an error of 10^{-6} moles may dominate a trace species. A common error test divides each component by its own physical absolute tolerance plus a relative contribution, for instance

$$e_i^{\mathrm{scaled}} = \frac{e_i}{a_i + r \max(|y_{n,i}|, |y_{n+1,i}|)}.$$

At a zero crossing, $a_i > 0$ keeps the requested absolute accuracy finite. If concentration units change from mol/L to mmol/L, multiply the matching absolute tolerance by 1000 along with the numerical concentrations. The relative tolerance is dimensionless. Leaving the old absolute tolerance in place silently imposes a thousandfold tighter accuracy requirement near zero.

Estimate error using one full step and two half steps

A solver without an embedded pair can estimate its local error by applying the same order-p stepper twice from a common starting state: take one full step for Y_c, and separately two successive half steps for Y_f. Suppose the local error for one smooth step, numerical minus exact, is $Ch^{p+1} + O(h^{p+2})$. The two half-step errors accumulate to $2C(h/2)^{p+1} + O(h^{p+2})$ because propagating the first over the short second half changes its leading coefficient only at higher order. Thus Y_f carries approximately 2^{-p} of the full-step error.

Subtracting the two approximations gives an estimate of the correction to Y_f: $(Y_f - Y_c)/(2^p - 1)$. Adding it gives a Richardson-extrapolated value. The exponent in a local step controller remains $1/(p+1)$ because the correction estimate scales as h^{p+1}. The denominator $2^p - 1$ comes from taking *two* shorter steps to the same endpoint. Confusing that denominator with a single half-step local error factor introduces a systematic error in the estimate.

Take Heun's second-order method on $y' = -y$, starting from one with $h = 0.4$. The full-step amplification is

$$1 - 0.4 + 0.4^2/2 = 0.68.$$

Each half step has amplification 0.82, so $Y_f = 0.6724$. The fine correction is

$$(0.6724 - 0.68)/3 \approx -0.0025333.$$

The extrapolated value is 0.6698667. The exact endpoint $e^{-0.4} \approx 0.6703200$ requires correction -0.0020800 to Y_f. The estimated sign and scale are correct while higher-order terms are still visible. Discontinuous forcing, a skipped event or inconsistent internal solver tolerances can spoil this comparison even when the two output times agree.

Practise on sheet 3 · Accept or retry a nonautonomous Euler step. Further reading · University of British Columbia · Numerical Solution of Ordinary Differential Equations, Joel Feldman · Variable Step Size Methods, Euler and Euler–2step, pp. 1–4.

Locate events between stored states

An event is a zero of $g(t, y(t))$, such as crossing a chosen temperature threshold. When its endpoint values have opposite signs, continuity brackets at least one event. Use a dense approximation to the trajectory or refine the time interval to locate the crossing. A linear interpolation between endpoints may be adequate for a smooth signal away from a turning point. If a value changes by a small amount near a flat crossing, a small field error can produce a much larger error in event time.

For $y = e^{-t}$, readings at 0.6 and 0.8 are 0.54881 and 0.44933. A straight interpolation to 0.5 gives $t \simeq 0.69813$, while the exact time is $\ln 2 = 0.69315$. The discrepancy arises from the curved trajectory even if the endpoints are exact. A tangential event such as $(t-1)^2 = 0$ never changes sign, and two crossings in one step can also be missed. Declare the crossing direction, the required timing accuracy and the physical effect of reaching it. After an event that changes the model, start the next segment using the event state and preserve the time at which the transition actually occurred.

Test the event interpolant with a known direction

At a crossing where a scalar signal has nonzero derivative, an approximate amplitude error δy produces a first-order timing error $\delta t \simeq -\delta y / y'$. For $e^{-t} = 0.5$, the derivative is -0.5 per second. A signal that is 0.002 too high near the crossing therefore appears to reach the threshold about 0.004 s late. This local estimate converts a solver tolerance on a state into a physically interpretable timing requirement.

A source that switches precisely at the threshold makes the event time part of the subsequent model. A timing error then affects all later states even if ordinary integration error within each regime is small. When comparing grids, compare both the located event times and the post-event states at shared physical times. Two traces that agree before a switch can diverge later purely because one entered the new law sooner.

Relate state accuracy to event-time accuracy

Let an event be defined by $G(t, y(t)) = 0$ and write the scalar event signal as $g(t)$. Its derivative along a smooth trajectory is $g' = G_t + G_y f$. At a simple crossing $g'(t_\star) \neq 0$, an error δg in evaluating the signal displaces the reported time by approximately $-\delta g / g'(t_\star)$. More rigorously, if the true signal derivative has magnitude at least m throughout the final bracket, the mean value theorem bounds time error by the event-signal error divided by m, in addition to any error left by the scalar root search.

Thus an event locator can solve its interpolated scalar equation to a tiny tolerance while returning a poor physical time. Errors in the interpolated state, uncertain event threshold and flat slope all propagate into the time estimate. At a tangential contact the derivative vanishes and the first-order estimate ceases to

apply. Two simple events can merge into one contact or disappear altogether under a perturbation that would merely shift a steep isolated crossing by a small amount.

Time two crossings around the top of a trajectory

Take vertical position $y(t) = 10t - 5t^2$ metres, with time in seconds. Its maximum is 5 m at one second. A detector at 4.8 m records two crossings, since $5 - 5(t-1)^2 = 4.8$ gives $t = 1 \pm 0.2$, namely 0.8 and 1.2 s. The respective vertical velocities are $+2$ and -2 m/s. A small threshold increase of 0.01 m therefore moves the rising crossing about 0.005 s later and the falling crossing about 0.005 s earlier.

Check the exact effect at 4.81 m. The half-separation becomes

$$\sqrt{0.19/5} = \sqrt{0.038} \approx 0.194936\,\mathrm{s}.$$

The actual time shifts have magnitude approximately 0.005064 s, so the local slope estimate is already informative. At a threshold of 4.999 m the speed is instead only $10\sqrt{0.001/5} \approx 0.14142$ m/s. Another increase of 0.01 m moves the threshold above the maximum and removes both crossings. Applying the previous linear timing estimate at this near-tangent setting would predict a displacement for an event that no longer exists.

A direction flag distinguishes the rising and falling events if each is bracketed, but it cannot recover two crossings hidden inside one large step with endpoint signals of the same sign. In this model the velocity event $y' = 0$ identifies the turning point and splits the time interval into monotone pieces. On each piece check the threshold against its attainable position range before starting a time root search. This strategy uses knowledge of the dynamics to ensure the event actually exists.

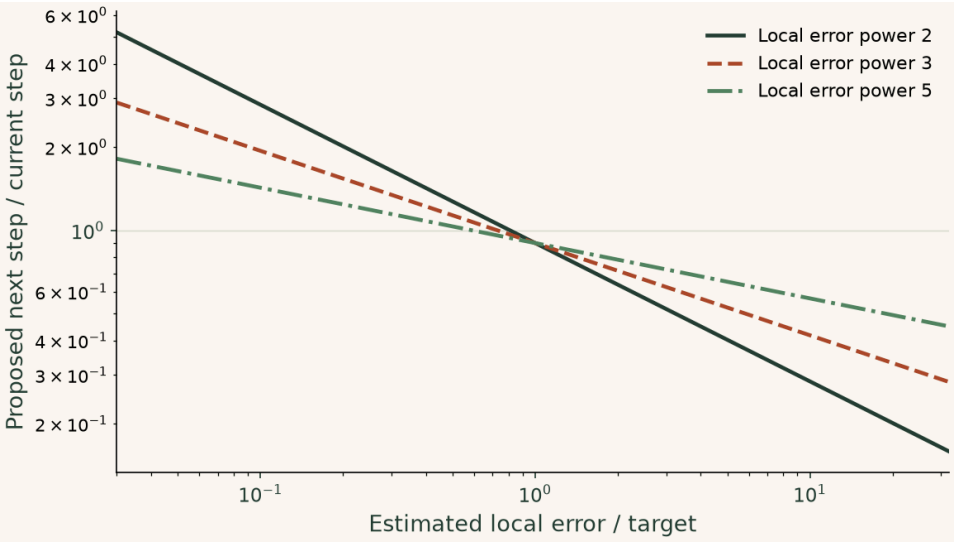
For example, a single step from 0.7 to 1.3 s gives position 4.55 m at both endpoints and therefore no sign difference relative to the 4.8 m detector. Inserting the peak at 1 s gives a positive signal there and immediately produces two valid brackets. Decreasing only the scalar root tolerance would have found neither event on the original unsplit step.

Choose the next step when a local error estimate misses its target

Suppose the local estimator behaves as $E \simeq Ch^p$. Matching the next error to tolerance by rescaling h, then applying a safety factor 0.9, predicts

$$h_{\mathrm{next}}/h = 0.9(\mathrm{tol}/E)^{1/p}.$$

Treat p as the power of this estimator.



With E eight times the target, a cubic estimator requests a new step 0.45 times the old one. The formula proposes a length; an acceptance test must separately decide whether to retry the current step. Sudden events can invalidate the assumed scaling.

▶ **Read values from the plot**

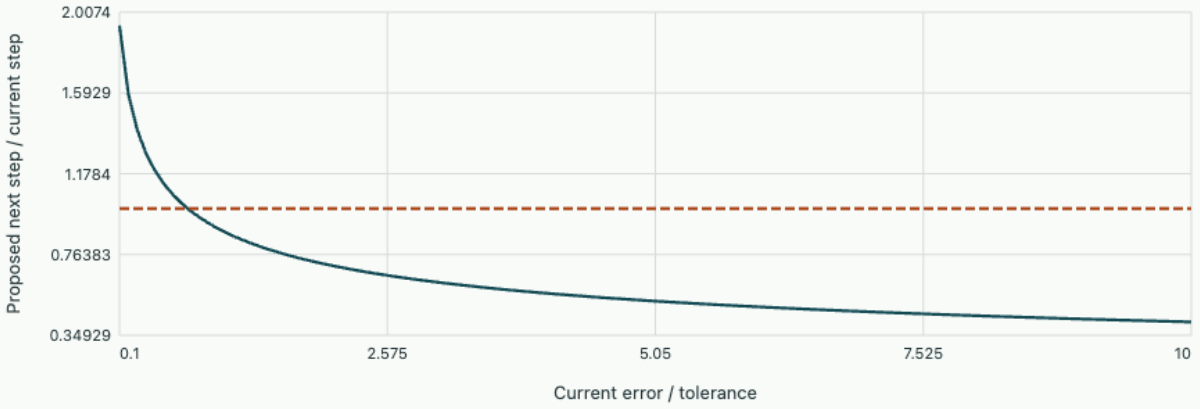
Change parameters · Choose the next step when a local error estimate misses its target

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

Estimator power p	3
Safety factor	0.9
Selected error / tolerance	8



Power-law step proposal No step change

Selected ratio requests hnext/h = 0.45. Accept/reject the current step separately; this rule assumes the estimator scales as h^3.

Study the full derivation

Find a crossing between accepted time steps · Dynamics and numerical time integration

An integrator can approximate a trajectory accurately yet report a poor switch time. Construct a continuous interpolant, relate a state error to an event-time error and inspect what happens when a trajectory merely touches a threshold.

Chapter 6 · Energy and geometric integration

Compare three time updates on a conservative oscillator and explain why a method can have small short-term state error yet distort an orbit over many periods.

FROM FIRST PRINCIPLES

Build an oscillator update out of two area-preserving shears

Take one undamped oscillator of mass $m > 0$ and stiffness $k > 0$. Its canonical position q and momentum p satisfy

$$\dot{q} = p/m,$$
$$\dot{p} = -kq.$$

Exact trajectories run along energy ellipses and preserve phase-plane area. If both slopes are taken at the left endpoint, the update and its area factor are

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & h/m \\ -hk & 1 \end{pmatrix} \begin{pmatrix} q_n \\ p_n \end{pmatrix}, \qquad \det = 1 + h^2 k/m.$$

A positive factor greater than one at each step makes the numerical orbit spiral outward. Instead update the momentum at fixed position, then move the position using that new momentum. Both operations are shears:

$$p_{n+1} = p_n - hkq_n,$$
$$q_{n+1} = q_n + hp_{n+1}/m.$$

The product matrix has diagonal entries $1 - h^2k/m$ and 1, with off-diagonal product $-h^2k/m$, giving determinant one. Its eigenvalues form a conjugate pair on the unit circle when $|2 - h^2k/m| < 2$. This calculation explains the bounded oscillatory behaviour for $0 < h\sqrt{k/m} < 2$. Preserving area alone would also allow one real mode to grow while another shrinks. It does not establish either stability or exact conservation of the original energy.

Background · Symplectic map

$$(D\Phi)^{\mathsf{T}} J (D\Phi) = J$$

This is preservation of the canonical symplectic form. It provides control of long-time structure without guaranteeing that every trajectory has the exact phase.

DERIVATION AND APPLICATIONS

Which quadratic energy does symplectic Euler preserve?

An undamped linear oscillator conserves the physical sum of kinetic and potential energy. A geometric time integrator may conserve a nearby quadratic form instead. Take $q'' = -\omega^2 q$ with constant ω and unit mass. Scale velocity by setting $u = \dot{q}/\omega$ so that both variables have units of length, and write $z = h\omega$. Use the kick-then-drift version of symplectic Euler. Its step is two triangular shears:

$$u' = u - zq,$$
$$q' = q + zu' = (1 - z^2)q + zu.$$

To find a preserved quantity, test a quadratic form $I = q^2 + u^2 + \alpha qu$. The changes of the ordinary sum and the product follow directly by multiplication and collection:

$$q'^2 + u'^2 - q^2 - u^2 = z(q'u' - qu),$$
$$I' - I = (z + \alpha)(q'u' - qu).$$

Choosing $\alpha = -z$ cancels the entire change at every step. Completing the square identifies the range where the invariant confines trajectories to closed ellipses:

$$I = q^2 + u^2 - zqu = (u - zq/2)^2 + (1 - z^2/4)q^2.$$

For $|z| < 2$ the form is positive definite. The physical energy is proportional to $q^2 + u^2$ and can oscillate, but the invariant bounds it. As $h \to 0$ the tilted ellipse approaches the exact energy circle. At the endpoint $|z| = 2$ the quadratic form ceases to bound both variables; preservation of an indefinite form alone does not guarantee stable trajectories.

WORKED EXAMPLE

Follow a moderate step exactly

Take $\omega = 2.00$ s⁻¹, $h = 0.250$ s, $q = 1.00$ m and $u = 0$. Then $z = 0.500$, $u' = -0.500$ m and $q' = 0.750$ m. The ordinary sum falls from 1.00 to 0.8125 m². But $I' = 0.8125 - 0.500(0.750)(-0.500) = 1.00$ m². A plot restricted to physical energy at one time could mistake the discrepancy for damping.

SECOND APPLICATION

Diagnose a seemingly exact conservation test outside stability

Choose $z = 2.50$ and the same initial $(q, u) = (1, 0)$. One step gives $(-5.25, -2.50)$. The ordinary sum is 33.8125, whereas the cross term is $2.50(13.125) = 32.8125$, so $I' = 1$. The invariant remains exact while the amplitude grows. Testing positivity of the conserved quadratic form therefore supplies information absent from a conservation-only check.

Choose the diagnostic before choosing the method

A mass on an ideal spring with unit mass and stiffness obeys $\dot{q} = p,\ \dot{p} = -q$. The state moves on circles with conserved energy $H = (q^2 + p^2)/2$. At short times a state norm measures trajectory error, but over many periods there are independent questions about amplitude, phase and energy. A time-stepping method that systematically increases the orbit radius can produce an artificial runaway even on this simplest stable physical system.

Forward Euler updates $q_{n+1} = q_n + hp_n,\ p_{n+1} = p_n - hq_n$. Square and add: the cross terms cancel, leaving $H_{n+1} = (1 + h^2)H_n$. The phase space area factor of the update matrix is its determinant $1 + h^2$, also greater than one. Backward Euler instead contracts energy by $1/(1 + h^2)$. The second method is stable in the sense of bounded states, but it removes mechanical energy from an undamped model. The sign of drift explains much more here than a solitary endpoint comparison.

Update momentum before position

The momentum-first symplectic Euler scheme sets $p_{n+1} = p_n - hq_n$ and $q_{n+1} = q_n + hp_{n+1}$. Its update matrix in (q, p) has rows $(1 - h^2, h)$ and $(-h, 1)$. Its determinant is one and trace is $2 - h^2$. For $0 < h < 2$, the two eigenvalues are a complex conjugate pair of unit magnitude because their product is one and the trace lies strictly between –2 and 2. The resulting numerical frequency satisfies

$$2\cos\theta = 2 - h^2.$$

Thus phase error still accumulates even if the amplitudes remain bounded.

Direct substitution also verifies an exact discrete invariant $\tilde{H} = (q^2 + p^2 - hqp)/2$. For $h < 2$ the corresponding quadratic form is positive definite, so constant $\tilde{H}$ confines the state to an ellipse. The exact physical energy can fluctuate around its initial level as the point moves around that ellipse. Area preservation alone would not guarantee this behaviour: for $h > 2$ the determinant remains one but real reciprocal eigenvalues permit growth. The eigenvalue and invariant calculations supply the missing qualification.

Check a long-time forecast with numbers

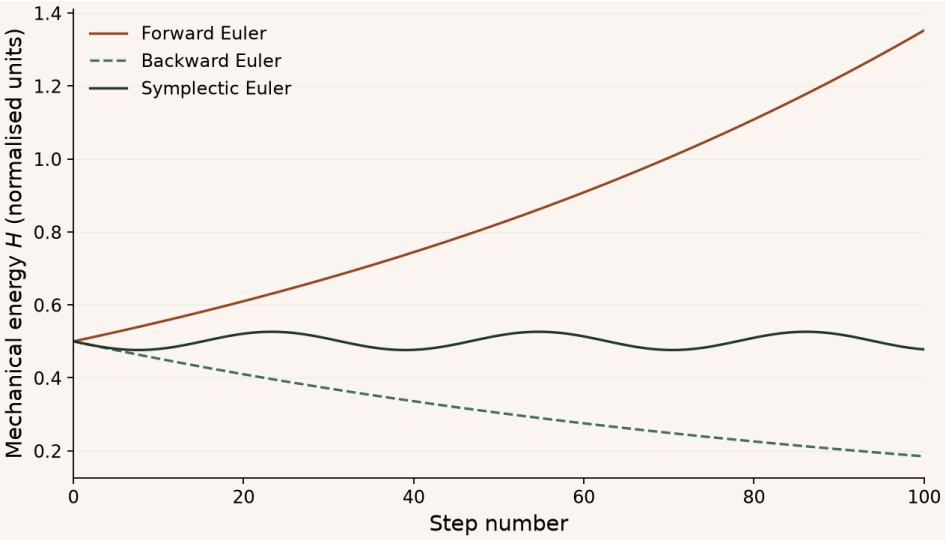
Start at $(q, p) = (1, 0)$ with $h = 0.1$. Forward Euler gives $(1, -0.1)$ and $H = 0.505$ after one step. After 100 steps the exact multiplicative law predicts

$$H = 0.5(1.01)^{100} = 1.3524.$$

Symplectic Euler instead gives $(0.99, -0.1)$, $H = 0.49505$, but $\tilde{H} = 0.5$ again. The first step alone might suggest that the second scheme dissipates energy. The invariant predicts that the decrease cannot continue monotonically over many periods.

Implicit midpoint, $(y_{n+1} - y_n)/h = J(y_n + y_{n+1})/2$ with the skew-symmetric oscillator matrix J, preserves the exact quadratic energy. Dot both sides with the midpoint state. The left side is

proportional to $H_{n+1} - H_n$ and the right side is zero by skew symmetry. Its rotations need not accumulate the exact physical phase, since the rotation angle per step is $2\arctan(h/2)$. A meaningful long-time study therefore records measured period, energy range and reversibility alongside any state error after a specified number of cycles.



An ideal oscillator sampled with step 0.1. Checking only the first few updates obscures the systematic growth or decay visible after 100 steps.

Test an algorithm by reversing the clock

A method is time symmetric if taking a step of length h and then a step of length $-h$ restores the starting state, up to arithmetic and solve tolerance. For the midpoint update, exchanging the old and new values leaves their average unchanged and changes both sides of the difference equation consistently. This gives exact reversibility for an accurately solved implicit step. For forward Euler on the oscillator, the product of the forward and backward matrices is $(1 + h^2)I$, so reversal actually scales the state.

At $h = 0.1$, the simple reversal from $(1, 0)$ therefore returns $(1.01, 0)$ for forward Euler, making the failure visible in one pair of steps. This diagnostic is useful only for a reversible underlying evolution paired with its time-reversed integration law. A dissipative or dissipatively constrained physical system has different requirements. Reversibility, energy preservation and phase accuracy probe separate mathematical properties, each with a different operational test.

Test a flow map on a small cell of phase space

For canonical coordinates $z = (q, p)$, a differentiable Hamiltonian update $z_+ = \Psi(z)$ is symplectic when its Jacobian $D\Psi$ preserves the matrix $J = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ through $D\Psi^T J D\Psi = J$. Direct multiplication shows that for any real two-by-two matrix M, $M^T J M = (\det M) J$. Thus in one degree of freedom symplecticity is equivalent to preserving oriented area. A small rectangle of nearby initial conditions should map to a region of the same signed area even when its sides stretch or tilt.

For more than one degree of freedom, preserving ordinary volume is necessary but no longer sufficient. A map may stretch one canonical (q, p) plane and compress another by the reciprocal factor. Its full determinant remains one while the area pairing in each plane changes. Energy is a further, independent diagnostic: an area-preserving map can move individual points among nearby energy contours. The determinant calculation tests a geometric property of the map across neighbouring trajectories, rather than just the energy at one chosen state.

Compare the phase volume of an oscillator ensemble

For $q' = p,\ p' = -q$ with unit frequency, forward Euler gives $M = \left(\begin{smallmatrix} 1 & h \\ -h & 1 \end{smallmatrix}\right)$, of determinant $1 + h^2$. Backward Euler gives $(I - hA)^{-1}$, where A has off-diagonal entries $1, -1$, so its determinant is $1/(1 + h^2)$. A kick followed by a drift gives the symplectic Euler matrix with rows $(1 - h^2, h)$ and $(-h, 1)$, whose determinant is exactly one. These facts hold for every input rectangle, so they can be checked by integrating its corners rather than relying on a single solution.

Choose $h = 0.2$ and an initial ensemble occupying a rectangle of area 0.01 in normalised phase coordinates. After 25 steps forward Euler predicts area $0.01(1.04)^{25} \approx 0.02666$. Backward Euler gives $0.01/(1.04)^{25} \approx 0.00375$. Symplectic Euler preserves 0.01. Repeated steps compound even a four percent per-step area change into a substantial distortion. For the oscillator all three approximate individual paths can initially look plausible, so this ensemble diagnostic reveals information inaccessible in a short plot of one coordinate.

As a counterexample in four dimensions, scale the first pair of coordinates by $(2, 1)$ and the second pair by $(1/2, 1)$. The volume multiplier is one, but the first canonical area doubles while the second halves. A full symplecticity test must check $M^T J M - J$ using the correctly ordered canonical coordinate pairs. An algorithm preserving volume in an arbitrary collection of engineering state units is not automatically qualified for long-time Hamiltonian integration.

Area preservation also survives composing individual canonical shears: the determinants multiply and every shear has determinant one. This explains why reversing the order of a kick and a drift changes the phase error but preserves the two-dimensional area test. The two compositions may therefore produce different numeric positions while agreeing precisely on one geometric invariant. A useful verification suite includes both ensemble area and a trajectory observable with a known time dependence.

Change parameters · Energy and geometric integration

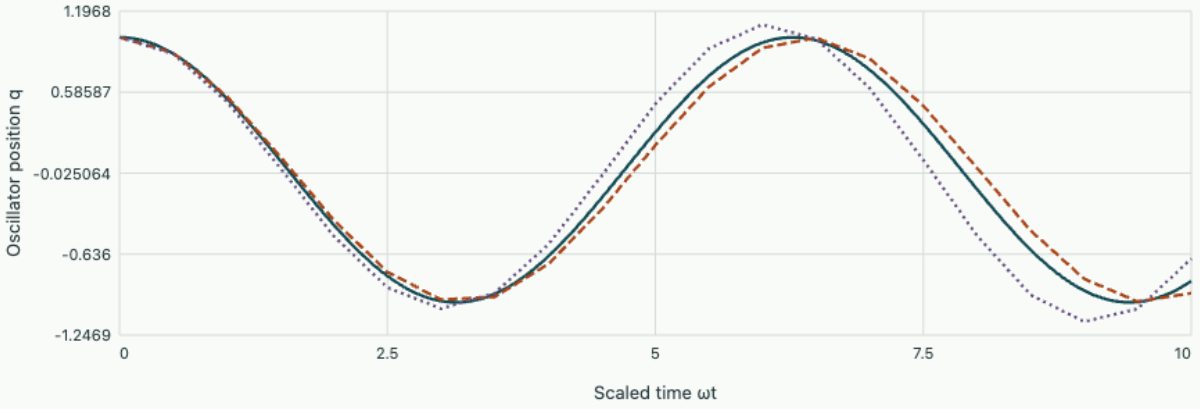
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Time step h (ω = 1) **0.5**

Number of forward steps **20**



Implicit midpoint phase lag 0.2009 rad. Forward-backward distance from the start is 8.951e-16 for implicit midpoint and 0.3635 for explicit midpoint.

Problem sheet 3 · Events and long trajectories

Adapt a time step to a tolerance, analyse a grazing event and estimate phase drift on an energy-preserving orbit.

Reject and retry

A controller compares an Euler–Heun difference of 0.018 m with a scale $0.002 + 0.001 \max(|y_n|, |y_H|)$ m. The larger state magnitude is 2 m and the attempted step is 0.3 s. Determine acceptance, the next proposed step with safety factor 0.9, and the expected new error ratio if curvature stays constant.

▼ **Worked solution**

The scale is 0.004 m, so $E = 4.5$, which requires rejection. The proposal is

$$0.3(0.9)/\sqrt{4.5} = 0.1273\,\mathrm{s}.$$

Since the difference scales quadratically, the expected new ratio is $4.5(0.1273/0.3)^2 \simeq 0.81$, assuming the scale stays unchanged.

Recompute both stages from the state and time at the start of the rejected trial. The accepted time remains unchanged until a valid trial has finished.

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.

Bound timing uncertainty

A threshold measurement has absolute state error at most 0.004 and is crossed with speed of magnitude at least 0.2 per second throughout a verified bracket. Bound the time uncertainty due solely to state error. Now compare with a tangential zero $g(t) = (t - 1)^2$ and determine when an apparent residual of 0.004 can occur.

▼ **Worked solution**

The mean value theorem applied to the monotone crossing gives time uncertainty at most $0.004/0.2 = 0.02$ s. Interpolation or time-discretisation errors need their own contributions if they are not already included in the state error.

At the tangency, $g = 0.004$ occurs at $1 \pm \sqrt{0.004}$, about 0.0632 s from the exact event. The event derivative vanishes at the root and the signal has the same sign on both sides, so a standard sign-change trigger can miss it.

Explore the related calculation with adjustable inputs. Decrease the crossing speed in the threshold example and inspect the growth of timing sensitivity.

Check energy separately from phase

Implicit midpoint on the unit oscillator rotates by $\theta = 2\arctan(h/2)$ per step. With $h = 0.2$, evaluate the phase error after 100 steps and compare the exact and numerical energies if both start at $H = 0.5$.

▼ **Worked solution**

The step angle is $2\arctan(0.1) = 0.1993373$ rad, compared with the physical 0.2 rad. After 100 steps the numerical solution lags by approximately 0.06627 rad.

Both energies are 0.5 in exact arithmetic because implicit midpoint preserves this quadratic invariant. Comparing energy histories alone therefore hides a positional error that continues to accumulate with elapsed time.

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.

Accept or retry a nonautonomous Euler step

Use Euler for $y' = t + y,\ y(0) = 1$ over $h = 0.4$. Compare a full step with two half steps, estimate the correction of the finer answer and its extrapolated value. For absolute local tolerance 0.02 and safety factor 0.9, determine the proposed retry step.

▼ **Worked solution**

The full step gives $1 + 0.4 = 1.4$. The first half step gives 1.2 at time 0.2. Its new slope is $0.2 + 1.2 = 1.4$ and the second half step yields

$$1.2 + 0.2(1.4) = 1.48.$$

With $p = 1$ the correction estimate is 0.08, so the extrapolated value is 1.56. The exact endpoint $2e^{0.4} - 1.4 \approx 1.58365$ shows the size of the remaining higher-order contribution.

The estimated fine error magnitude exceeds 0.02. The proposed step is $0.9(0.4)(0.02/0.08)^{1/2} = 0.18$, retried from time zero. Using a slope evaluated at time zero for the second half would skip the explicit time dependence and invalidate the estimate.

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

Revisit chapter 5 · Estimate error using one full step and two half steps.

Chapter 7 · Variation of constants and exponential steps

Integrate a stiff linear decay exactly while approximating only its slowly changing forcing term.

FROM FIRST PRINCIPLES

Integrate the linear part exactly before approximating the source

Suppose $y' = Ay + g(t, y)$, with a constant matrix A. It may carry a large decay rate even though the applied source g varies smoothly. Multiply by e^{-At} and use the product rule. The derivative of the exponential cancels the linear term:

$$\frac{d}{dt}(e^{-At}y(t)) = e^{-At}g(t, y(t)).$$

Integrate from t_n to $t_n + h$ and multiply back by the final exponential. The exact formula assigns each source contribution the propagation corresponding to its remaining time in the step:

$$y(t_n + h) = e^{hA}y(t_n) + \int_0^h e^{(h-s)A}g(t_n + s, y(t_n + s))\, ds.$$

Approximate only the source by its starting value. Integrating the exponential weight yields

$$y_{n+1} = e^{hA}y_n + h\varphi_1(hA)g(t_n, y_n),$$
$$\varphi_1(z) = \sum_{j=0}^{\infty} z^j/(j+1)!.$$

For nonzero scalar z the series equals $(e^z - 1)/z$; at zero it equals one. The series remains meaningful for a singular matrix. A constant source is integrated exactly by this step. Variation in the source, including its dependence on the moving state, determines its remaining time-discretisation error.

DERIVATION AND APPLICATIONS

Evaluate exponential weights without subtracting almost equal numbers

The integrating-factor solution of $y' = \lambda y + g(t)$ gives useful one-step formulas when λ is constant and the source changes slowly. Approximate the source within a step by the straight line from g_n to g_{n+1}. Set $z = h\lambda$ and use the unit interval variable $\theta = (s - t_n)/h$. Variation of constants first yields

$$y_{n+1} = e^z y_n + h\int_0^1 e^{(1-\theta)z}[g_n + \theta(g_{n+1} - g_n)]\, d\theta.$$

Integrate the constant and linear source terms separately. For $z \neq 0$ direct evaluation gives the two weights and their connection:

$$\varphi_1(z) = \frac{e^z - 1}{z},$$

$$\varphi_2(z) = \frac{e^z - 1 - z}{z^2},$$

$$y_{n+1} = e^z y_n + h[\varphi_1 g_n + \varphi_2(g_{n+1} - g_n)].$$

The apparent poles at zero are removable. Expanding the exponential before subtracting gives stable small-argument evaluations:

$$\varphi_1 = 1 + z/2 + z^2/6 + \cdots,$$

$$\varphi_2 = 1/2 + z/6 + z^2/24 + \cdots.$$

An accurate routine for $e^z - 1$ helps with the first weight, but subtracting z to compute the second can still lose digits. Use its own series near zero. At $z = 0$ the formula reduces to trapezoidal integration of the source, which is exact for the chosen linear interpolant. This limit tests simulated components whose linear decay is zero.

WORKED EXAMPLE

Check a strongly damped ramp

Set $\lambda = -2.00$ s⁻¹, $h = 0.500$ s, $y_n = 0$ and let the source rise from zero to 4.00 units/s over the step. Here $z = -1$ and $\varphi_2(-1) = e^{-1} = 0.367879$. The calculated rise is $0.500(4.00)(0.367879) = 0.73576$ units. Without decay the same ramp deposits exactly 1.00 unit. Both predictions can be checked directly by integrating the specified source history.

SECOND APPLICATION

Test a nearly neutral component

Take $z = -10^{-8}$ and evaluate $e^z - 1 - z$. Its correct numerator is about 5.00×10^{-17}, comparable to the rounding uncertainty from subtracting one from a number near one. The series instead gives $\varphi_2 \approx 0.499999998333$. For unit-duration integration of a ramp from zero to two units/s the response is about 0.9999999967 units. A returned zero here signals cancellation in weight evaluation, not suppression of the source by the model.

Start from this chapter. Here the exponential propagator is available exactly. Use it as a reference and check how much accuracy the explicit methods achieve at equal physical time.

OneSimulation laboratory · Compare time-step error with conservation

An exact decay factor can hide an aliased heat input

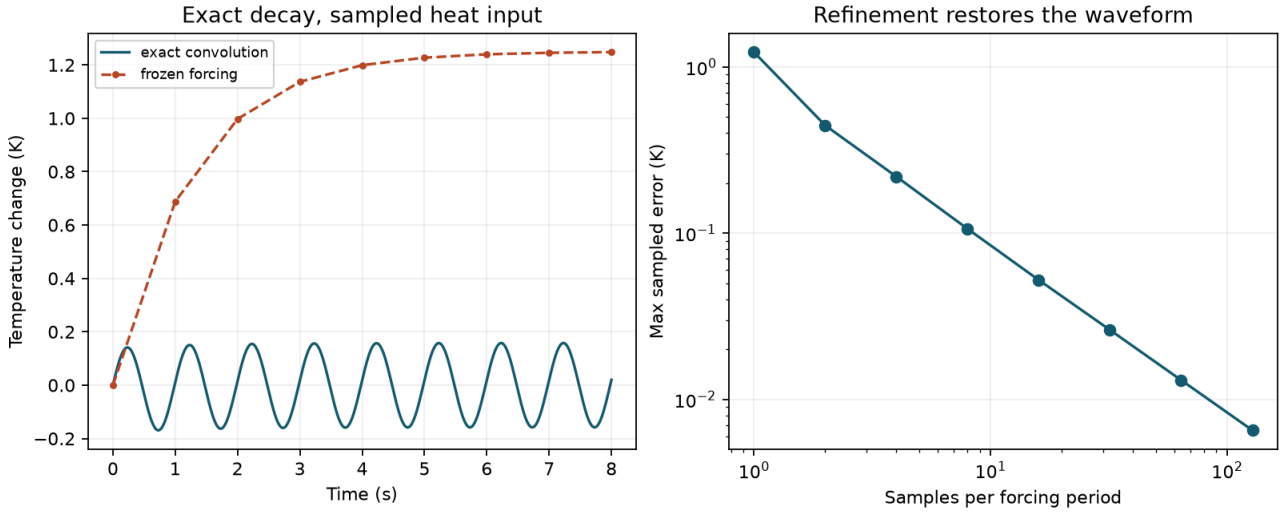
A thermal sensor obeys $y' = -ay + B\cos(\omega t)$ with $a > 0$, $y(0) = 0$. The state is a temperature change; B has units of temperature per time. Multiplying by e^{at} yields an exact integral of the applied force over the step. In complex notation the real part of $B(e^{i\omega t} - e^{-at})/(a + i\omega)$ gives the full continuous solution. This reference tests the response of the sensor, including the initial transient, at any requested time.

An exponential step advances the homogeneous decay by its exact factor e^{-ah}. A tempting low-cost rule freezes the heat input at its sampled left endpoint. Integrating that constant force gives

$$y_{n+1} = e^{-ah} y_n + B(1 - e^{-ah})\cos(\omega t_n)/a.$$

The gain of the homogeneous part is smaller than one for every positive time step. That stability property does not guarantee that the sampled input represents the actual periodic heater.

Choose $h = 2\pi/\omega$. The solver now observes the heater only when its cosine equals one. It therefore predicts $y_n = B(1 - e^{-at_n})/a$, exactly the response to constant heating. At these very same times the analytic periodic experiment gives $Ba(1 - e^{-at_n})/(a^2 + \omega^2)$. Neither numerical roundoff nor an approximation to the decay exponential is responsible for the difference. An entire period of alternating input was replaced by a single repeated measurement.



The default points all observe the same phase. The refinement experiment keeps both the physical signal and the end time fixed while altering how much of the input the step resolves.

In the notebook the default $a = 0.8\,\mathrm{s}^{-1}$, $B = 1\,\mathrm{K/s}$ and frequency one cycle per second make the long-time sampled responses approach 1.25 K and about 0.01994 K respectively. With more samples per period the frozen-input answer begins to recover the oscillation. For very short steps, evaluating its gain as $-\mathrm{expm1}(-ah)/a$ also avoids cancellation between nearly equal floating-point numbers. The two numerical issues call for different remedies: a stable elementary function helps with arithmetic, while additional observations or a better forcing interpolant address temporal aliasing.

A useful alternative test changes the phase of the forcing while holding the step fixed. Starting a full-period sampler at a cosine maximum produces apparent constant heating. Starting at a zero crossing

produces apparently zero heating at every step. The true steady response in either case has the same amplitude, shifted in time. Large sensitivity of the computed amplitude to this arbitrary shift provides another experimental diagnosis of inadequate forcing resolution.

A control engineer halves the step to half a forcing period and finds that successive heat readings alternate between $+B$ and $-B$. Does their zero mean guarantee the correct sampled temperatures?

▼ **Work through the solution**

It restores the correct arithmetic mean of the samples for complete periods. The resulting heat input is still represented internally as a piecewise constant square wave. The convolution weights recent heat input more heavily than earlier input, so matching an unweighted mean leaves amplitude and phase errors. Compare the exact solution at several phases and refine again before reporting either observable.

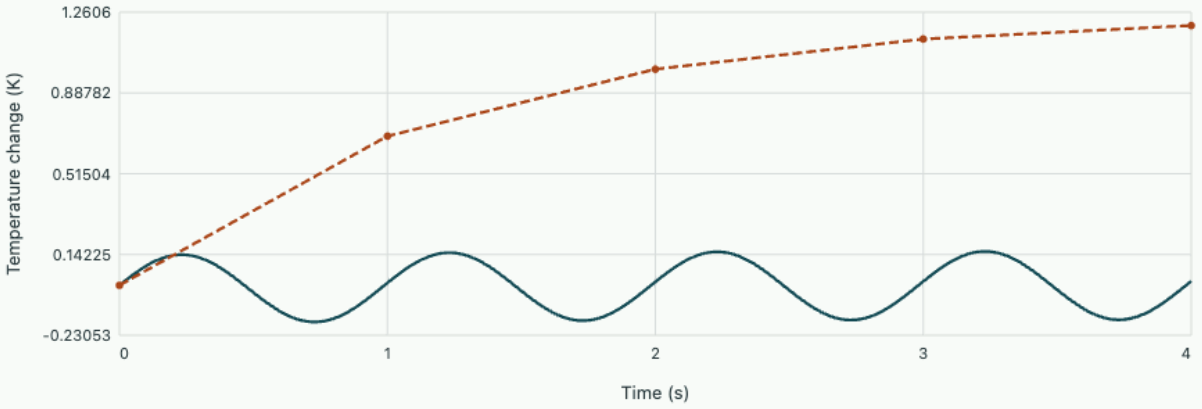
Change parameters · Explore this notebook example

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

Decay rate (1/s)	**0.8**
Forcing frequency (Hz)	**1**
Samples per forcing period	**1**



At 1 samples per cycle, endpoint error 1.1799 K. A single sample per period turns the oscillating drive into an apparently constant input.

Open editable notebook Download .ipynb

Preserve the linear response

For $y' = Ay + g(t, y)$ with constant A, multiply by e^{-At}. The derivative of the product is $e^{-At}g(t, y(t))$, because the two terms containing A cancel. Integrating between t_n and $t_n + h$, then multiplying back, gives the exact variation-of-constants formula

$$y(t_n + h) = e^{hA}y(t_n) + \int_0^h e^{(h-s)A}g(t_n + s, y(t_n + s))\, ds.$$

It separates the free propagation of the initial state from the accumulated, propagated forcing. Freezing g at the beginning of the step gives exponential Euler. Its forcing weight is $h\varphi_1(hA)$, where $\varphi_1(z) = (e^z - 1)/z$ has the continuous value 1 at zero. The defining integral or power series applies to a singular matrix without computing an inverse.

Derive the leading defect

Assume along the solution that the change in g over s is bounded by Ls, and that the linear propagator has norm at most M during the step. Subtract the frozen approximation from the exact integral. Its norm is bounded by $ML\int_0^h s\,ds = MLh^2/2$. This is a local error bound. To obtain a global first-order estimate one must also bound propagation of earlier state errors, using the stability of the linear propagator and the Lipschitz dependence of the remaining forcing on the state. A stiff term hidden inside g can still impose a severe practical restriction.

Heat loss with a steady input

Let $y' = -100y + 20$ with $y(0) = 1$ and time in seconds. The equilibrium is 0.2. At $h = 0.1$, exponential Euler gives $y_1 = e^{-10} + 0.2(1 - e^{-10}) \approx 0.20003632$, exactly the analytical solution at that time. Forward Euler would give -7, whereas backward Euler gives $3/11 \approx 0.27273$. Both the exponential step and backward Euler damp the fast mode here, but the computed amount of transient that survives differs substantially. A large step that accurately reproduces its endpoint can still leave an output trace too sparse to measure the early decay time.

Change the input to $20 + 10t$. During the first step exponential Euler ignores its linear growth. The missing term is

$$10\int_0^h e^{-100(h-s)}s\,ds = 10[h/100 - (1 - e^{-100h})/10000].$$

At 0.1 s it is approximately 0.00900005. Including a linear approximation to the forcing would recover this term exactly. This calculation isolates the effect of freezing a changing input while the stiff decay remains exact.

Evaluating the weights

For small $|z|$, direct subtraction of one from e^z loses significant digits. A dedicated exponential-minus-one function or the series $1 + z/2 + z^2/6 + \cdots$ yields a reliable φ_1. Large sparse systems often require only the action of an exponential on a vector. Computing this action introduces another error budget separate from time approximation; a constant-forcing test with a known equilibrium checks both parts together.

Constant states as an implementation test

If $Ay_* + g_* = 0$, an exponential step starting at y_* must reproduce the same state for every step length. The series identity $h\varphi_1(hA)A = e^{hA} - I$ makes the cancellation exact mathematically. Test this condition with a matrix containing both a zero eigenvalue and a strongly negative eigenvalue: the zero mode should accumulate a constant source linearly, while the decaying mode approaches its forced equilibrium. This test distinguishes a faulty treatment of the zero limit from ordinary temporal error caused by a varying input.

Integrate a linearly varying source with a second exponential weight

For $y' = Ay + b_0 + (t - t_n)b_1$ during a step, the variation-of-constants integral contains a constant and a linear source. Expand $e^{(h-s)A}$ in its power series inside $\int_0^h e^{(h-s)A} s\, ds$. Integrating each term gives $h^{j+2}A^j/(j+2)!$. Consequently the appropriate weight for the slope is $h^2\varphi_2(hA)$ with $\varphi_2(z) = \sum_{j\ge 0} z^j/(j+2)!$. Together with the constant-source φ_1 term this gives the exact step for a source genuinely affine in time.

The identity $z\varphi_2(z) = \varphi_1(z) - 1$ follows from the series. At zero the continuous value is one half, matching the elementary time integral of a ramp. Computing this weight by dividing two nearly equal numbers can lose digits near zero and fails as an ordinary matrix division when A is singular. The series definition and augmented-system formulations give well-defined actions in both cases. If the source depends on the unknown state, an independently known time slope is no longer enough: its total derivative includes the state sensitivity carried along the current trajectory.

Propagate a linearly increasing acceleration through a coupled system

Let $y = (q, v)^T,\ q' = v,\ v' = 2 + 6t$, with $q(0) = 1$ m and $v(0) = 0$. The matrix $A = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ is singular and satisfies $A^2 = 0$. The matrix series therefore terminate:

$$e^{hA} = I + hA,\ h\varphi_1 = hI + h^2A/2,\ h^2\varphi_2 = h^2I/2 + h^3A/6,$$

with arguments hA understood. Multiplying $b_0 = (0, 2)^T$ by the first source weight gives $(h^2, 2h)^T$. Multiplying $b_1 = (0, 6)^T$ by the ramp weight gives $(h^3, 3h^2)^T$.

At $h = 0.5$ s, the freely propagated initial state is still $(1, 0)^T$. The constant-source contribution is $(0.25, 1)^T$ and the ramp contribution is $(0.125, 0.75)^T$. The final state is therefore $(1.375, 1.75)^T$. Independently integrating the acceleration gives $v = 2t + 3t^2$ and $q = 1 + t^2 + t^3$, confirming both numbers. A method freezing acceleration at its initial value would report $q = 1.25,\ v = 1$ despite propagating the coupling $q' = v$ exactly.

This example shows how information travels through a seemingly empty entry of the source: the position receives a ramp contribution although the forcing acts directly only on velocity. Treating the

exponential as an entrywise operation would miss it. For larger sparse systems the same action can be obtained by augmenting the state with coordinates representing the constant and linear input, then applying one matrix exponential to the augmented initial vector.

For the stated ramp, define two extra scalar coordinates $c = 1$ and $s = t$ with $c' = 0,\ s' = c$. Then the augmented linear equation has $y' = Ay + b_0 c + b_1 s$. Powers of this single augmented matrix transfer c successively into s, acceleration, velocity and position. Its exponential is therefore a finite polynomial for the present nilpotent structure. This explicit construction supplies a second implementation that can test the individual φ actions without forming a matrix inverse. If the ramp is restarted at each step, initialise the auxiliary coordinate consistently with whether b_0 is the original source or its value at the new time.

Initialising both auxiliary coordinates anew with their original values while retaining a global-time ramp would repeatedly turn off the accumulated increase of the input.

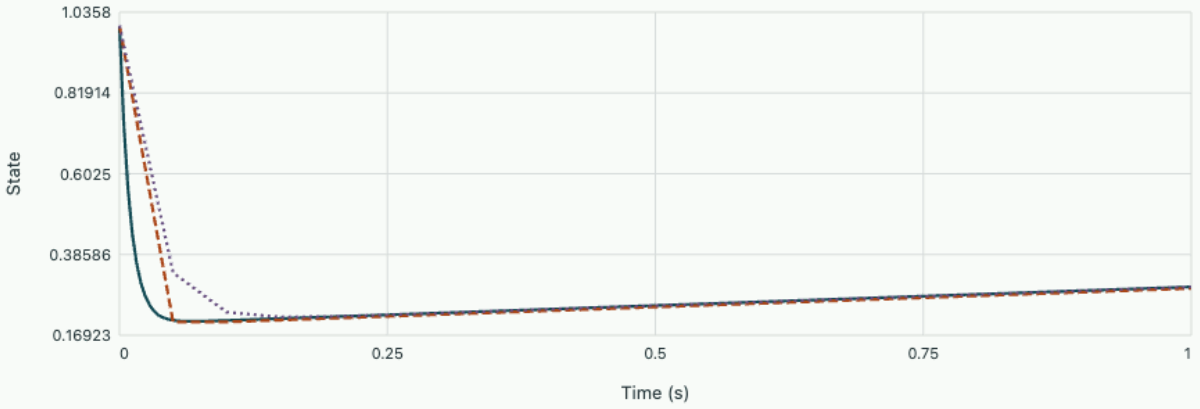
Change parameters · Variation of constants and exponential steps

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

Decay rate k (1/s)	100
Input ramp slope	10
Steps to time 1	20



Frozen-input endpoint error 0.0040339. With zero ramp, exponential Euler reproduces the complete solution at its time nodes.

Chapter 8 · Operator splitting and noncommuting flows

Determine how the order of two exact substeps affects a coupled evolution and recover second order by a symmetric composition.

FROM FIRST PRINCIPLES

Identify what goes wrong when two exact substeps are composed

Suppose two linear processes act simultaneously, $y' = (A + B)y$, with constant matrices and readily computed separate flows. The exact joint step is $e^{h(A+B)}$. Advancing first with A and then with B gives $e^{hB}e^{hA}$. Multiply their Taylor expansions through second order:

$$e^{hB}e^{hA} = I + h(A + B) + h^2(A^2/2 + BA + B^2/2) + O(h^3).$$

In the true joint exponential the mixed term is $(AB + BA)/2$. Subtract it, keeping the order of matrix multiplication:

$$e^{hB}e^{hA} - e^{h(A+B)} = \tfrac{1}{2}h^2(BA - AB) + O(h^3).$$

Each separate step can be exact and the composition can still have an error. If $AB = BA$ the linear flows commute and the composition is exact for all h. In the general case a symmetric schedule places half of the first process on either side of a full step of the second:

$$S_h = e^{hA/2}e^{hB}e^{hA/2} = I + h(A + B) + \tfrac{1}{2}h^2(A + B)^2 + O(h^3).$$

Both mixed products now acquire the exact coefficient $1/2$. Under smooth, stable propagation the resulting global error over a fixed interval is second order. Boundary conditions or an approximate internal subsolver can add other defects.

DERIVATION AND APPLICATIONS

Measure the error created by splitting a source from a decay

Two individually exact updates need not produce the exact combined dynamics. Examine $y' = s - \lambda y$ with constant source s and $\lambda > 0$. Both subproblems have explicit flows: source alone translates by sh, and decay alone multiplies by $e^{-\lambda h}$. The full equation instead admits a stationary state $y_* = s/\lambda$. Subtracting this state and integrating homogeneous decay gives the exact update

$$y_{\text{exact}} = e^{-z}y_n + y_*(1 - e^{-z}),$$
$$z = \lambda h.$$

If the source is applied first, its entire increment is damped for duration h. If decay is applied first, its entire increment escapes damping. A symmetric source-half, decay-full, source-half sequence averages these placements:

$$y_{SD} = e^{-z}y_n + she^{-z},$$
$$y_{DS} = e^{-z}y_n + sh,$$
$$y_{\mathrm{sym}} = e^{-z}y_n + \tfrac{1}{2}sh(1 + e^{-z}).$$

The exact source increment is the integral of decay over all arrival times. The two one-sided approximations are rectangles and the symmetric one is a trapezoid. Expanding the source increments around $h = 0$ gives their local errors:

$$y_{SD} - y_{\mathrm{exact}} = -s\lambda h^2/2 + O(h^3),$$
$$y_{\mathrm{sym}} - y_{\mathrm{exact}} = s\lambda^2 h^3/12 + O(h^4).$$

The opposite one-sided order has leading error of the opposite sign. With a fixed final time, repeated errors usually reduce the respective global orders by one. If $\lambda = 0$ or $s = 0$, all orders coincide exactly. A stationary state in each individual subproblem would be preserved by composition, whereas this full-system equilibrium arises from cancellation of two nonzero rates.

WORKED EXAMPLE

Compare a single step at the exact equilibrium

Let $s = 2.00$ units/s, $\lambda = 2.00$ s⁻¹, $h = 0.500$ s and $y_n = y_* = 1.00$. Exact evolution remains at one. The two split orders give

$$2e^{-1} = 0.73576,$$
$$1 + e^{-1} = 1.36788.$$

The symmetric sequence gives

$$0.5 + 1.5e^{-1} = 1.05182.$$

Symmetry reduces the drift markedly without eliminating it.

SECOND APPLICATION

Predict the long-time state of the discrete map

Iterating the symmetric step eventually reaches its own fixed point $y_h = (sh/2)(1 + e^{-z})/(1 - e^{-z})$. With the preceding data this is 1.08198, not the true value 1.00. Halving h gives

$$0.25(1 + e^{-0.5})/(1 - e^{-0.5}) = 1.02075.$$

The excess falls by a factor close to four, consistent with a second-order bias in a long-running equilibrium calculation.

Two solvable subproblems

Suppose an evolution has the form $u' = (A + B)u$, where the subproblems involving just A or just B have inexpensive exact flows. First applying A for one nominal step and then B gives $e^{hB}e^{hA}u_n$. Each operator is advanced over the same physical interval. When A and B commute, their exponential product equals the exponential of the sum. For general matrices the missing interaction between simultaneous changes controls the error.

Expand the factors through h^2. The mixed term of their product is h^2BA; that of the exact exponential is $h^2(AB + BA)/2$. The difference is $h^2(BA - AB)/2$, the signed commutator term. Reversing the order reverses its sign. Composing a half step of A, a full step of B and another half step of A instead balances both mixed terms. The local discrepancy is then order h^3 for fixed bounded matrices, giving global second order over a fixed interval under stable propagation.

Work out two shears

Let $u = (q, p)^T$, $A(q, p) = (p, 0)$ and $B(q, p) = (0, -q)$, so the exact motion is a unit-frequency oscillator. Each separate flow is a shear. From $(1, 0)$, a drift of length 0.2 followed by a kick produces $(1, -0.2)$. A kick followed by a drift produces $(0.96, -0.2)$. The exact result is $(\cos 0.2, -\sin 0.2) \approx (0.980067, -0.198669)$. The positions from the two first-order compositions bracket the exact one in this example. Their leading positional deviations are almost equal and opposite.

The symmetric drift–kick–drift composition first leaves the position at 1, changes the momentum to -0.2, then drifts by -0.02. Its final state is $(0.98, -0.2)$. The position error is about 6.66×10^{-5} while the momentum error is about 1.33×10^{-3}. Halving the step in a one-step test scales the leading momentum discrepancy roughly by eight. When measuring at a fixed final time instead, the number of steps doubles and the expected asymptotic error reduction is approximately four.

What the substeps must share

Both subflows need compatible state variables and boundary data. If each independently applies an external source intended to appear once in $A + B$, the combined model advances a different equation and cannot converge to the desired solution even with arbitrarily small steps. Conversely a source assigned to neither flow disappears. Check the combined first derivative of the composition at $h = 0$; it must reproduce the complete original right-hand side. This low-cost test catches modelling errors before comparing convergence slopes.

An exact limiting case for a code

Take two diagonal decay matrices with first-coordinate rates –1 and –3 and second-coordinate rates –2 and zero. Each separate flow rescales the coordinates. Their products commute, so starting from $(1, 1)$ and integrating to $t = 0.5$ must give (e^{-2}, e^{-1}) in either order. The equality holds for a single full step as well as any subdivision. This commuting case tests time scaling and state transfer without any expected splitting error. The shear example then tests the distinct question of whether the implementation achieves its intended order when a commutator is present.

Resolve the direction of a nonlinear splitting error

Consider logistic growth $u' = au - bu^2$ with $a, b > 0$. The separate linear growth flow is $G_h(u) = e^{ah}u$. The saturation flow solves $u' = -bu^2$ and is $D_h(u) = u/(1 + bhu)$ for positive initial states. Each flow is exact, positivity preserving and inexpensive. But the saturation strength depends on the current population, which growth changes during a step. Therefore applying growth before saturation generally produces a different result from applying saturation before growth.

Write the two full-step maps explicitly. Growth then saturation yields $e^{ah}u/[1 + bhe^{ah}u]$; saturation then growth gives $e^{ah}u/(1 + bhu)$. For positive u, h and a the first denominator is larger and the final value smaller. The exact nonlinear flow has denominator $1 + (b/a)(e^{ah} - 1)u$, obtained by solving the linear equation for $1/u$. Since $h < (e^{ah} - 1)/a < he^{ah}$, the two orderings bound the exact positive answer on opposite sides. This ordering result uses the particular signs and monotonicity of logistic growth; it is not a generic enclosure theorem for split systems.

Test a physical equilibrium that neither subflow preserves

Set $a = b = 1,\ u_0 = 1$. The complete model is exactly stationary because growth and saturation cancel. Choose $h = \log 2$, so an isolated growth step doubles the state. Growth followed by saturation gives $2/(1 + 2\log 2) \approx 0.8381$. Saturation followed by growth gives $2/(1 + \log 2) \approx 1.1812$. Neither result is the stationary true value. Powering either map through many long intervals would move the solution toward a numerical equilibrium different from the physical one, although each individual substep is solved without approximation.

A symmetric growth–saturation–growth step gives

$$u_1 = 2/(1 + \sqrt{2}\log 2) \approx 1.0100.$$

Symmetry has cancelled the leading ordering discrepancy without enforcing exact preservation of the steady state. To predict the long-run drift of the first ordering, solve its nonzero fixed-point equation $u = 2u/(1 + 2hu)$, giving $u = 1/(2h) \approx 0.7213$. For the opposite ordering the numerical equilibrium is $1/h \approx 1.4427$. These values expose a persistent bias that a one-step comparison can understate.

A steady physical reference is particularly informative when a splitting method is used to integrate dynamics near a balance of large opposing processes. Compare the numerical equilibrium and the

relaxation rate toward it. An implementation can improve its order by a symmetric composition while still requiring much finer steps, or a different assignment of source terms, to maintain the equilibrium accuracy a physical application demands.

A refinement study can use the exact nonzero fixed point formulas to separate local error from accumulated drift. For arbitrary positive h, the growth-first numerical fixed point is $(1 - e^{-h})/h = 1 - h/2 + O(h^2)$, while the saturation-first fixed point is

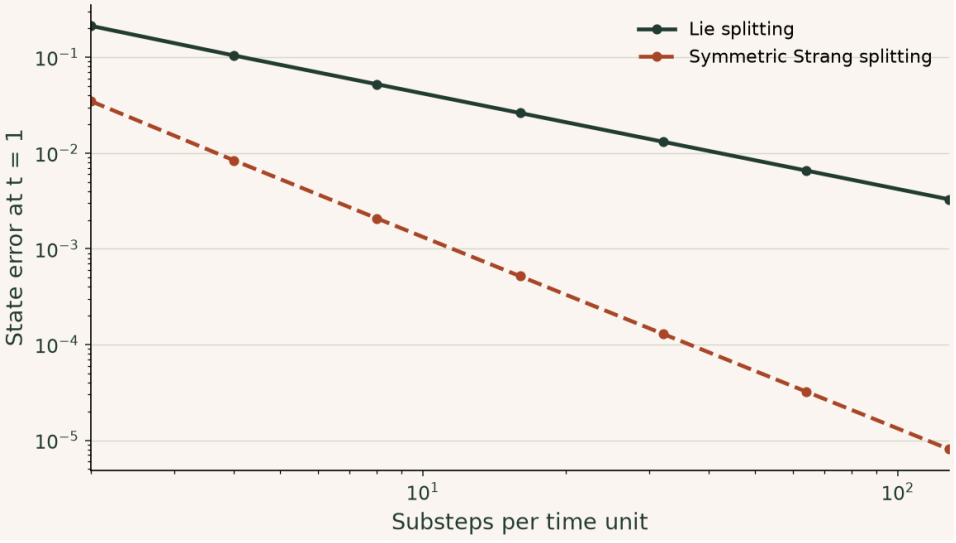
$$(e^h - 1)/h = 1 + h/2 + O(h^2).$$

Thus their steady-state biases have opposite signs and are first order even though the individual one-step defects from the physical equilibrium start at second order. An empirical comparison based on the first step and another measured after equilibration should not be expected to return the same power of the step size.

The two steady-state formulas also converge toward the same physical value as the step tends to zero.

Measure order lost by composing two exact subflows

For the unit harmonic oscillator split the velocity and acceleration updates into exact shears. Start from position one and momentum zero. Compose a full shear in each coordinate, or sandwich the momentum update between two half position updates.



Both approaches solve each individual shear exactly. Halving the step approximately halves the Lie error but quarters the symmetric error once in the asymptotic regime. The discrepancy comes from noncommuting subflows.

▶ **Read values from the plot**

Change parameters · Measure order lost by composing two exact subflows

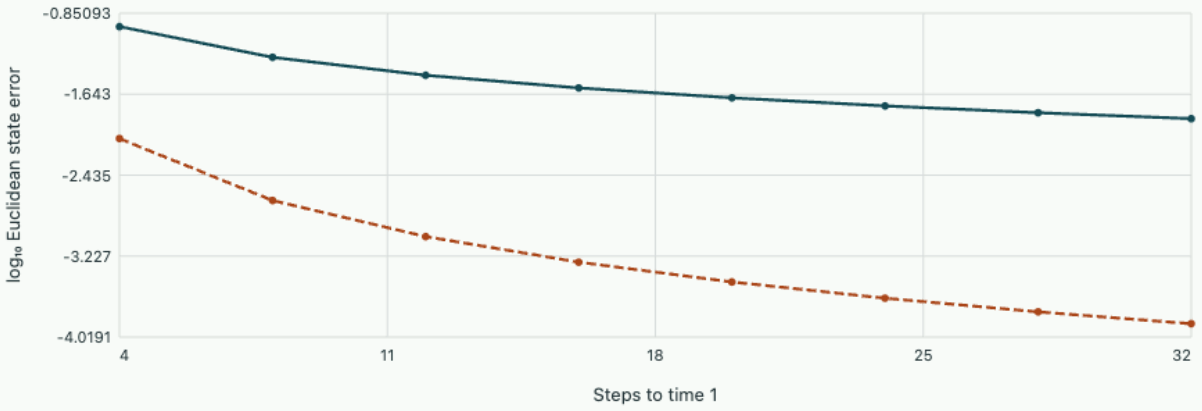
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Oscillator frequency ω **1**

Maximum steps to t = 1 **32**



Lie composition Symmetric Strang composition

Each shear substep is exact. The error comes from composition of noncommuting flows; expect first versus second global order.

Problem sheet 4 · Linear response and split flows

Separate errors caused by a changing source from errors caused by composing noncommuting dynamics.

Preserve an input equilibrium

Take $y' = -8y + 12$, $y(0) = 0$, step $h = 0.5$. Compute a first exponential Euler and backward Euler step. Check the equilibrium under both methods and compare their decay multipliers to it.

▼ **Worked solution**

The equilibrium is 1.5. The exponential step gives $1.5(1 - e^{-4}) \approx 1.47253$. Backward Euler solves $5y_1 = 6$ and gives 1.2.

Starting exactly at 1.5 leaves both methods there. An initial perturbation is multiplied by $e^{-4} \approx 0.01832$ for the first method and $1/5 = 0.2$ for the second. Agreement on steady state leaves a sizable difference in transient damping.

Explore the related calculation with adjustable inputs. Change the decay rate and forcing ramp, and compare the exact convolution with a frozen-input step.

A ramp with no decay

Take the limit of exponential Euler for $y' = t$, starting at $t = 0, y = 0$. Compute one step and the exact result. Then apply two half steps and derive the general global error on $[0, T]$ for uniform step $h = T/N$.

▼ **Worked solution**

With $A = 0$, $\varphi_1(0) = 1$. Freezing the input at zero gives zero after one step of length h; the exact value is $h^2/2$. Two half steps instead give

$$0 + (h/2)(h/2) = h^2/4.$$

Over the fixed interval the method sums $h \sum_{j=0}^{N-1} jh = h^2 N(N-1)/2$. Subtracting this from $T^2/2$ gives $Th/2$. A quadratic local discrepancy has accumulated into a first-order final error.

Explore the related calculation with adjustable inputs. Change the decay rate and forcing ramp, and compare the exact convolution with a frozen-input step.

Reverse the shears

For $q' = p,\ p' = -q$ start at $(0, 1)$, $h = 0.4$. Compute drift–kick, kick–drift and symmetric kick–drift–kick. Compare their momentum to the exact value and identify which two first-order errors have opposite signs.

▼ **Worked solution**

Drift–kick gives $(0.4, 0.84)$; kick–drift gives $(0.4, 1)$. The half kick initially has no effect, then the drift gives $q = 0.4$, and the second half kick gives $p = 0.92$.

Exact momentum is $\cos(0.4) \approx 0.921061$. The first two momentum errors are about −0.0811 and +0.0789; the symmetric error is about −0.00106. All three positions agree in this initial state, which makes momentum the more discriminating one-step test.

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 · An exact decay factor can hide an aliased heat input

Chapter 9 · Störmer–Verlet and a discrete invariant

Derive the leapfrog update, identify its oscillatory stability interval and quantify the difference between bounded energy and correct phase.

FROM FIRST PRINCIPLES

Derive a reversible position update from Newton's law

For a force that depends only on position, $m\ddot{q} = F(q)$, the acceleration can be evaluated at a stored position without an implicit solve. Expand the exact position one step forward and one step backward about the same time t_n. The odd derivatives reverse sign and cancel when the expansions are added:

$$q(t_n + h) + q(t_n - h) = 2q(t_n) + h^2\ddot{q}(t_n) + O(h^4).$$

Replace acceleration by force divided by mass and solve for the forward position. The resulting relation is unchanged if the labels $n + 1$ and $n - 1$ are exchanged:

$$q_{n+1} = 2q_n - q_{n-1} + \frac{h^2}{m}F(q_n).$$

Introduce the staggered momentum $p_{n-1/2} = m(q_n - q_{n-1})/h$. The same algebra becomes an impulse followed by a drift,

$$p_{n+1/2} = p_{n-1/2} + hF(q_n),$$
$$q_{n+1} = q_n + hp_{n+1/2}/m.$$

A specified momentum at $t = 0$ starts the first half step with $p_{1/2} = p_0 + hF(q_0)/2$. For smooth forces the method has second-order position accuracy. It is important to distinguish the staggered momentum from momentum at the stored position time when checking energy.

DERIVATION AND APPLICATIONS

Why a stable Verlet trajectory can arrive at the wrong time

Long-term energy behaviour is important in reversible mechanics, but phase error can dominate a timing measurement. Consider the velocity-independent linear oscillator $q'' = -\omega^2 q$. Use a uniform step h, an accurately initialized first point and the position recurrence obtained by eliminating the half-step momenta in Verlet. Write $z = h\omega$:

$$q_{n+1} - 2q_n + q_{n-1} = -z^2 q_n.$$

Assume a discrete mode $q_n = e^{in\vartheta}$. Divide by the mode value at n and combine the two exponentials. The angle advanced per step obeys

$$2\cos\vartheta - 2 = -z^2,$$
$$\sin(\vartheta/2) = z/2.$$

For $0 < z < 2$ choose the branch continuous with the exact oscillation at small h. The discrete angular frequency and its expansion are

$$\widetilde{\omega} = \frac{2}{h}\arcsin(z/2) = \omega\left[1 + \frac{z^2}{24} + O(z^4)\right].$$

The numerical oscillator runs slightly fast even though its amplitude can remain bounded indefinitely. Its phase lead grows approximately linearly with elapsed time at fixed step. Above $z = 2$ the dispersion relation permits no real phase angle, signaling a separate stability failure. The formula concerns time discretization alone; incorrect initial half-step velocities can add an initial phase offset.

WORKED EXAMPLE

Estimate a ten-period timing error

For $\omega = 5.00$ rad/s and $h = 0.100$ s, $z = 0.500$ and $\widetilde{\omega} = 20\arcsin(0.250) = 5.054$ rad/s. Ten true periods last 12.57 s. Over that duration the numerical phase gains about $(5.054 - 5.00)(12.57) \simeq 0.674$ rad. The estimated time at which the numerical oscillator completes its tenth period is $20\pi/5.054 = 12.43$ s, about 0.13 s too soon.

SECOND APPLICATION

Select a step from a phase requirement

Suppose 100 periods should accumulate less than 0.100 rad of phase error. The leading estimate gives $2\pi(100)z^2/24 < 0.100$, so $z < 0.0618$. For $\omega = 5.00$ rad/s this means $h < 0.0124$ s. A choice based only on the stability requirement $h < 0.400$ s would be far too coarse. Halving a small step should reduce the accumulated phase error by approximately four at fixed physical duration.

Position and momentum at staggered times

For unit mass and a time-independent potential $V(q)$, the equations are

$$q' = p,\; p' = -V'(q).$$

Treat the potential force over a half step, drift with the updated momentum for a full step, then use the new position for the last half kick. This yields

$$p_{n+1/2} = p_n - \tfrac{h}{2} V'(q_n),$$
$$q_{n+1} = q_n + h p_{n+1/2},$$
$$p_{n+1} = p_{n+1/2} - \tfrac{h}{2} V'(q_{n+1}).$$

Successive half kicks at the same position share a force evaluation. Reversing the sign of h exactly undoes the steps in reverse order, in exact arithmetic. Subtracting consecutive position increments gives

$$q_{n+1} - 2q_n + q_{n-1} = -h^2 V'(q_n).$$

The starting position and velocity must initialise the very first half step correctly; using the full velocity as the half-step velocity injects an initial phase error.

The oscillator has an exactly computable modified energy

For $V = \omega^2 q^2/2$ define $z = h\omega$. The position coefficient of q_n after one step is

$$a = 1 - z^2/2;$$

the momentum coefficient of q_n is $-h\omega^2(1 - z^2/4)$. The full map has equal diagonal entries a and off-diagonal entries $h, -hc$ with $c = \omega^2(1 - z^2/4)$. Since $a^2 + h^2 c = 1$, expanding squares shows $p_{n+1}^2 + c q_{n+1}^2 = p_n^2 + c q_n^2$.

When $0 < z < 2$, $c > 0$ and level sets of this invariant are bounded ellipses. The eigenvalues are $e^{\pm i\theta}$ with $\cos\theta = a$. Beyond two, one real eigenvalue has magnitude greater than one and trajectories generically grow. At exactly two the invariant loses control of position and a repeated eigenvalue can cause linear growth. An apparently bounded run starting from specially chosen data at the threshold cannot certify stability for arbitrary perturbations.

Follow a step and its timing

For $\omega = 1, h = 1, q_0 = 1, p_0 = 0$, the half kick gives -0.5, the drift gives $q_1 = 0.5$, and the final half kick gives $p_1 = -0.75$. The physical energy decreases from 0.5 to 0.40625 in this first step. The modified energy $(p^2 + 0.75q^2)/2$ equals 0.375 at both points. The discrete phase is $\theta = \pi/3$, so the computed trajectory returns after six steps. The true oscillator returns after $2\pi \approx 6.283$ seconds. A stable trajectory with a bounded energy error can therefore accumulate a large timing error over many cycles.

For small z, expanding $\theta = 2\arcsin(z/2)$ gives

$$\theta = z + z^3/24 + O(z^5).$$

At a fixed final time the accumulated phase error is therefore second order in h. A period estimate and an energy trace probe different aspects of long-time integration and should be reported separately.

Why the initial velocity test matters

A test with only the initial position nonzero may conceal the borderline instability at $h\omega = 2$. At that boundary the update is

$$q_{n+1} = -q_n + hp_n,\ p_{n+1} = -p_n.$$

For $p_0 = 0$ the positions merely alternate sign forever. With $q_0 = 0$ and $p_0 = 1$, however, explicit iteration gives

$$p_n = (-1)^n,$$

$$q_n = (-1)^{n-1} nh.$$

The position magnitude grows linearly. A stability test must account for all independent starting directions or small perturbations of the chosen one, especially when an eigenvalue of modulus one is repeated.

Prove conservation of angular momentum by individual substeps

For a particle of constant mass in a radial potential $V(\|q\|)$, the force is parallel to the position vector. Its angular momentum is $L = q \times p$. During a kick of duration τ, momentum changes by $\tau F(q)$ at fixed position, so the change in L is

$$\tau q \times F(q) = 0.$$

During a drift, position changes by $\tau p/m$ at fixed momentum, giving change $(\tau/m)p \times p = 0$. The full kick–drift–kick map therefore preserves angular momentum exactly in arithmetic without rounding, for any step length for which these evaluations are defined.

The proof uses radial force and the relationship between velocity and momentum. An external noncentral force exerts a physical torque, while a coordinate-dependent mass matrix changes the drift subflow. Under either modification the two zero cross products need rechecking. The argument also gives no bound on energy or phase. Exact conservation of angular momentum constrains orbital geometry but leaves freedom to move from the correct energy level onto a nearby one or to return to a point on the orbit at the wrong time.

Compare three diagnostics on an elliptical oscillator trajectory

Use unit mass and two-dimensional potential $V = (x^2 + y^2)/2$. Start at $q = (1, 0)$, $p = (0, 2)$, so $L = 2$ and physical energy is

$$E = (1 + 4)/2 = 2.5.$$

At step $h = 0.5$, the first half kick gives

$$p_{1/2} = (-0.25, 2).$$

The drift leads to $q_1 = (0.875, 1)$. The last half kick subtracts $0.25 q_1$, giving $p_1 = (-0.46875, 1.75)$. The resulting planar angular momentum is $0.875(1.75) - 1(-0.46875) = 2$, exactly as proved.

Calculate energy separately. The squared position is $0.875^2 + 1 = 1.765625$, and the squared momentum is

$$0.46875^2 + 1.75^2 = 3.2822265625.$$

Hence the physical energy after this step is 2.52392578125, about 0.96 percent above the initial value. The exact continuous position at this time is $(\cos 0.5, 2 \sin 0.5) \approx (0.87758, 0.95885)$, so the position error is concentrated in the second coordinate. Angular momentum, physical energy and the individual trajectory coordinates expose three distinct properties of the same update.

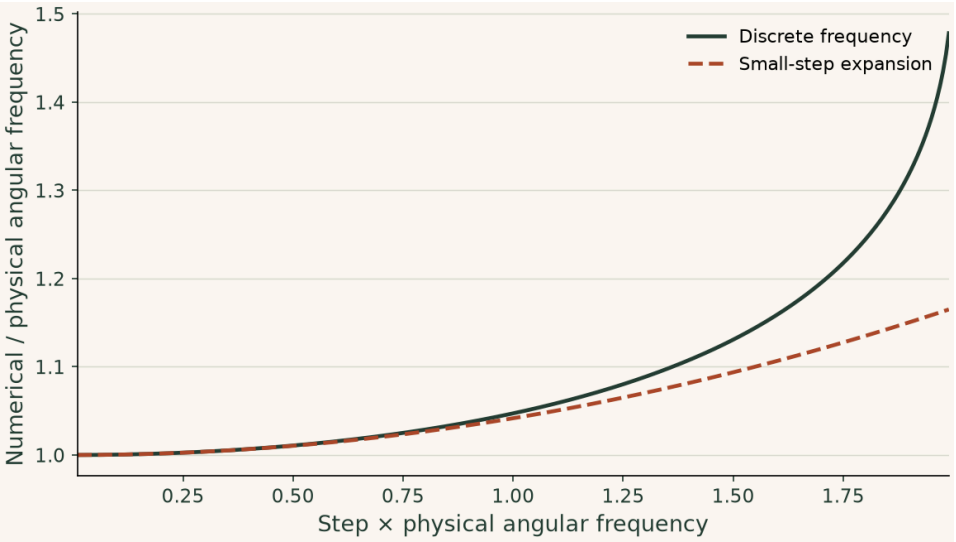
Because the harmonic force is separable by coordinate, the modified invariant from the scalar oscillator adds across the two directions: $[\|p\|^2 + (1 - h^2/4)\|q\|^2]/2$ is constant. Substitution gives 2.46875 at both steps, independently checking the arithmetic. For a general radial potential this particular quadratic modified invariant no longer follows, while the angular momentum proof remains valid. A conservation claim should identify the symmetry responsible rather than assuming every observed invariant survives a change of potential.

If there were a small constant transverse force, the physical change in angular momentum would be the time integral of $q \times F_{\text{ext}}$. A better diagnostic would subtract the accumulated applied torque from the measured change in L, rather than demand that L remain fixed. This distinction also guards against a subtle implementation error: zeroing a real torque to improve a conservation plot would change the physical problem and could hide a significant trajectory bias.

Measuring the signed torque makes its direction as well as its magnitude available for this balance check.

Estimate the phase error of a stable Verlet orbit

Eliminate momentum from velocity Verlet for the harmonic oscillator. A discrete sinusoid obeys $\cos\theta = 1 - (h\omega)^2/2$. Its angular frequency is θ/h; divide by ω before comparing steps.



At hω = 1 the numerical frequency exceeds the true one by about 4.72%. The scheme can maintain a bounded orbit yet acquire a substantial phase error after many cycles. The chart stops below the stability boundary at two.

▶ **Read values from the plot**

Change parameters · Estimate the phase error of a stable Verlet orbit

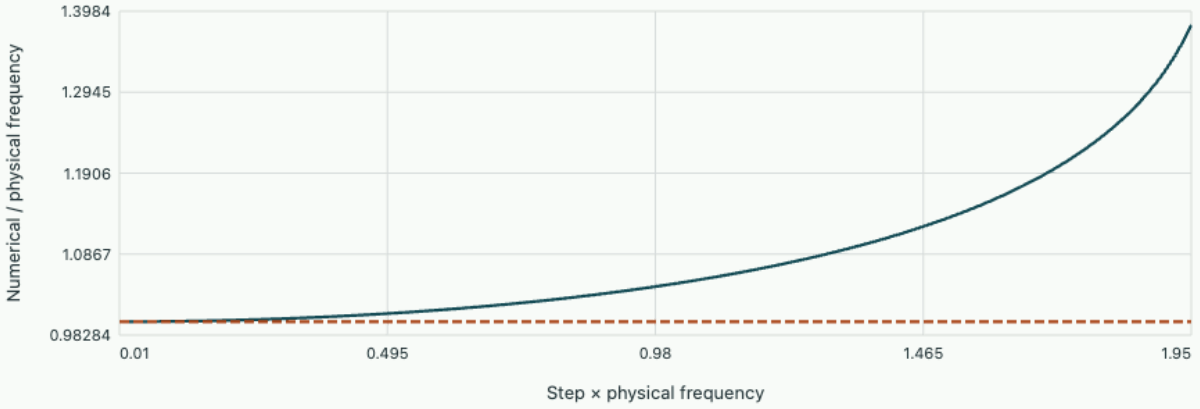
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Selected hω 1

Physical cycles 10



Velocity Verlet Exact frequency

Selected frequency ratio 1.0472; phase lead after 10 exact cycles is 2.9655 rad. The stability boundary is hω = 2.

Chapter 10 · Tangent equations and adjoint outputs

Compute how an observed final state changes when an initial state or a model parameter is perturbed.

FROM FIRST PRINCIPLES

Differentiate the trajectory that your instrument measures

Let a scalar parameter θ enter both a smooth rate law $y' = f(t, y, \theta)$ and the initial data $y(0) = g(\theta)$. Compare two solutions at the same physical time. Their difference divided by a small parameter increment tends to the sensitivity $S = \partial y / \partial \theta$. Differentiating the rate law at fixed time supplies a term from the moving state and a direct parameter term:

$$\dot{S} = f_y S + f_\theta,$$
$$S(0) = g'(\theta).$$

For a scalar terminal measurement $J = \psi(y(T), \theta)$ the chain rule gives

$$J' = \psi_y S(T) + \psi_\theta.$$

When many parameters enter a single output, a backward equation can reuse their common output weighting. Choose $\lambda(T) = \psi_y^T$ and $\dot{\lambda} = -f_y^T \lambda$. Then the terms containing the unknown sensitivity cancel in the derivative of their scalar product:

$$\frac{d}{dt}(\lambda^T S) = \lambda^T f_\theta,$$
$$J' = \psi_\theta + \lambda(0)^T g'(\theta) + \int_0^T \lambda^T f_\theta \, dt.$$

Evaluate all derivatives on the original trajectory. For $y' = -\theta y$ with fixed $y(0)$, direct differentiation gives $S(T) = -T y(T)$, the sign expected when increasing the decay rate. A parameter-dependent observation time would add an endpoint term.

DERIVATION AND APPLICATIONS

Differentiate the numerical update that produced the observed output

A gradient used in fitting a parameter should agree with the discretized objective actually evaluated. Consider explicit Euler on a scalar equation $y' = f(y, \theta, t)$ with fixed, parameter-independent time steps and initial state. A single final observation defines $J = \Phi(y_N, \theta)$. Treat the computed states as a chain of differentiable maps rather than replacing their dependence by that of the exact differential equation.

For the step $F_n(y, \theta) = y + hf(y, \theta, t_n)$ a small perturbation satisfies

$$\begin{aligned} \delta y_{n+1} &= A_n \delta y_n + B_n \delta\theta, \\ A_n &= 1 + h f_y, \\ B_n &= h f_\theta \end{aligned}$$

with derivatives evaluated on the stored numerical trajectory. Start from $\delta J = \Phi_y \delta y_N + \Phi_\theta \delta\theta$ and substitute the last step. Repeating this substitution backward defines a weight on each earlier state:

$$\begin{aligned} \lambda_N &= \Phi_y, \\ \lambda_n &= A_n \lambda_{n+1}. \end{aligned}$$

Each stage contributes its local parameter derivative weighted by the future effect of that stage. For a fixed initial condition the total derivative is

$$dJ/d\theta = \Phi_\theta + \sum_{n=0}^{N-1} \lambda_{n+1} B_n.$$

With vector states the reverse update uses the transpose Jacobian. An initial condition depending on θ adds $\lambda_0 \, \partial_\theta y_0$. The resulting gradient is exact for the discrete map up to arithmetic precision, though it can differ from the gradient of the continuous solution. Adaptive control decisions need special care if differentiating through changes in the time grid.

WORKED EXAMPLE

Differentiate a two-step fitted decay

Let $y' = -\theta y$, $y_0 = 2.00$, $h = 0.250$, $\theta = 1.00$, and $J = (y_2 - 1)^2/2$. The Euler states are 1.50 and 1.125, giving $\lambda_2 = 0.125$, $\lambda_1 = 0.09375$. The local parameter derivatives are -0.500 and -0.375. Thus

$$J' = 0.09375(-0.500) + 0.125(-0.375) = -0.09375.$$

Increasing the decay rate locally reduces the positive final residual.

SECOND APPLICATION

Compare with the continuous sensitivity at the same time

At $t = 0.500$ the exact value is $2e^{-0.5} = 1.21306$ and its derivative is $-0.500(1.21306) = -0.60653$. The continuous objective therefore has derivative -0.1292. That difference is expected at this coarse step. A centred parameter perturbation of the two-step numerical objective should approach -0.09375, providing a direct implementation check without assuming that the time integrator is already accurate.

Differentiate the evolution

Let $y' = f(t, y, \alpha)$ with smooth f and $y(0) = y_0(\alpha)$. Assume that the solution exists and depends differentiably on α throughout the observation window. Differentiating the equation along the solution gives

$$S' = f_y S + f_\alpha,$$
$$S(0) = y_0'(\alpha),$$

where $S = \partial y / \partial \alpha$. The first term propagates state changes through the dynamics and the second injects the direct parameter effect. Losing the initial derivative gives the wrong sensitivity whenever a parameter also changes the initial preparation.

For an output $Q = \phi(y(T), \alpha)$, the chain rule gives

$$Q_\alpha = \phi_y S(T) + \phi_\alpha.$$

Each parameter requires its own tangent column. Many observations can then reuse the propagated columns. Scaling matters because a derivative with respect to milliseconds is numerically different from the derivative with respect to seconds. Multiplying by a stated parameter perturbation converts the derivative back into a predicted observation change that can be checked directly.

A rate parameter and a final measurement

Take $y' = -\alpha y,\ y(0) = 2$. Then

$$S' = -\alpha S - y,\ S(0) = 0.$$

Multiplying by $e^{\alpha t}$ leaves $(e^{\alpha t} S)' = -2$, hence $S = -2te^{-\alpha t}$. At $\alpha = 0.5$ per second and $T = 2$ s the sensitivity of final amplitude is $-4/e \approx -1.47152$ amplitude units per inverse second. Raising the decay rate by 0.01 per second predicts a decrease of 0.014715. The exact change is $2e^{-1}(e^{-0.02} - 1) \approx -0.01457$; the discrepancy is quadratic in the parameter perturbation.

Testing a tangent code by differences requires controlling both the ODE integration error and subtraction error. Use several perturbation sizes. With a forward difference the truncation error normally decreases linearly at first, until numerical noise in subtracting two nearly equal outputs dominates.

Run one dual equation for a scalar output

Define a column vector z backward in time by $-z' = f_y^T z,\ z(T) = \nabla_y \phi$. Differentiating $z^T S$ cancels the two terms containing f_y. Integrating what remains gives

$$Q_\alpha = \phi_\alpha + z(0)^T y_0'(\alpha) + \int_0^T z^T f_\alpha \, dt.$$

For the preceding final-amplitude observation $z = e^{-\alpha(T-t)}$. Its product with $f_\alpha = -y$ is the constant $-2e^{-\alpha T}$, whose integral reproduces the tangent result. One adjoint trajectory can provide gradients with respect to many parameters for this single output. Its coefficients require the forward trajectory at earlier times, so storing or reconstructing that trajectory is part of the numerical cost. If the desired derivative is of the discrete numerical update itself, differentiating its specific steps and interpolation choices can matter at finite time resolution.

A cancellation at a particular observation time

Compare observing the final amplitude with observing the lost amplitude $Q = y(0) - y(T)$. If the rate is the only parameter, changing the observation reverses the sign of the computed rate sensitivity. If instead the unknown is the common initial amplitude, the direct derivative of $y(0)$ is one and the propagated derivative is $e^{-\alpha T}$. The final derivative of the loss is $1 - e^{-\alpha T}$, tending to zero as the observation window shrinks. Omitting the direct term would predict a large negative initial sensitivity for a loss that is identically zero at zero time.

Differentiate the time at which a model switches

If a trajectory switches dynamics at a threshold, the event time itself depends on the parameter. Let a pre-event solution $y^-(t, p)$ meet $G(t, y, p) = 0$ at $\tau(p)$. Denote by S^- its parameter derivative at fixed time, evaluated on the nominal event. Differentiating the event equation gives

$$\tau_p = -\frac{G_y S^- + G_p}{G_t + G_y f^-}.$$

The denominator is the rate at which the trajectory crosses the threshold. This formula requires a transversal crossing and a locally unchanged event sequence. At a grazing contact the derivative may diverge or cease to exist, and a perturbed trajectory can encounter a different number of events.

Assume the state itself is continuous while the right-hand side changes from f^- to f^+. The total parameter change of the event state is $S^- + f^- \tau_p$. But the post-event tangent at a fixed clock time must remove the amount of post-event evolution gained or lost by shifting the start time, giving $S^+ = S^- + (f^- - f^+)\tau_p$. This jump in a derivative of a continuous state is sometimes surprising. Comparing trajectories taken at the same physical clock time is different from comparing their states at their respective event times.

Estimate the final position after a velocity switch

Let a cart start at zero and travel at constant positive speed p until it reaches position one. It then travels at fixed speed 0.5. Observe its position at $T = 2$, and examine $p = 1$, so the event occurs at $\tau = 1$. The pre-event position is pt and its tangent at fixed time is

$$S^- = t = 1.$$

Using $G = y - 1$ gives

$$\tau_p = -1/p^2 = -1.$$

Since $f^- = 1$ and $f^+ = 0.5$, the tangent immediately after the switch becomes

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

Subsequent evolution at the fixed speed adds no further parameter source.

Check directly: as long as the event precedes the observation, $y(T, p) = 1 + 0.5(2 - 1/p)$. Its derivative at one is indeed 0.5. At $p = 1.01$ the exact final position is approximately 1.5049505 instead of 1.5. The first-order prediction is 1.505. Carrying the pre-event tangent unchanged through the switch would predict 1.51, nearly doubling the response. Resetting it to zero because the event position is fixed would miss the change entirely.

At $p = 0.5$ the switch is exactly at the observation endpoint. Parameters below this value give $y(T) = 2p$, whereas parameters above it use the switched expression. Both one-sided derivatives happen to equal two here because the two velocities match at this specific threshold value. Their second derivatives differ. Changing the post-event speed to 0.25 makes even the first derivatives disagree at the boundary. An apparent smooth result in one carefully chosen switching case cannot justify ignoring event timing throughout a parameter study.

A finite-difference sensitivity test should consequently check which events occurred on both perturbed trajectories. Agreement in total simulated duration and integrator tolerance is insufficient if the jump sequence changed. Recording event times together with the reported output makes it possible to distinguish an error in the reset law from a derivative that has ceased to exist across the selected parameter interval.

Propagate a reset depending on the event location

If a switching event resets the state by $y^+ = R(t, y^-, p)$, differentiate the reset along the moving event before returning to a fixed clock time. The result is

$$S^+ = R_y S^- + R_p + (R_t + R_y f^- - f^+)\tau_p.$$

This reduces to the continuous-state formula when R is the identity. The occurrence time must still come from the pre-event trajectory, since the post-event state generally does not satisfy the original trigger condition.

Modify the cart example so the trigger position is $H(p) = 1 + 0.2(p - 1)$, and the indicated position is reset to $R = 0.3y^- + 0.2$ before it increases at speed 0.5. At $p = 1$, $\tau = H/p = 1$ and $\tau_p = 0.2 - 1 = -0.8$. The pre-event fixed-time tangent is one. The reset gives

$$S^+ = 0.3 + (0.3 - 0.5)(-0.8) = 0.46.$$

The indicated final position at time two is $0.3H + 0.2 + 0.5(2 - H/p)$, exactly one at the nominal parameter. Differentiating this expression gives $0.3(0.2) - 0.5(0.2 - 1) = 0.46$, independently verifying the sign and the event-time correction. Simply differentiating the reset map at the nominal crossing position would give 0.3 and lose the varying length of the post-event trajectory.

Change parameters · Explore the worked calculation

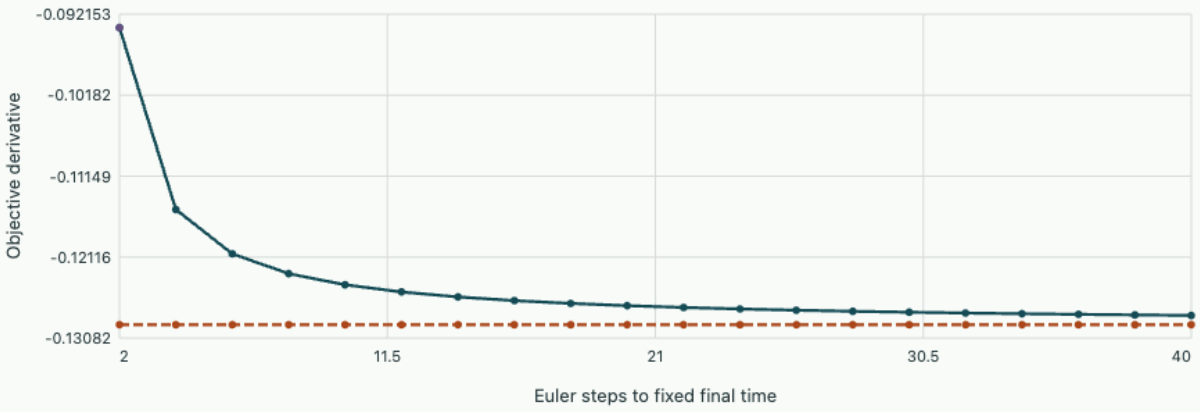
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Decay parameter θ 1

Euler steps to t = 0.5 2



Exact derivative of discrete objective | Continuous objective derivative | Chosen grid

At 2 Euler steps the derivative is -0.09375 versus continuum -0.12923. A correct discrete adjoint differentiates the numerical trajectory actually computed.

Problem sheet 5 · Long trajectories and sensitivities

Check numerical phase, derive the tangent contribution from initial data, and verify an adjoint for an accumulated observation.

Predict a numerical period

For velocity Verlet with $h = 0.5$, $\omega = 2$, compute the stability indicator and period in time units. Starting with $q = 0, p = 2$, give the first step and verify the modified energy.

▼ **Worked solution**

$z = h\omega = 1$ gives $\cos\theta = 1 - 1/2 = 1/2$, hence $\theta = \pi/3$. Six steps make a period of 3 time units, versus the exact period π.

The first half kick is zero, the drift gives $q = 1$, and the last half kick leaves $p = 1$. With $c = 4(1 - 1/4) = 3$, the modified energy $(p^2 + 3q^2)/2$ takes the value 2 at both states.

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.

Initial data also carry the parameter

For $y' = -\alpha y,\ y(0) = \alpha$, find $S = \partial y/\partial\alpha$ directly and from its differential equation. At what positive observation time is the first-order sensitivity zero if $\alpha = 2$?

▼ **Worked solution**

Direct differentiation of $y = \alpha e^{-\alpha t}$ gives

$$S = e^{-\alpha t}(1 - \alpha t).$$

The tangent equation is

$$S' = -\alpha S - y,$$

with $S(0) = 1$. Its integrating factor produces

$$e^{\alpha t} S = 1 - \alpha t.$$

At $t = 1/\alpha = 0.5$, the initial-amplitude increase and the increased decay rate cancel to first order. Setting the initial tangent to zero would miss this cancellation completely.

Explore the related calculation with adjustable inputs. Compare the derivative of the actual discrete trajectory with the continuous derivative as the time grid changes; include any initial-state derivative required by the present problem.

An accumulated output

Let $y' = \alpha$, $y(0) = 0$, $Q = \int_0^T y(t)\,dt$. Compute $dQ/d\alpha$ directly. Derive an adjoint with terminal value zero by integrating zS' by parts, and check $T = 3$.

▼ **Worked solution**

The trajectory is αt, so $Q = \alpha T^2/2$. The tangent satisfies $S' = 1$, $S(0) = 0$. To recover $\int S$, choose $-z' = 1,\ z(T) = 0$, giving $z = T - t$.

Integration by parts yields $\int_0^T S = \int_0^T zS'$ because the boundary product vanishes. Integrating $T - t$ gives $T^2/2 = 4.5$ for $T = 3$, the same gradient.

Explore the related calculation with adjustable inputs. Compare the derivative of the actual discrete trajectory with the continuous derivative as the time grid changes; include any initial-state derivative required by the present problem.

Chapter 11 · Shooting for two-point boundary values

Turn a missing initial slope into a scalar root, measure the sensitivity of endpoint matching and recognise an unstable propagation direction.

FROM FIRST PRINCIPLES

Convert a missing boundary slope into a measurable residual

Consider a twice differentiable scalar field on $0 \le x \le L$ satisfying $y'' = f(x, y, y')$, $y(0) = a$, $y(L) = b$. A forward initial-value solver needs both the starting value and its slope. Guess a slope s, integrate the two first-order equations $y' = v,\ v' = f(x, y, v)$ from (a, s), then record the signed endpoint discrepancy:

$$R(s) = y(L; s) - b.$$

Changing the guess by a small amount changes the residual by its derivative times the change. To compute that derivative along the same trajectory define $z = \partial_s y$, $w = \partial_s v$. Differentiate both the equations and their initial conditions:

$$\begin{aligned} z' &= w, \\ w' &= f_y z + f_v w, \\ z(0) &= 0, \\ w(0) &= 1. \end{aligned}$$

Now $R'(s) = z(L)$. Setting the linearised next residual to zero gives the Newton correction:

$$s_{\text{new}} = s - \frac{y(L; s) - b}{z(L)}.$$

For $f = 0$ the propagated sensitivity is exactly $z(x) = x$. The correction recovers the slope of the line connecting the boundary values in one step. A nearly zero $z(L)$ signals poor slope identifiability from that endpoint. A large growing sensitivity instead makes forward propagation vulnerable to numerical contamination between the boundaries.

Background · Shooting residual

$$F(s) = B(y(t_f; s))$$

Here s encodes missing initial data and B the final boundary condition. An unstable trajectory map can make the residual badly conditioned even with an accurate initial-value solver.

DERIVATION AND APPLICATIONS

Recognize two shooting solutions and the fold between them

A successful endpoint residual can describe more than one solution. Consider the dimensionless steady boundary-value model $u'' + \lambda e^{u} = 0$ on $0 < x < 1$, with $u(0) = u(1) = 0$ and positive λ. It represents a simplified nonlinear source balanced by diffusion. A shooting code would vary the unknown initial slope until the final value vanishes. To understand the possible roots, consider a symmetric trial family centred at $x = 1/2$:

$$u_{\mu}(x) = 2\ln\frac{\cosh(\mu/2)}{\cosh[\mu(x-1/2)]}, \qquad \mu > 0.$$

This family automatically enforces both boundary conditions. Differentiating twice gives

$$u_{\mu}'' = -2\mu^{2}\operatorname{sech}^{2}[\mu(x-1/2)].$$

The exponential source has the same spatial factor. Cancel it to determine exactly which problem parameter each profile solves:

$$\lambda(\mu) = 2\mu^{2}\operatorname{sech}^{2}(\mu/2),$$
$$u_{\mu}'(0) = 2\mu\tanh(\mu/2).$$

As μ increases, λ initially grows quadratically, then decays toward zero. Differentiate its logarithm to locate the turning point that separates the two branches of initial slopes:

$$\frac{d\ln\lambda}{d\mu} = \frac{2}{\mu} - \tanh(\mu/2) = 0.$$

Numerical solution of $\mu\tanh(\mu/2) = 2$ gives $\mu \simeq 2.3994$ and $\lambda_{\mathrm{fold}} \simeq 3.5138$. Below this value and above zero the family gives two distinct symmetric positive solutions. Close to the fold a small change in λ can produce a large change in the solution branch. Numerical residual alone cannot tell which branch a shooting iteration found.

WORKED EXAMPLE

Generate a known solution for checking a solver

Choose $\mu = 2$. Then $\lambda = 8/\cosh^{2}(1) = 3.3598$, the initial slope is $4\tanh(1) = 3.0464$ and the centre value is $2\ln\cosh(1) = 0.8676$. Integrating from $(u, u') = (0, 3.0464)$ should return near zero at $x = 1$ and near 0.868 at the midpoint. The second observation checks the shape, not only the enforced endpoint.

SECOND APPLICATION

Identify the branch from an additional reading

At the fold the maximum centre value on the lower-μ branch is

$$2\ln\cosh(2.3994/2) = 1.187.$$

A computed solution at $\lambda = 3.00$ with centre value 1.98 therefore lies on the higher-μ branch. Solving $\lambda(\mu) = 3$ on that side gives $\mu \approx 3.29$ and centre value about 1.98. A different initial slope producing the same endpoint can represent another valid solution.

Replace a missing slope by a parameter

Consider $y'' = f(t, y, y')$, $y(0) = a,\ y(T) = b$. Integrate as an initial-value problem with trial slope $s = y'(0)$ and define $R(s) = y(T; s) - b$. A root enforces the second boundary. If the initial-value solution is smooth in s, its derivative satisfies $v'' = f_y v + f_{y'} v'$ with $v(0) = 0,\ v'(0) = 1$. The derivative of the shooting residual is $v(T)$. Newton therefore updates the missing slope by $-R/v(T)$; a vanishing endpoint sensitivity calls for a different parameterisation or further analysis of existence and uniqueness.

Consider $y'' = -y,\ y(0) = 0,\ y(T) = 1$, for which $y(t; s) = s\sin t$. At $T = \pi/2$ the solution has $s = 1$. At $T = \pi$ no slope reaches the required endpoint. Changing the terminal boundary to zero at the same time admits every slope. An integrator returning a tiny nonzero value for $\sin\pi$ might suggest an enormous numerical slope, so inspect endpoint sensitivity and the mathematical boundary conditions before trusting a root.

A well-defined solution with a fragile shooting trajectory

For $y'' = k^2 y,\ y(0) = 1,\ y(1) = 0$, direct propagation gives

$$y(t) = \cosh(kt) + (s/k)\sinh(kt),$$

with exact slope $-k\coth k$. The required boundary-value solution is equivalently $\sinh(k(1-t))/\sinh k$, a bounded positive curve. A slope perturbation δs produces endpoint error $\sinh(k)\delta s/k$. For $k = 10$, this multiplier is approximately 1101.32. A slope error of 10^{-5} thus creates endpoint error 0.0110 even though the exact profile stays between zero and one.

This large number describes forward amplification of slope error. The inverse sensitivity of the fitted slope to the terminal boundary value is its reciprocal. At large k, direct propagation must subtract two growing numbers to recover the small endpoint. Tight initial-value tolerances can postpone the resulting loss of accuracy, while leaving the underlying direction of exponential growth intact.

Match shorter intervals

In multiple shooting introduce independent state guesses at intermediate nodes. Integrate each segment and impose matching equations equating its terminal state to the next guess. Together with the boundary conditions these equations form a sparse coupled nonlinear system. On a half interval the growing factor for $k = 10$ is e^{5} instead of e^{10}. Requiring continuity couples the segments while reducing the range of each forward propagation. Cancelling all the intermediate states algebraically can reintroduce the original long product.

Check the final curve against both endpoint conditions, the maximum matching defect and the differential equation. Small endpoint mismatch alone says little about the interiors of independently matched segments. For highly unstable propagation a simultaneous boundary-value discretisation or collocation method can be a more suitable formulation.

Separate root iteration error from integration error

Suppose the propagated terminal state carries absolute numerical error at most ϵ. Near a simple shooting root the induced slope uncertainty is approximately $\epsilon / |R'(s)|$. For the oscillator with $T = \pi/6$, the sensitivity is one half, so endpoint uncertainty 10^{-5} leaves slope uncertainty roughly 2×10^{-5}. Solving the numerical shooting equation to a terminal residual of 10^{-12} would not establish twelve accurate digits of slope. Change the propagation tolerance independently and compare the recovered slopes before interpreting the outer root solver's convergence.

Use shooting to locate an eigenvalue

For a string fixed at both ends, the eigenvalue problem $-u'' = \lambda u,\ u(0) = u(L) = 0$ has the trivial solution for every parameter. A shooting formulation that merely adjusts the initial slope while allowing it to vanish can therefore converge to an uninformative zero mode. Fix the amplitude scale instead by prescribing $u'(0) = 1$. For each trial positive λ define the residual $R(\lambda) = u(L; \lambda)$. A root then enforces the far boundary for a nonzero solution because the prescribed starting slope excludes the trivial state.

On the unit interval the solution is $u(t; \lambda) = \sin(\sqrt{\lambda}\, t)/\sqrt{\lambda}$, so the residual vanishes at $\lambda_j = j^2 \pi^2,\ j = 1, 2, \ldots$. The limiting solution at zero is $u = t$, which fails the far boundary. At $\lambda < 0$ the sine becomes a hyperbolic sine and again cannot vanish away from the origin. These facts locate the whole physical spectrum before considering numerical root searches. For a variable material coefficient the simple formula disappears, but amplitude normalisation and endpoint residual remain well-defined.

Bracket a mode and check its identity inside the interval

It is convenient to solve for wave number $k = \sqrt{\lambda} > 0$. At $k = 3$ the residual is $\sin 3 / 3 \approx 0.04704$, and at $k = 3.2$ it is $\sin 3.2 / 3.2 \approx -0.01824$, bracketing the first eigenvalue. Its derivative is $(k \cos k - \sin k)/k^2$. Starting with $k = \pi$ in the exact formula gives $R' = -1/\pi$. An absolute

integration error of 10^{-5} in the endpoint displacement can therefore shift the fitted wave number by roughly 3.14×10^{-5}. Propagating through $\lambda = k^2$ gives eigenvalue uncertainty about 1.97×10^{-4}.

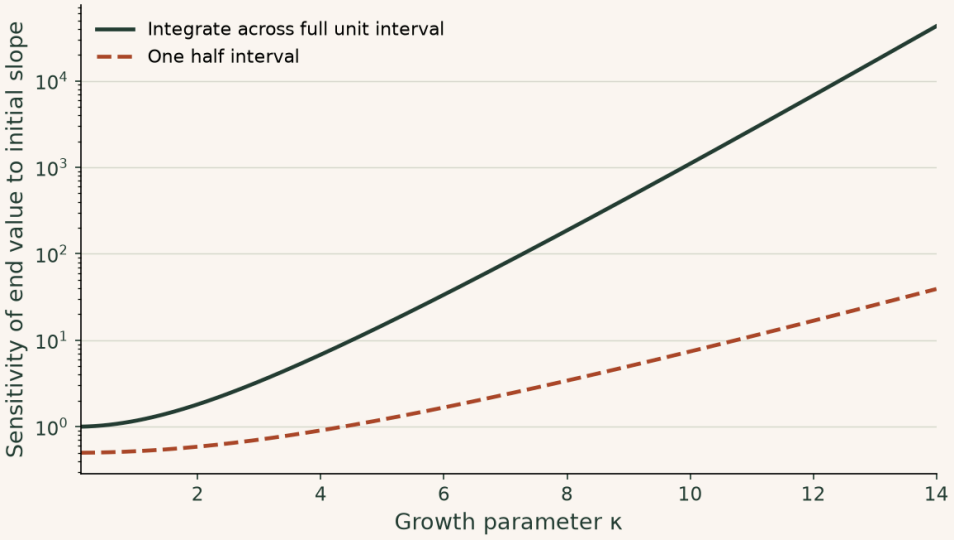
A residual search on $k \in [2, 7]$ could instead overlook both $k = \pi$ and 2π, since the residuals at the endpoints have the same positive sign. Counting interior nodes provides an independent mode label. With normalisation $u'(0) = 1$, the first eigenfunction has no internal zero and the second has exactly one at $t = 1/2$. Check these zero locations on the computed trajectory at a resolution sufficient to distinguish oscillations. A small endpoint residual alone confirms neither the desired mode index nor the absence of missed modes.

For high modes the physical wavelength decreases and the initial-value integrator must resolve phase over many cycles. A frequency estimate can drift substantially even while the trajectory's amplitude remains bounded. Refining the propagation tolerance, refining the scalar root tolerance and counting the interior nodes are three different checks. Changing all three simultaneously can hide which part of an apparent convergence failure needs attention.

A second normalisation issue appears when comparing eigenfunctions with sensor data. Multiplying any computed eigenfunction by a nonzero constant or by minus one leaves both homogeneous boundary conditions and the eigenvalue unchanged. Matching raw amplitudes or signs between two independently found modes can therefore suggest an error where only conventions differ. Normalise both functions in the same mass-weighted norm and align their signs before checking their shapes; keep the nodal count as a separate check on the mode index.

Measure the conditioning of a shooting residual

For $y'' = \kappa^2 y$ with fixed initial position, perturb the initial slope by one. The resulting change at distance L is $\sinh(\kappa L)/\kappa$. Compare one shooting integration with a single segment of a multiple-shooting formulation.



At $\kappa = 10$ a slope error of 10^{-6} becomes roughly 1.10×10^{-3} across the full interval, versus 7.42×10^{-6} over one half. Multiple shooting still has to solve its matching equations; the half-interval curve measures only the local propagation.

▶ **Read values from the plot**

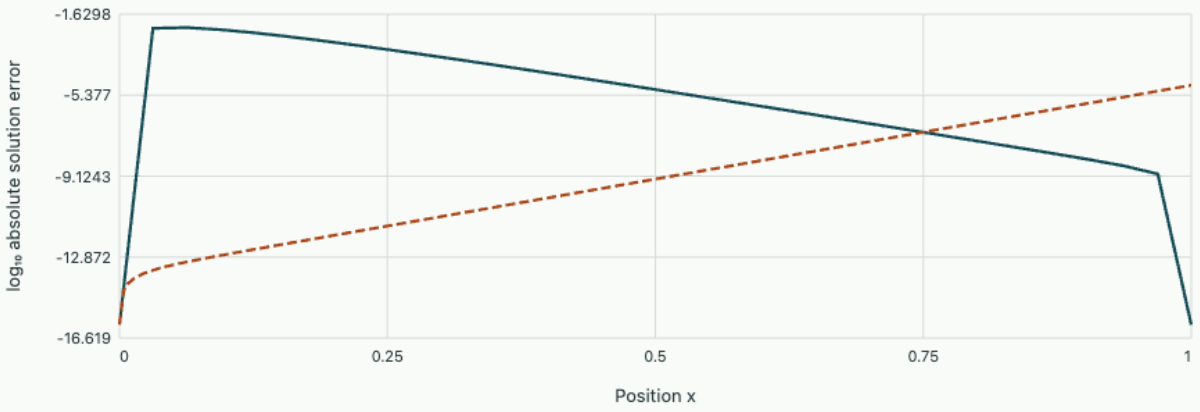
Change parameters · Measure the conditioning of a shooting residual

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

Layer parameter λ	20
Number of mesh intervals	32



λh = 0.625. Maximum collocation error 0.005642. The slope perturbation produces 0.00001213 right-face error.

Study the full derivation

Solving a two-point boundary problem without guessing the start slope · Dynamics and numerical time integration

When endpoint conditions act on different sides of an interval, shooting can amplify a tiny initial error. Construct a cubic collocation system, test its continuous differential residual and quantify which boundary-value solution the residual actually certifies.

Chapter 12 · Differential–algebraic equations and consistency

Derive a reduced evolution when an algebraic variable can be eliminated and find the hidden velocity constraint in a mechanical example.

FROM FIRST PRINCIPLES

Determine which initial states can satisfy an algebraic constraint

A differential–algebraic model may evolve y while determining z instantaneously through a constraint. In a semi-explicit model, $y' = f(t, y, z)$ specifies a rate and $g(t, y, z) = 0$ specifies admissible states. Initial data must satisfy the latter before the first time step. A differentiable solution must remain on it, hence the time derivative along the trajectory vanishes:

$$0 = \frac{d}{dt} g(t, y, z) = g_t + g_y f + g_z \dot{z}.$$

If the square matrix g_z is invertible near the initial state, the constraint locally determines z from t, y, and the hidden rate is

$$\dot{z} = -g_z^{-1}(g_t + g_y f).$$

For example impose the unit circle $y^2 + z^2 = 1$ with $y' = u(t)$. On its upper half $z = \sqrt{1 - y^2}$, giving $\dot{z} = -yu/z$. Algebraic consistency fixes $z(0)$ once the branch and $y(0)$ are known. At $z = 0$ this coordinate elimination becomes singular even though the circle itself is smooth. A constrained implicit time step imposes the original algebraic equation at the new time, which prevents the numerical drift that can accumulate from integrating only its differentiated version.

DERIVATION AND APPLICATIONS

Derive the constraint force of a pendulum without eliminating coordinates

A solver using Cartesian coordinates for a pendulum must satisfy geometric, velocity and acceleration constraints. Let $\mathbf{r} = (x, y)$ be the displacement from a fixed pivot, $\mathbf{v} = \dot{\mathbf{r}}$, m the bob mass, and L the fixed link length. The applied force $\mathbf{F}$ excludes the constraint reaction. Write the reaction as $-\lambda \mathbf{r}$, so positive λ corresponds to an inward force and has units of N/m. The holonomic constraint and its first derivative are

$$g = \tfrac{1}{2}(\mathbf{r} \cdot \mathbf{r} - L^2) = 0,$$
$$\dot{g} = \mathbf{r} \cdot \mathbf{v} = 0.$$

Differentiating once more shows that the acceleration is generally not tangent: maintaining curvature of the trajectory requires an inward normal component. Insert Newton's law and keep the known tangential speed:

$$\ddot{g} = |\mathbf{v}|^2 + \mathbf{r}\cdot\dot{\mathbf{v}} = 0,$$
$$m\dot{\mathbf{v}} = \mathbf{F} - \lambda\mathbf{r}.$$

Solving for the scalar multiplier gives both the required magnitude and the tangentially projected forcing:

$$\lambda = \frac{m|\mathbf{v}|^2 + \mathbf{r}\cdot\mathbf{F}}{L^2},$$
$$\dot{\mathbf{v}} = \frac{\mathbf{F}}{m} - \frac{\lambda}{m}\mathbf{r}.$$

The multiplier can be found only after compatible position and velocity are supplied. Satisfying the position constraint alone at the initial time does not remove radial velocity. For a rigid rod either sign of λ is admissible. A string goes slack when the bilateral model would require compression.

WORKED EXAMPLE

Calculate a reaction at the bottom of a swing

Choose y upward, $m = 0.200$ kg, $L = 0.500$ m, $\mathbf{r} = (0, -0.500)$ m and $\mathbf{v} = (1.00, 0)$ m/s. Gravity is $(0, -1.962)$ N. Thus $\lambda = [0.200 + 0.981]/0.250 = 4.724$ N/m, giving tension $\lambda L = 2.36$ N. The net upward acceleration is $(2.362 - 1.962)/0.200 = 2.00$ m/s², independently equal to v^2/L.

SECOND APPLICATION

Detect an inconsistent input and a physically slack string

At the same bottom position, a supplied velocity $(1.00, 0.100)$ m/s has $\mathbf{r}\cdot\mathbf{v} = -0.0500$ m²/s. Removing its radial projection gives $(1.00, 0)$ m/s before evaluating the multiplier. At the top, instead take $\mathbf{r} = (0, 0.500)$ m and tangential speed 1.00 m/s. The bilateral equation predicts

$$\lambda L = 0.400 - 1.962 = -1.56\,\mathrm{N}.$$

A string cannot provide that outward push, so a taut-string simulation must release the constraint there.

A state has fewer independent coordinates than entries

A semi-explicit system has $y' = f(t, y, z)$ and $0 = g(t, y, z)$. The algebraic equation supplies a condition at every instant; choosing every initial coordinate independently can violate it before time integration starts. If g_z is square and nonsingular locally, the implicit function theorem determines $z = \psi(t, y)$ and substitution gives an ordinary differential equation for y. Differentiating the constraint

gives $g_z z' = -g_t - g_y f$, so one differentiation determines the missing derivative. This is the usual semi-explicit index-one situation.

For $y' = -y + z, \quad 0 = y + 2z - t$, one has $z = (t - y)/2$ and $y' = (t - 3y)/2$. If $y(0) = 2$, consistency forces $z(0) = -1$ and $y'(0) = -3$. Differentiating the constraint gives

$$z'(0) = (1 - y'(0))/2 = 2.$$

Entering both initial coordinates as zero and two in the wrong order would describe a different trajectory even if a numerical solver silently repaired their algebraic inconsistency.

Discretise both conditions at the same time

A backward Euler step at $h = 0.2$ enforces $y_1 - 2 = 0.2(-y_1 + z_1)$ and $y_1 + 2z_1 = 0.2$. Eliminating z_1 gives $1.3y_1 = 2.02$, so $y_1 \approx 1.553846$ and $z_1 \approx -0.676923$. Computing the second variable from the new time matters. Using z_0 throughout the update and reporting it as the new algebraic variable would violate the current constraint. Report its residual together with the local or global accuracy estimate for y.

Hidden constraints and the force multiplier

For a unit-mass pendulum with Cartesian position q, velocity v, rod length L and downward gravity vector g, take $q' = v, \; v' = g - \lambda q$ and $q^T q = L^2$. Differentiating the rod constraint once imposes $q^T v = 0$; velocity must be tangent to the circle. Differentiating again gives $v^T v + q^T g - \lambda L^2 = 0$, determining the multiplier $\lambda = (v^T v + q^T g)/L^2$. The original position constraint contains no direct multiplier, so the index-one elimination used above does not apply.

At the bottom of a 1 m pendulum take $q = (0, -1)$ m, $v = (2, 0)$ m/s and $g = (0, -9.81)$ m/s². The multiplier is 13.81 s⁻² and acceleration is $(0, 4)$ m/s², the required inward centripetal acceleration. A velocity with a nonzero vertical component would already violate tangency at this position. Advancing differentiated equations alone can let position and tangent constraints drift because rounding and discretisation accumulate. Check both $q^T q - L^2$ and $q^T v$ over time; projecting position alone may leave a velocity pointing slightly through the rod constraint.

An algebraic residual can hide an inaccurate state

For the linear example, change a consistent state at fixed time by $\delta y = 0.02, \; \delta z = -0.01$. Its constraint residual stays exactly zero, since those changes cancel in $y + 2z$. Its differential right-hand side changes by -0.03. The perturbation lies along the constraint manifold and is invisible to any diagnostic that checks only the algebraic equation. Conversely a perturbation normal to the manifold can be easy to see in the residual even if y is accurate. Estimate time-discretisation error by a separate comparison, such as refining the time step at a common physical output time.

Project an inconsistent state in a physically weighted metric

Suppose linear constraints require $Bq = g$, while an unconstrained time step produces q^* with a small violation. A correction can be selected by minimising $(q - q^*)^T M(q - q^*)/2$ subject to the constraint, with M symmetric positive definite and B of full row rank. Stationarity gives

$$M(q - q^*) + B^T\lambda = 0.$$

Inserting the resulting $q = q^* - M^{-1}B^T\lambda$ into the constraint yields the smaller system $BM^{-1}B^T\lambda = Bq^* - g$. This formula corrects exactly the component normal to the feasible plane in the chosen metric.

If M is a mass matrix for velocity variables, minimising this norm has a kinetic-energy interpretation. For position variables it is still a chosen distance measure and should be justified by the desired correction. Dependent constraint rows can make the multiplier system singular even when the geometrically projected state is unique. Nonlinear constraints require updating their Jacobian during a projection, and a large correction may find a different feasible branch. Satisfying the algebraic constraint alone does not guarantee a time-accurate differential state.

Position and velocity require two distinct projections

Let $M = \mathrm{diag}(1, 4)$, $B = (1, 1)$ and require $q_1 + q_2 = 2$. An unconstrained update gives

$$q^* = (1.2, 1.0)^T,$$

with gap 0.2. The scalar multiplier matrix is $1 + 1/4 = 1.25$, so $\lambda = 0.16$. Since $M^{-1}B^T = (1, 1/4)^T$, the correction is $(-0.16, -0.04)^T$, leaving $q = (1.04, 0.96)^T$. A Euclidean projection would subtract 0.1 from each coordinate. Both restore the same constraint, but the mass-weighted correction moves the heavier coordinate less.

For a constant constraint value, consistent velocities must also satisfy $Bv = 0$. Suppose $v^* = (2, -1)^T$, with violation one. The velocity multiplier is $1/1.25 = 0.8$, giving $v = (1.2, -1.2)^T$. Its kinetic energy is $(1.2^2 + 4 \cdot 1.2^2)/2 = 3.6$, compared with $(2^2 + 4)/2 = 4$ before correction. The lost 0.4 is exactly half the squared mass norm of the removed normal component $(0.8, 0.2)^T$. Orthogonality eliminates the cross term in this energy decomposition.

This reduction illustrates why projecting velocities can suppress energy even if the underlying unconstrained update had no dissipative force. In a simulation of sustained motion, report correction norms or constraint impulses along with the residual after projection. With a moving constraint $Bq = g(t)$, the velocity condition becomes $Bv = g'(t)$, so forced boundary motion can do physical work. Applying a zero-velocity projection to that case would remove a legitimate component of the solution.

For the linear position example the chosen correction is also the smallest in the prescribed metric by direct comparison. Every other feasible correction has the form $(-0.16, -0.04) + t(1, -1)$. Its squared mass norm equals $0.032 + 5t^2$ because the mass-weighted cross term vanishes. The Euclidean

correction corresponds to $t = 0.06$, giving squared mass norm 0.05 rather than 0.032. This calculation makes the meaning of the multiplier solution visible without invoking a general optimisation theorem. If B varies with time, differentiating $B(t)q = g(t)$ also introduces the term $B'(t)q$ in the velocity condition, even for a constant prescribed value.

The derivative of the constraint must therefore be evaluated at the same time as the corrected velocities.

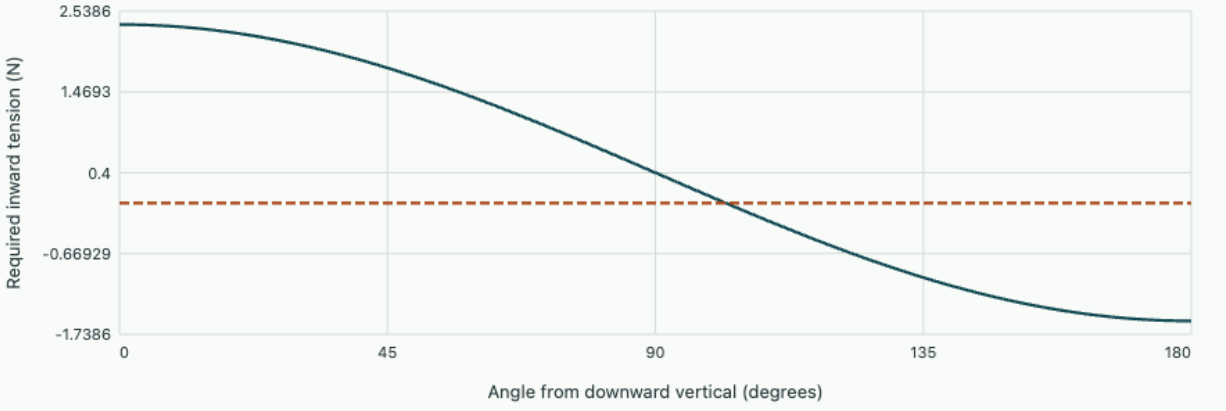
Change parameters · Explore the worked calculation

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

Tangential speed (m/s)	**1**
Length (m)	**0.5**
Mass (kg)	**0.2**



At the bottom the required tension is 2.362 N; at the top -1.562 N. A string cannot supply negative tension: that part of the bilateral prediction corresponds to slack, not constrained motion.

Problem sheet 6 · Boundary and algebraic conditions

Recover missing initial data, measure shooting amplification, and calculate a consistent implicit step for a constrained state.

A slope from an endpoint

Solve $y'' = 2$, $y(0) = 1,\ y(2) = 3$ by shooting. Write the endpoint residual and derivative, take a Newton update from zero, and evaluate the fitted midpoint value.

▼ **Worked solution**

With initial slope s, $y(t) = 1 + st + t^2$. The residual is

$$1 + 2s + 4 - 3 = 2 + 2s.$$

Its derivative is 2.

The update from zero gives $s = -1$ exactly because the residual is affine. At $t = 1$ the fitted value is 1; at $t = 2$ it returns 3. The trajectory initially decreases before turning upward.

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

Quantify an unstable direction

For $y'' = 16y,\ y(0) = 0$, express $y(1)$ as a function of the initial slope. Calculate the slope needed for $y(1) = 2$, and estimate the change in terminal value caused by an absolute slope error of 0.001.

▼ **Worked solution**

Superposition gives

$$y(t) = (s/4)\sinh(4t).$$

Hence $s = 8/\sinh 4 \approx 0.29315$ and the endpoint sensitivity is $\sinh 4/4 \approx 6.8225$.

A slope error of 0.001 causes terminal error about 0.00682. This is exact up to rounding in the multiplier because the differential equation and its slope dependence are linear.

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

Enforce a time-dependent constraint

For $y' = z,\ y + z = 1 + t,\ y(0) = 0$, find consistent initial values and one backward Euler step of length 0.5. Solve the reduced equation and assess why the step has this particular error.

▼ **Worked solution**

The initial constraint gives $z(0) = 1$ and therefore $y'(0) = 1$. At the new time $y_1 = 0.5 z_1$, $y_1 + z_1 = 1.5$, yielding $y_1 = 0.5, z_1 = 1$.

The reduced equation $y' = 1 + t - y$ has exact solution $y = t$ with this initial value. Its derivative is constant along the trajectory, so backward Euler reproduces it exactly here. Different initial data introduce an exponential transient for which the step would have time-discretisation 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.

Separate conservation, stability and accuracy in a decay calculation

A transfers to B with rate 0.8 s^{-1} and A+B = 1 mol/m^3. Start from pure A. Use Euler with a 1.5 s step, backward Euler with the same step, and the exact solution to find A and B after one step. Which tests pass? Predict the fine-grid error ratio for RK4 in the downloadable run, and identify a meaningful comparison of computational effort.

▼ **Worked solution**

The exact A concentration is exp(–1.2) ≈ 0.3010, leaving B ≈ 0.6990. Forward Euler multiplies A by 1–1.2 = –0.2, giving A = –0.2 and B = 1.2. Total concentration is conserved. The amplification modulus is less than one, so repeated steps decay, but every other A value is negative. Absolute stability and positivity test different properties.

Backward Euler gives A = 1/(1+1.2) ≈ 0.4545 and B ≈ 0.5455. Both are positive and their sum is one, yet the A error is about 0.1535. These invariant checks detect important failures, but their success cannot quantify time-discretisation error.

For a smooth solution on a fixed interval the global RK4 error should fall by approximately 2^4 when the step is halved in its asymptotic range. Use the chapter 2 OneSimulation plot to estimate the order from the last two mesh sizes. For a cost comparison count four right-hand-side calls per RK4 step versus one per Euler step. This model is deliberately small enough that wall-clock startup time would obscure that comparison.

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

Chapter 13 · Implicit midpoint, reversibility and the phase of an oscillator

Imagine integrating a mechanical trajectory forward, saving only its end state and then running the same method backward. Recovering the start is a stronger demand than obtaining a plausibly small one-step error. Even exact recovery does not guarantee that an oscillation peaked at the correct time.

Build a method from a centred difference

Integrate $y' = f(t, y)$ over an interval of length h. Midpoint quadrature approximates the integral by h times its integrand at $t_n + h/2$. The arithmetic mean of the unknown endpoint states approximates the exact midpoint state up to order h^2 on a smooth trajectory. Using this mean as the derivative's state argument gives

$$y_{n+1} - y_n = h f(t_n + h/2, (y_n + y_{n+1})/2).$$

Errors from midpoint quadrature and from evaluating f at the approximate midpoint both contribute $O(h^3)$ locally, giving global order two on a fixed interval under the usual regularity and stability conditions. In general the unknown appears inside f, so an implicit equation must be solved. If f is Lipschitz with constant L_f, direct fixed-point iteration contracts when $|h|L_f/2 < 1$. Newton iteration can converge over a wider range. Neither a failed nonlinear solve nor loose tolerances inherit the exact algebraic properties of the method.

Exchange y_n, y_{n+1} and replace h by $-h$. The left side changes sign, the arithmetic midpoint is identical and the midpoint time is identical if the reverse step starts at t_{n+1}. Thus, on the same locally unique solution branch,

$$\Phi_{-h} = \Phi_h^{-1}.$$

A round trip using variable steps must repeat the accepted steps in reverse order. Independently choosing a new adaptive mesh on the way back need not retrace the states.

Solve the step for a harmonic oscillator

Write $q' = \omega p,\ p' = -\omega q$, with scaled momentum p in the same units as q, and $z = q + ip$. The ODE becomes

$$z' = -i\omega z.$$

Solving the linear midpoint equation gives the scalar multiplier

$$M(u) = \frac{1 - iu/2}{1 + iu/2} = \exp[-2i \arctan(u/2)],$$

$$u = \omega h.$$

Its numerator and denominator are complex conjugates. Hence it preserves $q^2 + p^2$, proportional to the oscillator energy, exactly in real arithmetic. For any radius or step size $M(-u)M(u) = 1$. The rotation per step is

$$\theta = 2\arctan(u/2) = u - u^3/12 + O(u^5).$$

The numerical phase lags by $u - \theta$ per step. After N steps the accumulated lag is $N(u - \theta)$, increasing linearly in elapsed time even while the energy remains exactly constant.

Compare explicit midpoint Runge–Kutta. A half Euler predictor leads to

$$M_{\mathrm{RK2}} = 1 - iu - u^2/2.$$

It also has order two, but $|M_{\mathrm{RK2}}|^2 = 1 + u^4/4$. The amplitude grows slowly at every nonzero step. A forward and reverse pair multiplies the starting amplitude by $1 + u^4/4$. In a conservative problem these two second-order procedures can produce visibly different long records.

Inspect the return and the lag separately

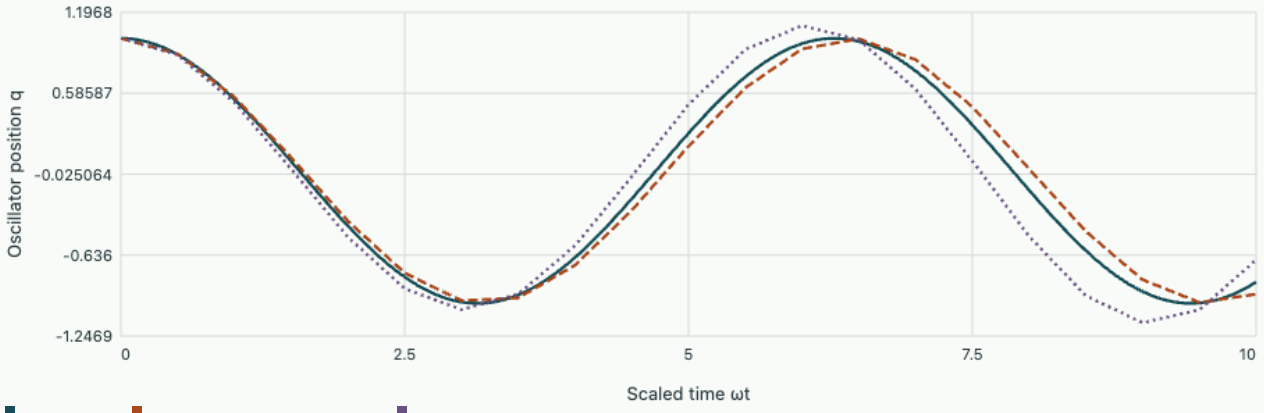
With $\omega = 1$, $h = 0.5$, $N = 20$, exact accumulation is 10 radians. Midpoint gives $40\arctan(0.25) = 9.7991$ radians, a lag of 0.2009 radians. The explicit round trip multiplies the radius by $(1 + 0.5^4/4)^{20} \simeq 1.364$. The implicit round trip returns the original radius and phase to roundoff. Run both updates below. Compare results at equal elapsed time by halving h and doubling N. The phase error should drop by about a factor of four. The graph uses computed steps for the methods and a more finely sampled exact reference.

Open this experiment online

Try the calculation

Time step h (ω = 1) **0.5**

Number of forward steps **20**



Implicit midpoint phase lag 0.2009 rad. Forward-backward distance from the start is 8.951e-16 for implicit midpoint and 0.3635 for explicit midpoint.

Tutorial on designing an accuracy criterion

A solver preserves the harmonic energy exactly over a simulation of 100 time units with $\omega = 1$. A measurement also requires phase error below 0.01 radians. Use the leading midpoint error to estimate an admissible step. How would you verify it?

▼ **Work through the solution**

The accumulated phase error is approximately $T\omega^3 h^2/12$. Solving $100h^2/12 < 0.01$ gives $h < 0.0346$. Verify at a chosen integer step count using $T - 2N\arctan[T/(2N)]$, or by convergence against a refined trajectory that measures the phase continuously. Energy alone has zero error for all these steps, so it cannot certify the phase target.

Further reading · Hairer's TUM geometric numerical integration course, symmetric methods.

Chapter 14 · Splitting a stiff relaxation from a slowly varying drive

A sensor tries to follow a changing reference while a fast relaxation damps departures. An explicit method may oscillate wildly even though the underlying transient decays monotonically. Treating the restoring term implicitly can resolve the slow motion using steps longer than the transient time. Whether those steps measure the transient accurately is a different question.

Construct an exact benchmark with two time scales

Let $s(t) = \sin t$ and choose a relaxation time $\varepsilon > 0$. Consider $y' = s'(t) + (s(t) - y)/\varepsilon$, with $y(0) = 1$. Set $e = y - s$. Subtraction of s' gives $e' = -e/\varepsilon$, hence the reference solution $y(t) = \sin t + e^{-t/\varepsilon}$. The slow reference changes over times of order one, while the initial mismatch decays significantly over ε. We know both pieces exactly and can diagnose errors even when a stable result looks visually smooth.

Explicit Euler applied to a perturbation of the state multiplies that perturbation by $1 - r$, with $r = h/\varepsilon$. It damps monotonically only for $0 < r \le 1$, changes sign while damping for $1 < r < 2$, and grows for $r > 2$. These statements compare nearby solutions at equal drive. The full signal also contains the time variation of s. At the boundary $r = 2$ a state perturbation alternates without decaying. Refining the step to follow a slow input alone can therefore give the wrong choice of h.

Solve only the fast part implicitly

Split $f_{\text{slow}}(t) = \cos t$ from $f_{\text{fast}}(t, y) = (\sin t - y)/\varepsilon$. Evaluate the slow part at t_n and the fast part at t_{n+1}, y_{n+1}. This first-order additive implicit-explicit Euler step needs one scalar linear solve:

$$(1 + r)y_{n+1} = y_n + h\cos t_n + r\sin t_{n+1}.$$

The state perturbation now multiplies by $1/(1 + r)$. It is damped at every positive step and tends to zero as r grows. Subtract $\sin t_{n+1}$ from both sides after solving. The mismatch obeys $e_{n+1} = [e_n + h\cos t_n - (\sin t_{n+1} - \sin t_n)]/(1 + r)$. The bracketed forcing defect is $O(h^2)$ for smooth input. In the stiff limit at fixed h, the endpoint approaches the current reference. Moving the reference in the fast term to the old time would instead drag it one sample behind in that limit. The choice of what to treat implicitly includes the time at which a fast equilibrium is imposed.

Test early-time accuracy and late-time stability

For $\varepsilon = 0.02$, $h = 0.05$, $r = 2.5$. At the first step the exact value is

$$\sin(0.05) + e^{-2.5} = 0.1321.$$

Explicit Euler gives

$$1 + 0.05(1 - 50) = -1.45.$$

The implicit-explicit value is

$$[1 + 0.05 + 2.5\sin(0.05)]/3.5 = 0.3357.$$

The latter is stable but visibly overestimates the surviving transient. Its damping factor $1/3.5 = 0.286$ differs substantially from $e^{-2.5} = 0.0821$. A question about peak rate or initial settling requires shorter steps even if the late-time reference is recovered.

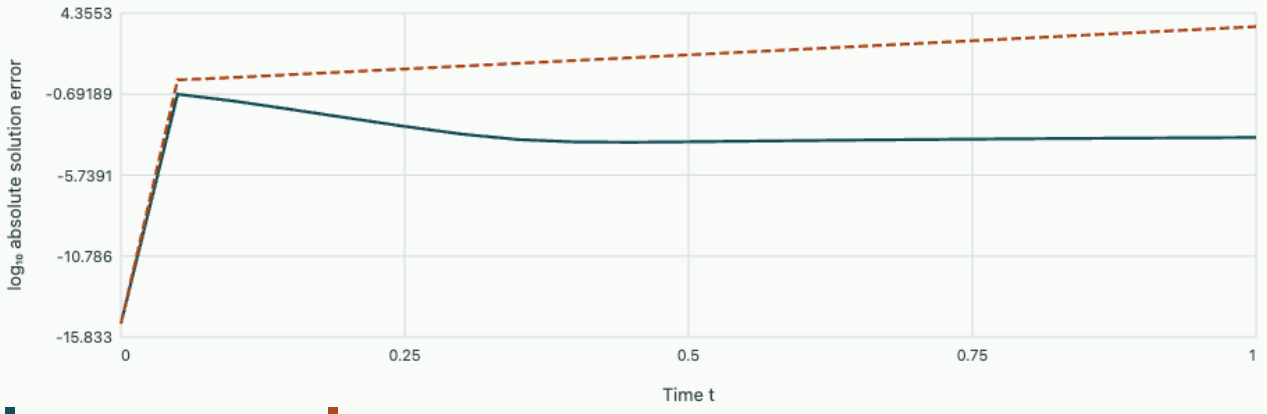
The experiment holds the final time at one while the number of steps changes. The vertical coordinate is the base-ten logarithm of absolute error, with values below 10^{-15} drawn at that floor. Negative heights therefore correspond to errors less than one. Track the maximum error separately from the final error as the transient dies out. Try an initially unstable explicit run, then double the number of steps. The stiffness criterion depends on the ratio h/ε, while accuracy on the reference also depends on h itself.

Open this experiment online

Try the calculation

Relaxation time ε **0.02**

Steps to final time 1 **20**



h/ε = 2.5. Explicit perturbation factor -1.5, implicit factor 0.2857. IMEX maximum error 0.2036, final error 0.0004054.

Tutorial on identifying which component is stiff

Keep $h = 0.04$ and reduce ε from 0.04 to 0.004. For each case compare exact, explicit and implicit amplification of a state perturbation. What quantity would force a reduction of h even with the implicit-explicit method?

▼ **Work through the solution**

At $r = 1$ the three factors are $e^{-1} = 0.368$, zero and $1/2$. At $r = 10$ they are 4.54×10^{-5}, -9 and $1/11 = 0.0909$. For a measurement of the initial decay rate the implicit discretisation substantially

misrepresents both exact factors. Bounding error during that transient requires smaller steps relative to ε, even though its numerical amplification is stable.

Further reading · Ascher, Ruuth and Wetton, implicit-explicit time discretisations.

Study the full derivation

Why stiff stability does not guarantee the expected order · Dynamics and numerical time integration

A differential equation is constructed to have a smooth, known solution and a rapidly damped error mode. Use it to distinguish stability of the homogeneous equation, tracking of a moving target and the effect of off-grid stage forcing.

Chapter 15 · Why a boundary value problem can be easier than its shooting equation

A concentration decays across a slab, with concentrations known at both faces. Computing a solution from the first face alone can bring in an exponentially growing mode. Enforcing both faces together gives a different numerical problem with the same differential equation.

Distinguish the boundary map from the physical solution

Use a nondimensional interval $0 \le x \le 1$ and write $y'' = \lambda^2 y$, with $y(0) = 1$, $y(1) = e^{-\lambda}$, $\lambda > 0$. The exact solution is $e^{-\lambda x}$, a smooth layer of thickness approximately $1/\lambda$. A shooting method prescribes a tentative slope $s = y'(0)$, integrates the initial value problem, and adjusts s to meet the second boundary. The exact initial-value solution is $\cosh(\lambda x) + (s/\lambda)\sinh(\lambda x)$. The correct slope is $-\lambda$. The small right boundary is recovered by subtracting two large nearly equal terms.

Perturb the exact slope by δs. Superposition gives the resulting error everywhere without a second initial-value solve:

$$\delta y(x) = \frac{\delta s}{\lambda}\sinh(\lambda x),$$

$$\frac{d\,y(1)}{ds} = \frac{\sinh\lambda}{\lambda}.$$

At $\lambda = 20$ the right-end amplification is 1.21×10^{7}. A slope error as small as 10^{-12} gives about 1.21×10^{-5} endpoint error, while the target value is just 2.06×10^{-9}. The actual boundary problem is much less sensitive to changes in prescribed concentrations. Its basis functions for arbitrary face values are $\sinh[\lambda(1-x)]/\sinh\lambda$ and $\sinh(\lambda x)/\sinh\lambda$, each bounded by one. A stable physical boundary problem can be paired with an ill-conditioned forward initial-value calculation. The particular derivative above measures sensitivity of the forward shooting evaluation.

Enforce the equation at the interior nodes

Take N equal intervals with $h = 1/N$. The centred second difference of any smooth exact solution equals $y''(x_j) + h^2 y^{(4)}(x_j)/12 + O(h^4)$, by Taylor expansion about x_j. Set that second difference equal to $\lambda^2 Y_j$. Incorporate the known boundary data Y_0, Y_N into the first and last right-hand sides. The interior equations are

$$-Y_{j-1} + [2 + (\lambda h)^2]Y_j - Y_{j+1} = 0,$$

$$j = 1, \dots, N-1.$$

Each row now couples neighbouring nodes and no row asks for a missing starting slope. The diagonal exceeds the sum of the off-diagonal magnitudes for $\lambda > 0$. Eliminate the lower diagonal from left to right and substitute back from the right. Both work and storage grow linearly with N. This finite-difference collocation approximates the differential equation at grid points. It still needs enough points inside the layer. The grid parameter λh measures its resolution.

At $\lambda = 5$ and just two intervals there is one unknown midpoint. Its equation gives $Y_1 = (1 + e^{-5})/(2 + 2.5^2) = 0.1220$, while the exact midpoint is

$$e^{-2.5} = 0.08208.$$

Both boundary values are satisfied exactly. Their residual alone cannot reveal the interior error. As h is halved at fixed λ, a well-resolved smooth solution exhibits second-order convergence.

Compare errors with a common physical target

The graph plots logarithmic absolute errors against position. The shooting curve isolates propagation of a controlled absolute initial slope perturbation of 10^{-12}, using the analytical propagation factor to avoid cancellation in the demonstration. The collocation curve solves the tridiagonal system numerically and compares each node to $e^{-\lambda x_j}$. Values below 10^{-16} are plotted at a floor. These are different sources of error. Increasing N improves collocation truncation error but cannot repair an error in the supplied shooting slope. At low λ the perturbed shooting solution may be much more accurate. Increase λ while keeping N fixed and identify where each method struggles.

Open this experiment online

Try the calculation

Layer parameter λ **20**

Number of mesh intervals **32**



λh = 0.625. Maximum collocation error 0.005642. The slope perturbation produces 0.00001213 right-face error.

Tutorial on a boundary tolerance

At $\lambda = 15$, require the error caused by the shooting slope alone to be less than 10^{-8} at the right face. What absolute accuracy is needed in s? Give an independent accuracy check for the collocation

solution.

▼ **Work through the solution**

Rearrangement gives

$$|\delta s| < 15 \times 10^{-8} / \sinh 15 = 9.18 \times 10^{-14}.$$

For collocation solve on a nested mesh with half the spacing and compare shared nodes. In the second-order regime the refined error is roughly one third of the coarse-to-fine difference there. Check that the estimate stabilises under another refinement and that the fine mesh resolves $1/\lambda$.

Further reading · Driscoll and Braun, finite-difference collocation for linear boundary problems · Driscoll and Braun, shooting and instability.

Study the full derivation

Solving a two-point boundary problem without guessing the start slope · Dynamics and numerical time integration

When endpoint conditions act on different sides of an interval, shooting can amplify a tiny initial error. Construct a cubic collocation system, test its continuous differential residual and quantify which boundary-value solution the residual actually certifies.

Chapter 16 · How two decaying modes produce a growing signal

Solve a triangular system exactly, derive the energy growth criterion and separate a real transient from numerical instability.

A coupling hidden by the eigenvalues

Let $y_1' = -y_1 + K y_2,\ y_2' = -2y_2$, with time in scaled units. The eigenvalues are –1 and –2 for every K. Their negative real parts guarantee asymptotic decay of this system. They do not say that $\|y(t)\|$ must decay monotonically. The second component acts as a temporary source in the first component before being damped away.

With initial state $(0, 1)$, $y_2 = e^{-2t}$. Multiply the first equation by e^t and integrate from zero to t. The result is

$$y_1 = K(e^{-t} - e^{-2t}).$$

The largest absolute excursion of the first component occurs when $2e^{-2t} = e^{-t}$, giving $t = \log 2$ and $y_1 = K/4$. At $K = 12$, a unit initial state generates a measured first component of 3.

Which initial directions grow at the initial instant?

Write $E = \|y\|_2^2/2$. Differentiation gives

$$E' = y^T S y,$$

with $S = (A + A^T)/2$. Here S has diagonal $(-1, -2)$ and off-diagonal $K/2$. A positive eigenvalue appears when the determinant $2 - K^2/4$ becomes negative. Thus some direction has positive instantaneous energy growth for $|K| > \sqrt{8}$. The specific initial vector above nevertheless has $E'(0) = -2$. Existence of a growing direction and immediate growth of every chosen initial condition are separate statements.

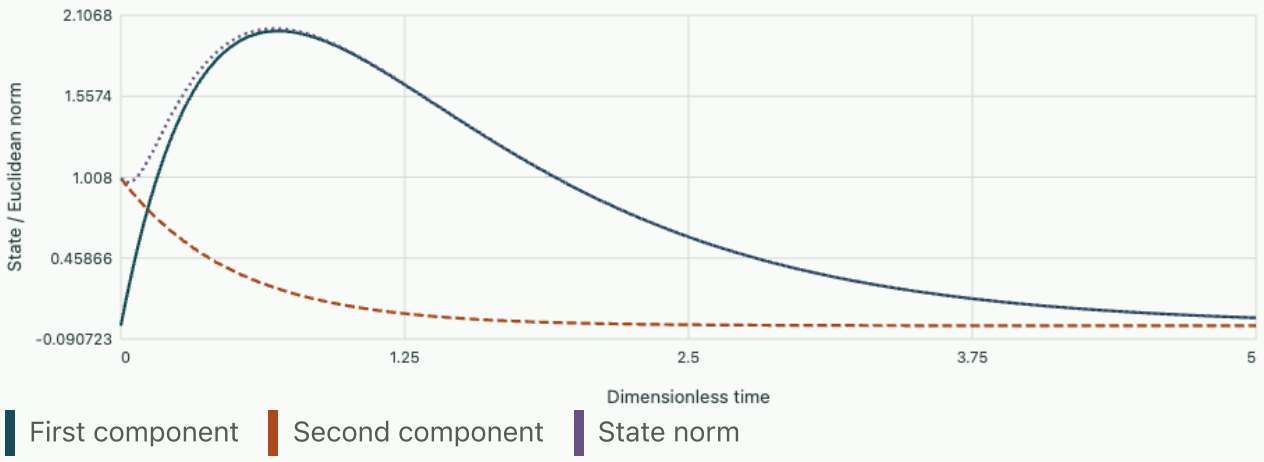
A numerical check with a false lead

Set $K = 8$. At $t = \log 2$ the exact components are $2, 1/4$. Forward Euler with a sufficiently small step will show an initial dip followed by a rise in the norm. Refining time steps changes the approximation to the peak but does not eliminate the physical growth. By contrast, using a coarse Euler step $h > 1$ puts the scalar mode –2 outside the forward-Euler stability interval and introduces genuine numerical growth.

Open this experiment online

Try the calculation

Coupling K 8



At t = ln 2 the components are (2, 0.25). Largest eigenvalue of the energy-rate matrix 2.5311.

The experiment plots the analytic components and norm. Use the signed coupling to reverse the first component. Explain why the norm is invariant under that sign change, and why a successful eigenvalue stability test alone cannot certify a bound on a short-time output.

Tutorial on transient amplification

Can a normal matrix with only negative real eigenvalues produce this Euclidean-norm amplification?

▼ **Worked solution**

A normal matrix is unitarily diagonalizable. Its evolution is a unitary change of independent decaying factors, giving a norm at most one for $t \geq 0$. Nonorthogonality of the damped directions matters here.

Change the initial state to $(1, 1)$ at $K = 6$. What is the initial energy slope?

▼ **Worked solution**

$$E' = -y_1^2 - 2y_2^2 + Ky_1y_2 = -1 - 2 + 6 = 3.$$

The instrumental first-component peak and the maximum total Euclidean norm generally occur at different times.

Chapter 17 · Locating a grazing event between time steps

Derive a sensitivity formula for event time, then measure how a harmless error in a trajectory can create or remove two threshold crossings.

A time is an implicitly defined output

Suppose an integrator advances $y' = f(t, y)$ and an event occurs at $g(t, y(t)) = 0$. Along the solution the derivative of the event function is $\dot{g} = g_t + \nabla_y g \cdot f$. A perturbation δg in the computed event function shifts a simple event time by $\delta t \simeq -\delta g / \dot{g}$. This estimate separates trajectory accuracy from event-time accuracy. At a tangency the denominator vanishes and the formula cannot be used.

> **Background · Event transversality**
>
> $$\dot{g} = \partial_t g + \nabla g \cdot f$$
>
> For a simple event this directional derivative is nonzero. Near grazing, a small state error can create a much larger error in the event time.

Resolve two crossings analytically

Take $g(t) = (t - t_0)^2 - d$, a local model for a turning point relative to a sensor threshold. For $d > 0$ the events are $t_0 \pm \sqrt{d}$. Their separation is $2\sqrt{d}$, and their slopes are $\pm 2\sqrt{d}$. For $d = 0$ there is only a tangent touch. If $d < 0$, the sensor is never triggered. An additive constant perturbation ϵ changes the effective parameter from d to $d - \epsilon$.

At $t_0 = 1$, $d = 0.01$ and $\epsilon = 0.001$, the first event moves from 0.9 to about 0.90513. The linear prediction is

$$0.001 / 0.2 = 0.005.$$

At $d = 0.0004$ the same perturbation removes both crossings entirely. Reporting more digits from a polished scalar rootfinder would not recover them.

What two endpoint samples fail to tell you

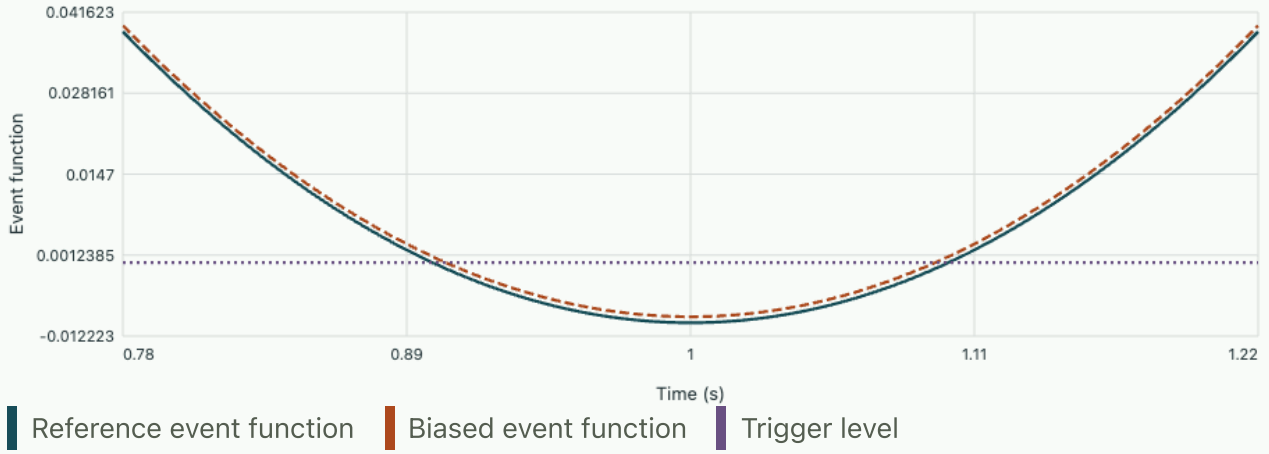
With $d = 0.0025$, a step from 0.9 to 1.1 has equal positive endpoint values of 0.0075 but crosses the threshold twice. Dense interpolation followed by an extremum check can detect this pair if the interpolant resolves the turning point. A sign-change test at accepted endpoints alone cannot. An interpolant is itself an approximation, so event feasibility should be tested against an error scale at the extremum.

Open this experiment online

Try the calculation

Threshold offset d **0.01**

Trajectory bias ε **0.001**



Reference: 0.9 s and 1.1 s. Perturbed: 0.90513 s and 1.0949 s.

Tutorial on event tolerances

For a simple crossing with slope magnitude 0.04 units/s, what error in the event function is consistent with a 1 ms time tolerance to first order?

▼ **Worked solution**

Multiply the slope by the desired time uncertainty. The event function should be accurate to about 4×10^{-5} units near the crossing.

A contact model needs the first event, even if it only touches the threshold and returns. How should the experiment be reported near grazing?

▼ **Worked solution**

Report the minimum event-function value and an estimated error on it, in addition to any time estimates. If zero lies within that uncertainty one cannot distinguish no crossing, tangency and a nearby pair from these data alone.

Study the full derivation

Find a crossing between accepted time steps · Dynamics and numerical time integration

An integrator can approximate a trajectory accurately yet report a poor switch time. Construct a continuous interpolant, relate a state error to an event-time error and inspect what happens when a trajectory merely touches a threshold.

Chapter 18 · A constrained integrator can get the circle right and the clock wrong

A frictionless bead moves on an ideal circular guide. Its exact path, speed and angular phase can all be calculated by hand. Use this unusually informative reference to derive RATTLE's two constraint corrections and see which numerical errors would escape a camera that measured position alone.

Recover the force exerted by a perfectly smooth guide

Use a circle of radius one metre, and measure mass in units of the bead mass. The position $q = (x, y)$ and velocity v are expressed in Cartesian coordinates. The guide exerts a reaction normal to its surface and no tangential force. Writing this reaction per unit mass as a scalar multiplier times q gives

$$\dot{q} = v,$$
$$\dot{v} = -\lambda q.$$

The sign convention makes positive λ point inward. The equation $q \cdot q = 1$ fixes the position but does not by itself describe which velocities are physically admissible.

Differentiate the constraint once to find

$$q \cdot v = 0.$$

Differentiate again to obtain

$$v \cdot v + q \cdot \dot{v} = 0.$$

Substituting the force yields

$$\lambda = (v \cdot v)/(q \cdot q).$$

Thus the required inward acceleration grows as the square of speed. The power supplied by the guide is $-\lambda q \cdot v = 0$, so the speed stays constant. The angular momentum per unit mass $x v_y - y v_x$ also stays constant because the guide force passes through the centre. These are independent diagnostics one can compute directly from a numerical trajectory.

Start at $q = (1, 0)$ with $v = (0, \omega)$. The exact position is $(\cos \omega t, \sin \omega t)$, the exact velocity is its derivative, and the exact angular momentum is ω in the unit-radius convention. An angle coordinate would make this example elementary. The point of the exercise is to keep Cartesian variables, as would be necessary for many connected bodies whose constraints cannot conveniently be eliminated in one global chart.

Solve for the first constraint impulse

Think of a velocity-Verlet step as an initial impulse, a drift, and a final impulse. With no external force the initial impulse is parallel to the old normal q_n. Absorb its unknown magnitude and factors of time into a dimensionless coefficient a. After the drift the candidate position must have the form

$q_{n+1} = aq_n + hv_n$. Since q_n has unit length and v_n is tangent, the norm of this expression has no cross term. Imposing the new position constraint gives a scalar quadratic equation and two possible roots.

$$a^2 + h^2\|v_n\|^2 = 1,$$

$$a = \sqrt{1 - h^2\|v_n\|^2}.$$

Choose the positive solution because the computed point must approach the old one as h goes to zero. The negative root approaches the opposite point on the circle. Its algebraic admissibility is not evidence that it represents the same initial-value problem. When $h\|v_n\| > 1$ the construction has no real solution at all. This is a geometric limit on the step, distinct from an estimate of integration accuracy. Even before reaching the limit the step can introduce a substantial angular error.

The velocity used during the drift is the chord velocity $v_{n+1/2} = (q_{n+1} - q_n)/h$. It contains a small inward term $(a - 1)q_n/h$. Expanding the square root yields

$$a = 1 - h^2\|v_n\|^2/2 + O(h^4).$$

This shows how a half-step of centripetal acceleration emerges from the constraint impulse. There was no need to insert the exact circular trajectory to obtain it.

Recover a tangent velocity at the new point

Although both ends of a chord lie on the guide, a chord velocity is generally not tangent at its destination. Apply the second normal impulse along q_{n+1}. To eliminate the normal component subtract its scalar projection, using the fact that the new normal again has unit length. This gives the velocity form of the second constraint correction.

$$v_{n+1} = v_{n+1/2} - (q_{n+1} \cdot v_{n+1/2})q_{n+1}.$$

The two impulses preserve angular momentum about the origin. The old-normal impulse makes no torque about the old position. The straight drift preserves the cross product of chord velocity with position because adding a parallel displacement does not change that cross product. The final impulse makes no torque about the new position. In this special problem tangency, unit radius and preserved angular momentum together also preserve speed. For a more general constrained Hamiltonian system, exact preservation of kinetic energy does not follow from the same argument.

Obtain the phase change directly from the new position. Its component along the old unit tangent is $h\omega$; its component along the old radius is a. For the branch chosen here the step advances the phase by $\arcsin(h\omega)$. The exact time increment advances by $h\omega$. Their difference is $h^3\omega^3/6 + O(h^5)$ for a small step. After T/h steps at fixed final time the total leading phase error is $Th^2\omega^3/6$. This is a precise second-order error prediction for an observable that energy and radius cannot diagnose.

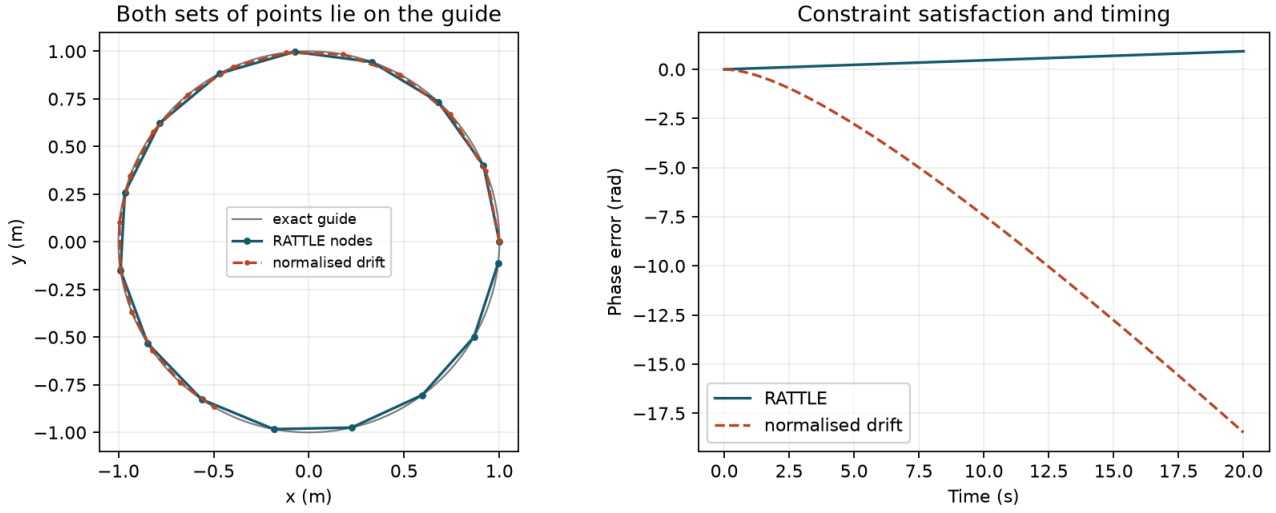
A comparison with a plausible geometric shortcut

Another algorithm proposes an unconstrained drift $q_n + h v_n$, rescales this position to the circle and then projects the old velocity onto the new tangent. Under this rule the angle change per step is $\arctan(h\|v_n\|)$ and the tangential speed changes by a factor $1/\sqrt{1 + h^2\|v_n\|^2}$. One finds the first expression by comparing the radial and tangential components before normalisation. The second follows by taking the dot product of the two successive unit tangents.

Thus this alternative delivers perfect radii and tangent velocities while steadily slowing the bead in a frictionless apparatus. Taking reciprocals of the squared-speed relation even gives

$$1/s_{n+1}^2 = 1/s_n^2 + h^2.$$

After many steps its speed can be predicted without running a single time loop. This is a useful second independent reference for the notebook. A method that projects state components separately need not realise the constraint force and its work correctly.



Before judging a circular trajectory, choose an observation time. On the circle plot a student can miss the phase displacement; the second panel reads it in radians at the same time.

With the default $\omega = 1.6\,\mathrm{s}^{-1}$ and $h = 0.25$ s, the product $h\omega$ is 0.4. The constrained step therefore advances by 0.411517 radians instead of 0.4. After 80 steps its lead is about 0.92135 radians. The analytical speed, radius and angular momentum are still reproduced to roundoff. This example makes the choice of reported output a scientific decision. A machine timed by successive crossings of a sensor would give a wrong timing signal even though a ruler would see the correct orbit radius.

Check an order of accuracy without changing the time horizon

Double the number of steps while halving their duration to retain the same final time. For the default run the lead drops from about 0.92135 to about 0.21727 radians. The quotient is a little more than four because the larger step is still influenced by higher powers of $h\omega$. Keep refining until the quotient settles near four. If instead the step were halved with the number of steps fixed, the run would end earlier and

both the elapsed time and the error per step would change. Such a comparison cannot isolate the global order at fixed final time.

An angle reported by an inverse tangent jumps from positive to negative pi across its branch cut. Differentiate or subtract such wrapped angles without care and the resulting spurious jumps can dominate a phase-error graph. Here each step moves through less than a quarter circle, so consecutive angle differences are unambiguous under phase unwrapping. Sampling a more rapidly moving system less often could reintroduce an ambiguity even with an accurate integrator internally.

A local OneSimulation installation permits another independent comparison. The optional notebook cell supplies its classical four-stage Runge–Kutta method with the continuous Cartesian acceleration obtained above. This method has no algebraic position correction; comparing its end radius and angular timing helps to distinguish the order of a differential equation solver from the constraint properties of the discrete update. For a distant endpoint one should test each diagnostic over the whole interval and repeat the comparison at a smaller step.

An additional time-reversal test can be made without knowing the exact phase. Integrate forward for a chosen number of steps, negate the final velocity and integrate the same number of positive steps. For the exact constrained update in this free problem, each second step sequence undoes the corresponding first rotation on the same branch. The final position returns to its initial value and the terminal velocity is the negative initial one. A return-to-origin test can therefore pass even when the intermediate phase differs visibly from the true physical clock. This is why a reversibility check complements, rather than replaces, a measurement at a specified observation time.

Open editable notebook Download .ipynb

Tutorial on enforcing a shape and measuring a time

1. Start at unit radius with tangential speed 1 m/s. Use a half-second constrained step to calculate the new position and the discrete phase advance.

▼ **Work through the solution**

In the chosen coordinates the new point is $(\sqrt{3}/2, 1/2)$. Its polar angle is $\pi/6 = 0.52360$ radians. The exact solution at half a second has angle 0.5 radians. A single step is already ahead by approximately 0.02360 radians.

---

2. For the same initial data and step, find the speed after four separately normalised drifts. What would a kinetic-energy sensor read?

▼ **Work through the solution**

The reciprocal squared speed starts at one and increases by 0.25 per step. After four steps the speed is $1/\sqrt{2}$ m/s. With unit mass the reported energy is 0.25 J instead of the original 0.5 J. All four reported positions still lie on the circle.

---

3. A student proposes accepting a step whenever its radius error is below 10^{-10} m. Supply a quantitative additional criterion for a phase error below 0.01 radians over ten seconds at $\omega = 1$.

▼ **Work through the solution**

The leading estimate gives $10h^2/6 < 0.01$, hence $h < 0.0775$ s approximately. Use the exact error $10[\arcsin(h) - h]/h$ or a refinement comparison for the final check; the asymptotic bound alone is not a rigorous guarantee at its edge. Tangency and torque balance remain useful complementary diagnostics.

Further reading · Hans Andersen, the RATTLE velocity algorithm for constrained motion

Change parameters · A constrained integrator can get the circle right and the clock wrong

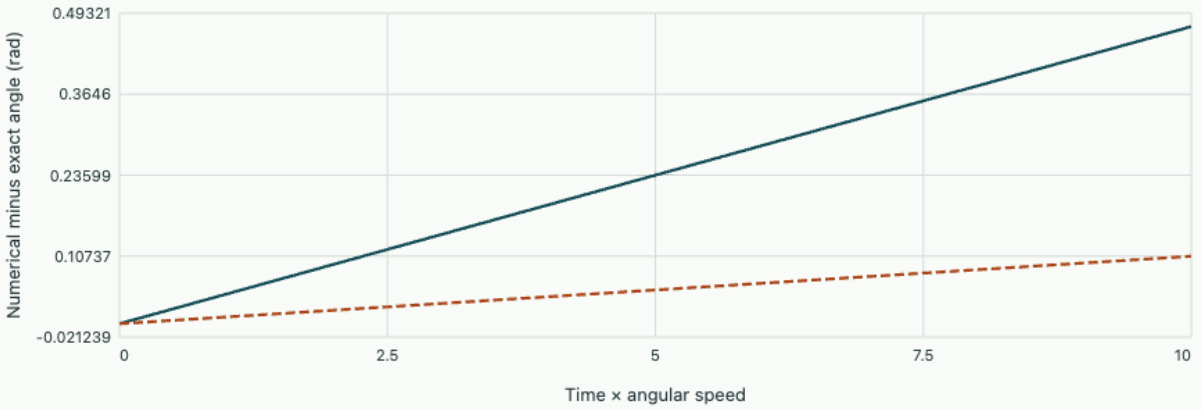
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Step × angular speed **0.5**

Steps **20**



RATTLE | Half the step, same elapsed time

The constrained radius and speed are exact, but each full step advances angle asin(hω). The full-step terminal phase lead is 0.47198 rad; no real constraint update exists when hω > 1.

Problem sheet 7 · Project · Choose a time integrator from measurements

A velocity sensor monitors an oscillator with one nearly conservative coordinate and one rapidly relaxing auxiliary variable. Design distinct numerical checks for phase, damping and the conditioning of a boundary measurement.

A stable calculation with the wrong period

The ideal oscillator has angular frequency 2 s⁻¹. Use an implicit midpoint step of 0.5 s. Derive the discrete rotation angle by acting on the complex scalar $u' = i\omega u$. Estimate how far the numerical phase lags after 20 s. Compare with the midpoint experiment.

▼ **Work through the solution**

One step multiplies by $(1 + i\omega h/2)/(1 - i\omega h/2)$. Its magnitude is one and argument $2\arctan(0.5) = 0.927295$ rad. In forty steps the exact phase advances 40 rad and the computed phase about 37.092 rad. The lag is about 2.91 rad even though the amplitude stays constant. A plot of energy alone cannot certify the period.

Relax a sensor without overdrawing the step budget

The auxiliary measurement follows $z' = -(z - u)/\tau$, with $\tau = 0.02$ s. During one short window, regard $u = 1$ as constant and start from $z = 0$. Compare forward and backward Euler and the exact value at $h = 0.1$ s. Which qualitative property matters when the measurement represents a bounded response?

▼ **Work through the solution**

Forward Euler gives $z_1 = 5$. Backward Euler gives $z_1 = (z_0 + 5u)/(1 + 5) = 5/6$. The exact value is $1 - e^{-5} \approx 0.99326$. The implicit step is bounded and monotone but its transient error is substantial. Use the split relaxation experiment to distinguish resolving a transient from staying stable after it.

Diagnose a spatial boundary sensor

A separate steady calibration solves $y'' = k^2 y$ on $[0, 1]$ with specified endpoint values. Starting a shooting solution with slope error δs at zero and exact starting value, derive the right-end error. Evaluate it at $k = 10$, $\delta s = 10^{-5}$. Why might a global boundary solve help?

▼ **Work through the solution**

The error solves the homogeneous equation with $e(0) = 0$ and $e'(0) = \delta s$. Therefore $e(1) = \delta s \sinh(k)/k \approx 0.01101$. A trial slope that looks accurate in isolation gives a visibly wrong right

boundary. A collocation matrix imposes both boundaries together; its own error and conditioning still need verification. Compare the two boundary solvers.

Design a reporting table

Name a distinct scalar check for each part of this instrument. State when you would halve the time step and when you would change the mathematical formulation.

▼ **Work through the solution**

Record accumulated phase at a fixed physical time, range and time-to-90% for the sensor response, and endpoint residual plus interior profile error for the spatial calibration. Halve the time step to test temporal truncation error. Change the boundary solver when initial-slope perturbations dominate the measurement despite a converged integrator.

Problem sheet 8 · Differentiate a bouncing trajectory through an impact

A camera records height at a fixed time after an object hits the floor. A nearby release height changes the impact time as well as the rebound speed. Compute the sensitivity that an estimator needs, and see why differentiating just the reset of velocity gives an incorrect answer even with an accurately located impact.

Distinguish a fixed observation time from an event time

A point mass falls from rest at height h under acceleration $-g$. Let the rigid floor reset its incoming velocity v^- to $v^+ = -ev^-$, where $0 < e < 1$. Derive its first impact time and the derivative of that time with respect to h. Which derivative of height vanishes at the impact?

▼ Work through the solution

Before contact $y = h - gt^2/2$, $v = -gt$. Setting $y = 0$ gives

$$\tau = \sqrt{2h/g},$$

$$d\tau/dh = 1/(g\tau).$$

The height evaluated at its own impact is $y(\tau(h), h) = 0$ for every release, so its total derivative is zero. At a fixed time before impact the partial derivative $\partial y/\partial h$ is instead one. The chain rule relates them: $1 + v^- d\tau/dh = 0$, since $v^- = -g\tau$. A camera clock held fixed during the experiment asks for a fixed-time sensitivity. A switch that records exactly at contact uses the event-time derivative. They describe different observations of the same trajectory.

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 missing timing term in the tangent update

Write $X = (y, v)$, $F = (v, -g)$, $n = (1, 0)^T$ and the reset $R(y, v) = (y, -ev)$ on the guard $y = 0$. For a state perturbation z^- compared at the nominal event time, follow it to its own event and then synchronise it with the nominal post-impact time. What matrix maps it to z^+?

▼ Work through the solution

The perturbed impact is shifted by $\delta\tau = -n^T z^- / (n^T F^-)$. At its own event the perturbation of the incoming state is $z^- + F^- \delta\tau$. Applying the reset derivative gives $DR(z^- + F^- \delta\tau)$. To compare states at a common time, advance the reset state back by $\delta\tau$ using the outgoing field, subtracting $F^+ \delta\tau$. Thus $z^+ = [DR + (F^+ - DRF^-)n^T/(n^T F^-)]z^-$. The bracket is the saltation matrix. It includes the change in impact time and depends on both vector fields. A transversal crossing, $n^T F^- \neq 0$, is required for this calculation.

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.

With $DR = \operatorname{diag}(1, -e)$ and $F^{+} = (-ev^{-}, -g)$, the matrix has first row $(-e, 0)$ and second row $(-(1+e)g/v^{-}, -e)$. The height perturbation's sign immediately after the event can seem surprising. A higher release reaches the floor later; at the nominal impact the newly rebounding higher-release trajectory has had less time to rise. The saltation matrix compares neighbouring smooth branches extended to a common time. Exactly at the discontinuity of the velocity field the full fixed-time map can be nonsmooth; the one-sided propagated tangent on either side is what the matrix supplies.

Compare the tangent to the exact rebound trajectory

Set $h = 1.00\,\mathrm{m}$, $g = 9.81\,\mathrm{m/s^2}$, $e = 0.80$, and observe at $T = 0.600\,\mathrm{s}$. Compute the height and velocity sensitivity with respect to initial height in SI units. What would a reset-only implementation predict?

▼ **Work through the solution**

The impact occurs at $0.45152\,\mathrm{s}$, leaving $\Delta = T - \tau = 0.14848\,\mathrm{s}$ of rebound. The pre-impact tangent is $(1, 0)$, so just afterwards it becomes $(-e, (1+e)/\tau)$. Free transport to T gives

$$\partial_h y = -e + (1+e)\Delta/\tau \approx -0.208,$$

$$\partial_h v = (1+e)/\tau \approx 3.99\,\mathrm{s}^{-1}.$$

The units of the first derivative are metres per metre. A naive reset-only operation leaves $(1, 0)$ unchanged and gives both the wrong sign of the position response and zero velocity response.

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.

One can check the calculation without any tangent integration. The first-rebound solution is

$$y = eg\tau(T - \tau) - g(T - \tau)^2/2,$$

$$v = eg\tau - g(T - \tau).$$

Differentiate these expressions at fixed T, remembering τ depends on h, to obtain the same two results. At a known T shortly after rebound, a measured height alone may decrease when the release height increases. After sufficient time on the same rebound branch this derivative turns positive again. An inverse problem built around a single image should check whether its measurement time has landed close to a zero of the sensitivity.

Verify the derivative without crossing an unreported event

Use the graph to compare a two-sided finite difference of full trajectories with the analytic tangent. Why does disagreement form a narrow interval around impact, and why does unlimited step reduction fail at a grazing crossing?

▼ Work through the solution

Compute $D_\eta = [y(T; h+\eta) - y(T; h-\eta)]/(2\eta)$ at the same observation time in both runs. Away from the event, both follow the same smooth sequence of modes and the difference tends to the analytic tangent as η decreases. Near impact one run can still be falling while the other is already rebounding. Its finite difference mixes two branches and rounds off the jump between one-sided derivatives. Narrowing the perturbation shrinks this region. At exactly the nonsmooth observation time a unique classical derivative need not exist. The central value there cannot be taken as evidence that the analytic tangent on either smooth side is wrong.

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.

Start with a one-percent perturbation and reduce it while monitoring a time well after the first bounce and before the second. At sufficiently small perturbations event-location tolerance and cancellation in subtraction become visible in a floating-point solver. For a nearly tangent hit the denominator $n^T F^-$ approaches zero, and small perturbations can remove or insert an event entirely. The event sequence and its transversality are part of a sensitivity report, just as step size and integrator order are part of a trajectory calculation.

Open this experiment online

Try the calculation

Release height (m) **1**

Restitution **0.8**

Height perturbation (%) **1**



Event-aware one-sided tangent Centred trajectory difference

The nominal impact is at 0.4515 s. At the fixed camera time 0.6005 s the height sensitivity is -0.2060 m/m. For these parameters both perturbed trajectories stay on or before the first rebound throughout the plot.

Nathan Kong and colleagues derive saltation matrices for hybrid dynamics and discuss event timing.

Notebook practice · A constrained integrator can get the circle right and the clock wrong

Dynamics and numerical time integration

Turn a differential model into a time step. Track accuracy, stability and structural properties separately in a computable physical or mathematical example.

1. ## Find a crossing between accepted time steps

   An integrator can approximate a trajectory accurately yet report a poor switch time. Construct a continuous interpolant, relate a state error to an event-time error and inspect what happens when a trajectory merely touches a threshold.

2. ## A delayed feedback model needs a history, not one starting value

   A controller responds to the measured displacement one time delay ago. Construct its solution in successive intervals, track propagated derivative discontinuities and derive the point at which increasing delay can destabilise negative feedback.

3. ## Solving a two-point boundary problem without guessing the start slope

   When endpoint conditions act on different sides of an interval, shooting can amplify a tiny initial error. Construct a cubic collocation system, test its continuous differential residual and quantify which boundary-value solution the residual actually certifies.

4. ## Why stiff stability does not guarantee the expected order

   A differential equation is constructed to have a smooth, known solution and a rapidly damped error mode. Use it to distinguish stability of the homogeneous equation, tracking of a moving target and the effect of off-grid stage forcing.

5. ## How far does an energy-preserving rotation lag behind?

   Integrate uniform circular motion by the implicit midpoint rule. The points stay exactly on the energy circle in exact arithmetic. Derive the angle advanced at each step and measure the accumulated timing error.

6. ## From existence of a trajectory to the order of a time step

   Establish what an initial-value solver approximates, derive Runge–Kutta order conditions for a general smooth vector field and construct an alternative high-order step by extrapolation.

7. ## Why consistent multistep formulas can fail under refinement

   Derive the root condition for constant steps, propagate a local defect to a fixed end time and isolate a growing parasitic mode with a controlled starting error.

8. ## Why an accurate position can hide an inaccurate constraint force

   Derive an energy test for implicit Runge–Kutta stages, then recover missing derivatives in a constrained model and calculate the accuracy lost by a multiplier.

Dynamics and numerical time integration · 1. Find a crossing between accepted time steps

An integrator can approximate a trajectory accurately yet report a poor switch time. Construct a continuous interpolant, relate a state error to an event-time error and inspect what happens when a trajectory merely touches a threshold.

An event is a root of a composite function

Let $y' = f(t, y)$ and let a sensor or change of dynamics be described by $g(t, y) = 0$. Along an exact trajectory the scalar event function is $h(t) = g(t, y(t))$. A simple event at t_* satisfies

$$\dot{h}_* = g_t + g_y f \neq 0.$$

If approximate trajectory or sensor evaluation introduces error δh, linearisation around the crossing gives

$$\delta t \simeq -\delta h / \dot{h}_*.$$

A small state error can therefore imply a much larger time error if the crossing is nearly tangential. Reporting an event timestamp to machine precision after bisecting a low-order interpolant is not equivalent to knowing the physical crossing time that accurately.

Consider a dimensionless unit-frequency oscillator with $y = \cos t,\ y' = -\sin t$ during the first half cycle. Its first falling crossing of threshold $a \in (0, 1)$ occurs at $t_* = \arccos a$, with slope $-\sqrt{1 - a^2}$. At $a = 0.5$, a positive vertical interpolation error 10^{-3} can delay the crossing by about 0.00115 time units. At $a = 0.9999$ the same linear estimate is about 0.0707, comparable with or larger than the event time itself; the simple linear local approximation then also fails. At precisely $a = 1$ the surface is tangent at the maximum and sign-change tests can miss it altogether.

Construct a state and slope interpolant on one accepted interval

Suppose samples of state and derivative are available at $t_n, t_{n+1} = t_n + h$. Write $s = (t - t_n)/h$ and choose a cubic $p(s)$ matching $p(0) = y_n, p(1) = y_{n+1}, p_s(0) = h f_n, p_s(1) = h f_{n+1}$. Solving these four linear conditions gives

$$p = (2s^3 - 3s^2 + 1) y_n + (s^3 - 2s^2 + s) h f_n + (-2s^3 + 3s^2) y_{n+1} + (s^3 - s^2) h f_{n+1}.$$

For an exact scalar trajectory sufficiently smooth in time, the Hermite error is $y(t) - p(t) = y^{(4)}(\xi)(t - t_n)^2 (t - t_{n+1})^2 / 4!$, bounded by $h^4 \max |y^{(4)}| / 384$. Linear interpolation instead has error bounded by $h^2 \max |y''| / 8$. Combine either bound with a known lower bound on $|\dot{h}|$ in a bracket to bound event-time error. In an actual ODE solve, y_n, f_n are numerical, so interpolation and integrated state errors both contribute. A high-order interpolant cannot repair incorrect endpoint data or sample a discontinuity assumed smooth.

Bracket on the physical trajectory and restart after a transition

At each accepted interval compare event values at both ends. For a simple crossing of the requested direction, solve $g(t, p(t)) = 0$ inside that interval using a bracketed root finder. The bracket controls the algebraic root-finding error; independent time-step refinement measures trajectory error. If the dynamics change at the event, truncate the old step there, apply any specified reset $y^{+} = R(t_{*}, y^{-})$ and restart the integrator under the new right side. Propagating the old smooth interpolant through a change of dynamics can corrupt both state and subsequent event ordering.

If a surface stays exactly active over a finite interval, repeated detection at the same time can stall the integration. A model of contact, hysteresis or directional rearming may be physically appropriate, depending on what the event means. Merely nudging the time forward by an arbitrary tolerance can skip a nearby second event. Similarly two opposite crossings within one large accepted interval may yield equal endpoint signs. A maximum step guided by known physical timescales, supplementary extrema detection or an analytically enclosed event function can be needed. Accuracy control on the smooth state alone does not guarantee detection of every specified surface crossing.

Compare interpolation errors independently of the propagation error

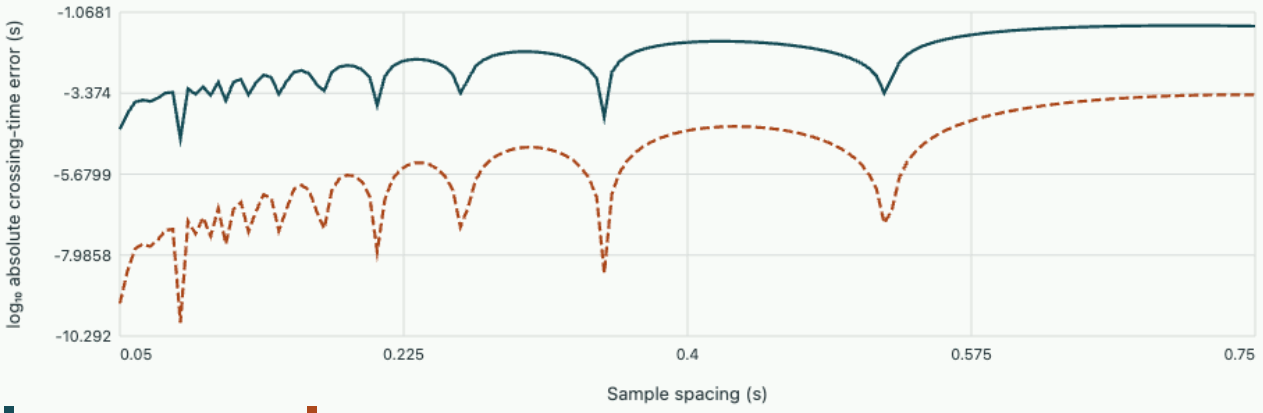
The calculator uses exact cosine endpoint samples and their derivatives, then locates the first falling crossing on a uniform time grid by linear or cubic Hermite interpolation. The curves show error against step spacing on a logarithmic vertical scale. Their irregular minima occur when a grid point happens to coincide with the event. Look for the error envelope over several spacings. At threshold 0.5 the exact result is 1.047198 s. Moving the threshold changes both the crossing slope and its position relative to the sampled grid.

Open this experiment online

Try the calculation

Crossing level **0.5**

Inspection step (s) **0.4**



Exact crossing 1.047198 s. At the selected step: linear 1.035331 s, Hermite 1.047161 s. Exact endpoints isolate interpolation error.

Distinguish uncertainty in the threshold from time discretisation

Two clocks can agree on an event and still report a poorly determined arrival time if the physical threshold is uncertain. For a calibrated level a the falling crossing solves $\cos t_* = a$, so implicit differentiation gives

$$dt_*/da = -1/\sqrt{1-a^2}.$$

A level error of 0.002 near $a = 0.5$ produces a timing uncertainty of about 0.00231 s for the unit-frequency benchmark. At $a = 0.9$ the same device gives 0.00459 s. Close to the turning point the trace is almost horizontal. Reducing the root finder's tolerance cannot remove that physical loss in sensitivity.

If the numerical trajectory also has a known absolute state error bound η, divide that bound by a lower bound on the crossing speed in the bracket. The result is a deterministic timing allowance. An independent random threshold error has a standard deviation and calls for a statistical interpretation. Adding those numbers in quadrature would need a probability model for both. Reporting the two separately preserves what the measurement and the solver actually established.

Tutorial · assess a reported event time

Get a crossing from just two samples

Use $y = \cos t,\ a = 0.5$ with samples at $t = 0.8, 1.2$. Compute the linearly interpolated time and its error. Why does its sign follow from curvature?

▼ **Work through the solution**

The endpoint values are 0.696707 and 0.362358; the fractional position is $(0.696707 - 0.5)/(0.696707 - 0.362358) = 0.5883$, giving $t_L \simeq 1.0353$. It is about 0.0119 s early. Over this interval $\cos t > 0$, so $y'' = -\cos t < 0$. A straight chord lies below the concave curve; the falling chord reaches the threshold before the true trajectory.

Bound a Hermite crossing-time error

On the same interval use $|y^{(4)}| \le 1,\ |\dot{y}| \ge \sin 0.8$. Give a simple guaranteed bound assuming exact endpoints and exact solution of the interpolating cubic root.

▼ **Work through the solution**

The state interpolation bound is

$$0.4^4/384 = 6.67 \times 10^{-5}.$$

Dividing by $\sin 0.8 = 0.7174$ gives 9.29×10^{-5} s. This uses monotonicity of the exact trajectory to convert a vertical defect at the computed root into a horizontal difference. It does not assume the cubic derivative equals the true derivative. Numerical endpoint error would need to be added in state units first.

Distinguish a tangency from a missed sign change

A gap sensor follows $h(t) = (t-1)^2 - \delta$. Compare event detection at $\delta = 0, 10^{-4}, -10^{-4}$ over the interval $[0.9, 1.1]$.

▼ **Work through the solution**

Both endpoints are positive in all three cases. At zero there is a repeated touching event. At positive δ there are two true crossings at 0.99 and 1.01 with opposite directions. At negative δ there is no root. Locating and evaluating the internal minimum distinguishes the cases when the uncertainty is below $|\delta|$. A sign test at just the endpoints cannot determine any of these distinctions, however accurately it evaluates those endpoints.

Further reading

- ETH Zürich · numerical methods for CSE, chapter 11 on numerical time integration
- SciPy solve_ivp · continuous output and sign-change event detection
- Course investigation · adaptive time stepping
- Compare time-step error with conservation

Dynamics and numerical time integration · 2. A delayed feedback model needs a history, not one starting value

A controller responds to the measured displacement one time delay ago. Construct its solution in successive intervals, track propagated derivative discontinuities and derive the point at which increasing delay can destabilise negative feedback.

Specify the additional data an ordinary initial value omits

For $t \geq 0$ consider $y'(t) = -ay(t-\tau)$ with positive constant gain a and delay $\tau > 0$. The derivative at the start depends on $y(-\tau)$; an initial value $y(0)$ alone is insufficient. Supply a history $y(t) = \phi(t)$ on $[-\tau, 0]$, continuous at zero for a continuous state. Two preparations sharing $y(0)$ but different earlier values can immediately develop different slopes. This is why treating a delay as a new algebraic coefficient in an ODE solver without a history buffer changes the physical problem.

Take constant history $\phi = 1$. For $0 \leq t \leq \tau$, the delayed input is one and $y = 1 - at$. For $\tau \leq t \leq 2\tau$, the derivative is $-a[1 - a(t-\tau)]$; integrate from $t = \tau$ to get

$$y = 1 - at + a^2(t-\tau)^2/2.$$

On the next interval another integral gives

$$y = 1 - at + a^2(t-\tau)^2/2 - a^3(t-2\tau)^3/6.$$

Each new interval solves an ordinary forced equation with known earlier right-side data. This is the method of steps. A smooth-looking initial history can nevertheless mismatch derivatives: here the left derivative at zero is zero while the equation supplies $-a$ on the right.

Trace how a step mismatch propagates through the delay

At $t = \tau$ continuity of y at zero makes y' continuous, but $y''(t) = -ay'(t-\tau)$ inherits the earlier jump in y'. At 2τ the discontinuity appears in the third derivative. High-order adaptive ODE error formulas assume smooth derivatives within an interval. Marking known breakpoints and integrating separately across the delay intervals avoids applying a high-degree interpolant through an unacknowledged loss of smoothness. With several delays, sums of delay lengths can generate additional event locations and their bookkeeping matters. A discontinuity in the physical forcing can also generate new fronts of reduced regularity.

Within a method-of-steps solver, values $y(t-\tau)$ usually lie between stored previous mesh points. Dense output supplies their interpolation. Its error feeds directly into the next interval's right side and can limit global accuracy even when the ODE time-step rule has higher formal order. Maintain enough history to cover the largest delay and requested outputs, and preserve the known one-sided behaviour at breakpoints. If a trial step is longer than the delay its stages may refer to the current unknown interval; an algorithm must either constrain that step or solve the resulting within-step interpolation problem consistently. A generic lookup of the last stored value instead creates an unintended zero-order-held feedback law.

Background · Can smooth data prevent the first derivative jump?

A constant history $\phi(t) = 1$ has zero left derivative at the start. The equation $y' = -a\,y(t-\tau)$ demands right derivative $-a$, creating an immediate mismatch. A history with $\phi(-\tau) = 1$ and $\phi'(0^-) = -a$ instead matches the first derivative. For example, choose $\phi(t) = 1 - a(t+\tau)$ on the history interval.

Matching a first derivative does not guarantee a smooth second derivative. Differentiating the equation once shows that the next matching condition would be

$$\phi''(0^-) = -a\,\phi'(-\tau^+).$$

For the linear history, the two sides are zero and a^2. The mismatch has moved up by one derivative. A high-order interpolant across the start may therefore still lose order even though the trajectory and its tangent look continuous.

Derive a delay-induced change of oscillatory stability

For a constant-coefficient retarded equation try $y = e^{st}$. Substitution gives

$$s + ae^{-s\tau} = 0.$$

Unlike a polynomial characteristic equation, this transcendental equation generally has infinitely many roots. A pair crosses the imaginary axis when $s = i\omega$, giving $a\cos(\omega\tau) = 0,\ \omega - a\sin(\omega\tau) = 0$. For positive ω, the first crossing is at $\omega = a,\ a\tau = \pi/2$. With negligible delay the model behaves like the stable ODE $y' = -ay$. The crossing indicates that negative feedback acting on old information can become destabilising; stability of the zero-delay ODE is no guarantee for arbitrary delay.

To check crossing direction, differentiate $s + ae^{-s\tau} = 0$ at constant a, obtaining $ds/d\tau = -s^2/(1+\tau s)$ after using the characteristic equation. At $s = ia$ its real part is $a^2/[1+(a\tau)^2] > 0$. Increasing delay moves this pair into the unstable half-plane. A numerically observed sign-changing transient alone does not prove instability, as stable delay systems can oscillate while decaying. Compare an envelope over successive periods with the characteristic roots and refine both time discretisation and history interpolation. An unstable numerical discretisation can otherwise imitate a feedback-induced growth.

Compare a stored history with the method-of-steps solution

The delay is the time unit and the history on the preceding interval is one. The calculator integrates an explicit Euler update with an integer number of steps per delay so it can read the delayed state exactly at a stored grid index. The comparison curve evaluates the exact piecewise polynomial. Its kth contribution starts after k minus one delay intervals and has coefficient given by the kth power of minus

the dimensionless feedback divided by k factorial. This follows by repeatedly integrating the previous interval's polynomial, so it is independent of the Euler recurrence.

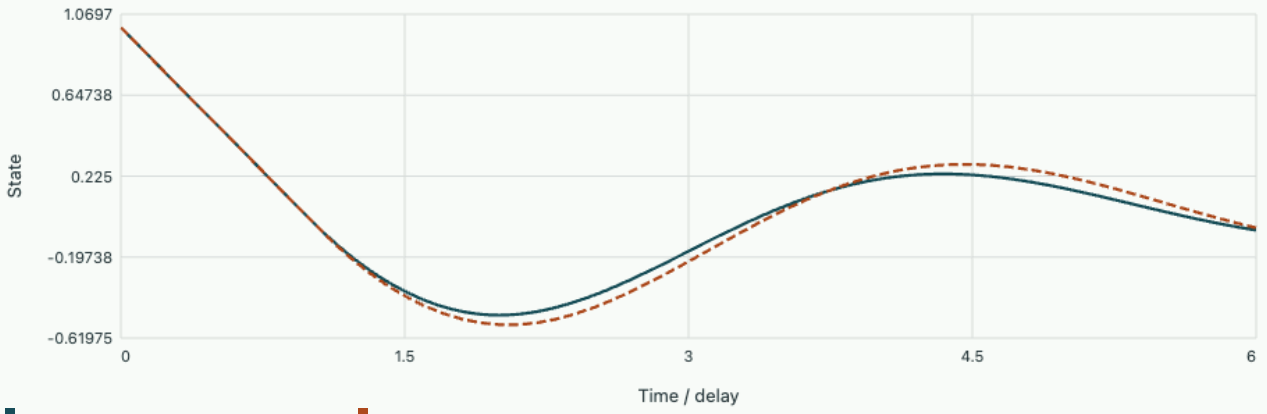
At feedback one, inspect the exact values after one, two and three delays before refining the grid. Euler already integrates the constant derivative on the first interval exactly, while errors enter on later intervals. Change the feedback across π/2 and compare the oscillation envelope. Six displayed delays are useful for checking the calculation but can be too short to measure slow growth close to threshold. A stable-looking short record does not replace the characteristic-root analysis. Because the grids here align with the delay, this experiment isolates stepping error and does not test a general off-grid history interpolator.

Open this experiment online

Try the calculation

Dimensionless feedback aτ **1**

Steps per delay **10**



Largest grid error 0.0642. First continuum instability at aτ = π/2 ≈ 1.571; current aτ = 1.00.

Tutorial · distinguish history dependence from numerical error

Compute three exact method-of-steps values

For $a = 1\,\mathrm{s}^{-1}$, $\tau = 1$ s and constant unit history, find $y(1), y(2), y(3)$. Does a negative response at two seconds prove instability?

▼ **Work through the solution**

The first value is zero. The second is

$$1 - 2 + 1/2 = -0.5.$$

The third is

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

A negative value reflects delayed feedback overshoot. With $a\tau = 1 < \pi/2$ the zero solution is asymptotically stable for this model even though the transient crosses zero. The exact interval polynomials provide stringent tests for a history interpolator at the first few breakpoints.

Construct incompatible histories with the same initial value

Keep $a = \tau = 1$ and compare $\phi_1(t) = 1$ with $\phi_2(t) = 1 + t$ on $[-1, 0]$. Obtain the distinct states on the first interval.

▼ Work through the solution

For the first history $y_1 = 1 - t$. For the second, $\phi_2(t-1) = t$, so $y_2' = -t$ and $y_2 = 1 - t^2/2$. At $t = 1$ their values are 0 and 0.5 although both began at 1. Their initial right derivatives differ by one. A restart record containing only time and last state cannot reconstruct the interrupted integration; retaining the relevant past function is part of the state representation.

Compare physical and discrete stability boundaries

A controller has $a = 4\,\mathrm{s}^{-1}$. Estimate the first critical delay and the oscillation period there. What numerical checks would separate it from a coarse-step instability?

▼ Work through the solution

The first critical delay is

$$\pi/(2a) = 0.3927\,\mathrm{s}.$$

Its frequency is $a/(2\pi) = 0.6366$ Hz and period $2\pi/a = 1.571$ s. Change the delay on both sides of the predicted value, resolve the oscillation with multiple step sizes and compare measured growth with the characteristic equation. Keep history interpolation accurate while refining; fixed low-order delayed lookup can determine the observed stability error even as the main integrator step is reduced.

Further reading

- Shampine and Thompson · Solving DDEs in MATLAB
- SciML · a worked calculation with a history function and the method of steps
- MATLAB dde23 · history functions and derivative discontinuities
- Course investigation · multistep integration and starting histories

- Compare time-step error with conservation

Dynamics and numerical time integration · 3. Solving a two-point boundary problem without guessing the start slope

When endpoint conditions act on different sides of an interval, shooting can amplify a tiny initial error. Construct a cubic collocation system, test its continuous differential residual and quantify which boundary-value solution the residual actually certifies.

Sensitivity of shooting need not be sensitivity of the physical solution

Take $y'' = \kappa^2 y$ on $0 \le x \le 1$, with $y(0) = 1,\ y(1) = 0$. The solution $y = \sinh[\kappa(1-x)]/\sinh\kappa$ decays from the left. To shoot from the left, introduce unknown slope s, yielding endpoint $y(1) = \cosh\kappa + (s/\kappa)\sinh\kappa$. The correct slope is $-\kappa\coth\kappa$ and its error is magnified by $\sinh\kappa/\kappa$ at the far end. For $\kappa = 20$ this factor is about 1.21×10^7. Even an initial slope accurate to 10^{-10} can cause a far-end mismatch of order 10^{-3}. The mathematical boundary solution can remain well behaved while this particular representation suffers cancellation of large growing and decaying components.

A collocation boundary solver instead treats approximate states at many locations as simultaneous unknowns and enforces both boundary conditions directly. Write the equation as first-order system $Y' = (v, \kappa^2 y)^T = F(x, Y)$, $Y = (y, v)$. On each panel $[x_i, x_{i+1}]$ choose a cubic Hermite polynomial using unknown endpoint states and endpoint slopes fixed by F_i, F_{i+1}. This already imposes the differential equation at both endpoints. Require the polynomial derivative to satisfy the ODE also at the midpoint. Matching yields two equations for each panel, with the boundary equations closing the overall algebraic system.

Derive a residual with only three right-side evaluations per panel

For panel width h, Hermite interpolation at the midpoint gives

$$Y_m = (Y_i + Y_{i+1})/2 + h(F_i - F_{i+1})/8.$$

Its derivative there is $p'_m = 3(Y_{i+1} - Y_i)/(2h) - (F_i + F_{i+1})/4$. Equating this with $F_m = F(x_i + h/2, Y_m)$ yields

$$Y_{i+1} - Y_i - \frac{h}{6}(F_i + 4F_m + F_{i+1}) = 0.$$

This is a three-point Lobatto collocation relation, resembling Simpson quadrature with a midpoint state determined by the same cubic. Neighbouring panels share endpoint states and slopes, constructing a continuously differentiable piecewise approximation for a smooth right side. The nonlinear algebraic Jacobian has local block structure apart from boundary conditions and a few global unknowns. Dense finite differencing of the entire state would discard that structure and can dominate a large adaptive solve.

As a simple independent test take the scalar ODE $y' = \lambda y$. With $z = h\lambda$, substitution of Y_m yields a one-panel transfer $R(z) = (1 + z/2 + z^2/12)/(1 - z/2 + z^2/12)$, agreeing with e^z through z^4. The symmetry $R(-z) = 1/R(z)$ checks the derivation. At large negative z, $R(z) \to 1$, so the local solution formula is not strongly damping for unresolved stiff modes. Global enforcement of boundary conditions and adaptive spatial resolution remain important: a stable algebraic solve need not resolve a very thin physical layer.

Interpret a continuous residual through a stability estimate

A collocation residual vanishes at the points where it was imposed, so checking only those values tests the nonlinear algebraic solve, not the differential approximation between points. Evaluate the continuous defect $d(x) = p'(x) - F(x, p(x))$ at independent points or use its suitably weighted panel norm. Refine where the residual is large, especially where the solution changes rapidly, and enforce the boundary conditions independently. A mesh with many uniform far-interior points can inefficiently resolve a solution having a layer of thickness $1/\kappa$ near a single wall.

For a scalar linear elliptic boundary problem $-y'' + \kappa^2 y = f$ with zero endpoint error, an approximation with smooth enough derivatives has error governed by $-e'' + \kappa^2 e = r$. If $\|r\|_\infty \le \epsilon$, a maximum-principle argument gives $\|e\|_\infty \le \epsilon/\kappa^2$ for $\kappa > 0$: at an interior positive error maximum, $-e'' \ge 0$, and similarly for the negative minimum. This bound links a differential residual in the same formulation to solution error under the stated boundary conditions. A residual in a differently scaled first-order system requires its own inverse stability estimate. Near a nonunique boundary solution an apparently tiny defect can correspond to a much larger state error.

Reconstruct the entire cubic boundary solution

The calculator solves the same linear collocation equations as the written panel derivation. For this constant matrix, its growing and shrinking eigendirections decouple and one panel applies the stated rational transfer to each. The numerical field shown between nodes is the actual cubic Hermite interpolant through those collocation states and their slopes. The reference is the independently evaluated hyperbolic-sine boundary solution. Increase the layer parameter at fixed panel count first and look just inside the left boundary. Both curves hit the imposed endpoints even on a visibly inadequate mesh.

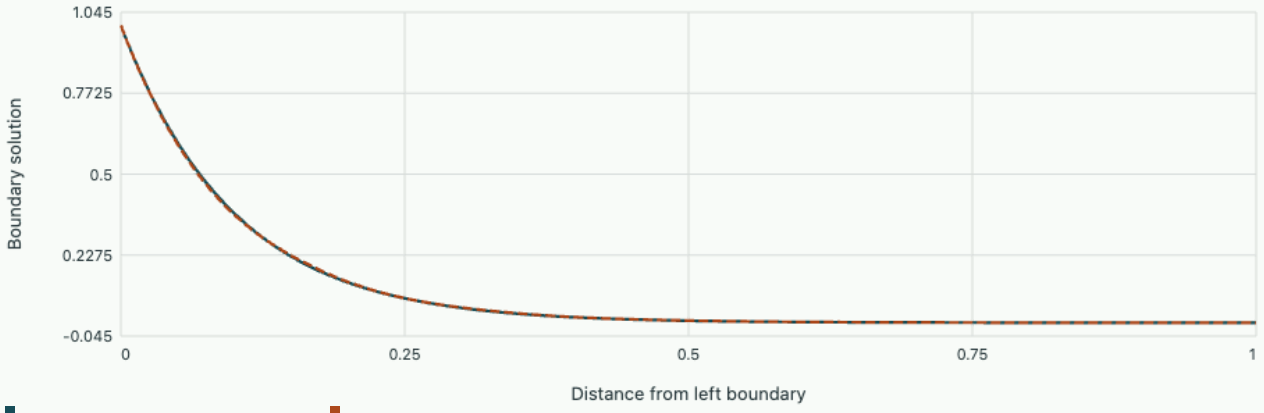
Then double the number of panels. The displayed maximum difference is measured on 241 independent evaluation locations and is a sampled diagnostic, not a guaranteed global norm. For large layer parameter the shooting endpoint amplification can be enormous even while the collocation solution becomes accurate upon mesh refinement. At only a few cells across the layer the cubic may bend below zero. A solution satisfying the nonlinear algebraic conditions of the chosen discretisation need not preserve positivity of the continuum field at every point.

Open this experiment online

Try the calculation

Layer parameter κ **10**

Uniform panels **6**



Exact boundary layer Cubic Lobatto solution

Sampled maximum field error 0.00674; shooting endpoint amplification sinh(κ)/κ = 1101.

Tutorial · distinguish mesh, boundary and algebraic errors

Check a cubic against an exactly representable solution

Choose $y' = 3x^2,\ y(0) = 0$ on a single unit panel. Verify the collocation residual using the candidate $y = x^3$ and identify the midpoint state from endpoint data.

▼ **Work through the solution**

Endpoint values are 0 and 1, slopes 0 and 3. The Hermite midpoint value is $1/2 - (3/8) = 1/8$, equal to $(1/2)^3$. The midpoint slope is $3/4$. Simpson's right side gives $[0 + 4(3/4) + 3]/6 = 1$, equal to the endpoint difference. Here the entire state polynomial, not just its grid values, is exact. This manufactured case checks signs and factors before using the solver on a stiff boundary layer.

Quantify endpoint sensitivity for one shooting trial

At $\kappa = 10$, an independent integration has a starting-slope error 2×10^{-7}. Estimate its endpoint error. Would measuring that endpoint residual accurately imply the underlying boundary problem is ill conditioned?

▼ **Work through the solution**

The amplification is $\sinh 10/10 \simeq 1101.3$, yielding endpoint error 2.20×10^{-4}. This exposes sensitivity of a one-way initial-value representation to the guessed slope. A boundary perturbation

instead propagates through bounded functions $\sinh[\kappa(1-x)]/\sinh\kappa$ and $\sinh(\kappa x)/\sinh\kappa$. Its effect on interior y need not contain the huge shooting factor. Distinguish sensitivity of the original boundary data from cancellation in one numerical strategy.

Spot a false residual certificate

A computation checks only $p' - F = 0$ at each panel's two ends and midpoint and reports error below 10^{-12}. What evidence is still needed for a claimed 10^{-6} physical solution tolerance?

▼ **Work through the solution**

Those are exactly the collocation equations; their tiny residual tests the solved discrete system. Evaluate the defect off those nodes or estimate its norm by an independent quadrature, verify boundary mismatches, and relate these defects to solution error using a suitable stability bound or independent reference refinement. For the second-order model with $\kappa = 5$ and zero endpoint error a uniform residual below 25×10^{-6} in $-p'' + 25p - f$ suffices for the stated magnitude bound. A reported first-order-system residual with different units cannot be inserted unchanged.

Further reading

- Kierzenka and Shampine · a boundary-value solver controlling residual and error
- MATLAB bvp4c · three-stage Lobatto collocation
- Course investigation · sensitivity of boundary-value shooting
- Compare time-step error with conservation

Dynamics and numerical time integration · 4. Why stiff stability does not guarantee the expected order

A differential equation is constructed to have a smooth, known solution and a rapidly damped error mode. Use it to distinguish stability of the homogeneous equation, tracking of a moving target and the effect of off-grid stage forcing.

Choose an exact solution before increasing stiffness

Let $\lambda < 0$, let $\phi(t)$ be a smooth specified function, and solve the Prothero–Robinson model $y' = \lambda[y - \phi(t)] + \phi'(t)$, $y(0) = \phi(0)$. The exact solution is $y(t) = \phi(t)$ for every λ. For an initial mismatch e_0, the error evolves $e(t) = e_0 e^{\lambda t}$. Thus the observed slow solution and fast relaxation timescale can be varied independently. This makes the model an incisive verification case for a stiff integrator: changing λ leaves the exact waveform fixed while altering how errors created by discrete stages are damped.

A stability function $R(h\lambda)$ describes propagation of errors for the homogeneous test $y' = \lambda y$. A-stability constrains $|R(z)| \leq 1$ over the left half-plane. L-stability additionally requires $R(z) \to 0$ as $|z| \to \infty$ in the stable stiff limit. Neither definition alone describes how a method evaluates a time-dependent forcing at its internal stages. Stage order and stiff accuracy can matter when those stage errors are multiplied by large $|\lambda|$. Observing bounded output on a homogeneous damping test does not establish a method's full formal order on a forced stiff problem.

Calculate the error recurrence of backward Euler and the trapezoidal rule

Write $e_n = y_n - \phi_n,\ z = h\lambda$. Backward Euler evaluates the whole right side at t_{n+1}. Collect terms after subtracting the exact target:

$$(1 - z)e_{n+1} = e_n + \phi_n - \phi_{n+1} + h\phi'_{n+1}.$$

The forcing defect begins at $h^2\phi''_{n+1}/2$. Its effect on the next error is divided by $1 - z$. The rapidly decaying homogeneous mode is also divided by that factor. This can make backward Euler track a particular smooth imposed target surprisingly well in the very stiff regime even though its classical nonstiff order is one. Its behaviour should be assessed in the desired observables and step–stiffness range, rather than ranked solely by the formal method label.

For the trapezoidal rule the corresponding equation is

$$(1 - z/2)e_{n+1} = (1 + z/2)e_n + \phi_n - \phi_{n+1} + \tfrac{h}{2}(\phi'_n + \phi'_{n+1}).$$

The defect is the trapezoidal quadrature error in the derivative and begins at order h^3. However, its homogeneous factor tends to -1 when $z \to -\infty$, so a pre-existing fast error can survive as a sign-alternating numerical component for many steps. With $z = -100$ the factors are $1/101 = 0.00990$ for backward Euler and $-49/51 = -0.96078$ for trapezoid. Both schemes are A-stable. Only the first

immediately damps this unresolved mode. Event resets, inconsistent initial preparation or changing external loads can seed it.

A midpoint constraint can force endpoint drift

Implicit midpoint uses $y_{n+1} - y_n = h[\lambda((y_n + y_{n+1})/2 - \phi_m) + \phi'_m]$. For fixed h and large negative λ, finite numerical states require the stage average to approach ϕ_m. Thus $y_{n+1} \simeq 2\phi_m - y_n$. The endpoints inferred from this interior constraint need not land on ϕ_{n+1}. Compare with trapezoid, which shares exactly the same scalar homogeneous stability function but evaluates two different times and states for nonlinear or time-dependent forcing.

With $y_n = \phi_n$ and $\phi(t) = t^2$, the limiting endpoint error is

$$2\phi(t_n + h/2) - \phi_n - \phi_{n+1} = -h^2/2.$$

The exact finite-z midpoint error starting from the target is $zh^2/[4(1 - z/2)]$. The trapezoidal derivative quadrature, in contrast, is exact for this quadratic target and keeps $e_n = 0$ at every step started exactly on it. Their common stability function did not predict their different forced response. In the midpoint example an alternating error can cancel at some output times, so inspecting only the final point at an even number of steps may obscure the largest per-step error. Compare all sampled times and separate propagated from newly injected defects.

Separate injected forcing error from a starting mismatch

This experiment uses the target t squared with step 0.1 s. It propagates the exact error recurrences of backward Euler, trapezoid and midpoint rather than estimating the difference by subtracting two large trajectories. The reference also includes the exponentially decaying initial mismatch. Move the stiffness control until a time step contains a hundred fast relaxation times. With a nonzero initial error, trapezoid and midpoint preserve a visible alternating component even though the exact mismatch essentially disappears after one step.

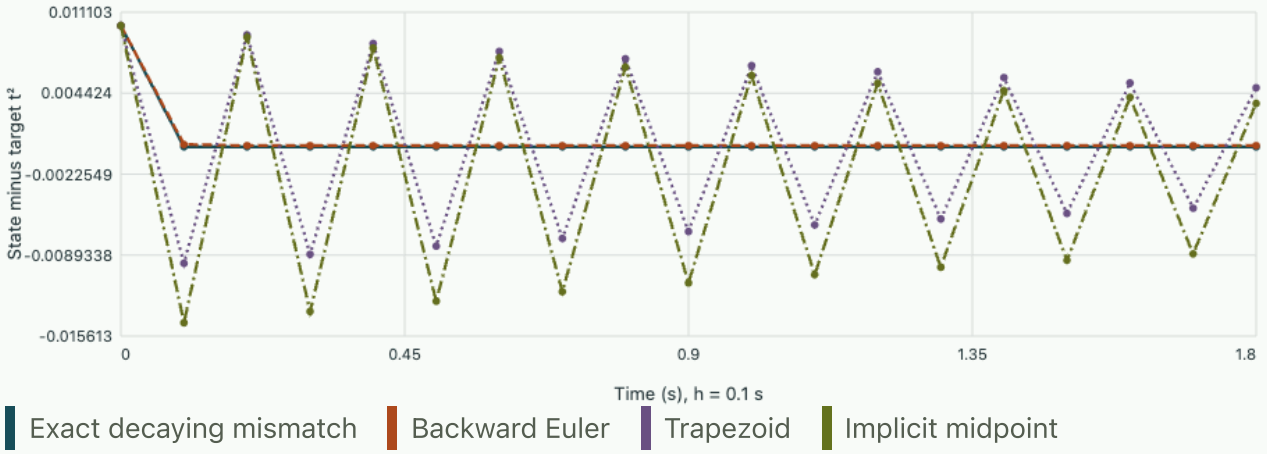
Set the initial mismatch to zero to isolate what an initially accurate trajectory gains from stage forcing. For this one quadratic target the trapezoidal derivative quadrature is exact and introduces no error, while midpoint still misses the endpoint because it constrains the value at the middle of the interval. Try a less stiff value and compare the peak and endpoint errors over all times. The demonstration is deliberately a manufactured exact target; substituting a cubic target would remove trapezoid’s special cancellation while keeping the same homogeneous stability function.

Open this experiment online

Try the calculation

log₁₀ of |hλ| 2

Initial mismatch 0.01



Exact decaying mismatch Backward Euler Trapezoid Implicit midpoint

hλ = -100, homogeneous midpoint/trapezoid factor -0.9608, backward Euler factor 0.0099. Initial mismatch is shown relative to the target.

Tutorial · compare error mechanisms at fixed work

Compute a forced stiff step exactly

Set $h = 0.10, \lambda = -1000, \phi = t^2,\ y_0 = 0$. Compute errors after the first step for backward Euler, trapezoid and midpoint.

▼ **Work through the solution**

Here $z = -100$, $\phi(h) = 0.01$. Backward Euler defect is h^2, giving $e_1 = 0.01/101 = 9.90 \times 10^{-5}$. Trapezoid yields zero error on this target. Midpoint gives $zh^2/[4(1 - z/2)] = -0.004902$, so its predicted y_1 is only 0.005098. The exact target is smooth in all three experiments; the fast stiffness multiplies the way its curvature is sampled within the step.

Check an unresolved initial-layer perturbation

Keep $z = -100$ and introduce an initial error of 0.01. How large is the propagated part after two steps, excluding new forcing defects?

▼ **Work through the solution**

The exact propagated part is $0.01e^{-200}$, negligible here. Backward Euler gives

$$0.01/101^2 = 9.80 \times 10^{-7}.$$

Both trapezoid and implicit midpoint give $0.01(49/51)^2 = 0.00923$ with restored positive sign after the second step. Negative sign at alternate steps can be invisible if the code saves every second state. Monitoring stage and consecutive-step output helps identify this persistence.

Design a refinement that distinguishes ordinary order from stiff forcing error

Why are a study at fixed λ with $h \to 0$ and one at fixed $h\lambda = -100$ different experiments? What should be plotted?

▼ **Work through the solution**

At fixed λ, sufficiently small h eventually resolves the fast timescale and standard smooth-solution order can reappear. Holding $h\lambda = -100$ changes the differential equation's error-relaxation scale at every refinement, deliberately keeping that fast mode unresolved while the exact target stays fixed. Plot maximum error over all physical sample times, the homogeneous amplification and representative stage defects for both sequences. Report the exact initial mismatch as well as ϕ. Otherwise a measured order can conflate target tracking, unremoved fast transients and cancellation at one favourable output time.

Further reading

- ETH Zürich · numerical methods for CSE, time integrators and stability
- Prothero and Robinson · stability and accuracy of one-step methods for stiff systems
- Course investigation · stiff numerical stability
- Compare time-step error with conservation

Dynamics and numerical time integration · 5. How far does an energy-preserving rotation lag behind?

Integrate uniform circular motion by the implicit midpoint rule. The points stay exactly on the energy circle in exact arithmetic. Derive the angle advanced at each step and measure the accumulated timing error.

From a midpoint balance to a rotation matrix

Write a two-component oscillator as $v' = \omega J v$, where J sends (x, y) to $(-y, x)$. Choosing scaled coordinates makes both components comparable; for a mechanical oscillator one may take displacement and a suitably scaled speed. The generator is skew-symmetric, so $d(v^T v)/dt = \omega v^T (J^T + J) v = 0$. The exact motion has angular advance $h\omega$ over an interval h.

Midpoint approximates the velocity at the midpoint of the segment joining the old and new states. This yields $(I - aJ)v_{n+1} = (I + aJ)v_n$ with $a = h\omega/2$. Since $J^2 = -I$, multiplication by $I + aJ$ eliminates the inverse and gives

$$R_h = (I - aJ)^{-1}(I + aJ) = \frac{(1 - a^2)I + 2aJ}{1 + a^2}.$$

The coefficients of I and J have squares summing to one. They are the cosine and sine of the angle $\theta_h = 2 \arctan a$. Thus every update preserves the squared radius. The angle differs from the exact $h\omega$. Expanding the arctangent gives

$$\theta_h = h\omega - (h\omega)^3/12 + O((h\omega)^5).$$

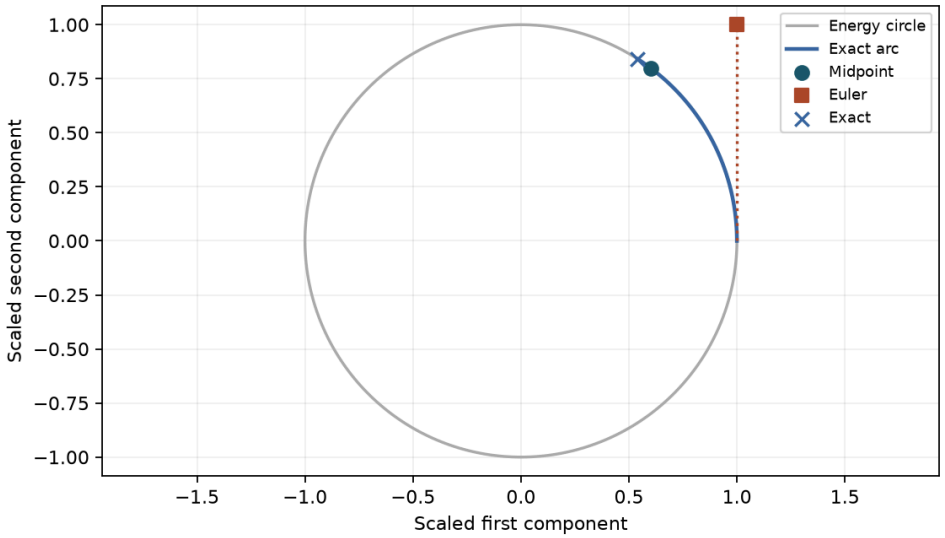
Over a fixed duration $T = nh$, the leading accumulated angle error is $-T\omega^3 h^2/12$. This phase lag is especially visible after many periods even when a graph of energy looks perfect.

For comparison, explicit Euler takes the tangent line to its end point. Its amplification is $\sqrt{1 + (h\omega)^2}$. Studying both amplitude and timing identifies different defects of the two updates.

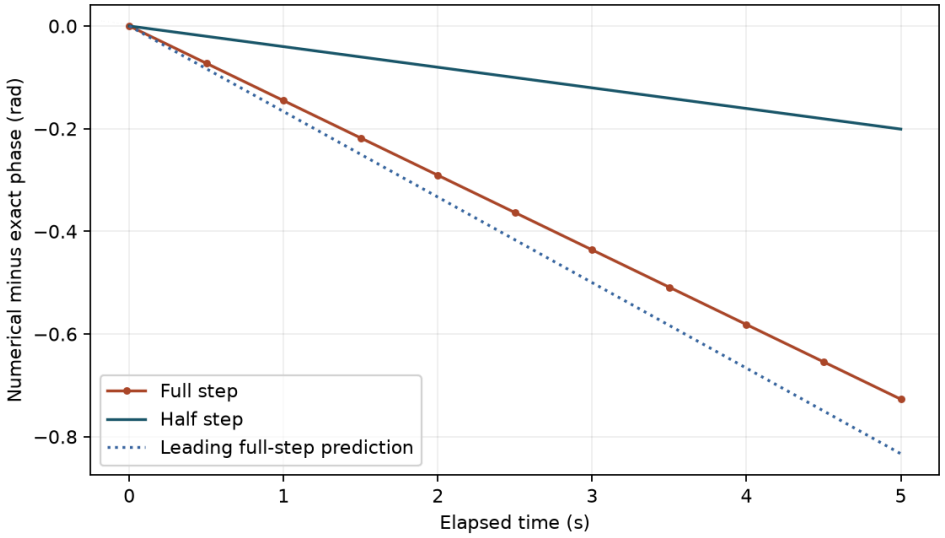
Put a clock on the unit circle

Let $\omega = 2$ rad/s, $h = 0.5$ s and $v_0 = (1, 0)$. The half-step parameter is 0.5. The rotation sends the first point to $(0.6, 0.8)$ instead of the exact $(\cos 1, \sin 1) \approx (0.5403, 0.8415)$. Both radii are one. The angular deficit is $1 - 2\arctan(0.5) \approx 0.072705$ rad per update. After ten updates, the numerical clock is behind by 0.72705 rad or about 0.364 s of physical time.

With $h = 0.25$ s over the same five seconds the deficit is $10 - 40 \arctan(0.25) \approx 0.20085$ rad. Its ratio to the coarse-step error is close to one quarter. In the notebook the rotation is built by a linear solve and its angle is measured from a dot product and an oriented area. These are independent ways to check the hand formula. Counting full turns requires retaining unwrapped increments; a final point alone cannot identify the number of completed turns.



The three markers correspond to one physical time step. The midpoint result lies on the circle at its own numerical angle.



The full and half steps cover the same physical interval. The dotted trace is the leading asymptotic error for the coarse step.

Exercise · reverse the sign of time

Show why a forward midpoint update followed by one equal backward update returns to the starting state. Does forward Euler have the same property here?

▼ Work through the solution

Replacing a by $-a$ interchanges the two factors in the Cayley formula, so $R_{-h} = R_h^{-1}$. Their product is the identity. For explicit Euler the product is $(I - h\omega J)(I + h\omega J) = (1 + (h\omega)^2)I$. The reversal increases every radius when the angular speed and the step are nonzero.

Exercise · set a phase budget

For $\omega = 2$ rad/s, approximate a time step that keeps the accumulated phase lag below 0.02 rad over six seconds. Give an exact way to check the choice.

▼ **Work through the solution**

The leading magnitude is $T\omega^3 h^2/12 = 4h^2$. Equating this to 0.02 gives about 0.0707 s. Choosing 90 equal steps over six seconds gives $h = 1/15$ s. The exact lag is $12 - 180\arctan(1/15)$ rad, about 0.0177. The exact check is useful when the step is large enough that the cubic series term has limited accuracy.

Change parameters · Explore this notebook model

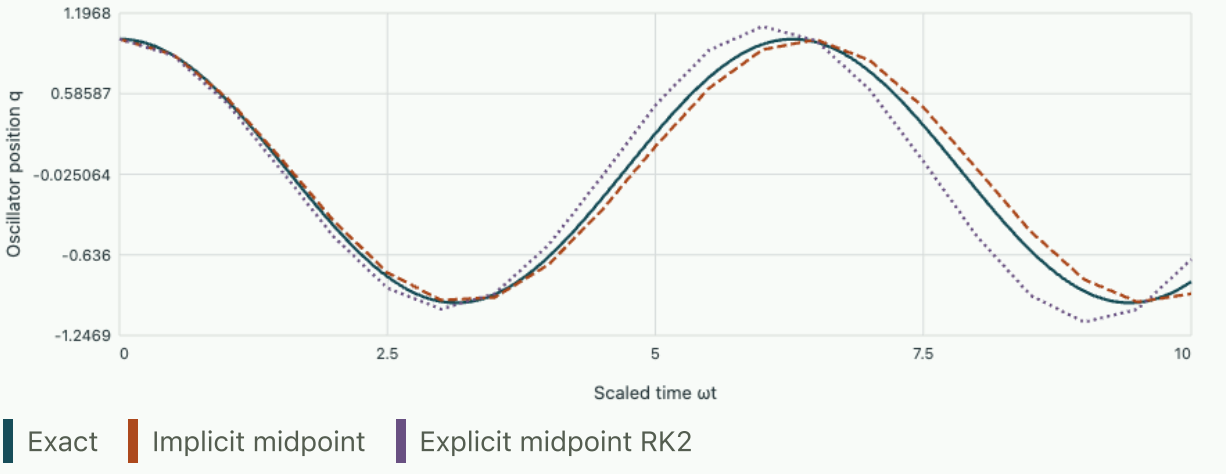
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Time step h (ω = 1) **0.5**

Number of forward steps **20**



Implicit midpoint phase lag 0.2009 rad. Forward-backward distance from the start is 8.951e-16 for implicit midpoint and 0.3635 for explicit midpoint.

Open editable notebook Download .ipynb



Dynamics and numerical time integration · 6. From existence of a trajectory to the order of a time step

Establish what an initial-value solver approximates, derive Runge–Kutta order conditions for a general smooth vector field and construct an alternative high-order step by extrapolation.

Construct the trajectory on a short interval

Let $f(t, y)$ be continuous on a time–state cylinder, bounded there by M and Lipschitz in y with constant L. For initial value y_0, seek a continuous curve in the closed sup-norm ball of radius R around y_0. On an interval of length H, apply

$$(Tz)(t) = y_0 + \int_{t_0}^{t} f(s, z(s))\, ds.$$

If $HM \le R$ and $HL < 1$, T maps the ball into itself and is a contraction. Iteration from the constant curve changes by at most HL times the preceding change. The geometric series of increments converges uniformly and the limit satisfies the integral equation. Differentiation gives the ODE. Two such limits have distance at most HL times their distance, hence coincide. Overlapping short cylinders extend the trajectory until continuation fails. For $y' = y^2,\ y(0) = 1$, the explicit $1/(1-t)$ illustrates why local existence gives no promise at $t = 1$.

The perturbed integral inequality $e(t) \le e_0 + \int_0^t (Le(s) + d(s))ds$ leads, by an integrating factor for the right-hand side, to $e(t) \le e^{Lt}e_0 + \int_0^t e^{L(t-s)}d(s)ds$. It explains why a finite-time global error bound for a numerical solver needs an appropriate domain and a perturbation-growth estimate.

Collect elementary derivatives without losing branching terms

For an autonomous smooth f, the first three elementary differentials are $f,\ f'f,\ f''(f,f)$, together with the third-order chain $f'(f'f)$. Represent a derivative acting on child differentials by a node with one branch per argument. A single node has order and density one. Recursively, a tree $\tau = [\tau_1, \dots, \tau_k]$ has $|\tau| = 1 + \sum |\tau_j|$ and density $\gamma(\tau) = |\tau| \prod \gamma(\tau_j)$. This factor follows by integrating the power of time once when a root is added. Multiplicities of identical branches give the same symmetry factor on the exact and numerical sides.

Define stage weights $\Phi_i(\bullet) = 1$. Taylor expanding each RK stage about the initial state replaces an integral for every child by a weighted stage sum. Consequently

$$\Phi_i([\tau_1, \dots, \tau_k]) = \prod_{\ell=1}^{k} \sum_j a_{ij} \Phi_j(\tau_\ell),$$

$$\sum_i b_i \Phi_i(\tau) = 1/\gamma(\tau).$$

Matching every tree through order p gives classical order p for general sufficiently smooth vector fields. The local defect is order $p+1$, yielding global order p under the perturbation bound above. For the

fourth-order tree consisting of a root, a child, and two leaves on that child, the condition is $b^T A c^2 = 1/12$, with $c = A\mathbf{1}$ and the square componentwise. For a root with three leaves it is $b^T c^3 = 1/4$. A scalar linear test equation supplies only chains.

Select accuracy by spending more substeps

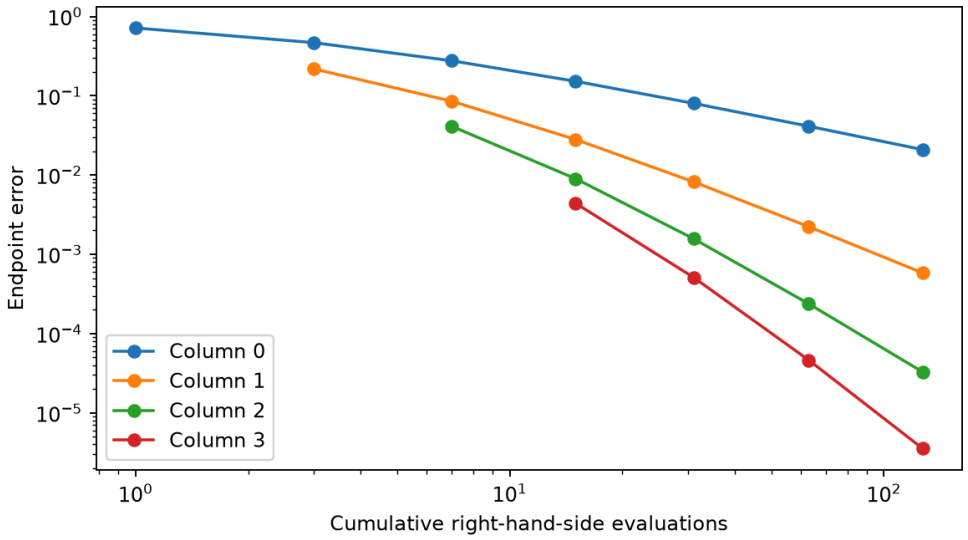
For a smooth problem over one macrostep H, let $T_{j,0}$ be its endpoint from $N_j = 2^j$ Euler substeps. At fixed H, assume the endpoint error has an asymptotic expansion in powers of $1/N_j$. This follows for smooth solutions by expanding the local error and propagating its coefficients. The leading term halves on doubling. Eliminate successive powers by

$$T_{j,k} = T_{j,k-1} + \frac{T_{j,k-1} - T_{j-1,k-1}}{2^k - 1}.$$

For $y' = y,\ y(0) = 1,\ H = 0.2$, the one- and two-substep estimates are 1.2 and 1.21, so the extrapolate is 1.22, close to $e^{0.2}$. Four substeps give 1.21550625. Extrapolating once gives 1.2210125 and twice gives 1.22135. Use the separation between successive columns to estimate the lower-order local error. For a column of global order $k + 1$, a local tolerance suggests a trial next H proportional to the tolerance ratio raised to $1/(k + 2)$ if using that column's error estimate. Compare accepted advance per function evaluation at candidate orders. Restart the expansion after a discontinuity and impose the stability restriction of the underlying method.

When can a correct extrapolation tableau be misleading?

An unresolved event or a nonsmooth forcing invalidates the common power expansion across grids. Agreement among columns can then mean that every grid missed the same feature. Split at a known event and test the reconstruction against the defect in the integral equation.



Each column cancels another inverse power of the substep count. Work includes all Euler subintegrations computed through the current row. The reference at time one is exp(1).

Tutorial on existence and order

Bound the Picard iteration for a linear cooling model

▼ **Work through the solution**

With $f = -2y + 1,\ y_0 = 1,\ R = 1, |f| \le 3$ and $L = 2$ on $[0, 2]$. Taking $H = 1/4$ gives $HM = 3/4$ and contraction ratio at most $1/2$. The sum of all remaining differences after an iterate is bounded by the last difference times $(1/2)/(1 - 1/2)$.

Distinguish two fourth-order calculations experimentally

▼ **Work through the solution**

A linear exponential tests the chain $b^T A^2 c$ and is blind to f''. To excite branching terms choose a smooth nonlinear vector field and compare to an independent fine solution at successively halved step sizes. A nonlinear test can reveal an incorrect coefficient, but the general order proof needs every independent tree condition.

Hairer, Nørsett and Wanner · nonstiff problems

Change parameters · From existence of a trajectory to the order of a time step

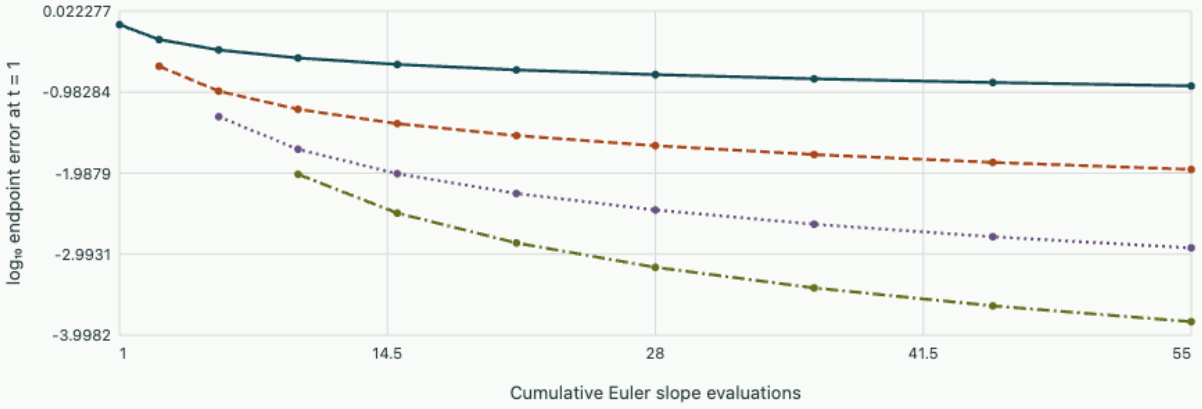
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Growth rate λ 1

Extrapolation columns 4



The exact endpoint is exp(λ) = 2.7183. Each Euler subdivision has an expansion in 1/n; extrapolation cancels successive terms while reusing the computed table.

Dynamics and numerical time integration · 7. Why consistent multistep formulas can fail under refinement

Derive the root condition for constant steps, propagate a local defect to a fixed end time and isolate a growing parasitic mode with a controlled starting error.

Turn the recurrence into a propagation matrix

Write a k-step scheme as $\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}$ with $\alpha_k = 1$. Let $\rho(z) = \sum \alpha_j z^j$ and $\sigma(z) = \sum \beta_j z^j$. Substituting exact constant and linear solutions shows consistency requires $\rho(1) = 0,\ \rho'(1) = \sigma(1)$. For continuous differentiable exact solutions these conditions make the defect divided by h tend to zero. Higher derivatives determine the local order.

For $f = 0$ stack the latest k errors. Their propagation matrix is the companion matrix C of ρ. A root r of multiplicity s contributes modes $n^{\ell} r^n,\ 0 \le \ell < s$ for nonzero roots. All powers of C stay bounded exactly when all characteristic roots lie in the closed unit disk and unit-modulus roots are simple. Interior repeated roots are harmless because exponential decay dominates their polynomial factors. This is null stability for constant step sequences.

Sum the defects over a fixed time

Let a convergent starter have maximum error E_0, the local discrete defect be $h\,\tau_n$, and f be Lipschitz with constant L on the solutions' domain. Repeated substitution using the bounded C^j first propagates the starter and each past perturbation. The Lipschitz terms contribute a sum of errors. For an implicit step take h sufficiently small to absorb its newest error term. With constants independent of h,

$$E_n \le C_1(E_0 + T \max |\tau_j|) + C_2 h \sum_{i<n} E_i \le C_1(E_0 + T \max |\tau_j|) \exp(C_2 T).$$

Thus consistency and null stability imply convergence. Conversely, convergence for arbitrary convergent starting values on smooth problems forces bounded homogeneous propagation. If a root lies outside the disk, a vanishing initial perturbation can grow exponentially over T/h steps. For a repeated unit root choose an initial perturbation proportional to h^{s-1} in the generalised mode so that its final error stays finite and nonzero. Consistency is forced by testing constants and linear-in-time solutions under convergence for all equations, taking an irreducible multistep formula and the usual normalisation. This equivalence concerns fixed-order, constant-step schemes. Changing step ratios changes the propagation matrices.

Control the perturbation before changing the grid

For $f = 0$, take $\rho = (z-1)(z-r)$, with $r = -1$ or $r = 2$. Choose $\sigma = 1 - r$ to make the associated formulas consistent. Initialise $y_0 = 1, y_1 = 1 + h^3$. The exact propagated endpoint error at $N = 1/h$ is

$$e_N = h^3(1 - r^N)/(1 - r).$$

For $r = -1$ every error stays $O(h^3)$. For $r = 2$ the growing factor eventually defeats every algebraic power. Exact starting data with no perturbations would hide this instability in the constant solution.

Transfer the experiment to the decaying state $y' = -y$ by choosing $\sigma(z) = (1 - r)z$. Now the middle coefficient in the recurrence is $1 + r - h(1 - r)$. A OneSimulation RK4 trajectory supplies a common starter and a separate full-step trajectory checked against e^{-1}. Inject h^3 into the starter and compare with the identical multistep solve without that injection. The difference isolates amplification along the recurrence.

OneSimulation experiment · Watch a vanishing starting error grow

Vary the number of steps and the injected starting error. Compare both perturbation amplification and the error of the reference trajectory.

Download the executable notebook · Python source

Adjust the experiment in the browser. Vary the parasitic root, the step count and the injected starting error independently. The first step is the RK4 polynomial for $y' = -y$; subsequent steps use the two-step recurrence above. The browser follows the perturbation in its own recurrence so that subtractive cancellation does not hide a tiny change. Compare $r = -1, 0, 1, 2$. At $r = 1$, the repeated root and vanishing differential term make this a degenerate, unstable formula. Hold the perturbation fixed to inspect amplification; then lower it by a factor of eight when doubling the number of steps to reproduce an h^3 perturbation. The notebook uses OneSimulation RK4 for its starter and reference trajectory.

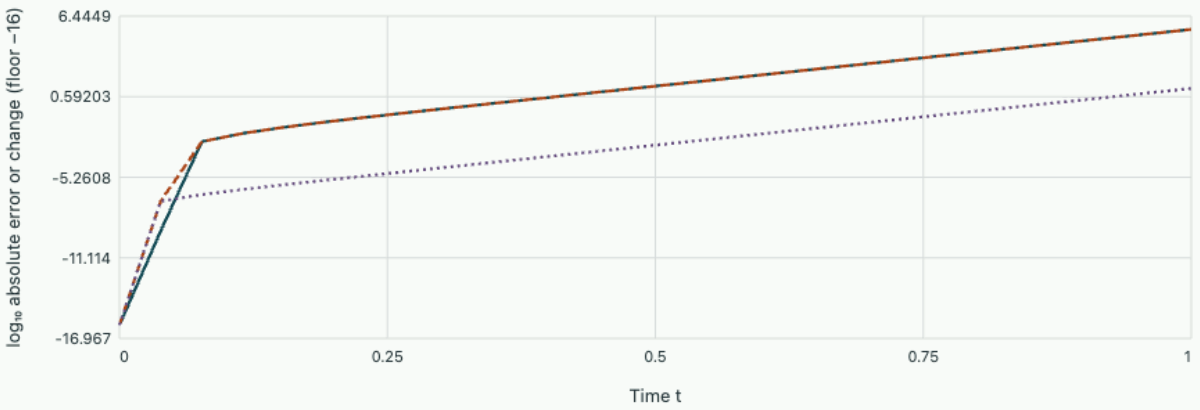
Open the live experiment

Try the calculation

Parasitic root r **2**

Steps to t = 1 **26**

$\log_{10}$ starting perturbation **-7**



Unperturbed discretisation error

Error with perturbed start

Propagated starting change

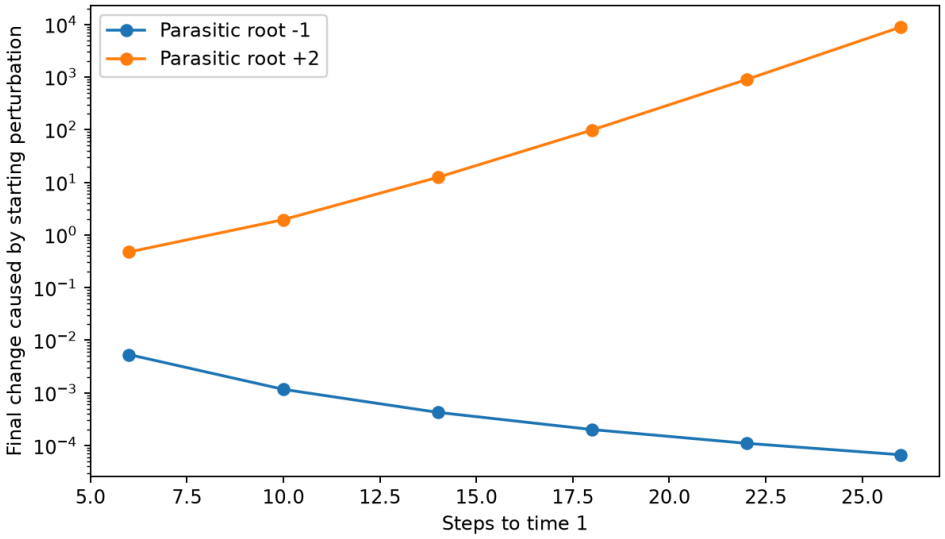
The characteristic roots at h = 0 are 1 and 2. The parasitic root lies outside the unit circle. A starting change of 1e-7 produces 15.6 change at t = 1. The error-change recurrence is evaluated separately to avoid subtracting two nearly equal trajectories.

```
"""Resolve a decaying signal with two consistent families of two-step methods."""
import matplotlib.pyplot as plt
import numpy as np
from onesim.numerics.time import ClassicalRungeKutta4TimeIntegrator

def decay(state, time):
    return (-state[0],)

def endpoint(parasitic, starter, h, n):
    previous, current = 1.0, starter
    for _ in range(1, n):
        previous, current = current, (1+parasitic-h*(1-parasitic))*current-parasitic*previous
    return current

method = ClassicalRungeKutta4TimeIntegrator()
counts = [6, 10, 14, 18, 22, 26]
fig, ax = plt.subplots(figsize=(7, 4), layout="constrained")
for parasitic in [-1, 2]:
    errors = []
    for n in counts:
        h = 1/n
        reference = (1.0,)
        starter = method.advance(reference, decay, h, time=0.)[0]
        for j in range(n):
            reference = method.advance(reference, decay, h, time=j*h)
        unperturbed = endpoint(parasitic, starter, h, n)
        perturbed = endpoint(parasitic, starter+h**3, h, n)
        errors.append(abs(perturbed-unperturbed))
```

```
        assert abs(reference[0] - np.exp(-1)) < 1e-5
    ax.semilogy(counts, errors, "o-", label=f"Parasitic root {parasitic:+d}")
    print(f"Root {parasitic:+d}: change at T=1 with 26 steps: {errors[-1]:.6g}")
print(f"Reference RK4 error with 26 steps: {abs(reference[0]-np.exp(-1)):.3g}")
ax.set(xlabel="Steps to time 1", ylabel="Final change caused by starting perturbation")
ax.legend()
fig.savefig("numerics-parasitic-root.png", dpi=170)
```

Output from the reference run

```
Root -1: change at T=1 with 26 steps: 6.6793e-05
Root +2: change at T=1 with 26 steps: 8875.58
Reference RK4 error with 26 steps: 6.93e-09
```



Terminal errors for two recurrences with distinct parasitic roots.

Tutorial on multiplicity and error order

Can two roots at one be tolerated if every local defect is tiny?

▼ **Work through the solution**

For $\rho = (z-1)^2$ and $f = 0$, choose $e_0 = 0, e_1 = h$. Differences stay equal to h, so $e_N = Nh = T$. The local residual is exactly zero while the starter converges and the endpoint fails to converge.

Separate the order of the starter from that of the method

▼ **Work through the solution**

A null-stable order-three formula has $\tau = O(h^3)$. If its starter has only $O(h^2)$ accuracy, the bound at fixed T becomes $O(h^2)$. Raising the order of the multistep formula alone cannot recover the lost order.

Hairer, Nørsett and Wanner · multistep methods

Dynamics and numerical time integration · 8. Why an accurate position can hide an inaccurate constraint force

Derive an energy test for implicit Runge–Kutta stages, then recover missing derivatives in a constrained model and calculate the accuracy lost by a multiplier.

Derive a nonlinear contractivity test

Suppose $\langle u - v, f(u) - f(v) \rangle \leq 0$ for all relevant states. For two trajectories advanced by the same RK step, use D_i for each stage-state difference and g_i for the stage-slope differences. Substitute the stage identity $D_i = d_0 + h \sum_j a_{ij} g_j$ into the square of $d_1 = d_0 + h \sum_i b_i g_i$. Collecting symmetric products gives

$$\|d_1\|^2 - \|d_0\|^2 = 2h \sum_i b_i \langle D_i, g_i \rangle - h^2 \sum_{ij} (b_i a_{ij} + b_j a_{ji} - b_i b_j) \langle g_i, g_j \rangle.$$

If $b_i \geq 0$ and the displayed symmetric coefficient matrix is positive semidefinite, the step cannot increase distances, assuming its nonlinear stage equations have a solution. This sufficient algebraic stability condition applies to nonlinear dissipative fields. For one-stage midpoint $a = 1/2, b = 1$ the matrix is zero. Backward Euler $a = b = 1$ gives one. Both are contractive, but their scalar amplification limits as $h\lambda \to -\infty$ are -1 and 0 respectively. Only the latter damps an unresolved fast mode completely.

Recover the derivatives needed to see the multiplier

For index one, $y' = f(y, z),\ 0 = g(y, z)$ with g_z invertible, the implicit function theorem eliminates z. Perturbing the algebraic constraint at fixed y by ϵ gives $\delta z = O(\epsilon / \sigma_{\min}(g_z))$. A mechanical position constraint has a different structure:

$$q' = v,$$
$$M v' = F(q, v) - G(q)^T \lambda,$$
$$g(q) = 0,$$
$$G = g_q.$$

Differentiating once gives $Gv = 0$. Differentiating again gives $Gv' + g_{qq}[v, v] = 0$. For positive definite mass and independent constraints the multiplier matrix $G M^{-1} G^T$ is invertible and

$$(G M^{-1} G^T) \lambda = G M^{-1} F + g_{qq}[v, v].$$

The position formulation is a differentiation-index-three system in first-order form. Consistent initial data require both position and velocity constraints. A small position residual alone places no bound on multiplier error without regularity of its time dependence.

Derive a concrete order loss before interpreting an index

A manufactured constraint $q = \sin t$ with $q' = v,\ v' = \lambda$ has $v = \cos t, \lambda = -\sin t$. Suppose a position solver introduces an alternating error $e_n = (-1)^n h^p$. A backward difference for velocity amplifies it to $2(-1)^n h^{p-1}$, and a second difference for the multiplier amplifies it to $4(-1)^n h^{p-2}$. A smooth error series instead can gain cancellations. This distinction prevents assigning every index-three method a universal loss of two orders. Consistent stage order, constraint enforcement, regularity, initial conditions and the reconstruction used for the multiplier all affect the theorem.

Consider a converged Radau collocation step for the simpler index-two constraint $q' = z,\ q = g(t)$. Its stages enforce the exact sampled $g(t_n + c_i h)$. Starting from exact q_n, the derivative stages equal those of the degree-s interpolant through the start and the s stage nodes. Stiff accuracy identifies the final stage as the output. Interpolant differentiation gives endpoint derivative error $O(h^s)$ for smooth g. The constrained position is exact here, illustrating precisely how classical differential-variable order fails to predict a multiplier measurement.

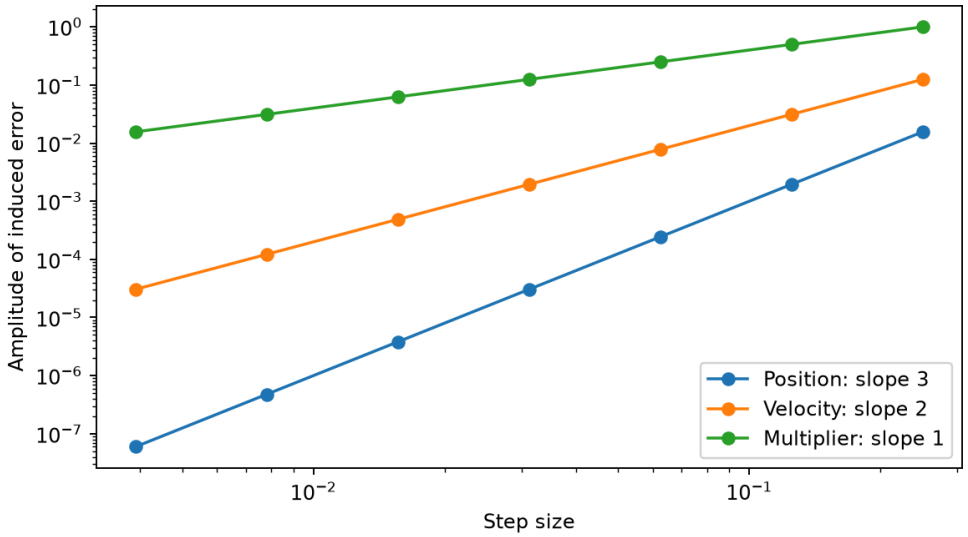
Derive an index-dependent rate in the constant-coefficient case

For a regular linear differential-algebraic pencil with constant coefficients, the algebraic part can be written as $Nz' = z - g(t)$, where $N^\nu = 0$ and $N^{\nu-1} \neq 0$. Repeated substitution gives $z = g + Ng' + \cdots + N^{\nu-1} g^{(\nu-1)}$. For a single Jordan chain, its last component specifies the input directly while the first can depend on its derivative of order $\nu - 1$. The highest derivative actually seen depends on the input direction and the reported component.

Interpolate g through $s + 1$ distinct nodes at fixed relative positions on an interval of length h. Taylor expansion at the endpoint through degree s is exact for the interpolant. The residual nodal values are $O(h^{s+1})$. Differentiating the fixed reference cardinal polynomials ℓ times introduces $h^{-\ell}$, giving $O(h^{s+1-\ell})$ for derivative order $\ell \leq s$. Smoothness through degree $s + 1$ is required. If nodal observations themselves carry $O(h^p)$ disturbances, their differentiated contribution is $O(h^{p-\ell})$. The worst reconstructed component on an index-ν chain therefore has order $\min(s + 1, p) - (\nu - 1)$, when $\nu - 1 \leq s$ and the forcing couples to that last derivative. This is an error bound for this specified reconstruction. A different integrator's stage order and enforcement of the constraints need a separate analysis.

What does a nonlinear stage tolerance control?

A stage solve tolerance on position has the units of position. When converting residuals to a velocity or acceleration tolerance, budget the factors $1/h$ and $1/h^2$ and the conditioning of the constraint matrix. Plot force error on a manufactured trajectory as well as the position-constraint residual.



Differencing an alternating position disturbance of amplitude h³ yields 2h² for velocity and 4h for the multiplier. These are the contributions from the disturbance alone. Smooth truncation errors need not follow these slopes.

Tutorial on contractivity and constraint data

Check the algebraic stability test on forward Euler

▼ **Work through the solution**

With $a = 0, b = 1$, the matrix coefficient is -1. For $y' = -y$, the amplification $1 - h$ has magnitude above one for $h > 2$, confirming that increasing the step may violate contractivity.

What force holds a unit mass on the unit circle at speed three?

▼ **Work through the solution**

Choose $g = (q^T q - 1)/2$, so $G = q^T$, $g_{qq} = I$, and $F = 0$. With $\|q\| = 1,\ q^T v = 0$, the multiplier is $\|v\|^2 = 9$ and the constraint force is $-9q$. An outward initial velocity violates the hidden constraint even if the position is exact.

Hairer and Wanner · stiff and differential-algebraic equations

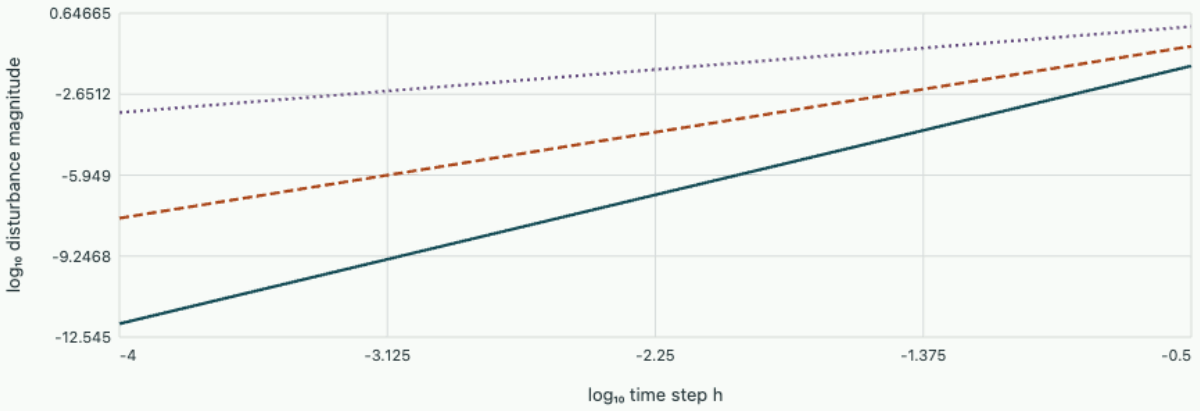
Change parameters · Why an accurate position can hide an inaccurate constraint force

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

Position disturbance order p	3
Disturbance coefficient	1



An alternating disturbance loses one power of h under each discrete derivative. For p = 3, the induced multiplier contribution scales as h^1.

Conserving motion in a physical system

Track an invariant in phase space. Compare a rotating charge, parametrically driven oscillator and magnetic orbit. Use phase and energy as independent error measurements.

1. Hamiltonian motion in phase space
2. Nonlinear stiffness and amplitude-dependent period
3. Liouville flow and what an ensemble can remember
4. Parametric resonance and Floquet growth
5. A magnetic orbit that spirals outward in a perfect vacuum
6. Relativistic particle dynamics and the field tensor
7. Euler–Poinsot motion, torque-free nutation and the intermediate axis
8. Symplectic transformations, Poisson brackets and the geometry of measurements
9. Canonical perturbations, small denominators and resonant capture
10. Distinguish a physical precession from numerical drift

Hamiltonian motion in phase space

Classical mechanics

Construct a Hamiltonian, read a phase portrait and distinguish bounded motion from rotation at a separatrix.

FROM FIRST PRINCIPLES

Construct Hamilton's equations by exchanging velocity for momentum

Begin with a smooth Lagrangian $L(q, \dot{q}, t)$ for a conservative mechanical system. Assume the momentum definition $p = \partial_{\dot{q}} L$ can locally be inverted to give $\dot{q}$ as a function of (q, p, t). This regularity is what lets position and momentum serve as independent state coordinates. Define the Hamiltonian and differentiate the product before substituting any equations of motion:

$$H = p\dot{q} - L,$$
$$dH = \dot{q}\, dp + p\, d\dot{q} - L_q\, dq - L_{\dot{q}}\, d\dot{q} - L_t\, dt.$$

The two $d\dot{q}$ terms cancel by the momentum definition. Comparing the coefficients of the independent differentials gives

$$H_p = \dot{q},$$
$$H_q = -L_q.$$

Hamilton's principle previously gave the Lagrange equation $\dot{p} = L_q$, so

$$\dot{q} = \partial_p H,$$
$$\dot{p} = -\partial_q H,$$
$$\partial_t H = -\partial_t L.$$

Along a solution the q and p contributions to dH/dt cancel. For a time-independent Hamiltonian H is constant, so planar trajectories lie on its level sets. In the elementary case $L = m\dot{q}^2/2 - U(q)$, the transformation yields

$$H = p^2/(2m) + U(q).$$

Background · Generalised momentum

$$p_i = \partial L / \partial \dot{q}_i$$

Its units depend on the coordinate. Conjugacy to an angle gives an angular momentum. Conservation requires the associated coordinate to be cyclic in the chosen Lagrangian.

DERIVATION AND APPLICATIONS

How Hamiltonian flow preserves phase-space area

An energy contour shows where a single trajectory may lie. To learn what the dynamics does to nearby initial conditions, follow a small patch of the canonical (q, p) plane. Consider one degree of freedom and a twice continuously differentiable Hamiltonian. Let the system advance by a short time δt. To first order the coordinate map is a displacement along the Hamiltonian velocity field:

$$q' = q + H_p\,\delta t,$$
$$p' = p - H_q\,\delta t.$$

The area of a sufficiently small parallelogram changes by the determinant of this map's derivative. Differentiate with respect to both initial coordinates. The product of off-diagonal elements enters only at second order:

$$\det \frac{\partial(q', p')}{\partial(q, p)} = (1 + H_{pq}\delta t)(1 - H_{qp}\delta t) + H_{pp}H_{qq}\delta t^2 = 1 + O(\delta t^2).$$

For the exact continuous map its area rate is the divergence of the phase velocity. Equality of the mixed second derivatives gives

$$\frac{1}{A}\frac{dA}{dt} = \partial_q \dot{q} + \partial_p \dot{p} = H_{pq} - H_{qp} = 0$$

for an infinitesimal material area A. Integrating adjacent patches gives the same conclusion for a finite region. The outline can stretch and rotate. Even explicit time dependence of H does not change this volume result. A friction force imposed independently of the Hamiltonian can change it. The second-order remainder above belongs to a short expansion; repeatedly taking finite steps with an approximate rule can accumulate a spurious change.

WORKED EXAMPLE

Track a cloud in a harmonic trap

Change the inputs and investigate this derivation.

Prepare a small rectangular set of initial states, 2.00 mm wide in position and 0.0100 kg m/s wide in momentum. Its area is 2.00×10^{-5} J s. In a trap of mass 0.500 kg and frequency $\omega = 4.00$ rad/s, the exact quarter-period map is $q' = p/(m\omega)$, $p' = -m\omega q$. The new widths are 5.00 mm and 0.00400 kg m/s. Their product retains the original area even though neither width is conserved.

SECOND APPLICATION

Diagnose a numerical phase portrait

Change the inputs and investigate this derivation.

An explicit Euler step in the same trap uses $q' = q + p\delta t/m$ and $p' = p - m\omega^2 q\delta t$. Its determinant is exactly $1 + \omega^2\delta t^2$. At $\delta t = 0.0500$ s it expands area by 1.04 per step, giving $1.04^{50} = 7.11$ times the initial area after 50 steps. A graph of the growing cloud could look like physical heating. The determinant exposes a discretization error without requiring an exact trajectory for each point.

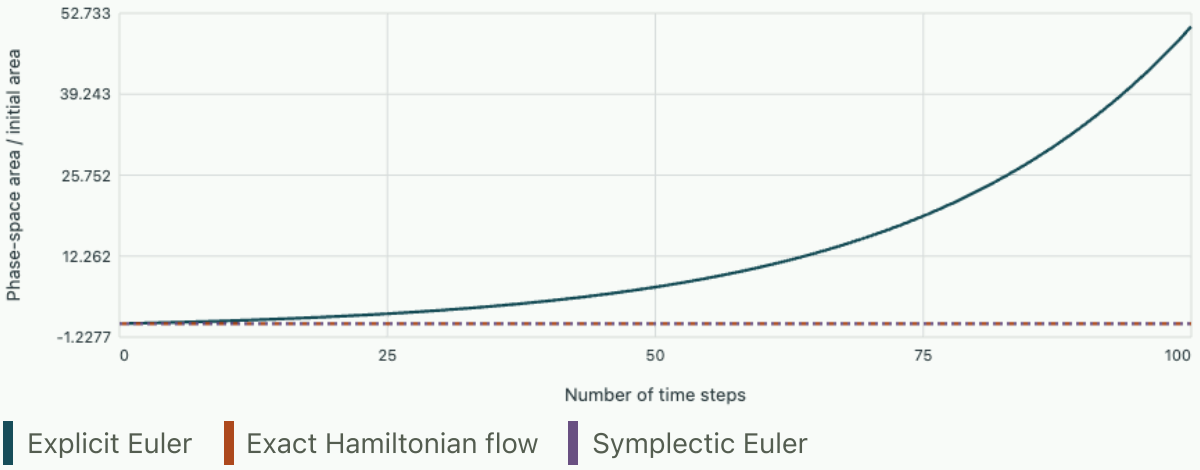
Try it · Phase area

Vary the physical parameters in the derived relation. Use the units and assumptions in the readout to connect the curves to the two worked examples. A parameter study can test the forward prediction; the same relation can then be inverted to infer a parameter from a measurement.

Open this live calculation

Try the calculation

Scaled step ωΔt **0.2**



Euler determinant 1+(ωΔt)² = 1.04. A symplectic scheme preserves area but still needs a step-convergence check for phase accuracy.

Position needs a conjugate momentum

The position of a pendulum at one instant does not say which way it will swing. A state needs both a configuration coordinate and a momentum. In Lagrangian mechanics define the canonical momentum by $p = \partial\mathcal{L}/\partial\dot{q}$. It can have different physical dimensions depending on the coordinate. For a particle on a line it has units of linear momentum. For a pendulum angle it has units of angular momentum. The pair (q, p) forms a two-dimensional phase space for one degree of freedom.

Provided the relation between p and $\dot{q}$ can be inverted, define $H(q, p, t) = p\dot{q} - \mathcal{L}(q, \dot{q}, t)$ with the velocity expressed in terms of q, p. Take a differential. The terms proportional to $d\dot{q}$ cancel by the definition of p. Using Lagrange's equation $\partial_q\mathcal{L} = \dot{p}$ gives

$$dH = \dot{q}\,dp - \dot{p}\,dq - \partial_t\mathcal{L}\,dt.$$

Matching the coefficients produces Hamilton's equations

$$\dot{q} = \partial_p H,$$
$$\dot{p} = -\partial_q H.$$

For a system with ordinary quadratic kinetic energy, a velocity-independent potential and fixed constraints, $H = K + U$. Its change along a solution is $dH/dt = \partial_t H$ because the two terms containing the derivatives with respect to q, p cancel. A time-independent Hamiltonian confines the trajectory to a level set.

Use Poisson brackets to test conservation

Hamilton's two equations specify how q_i and p_i change. To follow an observable $A(q, p, t)$, apply the chain rule and substitute $\dot{q}_i = \partial H/\partial p_i$ and $\dot{p}_i = -\partial H/\partial q_i$. This defines the Poisson bracket,

$$\frac{dA}{dt} = \frac{\partial A}{\partial t} + \{A, H\},$$
$$\{A, B\} = \sum_i \left(\frac{\partial A}{\partial q_i}\frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i}\frac{\partial B}{\partial q_i} \right).$$

Since $\{q_i, p_j\} = \delta_{ij}$, a candidate change of phase-space coordinates must preserve these brackets to be canonical. A vanishing bracket with H certifies conservation for an observable having no explicit time dependence. It is a local calculation and does not require solving an orbit.

For a particle in a two-dimensional trap set $H = (p_x^2 + p_y^2)/(2m) + (k_x x^2 + k_y y^2)/2$ and $L_z = x p_y - y p_x$. The momentum derivatives in the bracket cancel the two kinetic contributions. The remaining coordinate terms give $\dot{L}_z = (k_x - k_y) x y$, exactly the torque $x F_y - y F_x$. With $k_x = 4.00$ N/m, $k_y = 6.00$ N/m and instantaneous position $(0.030, 0.040)$ m, angular momentum decreases at 2.40×10^{-3} N m. At a point on either axis the instantaneous derivative vanishes even for the anisotropic trap. For a conservation law holding on every trajectory the expression must vanish throughout phase space, as it does when $k_x = k_y$.

Draw energy contours before integrating

A rigid pendulum of mass m and length ℓ has $p = m\ell^2\dot{q}$ and $H = p^2/(2m\ell^2) + mg\ell(1 - \cos q)$. At the bottom $q = 0, p = 0$ is a stable equilibrium. Low-energy contours are ellipses to leading order because the cosine potential is quadratic there. Each closed contour returns to its original phase point after one full oscillation. Points of greatest $|p|$ are fastest and occur at the bottom.

The top $q = \pi, p = 0$ has energy $2mg\ell$. Its contour separates swinging trajectories from trajectories that keep rotating in one direction. This separating curve is the separatrix. Since angles differing by 2π label the same physical position, the full portrait lives on a cylinder. A rotation winds around that cylinder. As a trajectory approaches the separatrix, it spends a long time near the unstable top, so the small-angle period no longer estimates its return time accurately.

Compare thresholds and phase velocities

For $m = 0.50$ kg, $\ell = 0.80$ m and $g = 9.81$ m/s², the potential barrier is 7.85 J. A start from the bottom at 5.0 rad/s supplies kinetic energy $0.5(0.50)(0.80)^2(5.0)^2 = 4.00$ J. At the turning point $\cos q = 1 - 4.00/(0.50 \times 9.81 \times 0.80) = -0.0194$. The maximum angle is about 91.1°. Energy classifies this as a swing even though a small-angle expansion would be inaccurate at the turning point.

Continuous rotation requires bottom speed above $\sqrt{4g/\ell} = 7.00$ rad/s for this ideal pendulum. At precisely that value the system approaches the upper equilibrium asymptotically. This statement assumes a rigid link that can sustain compression. A pendulum made from a flexible string may lose tension before reaching the top, in which case the assumed circular constraint fails.

Check a numerical phase portrait

For any smooth one-degree Hamiltonian the divergence of the phase velocity is

$$\partial_q \dot{q} + \partial_p \dot{p} = H_{pq} - H_{qp} = 0.$$

A small patch of initial states can shear and stretch while preserving its phase-space area. Viscous damping would instead contract the patch. Numerical time integration can also produce apparent contraction or growth, so refining the step is essential before interpreting a spiral as evidence of friction.

For the harmonic oscillator, ordinary forward Euler gives

$$q' = q + \Delta t\, p/m,$$
$$p' = p - \Delta t\, kq.$$

The determinant of this linear update is $1 + \Delta t^2 k/m$, larger than one at every step. The spurious outward spiral follows from the numerical map alone. Updating momentum first and then using the new momentum for the position gives determinant one. That area property preserves the geometric structure, though it does not by itself guarantee accurate phase for a coarse step. Keep both geometry and timing in view when judging numerical trajectories.

The area inside a closed harmonic-oscillator contour provides another numerical test. Its semiaxes are $\sqrt{2E/k}$ in position and $\sqrt{2mE}$ in momentum. Their product times π gives $2\pi E/\omega$. Increasing oscillator energy therefore expands the area in direct proportion. A phase portrait plotted using velocity replaces every momentum value by its value divided by mass, changing numerical areas and their units. Specify the vertical coordinate before comparing this diagnostic between models with different masses.

A slowly changing spring transports a phase-space area

Suppose an oscillator's spring stiffness can be varied externally during the motion. Its Hamiltonian $H = p^2/(2m) + m\omega(t)^2 q^2/2$ now depends explicitly on time, so ordinary energy conservation does not apply. At an instantaneous point on a trajectory Hamilton's equations make the terms involving $\dot{p}$ and $\dot{q}$ cancel in dH/dt. The remaining change is

$$dE/dt = \partial H/\partial t = m\omega\dot{\omega}q^2.$$

Energy is supplied when the imposed parameter raises the potential at the current displacement.

Take $|\dot{\omega}| \ll \omega^2$ so that many cycles occur while the frequency changes by an appreciable fraction. Over one cycle ω and E may then be treated as nearly fixed. Because average potential and kinetic energies of a harmonic oscillator are equal,

$$\langle q^2 \rangle = E/(m\omega^2).$$

Inserting this mean into the work rate gives the averaged relation

$$\frac{dE}{dt} \simeq \frac{E}{\omega}\dot{\omega}, \qquad \frac{d}{dt}\left(\frac{E}{\omega}\right) \simeq 0.$$

The quantity $J = E/\omega$ is the area enclosed by the instantaneous constant-energy ellipse divided by 2π. Calling it an adiabatic invariant states both the approximation and the observable. A slowly changing system migrates between contours of different energies in a way that preserves this action to leading order. This mechanical meaning of adiabatic concerns a timescale ratio. The derivation does not assume heat exchange, dissipation or a thermodynamic temperature.

Compare a slow compression with a sudden change

A 0.500 kg oscillator has initial angular frequency 20.0 rad/s and amplitude 0.0400 m. Its energy is $m\omega_i^2 A_i^2/2 = 0.160$ J and initial action is 0.00800 J s. Raise the frequency smoothly to 40.0 rad/s over many periods. Action conservation predicts $E_f = 0.320$ J and final amplitude $\sqrt{2E_f/(m\omega_f^2)} = 0.0283$ m. The external control supplies work even as the maximum displacement decreases.

For comparison, change the stiffness instantaneously at a turning point, retaining the same initial and final frequencies. Position and momentum are continuous, but the potential energy at that position quadruples immediately. The new energy is 0.640 J and the amplitude remains 0.0400 m. Make the same switch at the zero crossing instead and the energy initially remains 0.160 J because the potential vanishes there. These two outcomes distinguish a rapid change tied to an oscillator phase from the nearly phase-independent slow ramp. Recording both position and velocity after the ramp allows energy, action and phase to be assessed separately.

A time-dependent Hamiltonian still carries each small phase-space area element into one of equal area, even during a sudden change of curvature. This exact conservation for an ensemble is distinct from slow conservation of the contour area assigned to a single orbit. A group of uniformly distributed initial phases illustrates the difference: immediately after a sudden spring switch all points occupy the same locations and enclose their original region, but their new energies depend on phase. The switch maps the original single-energy contour onto points that subsequently occupy many different energy contours of the new Hamiltonian.

Nonlinear stiffness and amplitude-dependent period

Classical mechanics

Derive the leading frequency shift of a symmetric nonlinear oscillator and test whether a measured amplitude shift is consistent with a cubic spring.

FROM FIRST PRINCIPLES

Derive the first amplitude correction to a spring's frequency

A symmetric elastic restoring force has no quadratic displacement term; the first correction to a linear spring can therefore be cubic. Suppose a conservative oscillator satisfies $m\ddot{x} + kx + \alpha x^3 = 0$ with $k > 0$ and $|\alpha A^2/k| \ll 1$ at measured peak amplitude A. Leading motion is approximately $A\cos\Omega t$. Cube this motion to find which frequencies enter the perturbing force:

$$x^3 \simeq A^3[\tfrac{3}{4}\cos\Omega t + \tfrac{1}{4}\cos 3\Omega t].$$

To maintain a bounded periodic solution, its fundamental component must balance at first order. Otherwise the remaining fundamental force would resonantly grow the correction. Combine the inertia and restoring-force coefficients at $\cos\Omega t$:

$$-m\Omega^2 A + kA + \tfrac{3}{4}\alpha A^3 \simeq 0, \qquad \Omega \simeq \sqrt{k/m}\left(1 + \frac{3\alpha A^2}{8k}\right).$$

The third-harmonic component also distorts the waveform, but reinserting that order-α distortion into the cubic force affects the fundamental only at the next perturbative order. Positive α raises frequency with amplitude; negative α lowers it within the local range where the stable weakly nonlinear model applies.

DERIVATION AND APPLICATIONS

Find the third harmonic generated by a weak cubic spring

A change in oscillation period is one signature of a nonlinear restoring force. The same nonlinearity also changes the shape of the recorded motion. Let an undamped mass obey $m\ddot{x} + kx + \alpha x^3 = 0$, with $k > 0$. Release it from rest at the measured peak displacement A. Define $\omega_0 = \sqrt{k/m}$, the small signed parameter $\epsilon = \alpha A^2/k$, and a stretched time $\tau = \Omega t$. Write $x/A = \cos\tau + \epsilon u(\tau)$ with $u(0) = u'(0) = 0$ so that A remains the physical turning point.

Allow the frequency to shift as $\Omega = \omega_0(1 + \epsilon\nu)$. Substitute the expansion into the equation and retain terms only to first order in ϵ. Primes below denote differentiation with respect to τ:

$$u'' + u = 2\nu\cos\tau - \cos^3\tau = (2\nu - 3/4)\cos\tau - \tfrac{1}{4}\cos 3\tau.$$

A term driven at the unit frequency would grow proportionally to $\tau \sin \tau$, contradicting a bounded, periodic correction. Choose $\nu = 3/8$ to remove it. The remaining forcing is at three times the fundamental. A trial $C \cos 3\tau$ has $u'' + u = -8C \cos 3\tau$. Add the required homogeneous term to satisfy the turning-point conditions:

$$u(\tau) = \frac{\cos 3\tau - \cos \tau}{32}.$$

The predicted displacement therefore contains two lines with linked amplitudes and a shifted fundamental:

$$x(t) \simeq A[(1 - \epsilon/32)\cos \Omega t + (\epsilon/32)\cos 3\Omega t], \qquad \Omega/\omega_0 \simeq 1 + 3\epsilon/8.$$

The coefficient sign gives the third harmonic relative to the phase of a release at positive displacement. Squaring an amplitude to obtain a power spectrum erases that sign. The approximation assumes $|\epsilon| \ll 1$ and excludes appreciable dissipation during the observation.

WORKED EXAMPLE

Estimate a micrometre-scale distortion

Change the inputs and investigate this derivation.

Take $A = 12.0$ mm, $k = 20.0$ N/m and $\alpha = 1.00 \times 10^{4}$ N/m³. Then $\epsilon = 0.0720$, so the third-harmonic displacement amplitude is

$$12.0(0.0720)/32 = 0.0270\,\mathrm{mm}.$$

The fundamental frequency should increase by 2.70%. With $\omega_0 = 10.0$ rad/s, the third-harmonic line lies close to 30.8 rad/s.

SECOND APPLICATION

Test whether a measured line comes from the spring

Change the inputs and investigate this derivation.

At $A = 6.00$ mm on the same apparatus, ϵ falls by four and the third-harmonic amplitude falls by eight to 3.38 micrometres. The fractional frequency shift becomes 0.675%. A third-harmonic signal that instead halves with amplitude could arise from an already distorted input or readout. The two amplitude scalings plus an independent frequency measurement test the cubic model more strongly than observing a line at three times the frequency.

Try it · Duffing harmonic

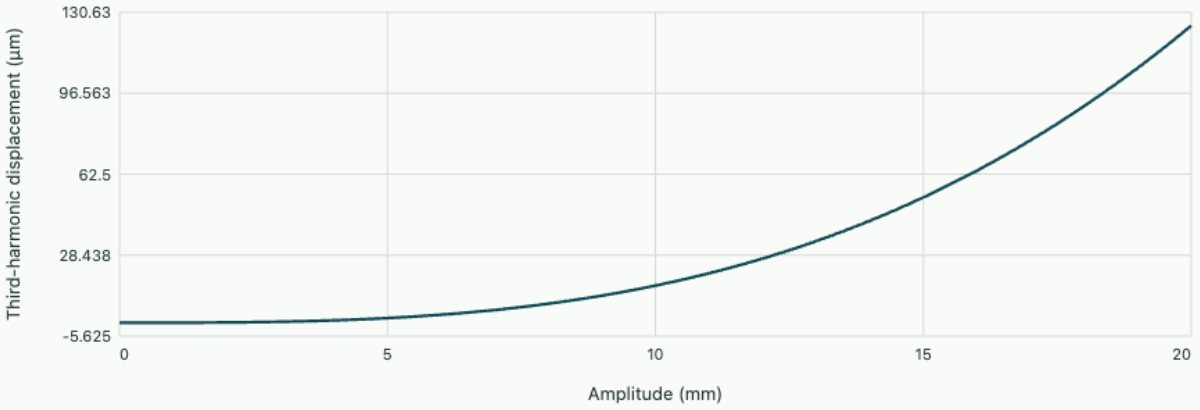
Vary the physical parameters in the derived relation. Use the units and assumptions in the readout to connect the curves to the two worked examples. A parameter study can test the forward prediction; the same relation can then be inverted to infer a parameter from a measurement.

Open this live calculation

Try the calculation

Amplitude (mm) **12**

Cubic stiffness (N/m³) **10000**



First perturbative harmonic

Linear stiffness 20 N/m. At chosen amplitude ε=0.072, third harmonic 27 µm; first frequency shift 2.7%. Requires ε≪1.

Start from a reversible force law

A spring that looks linear under small deflection can stiffen as it is displaced further. For an oscillator symmetric under $x \mapsto -x$, the restoring force expanded about equilibrium has no even powers. Keep the leading correction and write $m\ddot{x} + kx + \alpha x^3 = 0$, with $k > 0$ and α measured in N/m³. Assume the displacement is small enough that neglected fifth-order terms remain unimportant. Positive α gives a hardening spring. A negative cubic coefficient describes local softening and may need higher-order terms to represent large displacements safely within the model.

The potential is $kx^2/2 + \alpha x^4/4$. Without damping the total energy is conserved and a release at amplitude A fixes that energy. Unlike a harmonic oscillator, the quadratic curvature at the origin no longer fixes the period for every possible energy. We want a quantitative connection between the amplitude of the motion and its zero-crossing interval.

Obtain the first frequency correction

Try $x \simeq A\cos\Omega t$ and substitute it into the cubic equation. The identity $\cos^3 u = (3\cos u + \cos 3u)/4$ shows that the nonlinearity introduces a component at the original frequency and a third harmonic. Balance the fundamental coefficient to get $-m\Omega^2 + k + 3\alpha A^2/4 \simeq 0$. Expanding the square root gives

$$\Omega \simeq \omega_0 \left(1 + \frac{3\alpha A^2}{8k}\right),$$

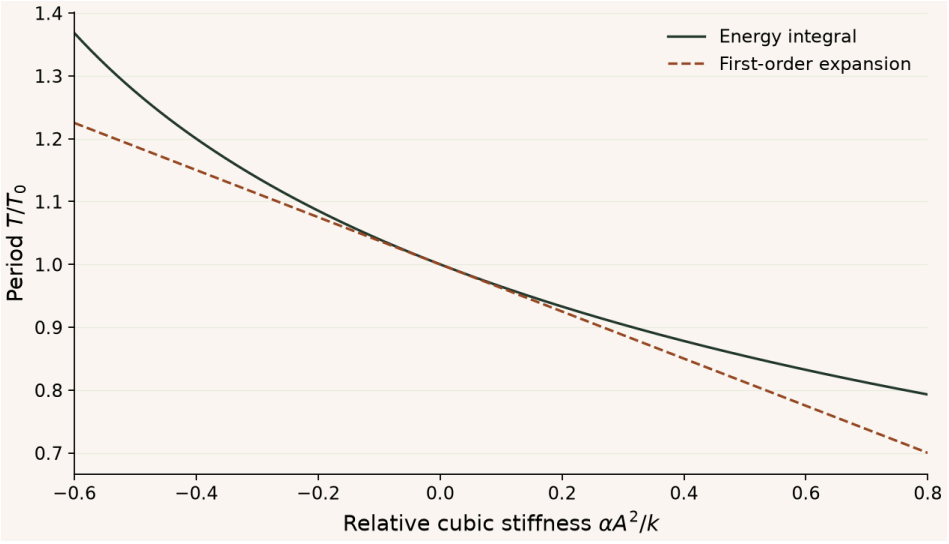
$$\omega_0 = \sqrt{k/m}.$$

The control parameter is $|\alpha| A^2/k$, the ratio of cubic to linear force at maximum displacement. The small additional third harmonic changes the waveform at first order but only feeds back into this leading frequency result at higher order. Fitting a perfect sinusoid to a high-precision trace may therefore show a residual even when the measured average period follows this formula.

Check the result using the conserved energy

The quarter-period is the time to move from equilibrium to A. Substitution of the energy at the turning point gives an integral over $dx/\sqrt{2[E - U(x)]/m}$. Let $x = A\sin u$ to remove the square-root endpoint. The remaining integrand is proportional to $\{1 + [\alpha A^2/(2k)](1 + \sin^2 u)\}^{-1/2}$. Expanding and averaging over the interval from zero to $\pi/2$ uses mean $\sin^2 u = 1/2$, yielding $T \simeq (2\pi/\omega_0)[1 - 3\alpha A^2/(8k)]$. This agrees with the reciprocal of the harmonic-balance result.

The integral approach generalises to experimentally tabulated conservative forces by integrating them first to obtain a potential. It would need modification for strong damping, where energy varies significantly between the two turning points of one cycle.



A positive cubic stiffness shortens the period. The first-order expansion is tangent to the energy-integral curve at zero and gradually loses accuracy away from that limit.

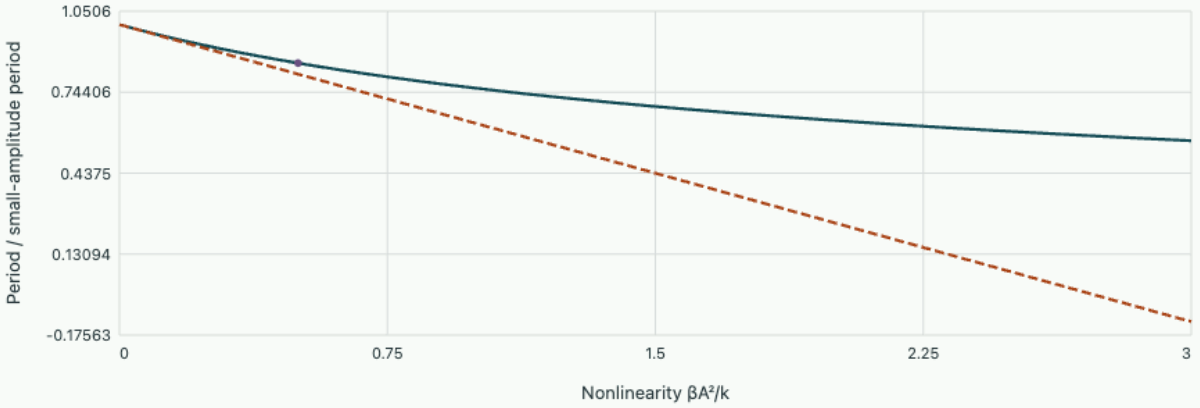
Try it · Duffing

Change the inputs and compare the recalculated curves with the preceding derivation and baseline figure. The axes give the units; the numerical readout states the model assumptions and any limiting cases. Reset the inputs to return to the reference setting.

Open this live calculation

Try the calculation

Cubic stiffness × amplitude² / linear stiffness **0.5**



Energy integral · First-order expansion · Selected amplitude

Exact period ratio 0.85413. The expansion is local to βA²/k ≪ 1; it becomes negative if extrapolated far enough.

Infer stiffness from period data

For $m = 0.50$ kg, $k = 200$ N/m and $\alpha = 4.0 \times 10^4$ N/m³, $\omega_0 = 20.0$ rad/s. At $A = 0.020$ m the force ratio is 0.080, giving a leading frequency increase of 3.0% to 20.6 rad/s. At half this amplitude the predicted increase is 0.75%. Measure the period over several cycles at each amplitude, keeping the amplitude change due to loss small. A plot of frequency against the square of amplitude should be approximately straight near the origin. A nonzero intercept shift under a change of temperature could indicate a change in ordinary linear stiffness instead of a new nonlinear coefficient.

Compute the entire amplitude dependence of a pendulum

A cubic expansion is useful near equilibrium. For a pendulum it is possible to carry the energy calculation beyond every finite Taylor approximation. Take a point mass with fixed length ℓ, no friction and a release from rest at angle $0 < \theta_0 < \pi$. Its conserved energy gives

$$\dot\theta^2 = (2g/\ell)(\cos\theta - \cos\theta_0).$$

Use $\cos\theta = 1 - 2\sin^2(\theta/2)$. Defining $k = \sin(\theta_0/2)$ reduces the speed magnitude to $2\sqrt{g/\ell}\sqrt{k^2 - \sin^2(\theta/2)}$.

One quarter period is the time to travel between zero angle and the turning point. The reciprocal speed appears to diverge at the latter, but it has an integrable square-root singularity for every ordinary turning angle. To evaluate it stably, set $\sin(\theta/2) = k\sin u$. The factor $\cos u$ from differentiating cancels the same factor inside the square root. Four equivalent quarters give

$$T = 4\sqrt{\ell/g}\int_0^{\pi/2} \frac{du}{\sqrt{1 - k^2\sin^2 u}} = 4\sqrt{\ell/g}\,K(k).$$

Here K denotes the complete elliptic integral with modulus k. Expanding the integrand at small k and integrating the first correction gives

$$T/T_0 = 1 + k^2/4 + \cdots = 1 + \theta_0^2/16 + \cdots,$$

where $T_0 = 2\pi\sqrt{\ell/g}$. The positive correction is the familiar softening: increasing amplitude lengthens the period because the restoring torque grows more slowly than its linear extrapolation.

Decide when a small-angle clock loses accuracy

A pendulum of length 0.800 m has $T_0 = 1.794$ s at $g = 9.81$ m/s². At a release of 60.0°,

$$k = \sin 30^\circ = 0.500.$$

Evaluating the integral gives $K(0.500) = 1.68575$ and $T/T_0 = 2K/\pi = 1.0732$, hence $T = 1.926$ s. The weak-angle approximation gives

$$1 + (\pi/3)^2/16 = 1.0685.$$

It misses the exact period by roughly 0.008 s per cycle; a many-cycle timing record resolves the accumulated discrepancy.

As θ_0 tends to π, the integrand near $u = \pi/2$ approaches the reciprocal of a vanishing cosine. The integral diverges logarithmically, expressing the long time spent close to the unstable upper equilibrium. No finite polynomial in amplitude can reproduce that divergence. A numerical timing experiment should therefore locate regular zero crossings rather than demand a turning-point velocity of exactly zero in sampled data. Comparing the crossing intervals with the regularized integral distinguishes numerical resolution error from the genuine increase in physical period.

A way to evaluate the elliptic integral without a special-function library is to form the arithmetic–geometric mean. Begin with $a_0 = 1$ and $b_0 = \sqrt{1 - k^2}$, then replace the pair by their arithmetic and

geometric means. Both converge rapidly to a common M, and $K(k) = \pi/(2M)$. For the 60° release the first pair is 1 and 0.86603, the next pair is approximately 0.93301 and 0.93060, and the subsequent pair agrees near 0.93181. Taking the reciprocal of the common value directly gives the period ratio. This procedure supplies an accurate independent benchmark even when the small-amplitude expansion ceases to converge quickly.

Numerical quadrature and the arithmetic–geometric mean should agree at a fixed amplitude. If a period extracted from actual trajectories differs while these two evaluations agree, examine the dynamical time step, how release angle is measured, and whether the observed oscillations have suffered appreciable damping.

Liouville flow and what an ensemble can remember

Classical mechanics

A cloud of identical noninteracting beads is released with slightly different initial positions and momenta. Its position distribution widens. Does the system occupy more phase space, or has its uncertainty merely changed shape?

Follow a small phase-space element

For one canonical degree of freedom the velocity of a point (q, p) is $(\partial_p H, -\partial_q H)$. A short-time map takes $\boldsymbol{z}$ to $\boldsymbol{z} + \boldsymbol{v}(\boldsymbol{z}, t)\, dt$. Its Jacobian matrix is $I + (\partial \boldsymbol{v}/\partial \boldsymbol{z})dt$; expanding the determinant gives $1 + (\nabla_z \cdot \boldsymbol{v})dt + O(dt^2)$. Hamilton's equations make the divergence zero because the two mixed second derivatives of a smooth Hamiltonian cancel. Each additional canonical pair contributes a similar cancellation. A time-dependent Hamiltonian may change energy but still preserve the instantaneous phase-space volume along the flow.

Conservation of the number of ensemble members in a moving phase-space box also gives the continuity equation $\partial_t f + \nabla_z \cdot (f\boldsymbol{v}) = 0$. The vanishing divergence leaves $df/dt = 0$ on each trajectory. A stationary mechanical system can admit $f = F(H)$ as a stationary distribution because trajectories stay on energy surfaces. Such stationarity alone does not imply that one trajectory explores an entire energy surface. For example, further integrals of motion restrict regular multidimensional trajectories.

Propagate a cloud released into free flight

For mass m, the exact map is $p = p_0,\ q = q_0 + tp_0/m$. Its determinant is one regardless of time. An initially uncorrelated cloud with variances s_q^2, s_p^2 has evolved covariance $C_{qp} = ts_p^2/m$. The new position variance is $s_q^2 + (ts_p/m)^2$, while the momentum variance stays fixed. Compute the determinant of this covariance matrix:

$$\sigma_q^2 \sigma_p^2 - C_{qp}^2 = s_q^2 s_p^2.$$

For a Gaussian the ellipse at a fixed multiple of the covariance radius thus retains its area while it shears. With $m = 1\,\mathrm{kg}$, $s_q = 1\,\mathrm{m}$, $s_p = 0.5\,\mathrm{kg\,m/s}$, and $t = 4\,\mathrm{s}$, the observed position standard deviation is $\sqrt{5} = 2.236\,\mathrm{m}$. The momentum standard deviation remains 0.5 and the covariance becomes $1.0\,\mathrm{kg\,m^2/s}$. Their uncorrected width product is 1.118 in action units. The square root of the covariance determinant is still 0.5. Multiplication of measured marginal widths alone falsely suggests an area increase.

Compare reversible shear and real drag

Now apply a linear drag $\dot{p} = -\gamma p$ to each bead with constant $\gamma > 0$. Its phase-space divergence is $-\gamma$, hence the map determinant after time t is $e^{-\gamma t}$. Direct integration gives

$$p = e^{-\gamma t} p_0,\; q = q_0 + (1 - e^{-\gamma t}) p_0 / (m\gamma).$$

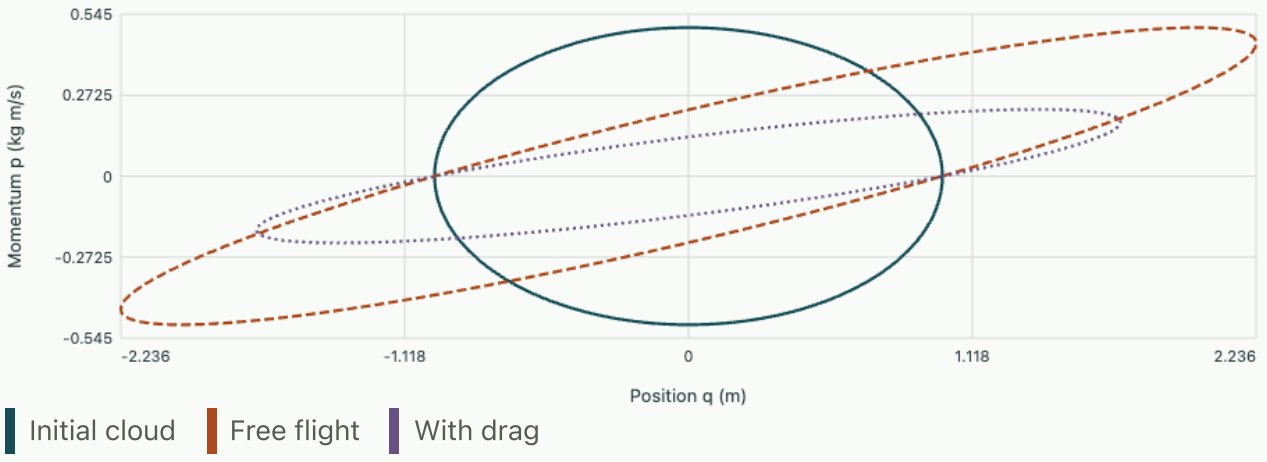
The cloud now genuinely loses area in the reduced bead description. A larger Hamiltonian model of bead plus environment can remain volume preserving. Choose a drag rate in the graph and compare the original, freely sheared and damped ellipses for the same observation time. Check there is a smooth zero-drag limit rather than substituting zero into a quotient.

Open this experiment online

Try the calculation

Time after release (s) **4**

Drag rate γ (1/s) **0.2**



Free position spread 2.236 m. The damped ellipse retains 44.93% of its initial area and has position spread 1.702 m.

Tutorial on distinguishing a shear from dissipation

A harmonic oscillator maps its scaled quadratures by a rotation. Show why an initially anisotropic cloud can periodically sharpen in position without phase-space cooling.

▼ **Work through the solution**

Write its covariance as $R\Sigma_0 R^T$. Since $\det R = 1$, its determinant is unchanged. A rotation of one quarter turn exchanges the original widths, so position can sharpen when the original momentum width in scaled units was smaller. Energy extraction is unnecessary.

At $t = 3\,\mathrm{s}$ the measured ellipse area is 55 percent of the initial area in the linear-drag model. Infer γ. Which independent observable checks the inference?

▼ **Work through the solution**

$$\gamma = -\ln(0.55)/3 = 0.199\,\mathrm{s}^{-1}.$$

The ratio of final to initial momentum standard deviations should also be 0.55 for drag without random forcing. Added environmental fluctuations would modify that prediction.

Further reading · Cambridge classical dynamics, Hamiltonian flow and Liouville's theorem.

Parametric resonance and Floquet growth

Classical mechanics

A pendulum can gain energy when its length varies periodically even if no additive force acts on it. A modulation near twice a natural frequency amplifies one phase of the oscillation and attenuates another. The stability test concerns the growth of a perturbation over a complete modulation period.

Linearise the timing mechanism

Consider $\ddot{x} + 2\gamma\dot{x} + \omega_0^2[1 + h\cos(\Omega t)]x = 0$. Here h is the fractional modulation of the stiffness and γ is the amplitude damping rate. If $x = \dot{x} = 0$ exactly, the solution stays zero. Small fluctuations or an initial condition provide the seed. Direct differentiation of $E = [\dot{x}^2 + \omega_0^2(1 + h\cos\Omega t)x^2]/2$ yields

$$\dot{E} = -2\gamma\dot{x}^2 - h\omega_0^2\Omega x^2\sin(\Omega t)/2.$$

Energy transfer depends on the phase of x^2, which naturally contains a component at twice the oscillation frequency.

Near $\nu = \Omega/2 \simeq \omega_0$, write $x = A(t)e^{i\nu t} + A^*(t)e^{-i\nu t}$ with slowly varying A. Multiplication by $\cos(2\nu t)$ mixes the negative-frequency term back into the positive-frequency term with coefficient $h\omega_0^2/2$. There is also a component at 3ν, rapidly oscillating in this first-resonance approximation. Retain terms proportional to $e^{i\nu t}$, neglect $\ddot{A}$ and products of small damping with $\dot{A}$. This gives

$$2i\nu\dot{A} + (\omega_0^2 - \nu^2 + 2i\gamma\nu)A + (h\omega_0^2/2)A^* = 0.$$

Find the first instability tongue

Set $\Delta = \nu - \omega_0$ and keep leading terms in the small ratios $h, \gamma/\omega_0, |\Delta|/\omega_0$. The slow equations are $\dot{A} = (-\gamma - i\Delta)A + igA^*$ and their conjugate, where $g = h\omega_0/4$. The two-by-two characteristic equation is

$$(s + \gamma)^2 = g^2 - \Delta^2.$$

The faster exponent is

$$\operatorname{Re} s_+ = -\gamma + \operatorname{Re}\sqrt{(h\omega_0/4)^2 - \Delta^2}.$$

The oscillation grows when $g^2 > \gamma^2 + \Delta^2$. Exactly on resonance this gives $h > 4\gamma/\omega_0$. With $\omega_0 = 10\,\mathrm{s}^{-1}$, $\gamma = 0.10\,\mathrm{s}^{-1}$ and $h = 0.08$, $g = 0.20\,\mathrm{s}^{-1}$, giving an on-resonance doubling time $\ln 2/0.10 = 6.93\,\mathrm{s}$. The approximate growing interval in drive angular frequency is $20 \pm 0.346\,\mathrm{s}^{-1}$. A large-amplitude real oscillator can saturate as its natural frequency changes with amplitude, a mechanism absent from this linear equation.

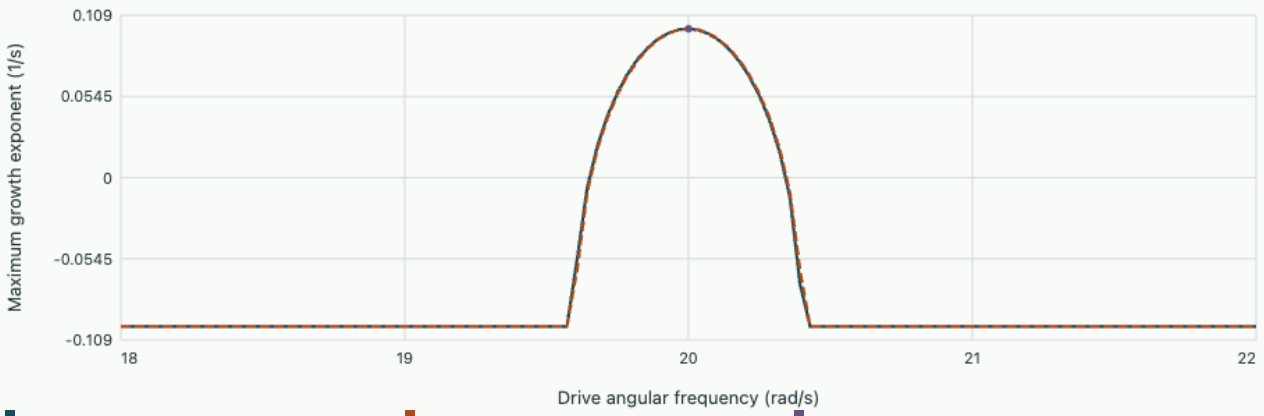
Compute the period map without averaging

Integrate the original linear equation over one drive period from two independent initial states $(1, 0)$ and $(0, 1)$. Their final columns form a matrix M, which maps every initial perturbation over the same period T. After j periods the map is M^j; the maximum asymptotic amplitude exponent is $\ln \max |\lambda_i(M)|/T$. Its determinant also has an exact check $e^{-2\gamma T}$ from the divergence in the $(x, \dot{x})$ plane. The calculation below integrates both columns by fourth-order Runge–Kutta. Compare the numerical Floquet exponent with the averaged prediction as the modulation depth increases. Both curves describe growth rates, including negative rates for decay.

Open this experiment online

Try the calculation

Fractional stiffness modulation	**0.08**
Damping γ (1/s)	**0.1**
Drive frequency Ω (rad/s)	**20**



At Ω = 20 rad/s, numerical exponent 0.09997 s⁻¹ and averaged exponent 0.1 s⁻¹. Amplitude doubling time 6.934 s.

Tutorial on phase-sensitive amplification

At exact resonance and zero loss, substitute $A = X + iY$ into $\dot{A} = igA^*$. Identify the two special quadrature directions.

▼ **Work through the solution**

Equating real and imaginary parts yields

$$\dot{X} = gY,\ \dot{Y} = gX.$$

Along $Y = X$ both grow at g, whereas along $Y = -X$ they decay at g. These directions reflect how timing the stiffness change can add or remove work on each cycle.

On resonance an oscillator with $\omega_0 = 40\,\mathrm{s}^{-1}$ first grows near $h = 0.03$. Estimate its small-signal damping and its free amplitude e-folding time.

▼ **Work through the solution**

From the weak-modulation threshold $\gamma = h\omega_0/4 = 0.30\,\mathrm{s}^{-1}$. Without modulation the underdamped free amplitude falls as $e^{-\gamma t}$, giving $3.33\,\mathrm{s}$. Threshold uncertainty or nonlinear frequency shifts propagate into this inference.

Further reading · Verhulst, Utrecht University, perturbation analysis of parametric resonance.

A magnetic orbit that spirals outward in a perfect vacuum

Numerical field laboratory

Track a charged particle in a prescribed uniform field, derive a rotation update and separate numerical energy growth from a real magnetic force.

Use the work law before solving the trajectory

A nonrelativistic charged particle of mass m and charge q moves perpendicular to a uniform, time-independent magnetic field $B\hat{\mathbf{z}}$. There is no electric field and the dilute test particles do not affect the applied field. The Lorentz force gives $\dot{v}_x = \omega v_y,\ \dot{v}_y = -\omega v_x$ for $\omega = qB/m$. Dotting with $\mathbf{v}$ gives zero, so the speed should stay constant. This property holds for either sign of the charge; the direction of rotation changes. It is a better initial test of a trajectory solver than how circular a short segment looks on screen.

Define $J(v_x, v_y) = (v_y, -v_x)$, so $J^T = -J$ and $J^2 = -I$. The exact velocity over a time step h is $[\cos(\omega h)I + \sin(\omega h)J]\mathbf{v}$. Integration in time gives a circular orbit of radius $v_\perp/|\omega|$ once the guiding centre is chosen. A magnetic field alone bends a trajectory without changing its kinetic energy; a slowly growing orbit radius at fixed B and fixed initial speed demands an explanation.

Why forward Euler adds energy at every step

Euler gives $\mathbf{v}_+ = (I + dJ)\mathbf{v}$ with $d = \omega h$. Squaring the norm eliminates the cross term because $v^T J v = 0$, leaving $\|\mathbf{v}_+\|^2 = (1 + d^2)\|\mathbf{v}\|^2$. No nonzero step has exact unit gain. A geometric picture makes the bias tangible: a tangent step leaves the initial velocity circle on its outside, whatever the point of departure.

After N steps the speed factor is $(1 + d^2)^{N/2}$. With 24 steps per physical orbit, $d = 2\pi/24 \simeq 0.2618$, so a full orbit multiplies speed by roughly 2.2. At shorter times or smaller steps this can resemble weak physical acceleration. Reducing h at fixed physical duration reduces the growth. Testing one step at fixed count, rather than comparing over a fixed physical interval, underestimates the cumulative error.

Solve a centred magnetic step exactly

Use the time-centred velocity in the Lorentz force:

$$\mathbf{v}_+ - \mathbf{v} = dJ(\mathbf{v}_+ + \mathbf{v})/2.$$

Premultiply by the transpose of $\mathbf{v}_+ + \mathbf{v}$. The left side is the difference of squared speeds and the right side vanishes, so the update preserves speed for every d in exact arithmetic. Rearranging yields a two-dimensional linear solve; using $J^2 = -I$ turns it into a rotation,

$$\mathbf{v}_+ = \frac{1 - t^2}{1 + t^2}\mathbf{v} + \frac{2t}{1 + t^2}J\mathbf{v},$$

$$t = d/2.$$

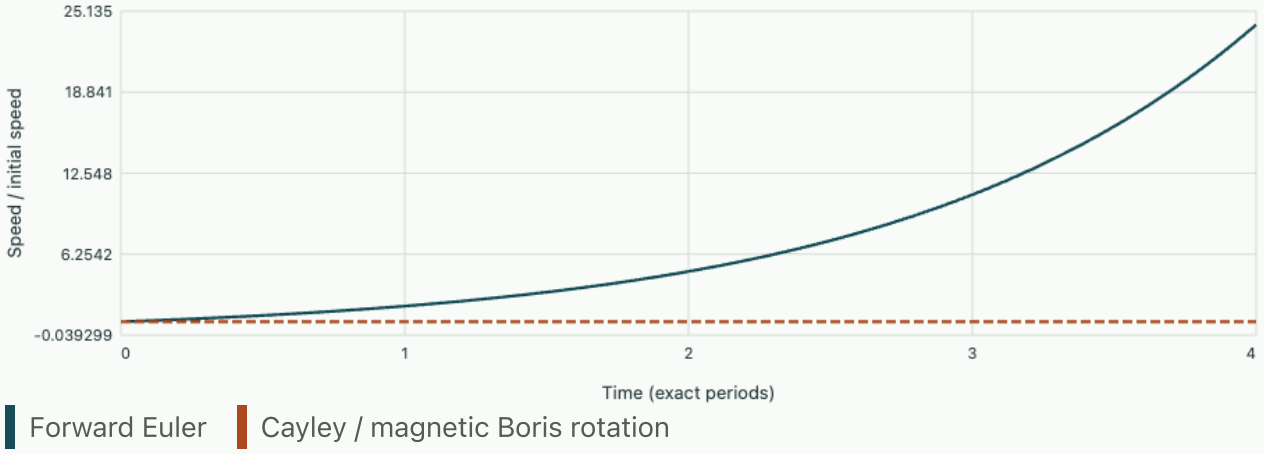
These coefficients are $\cos\alpha$ and $\sin\alpha$ for $\alpha = 2\arctan(d/2)$. This is the magnetic rotation used in the nonrelativistic Boris integrator. For simultaneous electric fields, the usual split step adds two half electric impulses around the magnetic rotation. The current experiment isolates the magnetic part; it makes no claim about a nonuniform electric field or the accuracy of interpolation from a field mesh to particles.

Open the live experiment

Try the calculation

Steps per exact orbit 24

Number of orbits 4



After 4 orbit(s), Euler speed 24.096, rotation speed 1. Rotation still lags the exact phase by 8.1411°.

Speed conservation does not fix arrival time

At small d,

$$\alpha = d - d^3/12 + O(d^5).$$

A speed-preserving orbit lags in gyrophase, even when its energy plot stays perfectly flat. After four true orbits at 24 steps per orbit the predicted lag is about 8.14 degrees. The browser advances the Euler and rotation velocities and reports both norm and lag. The y-axis rescales to fit both curves, so a sharply growing Euler error can flatten the appearance of the well-behaved curve. Inspect numerical values as well as the picture.

Suppose an optical pickup observes the time at which a dilute ion tracer first crosses a known line of sight after several orbits. A phase error can be an important measurement error even when the speed is exact. To use the trajectory as a reference, refine the step until the crossing time changes by less than the experimental timing resolution. If the required integration spans many revolutions, a uniform field

permits exact substeps. In realistic fields one can retain a structure-preserving rotation while resolving spatial variation independently.

A separate diagnostic is time reversal. Starting from the final velocity and applying the rotation with $-d$ exactly inverts a centred step, up to rounding. Forward Euler with $-d$ after $+d$ gives

$$(I - dJ)(I + dJ) = (1 + d^2)I.$$

A forward/backward test detects the lost reversibility without knowing a single experimental constant.

Tutorial

Given an orbit with 48 steps per period, estimate the gyrophase error after ten periods using the leading asymptotic term.

▼ **Work through the solution**

There are 480 steps with $d = 2\pi/48$. The accumulated lag is about $480d^3/12 \simeq 0.0897$ rad, or 5.14 degrees. Doubling the number of steps per period reduces the leading accumulated error fourfold.

A computation preserves speed but reverses the sign of its orbit relative to observation. Would the magnetic work test detect the error?

▼ **Work through the solution**

Both field signs give zero work. Check a signed initial acceleration component $a_y = -qBv_x/m$ for a positive initial v_x, or observe the oriented trajectory. A scalar invariant cannot detect all vector-sign errors.

What changes in the relativistic case with appreciable kinetic energy?

▼ **Work through the solution**

Evolve momentum $\mathbf{p} = \gamma m\mathbf{v}$ and infer velocity from momentum. In a uniform magnetic field without electric work γ stays constant, and the orbital frequency is reduced to $|q|B/(\gamma m)$. An energy-preserving nonrelativistic integration at the wrong frequency would still give the wrong measured period.

Relativistic particle dynamics and the field tensor

Waves, radiation and relativity

A circular track measures momentum per charge. A clock following the particle registers proper time, and a detector records how much kinetic energy it has gained. We will put these observations in one dynamical equation, solve that equation in uniform fields and exhibit a numerical pusher that can get the energy right while its orbit phase slowly slips.

What the track, clock and calorimeter measure

Consider a particle of invariant mass m, signed charge q and ordinary velocity $\mathbf{v}$. Inertial observers agree on the proper time between neighbouring events on its timelike worldline,

$$d\tau = dt\sqrt{1 - v^2/c^2}.$$

Adopt coordinates $x^\mu = (ct, x, y, z)$ and metric $g = \mathrm{diag}(1, -1, -1, -1)$. Dividing the displacement by proper time produces a vector of fixed Minkowski norm. Multiplying it by invariant mass gives four-momentum.

$$u^\mu = \gamma(c, \mathbf{v}),$$
$$p^\mu = m u^\mu = (\mathcal{E}/c, \mathbf{p}),$$
$$\mathcal{E}^2 - c^2|\mathbf{p}|^2 = m^2 c^4.$$

Energy includes the rest term. Momentum is $\gamma m \mathbf{v}$. Differentiating the last equation on a trajectory gives

$$d\mathcal{E}/dt = \mathbf{v} \cdot d\mathbf{p}/dt.$$

This identity replaces the nonrelativistic step of dotting $m\mathbf{a}$ with velocity. A magnetic field can rotate momentum with fixed magnitude and energy, whereas an electric field can change energy by doing work.

Suppose an electron begins at rest and gains 10.0 keV of kinetic energy across an electrostatic potential difference in vacuum. Its rest energy is 511 keV, giving $\gamma = 1.01957$ and $v/c = \sqrt{1 - 1/\gamma^2} = 0.1950$. The nonrelativistic estimate is 0.1978. In a modest energy measurement the discrepancy could lie below the experimental spread. For a precise orbit calculation the 1.96 percent change in relativistic cyclotron period is measurable independently of that discrepancy in speed.

A linear map that produces the Lorentz force

Seek a local equation linear in field and four-velocity and covariant under boosts. For the energy component to give $q\mathbf{E} \cdot \mathbf{v}$, and for the spatial components to give the electric force and the oriented cross product, construct the antisymmetric field tensor. With the stated metric and $\epsilon^{123} = 1$, set

$$F^{0i} = -E_i/c,$$

$$F^{i0} = E_i/c,$$

$$F^{ij} = -\epsilon^{ijk} B_k,$$

$$\frac{dp^\mu}{d\tau} = qF^{\mu\nu} u_\nu.$$

Lowering the index on u reverses the signs of its spatial components. The spatial product becomes $q\gamma(\mathbf{E} + \mathbf{v} \times \mathbf{B})$, and division by γ converts the proper-time derivative to $d\mathbf{p}/dt$. The time component similarly yields

$$d\mathcal{E}/dt = q\mathbf{E} \cdot \mathbf{v}.$$

Contract the force with u_μ. The symmetric product $u_\mu u_\nu$ annihilates the antisymmetric tensor, which proves that the evolution preserves the mass shell.

To recover Maxwell's kinematic constraints from potentials, put $A^\mu = (\phi/c, \mathbf{A})$ and $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$. Then $F^{0i} = (\partial_t A_i + \partial_i \phi)/c = -E_i/c$, while the spatial-spatial components are the magnetic curl. Adding a four-gradient to A^μ leaves the tensor unchanged because ordinary derivatives commute. The particle action $S = -mc^2 \int d\tau - q \int A_\mu dx^\mu$ gives the same force after integrating variations by parts with fixed endpoints. Gauge changes add only an endpoint term to this action. A trajectory measured with local forces cannot distinguish such potential representatives.

Compare the potentials and field-based description in the field-tensor chapter

Constant force, variable coordinate acceleration

Put a particle initially at rest in a uniform static electric field along x. Let $a_0 = qE/m$ be signed. Integration of $dp_x/dt = qE$ gives $p_x = ma_0 t$, so the speed tends to c even while momentum increases linearly. Proper-time evolution is simplest in terms of rapidity η, where $u^0 = c\cosh\eta$, $u^x = c\sinh\eta$. The two tensor equations imply $d\eta/d\tau = a_0/c$.

$$t = \frac{c}{a_0}\sinh(a_0\tau/c),$$

$$x = \frac{c^2}{a_0}\{\cosh(a_0\tau/c) - 1\},$$

$$\frac{dv_x}{dt} = \frac{a_0}{\gamma^3}.$$

The initial time and position are zero. The work $qEx = mc^2(\cosh\eta - 1)$ equals kinetic energy, which checks both the integration constant and the dimensional factor in the displacement. For an electron use the direction of the electric force as positive to simplify signs. At 20.0 kV/m it gains 1.0 keV in 5.0 cm. The energy gain follows from the static potential even though the speed does not grow exactly as

the square root of distance. The field region must last long enough for the particle to cross it and fringe fields can alter the entry trajectory.

Compare momentum, coordinate time and proper time in the charged-worldline experiment

Helical motion, radius and period

A uniform magnetic field along z keeps p_z, energy and $|\mathbf{p}_\perp|$ fixed. Write $p_x + ip_y$ as a complex number. The Lorentz force reduces to a first-order rotation equation and a single time integration gives the corresponding position. With positive B, positive charges starting along $+x$ initially turn towards $-y$.

$$\frac{d}{dt}(p_x + ip_y) = -i\frac{qB}{\gamma m}(p_x + ip_y),$$

$$R = \frac{p_\perp}{|q|B},$$

$$T = \frac{2\pi\gamma m}{|q|B}.$$

The helical pitch is $v_z T$. Radius gives transverse momentum and a calibrated period gives total energy through γ. A track seen only in projection along the field gives no direct evidence that longitudinal momentum vanishes. A second view or a timing measurement can resolve this ambiguity.

At 10.0 keV the electron has $pc = \sqrt{K(K + 2mc^2)} = 101.6$ keV. With entry perpendicular to a 2.0 mT field its radius is 0.169 m and its period 18.2 ns. Doubling the field halves both radius and period at fixed injection energy. Doubling momentum at fixed field doubles radius but changes period by the ratio of Lorentz factors. These are different tests of a track calibration. A beam with a spread of energies will dephase relative to a narrow timing gate even when every single particle remains in a perfectly uniform field.

A drift frame gives an exact crossed-field solution

For constant crossed fields with $E < cB$, transform to the frame moving at $\mathbf{v}_d = \mathbf{E} \times \mathbf{B}/B^2$. There $E' = 0$ and $B' = \sqrt{B^2 - E^2/c^2}$; calculate the transformed initial momentum and integrate a helix. Transform two successive equal-phase events back to the lab. Their perpendicular displacement in the electric-free frame is zero. If the motion has no component parallel to $\mathbf{B}$, the entire displacement there is zero, and the average lab cycle velocity is exactly $\mathbf{v}_d$. The detailed lab trace need not be circular and the instantaneous lab energy may oscillate.

Take $E = 30.0$ kV/m along y and $B = 0.50$ mT along z. The drift is along $+x$ at 6.00×10^7 m/s. A particle injected at that velocity crosses the uniform fields on a straight line because the forces cancel. An initially stationary positive particle instead begins by accelerating along $+y$. These observations can use the same field with different injected particles. If E exceeds cB, this electric-free frame is unavailable. Transform to a magnetic-free frame and integrate electric acceleration instead.

Continuous rotation and a discrete experiment

A naive forward step $\mathbf{p}_{n+1} = \mathbf{p}_n + q\Delta t\, \mathbf{v}_n \times \mathbf{B}$ adds a vector perpendicular to the previous momentum. Its squared norm increases by the squared magnitude of that kick. The apparent heating will persist in perfectly static uniform magnetic fields. For a purely magnetic update, evaluate the direction using the mean of old and new momenta. At fixed γ, the resulting implicit two-dimensional linear map can be solved algebraically and is an exact rotation by a modified angle.

$$\delta = 2\arctan\left(\frac{\omega_c \Delta t}{2}\right),$$

$$\omega_c = \frac{qB}{\gamma m}.$$

The sign of rotation follows the earlier complex equation. Dotting the update with the mean momentum gives zero change in squared norm, which verifies the energy property even for relatively large time steps. A common electromagnetic pusher surrounds the magnetic rotation by two electric half-kicks. In spatially varying fields, specify at what time and position each field is sampled, and check spatial interpolation and time convergence. For ultrarelativistic nearly cancelling electric and magnetic fields, preserving rotation in a pure field alone does not ensure cancellation is represented accurately.

Solve the constant tensor without transforming every time step

When uniform electric and magnetic fields are constant in spacetime, the Lorentz equation is a linear constant-coefficient equation for four-velocity as a function of proper time. Nonlinearity in ordinary time hides that simple structure, because the rate at which proper time accumulates changes along the orbit. Define the mixed tensor $M^{\mu}{}_{\nu} = (q/m)F^{\mu\alpha}g_{\alpha\nu}$. In a uniform field it does not vary along the trajectory. A formal solution is then

$$u(\tau) = e^{M\tau}u(0),$$

$$x(\tau) - x(0) = \int_0^{\tau} e^{Ms}u(0)\, ds.$$

The coordinate formula is written as an integral since M can be singular, as it is in a purely magnetic field. Blindly replacing this integral by $M^{-1}(e^{M\tau} - I)$ fails in exactly such simple cases. A power series for the integral or a block matrix exponential treats zero eigenvalues without an artificial inverse. Antisymmetry of the tensor implies $M^{\mathsf{T}}g + gM = 0$. Differentiating $e^{M^{\mathsf{T}}\tau}ge^{M\tau}$ shows that it stays equal to g, making preservation of $u^{\mu}u_{\mu} = c^2$ a structural property of the exact mapping.

For parallel constant fields along z, the matrix decomposes into a boost in the ct, z subspace and a rotation in x, y. Write $\alpha = qE/(mc)$, $\Omega = qB/m$, and denote the initial four-velocity components by u_0^{μ}. The proper-time solution is

$$u^0 = u_0^0 \cosh \alpha\tau + u_0^z \sinh \alpha\tau,$$
$$u^z = u_0^z \cosh \alpha\tau + u_0^0 \sinh \alpha\tau,$$
$$u^x + i u^y = (u_0^x + i u_0^y) e^{-i\Omega\tau}.$$

The longitudinal and transverse operations commute because they act in different subspaces. In lab time the transverse rotation slows as the electric field increases energy, although the proper-time rotation frequency stays constant. A particle therefore traces a helix whose transverse radius $p_\perp/(|q|B)$ is fixed while the pitch evolves. A single measured turn time without a tagged position in the accelerating region is insufficient for an energy reconstruction.

As a numerical test use an electron, quoting magnitudes, with $B = 1.0$ mT, $E = 10.0$ kV/m and initial purely transverse speed $0.10c$. One magnetic turn takes 35.72 ns in proper time and accumulates electric rapidity magnitude $|\alpha|\tau \simeq 0.2096$. The final Lorentz factor is $\gamma_i \cosh(0.2096) = 1.0272$, giving final kinetic energy about 13.9 keV, compared with 2.57 keV initially. The lab-time turn duration is $\gamma_i \sinh(|\alpha|\tau)/|\alpha| \simeq 36.2$ ns. Reversing the electric field changes the direction of the axial displacement but leaves final energy and turn duration unchanged for this initially transverse particle. Reversing the initial longitudinal velocity would generally break that energy symmetry.

Infer uncertainty and test units in a momentum reconstruction

A track fitter commonly reports field, radius, pitch, time and a fitted covariance matrix. The dynamical relation turns these observed quantities into momentum. Let L_z denote a signed pitch per turn. With known signed direction of circulation and calibrated absolute field magnitude,

$$|\mathbf{p}| = |q| B \sqrt{R^2 + (L_z/2\pi)^2}.$$

For small errors, logarithmic differentiation gives weights for fractional radius and pitch errors:

$$d\log p = d\log B + w_R\, d\log R + w_L\, d\log |L_z|,$$
$$w_R = \frac{R^2}{R^2 + (L_z/2\pi)^2},$$
$$w_L = 1 - w_R.$$

For $R = 0.20$ m and $L_z = 0.30$ m the radius accounts for 94.6 percent of the squared transverse-plus-longitudinal geometric factor. Independent relative uncertainties of 0.20 percent in field, 0.50 percent in radius and 2.0 percent in pitch yield about 0.525 percent uncertainty in total momentum. A common detector scale controlling both radius and pitch would enter with correlated signs and should not be counted as two independent errors. When the pitch tends to zero, use the original derivatives with respect to L_z; a reported fractional error on a quantity consistent with zero is ill defined.

Differentiating $K = \sqrt{m^2c^4 + p^2c^2} - mc^2$ gives

$$d\log K = (\gamma + 1) d\log p / \gamma.$$

The coefficient approaches two at low energy and one at very high energy. At the approximately 3.70 keV energy of the tutorial the above momentum error would imply about a 1.05 percent energy error. A kinetic-energy estimate from a small difference of nearly equal numbers can also lose numerical precision at much lower energies. Rationalise it as $K = p^2c^2/(\sqrt{m^2c^4 + p^2c^2} + mc^2)$ when working near rest.

Several independent checks use this reconstruction without assuming the track fitter is correct. The turn time inferred from the fitted γ can be compared with a separate timing channel. Reverse the magnetic field to reverse the direction of curvature while keeping the speed and energy approximately fixed. Translate the beam within a mapped uniform field and look for changing fitted radius that correlates with position, an indication of geometric or field-map bias. Finally plot the residual of the magnetic force law separately for transverse magnitude and angle. A time integrator may conserve momentum magnitude for symmetry reasons even when its field interpolation biases the phase or shifts the guiding centre systematically.

Advance a charge using relativistic momentum

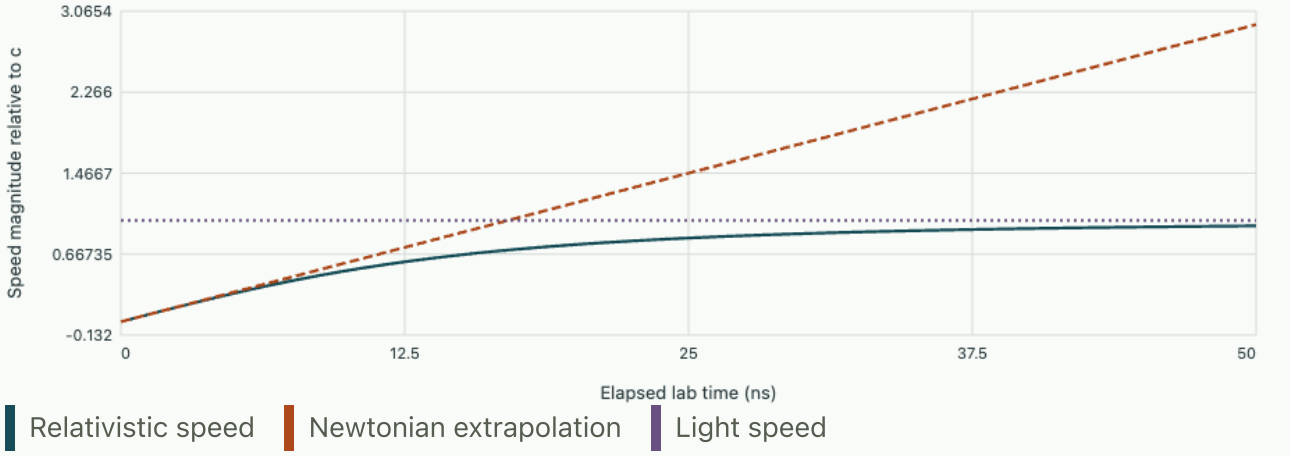
Increase the impulse while following speed, energy and proper time. Identify the range where a Newtonian momentum update becomes inaccurate.

Open this experiment online

Try the calculation

Electric field magnitude (kV/m) **100**

Lab time for a reading (ns) **17**



At 17 ns laboratory time the electron speed is 0.706 c. Its clock has advanced 15 ns and it has gained 211 keV.

Worked timing problem · infer an accelerating voltage without a radius

A tagged electron bunch completes ten nearly circular turns in a calibrated 5.0 mT magnetic field. Two time stamps indicate an elapsed 73.0 ns. The field is perpendicular to the orbit. Assuming complete turns and negligible electrical work during the timed interval, estimate the injection energy. What timing precision would distinguish a one-percent change in that energy?

▼ Use the relativistic period sensitivity rather than the energy fraction directly

The nonrelativistic turn period is

$$2\pi m/(eB) = 7.145\,\mathrm{ns}.$$

The measured period is 7.30 ns, giving $\gamma \simeq 1.0217$ and $K = (\gamma - 1)mc^2 \simeq 11.1$ keV. The relative energy sensitivity is much larger than the relative period change when γ is close to one. A one-percent change in K corresponds to $\Delta\gamma \simeq 2.17 \times 10^{-4}$, or relative period change 2.12×10^{-4}. Over the ten turns this gives approximately 15.5 ps.

To resolve that change spectroscopically, uncertainties in the elapsed time and field scale must together be significantly smaller than the equivalent change. A constant pickup cable delay cancels if the same electronics register both ends of the interval, while a drift of its timebase scale does not. Timing over more turns improves the total accumulated signal but also increases exposure to field inhomogeneity, energy loss and dephasing among particles with different injection energies. A narrower timing histogram does not automatically eliminate a systematic field calibration.

Finally verify the integer turn count independently. Missing a pickup crossing would cause a much larger change in inferred energy than the desired resolution. The measured wave packet may broaden so strongly that later crossings overlap. Recording the full pickup waveform and tracking ordered crossings, along with an independent rough radius estimate, can constrain the turn assignment before fitting the small relativistic shift. A perfect circular image with no time labels cannot supply that missing count.

Why momentum and kinetic energy do not gain fixed fractions together

In an electrostatic accelerating gap, the kinetic energy increment for a particle passing between two equipotential surfaces is its charge times the potential drop, regardless of injection energy. This statement follows by integrating the Lorentz power law over the actual path. A magnetic guiding field can alter the path and transit time while doing no direct work. A time-dependent electric field, a missed equipotential endpoint or appreciable radiation changes the simple inference and must be addressed separately.

For an electron entering with 10 keV and acquiring an additional 10 keV, kinetic energy doubles. Its initial momentum times c is about 101.6 keV and its final value about 144.4 keV. The radius in the same subsequent perpendicular magnetic field increases by a factor approximately 1.42. Doubling the radius would instead require doubling momentum, which demands a substantially greater energy increment at this low initial energy. Comparing radii before and after a known voltage supplies an independent check on the sign and magnitude of a particle's charge-to-mass model.

A measured energy gain below the nominal voltage does not by itself diagnose radiation. The potential actually traversed, finite angular acceptance, entrance or exit fringe fields, collision losses and biases in reconstructing longitudinal momentum can all matter. Repeat with reversed field polarity or several injection energies and compare the kinetic-energy difference rather than the two radii directly.

Tutorial · Reconstruct a track with two measurements

A particle known to be an electron or positron crosses a uniform $+z$ field of 1.0 mT. The projected radius is 0.200 m and the pitch magnitude is 0.300 m. Recover the transverse and longitudinal momentum, energy and period. What else must the detector record to determine the sign of charge?

▼ **Reconstruct both momentum components**

The transverse momentum is $|q|BR = 3.20 \times 10^{-23}$ kg m/s. During a cycle the parallel displacement is $p_z T/(\gamma m) = 2\pi p_z/(|q|B)$, so $|p_z| = 7.65 \times 10^{-24}$ kg m/s. The quadrature total is 3.29×10^{-23} kg m/s, yielding $p/(mc) = 0.1206$, $\gamma = 1.00724$, and approximately 3.70 keV kinetic energy. The period is about 36.0 ns.

Neither radius nor pitch magnitude carries the signed circulation. Record the direction of traversal relative to the calibrated direction of magnetic field. Reversing both charge and the direction of traversal yields the same unlabelled transverse ring. If timing or two ordered detectors supply a direction, the sign follows from the first Lorentz kick.

Tutorial · Can a constant-energy circle be trusted?

A pusher uses 20 uniform steps per exact cyclotron period in a uniform magnetic field. It reproduces initial kinetic energy to machine precision for 50 turns. Estimate the phase drift from the midpoint rotation formula. Design a refinement diagnostic that distinguishes a time-discretisation bias from a field calibration error.

▼ **Compare the exact and discrete angles**

The exact unsigned angle per step is

$$h = 2\pi/20 = 0.31416.$$

The discrete one is

$$2\arctan(h/2) = 0.31161.$$

The accumulated deficit over 1,000 steps is about 2.55 rad, roughly 0.41 turn. The method can keep every track point on a closed momentum circle yet miss the pickup electrode by a large fraction of a revolution. Coordinate radius may depend on how positions were staggered, so perform the angular comparison on momentum first.

Double the number of steps at fixed physical end time and take the difference in unwrapped phases. The leading angular error $h^3/12$ per step becomes a quadratic global error, so halving the step should reduce it by about four. A calibration error in $qB/(\gamma m)$ persists under time refinement. Track both the expected frequency and the convergence of the time series; comparing the coordinates at a single late time can confuse complete turns with near agreement.

Cambridge Electromagnetism · four-vectors, fields and charged matter

Euler–Poinsot motion, torque-free nutation and the intermediate axis

Analytical mechanics from constraints to invariants

A freely rotating body has fixed angular momentum in space, yet the components recorded by three body-mounted gyroscopes can vary periodically or linger near an unstable rotation. Solve the nonsymmetric top from its invariants, distinguish its body-component period from its spatial orientation, and extract a stability time scale from a measured inertia tensor.

Derive the body equations from a spatial conservation law

Choose principal axes fixed in a rigid body and label its moments $I_1 < I_2 < I_3$. In the co-rotating basis the angular-momentum components are

$$L_i = I_i \omega_i .$$

For any vector expressed in that basis, its inertial derivative is its component derivative plus $\boldsymbol{\omega} \times \mathbf{L}$. With no external torque the inertial derivative is zero. Taking components gives

$$I_1 \dot{\omega}_1 = (I_2 - I_3)\omega_2 \omega_3 ,$$
$$I_2 \dot{\omega}_2 = (I_3 - I_1)\omega_3 \omega_1 ,$$
$$I_3 \dot{\omega}_3 = (I_1 - I_2)\omega_1 \omega_2 .$$

Multiply these equations by the corresponding ω_i and sum: the coefficients of $\omega_1\omega_2\omega_3$ cancel, giving constant $E = \sum I_i \omega_i^2 / 2$. Multiplying instead by $I_i \omega_i$ gives constant $L^2 = \sum I_i^2 \omega_i^2$. In body angular-momentum space the motion lies on the intersection of a sphere and an energy ellipsoid. At fixed L, the possible energies range from $L^2/(2I_3)$ to $L^2/(2I_1)$. The intermediate value $L^2/(2I_2)$ divides two families of closed curves and contains the unstable principal-axis orientations.

In Poinsot's construction the inertia ellipsoid in angular-velocity space rolls on an invariable plane normal to the fixed spatial angular momentum. The ellipsoid corresponding to a given energy is

$$\sum I_i \omega_i^2 = 2E .$$

Its normal at the instantaneous angular velocity is proportional to $\mathbf{L}$; the tangent plane obeys $\mathbf{L} \cdot \mathbf{x} = 2E$. In space that plane therefore has fixed distance $2E/L$ from the origin. Its instantaneous point in the body along $\boldsymbol{\omega}$ has zero instantaneous rotational velocity because the vector cross product of angular velocity with itself is zero. These facts underlie the rolling construction. A photographed body axis traces a different curve from the moving contact point or from a body-frame gyro trace.

Test all three principal rotations locally

For rotation $(0, 0, \Omega)$, linearise the first two Euler equations in transverse perturbations. Eliminating one component gives

$$\ddot{\omega}_1 = -\Omega^2[(I_3 - I_2)(I_3 - I_1)/(I_1 I_2)]\omega_1.$$

Small transverse deviations oscillate. Performing the analogous process about I_1 also gives a restoring sign. About $(0, \Omega, 0)$ the sign reverses and the growing exponent is $\lambda = |\Omega|\sqrt{(I_2 - I_1)(I_3 - I_2)/(I_1 I_3)}$. The unstable rotation is still an exact solution when both perturbations vanish identically. Instability describes the evolution of arbitrarily small disturbances, not a spontaneous violation of the torque-free equations.

For moments proportional to $1, 2, 3$ and intermediate-axis rotation $\Omega = 6.0\,\mathrm{rad/s}$, $\lambda = 3.46\,\mathrm{s}^{-1}$ and the linear e-folding time is $0.289\,\mathrm{s}$. A perturbation with a growing-mode component 10^{-3} of the spin reaches roughly 10^{-1} in $\ln(100)/\lambda = 1.33\,\mathrm{s}$. Its nonlinear fate remains bounded by the momentum sphere and energy ellipsoid. An exponential growth fit carried past the first large reorientation would therefore misrepresent the global motion. Initial perturbations projected entirely on the decaying linear eigenvector can delay the first visible flip.

Solve the low-energy family with elliptic functions

For $L^2/(2I_3) < E < L^2/(2I_2)$ the trajectory surrounds the largest-inertia axis. Start at a crossing with $\omega_2 = 0, \omega_1 = A > 0, \omega_3 = C > 0$. Solving the two invariants gives

$$A^2 = (2EI_3 - L^2)/[I_1(I_3 - I_1)],$$
$$C^2 = (L^2 - 2EI_1)/[I_3(I_3 - I_1)].$$

When ω_1 reaches zero the other transverse component has maximum $B^2 = (2EI_3 - L^2)/[I_2(I_3 - I_2)]$. Use $s = \omega_2/B$. Eliminating the other squares with the invariants gives

$$\omega_1^2 = A^2(1 - s^2),$$
$$\omega_3^2 = C^2(1 - k_e^2 s^2).$$

Euler's second equation now reduces to

$$\dot{s}^2 = \nu^2(1 - s^2)(1 - k_e^2 s^2),$$
$$\nu^2 = \frac{(I_3 - I_2)(L^2 - 2EI_1)}{I_1 I_2 I_3},$$
$$k_e^2 = \frac{(I_2 - I_1)(2EI_3 - L^2)}{(I_3 - I_2)(L^2 - 2EI_1)}.$$

Separating variables defines $u = \nu t = \int_0^s dw/\sqrt{(1 - w^2)(1 - k_e^2 w^2)}$. Its inverse is Jacobi's $\mathrm{sn}(u, k_e)$. The continuously signed square roots are cn and dn, giving $(\omega_1, \omega_2, \omega_3) = (A\,\mathrm{cn}, B\,\mathrm{sn}, C\,\mathrm{dn})$. This construction defines the special functions from the mechanical quadrature instead of assuming them as a table of answers. At $k_e = 0$ the inverse is a sine and the stabilising motion is harmonic. As $k_e \to 1$, the integral to $s = 1$ diverges logarithmically and the period grows without bound.

The body-frame period is

$$T_b = 4K(k_e)/\nu,$$

with K the same integral evaluated at one. Some software defines elliptic functions in terms of parameter $m = k_e^2$ rather than modulus k_e. Supplying one to an implementation expecting the other produces a plausible curve with the wrong period. Check the derivative at the zero crossing $\dot{\omega}_2(0) = B\nu$ and the time of its first maximum K/ν before fitting a long trace. The third component dn repeats after half the full period, so a single principal-axis gyro can report a dominant frequency twice that of the full body vector.

Reconstruct the observed nutation angle

Orient the inertial vertical along $\mathbf{L}$. The angle between body axis three and this fixed direction obeys $\cos\theta = L_3/L = I_3\omega_3/L$. For $I_1, I_2, I_3 = (1, 2, 3)\,\mathrm{kg\,m^2}$ and initial $\boldsymbol{\omega} = (0.50, 0, 1.00)\,\mathrm{rad/s}$, the invariants give $L^2 = 9.25$, $2E = 3.25$ in SI, $A = B = 0.50\,\mathrm{rad/s}$, $C = 1.00\,\mathrm{rad/s}$, $\nu = 1.00\,\mathrm{s^{-1}}$, and $k_e^2 = 1/12$. Thus the nutation angle changes from $\arccos(3/\sqrt{9.25}) \approx 9.46^\circ$ to $\arccos[3\sqrt{11/12}/\sqrt{9.25}] \approx 19.2^\circ$. A torque is unnecessary for this change in the angle of a body axis relative to fixed $\mathbf{L}$.

The principal-axis components alone determine body orientation only when supplemented with an initial attitude and the kinematic equation. For the Euler-angle convention with polar axis $\mathbf{L}$, the azimuthal rate of body axis three is $\dot{\phi} = L(L_1^2/I_1 + L_2^2/I_2)/(L_1^2 + L_2^2)$ away from the polar singularity. It follows by differentiating the body-axis unit vector in space and projecting onto its azimuthal direction. A complete period of the body components therefore leaves a finite spatial rotation about fixed $\mathbf{L}$. Return of the three gyro readings does not imply that a painted mark has returned to its starting position in the laboratory.

Solve the high-energy family and join both sides at the separatrix

For $L^2/(2I_2) < E < L^2/(2I_1)$ the closed curves on the angular-momentum sphere surround axis one. The component ω_1 now keeps its sign while ω_3 changes sign through a cycle. As in the first family, choose a phase with $\omega_2 = 0$ and both other components positive. The two invariants give the same initial $A^2 = (2EI_3 - L^2)/[I_1(I_3 - I_1)]$ and $C^2 = (L^2 - 2EI_1)/[I_3(I_3 - I_1)]$. Now the maximum of ω_2 occurs when $\omega_3 = 0$, so

$$B^2 = (L^2 - 2EI_1)/[I_2(I_2 - I_1)].$$

Which component reaches zero is the physical branch choice that determines the arrangement of elliptic functions.

Let $s = \omega_2/B$. The invariant equations give

$$\omega_3^2 = C^2(1 - s^2),$$
$$\omega_1^2 = A^2(1 - k_h^2 s^2).$$

Substitution in the second Euler equation yields the same elliptic quadrature but with $\nu_h^2 = (I_2 - I_1)(2EI_3 - L^2)/(I_1 I_2 I_3)$ and $k_h^2 = (I_3 - I_2)(L^2 - 2EI_1)/[(I_2 - I_1)(2EI_3 - L^2)]$. All components are given by $(A\,\mathrm{dn}, B\,\mathrm{sn}, C\,\mathrm{cn})(\nu_h t, k_h)$, where the notation means each function is evaluated at the same time and modulus. On this energy side $0 < k_h < 1$. The low-energy modulus formula evaluated here would exceed one, signalling that the original choice of component scaling and physical extrema must be changed.

For the same inertias $1, 2, 3$, choose initial $(1, 0, 0.20)\,\mathrm{rad/s}$. Then $L^2 = 1.36,\ 2E = 1.12$, and the high-energy inequality holds since $2E > L^2/I_2 = 0.68$. The formulas give

$$A = 1,\ B = \sqrt{0.12} = 0.3464,\ C = 0.20,\ \nu_h = \sqrt{1/3} = 0.5774\,\mathrm{s}^{-1},\ k_h^2 = 0.12.$$

The minimum ω_1 is $A\sqrt{1 - k_h^2} = 0.9381\,\mathrm{rad/s}$, so that principal gyro never reverses even though the third one does. A body-marking aligned with axis three can undergo a conspicuous reversal while the dominant axis-one spin changes only modestly. These component predictions are more specific than an energy or angular-speed trace.

At $2E = L^2/I_2$, both moduli approach one. The definitions give

$$\mathrm{sn}(u, 1) = \tanh u,$$
$$\mathrm{cn}(u, 1) = \mathrm{dn}(u, 1) = \mathrm{sech}\, u.$$

One separatrix solution therefore has $\omega_2 = B\tanh(\nu t)$ with $\omega_1 = A\,\mathrm{sech}(\nu t)$ and $\omega_3 = C\,\mathrm{sech}(\nu t)$, up to signs and time shifts consistent with Euler's equations. It connects opposite steady orientations along axis two as t goes from negative to positive infinity. Its asymptotic approach is exponential, with ν equal to the small-disturbance growth exponent at the intermediate-axis equilibrium after inserting $B = L/I_2$. A nearby periodic orbit spends much of its time near the unstable spin and flips more rapidly between those intervals.

Reconstruct spatial attitude without a singular azimuth

Let $R(t)$ map body components into a laboratory basis. For body angular velocity $\boldsymbol{\omega}_b$, its kinematic equation is

$$\dot{R} = R[\boldsymbol{\omega}_b]_\times,$$

where the skew matrix acts by a cross product. Starting from measured $R(0)$, propagate by multiplying by the rotation appropriate to each short time interval. Exponentiating the skew matrix for an interval with constant estimated body angular velocity preserves orthogonality and determinant exactly.

Applying a naive forward-Euler matrix addition $R + hR[\omega]_\times$ generally stretches the basis, even if the angular velocity components came from an exact elliptic solution.

For a constant rotation about axis three at one radian per second, a single Euler step of duration $0.10\,\mathrm{s}$ takes a transverse unit vector to one of length $\sqrt{1 + 0.10^2} = 1.00499$. After one hundred such steps the transverse length would grow by $(1.01)^{50} \approx 1.64$. A quaternion or matrix exponential avoids that artificial stretch. Normalising each axis separately is insufficient if the axes lose orthogonality; check $R^T R$ against identity and $\det R$ against one. Also compute $R(I_1\omega_1, I_2\omega_2, I_3\omega_3)^T$ at every time. It should give the same spatial vector in a torque-free run.

A measured attitude time series allows a different inference from fitting Euler angles separately. Use relative rotations between successive images to estimate angular velocity in the correct body or spatial frame. A camera can momentarily lose sensitivity to yaw when a marked axis points towards it; simultaneous orthogonal views or multiple non-collinear marks avoid that degeneracy. The gyro and attitude measurements together distinguish a true physical precession from an Euler-angle wrap by 2π. Synchronise their clocks before comparing the phase of a nutation extremum, since even a small time delay produces a plausible-looking mismatch near a fast flip.

Infer inertia ratios from a free-rotation record

A calibrated triad records $\omega_i(t)$ in axes thought to be principal. Euler's equations predict that $\dot\omega_1/(\omega_2\omega_3)$ is the constant $(I_2 - I_3)/I_1$ wherever the denominator is appreciable, and similarly for cyclic indices. Instead of dividing noisy derivatives near zeros, integrate each component over a time interval. For example $\omega_1(t_2) - \omega_1(t_1) = a_1 \int_{t_1}^{t_2} \omega_2\omega_3\,dt$. Fit a_1 over several intervals and repeat for a_2, a_3. Select intervals with different angular configurations so that products do not all average nearly to zero. The ratios depend on inertia shape; multiplying every moment by a common constant leaves a torque-free angular-velocity record unchanged.

For a record giving $a_1 = -1.0$, $a_2 = 1.0$ and $a_3 = -1/3$, solve the homogeneous equations $I_2 - I_3 = -I_1$, $I_3 - I_1 = I_2$, $I_1 - I_2 = -I_3/3$. They give $I_1 : I_2 : I_3 = 1 : 2 : 3$. An independent angular-momentum or torque calibration is needed to set the absolute scale. A known short applied torque impulse that changes one principal angular velocity by a measured amount can do this, provided the other gyroscopic contributions over that impulse are small or accounted for. A fit that estimates all absolute inertias from angular velocity alone can have an apparently excellent residual and still be structurally underdetermined.

If the sensors are not along principal axes the inertia matrix contains off-diagonal entries. The coordinate-free body equation is

$$I\dot{\boldsymbol{\omega}} + \boldsymbol{\omega} \times (I\boldsymbol{\omega}) = 0.$$

Each measured time sample gives equations linear in the constant entries of the symmetric inertia tensor. With sufficiently rich motion one may estimate these entries up to scale, then diagonalise the fitted tensor to recover principal directions. Pure rotation along one axis gives insufficient excitation to

reveal three distinct directions. Comparing a family of initial spins with a single tensor provides a stronger validity test than fitting a different set of effective principal inertias to each record.

Relate a body-frame period to a measured geometric phase

After a closed polhode in body angular-momentum space the laboratory orientation differs by a rotation about the fixed spatial momentum. That phase can be separated into a term accumulated at a dynamical rate and a term depending on the oriented solid angle of the body momentum path. For a fixed L, the angular-velocity component along the momentum is $\boldsymbol{\omega} \cdot \hat{L} = 2E/L$, since $\mathbf{L} \cdot \boldsymbol{\omega} = 2E$. Over one body period it contributes $2ET_b/L$. Parallel transport of a transverse reference direction along the unit momentum curve contributes minus the oriented solid angle enclosed, up to whole turns and the chosen orientation of the sphere. The resulting net attitude phase is $2ET_b/L - \Omega_{\mathrm{solid}} \pmod{2\pi}$.

The geometric part can be visualised by carrying a local tangent basis along the closed curve. Its two basis vectors are constrained to stay perpendicular to the moving unit momentum and are transported with no imposed twist around that direction. A return to the initial unit momentum need not return that basis to its starting orientation. For a latitude circle on the unit sphere the enclosed cap area is $2\pi(1 - \cos\theta)$; shrinking the path to a point makes the geometric phase vanish modulo 2π. This is why comparing a painted mark after a nutation only to an integrated scalar spin speed can give a discrepancy even with exact torque-free motion.

In practice the total phase can be measured without choosing a cap convention. Fit the attitude at two times separated by the full gyro period, form $R(t + T_b)R(t)^T$, and extract the relative rotation around the known fixed spatial momentum. Independently integrate $2E/L$ over that interval and compare the difference modulo 2π with the signed spherical area traced by normalised body momentum. A factor-of-two error in the identified period, common if only ω_3 is recorded in the low-energy family, generally leaves the relative attitude rotation axis different from $\mathbf{L}$. The axis check can reveal the error before assigning a numerical geometric phase.

Identify which inertias can exist in an actual rigid body

Three positive principal moments inserted into Euler's equations define a mathematical top, but the moments of a physical nonnegative mass distribution also satisfy triangle inequalities. In principal coordinates $I_1 = \int (y^2 + z^2)\, dm$,

$$I_2 = \int (z^2 + x^2)\, dm,$$

$$I_3 = \int (x^2 + y^2)\, dm.$$

Therefore $I_1 + I_2 - I_3 = 2\int z^2\, dm \geq 0$, and likewise by cyclic permutation. Equality means all mass is confined to the corresponding coordinate plane. The convenient example $1, 2, 3$ lies on that boundary and can be realised by an ideal planar lamina. A fitted ratio $1, 2, 3.5$ would imply negative mean squared distance along one direction and should prompt a calibration or modelling check.

A uniform thin rectangular plate of mass M and side lengths b along body axis one and a along body axis two has $I_1 = Ma^2/12,\ I_2 = Mb^2/12,\ I_3 = M(a^2+b^2)/12$. Choosing $b = \sqrt{2}\,a$ gives exactly $1:2:3$. For $M = 0.60\,\mathrm{kg}$, $a = 0.20\,\mathrm{m}$, $b = 0.283\,\mathrm{m}$, the moments are approximately $0.0020, 0.0040, 0.0060\,\mathrm{kg\,m^2}$. The component equations for these moments give the same time histories as the scaled numerical example for identical initial angular velocities, while the actual angular momentum and energy are smaller by a factor 500. This gives a physical specimen for checking principal-axis stability without requiring a massive rotor.

Adding a small mass at a known location changes both inertia and possibly the centre of mass. If a symmetric pair of equal small masses is attached at $\pm d\hat{e}_3$, their combined mass m_a adds $m_a d^2$ to I_1 and I_2, nothing to I_3, and leaves the centre fixed. For the plate above and $m_a = 0.10\,\mathrm{kg}, d = 0.10\,\mathrm{m}$, the new moments are $0.0030, 0.0050, 0.0060\,\mathrm{kg\,m^2}$. The intermediate-axis growth exponent at fixed Ω changes from $|\Omega|/\sqrt{3}$ to $|\Omega|\sqrt{(0.005-0.003)(0.006-0.005)/(0.003\cdot 0.006)} = |\Omega|/3$. A reversible mass-placement experiment tests the inertia dependence of instability, not merely the existence of a visually dramatic flip.

Test intermediate-axis stability with the full body equations

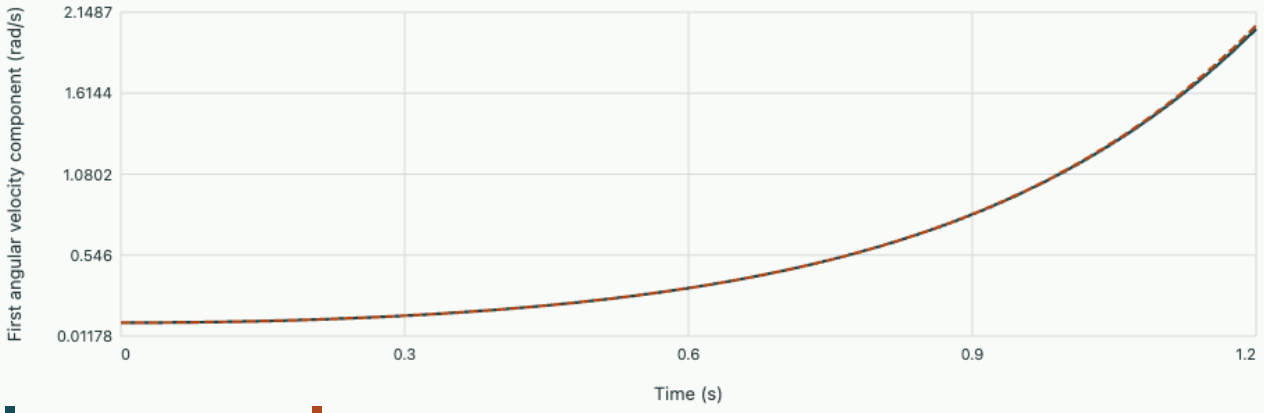
Change the initial tilt and spin. Compare the linear perturbation and the complete torque-free motion while monitoring energy and angular momentum in the OneSimulation output.

Open this experiment online

Try the calculation

Nominal middle-axis speed (rad/s) **12**

Initial transverse speed (rad/s) **0.1**



Initial linear growth rate 3.0984 s⁻¹. With zero initial transverse slope the predicted first doubling time is 0.42505 s.

Download the OneSimulation model

In your OneSimulation environment, run the model with these inputs.

```
python spectra-and-orbits.py rotor --spin 12 --seed 0.1
```

Estimate distance from a separatrix using dwell time

Near the intermediate-axis energy, elliptic modulus approaches one on either side. Writing $d = 1 - k_e^2$ on the low-energy family gives $K \sim \ln(4/\sqrt{d})$. Thus $T_b \sim 4\ln(4/\sqrt{d})/\nu$ for a slowly varying scale ν. Changing d by a factor of one hundred increases the expected body period by approximately $4\ln 10/\nu$. A tiny difference in the initial inertia-weighted energy can produce a readily observed increase in the time between reorientation events while the maximum angular speed scarcely changes. The logarithm follows from passages near a hyperbolic equilibrium, where the signed unstable perturbation grows exponentially.

If $\nu = 3.0\,\mathrm{s}^{-1}$ and $d = 10^{-4}$, the leading estimate is

$$T_b = (4/3)\ln 400 \approx 7.99\,\mathrm{s}.$$

Reducing d to 10^{-6} gives $T_b \approx 11.06\,\mathrm{s}$. Inverting this relation makes estimates of d exponentially sensitive to a period offset. An uncertain start or stop event of a few tenths of a second, or a weak external torque during a long dwell, can affect an inferred separatrix energy considerably. Report the uncertainty of $E - L^2/(2I_2)$ as well as the fitted period when claiming close approach to the unstable family.

One experimental check is to use both the crossing speed of $\omega_2 = 0$ and the time between successive such crossings. At a low-energy upward crossing the slope has magnitude $B\nu$; it remains finite near the separatrix even while the period diverges. A simple global slowdown from increased bearing drag tends to modify the crossing speed as well as the dwell. A correctly prepared near-separatrix torque-free motion has relatively rapid crossings separated by long intervals close to intermediate-axis rotation. The distinction is visible in the time domain and does not depend on estimating a derivative at the slowest, noisiest part of the record.

A fit places a nominally torque-free experiment slightly above the separatrix when E, L are computed from gyros, but below it when inferred from the oscillating component pattern. What should be checked first?

▼ Work through the solution

Recalculate both invariants from the same time stamps, calibrated orientation and inertia axes. The energy classification subtracts comparable quantities and can be sensitive to cross-axis bias. The low-energy family keeps its third component's sign, whereas the high-energy family keeps its first component's sign with the stated ordering. Verify that a full body period was actually recorded and that sign changes were resolvable. Persistent disagreement after uncertainty and sampling checks can indicate a slow torque, changing mass distribution or a misspecified principal-axis alignment.

Quantify a small external torque before fitting a free-top period

Suppose the laboratory angular momentum has magnitude $3.0\,\mathrm{kg\,m^2/s}$ and an approximately constant transverse torque $0.010\,\mathrm{N\,m}$ acts for a $6.0\,\mathrm{s}$ body period. The accumulated transverse impulse is $0.060\,\mathrm{kg\,m^2/s}$, enough to tilt the spatial angular-momentum direction by roughly $0.020\,\mathrm{rad}$, or 1.15°. Its magnitude changes only to second order if the torque remains perpendicular to the initial momentum. A check using only L^2 could therefore show a small relative error while the invariable plane has moved perceptibly.

Use two marked inertial reference directions to track the spatial momentum vector and plot its Cartesian increments against the integrated measured torque. When gravity acts about a support whose line passes through the mass centre, its torque vanishes ideally, but a small offset can supply just such a transverse disturbance. Reversing the initial spin while retaining the support geometry helps distinguish a fixed parasitic torque from gyro bias. If the spatial direction changes appreciably during one long near-separatrix dwell, fitting a torque-free elliptic period to the complete signal becomes inaccurate even if the early, fast gyro crossings agree with the free model.

Tutorial on a torque-free body's time scales

What period should be measured for the transverse gyro in the worked example? Use the small-parameter expansion $K = (\pi/2)[1 + m/4 + 9m^2/64 + \cdots]$ with $m = k_e^2$.

▼ **Work through the solution**

Substitution of $m = 1/12$ gives $K \approx 1.605$ and $T_b \approx 6.42\,\mathrm{s}$. The third-axis component repeats after about $3.21\,\mathrm{s}$, while the second component first reaches its positive maximum at $1.61\,\mathrm{s}$. Checking all three features tests the phase relation among the gyros and helps identify a sensor axis swap.

For a body with $I_1 = I_2 < I_3$, how do Euler's equations simplify and which feature of the asymmetric example disappears?

▼ **Work through the solution**

The third body component remains exactly constant, and the first two rotate sinusoidally at body-frame frequency $(I_3 - I_1)|\omega_3|/I_1$. The polar angle to fixed angular momentum is constant. A body axis can still precess in space, but the varying torque-free polar nutation caused by distinct transverse inertias disappears. Taking this limit in the low-energy solution gives $k_e \to 0$ consistently.

A gyro experiment reports slow amplitude decay along with fixed spatial angular momentum. Can that arise in a strictly rigid, torque-free model conserving kinetic energy?

▼ **Work through the solution**

The two rigid-body invariants fix the allowed intersection curve, so a persistent drift between such curves requires an effect beyond that ideal model or a changing calibration. Internal dissipation in a deformable body can decrease kinetic energy while conserving total spatial angular momentum, favouring rotation about the largest-inertia axis. An external torque could instead change angular momentum. Measuring both invariants helps distinguish the two explanations.

Compare principal-axis stability before fitting an asymmetric-top gyro trace.

David Tong derives the Euler equations, Euler angles and free-top motion.

Symplectic transformations, Poisson brackets and the geometry of measurements

Analytical mechanics from constraints to invariants

Two phase-space plots can enclose the same volume while predicting different responses to an impulse. Derive the stronger condition that makes a change of variables canonical, test it on measured linear maps and use Poisson brackets to propagate observables without solving separately for every coordinate.

Derive the matrix condition from Hamilton's equations

Collect n canonical positions and momenta into $z = (q_1, \dots, q_n, p_1, \dots, p_n)^T$. Write Hamilton's equations as $\dot{z} = J\nabla_z H$, where J has zero diagonal blocks, identity upper right and minus identity lower left. For a time-independent change $Z = f(z)$ with Jacobian $S = \partial Z/\partial z$, the chain rule gives

$$\dot{Z} = SJS^T \nabla_Z H.$$

To preserve the canonical form for every differentiable Hamiltonian require

$$SJS^T = J.$$

For an invertible square matrix this is equivalent to $S^T J S = J$. In geometric terms the bilinear antisymmetric form $\omega(u, v) = u^T J v$ has the same value on any transported pair of tangent vectors.

Taking determinants gives $(\det S)^2 = 1$, and symplectic maps have positive determinant equal to one. One way to see the sign is to take the top exterior power of the preserved two-form: ω^n fixes an oriented volume. In one canonical pair the identity $S^T J S = (\det S) J$ makes determinant one sufficient. In two or more pairs a single volume condition cannot enforce every entry of the matrix condition. For example in ordering (q_1, q_2, p_1, p_2), scaling by diagonal entries $(2, 1, 1, 1/2)$ preserves four-dimensional volume. But $\{Q_1, P_1\} = 2$ and $\{Q_2, P_2\} = 1/2$, so neither conjugate pair has the required bracket. A volume-preserving histogram test alone would miss this failure.

Identify brackets of observables and what they measure

Define $\{A, B\} = (\nabla A)^T J \nabla B = \sum_i (A_{q_i} B_{p_i} - A_{p_i} B_{q_i})$. For any explicitly time-dependent observable $A(z, t)$, differentiation along Hamiltonian trajectories gives

$$\dot{A} = \{A, H\} + \partial_t A.$$

The bracket is antisymmetric and obeys the product rule. Computing it with the chain rule in the new variables proves that canonical transformations preserve brackets, even between nonlinear observables. Conversely the fundamental brackets $\{Q_i, P_j\} = \delta_{ij}$, $\{Q_i, Q_j\} = \{P_i, P_j\} = 0$ are precisely the entries of $SJS^T = J$.

For a particle set $L_z = xp_y - yp_x$. Its Hamiltonian flow as a generator sends x to $x - \varepsilon y$, y to $y + \varepsilon x$ and rotates momenta by the same infinitesimal angle. The angular momentum components satisfy $\{L_x, L_y\} = L_z$ with cyclic permutations. Derive one entry directly for $L_x = yp_z - zp_y$ and $L_y = zp_x - xp_z$. The z, p_z pair contributes $xp_y - yp_x$, with the other terms zero. Rotations about two different axes generally fail to commute. The resulting commutator has the form of a rotation about the third axis even though each underlying infinitesimal motion preserves every canonical bracket.

Jacobi's identity $\{A, \{B, C\}\} + \{B, \{C, A\}\} + \{C, \{A, B\}\} = 0$ implies that the bracket of two time-independent conserved quantities is conserved under the same Hamiltonian. If the Hamiltonian is invariant under rotations about x and y, its invariance about z follows from this relation and $\{L_x, L_y\} = L_z$. Conservation alone need not make two observables commute. A central force can conserve all three angular momentum components, but their mutual brackets remain nonzero. For independent commuting integrals in an integrability argument one must check the involution condition explicitly.

Construct a canonical map for any smooth point transformation

Suppose new coordinates depend only on old coordinates, $Q = f(q)$. Preserve the canonical one-form exactly,

$$p^T dq = P^T dQ = P^T Df\, dq.$$

Thus $p = (Df)^T P$ and $P = (Df)^{-T} p$. Applying the same coordinate matrix to momenta works only for a restricted case such as an orthogonal change of coordinates. An extension of position must be accompanied by a reciprocal rescaling of its conjugate momentum. For $Q = q^2$ on $q > 0$, for example, $P = p/(2q)$. The free Hamiltonian becomes $2QP^2/m$; even a physically free particle can acquire a coordinate-dependent expression for its kinetic energy.

Polar coordinates demonstrate why a momentum variable need not equal a measured tangential component. In the plane $x = r\cos\varphi, y = r\sin\varphi$. Substituting into $p_x dx + p_y dy$ gives

$$p_r = \mathbf{p}\cdot\hat{r},$$

$$p_\varphi = r\,\mathbf{p}\cdot\hat{\varphi} = L_z.$$

The physical tangential momentum is p_φ/r. At radius $0.40\,\mathrm{m}$ and transverse momentum $0.60\,\mathrm{kg\,m/s}$, the canonical angular momentum is $0.24\,\mathrm{kg\,m^2/s}$. Treating the transverse component itself as conjugate to angle would make $\{\varphi, p_\varphi^{\mathrm{wrong}}\} = 1/r$. The coordinate singularity at $r = 0$ also prevents these canonical polar variables from covering the entire original phase space with one chart.

Generate shears and test a numerical map

A type-two generating function $F_2(q, P, t)$ defines $p = \partial_q F_2$, $Q = \partial_P F_2$, and $K = H + \partial_t F_2$ where the transformation can be locally inverted. The differential $d(F_2 - QP)$ equals $p\,dq - P\,dQ + (K - H)\,dt$. Thus the actions differ only by an endpoint term; taking another exterior

derivative at fixed time gives the symplectic form condition. For $F_2 = qP + sP^2/(2m)$, the map is $p = P, Q = q + sp/m$, free drift over time s. For $F_2 = qP + sV(q)$, the map is $P = p - sV'(q), Q = q$, a potential kick. Their composition is canonical for any step size even though an alternating drift and kick need not give the exact trajectory.

Take a unit-mass, unit-frequency oscillator with $h = 0.5$. A kick followed by drift gives $P = p - hq$, $Q = q + hP = (1 - h^2)q + hp$. Its amplification matrix has determinant one and trace $2 - h^2 = 1.75$, so its eigenvalues form a unit-modulus conjugate pair. The numerical phase per step is $\arccos(1 - h^2/2) = 0.50536$, slightly ahead of the exact phase 0.50000. Starting at $q = 1, p = 0$, the first step gives $Q = 0.75, P = -0.50$, and physical energy becomes 0.40625 instead of 0.5. Exact bracket preservation does not imply exact conservation of the original energy at every point. If $h > 2$ the same matrix remains symplectic but has a real expanding and contracting eigenvalue; geometric preservation is compatible with numerical instability.

For a numerical implementation with analytic equations unavailable, approximate the Jacobian of a one-step map at selected states by perturbing one canonical component at a time. Compute $S^T J S - J$ after nondimensionalising coordinates and momenta in conjugate pairs. Repeat with smaller perturbations and at different integration tolerances. A determinant close to one in four dimensions is insufficient. Testing the transported brackets of one position–momentum pair and one cross-pair can reveal a swapped column or misplaced mass factor before a long simulation exhibits visible energy drift.

Track canonical area for a transported experimental ensemble

A narrow group of trajectories lets an experimenter estimate a tangent map without applying a separate perturbation to each coordinate at the same time. Prepare a small covariance ellipse in one canonical pair $z = (q, p)$, propagate each member through a known linear Hamiltonian device, and compute $\Sigma_{\mathrm{out}} = S \Sigma_{\mathrm{in}} S^T$. The ellipse area is proportional to $\sqrt{\det \Sigma}$, and a symplectic map in one pair preserves that area. Its projections can grow or contract individually: a focusing device can reduce position spread while increasing momentum spread or correlation. Measuring only the two marginal widths ignores their covariance and can falsely suggest area growth.

For a free unit-mass drift of duration two in scaled units,

$$S = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}.$$

Start with $\Sigma = \mathrm{diag}(1, 1)$. The outgoing covariance has entries $\Sigma_{qq} = 5, \Sigma_{pp} = 1, \Sigma_{qp} = 2$. The apparent rectangle of marginal standard deviations grows by $\sqrt{5}$, yet the determinant remains $5 - 4 = 1$. A plot that suppresses the correlation has discarded exactly the information needed for the area test. Noise in the two channels also contributes to observed covariance, so calibrate sensor variances independently before attributing a small apparent growth to physical dissipation or numerical smoothing.

For two pairs it is possible for a perfectly canonical coupling to exchange projected emittances. Take a uniform rotation of the two positions and the same rotation of their conjugate momenta. It is canonical by construction. If the two incoming pairs have different covariance areas, a ninety-degree rotation

swaps them completely; at forty-five degrees each outgoing projected pair contains a mixture. The full four-dimensional symplectic structure remains intact. An experimentally available two-dimensional marginal then tests only part of the dynamics. Counting phase-space points in a volume and measuring cross-pair brackets answer different questions about a strongly coupled system.

The eigenvalues of $J\Sigma$ are invariant under the linear canonical covariance transport. Indeed $JS = S^{-T}J$ for symplectic S, so $J\Sigma_{\mathrm{out}} = S^{-T}(J\Sigma_{\mathrm{in}})S^{T}$, a similarity transformation. For a positive covariance these eigenvalues occur as conjugate imaginary pairs, and their positive magnitudes give symplectic covariance scales. This algebraic test can distinguish loss of correlations in a fitted transfer matrix from actual exchange between pairs. It assumes linear propagation of the ensemble. A nonlinear canonical map can stretch a finite ellipse into a curved shape whose covariance determinant changes while its exact enclosed area remains preserved.

Use generating functions to find a moving normal-mode basis

Consider a two-dimensional isotropic oscillator and rotate its coordinate axes by a prescribed time-dependent angle $\alpha(t)$. Let $Q = R(\alpha)^{T} q$, $P = R(\alpha)^{T} p$ for a proper planar rotation. At any fixed time the map is orthogonal and symplectic. To derive the Hamiltonian in these moving variables, use $F_2(q, P, t) = q^{T} R(\alpha) P$. Its time derivative at fixed old q and new P is $-\dot{\alpha}(Q_x P_y - Q_y P_x)$, with the sign corresponding to axes rotating counterclockwise in the laboratory. The transformed Hamiltonian is therefore

$$K = H_0(Q, P) - \dot{\alpha} L_z.$$

For constant rotation rate and an isotropic potential the new Hamiltonian has no explicit time dependence. Its value is the combination energy minus rotation rate times angular momentum encountered in the Noether argument. Hamilton's equations now include $-\dot{\alpha}\,\hat{z} \times Q$ in $\dot{Q}$ relative to the rotated physical velocity. Thus the transformed canonical momentum P equals the rotated laboratory momentum rather than $m\dot{Q}$. Confusing the two yields incorrect initial conditions for a rotating-frame simulation even if the transformation matrix at the starting instant is the identity.

Take $m = 1.0\,\mathrm{kg}$, a laboratory point initially at $q = (0.20, 0)\,\mathrm{m}$ with $p = (0, 0.30)\,\mathrm{kg\,m/s}$, and axes initially aligned but rotating at $2.0\,\mathrm{rad/s}$. The canonical P_y is $0.30\,\mathrm{kg\,m/s}$. The derivative of the rotating-frame coordinate is

$$0.30 - 2.0(0.20) = -0.10\,\mathrm{m/s}.$$

It moves in the negative rotating-frame y direction while its laboratory velocity points positive y. The angular momentum is $0.060\,\mathrm{kg\,m^2/s}$, so K lies $0.120\,\mathrm{J}$ below the laboratory Hamiltonian. Both position–momentum transformations preserve brackets; the extra term derives from the moving basis, not a failure of canonical geometry.

An oscillator with two nearly equal natural frequencies provides another reason to track the generating function. If one diagonalises a slowly changing stiffness matrix at each time and keeps only the instantaneous diagonal frequencies, the omitted rotation of the eigenvectors couples the nominal

modes through a term of this type. When the eigendirections turn rapidly near a degeneracy, the coupling can be appreciable even if both eigenvalues change little. Measuring the orientation as well as the eigenfrequencies allows the size of this effect to be estimated before fitting amplitude transfer to a spurious dissipation term.

Compare exact phase transport with a dissipative fit

A damped oscillator with $m\ddot{q} + c\dot{q} + kq = 0$ has equations $\dot{q} = p/m$, $\dot{p} = -kq - cp/m$. In the physical (q, p) plane their divergence is $-c/m$. A small material phase-space area therefore contracts as $A(t) = A(0)e^{-ct/m}$. No time-independent canonical coordinate change on the same two-dimensional isolated phase space can turn that physical contraction into a Hamiltonian flow preserving area. This conclusion concerns the reduced dissipative description. An expanded system including the degrees of freedom that absorb energy can still be Hamiltonian.

For $c/m = 0.40\,\mathrm{s}^{-1}$, a pair of independent initial perturbations spanning area one in calibrated coordinates spans only 0.449 after two seconds. A numerical step of $0.10\,\mathrm{s}$ fitted to that exact map should have determinant $e^{-0.04} = 0.9608$. Forcing its determinant to one during system identification would remove a real dissipative measurement. Conversely a conservative oscillator propagated by forward Euler at matched units and $h = 0.10$ has amplification determinant $1 + h^2 = 1.01$, corresponding to a spurious area growth. A geometric diagnostic is interpretable only after deciding what physical system and coordinate pair were actually observed.

The pair q, v differs from canonical q, p by a constant mass scaling for an ordinary single particle. In one pair with constant mass that scaling does not change a measured same-coordinate before-and-after determinant. For a position-dependent mass, constrained coordinates with a varying metric, or explicit time-dependent changes of measurement scale, volume in raw position–velocity coordinates need not equal canonical volume. For example a free particle expressed in $Q = q^2$ has canonical momentum $P = m\dot{Q}/(4Q)$. A detector estimating phase area in $(Q, \dot{Q})$ without the Q-dependent factor can infer contraction or expansion even when canonical phase area is exactly conserved.

Compute a bracket of mechanical momenta and recover the Lorentz force

For a charge e in a prescribed vector potential write mechanical momentum $\boldsymbol{\pi} = \mathbf{p} - e\mathbf{A}(\mathbf{q}, t)$. The canonical brackets are $\{q_i, p_j\} = \delta_{ij}$, but the measured mechanical components satisfy

$$\{\pi_i, \pi_j\} = e(\partial_i A_j - \partial_j A_i) = e\epsilon_{ijk} B_k.$$

At nonzero magnetic field they are not simultaneously canonical momenta conjugate to the Cartesian positions. This result is gauge invariant because it involves only the curl of the vector potential. It offers a physical example where replacing canonical by measured momenta in a software interface changes the matrix of fundamental brackets.

With $H = \pi^2/(2m) + e\phi$, the time derivative is

$$\dot{\pi}_i = \{\pi_i, H\} + \partial_t \pi_i.$$

The kinetic-energy bracket contributes $(e/m)(\boldsymbol{\pi} \times \mathbf{B})_i$; the scalar potential contributes $-e\partial_i \phi$; and the explicit derivative supplies $-e\partial_t A_i$. Their sum is $e[\mathbf{E} + (\boldsymbol{\pi}/m) \times \mathbf{B}]_i$. Dropping the last term would make time-dependent electric acceleration depend on gauge. In a static uniform B_z, the transverse mechanical momenta rotate although their energy is constant. Canonical bracket structure therefore contains observable information about the force even before integrating a trajectory.

A useful finite-difference test sets $A_y = Bx$ and samples the map from canonical to mechanical variables around one state. The measured bracket $\{\pi_x, \pi_y\} = eB$ should be constant in space. A code that insists on zero will attribute the entire magnetic response to an erroneous Hamiltonian gradient. A code using mechanical variables legitimately can instead use their noncanonical bracket matrix; its equations must then preserve the correct transformed structure, not the original constant J. Testing the resulting acceleration at zero velocity and at a transverse velocity distinguishes electric and magnetic terms and catches sign errors in the antisymmetric entries.

Derive a reversible drift–kick–drift scheme and its measured phase

For $H = p^2/(2m) + V(q)$, exact kinetic flow for a duration t is $D_t : (q, p) \mapsto (q + tp/m, p)$. Exact potential flow is $K_t : (q, p) \mapsto (q, p - tV'(q))$. Each is canonical. Compose $D_{h/2} K_h D_{h/2}$ in chronological order. Applying the same construction with $-h$ undoes each substep in reverse order, so the full map is time reversible. Symmetry eliminates the even local-error powers in the effective generator, giving a second-order method for smooth V. The individual Hamiltonian pieces need not commute; if they did, the split map would be exact at any step.

For a unit oscillator, carrying out the algebra gives

$$Q = (1 - h^2/2)q + h(1 - h^2/4)p,$$

$$P = -hq + (1 - h^2/2)p.$$

Verify directly that the determinant is one. The trace gives the same step angle $2 \arcsin(h/2)$ as the first-order kick–drift matrix in its stable range, but the new off-diagonal coefficients make the amplitude errors and time reversal different. Starting from $q = 1, p = 0$ at $h = 0.50$, one gets $Q = 0.875, P = -0.50$, energy 0.5078125. A negative step returns to the initial values apart from roundoff. Passing this one reversal test says nothing about absolute phase accumulated over hundreds of periods.

The map preserves $q^2 + (1 - h^2/4)p^2$ exactly for the harmonic test. At $h = 0.5$ the coefficient of p^2 is 0.9375, so the preserved ellipses differ slightly from the circles of physical energy. Their reversible oscillation in physical energy is bounded for $|h| < 2$. The per-step phase is about 0.50536 at this step, giving a phase gain 0.536 radians after one hundred steps. As h is halved while physical integration time stays fixed, the overall phase error falls approximately by four. Measuring an energy oscillation, a

reversal residual and a return time therefore probes three different properties of one apparently simple map.

Compare a symplectic map with the phase of a physical orbit

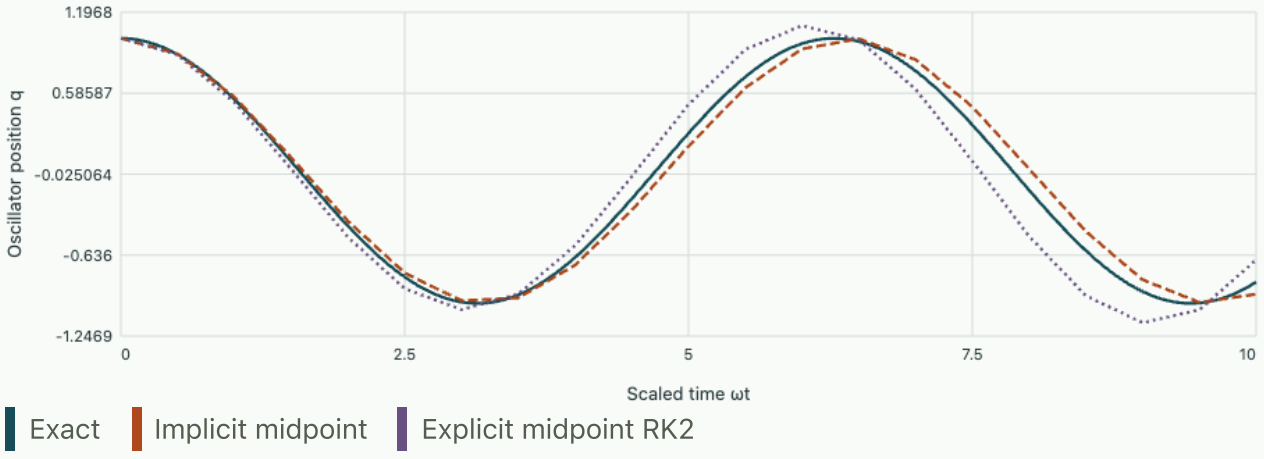
Change the step of the midpoint update. Distinguish preservation of phase-space structure from agreement of the measured oscillation frequency.

Open this experiment online

Try the calculation

Time step h (ω = 1) **0.5**

Number of forward steps **20**



Implicit midpoint phase lag 0.2009 rad. Forward-backward distance from the start is 8.951e-16 for implicit midpoint and 0.3635 for explicit midpoint.

Relate the order of two pulses to their Poisson bracket

A canonical transformation close to identity can be built by applying two brief Hamiltonian operations in different orders. Take one free drift and one potential kick, each of short duration s. A drift then a kick gives

$$q' = q + sp/m,\; p' = p - sV'(q + sp/m).$$

A kick then a drift gives

$$p'' = p - sV'(q),\; q'' = q + s[p - sV'(q)]/m.$$

To second order their coordinate difference is $q' - q'' = s^2 V'(q)/m$; their momentum difference is $p' - p'' = -s^2 p V''(q)/m + O(s^3)$. The difference vector equals the Hamiltonian vector generated by $G = pV'(q)/m = \{V, T\}$, multiplied by s^2. The bracket sets the leading effect of reversing the operation order.

For the harmonic oscillator at $q = 1, p = 0, m = k = 1$ with $s = 0.10$, a drift then kick gives $(1, -0.10)$; kick then drift gives $(0.99, -0.10)$. The coordinate difference 0.01 equals $s^2 V'/m$, and the leading momentum difference is zero for this initial condition. At $q = 0, p = 1$, the complementary experiment gives a momentum difference of -0.01 and no coordinate difference. Running both initial states tests the two components of the bracket-generated perturbation. A single apparently correct coordinate update would not detect a sign swap in the other component.

An experimental implementation might use a short force pulse as a kick and free propagation between pulses. The pulse must be short enough that motion during the kick is negligible; otherwise its own kinetic drift enters the comparison at the same order as the desired bracket. Extrapolate the measured difference divided by s^2 to small s, and separately measure acquisition noise, which grows in relative importance under that division. Alternating the pulse order cancels many static calibration shifts but does not magically remove timing offsets correlated with the order. Treat the small bracket signal as a physical difference between two specified protocols.

A transformation preserves all coordinate–momentum brackets but a time integration on the transformed variables produces the wrong laboratory trajectory. Can the bracket check alone identify the error?

▼ Work through the solution

Canonical coordinates guarantee the form of Hamilton's equations. Their actual trajectories still require the correct transformed Hamiltonian, including explicit time dependence of the generating function, consistent initial momenta and a sufficiently accurate integrator. Transform an independently known initial state, compare the initial velocity and acceleration of a measurable coordinate, and follow both descriptions back to the same laboratory frame. A bracket test can establish that a change of variables is admissible without establishing that the correct energy function or physical input was attached to it.

Set compatible units before inspecting a symplectic residual

Suppose position is recorded in millimetres and momentum in kilogram metres per second. Scaling each numerical coordinate to unit variance is convenient for fitting a matrix, but independent scaling of each coordinate and momentum can change the bracket matrix from its textbook form. If $z' = Dz$ with diagonal D, then $J' = DJD^T$. For one pair this multiplies its bracket by the product of the position and momentum scale factors. In two pairs these factors can differ. A residual computed against the original J can flag a harmless unit change or, conversely, mask an error if scales happen to compensate.

Choose a characteristic length q_0, momentum p_0, and common action scale $q_0 p_0$ for each pair if the dimensionless chart is to have a common symplectic scale. Otherwise carry the transformed bracket explicitly through the test. With two oscillators of different physical units, a canonical

nondimensionalisation may balance the dynamics better than using the same numerical length scale for both. Report the dimensional transformation along with any determinant or defect norm so another calculation can reconstruct which tangent vectors and measurement units were actually compared.

Tutorial on canonical tests

In two degrees of freedom try $Q_1 = q_1,\ Q_2 = q_2,\ P_1 = p_1 + \alpha q_2,\ P_2 = p_2$. Is this map canonical? What is the simplest modification that restores the condition?

▼ **Work through the solution**

The fundamental mixed momentum bracket is $\{P_1, P_2\} = \alpha$. Restore zero by also taking $P_2 = p_2 + \alpha q_1$, since then the extra term $\{p_1, \alpha q_1\} = -\alpha$ cancels. The added momentum vector is now the gradient of the scalar $\alpha q_1 q_2$, whose derivative matrix is symmetric. Both candidate maps preserve volume, reinforcing why volume alone cannot decide.

For the unit oscillator kick–drift map find a quadratic quantity it preserves exactly while its physical energy changes.

▼ **Work through the solution**

Direct substitution gives $Q^2 + P^2 - hQP = q^2 + p^2 - hqp$ for the kick–drift convention above. Multiplying by one half gives a positive modified quadratic invariant when $|h| < 2$. The first step at $h = 0.5$ gives $0.75^2 + (-0.50)^2 - 0.5(0.75)(-0.50) = 1.0$, equal to its initial value. At $h = 2$ the quadratic becomes degenerate and beyond it no longer bounds the motion.

A time-dependent momentum shift is generated by $F_2 = qP - a(t)q$. For a free physical particle find the transformed Hamiltonian and verify constant physical momentum.

▼ **Work through the solution**

The transformation gives

$$p = P - a,\ Q = q,$$
$$K = (P - a)^2/(2m) - \dot{a}\, Q.$$

Hamilton's equation gives $\dot{P} = +\dot{a}$, hence $\dot{p} = 0$. Omitting the explicit $\partial_t F_2$ term would predict constant P and a spurious acceleration of the original free particle. This test uses an observable whose physical behaviour is already fixed independently.

Compare generated canonical changes and numerical determinant tests.

David Tong develops canonical transformations and Poisson brackets in Hamiltonian dynamics.

Canonical perturbations, small denominators and resonant capture

Analytical mechanics from constraints to invariants

A weak periodic force can produce a bounded ripple on most trajectories and a large change on a narrow band of initial conditions. Derive the averaging transformation away from resonance, construct the slow resonant dynamics where that transformation fails, and estimate the physically observable width and time scale of a resonance.

Separate the unperturbed clock from a weak correction

Take an integrable Hamiltonian $H(I,\theta,t)=H_0(I)+\epsilon H_1(I,\theta,t)$ in action–angle variables. With $\epsilon=0$, actions are constants and angles advance at $\boldsymbol{\omega}(I)=\partial_I H_0$. A perturbation with Fourier component $h_{\mathbf{n}}(I)\cos(\mathbf{n}\cdot\theta-\Omega t)$ drives action changes proportional to its sine. If the phase sweeps quickly through all angles, positive and negative contributions largely cancel. If $\mathbf{n}\cdot\boldsymbol{\omega}-\Omega$ is small, the perturbation acts with a coherent sign over a long time. We want a transformation that makes both statements quantitative.

Use a near-identity type-two generating function $F_2(\theta,J,t)=\theta\cdot J+\epsilon S_1(\theta,J,t)$. Old actions are $I=J+\epsilon\partial_\theta S_1$, and new angles are $\psi=\theta+\epsilon\partial_J S_1$. Expand $H_0(I)=H_0(J)+\epsilon\boldsymbol{\omega}(J)\cdot\partial_\theta S_1+O(\epsilon^2)$. At first order the new Hamiltonian is $K=H_0(J)+\epsilon[H_1+\partial_t S_1+\boldsymbol{\omega}\cdot\partial_\theta S_1]$. Replacing old by new angles inside that bracket produces only higher-order errors. To leave the averaged part $\langle H_1\rangle$, solve the homological equation

$$(\partial_t+\boldsymbol{\omega}\cdot\partial_\theta)S_1=-\big(H_1-\langle H_1\rangle\big).$$

For an oscillating cosine with detuning $\Delta=\mathbf{n}\cdot\boldsymbol{\omega}-\Omega$, choose $S_1=-h_{\mathbf{n}}\sin(\mathbf{n}\cdot\theta-\Omega t)/\Delta$. The transformation is useful while the resulting action displacement $\epsilon|\partial_\theta S_1|$ is small on the relevant action scale and the derivatives entering the angle transformation remain controlled. Small ϵ alone is insufficient if Δ is even smaller. A singular denominator says that this choice to remove the harmonic is unsuitable, not that the underlying mechanical force becomes infinite.

Obtain a nonlinear frequency shift from the angle average

A free oscillator with mass m, small-amplitude frequency ω_0, and quartic correction $\beta q^4/4$ has unperturbed coordinates $q=\sqrt{2I/(m\omega_0)}\sin\theta$. Substituting gives the correction $\beta I^2\sin^4\theta/(m^2\omega_0^2)$. Since $\langle\sin^4\theta\rangle=3/8$, the first-order averaged Hamiltonian is $K=\omega_0 J+3\beta J^2/(8m^2\omega_0^2)$. Its derivative gives

$$\omega=\omega_0+3\beta J/(4m^2\omega_0^2).$$

Expressing $J\simeq m\omega_0 A^2/2$ in terms of leading oscillation amplitude yields the frequency shift $3\beta A^2/(8m\omega_0)$. Oscillatory harmonics at $2\theta,4\theta$ change the waveform but do not contribute directly to this averaged first-order clock.

For $m = 0.20\,\mathrm{kg}$, $\omega_0 = 10.0\,\mathrm{s}^{-1}$, $\beta = 80\,\mathrm{N/m^3}$, and $A = 0.10\,\mathrm{m}$, the predicted shift is $0.150\,\mathrm{s}^{-1}$, about 1.5%. The nonlinear-to-linear force ratio at maximum displacement is $\beta A^2/(m\omega_0^2) = 0.040$, consistent with the expansion. At twice the amplitude the shift quadruples and the expansion parameter grows to 0.16. Measuring period over several amplitudes and plotting the frequency against A^2 tests both the slope and the onset of curvature from higher orders. A positive β stiffens the spring; a negative quartic term is only a local approximation unless higher powers restore global confinement.

Solve the slow motion in the resonant frame

Consider a rotor $H = I^2/(2\mathcal{I}) - \epsilon V\cos(\theta - \Omega t)$, where $\mathcal{I}$ is moment of inertia and $\epsilon V > 0$ has energy units. Choose $F_2 = (\theta - \Omega t)P$, so $P = I$,

$$\phi = \theta - \Omega t,$$

$$K = P^2/(2\mathcal{I}) - \Omega P - \epsilon V\cos\phi.$$

Completing the square gives $K = (P - P_*)^2/(2\mathcal{I}) - \epsilon V\cos\phi + \text{constant}$ with $P_* = \mathcal{I}\Omega$. This is an exact pendulum in the rotating frame. It retains the resonant cosine instead of attempting to average it away. Phase locking means libration of ϕ around zero; outside its separatrix the relative phase circulates.

The stable fixed point has $P = P_*$, $\phi = 0$; its small libration frequency is

$$\omega_{\mathrm{lib}} = \sqrt{\epsilon V/\mathcal{I}}.$$

The saddle at $\phi = \pi$ has energy $+\epsilon V$ relative to the same constant. At $\phi = 0$ that energy is reached for $|P - P_*| = 2\sqrt{\mathcal{I}\epsilon V}$, the resonance half-width in action. Its complete width is twice that number. For $\mathcal{I} = 0.050\,\mathrm{kg\,m^2}$, $\Omega = 8.0\,\mathrm{s}^{-1}$ and $\epsilon V = 0.0020\,\mathrm{J}$, the central action is $0.40\,\mathrm{kg\,m^2/s}$; the half-width is $0.020\,\mathrm{kg\,m^2/s}$ and small libration period $31.4\,\mathrm{s}$. The rotor makes many laboratory turns while its slow relative phase completes one libration.

An initial action inside the quoted half-width is necessary but not sufficient for trapping at every phase. At $\phi = \pi$, any nonzero $P - P_*$ is above the saddle energy. Compute the actual rotating-frame energy of the initial state and compare with the saddle. For $\phi = 0$, $P = 0.415$, the kinetic detuning energy is $0.00225\,\mathrm{J}$; subtracting $0.0020\,\mathrm{J}$ gives $0.00025\,\mathrm{J}$, below the saddle. With the same action at $\phi = \pi$ the value is $0.00425\,\mathrm{J}$, above the saddle. Selecting trajectories by mean frequency alone can therefore mix trapped and circulating populations.

Retain a resonant harmonic in a general integrable system

For one action with $H_0'(I_*) = \Omega$, use the rotating frame and expand $H_0(I) - \Omega I$ about I_*. The linear term vanishes, leaving $H_0''(I_*)(I - I_*)^2/2$. Retain the slowly varying resonant cosine and average the rapidly varying harmonics. When $H_0'' \neq 0$, this gives the same pendulum normal form with an effective inertia and possibly a reversed sign of the stable phase. The action half-width scales as $2\sqrt{\epsilon|V/H_0''|}$.

Thus a weak perturbation can influence a band of width proportional to $\sqrt{\epsilon}$, parametrically larger than a typical nonresonant action ripple proportional to ϵ.

For an exactly harmonic oscillator $H_0 = \omega_0 I$ the curvature vanishes. One cannot replace H_0'' with zero in the width formula. An exactly tuned additive drive in a linear oscillator instead produces secular growth until damping, detuning, nonlinearity or a finite drive interval intervenes. In multiple actions, one first transforms to the resonant phase $\mathbf{n} \cdot \theta - \Omega t$ and complementary angles, preserving the canonical one-form. Other near-resonances can prevent isolation of a single pendulum. Apparent chaos inferred from a short irregular time trace alone needs additional tests, including resolved phase portraits, convergence under time refinement and identification of competing resonances.

Calculate a phase-locking threshold for a drive that sweeps in frequency

The exact rotating rotor also allows a tractable time-dependent drive frequency. Write the pattern angle as $\alpha(t)$, with $\Omega(t) = \dot{\alpha}$. In its frame $K = P^2/(2\mathcal{I}) - \Omega(t) P - \epsilon V \cos\phi$. Hamilton's equations give

$$\dot{\phi} = P/\mathcal{I} - \Omega(t),$$

$$\dot{P} = -\epsilon V \sin\phi.$$

Define the detuning momentum $p = P - \mathcal{I}\Omega(t)$. Even when $\dot{\phi} = 0$ at one instant, phase locking throughout a ramp requires $\dot{p} = 0$ and hence $\sin\phi_0 = -\mathcal{I}\dot{\Omega}/(\epsilon V)$. The restoring torque must be sufficient to accelerate the rotor along with the changing pattern.

For constant $\dot{\Omega}$ and $|\mathcal{I}\dot{\Omega}| < \epsilon V$, a stable locked phase with $\cos\phi_0 > 0$ exists. Perturbing the phase by η gives $\ddot{\eta} + (\epsilon V/\mathcal{I})\cos\phi_0\,\eta = 0$, so the libration softens as the required torque approaches its maximum. At threshold the stable and unstable stationary phases meet. Above it no fixed phase can provide the torque needed by the ramp. The mere existence of a stable locked solution still does not imply that an arbitrary initial phase and momentum will be captured in it. In a conservative system initial phase-space location and separatrix crossing are essential.

With the previous $\mathcal{I} = 0.050\,\mathrm{kg\,m^2}$ and $\epsilon V = 0.0020\,\mathrm{J}$, the largest trackable constant acceleration in this model is $0.040\,\mathrm{rad/s^2}$. At half that value the stable phase lags by $\pi/6$, and the small relative-phase frequency is

$$0.20\sqrt{\cos(\pi/6)} = 0.186\,\mathrm{s^{-1}}.$$

Increasing the drive frequency by $1.0\,\mathrm{rad/s}$ at this rate takes $50\,\mathrm{s}$, only about one and a half libration periods. Adiabatic capture estimates that additionally require many libration periods may therefore be unconvincing even though the elementary torque threshold is satisfied. A measured lag versus sweep rate assesses the static torque condition, while a full phase trajectory tests dynamic capture.

The same equations admit a tilted effective washboard potential in the accelerating frame, $U_{\mathrm{rel}} = -\epsilon V\cos\phi + \mathcal{I}\dot{\Omega}\,\phi$, as long as $\dot{\Omega}$ is constant. Its potential difference over a whole 2π turn is

nonzero, so unwrap the relative phase before using its energy. Mapping every phase back to $[-\pi, \pi)$ hides this difference and produces artificial jumps in the inferred energy. In the original variables the physical motor does work. There is no contradiction between a conserved local accelerating-frame energy for a constant ramp and the changing laboratory kinetic energy of a locked rotor.

Derive secular energy exchange in a pair of weakly coupled oscillators

Take two equal masses m with individual natural frequency ω_0 and potential coupling $k_c(q_1 - q_2)^2/2$. The normal coordinates $Q_\pm = (q_1 \pm q_2)/\sqrt{2}$ have frequencies $\omega_+ = \omega_0$,

$$\omega_- = \sqrt{\omega_0^2 + 2k_c/m}.$$

Release mass one at displacement A with both velocities and q_2 initially zero. Superposition gives

$$q_1 = A[\cos \omega_+ t + \cos \omega_- t]/2,$$
$$q_2 = A[\cos \omega_+ t - \cos \omega_- t]/2.$$

The slowly varying difference phase transfers oscillation amplitude between the masses over a time much longer than their rapid periods.

For $\omega_0 = 10\,\mathrm{s}^{-1}$ and $k_c/m = 1.0\,\mathrm{s}^{-2}$, the exact separation is

$$\sqrt{102} - 10 = 0.0995\,\mathrm{s}^{-1}.$$

The first near-complete exchange occurs at $t = \pi/|\omega_- - \omega_+| \approx 31.6\,\mathrm{s}$. Over a single fast period the displacements can look almost like uncoupled oscillations, despite large later energy redistribution. Perturbing two isolated oscillators while trying to average the coupling over both independent fast angles would discard a term depending on their slow difference phase. Retaining the resonant combination predicts the exchange; diagonalising the linear model supplies a reference solution for that approximation.

Detune the uncoupled springs so their natural frequencies have a separation large compared with $k_c/(m\omega_0)$. The coupled normal modes then align more closely with the original oscillators, and the energy exchange cannot reach its former depth. In a nonlinear problem the individual frequencies also depend on action, so the energy transfer can detune itself. To interpret an observed beating envelope, measure both normal frequencies and relative phases. An apparent loss of amplitude in oscillator one can be energy stored in oscillator two, distinguishable from damping through recovery at later times and a balance of the total Hamiltonian.

Estimate the accuracy of averaging over a finite observation time

For a nonresonant perturbation with initially slow action change, one can integrate the first-order drive along the unperturbed trajectory over the actual observation window instead of assuming an infinite-time average. For $H_1 = A\cos(\theta - \Omega t)$, $\dot{I} = A\sin(\Delta t + \phi_0)$ to leading order with $\Delta = \omega_0 - \Omega$. Thus $\Delta I(T) = A[\cos\phi_0 - \cos(\Delta T + \phi_0)]/\Delta$. For $\Delta \neq 0$ its absolute value remains at most $2|A/\Delta|$, and its largest possible value over initial phase is $2|A\sin(\Delta T/2)/\Delta|$. For $\Delta \to 0$ the phase-dependent result tends to $AT\sin\phi_0$. The finite-time expression is regular right at resonance and shows exactly when a secular term emerges.

For $A = 0.001\,\mathrm{J}$, $T = 10\,\mathrm{s}$ and $\Delta = 0.20\,\mathrm{s}^{-1}$, the maximum-amplitude change in action over initial phase is $0.00841\,\mathrm{J\,s}$. At zero detuning the leading largest change is $0.0100\,\mathrm{J\,s}$, only modestly larger because this observation window resolves a frequency band of order $1/T$. For $\Delta = 2\pi/T$ the change after exactly T vanishes, although the action swung during the interval. A final-minus-initial measurement alone can mistake a completed nonresonant excursion for an entirely unperturbed trajectory. Recording at an additional end time tests that cancellation.

At leading order a force of finite duration produces this sine-over-detuning envelope even in a linear system. It is an observation-time bandwidth, distinct from the dynamical resonance width proportional to $\sqrt{\epsilon}$ in the nonlinear rotor. Repeat the experiment with doubled drive duration at fixed amplitude. Features that narrow inversely with the duration are evidence of finite-time spectral selection. A persistent bounded island in a stroboscopic phase portrait, with the predicted amplitude-dependent width and libration frequency, is evidence of the phase-locking dynamics. If drive amplitude is large enough to change the natural frequency appreciably over the interval, the unperturbed phase-integration formula ceases to describe the trajectory.

Construct a stroboscopic section with the right clock

For a time-periodic Hamiltonian sample (I, θ) every period of the external drive at a fixed drive phase. This creates a two-dimensional area-preserving map for one periodically forced degree of freedom. A phase-locked solution can become a fixed point of this map, and librating trajectories around it can form an island. Sampling at each crossing of a moving coordinate threshold instead can produce a different return map and a different phase weight. A return plot can look irregular simply because the drive phase was not held fixed at the samples.

Consider a rigid rotor commanded at $\Omega = 8\,\mathrm{s}^{-1}$. Its drive period is

$$2\pi/8 = 0.7854\,\mathrm{s}.$$

An acquisition system sampling nominally every $0.7900\,\mathrm{s}$ accumulates approximately $8(0.7900 - 0.7854) = 0.0368\,\mathrm{rad}$ of phase error per sample. After one hundred samples the sampling phase has moved about $3.68\,\mathrm{rad}$, enough to smear a narrow island across a substantial fraction of the plot. Interpolating trajectories to timestamps from a synchronised drive encoder controls that effect. Keep unwrapped relative phase as a separate channel to count actual slips.

A numerical area test on the stroboscopic map should use canonical action and angle or an equivalent canonical coordinate pair. A scatter plot of polar angle against angular velocity for a state-dependent inertia can distort area. Changes of pixel density, trajectory dwell time and resolution near the separatrix further affect visual appearance. To identify a resonance quantitatively, measure the action extent at the stable phase, the local libration period and the saddle crossing energy on independently prepared trajectories. Agreement among those three observations is less easily mimicked by a sampling artefact than the existence of a visually closed loop alone.

Recover a frequency shift by removing a secular term

Canonical averaging yields a frequency correction for the quartic oscillator. An independent derivation makes clear why a small difference in frequency can become a large trajectory error at long times. Write the equation as $\ddot{q} + \omega_0^2 q + (\beta/m) q^3 = 0$. Put $q = A x$, $\tau = \omega t$, $\lambda = \beta A^2/(m\omega_0^2)$ and seek $\omega = \omega_0(1 + c\lambda + \cdots)$,

$$x = x_0 + \lambda x_1 + \cdots .$$

Leading initial conditions $x(0) = 1, x'(0) = 0$ give

$$x_0 = \cos\tau .$$

At first order one finds

$$x_1'' + x_1 = 2c\cos\tau - \cos^3\tau .$$

Using $\cos^3\tau = (3\cos\tau + \cos 3\tau)/4$, the right-hand side contains a term $(2c - 3/4)\cos\tau$ resonant with the leading motion. Choosing $c = 3/8$ removes the forcing that would otherwise generate a term proportional to $\tau\sin\tau$. The frequency then agrees with the action-angle calculation. The remaining particular solution is $\cos 3\tau/32$. Including the homogeneous term required by $x_1(0) = x_1'(0) = 0$ gives

$$x_1 = (\cos 3\tau - \cos\tau)/32 .$$

This separates the leading frequency shift from an actual third-harmonic distortion of the orbit. Fitting only frequency and overlooking the small third harmonic cannot generally reproduce a precise waveform.

In the earlier example $\lambda = 0.040$, so the displacement amplitude at the third harmonic is $A\lambda/32 = 0.125\,\mathrm{mm}$ at $A = 0.10\,\mathrm{m}$. The frequency shift is about $0.15\,\mathrm{s}^{-1}$. After twenty seconds the uncorrected harmonic prediction is nearly three radians out of phase, far larger than the instantaneous third-harmonic displacement. A small-amplitude correction to the right-hand side can therefore cause an order-one error in the long-time phase. Measuring both the third harmonic and frequency gives two independent checks of the same inferred quartic stiffness. A drive electronics harmonic can mimic the former without producing the amplitude-dependent free-oscillation frequency.

Inspect growth on a parametrically driven orbit

Move the drive away from resonance and examine stroboscopic growth. Determine whether changing phase, amplitude or forcing frequency is responsible.

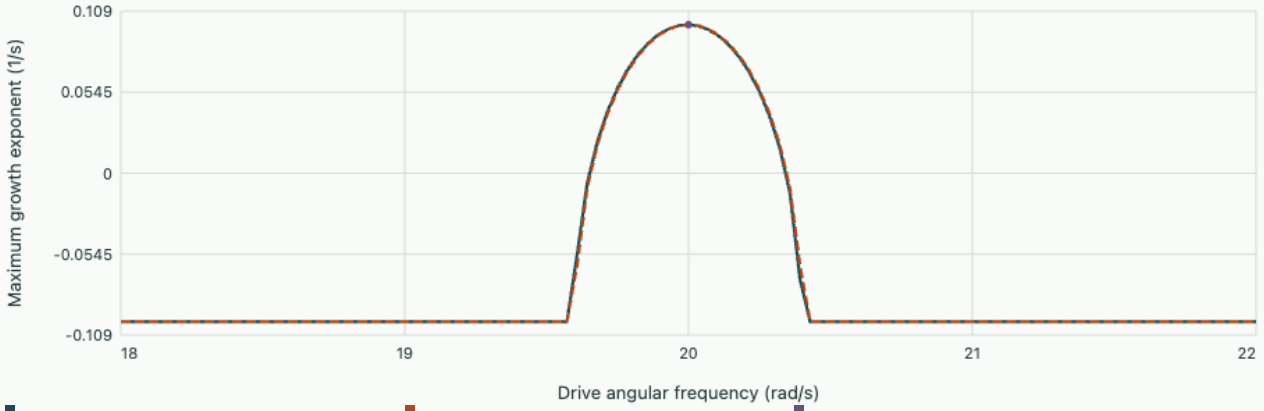
Open this experiment online

Try the calculation

Fractional stiffness modulation **0.08**

Damping γ (1/s) **0.1**

Drive frequency Ω (rad/s) **20**



One-period numerical map | Weak-modulation average | Selected exponent

At Ω = 20 rad/s, numerical exponent 0.09997 s⁻¹ and averaged exponent 0.1 s⁻¹. Amplitude doubling time 6.934 s.

Test overlap of nearby resonances without declaring chaos from one curve

Take a rotor driven by two weak periodic patterns with distinct angular speeds Ω_1, Ω_2. For an isolated pattern j, the earlier calculation gives a resonant centre $P_j = \mathcal{I}\Omega_j$ and half-width $\Delta P_j = 2\sqrt{\mathcal{I}\epsilon V_j}$. Comparing the sum of estimated half-widths with the separation $|P_2 - P_1|$ is a practical indication of when the isolated-island approximation may fail. It does not by itself prove every trajectory between the resonances is chaotic. The relative phases, other harmonics and the actual time dependence determine the global structure.

For the same inertia $0.050\,\mathrm{kg\,m^2}$, choose $\Omega_1 = 8.0\,\mathrm{s^{-1}}$, $\Omega_2 = 8.6\,\mathrm{s^{-1}}$, and forcing energies 0.0020 and $0.00050\,\mathrm{J}$. The centres differ by $0.030\,\mathrm{kg\,m^2/s}$; the individual half-widths are 0.020 and $0.010\,\mathrm{kg\,m^2/s}$. Their nominal intervals just touch. Halving both forcing energies shrinks both widths by $1/\sqrt{2}$, leaving a gap. Doubling both makes the estimated intervals overlap. The speed difference corresponds to a beat period $2\pi/0.6 = 10.47\,\mathrm{s}$, another clock that must be resolved when recording nearby motion.

Transforming into the first rotating frame makes the first perturbation stationary and leaves the second with time dependence proportional to $\cos[\phi - (\Omega_2 - \Omega_1)t]$. A stroboscopic map sampled at the beat period is therefore more informative than a map with unsynchronised lab sampling. Prepare nearby initial conditions, compare their separation over multiple beat periods, and repeat with a refined symplectic time step and a modestly varied observation interval. Short-term stretching near an ordinary hyperbolic fixed point is possible even in a fully integrable pendulum. Evidence for widespread chaotic transport requires more than the growth of one separation during one saddle passage.

A student finds an apparent threshold by plotting whether the momentum ever crosses the midpoint between the two resonant centres. What additional evidence distinguishes irregular transport from a single large regular libration?

▼ **Work through the solution**

A regular libration may cross that midpoint repeatedly inside one island, especially near its separatrix. Plot return pairs at a fixed beat phase, check for closed invariant curves or broader area filling, and track unwrapped relative phase to see whether it slips irregularly between the two patterns. Vary initial conditions and compare the statistical trend with the isolated-width estimate. Confirm that energy balance in the driven system includes explicit work by both patterns; a drift of the autonomous single-pattern energy is expected once the second drive is present.

Use phase reversal to isolate the resonant action transfer

For an initially weak sinusoidal drive, prepare two otherwise identical trajectories with initial relative phases differing by π. Their leading changes of action during a short resonant pulse have opposite signs, since $\dot{I}$ is proportional to $\sin\phi_0$. Half the difference isolates that phase-sensitive term, whereas half the sum can reveal phase-independent heating or higher-order effects. With drive coefficient $A = 0.001\,\mathrm{J}$, duration $2.0\,\mathrm{s}$ and initial phases $\pm\pi/2$, short-pulse first-order predictions differ by $0.0040\,\mathrm{J\,s}$. The test requires that detuning and nonlinear phase drift over the pulse remain small.

Perform the comparison at several durations. A linear growth of the early phase-sensitive difference is compatible with coherent transfer. Subsequent saturation, reversal or a slow oscillation can reflect the resonant pendulum dynamics once the action shift changes the phase. Friction can also saturate a response, but it produces mechanical energy loss and typically modifies the two phase preparations in different ways from conservative trapping. Measuring work from the applied torque and the full phase history adds the independent information needed to separate those explanations. A final amplitude alone gives neither the sign of the action transfer nor its phase dependence.

Tutorial on weak forcing

Far from resonance let a one-mode perturbation be $A\cos(\theta - \Omega t)$ with $A = 0.001\,\mathrm{J}$ and detuning $0.50\,\mathrm{s^{-1}}$. What first-order ripple scale should appear in the old action?

▼ **Work through the solution**

The generating function has amplitude $A/|\Delta| = 0.002\,\mathrm{J\,s}$. For harmonic number one its angle derivative has the same amplitude, so the old action oscillates about the averaged action on that scale. At detuning $0.01\,\mathrm{s}^{-1}$ the nominal ripple would be fifty times larger. Check its size against the action scale and the change of detuning across the ripple before trusting that continuation.

In the rotor example the forcing energy is multiplied by four. What happens to the resonant action, action width and small-amplitude libration period?

▼ **Work through the solution**

The central action is set by $H_0' = \Omega$ and stays $0.40\,\mathrm{kg\,m^2/s}$. The half-width doubles to $0.040\,\mathrm{kg\,m^2/s}$. The libration frequency doubles, so the period halves to $15.7\,\mathrm{s}$. These predictions apply to the exact sinusoidal rotor model; other harmonics of a physical drive may also change as its amplitude increases.

A slowly swept drive seems to carry a phase-locked rotor beyond its original resonance band. Is this inconsistent with the static separatrix estimate?

▼ **Work through the solution**

A varying $\Omega(t)$ moves the resonance centre $\mathcal{I}\Omega(t)$. In coordinates measured from that moving centre the equation for detuning momentum has an additional torque-like term $-\mathcal{I}\dot{\Omega}$. Locking can persist if the available restoring torque can track the ramp and the initial motion stays inside the evolving island. The static width describes a fixed-frequency section. Comparing a captured trajectory to its initial centre after the sweep tests the wrong phase portrait.

Compare a resonance driven by a varying stiffness with a rotating additive potential.

David Tong's Hamiltonian course develops canonical variables and adiabatic invariants.

Distinguish a physical precession from numerical drift

Mechanics in the laboratory

A bound orbit circulates many times about its centre. A small change in the angle of successive closest approaches could be physical or could be accumulated phase error. Construct an exactly solvable family of precessing orbits that lets the two effects be compared without assuming a reference trajectory from another code.

Replace time by the measured polar angle

A particle of mass m moves in a static central potential $V(r) = -k/r + \alpha/(2r^2)$ with $k > 0$. The radial force is $-k/r^2 + \alpha/r^3$. Angular momentum $L = mr^2\dot{\phi}$ is constant. Put $u = 1/r$ and differentiate with respect to ϕ, denoting this derivative by a prime. The identities $\dot{r} = -(L/m)u'$ and $\ddot{r} = -(L^2/m^2)u^2u''$ follow by the chain rule. Subtracting the centripetal term $r\dot{\phi}^2 = (L^2/m^2)u^3$ from radial acceleration gives Binet's equation.

$$u'' + (1+\delta)u = \frac{mk}{L^2},$$

$$\delta = \frac{m\alpha}{L^2}.$$

The sign of the perturbation matters. Positive α gives a repulsive short-range correction, increasing radial frequency per unit angle. A mildly negative correction decreases it. This family is a test potential; a measured nonzero advance in nature need not have this particular cause.

Read apses from a harmonic equation

For $1 + \delta > 0$, write $\nu = \sqrt{1+\delta}$, $u_c = mk/[L^2(1+\delta)]$ and impose $u(0) = u_c(1+e)$, $u'(0) = 0$ with $0 < e < 1$. Then $u = u_c[1 + e\cos(\nu\phi)]$. The radial minimum corresponds to a maximum of u. Its next occurrence is at angle $2\pi/\nu$, so the change of line of apsides per radial period is

$$\Delta\omega = 2\pi\left(\frac{1}{\sqrt{1+\delta}} - 1\right).$$

For $\delta = 0.04$, successive pericentres rotate backwards by approximately 7.0 degrees. At $\delta = 0$ the Kepler ellipse closes. For a strictly circular orbit the algebraic radial frequency has a well-defined small-oscillation limit, but the direction of pericentre of the observed circle has no unique value. A nearly circular orbit therefore has a genuine experimental difficulty even when the integration is accurate.

Compute the bias of a discrete orbit exactly

Shift $v = u - u_c$. On a uniform angular grid of spacing h, the centred second difference gives

$$v_{j+1} = 2v_j - v_{j-1} - (h\nu)^2 v_j.$$

Insert the trial $v_j = e^{ij\psi}$ and simplify:

$$\cos\psi = 1 - \tfrac{1}{2}(h\nu)^2,$$

$$\nu_h = \frac{2}{h}\arcsin(h\nu/2).$$

The numerical mode oscillates if $h\nu < 2$. Inside that range it is too fast, since the arcsine exceeds its argument. A bounded orbit can thus accumulate a negative apsidal error. For an unperturbed Kepler case and a 30-degree grid, the numerical angular frequency is roughly 1.012. Treating this as a physical result would infer a retrograde precession of roughly four degrees per revolution. Refining h halves the angular step and quarters the leading error, because $\nu_h = \nu + \nu^3 h^2/24 + \cdots$.

The experiment plots radial phase after multiple nominal revolutions and reports the implied apsidal angles. Vary the perturbation at fixed step first, then halve the step at fixed perturbation. Distinguish the discrepancy that persists in the continuum from the discrepancy that can be extrapolated away. Starting all curves at pericentre makes the phase convention identical.

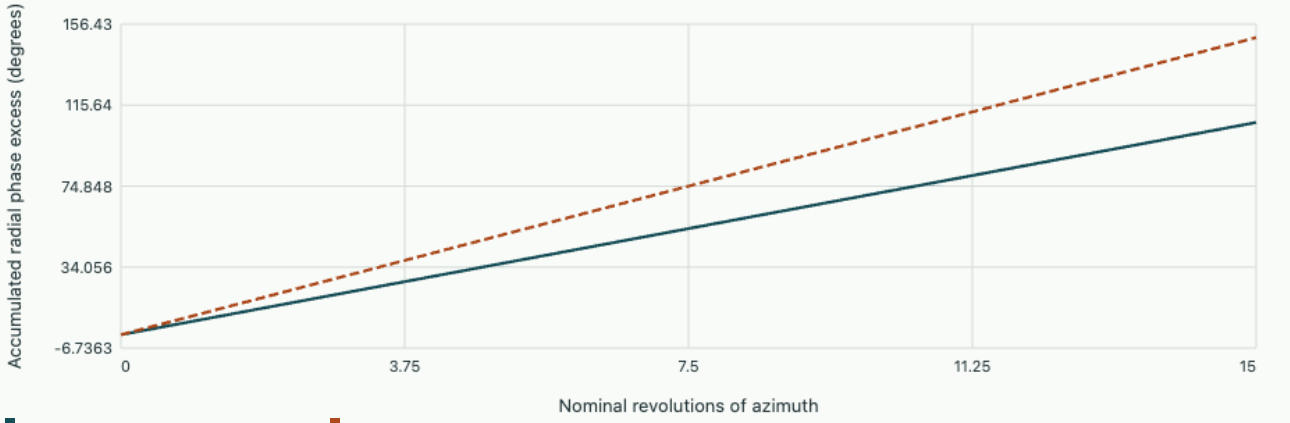
Hold the force exactly Newtonian. Compare the accumulated numerical phase on a 30° grid and 6° grid. The exact radial phase stays at zero in both runs. Use the exported values to estimate how the phase error scales when the angular step is divided by five.

Open this experiment online

Try the calculation

Relative force correction δ **0.04**

Angular grid step (degrees) **24**



Apsidal advance per radial period: exact -6.991°, numerical -9.7105°.

Tutorial

A measured radial period is 364 degrees of azimuthal advance. In this family, which sign of α is indicated and what is δ?

▼ Work through the solution

The radial cycle takes more than 360 degrees, so the radial frequency per unit angle is below one. The inferred correction is attractive,

$$\delta = (360/364)^2 - 1 \simeq -0.02186.$$

A simulation reports advances of −1.20, −0.60 and −0.45 degrees at successive halvings of the angular step. What leading continuum estimate follows?

▼ Work through the solution

With second-order phase bias, the final difference 0.15 equals three times the remaining magnitude of that bias. Adding 0.05 to −0.45 gives −0.40 degrees. The earlier pair gives the same estimate. An additional refinement tests the assumed scaling.

Can a long observation of a closed numerically integrated orbit rule out a time-step error in the orbital period?

▼ Work through the solution

No. Closure depends on a ratio of radial to azimuthal phase advances. Both may be biased in the same proportion. Independently time successive crossings and compare with an analytic period or a finer integration. Different observables can hide different errors.

The midpoint experiment shows the converse sign of phase error for another bounded integrator.

Multiple scales and identifiable dynamics

Derive a system with competing rates and ask which measurements can separate them. Resolve a decaying field and a coupled mechanical instrument before interpreting the numerical trajectory.

1. Closed-loop poles, damping and performance
2. Sampling, saturation and commissioning
3. Impulse response and finite-duration loading
4. Finite-amplitude oscillations and model validity
5. How a magnetic field relaxes into a finite conductor
6. Posing an electromagnetic evolution problem
7. A moving-coil galvanometer as an electrical and mechanical system
8. Eddy-current braking, magnetic penetration and the heat left behind
9. Estimate a cavity linewidth from a finite, biased record
10. Choose a frequency that identifies a coil's stored and dissipated energy
11. Master equations, detailed balance and probability currents
12. How an unobserved environment changes a quantum state

Closed-loop poles, damping and performance

Control engineering

Translate a time-domain requirement into acceptable poles and find when additional gain destabilises a higher-order plant.

FROM FIRST PRINCIPLES

Connect characteristic roots to decay and oscillation

Near an equilibrium, many closed-loop positioning stages reduce to a constant-coefficient second-order equation. Consider the free response after the reference has become constant. Let x be deviation from its final value, $\omega_n > 0$ the natural frequency and ζ the dimensionless damping ratio. Try an exponential response because time derivatives of an exponential reproduce the same waveform.

$$\ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x = 0,$$
$$x = e^{st} \implies s^2 + 2\zeta\omega_n s + \omega_n^2 = 0.$$

For $0 < \zeta < 1$ solving the quadratic produces conjugate roots. Their real part sets the exponential decay rate and their imaginary part sets the observed angular oscillation frequency.

$$s_{1,2} = -\zeta\omega_n \pm i\omega_n\sqrt{1-\zeta^2},$$
$$x = e^{-\zeta\omega_n t}(A\cos\omega_d t + B\sin\omega_d t).$$

Here $\omega_d = \omega_n\sqrt{1-\zeta^2}$ and the initial displacement and velocity set A, B. A measured period gives ω_d while the decaying peak envelope gives $\zeta\omega_n$. Stable poles describe decay of free modes. The numerator of a forced response also influences how prominently each mode appears.

DERIVATION AND APPLICATIONS

Separate a cubic stability boundary from oscillation frequency

Feedback gain can move an oscillatory pair across the imaginary axis even with every characteristic coefficient positive. Take a continuous linear unity-feedback loop whose open-loop transfer is $K/[s(s+a)(s+b)]$ with positive a, b, K. The integral mode can represent a position obtained from velocity; the other poles can represent actuator and measurement dynamics. Add unity to the loop transfer and clear its denominator to obtain the closed-loop polynomial.

$$p(s) = s^3 + (a+b)s^2 + ab\,s + K.$$

A marginal oscillation has $s = i\omega$ with real nonzero ω. Separate the real and imaginary parts. The imaginary equation sets the frequency without requiring the gain, while the real equation selects the gain that supports it.

$$K - (a+b)\omega^2 = 0,$$
$$\omega(ab - \omega^2) = 0,$$
$$K_{\text{crit}} = ab(a+b).$$

At this gain the cubic factors exactly into a real stable pole and a pair of purely imaginary roots. The cubic Routh first column determines which side of the boundary is stable. Its first two entries are positive, making the sign of the third entry decisive for positive gain.

$$p_{\text{crit}}(s) = (s + a + b)(s^2 + ab), \qquad \text{first column: } 1,\ a+b,\ ab - \frac{K}{a+b},\ K.$$

Below the threshold all entries have the same sign. Above it there are two sign changes and therefore two right-half-plane roots. The model treats actuator authority as unlimited and omits nonlinear friction; a finite observed oscillation amplitude requires additional physics beyond the pole crossing.

WORKED EXAMPLE

Predict the onset period of a positioning loop

Set $a = 2.00$ s⁻¹ and $b = 8.00$ s⁻¹. Marginal gain is $K_{\text{crit}} = 160$ s⁻³. The oscillation has frequency 4.00 rad/s, hence period $2\pi/4.00 = 1.57$ s. At $K = 120$ s⁻³ the third entry is $16 - 12 = 4.00$, giving stability. At $K = 180$ s⁻³ it is -2.00 and the pair grows.

Change the inputs and inspect this calculation below

SECOND APPLICATION

Recover the two lag rates from a marginal test

Another loop reaches marginality at $K = 60.0$ s⁻³ with a measured oscillation of $\sqrt{10}$ rad/s. The rates have product $ab = 10.0$ s⁻² and sum $K/(ab) = 6.00$ s⁻¹. These observations cannot describe two positive real lag rates because the maximum product at fixed sum is $6^2/4 = 9.00$ s⁻². The failed consistency check flags model structure or calibration before attributing a unique time constant to either stage.

Interactive calculation · Change the inputs of this example

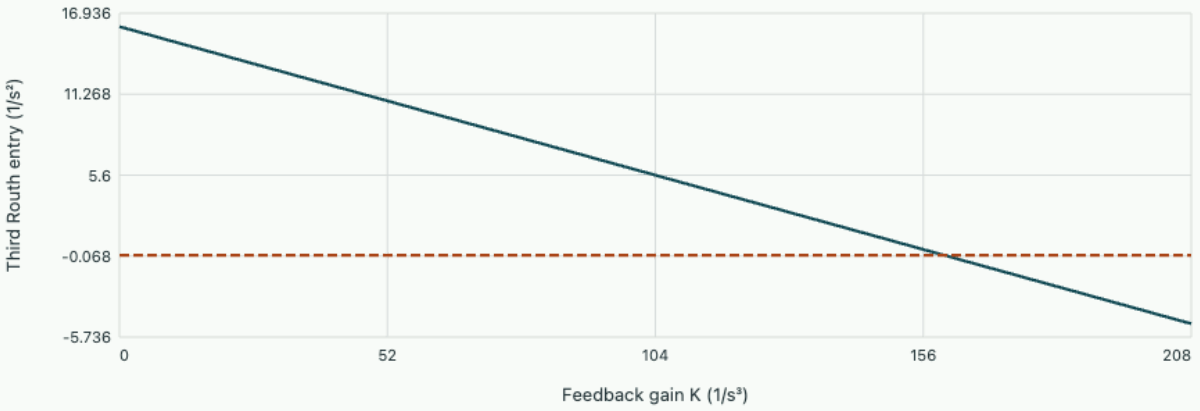
Use the adjustable quantities to test the two worked calculations above. The plot recomputes this derivation's balance or constitutive relation; its numerical readout reports the quantities needed for the forward calculation and the inverse or design question. Keep the fixed assumptions stated in the readout when comparing with the example values.

Open this live calculation

Try the calculation

First lag rate a (1/s) **2**

Second lag rate b (1/s) **8**



Critical gain 160.000 /s³; onset period 1.571 s. The third pole at onset is –10.000 /s.

Relate polynomial coefficients to observed motion

For a normalised second-order system, the free response satisfies $\ddot{y} + 2\zeta\omega_n\dot{y} + \omega_n^2 y = 0$. Substitution of e^{st} gives

$$s = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1}.$$

The real part sets the decay rate. For $0 < \zeta < 1$, the imaginary part $\omega_d = \omega_n\sqrt{1-\zeta^2}$ sets the time between successive extrema. Doubling ω_n at fixed ζ halves all characteristic times but leaves the fraction of overshoot unchanged.

Assume a unit DC gain, a constant numerator and zero initial displacement and velocity. The step response is

$$y = 1 - e^{-\zeta\omega_n t}[\cos(\omega_d t) + (\zeta\omega_n/\omega_d)\sin(\omega_d t)].$$

Differentiate the bracket together with its envelope. The cosine terms cancel and the remaining derivative is proportional to $e^{-\zeta\omega_n t}\sin(\omega_d t)$. The first peak occurs at $t_p = \pi/\omega_d$. Substituting that time gives the fractional overshoot

$$M_p = \exp\left(-\frac{\pi\zeta}{\sqrt{1-\zeta^2}}\right).$$

Design for a specified peak

Suppose a positioning axis may overshoot a 10 mm command by at most 1 mm. Solving the previous expression for the limiting damping ratio gives

$$\zeta = -\ln M_p / \sqrt{\pi^2 + (\ln M_p)^2} = 0.591.$$

At $\omega_n = 5$ rad/s the corresponding poles are approximately $-2.96 \pm 4.03i$ per second and the first peak occurs at 0.779 s. A nominal oscillatory second-order system with this damping just meets the 1 mm bound. A zero near the dominant poles, a saturated drive or another slow pole changes the transient, so the result should be checked against the full model after the poles are chosen.

A settling specification differs from a peak specification. The absolute normalised error stays below the envelope $e^{-\zeta\omega_n t}/\sqrt{1-\zeta^2}$. To guarantee a 2% settling band using this bound, solve $t \geq -\ln[0.02\sqrt{1-\zeta^2}]/(\zeta\omega_n)$. This conservative time for the example is about 1.40 s. The commonly used $4/(\zeta\omega_n)$ is a convenient estimate. The envelope bound spells out both the tolerance and its assumptions.

Follow a gain change on a third-order plant

A plant with two lags and an integrator has open-loop transfer $K/[s(s+1)(s+2)]$ in units normalised to a common time base. Close unity negative feedback. The characteristic polynomial is $s^3 + 3s^2 + 2s + K$. Its coefficients are all positive for $K > 0$, which rules out positive real roots but allows an unstable complex pair. Form the first column of the Routh array: $1,\ 3,\ (6-K)/3,\ K$. All entries must have the same sign for strict stability. Thus $0 < K < 6$.

At $K = 6$ the polynomial factors as $(s+3)(s^2+2)$. The step excites an undamped pair at angular frequency $\sqrt{2}$, alongside a decaying component. At $K = 7$ the first-column signs change twice, so two roots have crossed into the right half-plane. The input-output gain alone did not reveal this upper bound. Additional lags allow feedback to arrive out of phase with motion. Chapter 4 reaches the same type of conclusion by examining sine-wave responses, which can often be measured even when coefficients of a polynomial are poorly known.

Find the effect of a finite closed-loop zero

Pole locations specify the available exponential terms of a causal linear response. Their coefficients depend on the numerator and initial condition. To isolate this distinction, choose two positive times $\tau_1 \neq \tau_2$ and a transfer function $G(s) = (1 + T_z s)/[(1 + \tau_1 s)(1 + \tau_2 s)]$. It has unit static gain for every T_z, so changing the numerator neither moves the poles nor changes the final unit-step response. For zero initial state and a unit step $Y(s) = G(s)/s$. Write $Y = 1/s + A/(s + 1/\tau_1) + B/(s + 1/\tau_2)$.

The initial value $y(0^+) = 0$ gives $A + B = -1$. The initial-slope theorem gives

$$\dot{y}(0^+) = \lim_{s \to \infty} sG(s) = T_z/(\tau_1 \tau_2);$$

equivalently $-A/\tau_1 - B/\tau_2 = T_z/(\tau_1 \tau_2)$. Solving and transforming back gives

$$y(t) = 1 - \frac{\tau_1 - T_z}{\tau_1 - \tau_2} e^{-t/\tau_1} + \frac{\tau_2 - T_z}{\tau_1 - \tau_2} e^{-t/\tau_2}.$$

For $T_z = \tau_1$ one mode has zero coefficient in this particular input–output map. A physical internal state with that mode can still respond to a nonzero initial condition or another input. With $T_z < 0$ the zero lies in the right half-plane and the initial output slope opposes the eventual response. That inverse response can be a serious limitation when interpreting an early excursion as a sign error in the actuator.

Time the undershoot of two competing actuator paths

Consider an actuator that changes a measured quantity through a slow positive effect and an immediate faster opposing effect. Let $\tau_1 = 4$ s, $\tau_2 = 1$ s and $T_z = -2$ s. The poles are both stable, and the output approaches one under a unit command. Yet the response is

$$y = 1 - 2e^{-t/4} + e^{-t},$$

with initial slope $-0.50\,\mathrm{s}^{-1}$. Set $\dot{y} = 0.5e^{-t/4} - e^{-t} = 0$. It follows that $e^{-3t/4} = 0.5$, so the minimum occurs after $t = 4 \ln 2/3 = 0.924\,\mathrm{s}$. Its value is

$$1 - 2(0.794) + 0.397 = -0.191.$$

For a 10 mm final position change this model briefly produces about 1.91 mm motion in the wrong direction. A physical travel limit 1.0 mm behind the starting point would therefore be violated even though the real parts of both poles meet a settling-time target. Compare instead $T_z = 0$ with identical poles and DC gain: the initial slope vanishes and the response rises monotonically. Before changing controller gain to correct an observed undershoot, measure which pathway acts first. Inverting the right-half-plane zero with a causal stable feedforward filter is impossible, since exact inversion would introduce an unstable pole.

For this response the first crossing back through the initial position is later than the minimum. Let $u = e^{-t/4}$; solving $1 - 2u + u^4 = 0$ gives the trivial root $u = 1$ at release and a further root near

0.544. Thus the measured position remains on the wrong side for about $-4\ln(0.544) = 2.44$ s. A simple diagnostic that samples only once after two seconds would label the overall gain as negative. At one minute the same test reads nearly the full positive change. Recording the full waveform distinguishes an initial inverse response from an overall sensor polarity error. A disturbance entering the two paths at a different location can exhibit yet another zero, which is why a zero should always be quoted for a specified input and output.

OneSimulation · Which path is responsible for the initial undershoot?

Two stable first-order actuators contribute to the same output. Give the slow path gain $+2$ and a two-second time constant, the fast path gain -1 and a quarter-second time constant. Both start at rest and receive the same unit step. Predict the direction of the first and last movements from the initial slope and final value.

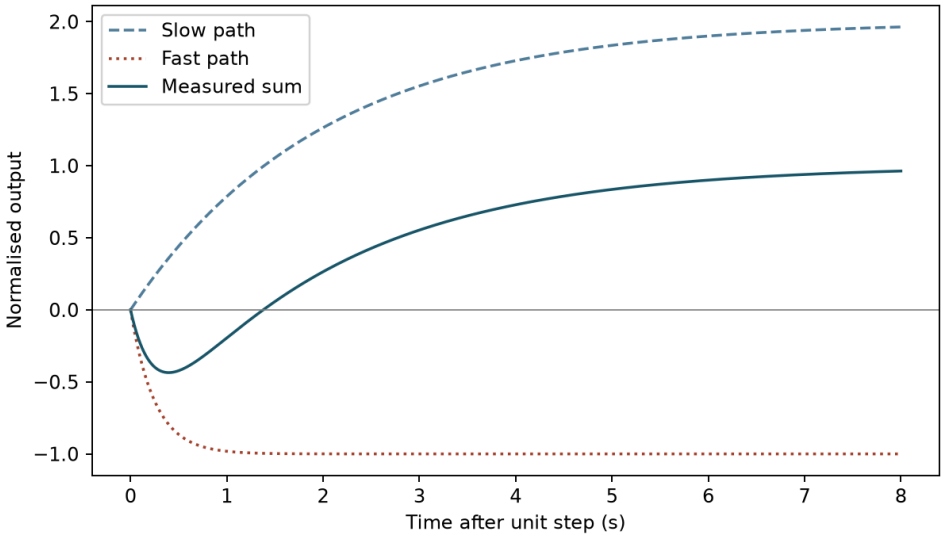
$$y(t) = 1 - 2e^{-t/2} + e^{-4t},$$

$$\dot{y}(0) = 1 - 4 = -3\ \mathrm{s}^{-1},$$

$$y(\infty) = 1.$$

Integrate the two state equations using the fourth-order OneSimulation time stepper and compare every sample with the exponential solution. The grid minimum is –0.436 at 0.40 s. The component curves identify the mechanism. Initially the negative path builds up fastest; once it has nearly settled the slow positive path continues to grow. Vary the two time constants while keeping their gains fixed to see when the initial slope changes sign.

```
"""A positive step drives two stable paths with opposite output signs."""
import matplotlib.pyplot as plt
import numpy as np
from onesim.numerics.time.methods.backends.fixed_step import FixedStepRungeKutta

tau = np.array([2., 0.25])  # seconds, slow and fast paths
weights = np.array([2., -1.])
solver = FixedStepRungeKutta("rk4")
run = solver.integrate(lambda t, x: (1-x)/tau, np.zeros(2), 0., 8., step=0.02)
t = run.times
terms = run.states * weights
output = terms.sum(axis=1)
reference = (1-np.exp(-t[:, None]/tau)) @ weights
assert np.max(abs(output-reference)) < 1e-6
index = np.argmin(output)
print(f"First minimum on the grid: {output[index]:.3f} at {t[index]:.2f} s")

fig, ax = plt.subplots(figsize=(7, 4), constrained_layout=True)
ax.plot(t, terms[:, 0], "--", color="#51809c", label="Slow path")
ax.plot(t, terms[:, 1], ":", color="#a5442e", label="Fast path")
ax.plot(t, output, color="#19596a", label="Measured sum")
ax.axhline(0, color="0.5", linewidth=0.7)
ax.set(xlabel="Time after unit step (s)", ylabel="Normalised output")
ax.legend()
fig.savefig("control-engineering.png", dpi=170)
```



The ability to plot both internal states separates an inverse response caused by competing paths from a sign mistake in the output sensor.

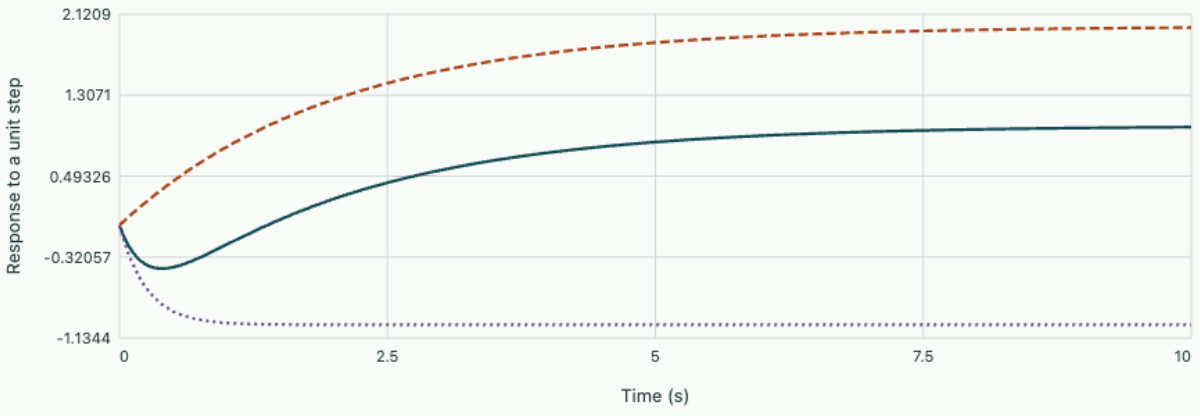
Download the Python example

Change the inputs of this exact example and compare the recalculated curves with the recorded OneSimulation output above. The browser evaluates the stated analytical or modal reference.

Open this live calculation

Try the calculation

Slow-path time constant (s)	2
Fast-path time constant (s)	0.25



Initial slope -3.000 s^{-1}; final value1. Minimum -0.436 at 0.396 s. This is the exponential reference for the two states integrated in the OneSimulation example.

Assign the poles of a spring-mounted positioning stage with PID

A positioning loop can measure displacement, accumulate position error and damp measured motion. Let a stage of mass m, viscous damping c and return spring k obey $m\ddot{y} + c\dot{y} + ky = u + d$, with actuator force u and disturbance force d. The controller uses $e = r - y$, $\dot{z} = e$, and $u = K_p e + K_i z - K_d \dot{y}$. Differentiating the measured position supplies damping without turning a step in commanded position into an impulsive force command.

To find the natural motion, set $r = d = 0$ and substitute $u = -K_p y + K_i z - K_d \dot{y}$. Move the feedback terms to the left and differentiate once to eliminate z, using $\dot{z} = -y$. The resulting characteristic polynomial is

$$m s^3 + (c + K_d) s^2 + (k + K_p) s + K_i .$$

The derivative gain adds viscous damping, the proportional gain adds stiffness, and closing an integral loop adds a state. At any stable equilibrium with a constant reference, $\dot{z} = 0$ requires $y = r$. The integrator can then supply whatever constant force the spring and disturbance require, within actuator limits. The transient still depends on how the reference enters the controller.

Take $m = 2.0\,\mathrm{kg}$,

$$c = 3.0\,\mathrm{N\,s\,m^{-1}},$$

$$k = 5.0\,\mathrm{N\,m^{-1}}.$$

Suppose the desired poles are $-2, -3, -4\,\mathrm{s^{-1}}$. Multiplying the three factors gives the desired polynomial $2(s+2)(s+3)(s+4) = 2s^3 + 18s^2 + 52s + 48$. Comparing coefficients produces $K_d = 15\,\mathrm{N\,s\,m^{-1}}$,

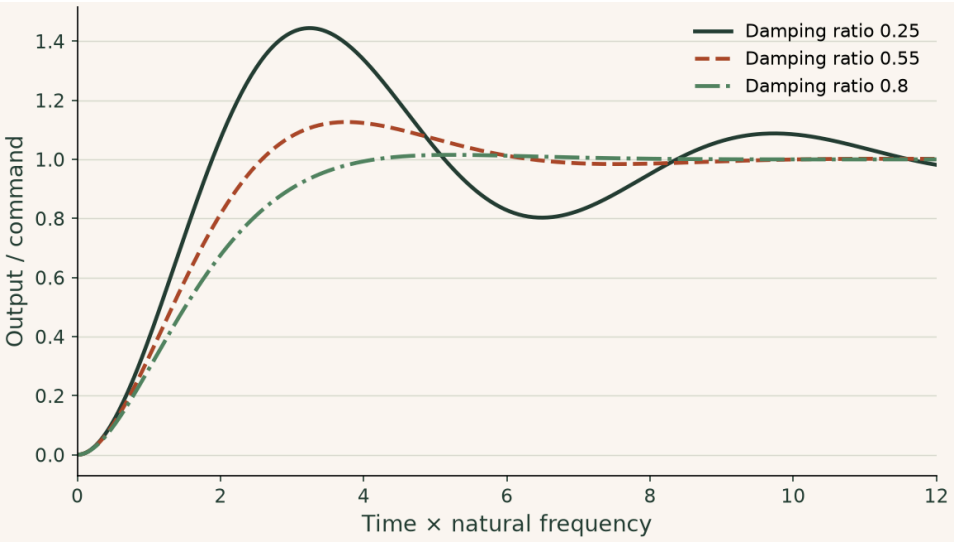
$$K_p = 47\,\mathrm{N\,m^{-1}},$$

$$K_i = 48\,\mathrm{N\,m^{-1}s^{-1}}.$$

For a 20 mm command from rest the initial proportional force is $0.940\,\mathrm{N}$. At the final resting position with $d = 0$, the spring needs $0.100\,\mathrm{N}$. $K_i z$ supplies it and the proportional contribution is zero. A filtered derivative introduces an additional pole, so this ideal coefficient match must be checked again against the implemented controller bandwidth.

Settle a tracking step without hiding its peak

Apply a unit step to a unity-DC second-order transfer function with no zeros. Keep natural frequency fixed. The first maximum occurs when $\omega_d t = \pi$; use the waveform both to check the overshoot formula and to see how late oscillations affect a settling band.



At $\zeta = 0.55$ the first overshoot is about 12.6%. A design with a smaller damping ratio may cross the command sooner while taking much longer to stay within a narrow tolerance.

▶ **Read values from the plot**

Interactive calculation · change the inputs of this example

Move through critical damping and compare overshoot, peak time and the duration of the tail.

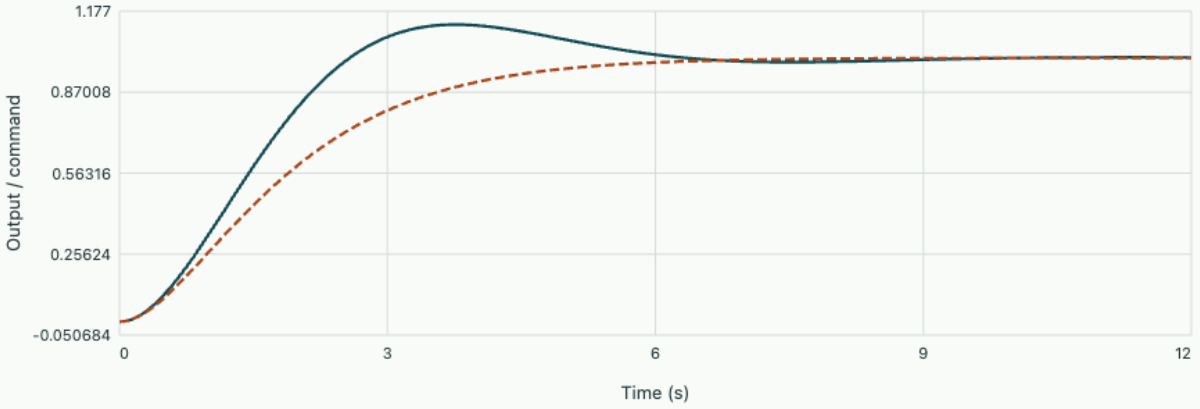
The curves below are recalculated from this section's equations in your browser. The fixed reference plot above remains available for comparison.

Open this live calculation

Try the calculation

Damping ratio **0.55**

Natural frequency (rad/s) **1**



Selected poles Critical damping

First overshoot 12.63%; first peak at 3.762 s.

Sampling, saturation and commissioning

Control engineering

Carry a continuous-time design onto a sampled actuator, predict where its operating assumptions fail, and distinguish an observed final value from an equilibrium.

FROM FIRST PRINCIPLES

Integrate a held command across one sample

A sampled actuator holds its command fixed until the next update. Consider a stable first-order plant $\tau\dot{y} + y = Ku$ with constant coefficients and sample interval h. Let y_k be the value at the start of interval k and u_k the command held throughout it. During that interval the equilibrium for the applied command is the constant Ku_k. Subtract it and integrate the homogeneous decay over exactly one interval.

$$\frac{d}{dt}(y - Ku_k) = -\frac{y - Ku_k}{\tau},$$
$$y_{k+1} - Ku_k = e^{-h/\tau}(y_k - Ku_k).$$

Set $a = e^{-h/\tau}$. The resulting discrete plant is an exact sample-to-sample model under a zero-order hold. For proportional feedback with unit reference scale, substitute $u_k = g(r_k - y_k)$ to see where the sampled free response pole lies.

$$y_{k+1} = ay_k + K(1 - a)u_k,$$
$$z_{\mathrm{cl}} = a - K(1 - a)g.$$

A sample-to-sample deviation is multiplied repeatedly by z_{cl}, so decay requires $|z_{\mathrm{cl}}| < 1$. This condition depends on the time at which the command is applied. A full sample of computational delay changes the state update and its characteristic equation.

Background · Discrete pole

$$z = e^{sT_s}$$

This maps a continuous modal exponent into a sampled multiplier over interval Ts. Different continuous frequencies can give the same sampled phase.

DERIVATION AND APPLICATIONS

Quantify the phase and attenuation introduced by a hold

A digital controller sends a piecewise-constant command to a continuous actuator. Even ideal sampling and instantaneous arithmetic do not make the resulting fundamental sinusoid identical to the requested one. Consider samples of a complex sinusoid at period h. A zero-order hold outputs each sample unchanged over the following interval. For times $t = kh + s$ within interval k, the ratio of held signal to the desired continuous sinusoid is periodic in the within-sample time s.

$$u_h(kh+s) = e^{i\omega kh},$$
$$\frac{u_h(kh+s)}{e^{i\omega(kh+s)}} = e^{-i\omega s}, \quad 0 \le s < h.$$

The zero-frequency Fourier coefficient of this periodic ratio gives the component at the original sinusoidal frequency. The remaining coefficients produce spectral images spaced by the sample frequency. Integrate the ratio over one hold interval and factor out the midpoint phase.

$$H_0(i\omega) = \frac{1}{h}\int_0^h e^{-i\omega s}\,ds = e^{-i\omega h/2}\,\frac{\sin(\omega h/2)}{\omega h/2}.$$

Below Nyquist the sine factor is positive. The fundamental therefore has phase equivalent to a delay of half a sample and a frequency-dependent gain. For small ωh, the sine expansion makes the leading attenuation quadratic while phase lag is linear.

$$\arg H_0 = -\frac{\omega h}{2},$$
$$|H_0| = 1 - \frac{(\omega h)^2}{24} + O((\omega h)^4).$$

This calculation concerns the fundamental component of a held command. Additional computation latency and sensor filtering accumulate their own phase. Frequencies approaching Nyquist also require examining the sampled closed-loop dynamics directly.

WORKED EXAMPLE

Check a servo's phase allowance

At 50.0 Hz crossover and 1.00 kHz sampling,

$$\omega h/2 = \pi(50/1000) = 0.1571.$$

The hold contributes -9.00° and its fundamental magnitude is 0.9959. A further full sample of computation delay contributes -18.0°, giving -27.0° combined at that frequency. The small gain

change would conceal this phase cost in a magnitude-only measurement.

Change the inputs and inspect this calculation below

SECOND APPLICATION

Choose a sampling interval from a phase budget

At a proposed crossover of 80.0 Hz, allocate at most 6.00° lag to the hold. Since the magnitude of the phase in degrees is $180fh$, the interval must satisfy $h \le 6/(180 \times 80) = 0.417$ ms, or a sample rate of at least 2.40 kHz. At this limit the fundamental magnitude is

$$\sin(6^\circ)/(6\pi/180) = 0.9982.$$

The resulting condition addresses the hold alone.

Interactive calculation · Change the inputs of this example

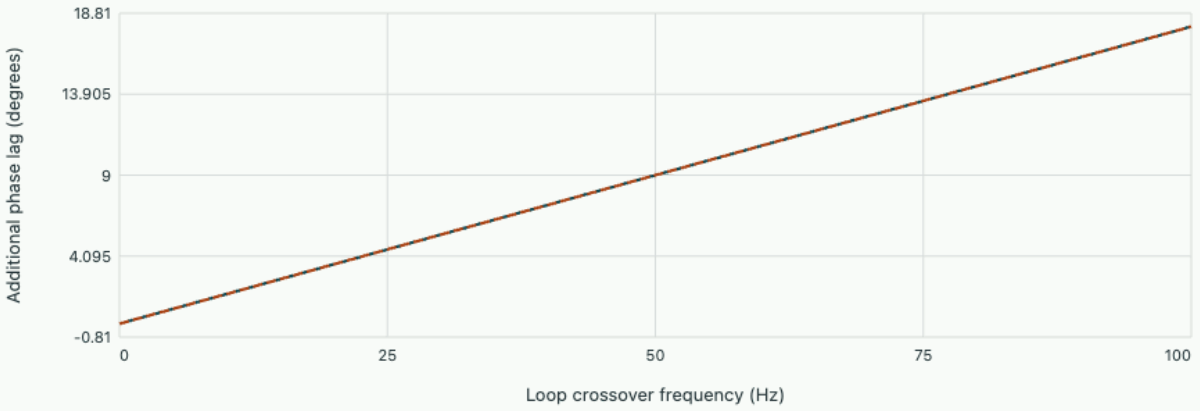
Use the adjustable quantities to test the two worked calculations above. The plot recomputes this derivation's balance or constitutive relation; its numerical readout reports the quantities needed for the forward calculation and the inverse or design question. Keep the fixed assumptions stated in the readout when comparing with the example values.

Open this live calculation

Try the calculation

Sampling frequency (Hz) **1000**

Computation delay / sample **0**



At50 Hz, additional lag=9.000° and hold magnitude=0.99589. Use only the first hold lobe for this phase interpretation.

Why one sample of latency can change the stability margin

A desktop position servo has a first-order response with time constant τ. A digital controller samples position every h seconds and holds each command constant until another is applied. Two firmware versions use the same sample period and proportional gain, yet one applies the command one whole sample later. Derive their closed-loop recurrences before attributing an observed oscillation to a different physical plant.

For the unit-gain plant $\tau\dot{x} + x = u$, solve the differential equation over a single interval with constant input u_k. The result is

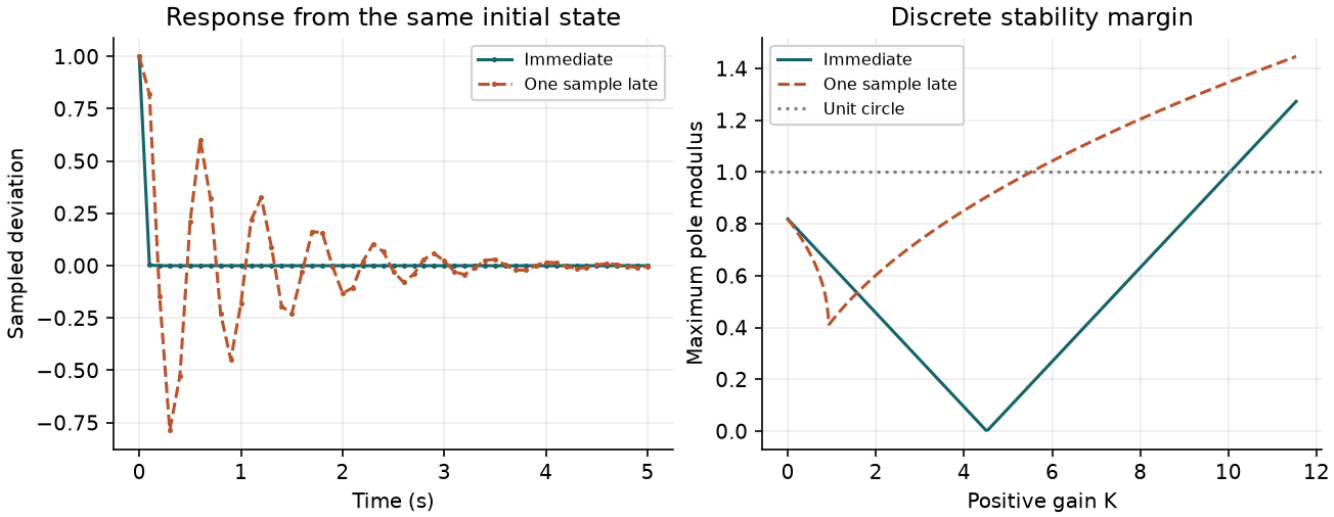
$$x_{k+1} = a x_k + b u_k,$$

where $a = e^{-h/\tau}$ and $b = 1 - a$. Under immediate regulation to zero the command is $u_k = -Kx_k$. Then there is one closed-loop pole, $a - bK$, and the positive gain range is $0 \le K < (1 + a)/b$. At equality the pole sits at minus one and a step-to-step sign alternation persists.

With a one-sample delay the command is $u_k = -Kx_{k-1}$. A trial solution $x_k = z^k$ now gives

$$z^2 - az + bK = 0.$$

For $0 < a < 1$ and $K \ge 0$ the two roots both lie inside the unit circle exactly when $bK < 1$. One can see the source of the upper limit by following the roots. At low gain they are real, lie between zero and a, and meet when $bK = a^2/4$. Beyond that value they are conjugate and their common modulus is $\sqrt{bK}$, since their product equals bK.



The nearly deadbeat immediate update has one small pole. Adding a remembered sample gives two oscillatory poles for exactly the same plant and feedback gain.

Take $\tau = 0.50$ s, $h = 0.10$ s and $K = 4.5$. The exact hold gives $a = 0.81873$, $b = 0.18127$. Immediate application produces a pole near 0.0030, while the delayed poles are about $0.4094 \pm 0.8051i$, of modulus 0.903. Starting at unit displacement with the old stored sample equal to zero, the delayed plant first decays naturally. If the old sample had instead already equalled one, the first two data points would differ even though the subsequent recurrence and its poles were unchanged.

The complex argument of a delayed pole supplies another quantity to measure. At the default gain the pole angle is approximately 1.10 radians per sample. The damped oscillation therefore takes $2\pi/1.10$ samples per cycle, approximately 0.57 s of physical time. Its envelope shrinks to about $0.903^{10} = 0.36$ of its original value over one second. Estimate both quantities from a simulated sequence. Divide the measured period by the sampling interval before comparing with the pole angle. When the response is sampled sparsely, looking only at the step-to-step change of sign can obscure an oscillation that takes nearly six samples to repeat.

Choose a gain for which immediate feedback is stable and delayed feedback is unstable, and quantify the growth.

▼ Work through the solution

The two limiting gains are 10.033 and 5.517. At $K = 7$, the immediate pole is

$$0.81873 - 7(0.18127) \simeq -0.450.$$

The delayed conjugate pair has modulus $\sqrt{7(0.18127)} \simeq 1.126$, so its envelope grows by about 12.6% per sample. In the notebook, compare these predictions using the same physical-time horizon. If an actual processor has a fixed time delay, shortening h changes the delay in samples; the one-sample derivation must then be replaced with the actual timing model.

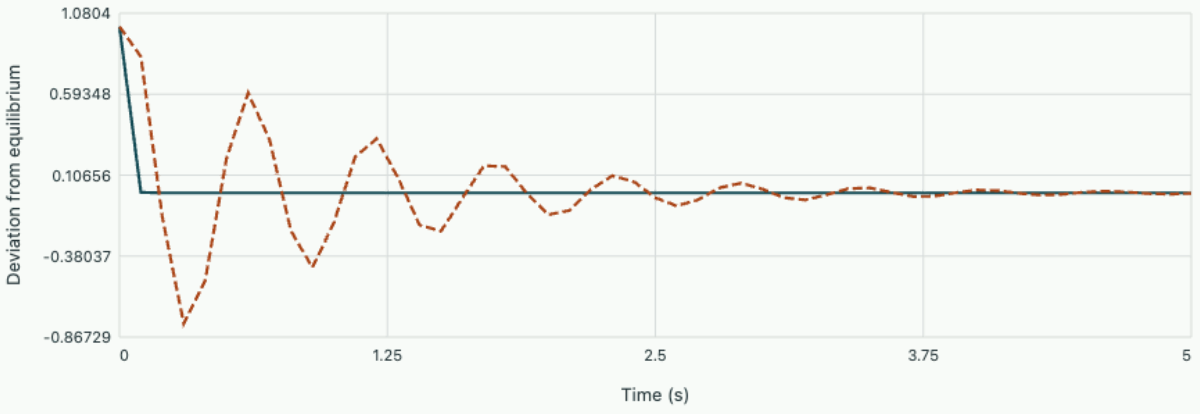
Interactive calculation · Sample delay motion

Vary the notebook inputs and compare both plotted observables. The same equations generate the browser curves and the notebook reference; fixed material, geometry and model assumptions are stated in the readout.

Open this live calculation

Try the calculation

Feedback gain	**4.5**
Plant time constant (s)	**0.5**
Sample period (s)	**0.1**



Selected pole moduli: immediate 0.003, delayed 0.903. Immediate gain limit 10.033; delayed limit 5.517. Both share the same hold map; one extra sample changes the characteristic equation.

Download the editable notebook for this example

Interactive calculation · Sample delay poles

Vary the notebook inputs and compare both plotted observables. The same equations generate the browser curves and the notebook reference; fixed material, geometry and model assumptions are stated in the readout.

Open this live calculation

Try the calculation

Feedback gain	4.5
Plant time constant (s)	0.5
Sample period (s)	0.1



Selected pole moduli: immediate 0.003, delayed 0.903. Immediate gain limit 10.033; delayed limit 5.517. Both share the same hold map; one extra sample changes the characteristic equation.

Download the editable notebook for this example

Open editable notebook Download .ipynb

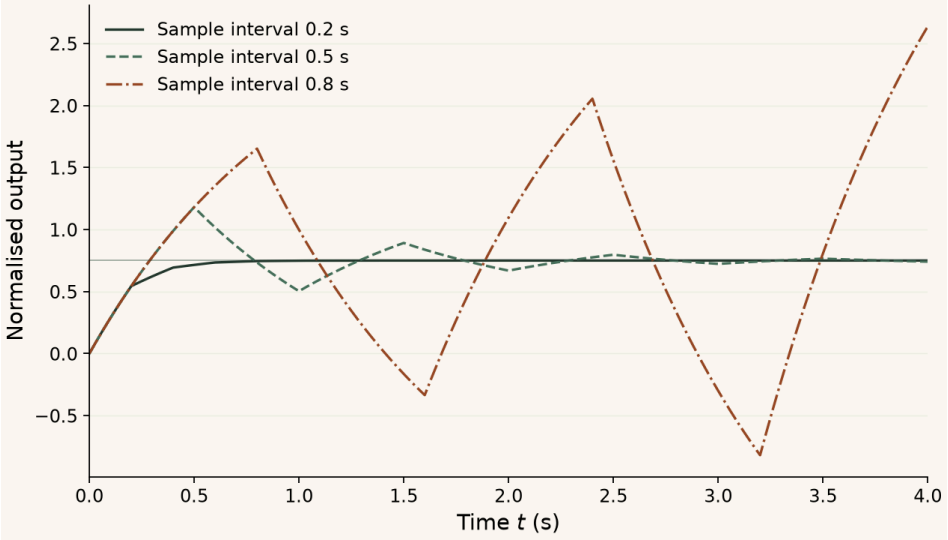
Derive the held-input model

A processor reads a sensor every h seconds and holds its most recent command until the next sample. Solve $\tau\dot{y} + y = Ku_k$ exactly from kh to $(k+1)h$ with this constant input. The result is

$$y_{k+1} = ay_k + K(1-a)u_k,$$

where $a = e^{-h/\tau}$. This recursion is exact at the sampling instants for the assumed hold. Taking a large Euler step through the continuous closed-loop equation describes a different experiment. A real controller has only the sampled information available when it computes u_k.

With proportional control $u_k = K_p(r_k - y_k)$ the discrete pole is $z = a - KK_p(1 - a)$. Stability requires $|z| < 1$. A negative pole of modulus less than one alternates sign each update while decaying. For $K = 1,\ \tau = 1$ s and $K_p = 3$, $h = 0.2$ s gives $z = 0.275$; $h = 0.8$ s gives $z = -1.203$. The same nominal continuous design gives radically different sampled records. For positive gain the upper gain limit is $(1 + a)/[K(1 - a)]$. Additional computation and transport delays lower the available margin further.



Same plant and gain, three sample intervals. The lines show the exact continuous output between updates; the held quantity is the actuator command.

Locate the sampling limit in a computed trace

Keep the plant integration at 40 RK4 substeps per hold. Predict which sample intervals give a monotone response, a decaying oscillation or a growing one. Compare the stored OneSimulation traces on either side of 0.69 s. Estimate the discrete pole from two successive output increments; their ratio removes the unknown equilibrium.

At 0.65 s the measured pole is −0.9118, while at 0.70 s it is −1.0137. The latter increasingly overshoots even though the internal integration resolves each hold. Try 0.69 s and 0.70 s with `python sampled-feedback.py --period 0.69`, then double `--substeps` to check that the stability change is a sampling effect. Download the runnable OneSimulation model.

Why a saturated integrator remembers the past

Let $v_k = K_p e_k + I_k$ be the requested command, and u_k its value clipped to the actuator range. An unconstrained integrator adds $hK_i e_k$ to I_k at every sample. While an unreachable command holds a heater at its upper limit, positive error keeps accumulating; when the command becomes reachable

again the stored integral may keep the heater fully on. This delayed recovery is called windup. Logging only actual heater power hides the growth of the internal demand. Log both signals and the integral state.

A back-calculation update feeds the clipping discrepancy back to the integrator:

$$I_{k+1} = I_k + h[K_i e_k + (u_k - v_k)/T_t].$$

When the request exceeds the upper stop, the second term is negative. At fixed error and fixed clipping level, this update attracts the requested command toward a value consistent with saturation, provided its discrete tracking dynamics are stable. For example, with $h = 0.1$ s, $K_i = 1$, $e = 2$, $u = 5$, $v = 8$, and $T_t = 0.5$ s, the integral increment is -0.4, compared with $+0.2$ without correction. Tuning T_t must account for the sample interval and actuator dynamics. Anti-windup cannot make an infeasible steady torque available.

Use a numerical result to distinguish two errors

Return to the OneSimulation tracking experiment. It models a continuous, unsaturated proportional loop and reports the error at a finite time. For a unit step from rest with loop gain $g = KK_p$, the error after N closed-loop time constants is $e_N = (1 + ge^{-N})/(1 + g)$. With $g = 8$ and $N = 6$, predict 0.11331. Its difference from the equilibrium $1/9$ is transient motion. Halving the plant speed changes the physical time needed to reach the same dimensionless sample, while this sampled error stays fixed.

Test the interpretation by recording at three and six time constants and taking the ratio of the two deviations from equilibrium. Predict e^{-3} for the later deviation divided by the earlier. A persistent discrepancy warrants checking the recorded timestamps, the implementation of the step and whether a limit was active. To examine sampling or clipping, explicitly add these mechanisms to the model before comparing with the equations above. The logged linear-loop runs answer the finite-record question directly.

Derive an exact sampled stability boundary with a delayed command

A digital controller may acquire a sample at the start of an interval but only deliver its computed command at the following update. The one-sample computation delay then introduces an additional dynamic state. Consider a first-order plant $\tau\dot{y} + y = Ku$ with an ideal zero-order hold and sampling period T. Its exact sample-to-sample relation is

$$y_{k+1} = a y_k + b u_k,$$

with $a = e^{-T/\tau}$ and $b = K(1 - a)$. Assume proportional negative feedback about a zero reference and express gains with dimensions chosen so that bg is dimensionless.

With negligible computation delay, $u_k = -g y_k$ yields one pole $a - bg$, so stability requires $|a - bg| < 1$. With a one-sample delay the applied command is $u_k = -g y_{k-1}$, giving $y_{k+1} = a y_k - bg y_{k-1}$. Substitution of a trial sequence z^k gives the characteristic polynomial $z^2 - az + bg$. For a real monic

quadratic $z^2 + c_1 z + c_0$, the unit-disc conditions are $1 + c_1 + c_0 > 0,\ 1 - c_1 + c_0 > 0,\ 1 - c_0 > 0$. For $0 < a < 1,\ b > 0,\ g > 0$ only the last condition limits increasing gain: $bg < 1$.

Once the delayed poles form a complex pair, their product equals their squared modulus. That makes the bound physically transparent: as bg reaches one the oscillation no longer decays between samples. This mechanism is invisible if a continuously implemented controller is analysed and merely evaluated at plotting intervals.

Compare two software schedules at the same sampling rate

Take $K = 1$,

$$\tau = 0.50\,\mathrm{s},$$
$$T = 0.10\,\mathrm{s}.$$

The exact coefficients are

$$a = 0.8187,\ b = 0.1813.$$

Without delay the positive gain limit is

$$(1 + a)/b = 10.03.$$

With a delayed command it is $1/b = 5.52$. A gain of 7.0 gives a stable immediate pole $0.8187 - 7(0.1813) = -0.450$. The negative sign predicts an alternating error whose sample values decrease in magnitude.

Apply the same gain through the delayed schedule. The characteristic polynomial becomes $z^2 - 0.8187z + 1.2689$. Its roots are complex with modulus $\sqrt{1.2689} = 1.126$, implying about 12.6% growth per sample in the envelope. An input limit may eventually bound the motion or produce a limit cycle, but it does not restore local linear stability around the unsaturated equilibrium. A controller test that records sensor samples and commanded values with separate timestamps can distinguish this delay from a sign error or an incorrectly discretised plant. Reduce the delayed gain to 4.0 and the predicted modulus becomes 0.852; the same clock can operate stably with a different response speed.

To recognise the delayed oscillation in a recorded trace, compute its angle as well as its modulus. At gain seven each real part is

$$a/2 = 0.4094.$$

Dividing by the modulus 1.126 gives $\cos\theta = 0.364$, so $\theta = 1.20\,\mathrm{rad/sample}$. The apparent oscillation period is about $2\pi(0.10)/1.20 = 0.52$ s. By comparison the immediate negative real pole produces a two-sample alternating pattern with period 0.20 s. A controller log showing roughly five samples per growing cycle accords with the delayed implementation, whereas alternating every sample points towards the immediate model operating too aggressively. This interpretation uses the synchronised

physical actuator output; a software variable computed today for application tomorrow may have a different timestamp from the hardware.

Why a fast vibration can appear as a slow feedback error

An analogue position sensor may detect a genuine low-frequency displacement and a much faster frame vibration. Sampling can make the two contributions indistinguishable. For a sampling interval $T_s = 1/f_s$, evaluate a complex sinusoid at sample n:

$$e^{i2\pi(f+kf_s)nT_s} = e^{i2\pi fnT_s}e^{i2\pi kn} = e^{i2\pi fnT_s}, \qquad k \in \mathbb{Z}.$$

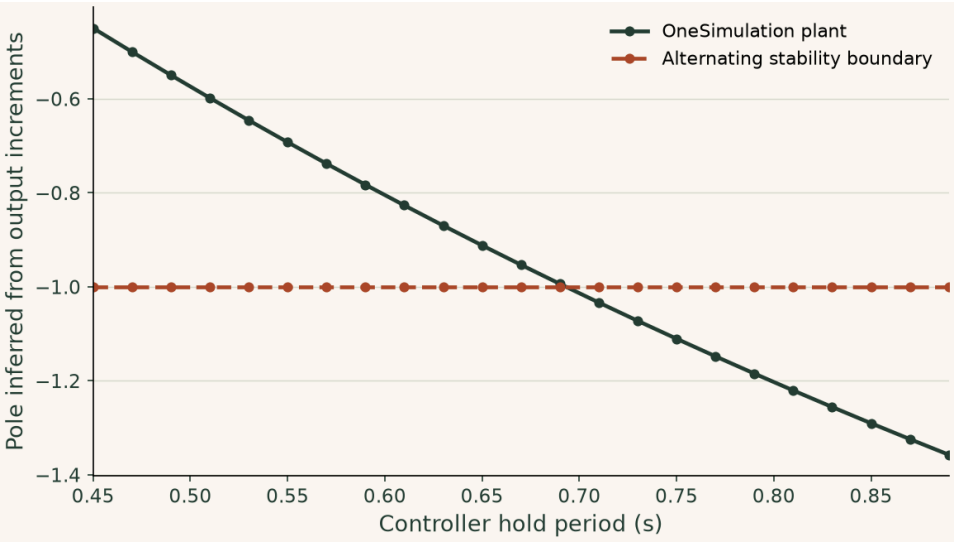
A real cosine is also unchanged by reversal of frequency, so its apparent frequency can be folded into $0 \le f_a \le f_s/2$. At $f_s = 100\,\mathrm{Hz}$, a 92 Hz cosine gives precisely the samples of an 8 Hz cosine with reversed phase angle. If both are present, their sampled amplitudes combine according to phase. Removing 8 Hz in software would remove the true 8 Hz motion as well. Filter the continuous sensor signal before sampling if the higher component must be attenuated independently.

Suppose an encoder sees an unwanted sinusoidal displacement of amplitude $0.50\,\mathrm{mm}$ at 92 Hz, while feedback must track motion through 8 Hz. A fourth-order analogue Butterworth low-pass filter with cutoff 20 Hz has magnitude $|H(f)| = [1 + (f/20\,\mathrm{Hz})^8]^{-1/2}$. At 92 Hz the ratio is 0.00223, or approximately $-53.0\,\mathrm{dB}$: the folded displacement amplitude is only $1.12\,\mu\mathrm{m}$. At 8 Hz the ratio is 0.99967, so a $0.10\,\mathrm{mm}$ desired component retains about $99.97\,\mu\mathrm{m}$. These are separate amplitude tests on the analogue filter. If both vibrations persist, the sampled 8 Hz component can vary with their relative phase.

A passband magnitude close to one does not bound the filter's phase lag. A controller designed with substantial gain near 8 Hz must also include the filter's complex frequency response in its phase-margin calculation. Aliasing, sample-and-hold delay and computational delay are three different contributions to the observed loop dynamics.

Find the stability boundary from three measured outputs

Sample the unit first-order plant three times with gain 3 and a constant reference. A one-pole response gives $\Delta y_{k+1} = z\Delta y_k$, so the ratio of increments identifies the discrete pole without knowing the final value. Here the plant uses forty RK4 substeps per controller interval.



The measured pole crosses -1 between 0.69 and 0.71 seconds. From the exact hold map $z = 4\exp(-h) - 3$, the critical period is $\ln 2$ seconds. A continuous plant integrated more accurately at the same hold period still sees the same delay in the controller update.

▶ **Read values from the plot**

Download the OneSimulation model (Python)

Interactive calculation · change the inputs of this example

Change controller gain and plant time constant, then find the sample interval where the discrete pole crosses minus one.

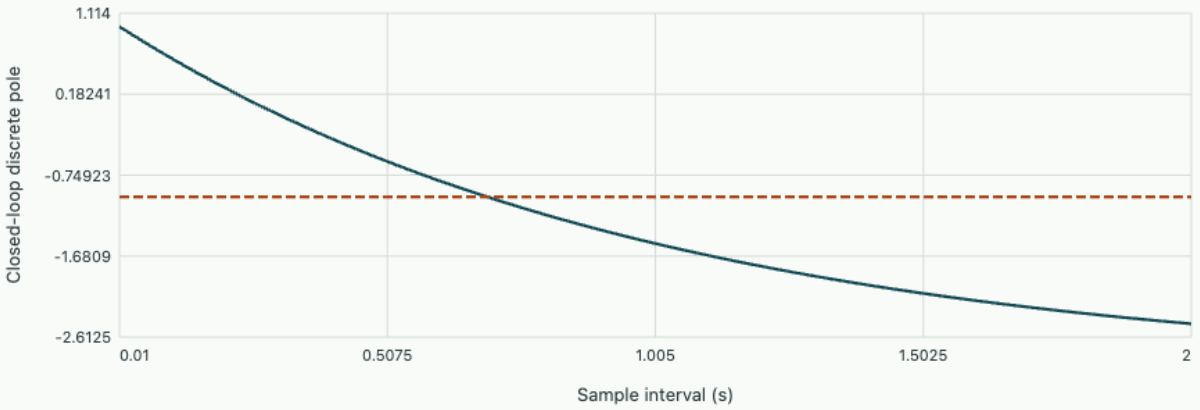
The curves below are recalculated from this section's equations in your browser. The fixed reference plot above remains available for comparison.

Open this live calculation

Try the calculation

Proportional gain **3**

Plant time constant (s) **1**



Exact zero-order hold Lower stability boundary

Stability requires sample interval below 0.693 s. More accurate plant integration does not remove the held-feedback limit.

Impulse response and finite-duration loading

Machine dynamics

Construct the response to a short pulse and determine when an impulse approximation preserves the peak motion.

FROM FIRST PRINCIPLES

Build a transient response by accumulating elementary impulses

An impulse J is the time integral of applied force. Across an ideal instantaneous impulse the displacement of mass m stays continuous, while integrating Newton's law gives a velocity jump J/m. Finite spring and viscous forces have vanishing impulse over that ideal interval. For an initially stationary undamped mount with stiffness k, the following motion is free vibration with angular frequency $\omega_n = \sqrt{k/m}$ and the new initial velocity.

$$x(0^+) = 0,$$

$$\dot{x}(0^+) = J/m,$$

$$x(t) = \frac{J}{m\omega_n}\sin(\omega_n t).$$

Now divide a finite force trace into short intervals. The interval starting at s contributes impulse $F(s)\,ds$. For a linear time-invariant system its later displacement is the impulse solution delayed by s and scaled by that impulse. Summing over all earlier intervals yields

$$x(t) = \frac{1}{m\omega_n}\int_0^t F(s)\sin[\omega_n(t-s)]\,ds.$$

An existing initial displacement or velocity contributes a separate free response by superposition. If the pulse is very short relative to the vibration period, the sine factor changes little during it and the area of the pulse controls the subsequent motion.

DERIVATION AND APPLICATIONS

Obtain the residual vibration after a ramped force is removed

The time integral of a force sets momentum change only when the support forces have little time to act. For longer pulses a linear undamped oscillator retains information about the shape and timing of the load. Consider mass m, stiffness k, natural frequency $\omega = \sqrt{k/m}$ and zero initial motion. For a force acting only from zero to T, the post-pulse displacement follows the convolution with the impulse response.

$$x(t) = \frac{1}{m\omega}\int_0^T F(s)\sin[\omega(t-s)]ds, \qquad t \geq T.$$

Expand the sine of the difference into sine and cosine of t. The two integrals multiply quadrature components of free vibration. Their resultant magnitude is the modulus of a single Fourier integral at the natural frequency. For a symmetric triangular pulse centred at $T/2$, its area is $J = F_p T/2$ and its transform factorises because the triangle is a convolution of two equal rectangles of duration $T/2$.

$$A = \frac{1}{m\omega}\left|\int_0^T F(s)e^{-i\omega s}ds\right|,$$

$$\int_0^T F(s)e^{-i\omega s}ds = Je^{-i\omega T/2}\left[\frac{\sin(\omega T/4)}{\omega T/4}\right]^2.$$

Each rectangle contributes the average complex phase over its duration; the product of these averages produces the squared factor. The impulse estimate is recovered for a very short triangle, while certain longer pulses leave no residual motion in the ideal model.

$$\frac{A}{J/(m\omega)} = \left[\frac{\sin(\omega T/4)}{\omega T/4}\right]^2,$$

$$T_{\text{first zero}} = \frac{4\pi}{\omega}.$$

WORKED EXAMPLE

Compare a triangular pulse with its impulse limit

For $m = 2.00$ kg and $\omega = 20.0$ rad/s choose $T = \pi/10$ s and $F_p = 4.00$ N. Pulse area is 0.628 N s and the impulse estimate is 15.7 mm. Here $\omega T/4 = \pi/2$ and the squared factor is

$$4/\pi^2 = 0.405.$$

Actual residual amplitude is 6.37 mm. The cancellation occurs because different increments of force reach different phases of the oscillator.

Change the inputs and inspect this calculation below

SECOND APPLICATION

Design a pulse with zero residual motion

Another lightly damped test rig is idealised as undamped at 5.00 Hz. The first zero requires a symmetric triangular pulse lasting two periods, $T = 0.400$ s. To deliver impulse 0.300 N s, choose $F_p = 2J/T = 1.50$ N. The mass can move during the pulse, but its final displacement and velocity are

both zero in the ideal model. A measured ring-down identifies sensitivity to frequency error, damping or a distorted force shape.

Interactive calculation · Change the inputs of this example

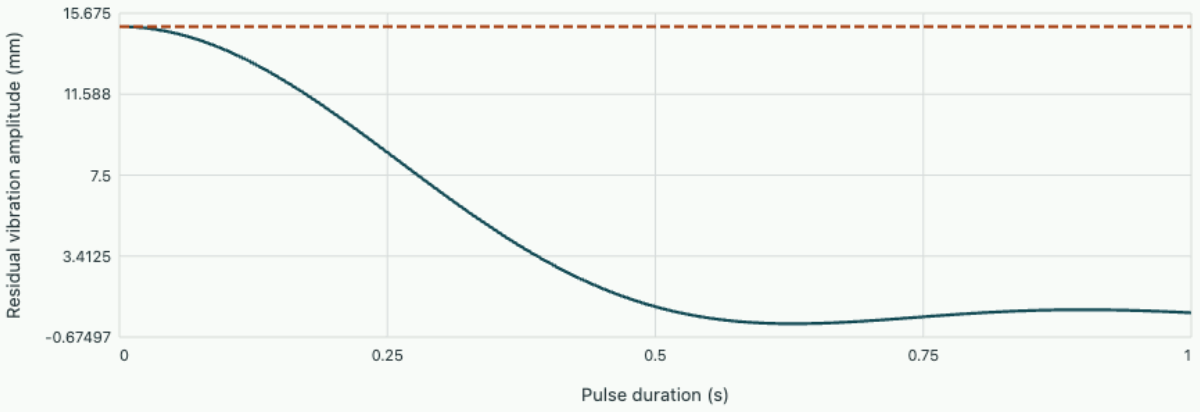
Use the adjustable quantities to test the two worked calculations above. The plot recomputes this derivation's balance or constitutive relation; its numerical readout reports the quantities needed for the forward calculation and the inverse or design question. Keep the fixed assumptions stated in the readout when comparing with the example values.

Open this live calculation

Try the calculation

Impulse (N s) **0.6**

Natural frequency (rad/s) **20**



Triangular force pulse Impulse estimate

The first cancellation duration is 0.628 s. Mass2 kg; equal pulse area does not imply equal Fourier content at the natural frequency.

A short event and a long response

A latch strikes a mounted instrument for only a few milliseconds, but its motion persists after contact has ended. Peak force alone gives an incomplete description of that motion. Duration, impulse and the system's natural time scale all matter. Define impulse $J = \int F(t)\,dt$, with units N·s. Integrating $m\ddot{x} + c\dot{x} + kx = F$ over an ideal instantaneous event leaves $m[\dot{x}(0^+) - \dot{x}(0^-)] = J$. The position remains continuous since no finite mass can jump a finite distance in zero time under a finite impulse.

For an initially resting underdamped oscillator, the subsequent response follows from the two initial conditions $x(0^+) = 0$ and $\dot{x}(0^+) = J/m$. With $\alpha = c/(2m)$ and $\omega_d = \sqrt{k/m - \alpha^2}$, the

displacement is $x = J e^{-\alpha t}\sin(\omega_d t)/(m\omega_d)$ for $t > 0$. The response to a unit impulse is the same expression with $J = 1$. Its units are displacement per unit impulse.

Build a solution from elementary impulses

Divide a general forcing record into short intervals at times s. Each interval contributes impulse $F(s)\,ds$ and creates a response delayed by s. Adding all contributions in a linear time-invariant system with zero initial state produces the convolution

$$x(t) = \int_0^t F(s)\,\frac{e^{-\alpha(t-s)}\sin[\omega_d (t-s)]}{m\omega_d}\,ds.$$

This construction makes causality explicit: later parts of the force record cannot influence the current response. Nonzero initial conditions add the free solution. In a multi-mode system each mode has its own impulse response, and the load’s location determines the force projection into it. Replacing the structure by its first mode may describe long-wave displacement well while missing a short, local acceleration peak.

Work a pulse of known duration

Neglect damping and apply constant force F_0 for $0 < t < t_p$. During the pulse integration gives

$$x = (F_0/k)[1 - \cos(\omega_n t)],$$
$$\dot{x} = F_0 \omega_n \sin(\omega_n t)/k.$$

After unloading the mass oscillates freely. Compute its amplitude from the end displacement and velocity:

$$A = \sqrt{x(t_p)^2 + \left[\dot{x}(t_p)/\omega_n\right]^2} = \frac{2F_0}{k}\left|\sin(\omega_n t_p/2)\right|.$$

Let $m = 4.0$ kg, $k = 1600$ N/m, $F_0 = 20$ N and $t_p = 0.050$ s. Then $\omega_n = 20$ rad/s and the impulse is 1.0 N·s. The formula predicts $A = 25\sin(0.50) = 12.0$ mm after unloading. Replacing the whole pulse by an instantaneous impulse gives

$$A_{\mathrm{imp}} = J/(m\omega_n) = 12.5\,\mathrm{mm}.$$

Some restoring force developed during the real pulse, so the impulse approximation is slightly high.

Find the criterion and its limitations

Divide the finite-pulse amplitude by the impulse approximation to obtain $|\sin z/z|$ with $z = \omega_n t_p/2$. Its small-z expansion is $1 - z^2/6 + \cdots$. At fixed impulse, a sufficiently short pulse therefore gives about

the same low-mode amplitude regardless of duration. A shorter event has a broader frequency content, potentially increasing the importance of higher modes.

A pulse lasting exactly one natural period gives zero residual amplitude in the ideal undamped oscillator, although the displacement during the pulse peaks at $2F_0/k$. Checking only the post-event motion would miss that maximum. In experiments record the force and response during contact as well as after it. Damping, uncertain pulse shape or nearby modes make exact cancellation unlikely, but the example shows why pulse duration must accompany a quoted peak force.

Compare pulses of equal impulse

A force transducer may report identical total impulse for a sharp tap and a broader pulse. For an undamped mode, the amplitude after the force has ended is the modulus of its frequency-weighted impulse divided by $m\omega_n$:

$$A = \left| \int F(t) e^{-i\omega_n t} dt \right| / (m\omega_n).$$

This follows by expanding the sine kernel of the convolution into two quadratures. For a rectangular pulse the ratio to the ideal instantaneous-impulse amplitude is $|\sin(\omega_n t_p/2)/(\omega_n t_p/2)|$. At a duration equal to half the natural period that factor is

$$2/\pi = 0.637.$$

Two force records with equal signed area can therefore have visibly different post-event amplitudes at one sensor if their durations differ. A short force pulse can consequently calibrate a low-frequency mode while underexciting a faster one in the same structure.

Derive an energy-consistent time update for a force record

A measured force history may be irregular enough that a closed-form convolution is inconvenient. The constant-average-acceleration Newmark update advances a linear oscillator one time step at a time. Let x_n, v_n, a_n be displacement, velocity and acceleration at t_n. Approximate the acceleration over the step through its endpoint average in the velocity update, and use the corresponding trapezoidal velocity update for displacement. With step h the relations are

$$v_{n+1} = v_n + \tfrac{h}{2}(a_n + a_{n+1}),$$

$$x_{n+1} = x_n + h v_n + \tfrac{h^2}{4}(a_n + a_{n+1}).$$

Enforce the actual equation $m a_{n+1} + c v_{n+1} + k x_{n+1} = F_{n+1}$ at the new time. Gathering unknown acceleration terms gives effective mass $m + ch/2 + kh^2/4$. The right side is $F_{n+1} - c(v_n + h a_n/2) - k(x_n + h v_n + h^2 a_n/4)$. Solve that scalar equation and back-substitute into both kinematic updates. For several coupled linear coordinates the same derivation gives a linear system with an effective matrix.

For undamped unforced systems, multiply the average endpoint equilibrium by $x_{n+1} - x_n = h(v_{n+1} + v_n)/2$. Differences of squares show that the discrete change in kinetic energy plus potential energy vanishes exactly. This property detects sign or unit errors in an implementation. It does not guarantee accurate phase at a coarse time step.

Check one step and distinguish stability from temporal resolution

Take $m = 2\,\mathrm{kg}$, $c = 0$, $k = 200\,\mathrm{N/m}$. Begin from $x_0 = 0.010\,\mathrm{m},\ v_0 = 0$ and release without applied force. Initial acceleration is

$$-kx_0/m = -1.00\,\mathrm{m/s^2}.$$

With $h = 0.10\,\mathrm{s}$, effective mass is

$$2 + 200(0.10)^2/4 = 2.50\,\mathrm{kg}.$$

Its right side is $-200[0.010 + 0.10^2(-1)/4] = -1.50\,\mathrm{N}$, giving $a_1 = -0.600\,\mathrm{m/s^2}$. The updates give

$$v_1 = -0.0800\,\mathrm{m/s}, x_1 = 0.00600\,\mathrm{m}.$$

The initial energy is

$$200(0.010)^2/2 = 0.0100\,\mathrm{J};$$

after one step it is

$$2(0.080)^2/2 + 200(0.006)^2/2 = 0.0100\,\mathrm{J}.$$

The exact displacement at 0.10 s is nevertheless $0.010\cos(1) = 0.00540\,\mathrm{m}$. The numerical oscillator advances through angle $2\arctan(\omega_n h/2) = 0.9273$ rad per step instead of 1.00 rad. Over many cycles this small period error accumulates. Halving the step is a direct temporal convergence test. For a measured force pulse, a second check is that its sampled impulse and timing converge as well; an accurate oscillator update cannot recover a force feature missed by the acquisition system.

It is possible to predict the amount of step refinement needed for a phase-sensitive comparison. For small $z = \omega_n h/2$, expand $2\arctan z = 2z - 2z^3/3 + \cdots$. Dividing the numerical angular advance by the true advance gives $1 - (\omega_n h)^2/12 + \cdots$. The leading relative frequency error therefore falls by a factor of four when the step is halved. A one-percent target suggests $\omega_n h \lesssim \sqrt{0.12} = 0.346$ for this single undamped mode. At 10 rad/s that means roughly 0.035 s or less, before checking the sampling of the force itself. This criterion addresses the fundamental frequency, not the fastest contributing mode in a multi-degree model. If a high-frequency mode contributes significantly at the measurement station, use its frequency to assess temporal phase resolution or test convergence of the complete measured signal directly.

Once these conditions are met, comparing the computed displacement and velocity at the end of loading also tests the predicted initial condition for the subsequent free ring-down.

Finite-amplitude oscillations and model validity

Machine dynamics

Use conservation of energy and a frequency approximation to quantify when a small-angle pendulum model loses timing accuracy.

FROM FIRST PRINCIPLES

Derive the pendulum period and identify the small-angle assumption

A point mass m suspended on a massless rigid link of length ℓ feels a gravitational torque $-mg\ell\sin\theta$ about the hinge. Assume a stationary pivot and negligible drag. Angular momentum balance gives an exact nonlinear equation. Replacing $\sin\theta$ by θ only for small angles in radians gives a linear oscillator and its predicted period.

$$m\ell^2\ddot{\theta} = -mg\ell\sin\theta,$$

$$T_{\text{small}} = 2\pi\sqrt{\ell/g}.$$

To track finite amplitude, release from rest at angle $A < \pi$ and use mechanical energy conservation. At any intermediate angle, loss of potential energy equals gained rotational kinetic energy.

$$\tfrac{1}{2}m\ell^2\dot{\theta}^2 = mg\ell(\cos\theta - \cos A),$$

$$T = 4\sqrt{\frac{\ell}{2g}}\int_0^A \frac{d\theta}{\sqrt{\cos\theta - \cos A}}.$$

The factor four uses the symmetry of the four quarter cycles. The mass cancels even at large amplitude. The exact restoring torque grows more slowly with deflection than its small-angle approximation, so finite-amplitude swings take longer. Measured variation of period with release angle directly tests the linearisation.

DERIVATION AND APPLICATIONS

Derive the period of an oscillator with purely cubic stiffness

A spring with almost no linear stiffness near equilibrium can have a period set primarily by release amplitude. Consider mass m on a horizontal guide with restoring force $-\beta x^3$ for $\beta > 0$. Ignore damping and assume the fitted force law holds between $-A$ and $+A$. Release from rest at $x = A$. Integrating the force law gives a quartic potential, and conservation of energy provides the speed anywhere on the inward leg.

$$\tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{4}\beta x^4 = \tfrac{1}{4}\beta A^4,$$

$$|\dot{x}| = \sqrt{\frac{\beta}{2m}}\sqrt{A^4 - x^4}.$$

By symmetry the time from an extreme position to the centre is one quarter of a full cycle. Write $dt = |dx|/|\dot{x}|$ and substitute $u = x/A$ in the integral. This isolates the amplitude dependence without guessing a sinusoidal shape.

$$T = \frac{4}{A}\sqrt{\frac{2m}{\beta}}\int_0^1 \frac{du}{\sqrt{1-u^4}}.$$

The dimensionless integral evaluates numerically to 1.31103. Its endpoint singularity is integrable because the speed near the turning point vanishes as the square root of the distance remaining. The resulting period and angular frequency are

$$T = \frac{7.4163}{A}\sqrt{\frac{m}{\beta}},$$

$$\omega = \frac{2\pi}{T} = 0.8472\,A\sqrt{\frac{\beta}{m}}.$$

The positive cubic restores the mass for displacements of either sign, but its tangent stiffness at zero is zero. The full nonlinear dynamics therefore predicts a period diverging with decreasing amplitude. Even a weak residual linear spring would eventually dominate in that limit.

WORKED EXAMPLE

Predict the timing of two releases

Take $m = 0.500$ kg, $\beta = 2.00 \times 10^5$ N/m³ and $A = 20.0$ mm. The predicted period is

$$7.4163\sqrt{0.500/(2.00\times10^5)}/0.0200 = 0.586\,\mathrm{s}.$$

Reducing the release to 10.0 mm doubles the predicted period to 1.17 s. At the larger amplitude the centre-crossing speed is

$$A^2\sqrt{\beta/(2m)} = 0.179\,\mathrm{m/s}.$$

Change the inputs and inspect this calculation below

SECOND APPLICATION

Infer stiffness from a measured period

A second fixture of mass 0.800 kg completes a cycle in 0.750 s after release at 15.0 mm. The implied coefficient is $\beta = m[7.4163/(AT)]^2 = 3.48 \times 10^5$ N/m³. Its predicted restoring-force magnitude at release is

$$\beta A^3 = 1.17\,\mathrm{N}.$$

An independent static force measurement there tests the dynamic estimate without assuming a sinusoidal motion.

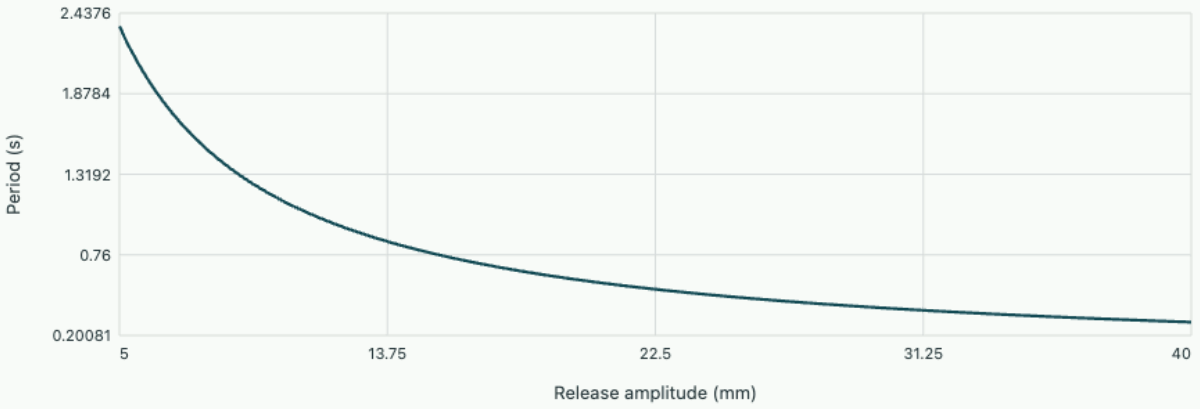
Interactive calculation · Change the inputs of this example

Use the adjustable quantities to test the two worked calculations above. The plot recomputes this derivation's balance or constitutive relation; its numerical readout reports the quantities needed for the forward calculation and the inverse or design question. Keep the fixed assumptions stated in the readout when comparing with the example values.

Open this live calculation

Try the calculation

Mass (kg)	0.5
Cubic force coefficient (kN/m³)	200



Energy-integral period

At20 mm amplitude, period=0.586 s. This pure cubic spring has no nonzero small-amplitude linear frequency.

Retain the actual restoring moment

A pendulum gives a practical test of a linear model because its length and release angle are easy to set while its period can be measured over many cycles. Approximate the moving body first by a point mass m at distance l from a frictionless pivot. With θ measured from vertically downwards, gravity produces torque $-mgl\sin\theta$. Dividing the rotational balance by ml^2 leaves $\ddot\theta + (g/l)\sin\theta = 0$. The linear small-angle frequency $\omega_0 = \sqrt{g/l}$ follows only after replacing $\sin\theta$ by θ in radians.

The leading correction has a definite sign:

$$\sin\theta \simeq \theta - \theta^3/6.$$

At a large positive angle the restoring force magnitude is smaller than the linear extrapolation. The pendulum should take longer to complete its swing. A sign test like this is useful before calculating a period correction.

Derive an integral valid at finite amplitude

For release from rest at amplitude $A < \pi$, conservation of mechanical energy gives

$$\tfrac{1}{2}ml^2\dot\theta^2 + mgl(1-\cos\theta) = mgl(1-\cos A).$$

During one quarter of the orbit the angle travels between zero and A. Substitute $\sin(\theta/2) = \sin(A/2)\sin\psi$ into the time integral. Cancelling common square roots gives

$$T = 4\sqrt{l/g}\int_0^{\pi/2} \frac{d\psi}{\sqrt{1 - k^2\sin^2\psi}},$$

$$k = \sin(A/2).$$

For modest amplitudes expand the integrand as $1 + \tfrac{1}{2}k^2\sin^2\psi + \cdots$. The average of $\sin^2\psi$ over the quarter interval is one half. Therefore $T/T_0 \simeq 1 + k^2/4 \simeq 1 + A^2/16$ through second order in amplitude. The oscillation period depends on energy, while in the linear spring model all amplitudes share the same period.

Estimate a measurable timing error

Take $l = 0.80\,\mathrm{m}$, $g = 9.81\,\mathrm{m/s^2}$ and release at $A = 30^\circ = 0.5236\,\mathrm{rad}$. The small-angle period is

$$T_0 = 2\pi\sqrt{0.80/9.81} = 1.794\,\mathrm{s}.$$

The leading correction is 1.71%, giving about 1.825 s. Over ten full oscillations the predicted delay relative to the small-angle model is roughly 0.31 s. Reducing the release angle to 15° reduces the

leading relative correction by a factor of four.

A timing series from a damped experimental pendulum may show the period drifting towards T_0 as its amplitude decays. Measuring only the first release angle while averaging many later periods mixes different amplitudes. To test the finite-amplitude law, use consecutive same-direction crossings to measure each local period and estimate its associated amplitude from neighbouring turning points. The correction above is an expansion around zero amplitude. As A approaches the inverted equilibrium, the exact integral grows without bound and a low-order polynomial is no longer an adequate approximation.

Check the scale of the neglected term

Retaining one further term in the integrand gives $1 + \frac{1}{2}k^2 \sin^2\psi + \frac{3}{8}k^4 \sin^4\psi$. The quarter-interval average of the fourth power of sine is $3/8$, so $T/T_0 \simeq 1 + k^2/4 + 9k^4/64$. At 30° this gives 1.01738, while the leading expansion in amplitude gave 1.01713. The small difference is measurable only if the uncertainty in pendulum length, release angle and crossing times supports the comparison.

Derive an amplitude-dependent resonance relation

A pendulum softens as its amplitude grows because the sine restoring law rises more slowly than its linear tangent. A mechanical flexure can instead harden through geometry. A simple modal model is $m\ddot{x} + c\dot{x} + kx + \beta x^3 = F\cos\omega t$. For positive β the restoring stiffness increases with excursion. Assume one dominant harmonic and take $x \approx A\cos(\omega t - \phi)$. Expanding the cubic gives

$$x^3 = A^3[3\cos(\omega t - \phi) + \cos(3\omega t - 3\phi)]/4.$$

Keeping only the fundamental turns the cubic term into an amplitude-dependent in-phase force.

Balance in-phase and quadrature components against the input, square and add to eliminate ϕ. The first-harmonic amplitude estimate satisfies

$$F^2 = A^2\left[\left(k + \tfrac{3}{4}\beta A^2 - m\omega^2\right)^2 + (c\omega)^2\right].$$

In a lightly damped free motion the same approximation gives a backbone frequency $\omega_b^2 \approx [k + (3/4)\beta A^2]/m$. This is an amplitude-dependent estimate of conservative oscillation frequency, not a new eigenvalue of the linearised equilibrium. At very small amplitude it recovers $\sqrt{k/m}$, while a negative cubic coefficient tilts the relation towards lower frequencies. If the cubic approximation is only a local expansion, do not extrapolate it until it predicts an unphysical loss of global stiffness.

Identify a hardening spring from two release amplitudes

A test component has moving mass 1.0 kg and small-amplitude stiffness 400 N/m. Low-level ring-down gives $\omega_0 = 20\,\mathrm{rad/s}$. At measured oscillation amplitude 20 mm the dominant free frequency is approximately 22 rad/s. Under the weakly nonlinear one-harmonic model, $\beta \simeq 4m(22^2 - 20^2)/[3(0.020)^2] = 2.80 \times 10^5\,\mathrm{N/m^3}$. At 10 mm the inferred backbone is $\sqrt{400 + 0.75(2.8 \times 10^5)(0.010)^2} = 20.52\,\mathrm{rad/s}$, a direct third reading against which to test the fit.

Now drive near 22 rad/s with viscous coefficient $c = 0.80\,\mathrm{N\,s/m}$. At the 20 mm backbone amplitude the spring and inertia fundamentals cancel, so maintaining this motion takes forcing $F = c\omega A = 0.352\,\mathrm{N}$ in quadrature with displacement. The neglected cubic third-harmonic force amplitude is

$$\beta A^3/4 = 0.56\,\mathrm{N};$$

its displacement effect can nevertheless be small because the oscillator is driven far from its fundamental resonance at 66 rad/s. A separate mode near that frequency would invalidate this simplification. In a forced frequency sweep the amplitude equation can admit several positive roots near resonance. Its algebra does not establish their stability; slow upward and downward sweeps, each with sufficient settling time, can reveal different stable responses and jump transitions.

An initial static measurement tests the sign of the fitted coefficient independently. At a displacement of 20 mm the linear restoring force is

$$400(0.020) = 8.0\,\mathrm{N};$$

the cubic contribution is

$$2.8 \times 10^5 (0.020)^3 = 2.24\,\mathrm{N}.$$

The predicted static force is therefore 10.24 N, and the tangent stiffness is

$$400 + 3(2.8 \times 10^5)(0.020)^2 = 736\,\mathrm{N/m}.$$

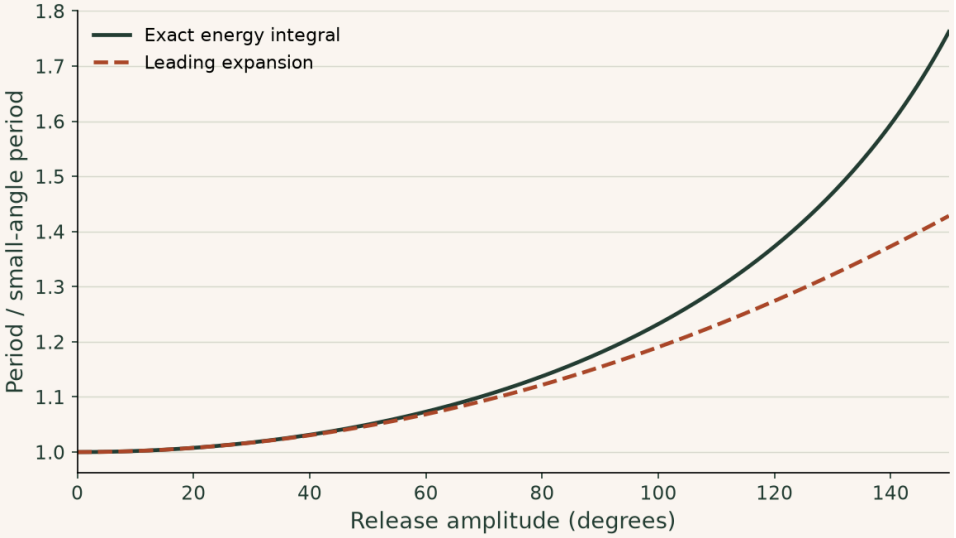
This tangent stiffness applies to small vibrations around a held biased equilibrium. Using it directly for an oscillation centred at zero would overestimate the measured amplitude-dependent frequency because the instantaneous stiffness changes during each cycle. Static bias and free oscillation amplitude are physically different tests of the same nonlinear restoring law.

Repeating a decay test around a deliberately shifted equilibrium distinguishes the two experimentally.

Set an amplitude limit for a pendulum-based timing model

For a rigid pendulum released from rest the exact period ratio is $2K(\sin^2(\theta_0/2))/\pi$, where the complete elliptic integral here uses its parameter rather than its modulus. Compare it with the first

amplitude correction $1 + \theta_0^2/16$, with angle in radians.



At 90° the exact period is about 1.180 times the linear prediction; the leading expansion gives 1.154. The series is useful at modest amplitudes, but it cannot reproduce the divergence on approach to an upright release.

▶ **Read values from the plot**

Interactive calculation · change the inputs of this example

Increase the maximum release angle and compare the energy integral with the first series correction. Changing length rescales both periods.

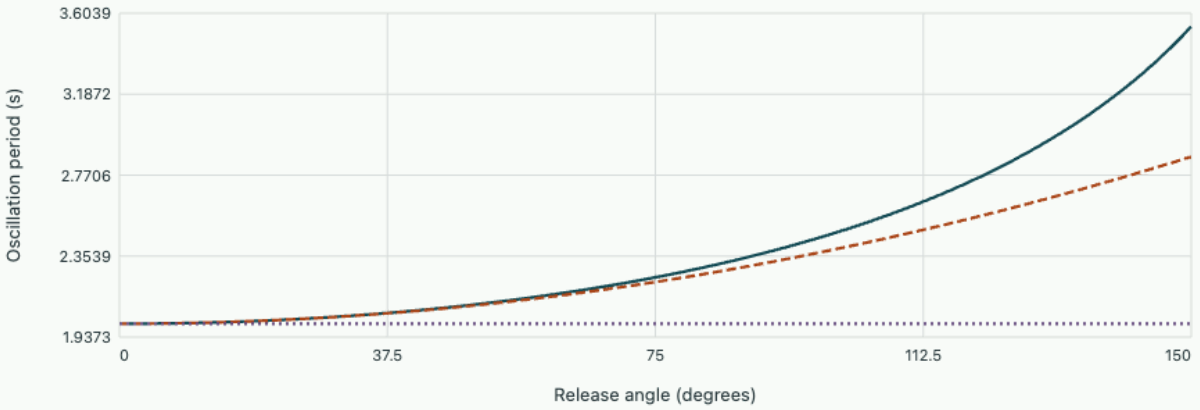
The curves below are recalculated from this section's equations in your browser. The fixed reference plot above remains available for comparison.

Open this live calculation

Try the calculation

Maximum release angle (degrees) **150**

Pendulum length (m) **1**



Energy integral via AGM · First amplitude correction · Small-angle limit

At 150° the exact period is 3.535 s, 76.22% above the linear estimate.

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

$$x = 2\,\mathrm{erfc}^{-1}(1/2)\sqrt{Dt} = 0.9539\sqrt{Dt}.$$

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.

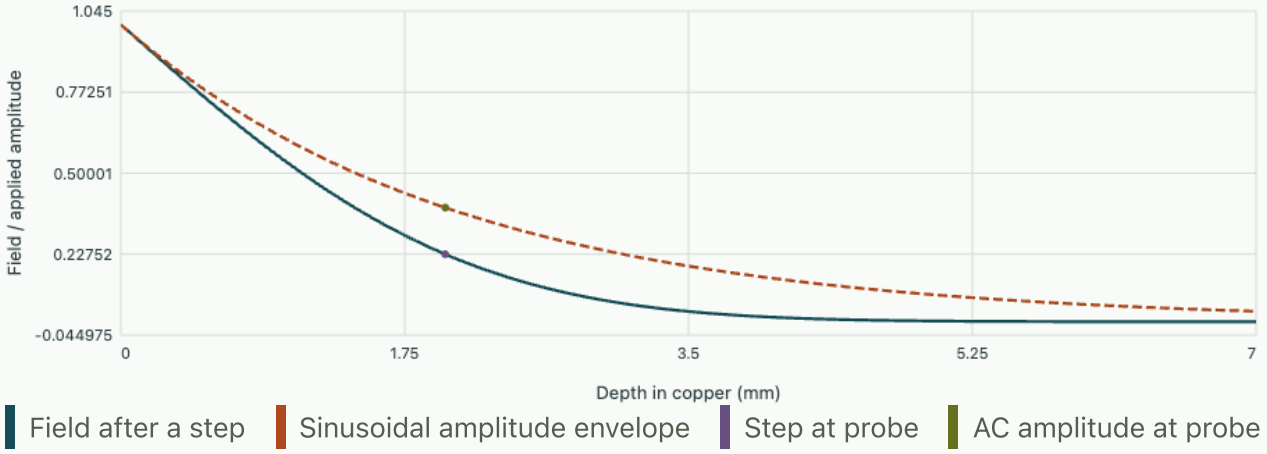
Open this experiment online

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 $J_y = B_s[\mu\sqrt{\pi Dt}]^{-1} \exp[-x^2/(4Dt)]$. 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

$$D = 4a^2/(\pi^2 \tau_0) = 0.00608\,\mathrm{m^2/s}.$$

Then

$$\sigma = 1/(\mu D) = 6.55 \times 10^7\,\mathrm{S/m}.$$

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 $\operatorname{erfc}[x/(2\sqrt{Dt})] - \operatorname{erfc}[x/(2\sqrt{D(t-T)})]$. 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

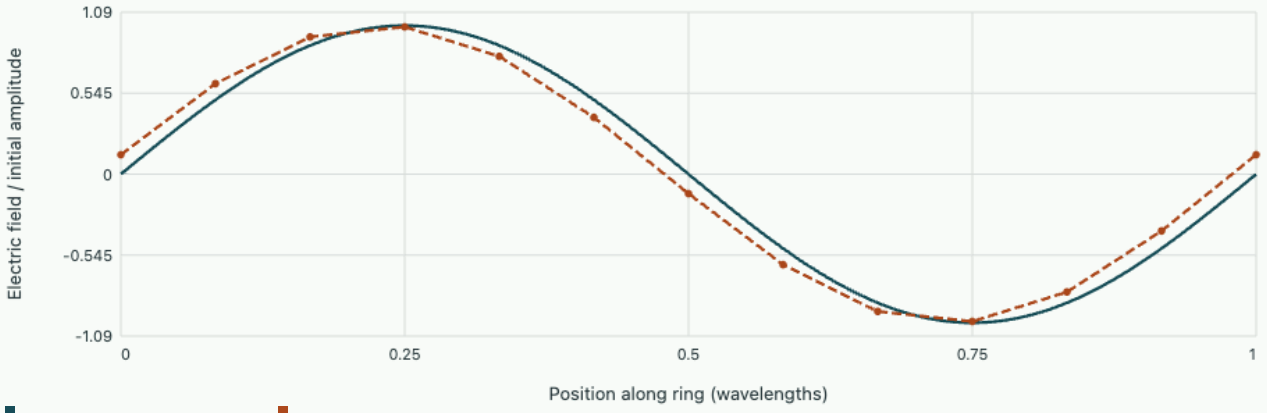
Keep the equations and initial mode fixed while refining the grid. Examine numerical phase delay alongside the physical propagation law.

Open this experiment online

Try the calculation

Points per wavelength **12**

Courant number **0.8**



Continuous wave Time-stepped nodes

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

$$F'' + c^2k^2F = 0.$$

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.

A moving-coil galvanometer as an electrical and mechanical system

Electromechanics, optics and measurement

A scale division records an angle, yet the quantity marked on the dial is electric current. A spring, a magnetic field and the resistance of the external measuring circuit connect the two. Derive the calibration, predict how a pointer approaches its reading, and find which experiment measures the charge of a short pulse.

Convert the force on two straight sides into a torque

A light rectangular coil has N turns, active length l, width w, and area per turn $A = lw$. It pivots about an axis parallel to its long sides. Shaped pole pieces confine an approximately radial magnetic field of magnitude B to an air gap. Over the working angular range the magnetic field is perpendicular to the active wire and lies approximately in the plane of the coil. An elementary length carrying conventional current i experiences $d\mathbf{F} = i\,d\mathbf{l} \times \mathbf{B}$. The two long sides carry current in opposite directions and experience forces of equal magnitude $NilB$. Acting with lever arms $w/2$, the forces make a couple whose torque is $NiAB$. Their net force is approximately zero. The short ends supply negligible torque about the pivot under the ideal geometry.

Write $K = NAB$, and choose the positive sense of current to give positive torque. Its units are $\mathrm{N\,m/A}$. In a uniform Cartesian field an ordinary rigid loop instead gives torque $NiAB\sin\alpha$, where α is the angle between its normal and the field. The radial gap maintains the factor equivalent to $\sin\alpha$ close to unity as the pointer turns. This is a design choice that produces a nearly linear scale. If a measured calibration curve flattens at a large angle, field nonuniformity and a nonlinear torsion spring are candidate causes that can be tested separately.

For a soft torsion spring with stiffness κ, the return torque is $-\kappa\theta$ at small deflections from the zero-current setting. At static equilibrium $\theta = Ki/\kappa$. With $N = 200$, $A = 1.00\,\mathrm{cm^2}$, $B = 0.250\,\mathrm{T}$ and $\kappa = 8.0 \times 10^{-4}\,\mathrm{N\,m/rad}$, the torque constant is $0.0050\,\mathrm{N\,m/A}$ and the current sensitivity is $6.25\,\mathrm{rad/A}$. A $5.0\,\mathrm{mA}$ current turns the pointer through $0.03125\,\mathrm{rad}$, about 1.79°. A mirror mounted on the same axle deflects a reflected laser through twice this angle. At a screen distance of $0.50\,\mathrm{m}$ the spot shifts $0.50\tan(2\theta) = 31.3\,\mathrm{mm}$, making the small mechanical rotation easy to resolve.

Recover the reciprocal voltage from work

Now rotate the coil through a small positive angle while it carries positive current. The electromagnetic torque does mechanical work $Ki\,d\theta$. The voltage associated with the motion must therefore draw electrical work $ie_{\mathrm{back}}\,dt$ of equal magnitude when the motor torque assists that motion. This gives the back-emf relation $e_{\mathrm{back}} = K\dot\theta$, measured as a positive drop in the motor sign convention. Its coefficient has units $\mathrm{V\,s/rad}$, numerically the same SI coefficient as the torque constant. One can also derive it from the motional emf on the two active sides. Each moves at speed $(w/2)\dot\theta$, crosses field B, and generates a voltage $Bl(w/2)\dot\theta$ along its wire. The two contributions add over every turn.

Let J denote the rotational inertia, b the mechanical viscous damping coefficient, L the coil inductance and R the sum of winding plus external series resistance seen by a voltage source $v(t)$. The electromechanical equations are

$$J\ddot{\theta} + b\dot{\theta} + \kappa\theta = Ki,$$

$$L\dot{i} + Ri + K\dot{\theta} = v(t).$$

Multiplying the first equation by $\dot{\theta}$ and the second by i eliminates the coupling power $Ki\dot{\theta}$. The result says that source power vi equals the rate of increase of $J\dot{\theta}^2/2 + \kappa\theta^2/2 + Li^2/2$, plus $b\dot{\theta}^2 + Ri^2$. This identity fixes the relative signs of the torque and back emf. If a sign mistake turns a passive closed coil into a self-accelerating pointer, the same mistake produces a residual coupling term in this power balance. Mechanical friction and winding heat represent distinct sinks even though both can damp the observed displacement.

Predict overshoot and find a critical external resistance

If the electrical relaxation time L/R is short compared with the oscillation time $1/\sqrt{\kappa/J}$, set $i \simeq (v - K\dot{\theta})/R$. Substituting into the mechanical equation increases its effective damping to $b_{\mathrm{eff}} = b + K^2/R$. Set $v = 0$ when discussing free relaxation with the electrical terminals connected through a resistor. A moving shorted coil generates a current that opposes its motion. Opening the circuit instead suppresses that current and leaves the mechanical damping. An ideal current source holding specified current through all pointer motions also changes the back-emf response because its output voltage adjusts to enforce that current.

The undamped angular frequency is $\omega_0 = \sqrt{\kappa/J}$ and the damping ratio is

$$\zeta = b_{\mathrm{eff}}/(2\sqrt{\kappa J}).$$

Take $J = 2.0 \times 10^{-6}\,\mathrm{kg\,m^2}$, $b = 1.6 \times 10^{-5}\,\mathrm{N\,m\,s/rad}$, and the preceding K, κ. Then $\omega_0 = 20.0\,\mathrm{rad/s}$, $2\sqrt{\kappa J} = 8.0 \times 10^{-5}\,\mathrm{N\,m\,s/rad}$, and an open circuit gives $\zeta = 0.20$. With $R = 1.00\,\Omega$ the electric contribution to damping is $2.50 \times 10^{-5}\,\mathrm{N\,m\,s/rad}$, giving $\zeta = 0.5125$. For a voltage step the ideal fractional overshoot in the reduced model is

$$\exp[-\pi\zeta/\sqrt{1-\zeta^2}] = 0.153.$$

The first peak occurs after $\pi/(\omega_0\sqrt{1-\zeta^2}) = 0.183\,\mathrm{s}$.

Critical damping occurs when $b + K^2/R = 2\sqrt{\kappa J}$, which requires a total resistance of about $0.391\,\Omega$. A winding resistance of $0.300\,\Omega$ therefore leaves $0.091\,\Omega$ for external source and leads. If the winding alone exceeds the target resistance, a simple passive external series resistor cannot reach this damping in the reduced model. A finite L must also be checked: with $L = 5.0\,\mathrm{mH}$ and $R = 1.0\,\Omega$, $\omega_0 L/R = 0.10$, so the approximation provides a useful starting estimate. Reducing R indefinitely eventually invalidates the elimination of current. The full pair of equations then predicts the damping and frequency dependence.

Distinguish a charge impulse from a slow current reading

Suppose a brief imposed current pulse ends long before the pointer has turned appreciably. Integrating the torque equation over that interval gives

$$J\Delta\dot{\theta} \simeq K\int i\,dt = Kq.$$

The initial signal is now angular velocity proportional to total transmitted charge. For weak subsequent damping, the first excursion is approximately $Kq/(J\omega_0)$. A dose of $20\,\mu\mathrm{C}$ gives $2.50\,\mathrm{mrad}$ with the instrument above, corresponding to about $2.50\,\mathrm{mm}$ on the mirror scale. Two current pulses of identical charge but different shapes give nearly the same reading if both are short on the mechanical time scale. Their peak currents can be entirely different.

For appreciable damping, use the impulse solution $\theta(t) = Kq\,e^{-\zeta\omega_0 t}\sin(\omega_d t)/(J\omega_d)$, with $\omega_d = \omega_0\sqrt{1-\zeta^2}$ for an underdamped pointer. Its maximum obeys $\tan(\omega_d t_p) = \omega_d/(\zeta\omega_0)$ in the first quarter cycle. Substitution corrects the ballistic calibration instead of treating the damped peak as an undamped amplitude. The state of the external circuit after the pulse matters because it determines ζ during the throw. A calibration pulse should leave the measuring circuit connected in the same way as the unknown pulse.

Measure the torque constant without assuming the magnetic field

The preceding design calculation multiplies turn count, active area and a nominal air-gap induction. An independent calibration can measure the same constant without knowing the field distribution along every wire. Move the pointer through a small controlled angle while measuring its open-circuit voltage with a high-impedance instrument. In the same working interval the voltage is $K(\theta)\dot{\theta}$ up to a polarity choice. Integrating the voltage over the sweep gives $\int_{\theta_1}^{\theta_2} K(\theta)\,d\theta$. If K is constant this is simply $K\Delta\theta$, independent of sweep speed. It is often easier to integrate a recorded voltage pulse accurately than to differentiate the angle trace to obtain instantaneous speed.

For a sweep through $0.0800\,\mathrm{rad}$, suppose the measured voltage integral is $0.401\,\mathrm{mV\,s}$. The generator constant is $5.01\times10^{-3}\,\mathrm{V\,s/rad}$. Independently add a known small torque to the spring and measure its static angle. A torque $32.0\,\mu\mathrm{N\,m}$ rotates the pointer by $0.0400\,\mathrm{rad}$, giving $\kappa = 8.00\times10^{-4}\,\mathrm{N\,m/rad}$. Combining the two measurements predicts a current sensitivity of $6.27\,\mathrm{rad/A}$. Splitting the sweep into smaller equal angular intervals tests whether the measured K depends on angle. Sweeping backwards should reverse the integrated voltage. A voltage offset instead contributes its offset times duration, so repeating at two sweep speeds helps separate it.

A dynamic cross-check uses free decay under two load conditions. Measure the exponential decay rate α_{open} with a high-impedance voltage probe and α_R with a known total resistance in the closed circuit. In the short-electrical-time limit, subtraction of the two damping laws yields

$$K^2 = 2JR(\alpha_R - \alpha_{\mathrm{open}}).$$

With $J = 2.0 \times 10^{-6}\,\mathrm{kg\,m^2}$, $R = 1.00\,\Omega$, and decay rates 4.0 and $10.25\,\mathrm{s^{-1}}$, this gives

$$K = 0.0050\,\mathrm{N\,m/A}.$$

The method needs two identifiable decay rates and independent inertia. If the loaded response is overdamped, the envelope of two oscillation peaks is no longer available; fit the full two-real-pole response to the measured displacement.

Keep the third dynamical variable in a frequency-response experiment

For a sinusoidal voltage at angular frequency ω the exact linear transduction law follows by eliminating the complex current after transforming both differential equations. The mechanical impedance in angular-displacement form is $\kappa - J\omega^2 + i\omega b$; the winding impedance is $R + i\omega L$. Using phasors with time dependence $e^{i\omega t}$ gives

$$\frac{\Theta}{V} = \frac{K}{(R + i\omega L)(\kappa - J\omega^2 + i\omega b) + i\omega K^2}.$$

At zero frequency the gain is $K/(R\kappa)$ and the phase tends to zero, as the static calibration requires. At a sufficiently high frequency both the electrical and inertial impedances grow. For $L > 0$ the magnitude approaches $K/(LJ\omega^3)$. The reduced model with the current eliminated algebraically falls only as $1/\omega^2$. The change of the asymptotic slope provides an experimental way to distinguish finite inductance from an incorrectly adjusted viscous damping constant. Beyond the working band of a real instrument, elastic modes of the pointer or parasitic capacitance can enter before that asymptote is resolved.

At the nominal undamped frequency $20\,\mathrm{rad/s}$, with $R = 1.0\,\Omega$ and $L = 5.0\,\mathrm{mH}$, the stiffness and inertia terms cancel. The denominator becomes $-\omega^2 L b + i\omega(Rb + K^2)$, or approximately $-0.000032 + i\,0.000820$ in consistent SI units. Its magnitude gives about $6.09\,\mathrm{rad/V}$. A sinusoid of $1.0\,\mathrm{mV}$ peak produces about $6.09\,\mathrm{mrad}$ peak, whereas the static response to $1.0\,\mathrm{mV}$ is $6.25\,\mathrm{mrad}$. The phases differ radically despite their similar amplitudes: near this frequency the pointer lags by about 92.2°. A gain-only calibration at a single frequency would miss the phase lag.

Change the electrical load and compare two step models

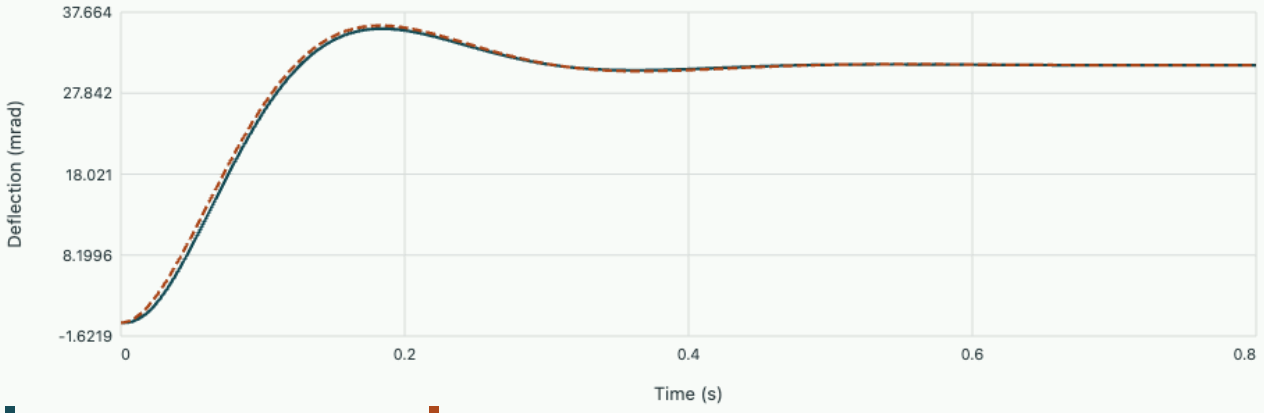
In the live graph the source voltage is chosen anew at each resistance to give a final current of $5.0\,\mathrm{mA}$ after the pointer settles. That convention keeps the final angle common to all curves and isolates the transient response of the load. The full model retains $i, \theta, \dot\theta$; the second curve sets $L\dot i = 0$. Predict first which trace should start flatter. With finite inductance the current initially vanishes and increases at a finite rate. The initial torque and angular acceleration therefore vanish, while the reduced model applies a finite torque immediately. Increasing the inductance makes that distinction easier to see, especially when the electrical response time becomes comparable to the period of mechanical oscillation.

Open the live experiment

Try the calculation

Total resistance (Ω) 1

Winding inductance (mH) 5



Coupled winding and pointer Short electrical time approximation

The reduced damping ratio is 0.512. The largest deflection in the first 0.80 s is 35.67 mrad. The common static limit is 31.25 mrad.

How could a student change the static range of the instrument without changing its suspension or its torsional stiffness?

▼ **Work through the solution**

For a dc ammeter, connect a calibrated shunt resistance in parallel with the coil, allowing only a known fraction of total current into it. With a coil winding resistance $0.300\,\Omega$ and a $0.0333\,\Omega$ shunt, the coil carries about ten percent of a steady terminal current and the indicated full-scale terminal current increases about tenfold. The total circuit seen by a motional voltage during a pointer oscillation depends on the shunt and source impedance, so the static range multiplier does not by itself give the new damping. For a voltmeter, a series resistance instead sets the current produced by an applied terminal voltage. In both cases record the resistance network whenever reporting a dynamic response.

Estimate the back action on the signal being measured

A galvanometer becomes part of the circuit it observes. Suppose a source with open-circuit voltage $10.0\,\mathrm{mV}$ and internal resistance $2.00\,\Omega$ drives a coil of resistance $0.300\,\Omega$. After settling, the current is $4.35\,\mathrm{mA}$ and the pointer turns by $27.2\,\mathrm{mrad}$. The undisturbed short-circuit current of the source would have been $5.00\,\mathrm{mA}$. Interpreting the indicated value as the current that would flow with a perfect zero-resistance meter underestimates that quantity by about thirteen percent. Measuring the open-circuit voltage instead requires a sufficiently large input resistance, another electrical boundary condition.

During pointer motion the disturbance includes the back emf as well as winding resistance. If the same coil is briefly turning at $0.40\,\mathrm{rad/s}$ it develops a $2.0\,\mathrm{mV}$ back emf. Neglecting inductance for that estimate, the resulting current is

$$(10.0 - 2.0)\,\mathrm{mV}/2.30\,\Omega = 3.48\,\mathrm{mA}.$$

The deflection therefore contains information about both the source and the pointer's earlier motion. A static calibration alone cannot recover a rapid source waveform. A known source resistance and the two coupled dynamic equations provide a way to estimate the waveform from synchronised angle and current measurements.

Tutorial on calibration and transient identification

A current source produces a static mirror shift of $12.5\,\mathrm{mm}$ on a screen $0.50\,\mathrm{m}$ away. Estimate the current and identify a reversible test for an angular zero offset.

▼ **Work through the solution**

Use $\theta = \tfrac{1}{2}\arctan(0.0125/0.50) = 0.01250\,\mathrm{rad}$ and $i = \kappa\theta/K = 2.00\,\mathrm{mA}$. Repeat with reversed current of equal magnitude. Half the difference of the two angles cancels a fixed angular zero offset and isolates the response odd in current. A changing temperature during reversal can alter the spring or winding resistance, so alternate the order and record the zero afterwards if the required uncertainty is small.

For an open-circuit free decay the ratio of two successive same-side positive peak angles is 3.60. Their time separation is $0.321\,\mathrm{s}$. Recover the mechanical parameters that can be inferred if J is known.

▼ **Work through the solution**

The logarithmic decrement is $\ln 3.60 = 1.281$. The decay rate is $\alpha = 1.281/0.321 = 3.99\,\mathrm{s}^{-1}$, and the damped angular frequency is

$$2\pi/0.321 = 19.57\,\mathrm{rad/s}.$$

Then $\omega_0 = \sqrt{\omega_d^2 + \alpha^2} = 19.97\,\mathrm{rad/s}$,

$$b = 2J\alpha = 1.60 \times 10^{-5}\,\mathrm{N\,m\,s/rad},$$

$$\kappa = J\omega_0^2 = 7.98 \times 10^{-4}\,\mathrm{N\,m/rad}.$$

This identification assumes viscous damping and a sufficiently high impedance to make the electrical damping negligible.

Explore how motion generates voltage in the induction chapter.

Eddy-current braking, magnetic penetration and the heat left behind

Electromechanics, optics and measurement

A conducting loop moves above a row of magnetic poles. The drag first increases with speed, but a rapid crossing need not generate proportionally greater braking. Trace the work done against the Lorentz force into circuit energy, heat and a measurable temperature change. Then decide which features of a solid plate the single-loop model can capture.

Generate a voltage from a spatially periodic field

Let a rectangular loop of height h along y and width w along x translate to the right above a stationary magnetic pattern. At the loop let the normal induction be $B_z(x) = B_0 \cos(kx)$, with $k = 2\pi/\lambda$. Its two active sides are centred symmetrically about the coordinate x_c. Assume the field varies negligibly along y over this height. Integrating through the loop gives

$$\Phi = h \int_{x_c - w/2}^{x_c + w/2} B_0 \cos(kx)\, dx = [2hB_0 \sin(kw/2)/k] \cos(kx_c).$$

A uniform translation $x_c = ut$ therefore produces an external emf with peak amplitude $E_{\mathrm{pk}} = Ku$, where $K = 2hB_0 |\sin(kw/2)|$.

This spatial averaging explains an initially surprising design feature. A loop spanning precisely one full spatial period always links zero net flux from the sinusoidal field and therefore generates no emf in this ideal geometry. A loop of half a period has maximum coupling at a specified height. If w is much smaller than the period,

$$\sin(kw/2) \simeq kw/2,$$

$$K \simeq hwB_0 k.$$

This is exactly the small-loop result from using its area times the local gradient of induction. Reversing velocity reverses the emf at a fixed position. At a fixed positive speed the temporal frequency is $f = u/\lambda$, so the same conductor experiences different electrical frequencies at different mechanical speeds.

Include finite inductance and derive the drag curve

Model the current path by a constant resistance R and self-inductance L. Its steady periodic current solves $L\dot{i} + Ri = \mathcal{E}(t)$. In phasor notation its magnitude is $E_{\mathrm{pk}}/\sqrt{R^2 + (\omega L)^2}$, lagging the emf by $\arctan(\omega L/R)$, where $\omega = ku$. The current develops field energy $Li^2/2$ and dissipates Ri^2. Multiplying the circuit equation by i and averaging over a period eliminates the derivative of stored energy. The average mechanical power delivered against the drag is therefore the mean Joule heating $P = E_{\mathrm{pk}}^2 R/[2(R^2 + \omega^2 L^2)]$. Dividing by the speed gives the drag magnitude

$$F(u) = \frac{K^2 u}{2R[1 + (u/u_c)^2]},$$

$$u_c = \frac{R}{kL}.$$

At low speed the drag is proportional to velocity. At the crossover $u = u_c$ it reaches $K^2 u_c/(4R)$, and at much larger speeds it falls approximately as $1/u$ within the assumed lumped model. The incoming mechanical power then tends to a constant rather than growing without bound. Why is there a maximum? Increasing speed raises the emf, but it also raises the reversal frequency. An inductive current increasingly lags the driving emf, and its component in phase with the emf does the net dissipative work. A measured turnover therefore tests both an electrical time scale and the spatial scale of the magnet pattern.

Use $h = 50\,\mathrm{mm}$, $w = 25\,\mathrm{mm}$, $\lambda = 0.100\,\mathrm{m}$, $B_0 = 0.20\,\mathrm{T}$, $R = 0.050\,\Omega$ and $L = 0.150\,\mathrm{mH}$. Then $K = 0.01414\,\mathrm{T\,m}$, $u_c = 5.31\,\mathrm{m/s}$, and maximum mean drag is $5.31\,\mathrm{mN}$. At $u = 2.0\,\mathrm{m/s}$ the field alternates at $20\,\mathrm{Hz}$, giving an emf of $28.3\,\mathrm{mV}$ peak. The current is $0.529\,\mathrm{A}$ peak and its phase lag 20.7°. The resistor receives approximately $7.00\,\mathrm{mW}$, and the motor maintaining translation supplies a mean force of $3.50\,\mathrm{mN}$. Using peak current in I^2R without its factor one half would overstate the mechanical force by a factor two.

Determine whether the current penetrates a real plate

A continuous metal plate contains many possible current paths, unlike a preformed loop. Their resistances, inductances and mutual couplings depend on the plate boundary, magnet geometry and frequency. For a good homogeneous conductor the magnetic diffusion equation gives a skin depth $\delta = \sqrt{2/(\mu\sigma\omega)}$ under sinusoidal excitation. At a depth δ a field entering a locally planar half-space has lost a factor e in amplitude and acquired a phase delay of one radian. The frequency to use is the frequency experienced by the moving material; the fact that the magnets are stationary in the laboratory does not make this frequency zero.

For copper with $\sigma = 5.8 \times 10^7\,\mathrm{S/m}$ and permeability close to μ_0, the skin depth is about $14.8\,\mathrm{mm}$ at $20\,\mathrm{Hz}$ and $1.48\,\mathrm{mm}$ at $2.0\,\mathrm{kHz}$. A $2.0\,\mathrm{mm}$ plate therefore sees little through-thickness attenuation in the first experiment but may need a diffusion calculation in the second. The plate's lateral current pattern can still matter when the through-thickness current density is nearly uniform. Slots interrupt wide current loops and usually reduce drag even when the mass and the bulk conductivity change little. Comparing a slotted and an unslotted plate directly tests the importance of current closure.

If a plate experiences sharply separated poles, its applied field in the moving frame contains a fundamental and higher harmonics. Harmonic n has a skin depth smaller by $\sqrt{n}$, so checking only the fundamental can miss surface concentration associated with narrow features. A Hall scan along the track supplies the actual spatial field spectrum. Multiplying each spatial wavenumber by the speed converts it into a temporal angular frequency for a constant-speed pass. Once the speed changes significantly over one pole spacing, a time-domain calculation using the measured position trace is preferable to this fixed-frequency conversion.

Close a mechanical and thermal energy budget

Take a rotating brake as a second experiment. A flywheel of known inertia $J_{\mathrm{rot}} = 0.020\,\mathrm{kg\,m^2}$ slows from 120 to $80\,\mathrm{rad/s}$ while no drive supplies power. The released rotational kinetic energy is

$$J_{\mathrm{rot}}(120^2 - 80^2)/2 = 80.0\,\mathrm{J}.$$

A separate run with the magnet removed measures bearing and air losses. Suppose the corresponding speed-dependent losses integrated over the magnetic braking trajectory total $8.0\,\mathrm{J}$. Then the electromagnetic channel receives $72.0\,\mathrm{J}$. Using the reference run's total energy over an arbitrary equal time could give a different answer, since the two speed histories affect the parasitic losses.

A calorimeter enclosing the conducting rotor has effective heat capacity $C = 60\,\mathrm{J/K}$ and a thermal link of conductance $G = 0.12\,\mathrm{W/K}$ to an environment at constant temperature T_a. Its balance is

$$C\,dT/dt = P_{\mathrm{deposited}} - G(T - T_a).$$

During a $20\,\mathrm{s}$ run, suppose the time-averaged rise above ambient is $0.50\,\mathrm{K}$. Then $1.20\,\mathrm{J}$ flowed through the thermal link. If all electromagnetic heat entered the calorimeter, its final rise from an ambient start is

$$(72.0 - 1.20)/60 = 1.18\,\mathrm{K}.$$

A measured rise near $0.80\,\mathrm{K}$ would leave an energy discrepancy of approximately $23\,\mathrm{J}$. Check conduction into a shaft, heat stored outside the instrumented volume, and the estimate of parasitic torque before adjusting the electromagnetic theory.

Find the spatial features that dominate a measured braking signal

A single Fourier component of field already produces an informative turnover. In a tracked experiment the field often has a measurable third harmonic. Let $B(x) = B_1 \cos kx + B_3 \cos(3kx + \varphi)$, with constant translation speed. Flux integration across the finite loop multiplies each harmonic by its own width factor. The corresponding emf peak is $2h|B_n|u|\sin(nkw/2)|$, and its temporal angular frequency is nku. The linear RL equation treats the two drives independently. Orthogonality over a complete fundamental period makes the mean of their product vanish, so total mean heat can be found by adding the individual resistive powers. Within one period the interference terms do affect instantaneous torque and power.

Suppose $B_3/B_1 = 0.20$ and keep the preceding $w = \lambda/4$. Here $\sin(kw/2)$ and $\sin(3kw/2)$ both have magnitude $1/\sqrt{2}$. At low speed the third harmonic therefore adds about four percent to mean drag. At $u = u_c$, however, its angular frequency sees inductive impedance $3R$ while the fundamental sees R. The third-harmonic share relative to the fundamental becomes $0.20^2(1 + 1)/(1 + 9) = 0.008$, only 0.8%. A peak-to-peak field reading can change visibly without a comparable change in mean drag. Conversely, if the loop width happens to null one spatial harmonic, another smaller harmonic can supply most of the signal. A field map and the loop geometry together determine what source drives the circuit.

In a solid plate, predicting force from field and speed needs a field solution that enforces current closure. In the slowly moving conductor, $\mathbf{J} = \sigma[-\nabla\phi + \mathbf{u} \times \mathbf{B}]$ when the imposed magnetic pattern is effectively time independent in the laboratory and the induced magnetic field can be neglected. Charge conservation then gives

$$\nabla \cdot [\sigma(-\nabla\phi + \mathbf{u} \times \mathbf{B})] = 0.$$

On an electrically insulated plate edge, enforce $\mathbf{n} \cdot \mathbf{J} = 0$. Solving this potential problem accounts for charge accumulation that redirects currents at a boundary. Simply setting $\mathbf{J} = \sigma\mathbf{u} \times \mathbf{B}$ can predict current flowing straight out of an insulating edge. The force follows from integrating $\mathbf{J} \times \mathbf{B}$ through the conducting volume, and Joule heat follows from $\int J^2/\sigma \, dV$.

A useful check on such a field calculation is a finite plate translating wholly inside an exactly uniform static magnetic field. Charges can rearrange until a uniform electrostatic field cancels $\mathbf{u} \times \mathbf{B}$. An isolated plate in this steady limit then has zero transport current and zero drag. Motion alone does not guarantee braking. At an entry edge, the moving conductor spans a spatial change of induction, and one scalar potential generally cannot cancel the motional drive around every closed path. Slits, thickness variations, a conducting contact at a boundary, and the induced field at higher speeds each change this test problem in an identifiable way.

Interpret a thermal transient without assuming it is instantaneous

If the brake runs at a stable speed and equilibrium electrical conditions, a local single-temperature model sees a nearly constant power. Its step response from ambient is $\Delta T = (P/G)[1 - \exp(-Gt/C)]$. Removing the drive gives an exponential decay with the same time constant C/G. Measuring that cooling tail fixes the ratio of heat capacity to conductance but does not set either one separately. A known heater power or an independent heat-capacity measurement supplies the missing absolute scale. A slope evaluated over a long interval without correcting leakage will underestimate the input power as the sample warms.

Take a small bench calorimeter enclosing the single test loop with $C = 1.20\,\mathrm{J/K}$ and $G = 0.020\,\mathrm{W/K}$. The $7.00\,\mathrm{mW}$ low-speed run gives a $0.350\,\mathrm{K}$ eventual rise and a $60\,\mathrm{s}$ time constant. After $30\,\mathrm{s}$ the rise is about $0.138\,\mathrm{K}$, whereas a thermally isolated calculation would give $0.175\,\mathrm{K}$. Its stored heat $0.166\,\mathrm{J}$ is less than the $0.210\,\mathrm{J}$ mechanical work over that interval; the thermal link accounts for the remainder. Loading the conductor thermally with an additional copper holder increases the response time and changes the initial slope without changing the final rise if conductance to the bath remains the same.

Compare predictions at a chosen travel speed

The live graph displays the force predicted by the finite-inductance loop alongside the resistive low-frequency extrapolation. Choose a speed, calculate the apparent excitation frequency, and compare the returned skin depth to the actual conductor thickness before interpreting the prediction as a plate measurement. Try choosing an inductance twice as large. The location of maximum force halves in

speed, but the low-speed slope at fixed geometry and resistance is unchanged. A real plate whose high-speed drag does not follow the single-loop curve can still agree with the model on its low-speed slope. Plotting the measured residuals against speed can reveal the point where extra current paths or through-thickness diffusion become important.

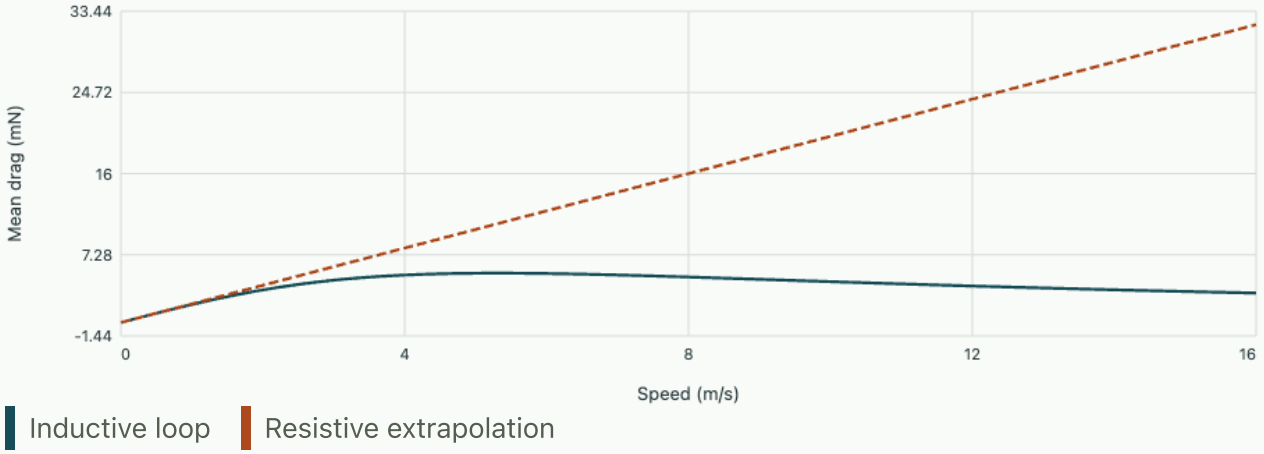
Open the live experiment

Try the calculation

Travel speed (m/s) **2**

Loop inductance (mH) **0.15**

Peak field (T) **0.2**



At 2.00 m/s the drag is 3.50 mN and the mean heat is 7.00 mW. The copper skin depth at 20.0 Hz is 14.78 mm. Peak drag occurs at 5.31 m/s.

A plate and holder cool from $0.240\,\mathrm{K}$ above ambient to $0.120\,\mathrm{K}$ in $42\,\mathrm{s}$ after a run. Can these two readings establish the braking power?

▼ **Work through the solution**

The half-life gives

$$C/G = 42/\ln 2 = 60.6\,\mathrm{s},$$

which agrees approximately with the small calorimeter example. It determines only a ratio. If an independent steady $10.0\,\mathrm{mW}$ heater then produces a $0.500\,\mathrm{K}$ temperature rise, $G = 0.0200\,\mathrm{W/K}$ and $C = 1.21\,\mathrm{J/K}$. A magnetic run that settles at $0.350\,\mathrm{K}$ now corresponds to $7.00\,\mathrm{mW}$, provided the thermal boundary and deposition pattern are sufficiently similar to the calibration.

Check the sign of instantaneous electromagnetic work

At a phase where the applied emf has just crossed zero, a finite-inductance loop may still carry considerable current. During part of a cycle the emf and current can have opposite signs. The product $\mathcal{E} i$ is then negative even though $R i^2$ remains positive. The difference is a negative derivative of the magnetic store $L i^2/2$. The loop briefly returns energy mechanically while its stored field decreases, and some of that release also becomes heat. A force transducer resolving separate pole passages can therefore see instantaneous variations that a mean braking-force estimate suppresses.

For the numerical $2.0\,\mathrm{m/s}$ case the peak field energy is $L I_{\mathrm{pk}}^2/2 \approx 21.0\,\mu\mathrm{J}$. The dissipated energy per full pole period is

$$7.00\,\mathrm{mW}/20\,\mathrm{Hz} = 350\,\mu\mathrm{J}.$$

These quantities serve different purposes. The first is a reversible instantaneous store; the second is net irreversible energy delivered each time the conductor moves one spatial period. Dividing the second by the spatial period of $0.100\,\mathrm{m}$ recovers the mean drag of $3.50\,\mathrm{mN}$ without using the translational speed explicitly.

Tutorial on braking measurements

At $u = 2.0\,\mathrm{m/s}$ the loop above experiences a mean drag of $3.50\,\mathrm{mN}$. Predict what halving the induction and doubling the speed does, keeping its geometry and electrical properties fixed.

▼ **Work through the solution**

Halving induction quarters K^2. The speed factor contributes $2[1+(2/5.31)^2]/[1+(4/5.31)^2] \approx 1.46$. The combined change gives a drag about 0.364 times the original, or $1.27\,\mathrm{mN}$. The mechanical power decreases less because speed has doubled. A linear-speed extrapolation would miss the inductive phase lag, already significant at $4\,\mathrm{m/s}$.

In a repeated flywheel test the brake becomes warmer with every run. Which electrical parameter change can explain a change of drag even with the magnets held fixed?

▼ **Work through the solution**

The resistance of an ordinary metal generally increases with temperature. At a fixed speed F is proportional to $R/[R^2 + (\omega L)^2]$. Raising resistance reduces drag on the low-frequency resistive side and can increase it on the high-frequency inductive side. The maximum occurs at $R = \omega L$ if electrical geometry and inductance stay fixed. Measuring the temperature and resistance before each run helps distinguish this effect from a weakening field or a changed gap.

Compare a step-driven diffusion profile with sinusoidal skin penetration.

Estimate a cavity linewidth from a finite, biased record

Numerical field laboratory

Connect the time and frequency descriptions of one damped mode, then diagnose why a weak detector background systematically biases the inferred quality factor.

Know whether the detector reads field or energy

After a drive switches off, let a single cavity mode have complex analytic amplitude $a(t) = a_0 e^{-\kappa t/2} e^{-i\omega_0 t}$ for $t \geq 0$. The energy is proportional to $|a|^2$, giving $U = U_0 e^{-\kappa t}$. Its instantaneous decay rate $-U'/U = \kappa$ implies

$$Q = \omega_0 U / P_{\mathrm{loss}} = \omega_0 / \kappa.$$

A phase-sensitive receiver measuring field amplitude sees a time constant twice as long as the energy lifetime. Before fitting the data, ask how the measurement transducer converts the field into a reported number.

Fourier transforming the causal decaying amplitude gives an integral proportional to $1/[\kappa/2 - i(\omega - \omega_0)]$, up to the time convention. Its squared magnitude has half-power points at $|\omega - \omega_0| = \kappa/2$. Therefore the full power linewidth in hertz is $\Delta f = \kappa/(2\pi)$. A measurement of either the energy decay time $\tau_U = 1/\kappa$ or the spectral width can test the other without having to calibrate the absolute detector gain.

At $f_0 = 2.0$ GHz and $\tau_U = 0.50\ \mu$s, we obtain $\kappa = 2.0 \times 10^6$ s^{-1}, $\Delta f \simeq 318.3$ kHz and $Q \simeq 6283$. A field receiver would report an amplitude time constant of 1.0 µs. Substituting that time directly into $Q = 2\pi f_0 \tau_U$ would double the apparent quality.

Derive the two-point fit with a detector offset

Let a power detector report $S(t) = A e^{-\kappa t} + B$, with a constant background $B > 0$. In the absence of background two samples separated by Δt give

$$\widehat{\kappa} = \log[S(t)/S(t+\Delta t)]/\Delta t.$$

With background, the logarithmic derivative is

$$-\frac{d}{dt}\log S = \kappa \frac{A e^{-\kappa t}}{A e^{-\kappa t} + B}.$$

It starts below the true rate and tends to zero in the tail. A straight line fitted to late log data can therefore look impressively precise while overestimating the lifetime. Increasing the acquisition duration can worsen this bias if the entire record is forced into a straight-line model.

In the experiment set $b = B/A = 0.04$. A two-point fit from one true lifetime to two has dimensionless rate $\log[(e^{-1} + 0.04)/(e^{-2} + 0.04)]$, approximately 0.844. Its apparent lifetime is about 18.5% high.

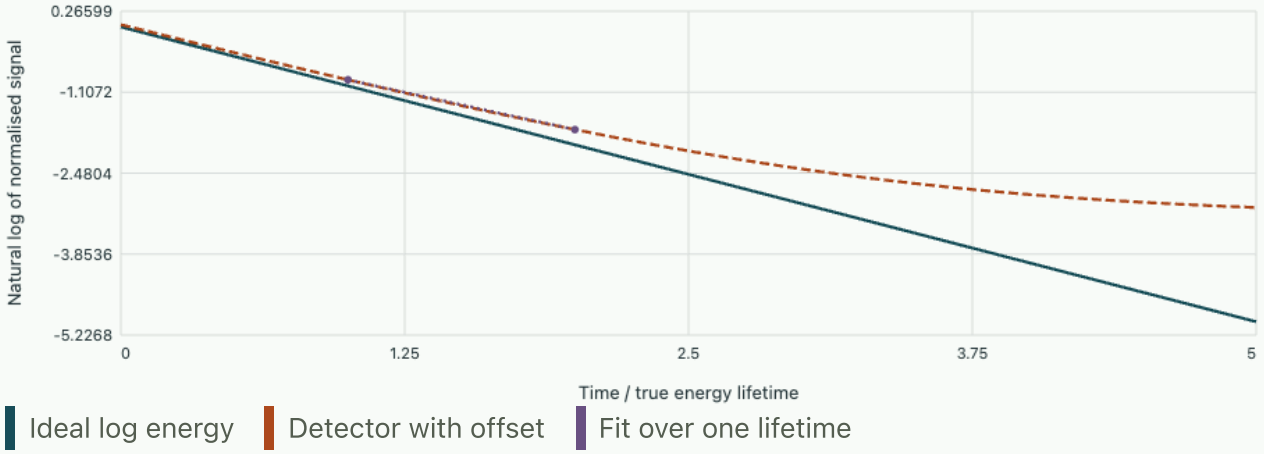
The source, walls and coupling have not changed. Only the interpretation of detector output has changed. Move the start of the fit towards three lifetimes and inspect the increasing gap from the ideal log line.

Open the live experiment

Try the calculation

Additive detector offset / initial energy **0.04**

Start of fitting window (lifetimes) **1**



Two-point inferred lifetime 1.1845 times the true lifetime. The same fractional bias transfers to Q for a known mode frequency.

Design a second independent inference

With three equally spaced, noiseless readings S_0, S_1, S_2 of one exponential plus a constant, subtract consecutive values before taking a logarithm. If $a = e^{-\kappa \Delta t}$, the differences are $A'(1-a)$ and $A'a(1-a)$, giving $a = (S_1 - S_2)/(S_0 - S_1)$. This cancels the constant background. At an interval of one lifetime, a first signal of unit energy and background 0.04 gives approximately $1.04, 0.407879, 0.175335$. Their ratio of differences recovers e^{-1}. In a real noise floor two nearly equal readings make this division unreliable. Use multiple windows or fit the raw signal with a measured noise model.

Finite spectral resolution creates a separate challenge. Multiplying the measured field by a rectangular window of duration T convolves its spectrum with a sinc-like response of width of order $1/T$. For the 318 kHz linewidth in the example, $T = 1$ µs gives frequency bins spaced by 1 MHz in a basic Fourier transform. Zero padding can interpolate its curve, but cannot restore the unmeasured late-time signal. A parametric fit may extract a sub-bin frequency or decay rate by using its model assumptions; its uncertainty then depends on signal-to-noise ratio and on whether the one-mode theory is adequate.

Check the single-mode model before refining its fit

If two nearby modes reach the same detector, their fields add before squaring. The result contains an interference term oscillating at their frequency difference on top of a decaying envelope. A scalar positive background cannot explain that alternating residual. Test by narrowing the excitation band, moving the pickup to a symmetry node of one mode, or examining a complex receiver's quadratures. A model with two exponentials has more degrees of freedom and can fit noise as well as signal; require consistency across independent records.

An honest linewidth report specifies the measured quantity, acquisition window, background treatment, coupling state, reference frequency and uncertainty. A fit to stored energy measures total loss. Inferring internal absorption alone additionally requires a known port coupling or a measurement that changes it separately, as in the course's cavity project.

Tutorial

A detector background is independently measured but subtracted too aggressively. What feature can appear in the logarithmic tail?

▼ **Work through the solution**

The corrected signal can approach zero and cross below it. The logarithm then becomes undefined. Even before that point its apparent local slope can become arbitrarily large. Inspect the raw signed residual and its uncertainty instead of dropping nonpositive points without accounting for selection bias.

A second experiment doubles the peak stored energy while the absolute detector offset is fixed. What should happen to the shape of the normalised log tail?

▼ **Work through the solution**

$b = B/A$ halves. Background contamination becomes important later by about $\tau_U \ln 2$. A real change in the physical lifetime need not follow that shift when the response remains linear.

Can a passive resonator's intrinsic losses be determined from a single ring-down measured at one port?

▼ **Work through the solution**

The observed decay is the sum of internal and external rates. A single accurately measured lifetime identifies their sum. A calibrated change in coupling or an independent power budget is needed to separate them.

Choose a frequency that identifies a coil's stored and dissipated energy

Numerical field laboratory

Design a small-signal experiment to estimate resistance and inductance, propagate amplitude and phase uncertainty and decide when a distributed field model is needed.

A measured impedance is one complex number

A laboratory sensing coil is driven by a small sinusoidal current at angular frequency ω. In its simplest usable range, a series model predicts $Z = R + i\omega L$. The voltage across a port defines the sign of current entering the positive terminal. With peak phasors, complex power is

$$S = VI^{*}/2 = |I|^{2}(R + i\omega L)/2.$$

Thus the real part measures average irreversible loss. Magnetic energy at peak current is $L|I|^{2}/2$; reactive power equals ω times this peak energy. These equalities supply two routes to the same parameters provided parasitic capacitance and frequency-dependent losses remain negligible.

At 100 Hz, a coil with $R = 20\ \Omega$ and $L = 40$ mH has reactance $2\pi(100)(0.040) = 25.13\ \Omega$, magnitude 32.12 Ω and phase about 51.49 degrees. A 10 mA peak current produces about 0.321 V peak and 1 mW average resistive loss. The peak magnetic energy is only $2\ \mu\mathrm{J}$. Check that the reactive power obtained from the phase and current is about 1.26 mvar. Confusion between RMS and peak phasors easily introduces a factor two while still leaving the impedance ratio correct.

Derive what an amplitude error and a timing error do

Suppose the readout gives magnitude M and phase ϕ. Then $R = M\cos\phi$, $L = M\sin\phi/\omega$. For small independent uncertainties and known ω, differentiate both expressions. The relative first-order perturbations are

$$\frac{\delta R}{R} = \frac{\delta M}{M} - \tan\phi\,\delta\phi,$$

\frac{\delta L}{L}=\frac{\delta M}{M}+\cot\phi\,\delta\phi .

The phase uncertainty must be expressed in radians. For independent random errors their variances add with the squared coefficients; the signed equations above describe individual perturbations. Near zero phase even a very accurate amplitude tells little about inductance, while near 90 degrees inferring resistance requires exceptional phase accuracy. A near-45-degree measurement balances these two sensitivities when the uncertainty budget treats them similarly.

If the phase uncertainty is 0.5° and relative magnitude uncertainty is 1%, the 45-degree standard deviation in either recovered parameter is approximately $\sqrt{0.01^{2} + 0.00873^{2}} = 0.0133$, or 1.33%. At five degrees the relative inductance contribution from phase alone is $\cot 5^{\circ} \times 0.00873 \simeq 10\%$. Taking data at more digits does not improve phase calibration.

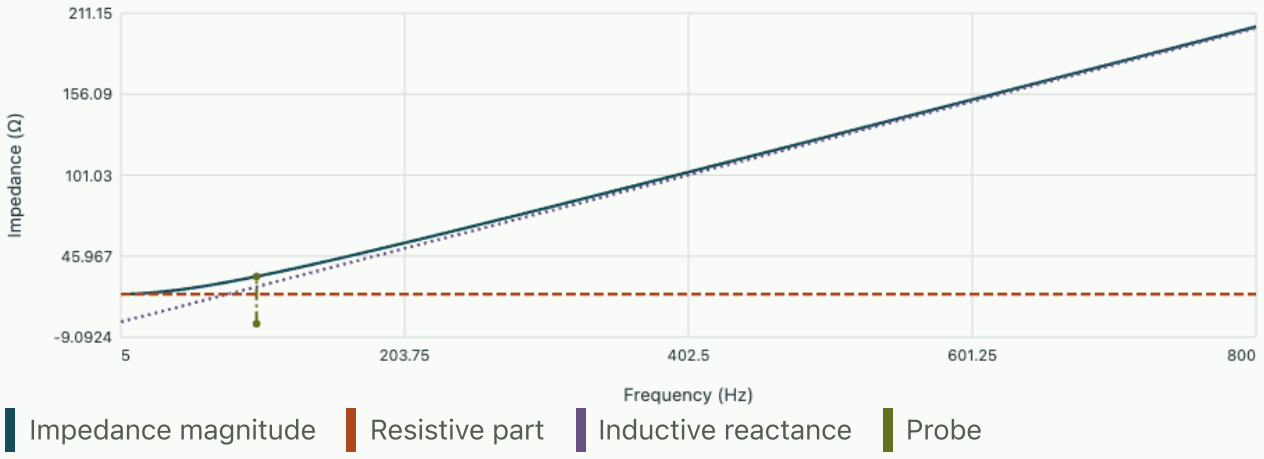
Choose the operating point, then challenge the model

Open the live experiment

Try the calculation

Probe frequency (Hz) 100

Coil inductance (mH) 40



At the probe: |Z| = 32.119 Ω, phase 51.488°. Equal resistive and reactive contributions at 79.577 Hz.

Vary frequency at fixed inductance and read where resistive and reactive contributions are equal. The crossover is $f_c = R/(2\pi L)$, about 79.58 Hz for the default coil. Now double the inductance without changing R; the crossover halves. The graph is a prediction with known parameters. In an actual inverse experiment you know only rough bounds initially. Begin with a small frequency sweep that brackets the crossover, then allocate repeated measurements where the target parameter is most sensitive.

Measure at two or more separated frequencies and reconstruct R and L independently at each. A frequency-dependent apparent R could be caused by skin effect or core loss. A falling apparent inductance near a self-resonance can signal stray capacitance. Simply allowing R, L to take arbitrary values at each frequency fits all data without explaining the mechanism. Add a candidate physical effect and check an independent observable such as loss under a different current amplitude, ring-down frequency or a change in lead geometry.

Close the loop from a fitted parameter to a field calculation

The lumped magnetic energy $LI^2/2$ should match the field integral $\int \mathbf{B} \cdot \mathbf{H}\, dV/2$ for a linear reversible system that includes the relevant surrounding field. This is why the exterior air region and its imposed boundary can affect a numerical estimate even when the measured coil is compact. Increase the domain and refine the mesh separately. A converged mesh in a small box can deliver a very

repeatable value for the wrong boundary problem. The earlier chapter on mutual-inductance matrices generalises this check to cross terms between currents.

Another independent measurement uses a slowly executed current ramp with a known winding resistance. Subtract the measured resistive voltage from terminal voltage and integrate over time. Under the linear model,

$$\int (V - RI)\, dt = L\Delta I.$$

This test weights offsets very differently from sinusoidal phase detection: a constant voltmeter offset integrates into a bias proportional to record duration, while a high-frequency phase error drops out. Agreement of the two methods is stronger evidence than repeating a single instrument setup at the same frequency.

Tutorial

Only voltage and current magnitudes are available, but two frequencies are accurately known. Can a constant R and L be recovered, and what makes the inference fragile?

▼ **Work through the solution**

$$M_i^2 = R^2 + \omega_i^2 L^2.$$

Subtract the equations to get

$$L^2 = (M_2^2 - M_1^2)/(\omega_2^2 - \omega_1^2).$$

Insert this in either equation for R^2 and take positive roots. Nearly equal frequencies produce a small denominator; nearly equal magnitudes amplify measurement noise. Under a misspecified model the inferred square can even become negative.

At a fixed frequency the current reverses direction in a repeat measurement but its reference voltage polarity is held fixed. What measurement error could this create?

▼ **Work through the solution**

Reversing the defined port-current orientation multiplies the inferred impedance by –1 unless the voltage reference is reversed consistently. An apparent negative resistance in passive hardware may be a port convention issue. Check the direction of measured power flow with a known reference resistor.

What extra evidence would justify carrying a fitted inductance into a time-domain prediction?

▼ **Work through the solution**

Make the prediction for a current waveform whose spectrum lies within the validated range of the measurements. Test at another amplitude for nonlinearity or hysteresis and record the residual voltage after subtracting $RI + L\,dI/dt$. A successful fit at one frequency alone does not establish a broadband constitutive law.

Master equations, detailed balance and probability currents

Statistical mechanics from microscopic counting to material response

A two-state detector can switch at random while displaying reproducible occupation fractions and relaxation times. A three-state detector can maintain exactly constant occupancies while carrying a persistent directed current. Derive the probability dynamics, infer transition rates from timed records and decide what a stationary histogram can and cannot reveal about equilibrium.

From short-time transition probabilities to a generator

Take discrete states i resolved by an instrument. In a short time dt, conditional on current state i, let the probability of one transition $i \to j$ be $w_{ij}dt + o(dt)$. The probability of remaining in i is one minus the sum of departures to this order. Assume time-independent rates and the Markov property, meaning the conditional law of future transitions is determined by the current state. Sum gains and losses to derive

$$\dot{p}_i = \sum_{j \neq i} (p_j w_{ji} - p_i w_{ij}),$$

$$\sum_i \dot{p}_i = 0.$$

Each gain in one state is a loss in another, so total probability is conserved. In a column-vector convention $\dot{\mathbf{p}} = L\mathbf{p}$, off-diagonal entry $L_{ij} = w_{ji}$ and columns sum to zero. In a row-vector convention transpose the matrix. This is not cosmetic in a numerical calculation: using a matrix assembled for column probabilities while evolving rows can lose normalisation or give an incorrect stationary distribution.

For a given state the exit rate is

$$r_i = \sum_{j \neq i} w_{ij}.$$

Its survival probability obeys $\dot{S}_i = -r_i S_i$, so $S_i(t) = e^{-r_i t}$. Conditional on exiting, the destination has probability w_{ij}/r_i. This permits direct event-driven simulation: draw an exponential waiting time, choose a destination with these probabilities, advance the clock and repeat. Sampling one transition per arbitrary fixed time step generally distorts waiting-time statistics unless a controlled approximation or exact finite-time propagator is used.

Determine two rates from one occupancy and one time

For a two-state system name the rates $a = w_{01}$, $b = w_{10}$. Eliminating $p_0 = 1 - p_1$ gives

$$\dot{p}_1 = a - (a + b)p_1.$$

Its stationary value $p_* = a/(a + b)$ is independent of the common rate scale, and deviations decay with time constant $\tau = 1/(a + b)$. Thus one stationary fraction gives a ratio, one relaxation time gives a sum, and the two together identify both rates.

A location sensor records an upper-state fraction 0.30 and relaxation time 0.20 seconds. Then $a = 1.50$ and $b = 3.50$ per second. Mean waiting times are $1/a = 0.667$ s in the lower state and $1/b = 0.286$ s in the upper one. Neither equals the population relaxation time of 0.20 s. A transition rate inferred by taking the reciprocal of the average of these two waiting times estimates a different quantity.

Define the indicator $X = 1$ in state one and zero otherwise. Conditional probabilities from the same linear solution give

$$\langle [X(t) - p_*][X(0) - p_*] \rangle = p_*(1 - p_*) e^{-|t|/\tau}.$$

Consequently a normalised equilibrium autocorrelation determines τ without needing a deliberate preparation away from stationarity. A bandlimited detector may miss short visits, reducing counts and distorting waiting-time distributions even if the mean occupation looks plausible. Compare time sampling with both exit times as well as the relaxation time, or model the measured signal with an explicit observation process.

Detailed balance is stronger than stationarity

Let π_i denote a stationary distribution. The net probability current on an oriented edge is $J_{ij} = p_i w_{ij} - p_j w_{ji}$. Stationarity requires only that currents sum to zero at each vertex. Detailed balance requires $\pi_i w_{ij} = \pi_j w_{ji}$ on every edge individually. For a system in equilibrium with one heat bath, with states even under time reversal and energies E_i, thermal rates satisfy $w_{ij}/w_{ji} = e^{-\beta(E_j - E_i)}$ for nondegenerate states. If states collect multiple microstates, degeneracies or coarse-grained free energies enter their equilibrium weights.

For the previous two-state process at equilibrium with equal state degeneracies, $p_*/(1 - p_*) = 3/7$ implies

$$(E_1 - E_0)/(k_B T) = \log(7/3) = 0.847.$$

Multiplying both rates by ten preserves the energy inference and speeds relaxation tenfold. Thermodynamic weights therefore do not determine the kinetic barrier. A physical sample may change its rate scale with solvent viscosity or transition geometry without changing the equilibrium ratio.

A convenient test on a connected graph with reciprocal edges is the product of rate ratios around each closed cycle. With detailed balance all stationary factors telescope to one. Conversely, if the product is one on every independent loop, consistent state weights can be reconstructed from any reference state by integrating ratios along a path. Path independence makes these weights a detailed-balance stationary distribution after normalisation. A missing reverse edge or unobserved channel requires additional care; a detector that cannot resolve two routes can hide directed currents.

A steady nonequilibrium system with the same state histogram

Consider three equally named sites arranged in a ring, with clockwise rate k_+ and anticlockwise rate k_- on every link. Symmetry gives stationary occupation $1/3$ at each site for all positive rates. Each oriented clockwise edge carries current $J = (k_+ - k_-)/3$. The cycle affinity in dimensionless units is $\mathcal{A} = \log[(k_+/k_-)^3]$. For $k_+ = 4$, $k_- = 1$ per second, each edge transports net probability one per second, and $\mathcal{A} = 3\log 4 = 4.159$. A perfectly flat histogram conceals the directed operation.

For Markov jumps with resolved reversible edges and their conventional time reversals, the total stochastic entropy production rate is

$$\dot{S}_{\mathrm{tot}} = \frac{k_B}{2}\sum_{ij}(p_i w_{ij} - p_j w_{ji})\log\frac{p_i w_{ij}}{p_j w_{ji}} \geq 0.$$

Each term is nonnegative under the product $(x-y)\log(x/y)$; the factor of one half prevents counting both directions separately. For the stationary ring this gives $k_B J\mathcal{A} = 4.159 k_B$ per second. An actual heat dissipation rate needs a physical local detailed-balance model identifying environments, work and entropy flows. A collection of counted transitions without that calibration can still yield evidence for broken time-reversal symmetry at its level of resolution.

Stationary Shannon entropy $-k_B\sum p_i \log p_i$ is constant in this ring. The positive total entropy production is compatible with that constancy because an environment sustains the directed current and carries compensating entropy. Equating production with just the derivative of stationary state entropy would predict zero for every driven stationary process.

Why a distribution relaxes toward a stationary law

For a finite irreducible time-independent chain let $\pi_i > 0$ be its unique stationary distribution. The relative entropy $D(p\|\pi) = \sum p_i \log(p_i/\pi_i)$ is nonnegative. To see its contraction over a finite observation interval, write the propagated probability as a stochastic mixture of earlier probabilities and apply the log-sum inequality to each output state. Summing gives

$$D(Pp\|P\pi) \leq D(p\|\pi).$$

Since $P\pi = \pi$, the relative entropy to stationarity cannot grow.

For a detailed-balance chain, substituting the master equation into this derivative gives $-k_B dD/dt = \dot{S}_{\mathrm{tot}}$ in the simple closed heat-bath model. For a driven chain the system may remain at π with constant relative entropy while producing positive steady entropy. Relaxation toward stationarity and dissipation required to maintain that stationary state are distinct measurable quantities.

Compare relaxation inferred from a time trace with the thermodynamic irreversibility experiment

Infer generator entries from residence times

An ideal continuously observed Markov record contains successive states and dwell lengths. Let T_i be accumulated time in state i and N_{ij} the number of resolved $i \to j$ jumps. Its log likelihood contains $\sum_{ij} N_{ij} \log w_{ij} - \sum_i T_i \sum_{j \neq i} w_{ij}$. Differentiation gives

$$\widehat{w}_{ij} = N_{ij}/T_i.$$

Dividing by total experiment time instead estimates the gross stationary flux $\pi_i w_{ij}$, not a conditional transition rate.

For a nearly stationary two-state record with $T_0 = 70$ s, $T_1 = 30$ s and 105 transitions in each direction the estimates are 1.50 and 3.50 per second. Counting uncertainty for each reasonably large Poisson-like count is roughly the inverse square root of the count in relative terms; precise interval estimates condition correctly on exposures and the observation process. Equality of opposite transition counts in a long two-state record is essentially forced by alternation. It cannot by itself establish a thermal energy gap or exclude multiple hidden driven pathways.

Turn a barrier diagram into two independently measurable rates

A thermally activated transition between two metastable configurations often has a forward rate $a = \nu_A e^{-(G^\ddagger - G_A)/(k_B T)}$ and reverse rate $b = \nu_B e^{-(G^\ddagger - G_B)/(k_B T)}$, where $G^\ddagger$ denotes a dividing-surface barrier under a chosen model. If the same transition-state construction applies in both directions with consistent partition factors, their ratio reproduces the equilibrium free-energy difference. The overall barrier controls how rapidly equilibrium is approached. Changing a barrier while leaving both wells unchanged can alter both kinetics by many orders of magnitude without changing stationary occupancy.

As an illustration let the higher-energy state cost $2k_B T$, and put a transition configuration $10k_B T$ above the lower state. With a common attempt rate of 10^6 per second, the upward rate is 45.4 per second and the downward rate 335 per second. The stationary upper fraction is approximately 0.119 and the relaxation time 2.63 ms. Lower the barrier by $2k_B T$ without changing the two state energies. Both rates increase by $e^2 = 7.39$, while occupancy is unchanged. A detector recording only the equilibrium level histogram could miss the intervention entirely.

For a temperature scan, a straight line in the log rate against inverse temperature is evidence for an approximately constant activation energy and prefactor over that interval. A free-energy barrier includes entropic as well as enthalpic contributions; its numerical value can vary with temperature even if the microscopic potential does not. A change of dominant pathway can bend the apparent Arrhenius plot. Measure forward and reverse rates independently, then compare their ratio with an independently determined equilibrium population. A barrier fit explaining one rate but violating the equilibrium ratio usually indicates inconsistent state definitions, missed short visits or an unmodelled pathway.

A useful control is a weak static bias that lowers one well relative to the other. The equilibrium log-odds shift measures the difference between the state couplings to the bias. The individual log-rate shifts additionally report the related dividing-surface response. If the state-occupation ratio moves as expected but only one rate changes appreciably, the effective transition configuration may be much

closer in bias response to the other state. Such statements depend on the activated single-path interpretation; the response slopes themselves are more directly observed than an inferred barrier shape.

Use uniformisation to propagate a generator without negative probabilities

The formal finite-time propagator of a time-independent jump process is $P(t) = e^{Lt}$. A forward Euler update $p(t + \Delta t) = (I + L\Delta t)p(t)$ conserves total probability if the generator is assembled correctly, but a large step can yield negative diagonal probabilities. An exact algebraic representation starts with a rate q at least as large as every total exit rate. Then $K = I + L/q$ is a stochastic matrix in the same column convention. Write $L = q(K - I)$; since K and the identity commute,

$$e^{Lt} = e^{-qt} \sum_{n=0}^{\infty} \frac{(qt)^n}{n!} K^n.$$

This can be read as drawing a Poisson number of opportunities at rate q, followed by n transitions of K. At a lightly connected state some opportunities become self transitions. The construction leaves real waiting-time statistics correct while eliminating negative transition probabilities. Its truncation error in total probability is bounded by the omitted Poisson tail because every matrix power is stochastic.

For a two-state model with rates two and six per second choose $q = 6$ per second. Starting in the lower state, the first opportunity after some waiting time moves upward with probability one third and otherwise does nothing. In the upper state the next opportunity moves downward with probability one. At 0.25 s the Poisson mean is 1.5. Retaining terms through order eight leaves a tail below 3×10^{-5}. Compare the resulting upper probability against $(2/8)[1 - e^{-8(0.25)}] = 0.2162$. A small discrepancy due solely to truncation should shrink with the computable tail probability.

The representation also suggests a measured-data sanity check. A fitted generator with negative off-diagonal rates cannot yield a valid K at any positive q. A snapshot transition matrix can itself be nonnegative and stochastic but need not arise from a time-homogeneous continuous-time generator. In the two-state case its nontrivial eigenvalue must be $e^{-(a+b)t} > 0$. A measured negative eigenvalue rules out that simple generator model under perfect sampling, though finite counts and hidden periodic driving can distort empirical estimates. Taking a matrix logarithm blindly may return complex or negative rates rather than a physical kinetic explanation.

Population relaxation and a first-arrival experiment answer different questions

A first-passage measurement stops its clock when a target is first reached, even if that target would subsequently be left many times. To calculate its mean, treat the target as absorbing. For any other state i, wait on average $1/r_i$, then add the mean remaining time from the next state weighted by its conditional transition probability. The backward equations are

$$-1 = \sum_j w_{ij}(t_j - t_i),$$

with $t_{\mathrm{target}} = 0$. Their solution depends on possible returns before the first successful passage. It generally differs from an inverse equilibrium occupation or from a slowest relaxation time.

Consider a three-site linear chain $0 \leftrightarrow 1 \leftrightarrow 2$ with equal rate k on each adjacent directed edge, starting at zero and stopping at two. From zero one must reach one after mean $1/k$. From one the mean wait is $1/(2k)$, then there is probability one half of returning to zero and one half of finishing. Solving $t_0 = 1/k + t_1$, $t_1 = 1/(2k) + t_0/2$ gives

$$t_0 = 3/k.$$

Ignoring returns would yield $1.5/k$, exactly half the correct first-passage mean.

The unmodified closed chain has equal stationary weights, but that fact alone says nothing about the numerical value of $3/k$. Speeding every link by ten leaves occupancies unchanged and reduces passage times tenfold. A detector that registers only visits longer than a threshold makes a brief first arrival invisible and can report even longer apparent passage times. Calibrate detector dead time and distinguish entrance to the target from persistence within it.

For a larger network, solve the sparse backward linear system and compare against many event-driven paths. Check boundary initial conditions by starting simulations exactly at the absorbing target: the first-passage time must then be zero under that convention. Some experiments instead ask for a first return after leaving, which requires a different initial condition. A zero-lag event in one definition is a physically different measurement in the other.

Estimate a cycle affinity with finite counting statistics

A clockwise ring trajectory contains gross counts in both directions on each edge. Estimating each conditional rate as count divided by origin-state exposure, the product of clockwise-to-anticlockwise rates around a complete closed loop cancels the exposures. The estimated cycle affinity is the sum of log forward counts minus log reverse counts, assuming every required count is nonzero and rates were stationary during the record. This is a useful cross-check independent of how the total time was partitioned among states.

If all three edges register roughly 120 clockwise and 80 anticlockwise events, the estimated affinity is

$$3\log(1.5) = 1.216.$$

For large independent Poisson-like gross counts the approximate variance of this log product is $3(1/120 + 1/80) = 0.0625$, giving standard error 0.25 as a first estimate. Actual path constraints couple counts; uncertainty can be assessed from independent records or a complete stochastic likelihood. The point estimate alone does not tell whether an observed directionality is statistically distinguishable from zero.

One missing reverse event makes the unregularised logarithm infinite, but a finite record cannot establish an infinite physical affinity. Report a confidence bound based on the exposure and upper limit on the unseen reverse rate. Extend the acquisition or change sampling speed to exclude missed short-lived reverse states. Numerical addition of an arbitrary count can help regularise a plotting routine but is not a substitute for a specified prior or measurement model when claiming physical entropy production.

A driven occupation loop can carry charge without changing concentrations

Place three sites in a cycle and drive their transition rates by two controls, for example energies and barriers. Even if each instantaneous frozen generator satisfies detailed balance with a well-defined equilibrium distribution, the physical occupations need not equal that moving equilibrium distribution at each moment. A slow lag can produce a small edge current at every instant and a finite integrated current over a cycle. The current through an edge is time-integrated transfer; it can be nonzero while all occupations return to their initial values after a complete drive period.

A simple piecewise illustration avoids subtleties of adiabatic geometric pumping. In stage one, connect only sites zero and one while strongly favouring one. In stage two connect only one and two while favouring two. In stage three connect only two and zero while favouring zero. Begin with nearly all probability at zero and wait long enough for each enabled pair to relax. Almost the entire population travels once around the loop per cycle. No site accumulates probability over repeated complete cycles, but each edge registers about one forward crossing net per cycle. External control changes the energy and barrier landscape and supplies the required work.

If each programmed transfer succeeds with probability 0.95, it is tempting to announce a current of 0.95^3 per cycle. This is the probability of one fully successful trajectory starting on the prescribed site in a simplified no-return preparation, not generally the long-run edge current of a periodically driven ensemble. Failed transfers leave probability on other sites in later stages. The correct periodic distribution is the unit-eigenvalue vector of the product of the three finite-time propagators. Compute edge currents during each stage using its evolving occupations and integrate, then compare a complete cycle after transients have decayed.

This construction supplies a statistical check on experimental counting. Over a complete cycle the integrated net inflow at a site equals its change in occupancy probability. In a periodically stationary state that change is zero, so all three consistently oriented integrated edge currents must agree. A disagreement between measured edges can reveal a missed transition channel, a start-up transient or incorrect gating. Agreement does not prove absence of drive or detailed balance. A torque-free stationary ring and a three-stage external pump can share nearly the same time-averaged site histogram yet differ clearly in time-resolved edge currents.

Finally compare the work and heat bookkeeping with the master-equation entropy-production expression. During an external change of a site's energy without a transition, the state's energy changes by work. During a transition at frozen controls, assigning heat requires the energy difference and the relevant bath. Lumping an entire driven cycle into a single time-averaged generator can conceal those distinctions and underestimate dissipation.

The ordered protocol can also be run backwards. If the three stages are reversed while preparations are equilibrated in each corresponding periodic state, the preferential current direction should reverse in the ideal transfer picture. A fixed instrumental preference for counting neighbouring sites in one direction would not generally reverse. This distinguishes controlled transport from a direction-dependent detection efficiency. Do not confuse this observed current reversal with exact microscopic time reversal of each individual stochastic trajectory: the reversed experiment has its own ensemble of random jumps. Comparing path probabilities, rather than pairing individual paths as though their noise histories were reproducible, is the appropriate starting point for entropy-production and fluctuation-relation tests. The distinction becomes particularly visible when a stage transfers almost all the probability in one direction while a few rare trajectories travel against the applied protocol.

Distinguish stationarity from detailed balance

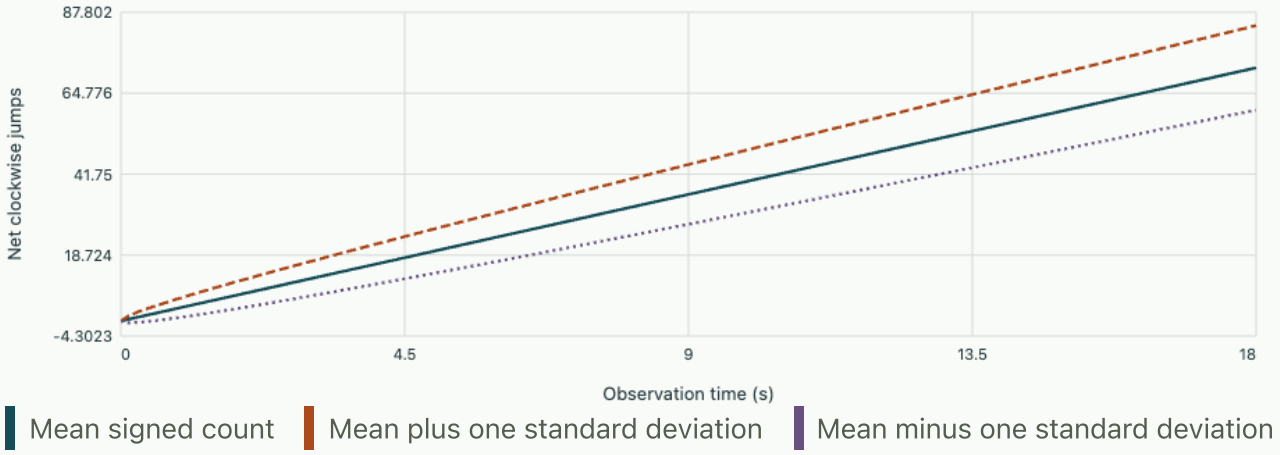
Keep the three populations identical while varying the directed transition rates. Identify the observable that reveals time-reversal asymmetry.

Open this experiment online

Try the calculation

Clockwise bias **0.5**

Total jump rate (s⁻¹) **8**



Every stationary population is 1/3. Net clockwise edge current 1.3333 s⁻¹. Entropy production 4.3944 kB/s.

Tutorial · Recover a generator from snapshots rather than full events

At sampling interval 0.10 s, a two-state record yields $P_{01} = 0.12$ and $P_{10} = 0.28$. Infer the continuous-time rates, and compare with directly dividing each transition probability by 0.10 s.

▼ **Use the exact finite-interval propagator**

The exact transition probabilities are $P_{01} = p_*[1 - e^{-(a+b)\Delta t}]$ and $P_{10} = (1 - p_*)[1 - e^{-(a+b)\Delta t}]$. Their sum is 0.40, so $a + b = -\log(0.60)/0.10 = 5.108$ per second. The stationary upper fraction is

$$0.12/0.40 = 0.30.$$

Thus $a = 1.532$ and $b = 3.576$ per second. Dividing by the interval gives only 1.2 and 2.8 because return transitions inside a sampling interval can erase evidence of the intermediate visit.

Tutorial · Can measured rates arise from one equilibrium potential?

For a three-state triangle suppose clockwise to counterclockwise conditional rate ratios are 2, 3 and 0.25 around the three distinct edges. Could these be equilibrium weights in a common bath? If a single uncertain ratio were corrected, what would it have to be on the third link?

▼ **Apply the closed-loop consistency condition**

The product is $2 \cdot 3 \cdot 0.25 = 1.5$, giving nonzero loop affinity $\log 1.5 = 0.405$. No consistent detailed-balance set of three state weights exists for these exact ratios. The third ratio would have to be $1/6$, giving a telescoping unit product. If counts are noisy, propagate uncertainty in the log ratios or compare a constrained detailed-balance likelihood against an unconstrained model. A value barely inconsistent with unity within one sparse record is not enough to establish driving.

Tutorial · Detect a hidden state with the same observed level

A detector reports two visible levels, but its lower level combines two unresolved substates. The measured lower-state survival probability is $0.5e^{-t} + 0.5e^{-9t}$ with time in seconds. Calculate the initial hazard and its long-time limit. Can a two-state Markov fit reproduce both?

▼ **Differentiate survival before interpreting the tail**

The hazard is minus the derivative of $\log S$. At zero time it is the initial weighted mean of rates, 5 per second. At long times the rapidly departing component has almost disappeared and the hazard tends to 1 per second. A single Markov exit rate would give a constant hazard for all dwell ages. The observed model cannot reproduce both limits with one rate.

A changing apparent hazard may also result from limited detector temporal resolution or variable preparation of the substates. Repeat with shorter acquisition intervals and condition the dwell distribution on the preceding visible transition or an independent preparation marker. A fitted sum of

exponentials identifies a need for a richer observation or state model; it does not uniquely identify microscopic structures absent in the measurements.

Cambridge Kinetic Theory · stochastic processes and approach to equilibrium

How an unobserved environment changes a quantum state

Quantum states, dynamics and measurement

Model a qubit and a two-state environment by a joint unitary. Derive channels, the continuous-time master equation and the relation between longitudinal and transverse relaxation. A recorded event and an unread detector produce different conditional predictions.

Derive loss of coherence from a larger unitary system

A qubit prepared in $\alpha|0\rangle + \beta|1\rangle$ interacts with an environment initially in $|e\rangle$. A joint unitary correlates $|0, e\rangle \to |0, e_0\rangle,\ |1, e\rangle \to |1, e_1\rangle$, with both environmental states normalised. The joint output is a pure state. Its qubit density matrix follows by tracing over a complete environment basis. Populations stay $|\alpha|^2, |\beta|^2$. The off-diagonal entry is $\alpha\beta^*\langle e_1|e_0\rangle$. Orthogonal environment records remove coherence from all local qubit observations; identical records leave its magnitude intact. Partial distinguishability gives a continuous range in between. Energy need not flow in this controlled dephasing model. An observed reduction in fringe contrast can arise from entanglement with unobserved degrees of freedom even when total evolution is unitary.

Consider instead amplitude loss from an excited qubit to an initially empty environment mode:

$$|0, 0_e\rangle \mapsto |0, 0_e\rangle, \qquad |1, 0_e\rangle \mapsto \sqrt{1-p}|1, 0_e\rangle + \sqrt{p}|0, 1_e\rangle.$$

The two output states remain orthogonal and unit-normalised, so this map on the prepared environment subspace can be extended to a unitary on the full joint space. If environment state k is unobserved, sum the associated unnormalised alternatives in density-matrix form:

$$\rho' = \sum_k K_k \rho K_k^\dagger.$$

Here $K_0 = |0\rangle\langle 0| + \sqrt{1-p}|1\rangle\langle 1|$ and $K_1 = \sqrt{p}|0\rangle\langle 1|$. Since $\sum K_k^\dagger K_k = I$, trace is preserved. The excited population obeys $\rho'_{11} = (1-p)\rho_{11}$ and the coherence $\rho'_{01} = \sqrt{1-p}\rho_{01}$. Averaging state vectors before forming outer products would introduce interference between orthogonal environmental records that have already been traced out.

Take a controlled infinitesimal limit

In a memoryless approximation each small interval dt meets a fresh environment record with vanishing probability of a previously lost excitation returning. Let jump operators L_j have units $1/\sqrt{\mathrm{s}}$, set $K_j = \sqrt{dt}\, L_j$ for events and $K_0 = I - dt(iH/\hbar + \sum L_j^\dagger L_j/2)$. Expanding the completeness relation shows trace preservation through first order in dt. Expand $\sum K_k \rho K_k^\dagger - \rho$, divide by dt and keep the finite terms:

$$\frac{d\rho}{dt} = -\frac{i}{\hbar}[H, \rho] + \sum_j \left(L_j \rho L_j^\dagger - \tfrac{1}{2} L_j^\dagger L_j \rho - \tfrac{1}{2} \rho L_j^\dagger L_j \right).$$

The first term of each dissipative bracket adds probability after an event and the anticommutator subtracts probability from the no-event branch. Their traces cancel. This construction is completely positive, meaning that the same local channel remains positive even when tensored with identity on an unused entangled ancilla. Positivity only on isolated input states would be too weak: transposition maps density matrices to density matrices on a single system, yet partial transposition can produce a negative eigenvalue when applied to half an entangled pair. The Lindblad form embodies the relevant stronger constraint for the assumed Markovian dynamics.

Independent loss and pure dephasing can be represented by $L_1 = \sqrt{\Gamma_1}|0\rangle\langle 1|$,

$$L_\phi = \sqrt{\Gamma_\phi/2}\,\sigma_z.$$

The factor one half in the dephasing operator fixes our convention for Γ_ϕ. Direct substitution gives $\dot\rho_{11} = -\Gamma_1\rho_{11}$, while $\dot\rho_{01}$ acquires $-(\Gamma_1/2 + \Gamma_\phi)\rho_{01}$ in addition to its Hamiltonian phase. Thus

$$T_1 = \Gamma_1^{-1},$$

$$T_2^{-1} = (2T_1)^{-1} + \Gamma_\phi.$$

At zero temperature with these mechanisms, $T_2 \le 2T_1$. This upper bound is tied to definitions and assumptions. A broad ensemble with static detuning spread, a driven nonequilibrium sample or a refocused pulse sequence may yield an observed envelope with a different shape and timescale. Fitting every contrast decay with one exponential and interpreting that number as the homogeneous T_2 can assign an unphysical dephasing rate.

Update a state when an individual event is recorded

If an ideal detector reports branch j, the probability is $p_j = \operatorname{tr} K_j\rho K_j^\dagger$, and the normalised conditional output is $K_j\rho K_j^\dagger/p_j$. The conditional evolution depends on which environment observable is measured; summing outcomes gives the same physical unconditioned channel. In the loss model, detecting one emission places the qubit in its ground state. Absence of an emission also carries information. At finite duration t, no emission acts by $K_0 = |0\rangle\langle 0| + e^{-\Gamma_1 t/2}|1\rangle\langle 1|$ for the simple undriven zero-temperature case.

From initial $(|0\rangle + |1\rangle)/\sqrt{2}$, at $t = T_1 \ln 2$ the no-event probability is

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

Conditional excited probability is

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

With the event branch in the ground state, the unconditioned excited probability is $(3/4)(1/3) = 1/4$, agreeing with exponential decay of its initial one half. Repeatedly renormalising only no-event trajectories and calling their mean the ensemble population would overestimate survival. If the event

detector misses photons, “no click” groups genuine no-emission with undetected emissions and its conditional state is generally mixed.

Use heating and cooling to infer the reservoir

Allow upward and downward transitions at rates $\Gamma_\uparrow, \Gamma_\downarrow$. Rate equations give

$$\dot{p}_1 = \Gamma_\uparrow(1 - p_1) - \Gamma_\downarrow p_1,$$

$$p_{1,\infty} = \frac{\Gamma_\uparrow}{\Gamma_\uparrow + \Gamma_\downarrow},$$

$$p_1(t) = p_{1,\infty} + [p_1(0) - p_{1,\infty}]e^{-(\Gamma_\uparrow + \Gamma_\downarrow)t}.$$

Both upward and downward rates contribute half their sum to coherence decay. If the reservoir is in thermal equilibrium, detailed balance gives

$$\Gamma_\uparrow/\Gamma_\downarrow = e^{-\hbar\omega/(k_B T)}.$$

The population relaxation time measures their sum and equilibrium population measures their ratio, allowing a substantive check of a proposed bath temperature. The relation is a thermodynamic assumption in addition to the two-state rate model. A pumped bath can have the same measured relaxation time with an entirely different stationary population.

For a 5.00 GHz transition, $\hbar\omega/k_B = h(5\,\mathrm{GHz})/k_B \simeq 0.240$ K. A stationary excited fraction 0.10 would imply $T = 0.240/\ln 9 = 0.109$ K under equilibrium. With population relaxation 20 µs the sum of rates is $5.0 \times 10^4\,\mathrm{s}^{-1}$, giving 5000 upward and $45\,000$ downward transitions per second for the respective source populations. Verified independent upward and downward preparations test more than fitting the decay of an initially excited state alone.

Discriminate static detuning, Markovian dephasing and coherent error

A Ramsey experiment prepares an equal superposition, allows free evolution for time t, then converts relative phase to a population by a final analysis pulse. If the detuning δ is fixed in each run but varies randomly between runs, then the ensemble coherence is multiplied by the characteristic function $C(t) = \int p(\delta)e^{-i\delta t}d\delta$. Here δ is an angular frequency. For a Gaussian detuning of mean δ_0 and standard deviation σ,

$$C = e^{-i\delta_0 t}e^{-\sigma^2 t^2/2}.$$

A uniform detuning offset shifts fringe phase without reducing the contrast of a fully coherent, identically repeated preparation. A Lorentzian distribution yields an exponential envelope which can

resemble homogeneous Markovian dephasing in a simple time scan. The fitted functional form alone may therefore fail to identify the physical cause of lost contrast.

An ideal instantaneous π pulse about x halfway through free evolution reverses the effect of a static z-detuning on relative phase: $e^{-i\delta t\sigma_z/4}\sigma_x e^{-i\delta t\sigma_z/4} = \sigma_x$, since σ_x anticommutes with σ_z. All static offsets refocus at the echo. Independent white Markovian dephasing accumulated in each half is uncorrelated and cannot be undone in this way. The deterministic Hamiltonian error and a rapidly fluctuating environment can share a short Ramsey contrast while giving different echo contrasts. If the pulse itself has finite duration or an angle error, the algebraic cancellation is no longer exact; repeating at two pulse powers tests that particular assumption.

Take $\sigma/2\pi = 20.0$ kHz and negligible damping. At $t = 10.0\,\mu$s the free Ramsey contrast from Gaussian detuning alone is

$$e^{-(1.257)^2/2} = 0.454.$$

An ideal echo returns unity. Add independent $T_1 = 80\,\mu\text{s},\ T_\phi = 50\,\mu\text{s}$. The homogeneous coherence time is

$$[1/(160\,\mu\text{s}) + 1/(50\,\mu\text{s})]^{-1} = 38.1\,\mu\text{s}.$$

The expected echo contrast is now $e^{-10/38.1} = 0.769$ and the ordinary Ramsey contrast 0.349. A single fitted T_2^* cannot display these three separate timescales. Measuring populations, Ramsey fringes and echoes on consistent ensembles gives three complementary observables with independently interpretable limits.

For a time-dependent classical detuning $\delta(t)$, pulse sequences choose a modulation $y(t) = \pm 1$ multiplying the noise phase. The phase error is

$$\phi = \int y(t)\delta(t)dt.$$

For stationary zero-mean Gaussian noise, $\langle e^{-i\phi}\rangle = e^{-\langle\phi^2\rangle/2}$ and the variance is a double integral of the noise correlation weighted by $y(t)y(t')$. Equivalently it is the spectral integral of the power density weighted by $|\int y(t)e^{i\omega t}dt|^2$, with the normalisation set by the chosen two-sided Fourier convention. A single echo has zero weight at $\omega = 0$ because its two half integrals cancel, while keeping finite-frequency sensitivity. Families of pulse spacings can therefore probe noise timescales rather than simply fitting larger and larger numbers to a nominal coherence time.

Find the damped saturation curve from optical Bloch equations

A continuously driven qubit has an additional route to separate population and coherence relaxation. In a rotating frame let $H = \hbar(\Omega\sigma_x + \Delta\sigma_z)/2$, with $z = p_e - p_g$, equilibrium $z_{\text{eq}} = -1$ at zero temperature, and transverse damping $1/T_2$. Choose a consistent Bloch basis so the commutator gives the cross product of $(\Omega, 0, \Delta)$ with (x, y, z). The equations are

$$\dot{x} = -\Delta y - x/T_2,$$
$$\dot{y} = \Delta x - \Omega z - y/T_2,$$
$$\dot{z} = \Omega y - (z - z_{\mathrm{eq}})/T_1.$$

At steady state the first equation gives

$$x = -\Delta T_2 y.$$

The second then gives

$$y = -\Omega T_2 z/(1 + \Delta^2 T_2^2).$$

Inserting into the third gives

$$z = z_{\mathrm{eq}}/[1 + \Omega^2 T_1 T_2/(1 + \Delta^2 T_2^2)].$$

Thus the excited fraction for $z_{\mathrm{eq}} = -1$ is

$$p_{e,\infty} = \frac{\Omega^2 T_1 T_2}{2[1 + \Delta^2 T_2^2 + \Omega^2 T_1 T_2]}.$$

A weak resonant drive gives a population proportional to its power, and a strong continuous drive saturates toward one half rather than preparing a permanently inverted state. Coherent π-pulse inversion is a transient unitary process; continuous steady balance with relaxation gives a different preparation. The full angular-frequency width at half of the on-resonance excitation is $2\sqrt{1 + \Omega^2 T_1 T_2}/T_2$, so increasing power broadens the line even at unchanged intrinsic T_2.

Choose $T_1 = 20$ µs, $T_2 = 16$ µs and resonant $\Omega/2\pi = 10$ kHz. The saturation parameter is

$$s = \Omega^2 T_1 T_2 = 1.26.$$

The stationary excited fraction is $s/[2(1 + s)] = 0.279$, and the ordinary-frequency full width is

$$\sqrt{1 + s}/(\pi T_2) = 29.9\,\mathrm{kHz}.$$

At very low drive it would be 19.9 kHz. A measured linewidth that increases approximately as the square root of power at high drive can have this coherent saturation origin without any increase in the environment's pure-dephasing rate. Comparing the zero-power linewidth extrapolation with an independent time-domain T_2 is a stronger test than fitting a single apparent line broadening.

Connect an individual event stream to an ensemble lifetime

The lifetime T_1 of one excited qubit is a distribution of observed emission times across many trials. With perfect detection, no drive and initially excited state, survival $S(t) = e^{-\Gamma t}$ implies waiting-time density $w(t) = -dS/dt = \Gamma e^{-\Gamma t}$. The mean waiting time and standard deviation are both $1/\Gamma$. A no-click trajectory stays in the excited state conditionally until its first emission when the initial preparation is certainly excited. Its unnormalised weight decays exponentially, but assigning every no-click trajectory a continuously decreasing normalised excitation would confuse conditional and unconditioned statements. For an initial superposition the absence record instead changes the normalised state as derived earlier.

Real acquisitions may stop each trial at a fixed deadline T_{stop}. Treating all no-click trials as though they emitted exactly at the deadline biases lifetime inference. For n observed emissions at times t_j and m censored trials with no emission through T_{stop}, the likelihood is proportional to $\prod_{j=1}^{n} \Gamma e^{-\Gamma t_j} \prod_{k=1}^{m} e^{-\Gamma T_{\mathrm{stop}}}$. Optimising gives

$$\hat{\Gamma} = n / (\sum_{j} t_j + m T_{\mathrm{stop}}).$$

The numerator is actual events and the denominator all time observed at risk. For 40 detections totalling 480 µs and 10 additional censored runs each observed 30 µs, the estimate is $40/780 = 0.0513\,\mu\mathrm{s}^{-1}$, or $T_1 \simeq 19.5\,\mu\mathrm{s}$. Discarding the censored runs instead would report 12 µs from the mean detected delay. This statistical correction follows directly from the same physical survival law used for the quantum channel.

Detection inefficiency can cause a different data problem. In a single-excitation undriven trial, a missed decay leaves a qubit in its ground state and unable to emit again. The probability of no registered event up to t is $1 - \eta[1 - e^{-\Gamma t}]$ at constant efficiency η, with a nonzero late plateau $1 - \eta$. It is not $e^{-\eta\Gamma t}$, which would describe thinning a renewable stationary Poisson event stream. Long-time survival of the recorded count and independent final-state readout can distinguish a missed emission from a genuinely long-lived excitation. This example shows why the detector model must specify whether events can repeat after one has been missed.

Audit a reconstructed Pauli noise channel

Suppose three independently prepared Bloch components appear to shrink by factors $\lambda_x, \lambda_y, \lambda_z$ after a calibrated waiting time, with no displacement of the centre. A diagonal unital qubit channel can be written $\mathcal{E}(\rho) = \sum_{\nu=0,x,y,z} p_\nu \sigma_\nu \rho \sigma_\nu$ for a physical Pauli mixture. Expanding its action on the three axes and inverting gives

$$p_0 = (1 + \lambda_x + \lambda_y + \lambda_z)/4,$$
$$p_x = (1 + \lambda_x - \lambda_y - \lambda_z)/4,$$

with cyclic partners for p_y, p_z. Complete positivity demands all four numbers be nonnegative. Checking each $|\lambda_i| \le 1$ only checks lengths along single Bloch axes and, for this diagonal map, positivity on a

single qubit. It does not guarantee that the reconstructed operation is consistent on half of an entangled system.

For $\lambda = (0.9, 0.9, 0.5)$, every axis appears to shrink, yet $p_z = (1 - 0.9 - 0.9 + 0.5)/4 = -0.075$. Thus the exact diagonal unital channel inferred from those numbers is not completely positive. Finite measurement noise, neglected affine offsets or changing preparation could be responsible. Compare the fit with uncertainty regions and a constrained process model; simply forcing each measured contraction coefficient into $[-1, 1]$ cannot repair this violation. As a constructive control, $\lambda = (0.8, 0.7, 0.6)$ gives $p_0 = 0.775,\ p_x = 0.125,\ p_y = 0.075,\ p_z = 0.025$, all valid and summing to one. A simulator using these four random Pauli operations on entangled inputs then supplies a direct check of the estimated channel beyond single-qubit positivity.

Test how a coherence measurement depends on an ensemble

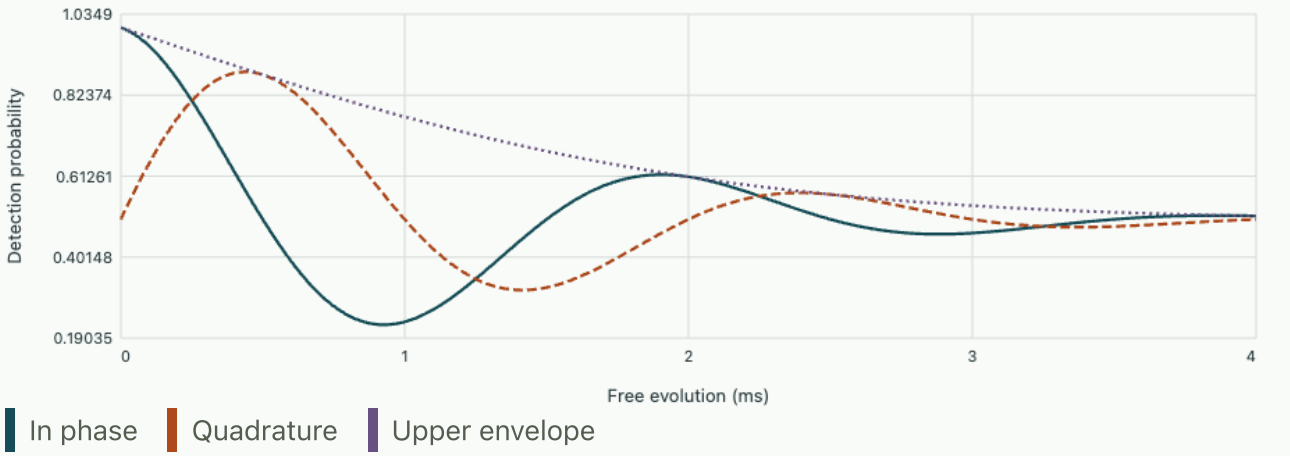
Compare quasistatic frequency spread with homogeneous decay in a Ramsey sequence. Identify which further measurement could distinguish the mechanisms.

Open this experiment online

Try the calculation

Mean cyclic detuning (Hz) **500**

Shot-to-shot rms spread (Hz) **80**



Contrast after 1 ms = 0.53455. The two probabilities at 0.50 ms are 0.5 and 0.87729.

Prove when two different event descriptions give the same channel

A given unconditioned quantum operation does not determine a unique set of environmental records. Suppose a set of Kraus operators K_j describes a channel. If an environment detector resolves an

alternative orthonormal basis related by unitary $U = (u_{\alpha j})$, its new system operators are $\tilde{K}_\alpha = \sum_j u_{\alpha j} K_j$. Summing the unnormalised conditional outputs gives

$$\sum_{\alpha,j,k} u_{\alpha j} u^*_{\alpha k} K_j \rho K_k^\dagger = \sum_j K_j \rho K_j^\dagger.$$

This calculation explains why environment-dependent state trajectories can be quite different while every unconditioned prediction on the system agrees. A simulator's choice of individual jump records only describes a particular monitoring scheme when the environment observable actually corresponds to those records.

Consider complete dephasing in the z basis. One representation is $K_0 = |0\rangle\langle 0|,\ K_1 = |1\rangle\langle 1|$, an ideal path measurement with its result forgotten. Rotate the two environment labels by the Hadamard matrix to obtain

$$\tilde{K}_+ = (I/\sqrt{2}),\ \tilde{K}_- = (\sigma_z/\sqrt{2}).$$

These describe equally likely identity or phase-flip operations. For input $|+x\rangle$, the first recorded instrument conditionally gives $|0\rangle$ or $|1\rangle$; the second gives $|+x\rangle$ or $|-x\rangle$. Both produce $I/2$ if their outcome is ignored. For initial $|0\rangle$ the first record is deterministic, while the second remains equally random and leaves the state unchanged up to a phase. Measuring only the system after discarding the labels cannot tell which set was used, but correlations with a retained environmental measurement can immediately tell them apart.

The same distinction is useful for a partially dephasing environment. If $|0\rangle, |1\rangle$ become correlated with environment states of real overlap c, the qubit's transverse Bloch vector contracts by c. For $-1 \le c \le 1$ a random phase-flip representation has probabilities $p_I = (1+c)/2,\ p_Z = (1-c)/2$. For $c < 0$ a fringe phase reversal accompanies partial contraction; a reported positive “visibility” loses that sign. One can separate a deterministic phase rotation from the magnitude of random or entangling dephasing if the phase reference is maintained. Over independent identical intervals a real c multiplies repeatedly; the coherence after n steps is c^n. If the same randomly chosen phase flip is instead held fixed throughout all intervals, successive errors are correlated and the even-step coherence can return to one. A single-step channel tomography cannot identify the environmental memory shared across steps.

Tell apart a memoryless phase-flip channel and a fixed hidden phase

Both devices individually implement $p_Z = 0.20$ over a calibrated interval for freshly prepared qubits. In device A the phase-flip draws at successive intervals are independent. In device B one draw is held fixed for both consecutive intervals. Predict $+x$ readout after one and two intervals for an initial $|+x\rangle$.

▼ Derive and check the result

One interval leaves $x = 1 - 2p_Z = 0.60$, so both devices give

$$P(+x) = 0.80.$$

Independence in A gives $x = 0.60^2 = 0.36$ after two intervals, or $P(+x) = 0.68$. For B, two identical phase flips multiply to identity whether the hidden choice was identity or Z, hence $P(+x) = 1$. In both devices the one-step process applied to an independently prepared ensemble is consistent with the measured completely positive map. Composing that map twice implicitly assumed fresh, independent environment conditions; a memory experiment tests exactly that assumption. Observing a revival thus need not imply an energy input, a negative probability or a nonunitary failure in the larger coupled system.

Distinguish dipole coherence from an excited-state population lifetime

Consider a single initially excited emitter in a memoryless zero-temperature environment with radiative population decay Γ and no additional dephasing. The probability density of emitting at time $t \geq 0$ is $\Gamma e^{-\Gamma t}$. A normalised single-photon temporal mode consistent with this ideal undistinguished radiative channel is $g(t) = \sqrt{\Gamma}\, e^{-\Gamma t/2} e^{-i\omega_0 t}$ on $t \geq 0$. Its field amplitude decays twice as slowly as the population probability. Use the Fourier convention $\tilde{g}(\omega) = (2\pi)^{-1/2} \int g(t) e^{i\omega t} dt$. The half-line exponential integral gives

$$|\tilde{g}(\omega)|^2 = \frac{\Gamma}{2\pi[(\Gamma/2)^2 + (\omega - \omega_0)^2]}.$$

It integrates to one in angular frequency. At a detuning of $\Gamma/2$ the intensity has half its peak, so the full angular-frequency linewidth is Γ, or $\Gamma/(2\pi)$ in hertz. Using the population decay directly as a field amplitude would predict a linewidth twice too large and the wrong wave-packet envelope. The transform-limited result requires a single radiative mode in the sense of a pure emitted temporal state and absence of additional phase randomisation. Extra pure dephasing broadens the spectrum while population decay may remain unchanged.

For $\Gamma^{-1} = 10.0$ ns the ideal linewidth is 15.9 MHz. Add independent homogeneous pure dephasing at $\Gamma_\phi = 2.0 \times 10^7\ \mathrm{s}^{-1}$. The first-order dipole correlation decays at $\Gamma/2 + \Gamma_\phi$, giving a full line width $(\Gamma/2 + \Gamma_\phi)/\pi = 22.3$ MHz. The photon detection time distribution can still have 10 ns decay. A measured line twice its nominal transform-limited width might indicate pure dephasing, but spatially varying frequencies, slow source drift or spectral resolution can also broaden the observed distribution. A lifetime measurement alone does not calibrate those effects.

The acquisition window has an independently calculable consequence. If only photons arriving before time T are retained, their normalised temporal mode is a truncated exponential divided by $\sqrt{1 - e^{-\Gamma T}}$ in this ideal pure-emission model. Its Fourier amplitude gains factor $1 - e^{-[\Gamma/2 - i(\omega - \omega_0)]T}$ in the numerator. At $T \ll 1/\Gamma$ the shape approaches the squared sinc of a short rectangular temporal window. Selecting only early photons can thus broaden the spectrum even though the emitter and its environment are unchanged. A time gate, an energy filter and unconditioned counting describe different optical preparations and different uncertainty budgets.

Infer the transform-limited fraction from a linewidth and test the model

A homogeneous line has measured weak-drive full width 25.0 MHz and population lifetime 10.0 ns. Compute T_2 and the implied pure dephasing rate assuming ideal Lorentzian broadening. Which second measurement would discriminate this inference from static sample inhomogeneity?

▼ **Derive and check the result**

With angular half width $1/T_2$, the frequency full width is $1/(\pi T_2)$, so $T_2 = 12.73$ ns. Subtracting $1/(2T_1) = 5.00 \times 10^7\ \mathrm{s}^{-1}$ from $1/T_2 = 7.85 \times 10^7\ \mathrm{s}^{-1}$ gives

$$\Gamma_\phi = 2.85 \times 10^7\ \mathrm{s}^{-1}.$$

The coherence relative to the radiative limit is

$$T_2/(2T_1) = 0.637.$$

Compare a time-domain echo, a single-emitter linewidth or spectra over short acquisition segments with the ensemble line. A recoverable component of contrast or slowly wandering line centres would indicate a static or slow contribution instead of assigning every megahertz to homogeneous phase damping.

Tutorial · distinguish different sources of coherence loss

Recover a pure-dephasing rate and assess a conflicting fit

A population measurement gives

$$T_1 = 30.0\,\mu\mathrm{s};$$

homogeneous Ramsey contrast gives

$$T_2 = 24.0\,\mu\mathrm{s}.$$

Find T_ϕ. What would a reported $T_2 = 80\,\mu$s imply in the same model?

▼ **Work through the solution**

Subtract $1/(60\,\mu\mathrm{s})$ from $1/(24\,\mu\mathrm{s})$ to get $\Gamma_\phi = 0.0250\,\mu\mathrm{s}^{-1}$, hence $T_\phi = 40.0\,\mu$s. Replacing 24 by 80 yields a negative pure dephasing rate, incompatible with the specified independent time-homogeneous channels. Check the fitted envelope, uncertainties, thermally shifted equilibrium population and whether

the two times were measured under comparable conditions. A negative fitted number is evidence about the model or data interpretation, not a physically negative random-event rate.

Calculate an inefficient no-click update

The initial excited probability is one half, the decay probability for an initially excited qubit during the window is 0.40, and a true emitted excitation is detected with efficiency 0.75. Find the probability of no click, its conditional excited fraction and the unconditioned population.

▼ **Work through the solution**

A real emission occurs with probability $0.5 \times 0.40 = 0.20$; a click is seen with probability 0.15, so no click has probability 0.85. Remaining excited and no click occurs with probability $0.5(1 - 0.40) = 0.30$. Conditional excitation is $0.30/0.85 = 0.3529$; the unconditioned excitation is 0.30 since clicked branches are ground. For an initially coherent preparation the no-click off-diagonal term must also be updated by grouping the undetected-emission and true-no-emission Kraus outcomes. Simply replacing p by detector efficiency times p would predict the wrong physical excitation loss.

Diagnose a nonphysical time step and repair it

Apply one forward-Euler step to $\dot{p}_1 = -\Gamma p_1$ with initially excited state and $\Gamma \Delta t = 1.2$. Compute the prediction and an exact completely positive alternative. Would enforcing unit trace fix Euler's failure?

▼ **Work through the solution**

Euler gives $p_1 = 1 - 1.2 = -0.2,\ p_0 = 1.2$, already with exact trace one. The finite-time amplitude-damping channel has $p = 1 - e^{-1.2} = 0.6988$, giving $p_1 = 0.3012$. For this simple model the exact Kraus step can be applied repeatedly at arbitrary duration. In a driven model an operator-splitting scheme with unitary Hamiltonian substeps and completely positive damping substeps maintains positivity while introducing a controllable splitting error. Check convergence by halving the step and comparing with an independent observable. Renormalising trace or clipping a negative eigenvalue changes the numerical trajectory without diagnosing the original time-step error.

Further reading

- Cambridge · open systems, measurements and the Lindblad equation
- Course investigation · density matrices and dephasing
- Course investigation · driven coherent response

Applied study

Distinguish stability, positivity and accuracy when stepping a model forward in time.

How does a short step change a long trajectory?

A well-mixed reactant A converts to B at rate kA. In a closed vessel $\dot{A} = -kA$ and $\dot{B} = kA$. For $A(0) = 1$, the exact concentration of A is e^{-kt} in the chosen units, while $A + B$ remains constant. A numerical scheme can preserve the total exactly and still put the wrong amount in each species.

Integrating the differential equation over one time interval gives

$$A_{n+1} = A_n - k \int_{t_n}^{t_{n+1}} A(t)\, dt.$$

Replacing the integrand by its left endpoint yields forward Euler, $A_{n+1} = (1 - z)A_n$ with $z = kh$. Using the unknown endpoint instead gives backward Euler, $A_{n+1} = A_n/(1 + z)$. Each method replaces the exact amplification factor e^{-z} by a simpler function $R(z)$.

Distinguish stability, positivity and accuracy

For the decay mode, forward Euler damps successive values when $|1 - z| < 1$, or $0 < z < 2$. Keeping an initially positive concentration nonnegative requires the stronger condition $z \le 1$. At $z = 1.5$ the forward Euler values alternate in sign as their magnitudes shrink. Backward Euler has $R(z) = 1/(1 + z) > 0$ for every positive step, although a large step can still give a poor quantity of reactant.

Taylor expansion gives

$$e^{-z} = 1 - z + z^2/2 + \cdots.$$

Euler misses a term of order h^2 per step. Over a fixed time interval the number of steps is proportional to $1/h$, producing first-order global error under stable propagation. Classical four-stage Runge–Kutta (RK4) evaluates intermediate slopes. On this linear problem its factor is

$$R_4 = 1 - z + z^2/2 - z^3/6 + z^4/24.$$

It first differs from the exponential at order h^5, giving fourth-order global convergence.

Test a hand calculation against the exact decay

Take $k = 0.8$ s⁻¹ and $h = 0.625$ s, giving $z = 0.5$. After two steps, the exact fraction remaining is

$$e^{-1} = 0.36788.$$

Forward Euler gives $0.5^2 = 0.25$ and backward Euler gives

$$(1/1.5)^2 = 0.44444.$$

Both give positive, decaying sequences, but one consumes too much A and the other too little. The signs follow from $1 - z < e^{-z} < 1/(1+z)$ for $z > 0$.

RK4 gives $R_4(0.5) = 0.6067708$, hence $A_2 \simeq 0.36817$. Comparing all methods at the same physical time separates time error from a simple difference in how long the reaction has proceeded.

Problem class

Keep the chemical interpretation

With $k = 0.8$ s⁻¹, find the forward Euler upper step for nonnegative concentrations and the upper stability boundary. What happens exactly at the first bound?

▼ Show the calculation

Nonnegativity gives

$$h \le 1/k = 1.25\,\mathrm{s}.$$

The decaying stability interval has $h < 2/k = 2.50$ s. At $h = 1.25$ s the method sends all remaining A to B in one step. The exact solution still contains e^{-1} of the starting A. Thus respecting a nonnegativity bound alone permits considerable error.

Measure an order from two runs

Suppose a step halving reduces the largest RK4 error over the same time interval from 2.00×10^{-8} to 1.22×10^{-9} mol/m³. Estimate the observed order. Explain what a mass-balance test would miss.

▼ Show the calculation

Assuming $E_h \approx Ch^p$, take $p = \log_2(E_h/E_{h/2}) \approx 4.03$. For these equations the stage increments of A and B are equal and opposite. Their sum can remain at the initial total throughout either run even when both concentrations have sizable individual errors.

ONESIMULATION LABORATORY

Compare time-step error with conservation

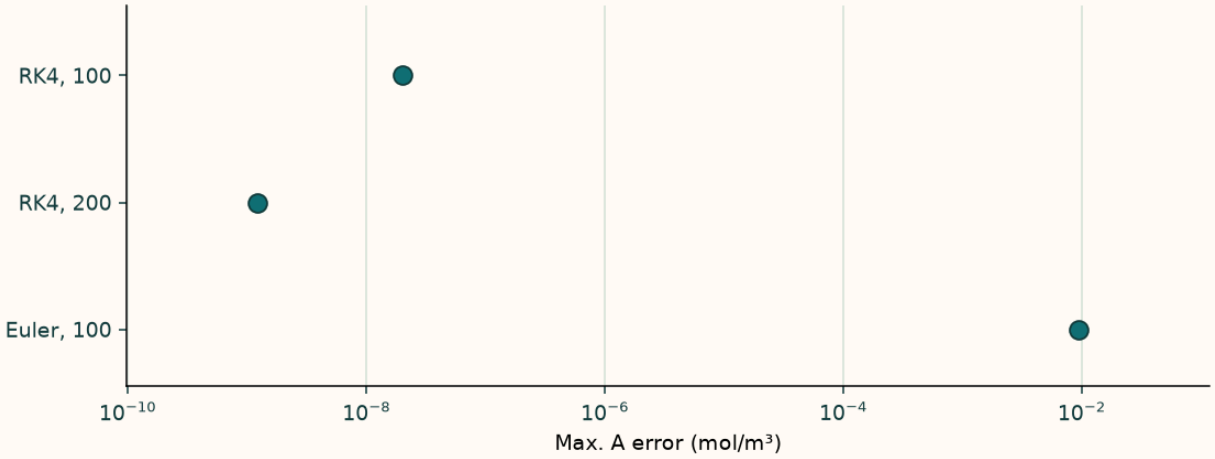
Compare recorded OneSimulation runs on this page. To test your own inputs, download the Python model and run the model in an activated OneSimulation environment.

Run the isothermal reaction from $A = 1,\ B = 0$ mol/m³ to $t = 5/k = 6.25$ s. Measure the maximum A error over all stored times and the drift of $A + B$. First double the number of RK4 steps. Then restore 100 steps and change the time integrator. Download the Python experiment · OneSimulation first-order decay model.

```
python ordinary-differential-equations.py
python ordinary-differential-equations.py --steps 200
python ordinary-differential-equations.py --method euler
```

Errors in species A and in the conserved sum over five decay times.

Method, steps	Final A (mol/m³)	Max. A error (mol/m³)	Max. drift (mol/m³)
RK4, 100	0.00673794883	2.00×10^{-8}	4.44×10^{-16}
RK4, 200	0.00673794711	1.22×10^{-9}	3.33×10^{-16}
Euler, 100	0.00592052922	9.39×10^{-3}	4.44×10^{-16}



Halving the RK4 step reduces the largest concentration error by about sixteen. The logarithmic axis also shows the much larger Euler error at 100 steps.

As a fourth run, try `--method euler --steps 5`. This lands exactly on the nonnegativity limit. Predict both the final concentration and the largest error before running it.

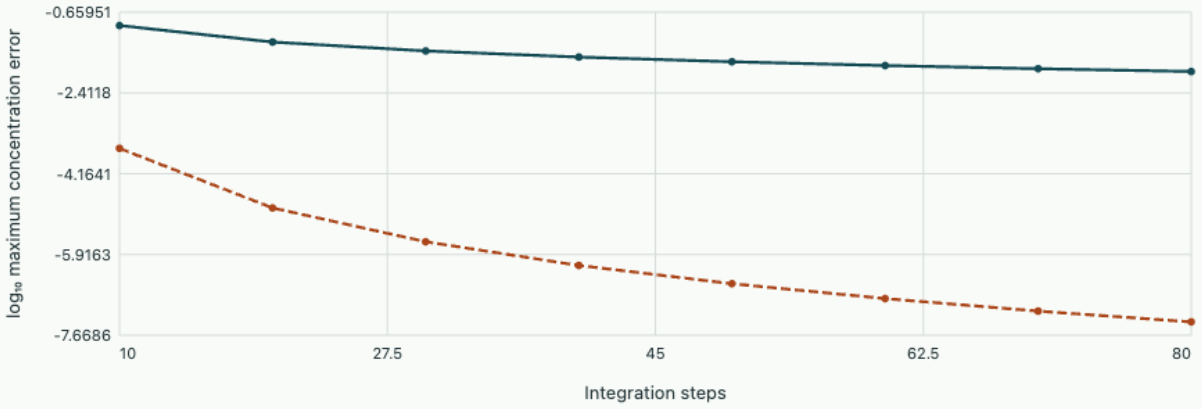
Change parameters · Change the laboratory inputs

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

Reaction rate k (1/s)	**0.8**
End time (s)	**6**
Maximum steps	**80**



The model transfers A to B with A + B = 1 and initial A = 1. Both methods conserve that sum; time accuracy differs. RK4 costs four slope evaluations per step.

Problem sheet 1 · Predict when a conservative calculation loses all reactant

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.

Embedded error estimate

$$e = y_{n+1}^{[p]} - y_{n+1}^{[p-1]}$$

Two formulas share stages and estimate a local error. The controller also needs absolute and relative scales for each component.

Derivation · Adaptive steps and event locations

Event transversality

$$\dot{g} = \partial_t g + \nabla g \cdot f$$

For a simple event this directional derivative is nonzero. Near grazing, a small state error can create a much larger error in the event time.

Derivation · Locating a grazing event between time steps

Local truncation error

$$d_{n+1} = y(t_{n+1}) - \Phi_h(y(t_n))$$

This is the one-step defect from exact starting data. Some conventions divide it by h. State the convention before comparing local and global orders.

Derivation · Initial values and propagated error

Shooting residual

$$F(s) = B(y(t_f; s))$$

Here s encodes missing initial data and B the final boundary condition. An unstable trajectory map can make the residual badly conditioned even with an accurate initial-value solver.

Derivation · Shooting for two-point boundary values

Stability function

$$y_{n+1} = R(h\lambda) y_n$$

On the scalar linear test equation, absolute stability depends on the point h times lambda. It gives only part of the story for a nonnormal matrix or nonlinear problem.

Derivation · Absolute stability and stiffness

Symplectic map

$$(D\Phi)^{\mathsf{T}} J (D\Phi) = J$$

This is preservation of the canonical symplectic form. It provides control of long-time structure without guaranteeing that every trajectory has the exact phase.

Derivation · Energy and geometric integration

Further reading

- Reconstruct the symplectic Euler and Störmer–Verlet updates in section 1. Sections 3–5 explain why reversing or composing their substeps changes the order. Ernst Hairer · TU München, Symplectic integrators (PDF)
- Work through chapter 6 on initial values, then turn to chapter 10 on boundary values. Reuse an exact solution to distinguish a poorly conditioned shooting map from an inaccurate time step. Driscoll and Braun · Fundamentals of Numerical Computation

Compare specific topics with the university sources