Nonlinear equations

Bracket and certify roots, derive Newton and secant updates for coupled equations, and follow branches and active constraints as parameters change.

Read the subject in depth

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

- **Nonlinear systems and iterative methods**
  Begin with a residual and ask which solution is accessible. Study how curvature, scaling and conditioning affect a computed root.

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. **Equilibria, branches and stability**
   Turn a physical stationarity condition into a nonlinear equation. Trace the solutions through loss of stability and decide which roots survive a perturbation.
2. **Nonlinear inference and implicit equations**
   Investigate what a residual and a local derivative measure in inverse fitting, discretised boundary equations and an implicitly advanced stiff mode.

How the course fits together

1. **Chapters 1–4**
   A sign-changing bracket certifies existence, a contraction supplies an error bound, and tangents or secants accelerate local convergence. Replace the scalar derivative by a Jacobian for simultaneous equations.
2. **Chapters 5–8**
   Globalise Newton steps, follow parameter-dependent roots, and examine polynomial sensitivity. Interval bounds can certify that a computed root is the only one inside a chosen box.
3. **Chapters 9–12**
   Approximate Jacobians or their linear solves when derivatives are expensive. Residual minimisation and complementarity address measurement mismatch and switching physical constraints.

Contents

1. Chapter 1 · Existence, brackets and error certificates

2. Chapter 2 · Fixed-point iteration and contraction
3. Problem sheet 1 · Isolating and iterating roots
4. Chapter 3 · Newton and secant updates
5. Chapter 4 · Newton’s method for coupled balances
6. Problem sheet 2 · Tangents and Jacobians
7. Chapter 5 · Line searches and trust regions
8. Chapter 6 · Continuation and folds
9. Problem sheet 3 · Safeguards and continuation
10. Chapter 7 · Polynomial roots and coefficient sensitivity
11. Chapter 8 · Root enclosures with interval Newton
12. Problem sheet 4 · Root sensitivity and certificates
13. Chapter 9 · Secant information for systems
14. Chapter 10 · Solving Newton equations inexactly
15. Problem sheet 5 · What an approximate Jacobian knows
16. Chapter 11 · Gauss–Newton and residual curvature
17. Chapter 12 · Inequalities and active constraints
18. Problem sheet 6 · Misfit and one-sided equilibrium
19. Chapter 13 · Why a flat zero slows Newton's method
20. Chapter 14 · Eliminating a fixed point's leading geometric error
21. Chapter 15 · Deciding how far to trust a linearised residual
22. Chapter 16 · A physically positive root across six decades
23. Chapter 17 · Finding a turning point without using the load as a coordinate
24. Chapter 18 · Newton's method for an unknown covariance square root
25. Problem sheet 7 · Project · Locate contact and calibrate a misleading residual
26. Problem sheet 8 · Find a periodic orbit when the starting point has no preferred phase
27. Nonlinear systems and iterative methods
28. Equilibria, branches and stability
29. Nonlinear inference and implicit equations
30. Applied study
31. Notation and assumptions
32. Further reading

Chapter 1 · Existence, brackets and error certificates

Decide whether a scalar balance has a root, isolate it with a sign-changing interval and translate an equation residual into a bound in physical units.

FROM FIRST PRINCIPLES

Turn a sign change into a location certificate

Suppose f is continuous on $[a, b]$ and the endpoint values have opposite signs. By the intermediate value theorem at least one zero lies between them. Evaluate the midpoint $m = (a + b)/2$. If $f(m) = 0$, the equation has been solved. Otherwise exactly one of the two half intervals still has opposite signs at its endpoints. Keep that half and repeat. From an initial width $w_0 = b - a$, the width after n such updates is

$$w_n = \frac{w_0}{2^n},$$
$$|m_n - r| \le \frac{w_n}{2} = \frac{w_0}{2^{n+1}}$$

for at least one enclosed root r, where m_n is the midpoint of the current interval. To target an absolute location error $\varepsilon > 0$, a sufficient number of halvings is

$$n \ge \max\left(0, \left\lceil \log_2 \frac{w_0}{2\varepsilon} \right\rceil\right).$$

The certificate uses reliable endpoint signs and continuity. A jump can change signs without crossing zero, while a tangential root can exist without changing signs. Monotonicity is optional for existence but can establish that the enclosed root is unique.

Background · Bracket

$$f(a)f(b) < 0$$

Continuity is needed to infer a zero between endpoints. A root of even multiplicity may exist without a sign change.

DERIVATION AND APPLICATIONS

Certify the location obtained by intersecting an endpoint secant

A sign bracket guarantees an enclosed root. If derivatives are also bounded, its endpoint secant can produce a sharper certified location without a sequence of additional function evaluations. Assume $f \in C^2[a, b]$, opposite endpoint signs, $|f'| \ge m > 0$, and $|f''| \le M$. The derivative cannot change sign

on the interval, so there is exactly one root r. Form the line joining the endpoint values and solve it for zero,

$$s = a - \frac{f(a)(b-a)}{f(b)-f(a)}, \qquad a < s < b.$$

The line vanishes at s, so the true residual there is exactly the error made by interpolating f by that line. The two-node interpolation remainder and the mean value theorem applied between s and r give successively

$$|f(s)| \le \tfrac{1}{2} M (s-a)(b-s), \qquad |f(s)| = |f'(\zeta)|\,|s-r| \ge m|s-r|.$$

Combining the inequalities gives a certified enclosure centred on the proposed root approximation,

$$|s-r| \le \frac{M}{2m}(s-a)(b-s) \le \frac{M}{8m}(b-a)^2.$$

Intersect this enclosure with the original bracket to retain both guarantees. The constants must hold over the whole interval; an observed derivative at one point supplies neither bound by itself. When the function is linear the curvature bound vanishes and the secant solves the problem exactly.

WORKED EXAMPLE

Enclose a square root from two evaluations

For $f(x) = x^2 - 2$ on $[1.4, 1.5]$, the endpoint values are -0.04 and 0.25. The secant gives $s = 1.4137931$. Since $M = 2$ and $m = 2.8$, the sharper bound is

$$(s - 1.4)(1.5 - s)/2.8 = 0.000425.$$

The true root 1.4142136 indeed differs by 0.0004205, and the location guarantee is much tighter than half the original width.

SECOND APPLICATION

Determine the bracket width needed for a certificate

Another calibration guarantees $m = 4$ and $M = 3$ throughout the working interval. To certify secant location error 10^{-4}, the uniform width criterion gives

$$b - a \le \sqrt{8(4)10^{-4}/3} = 0.0327.$$

Starting at width 0.20, three bisections give 0.025. Its worst secant error is then at most $3(0.025)^2/32 = 5.86 \times 10^{-5}$.

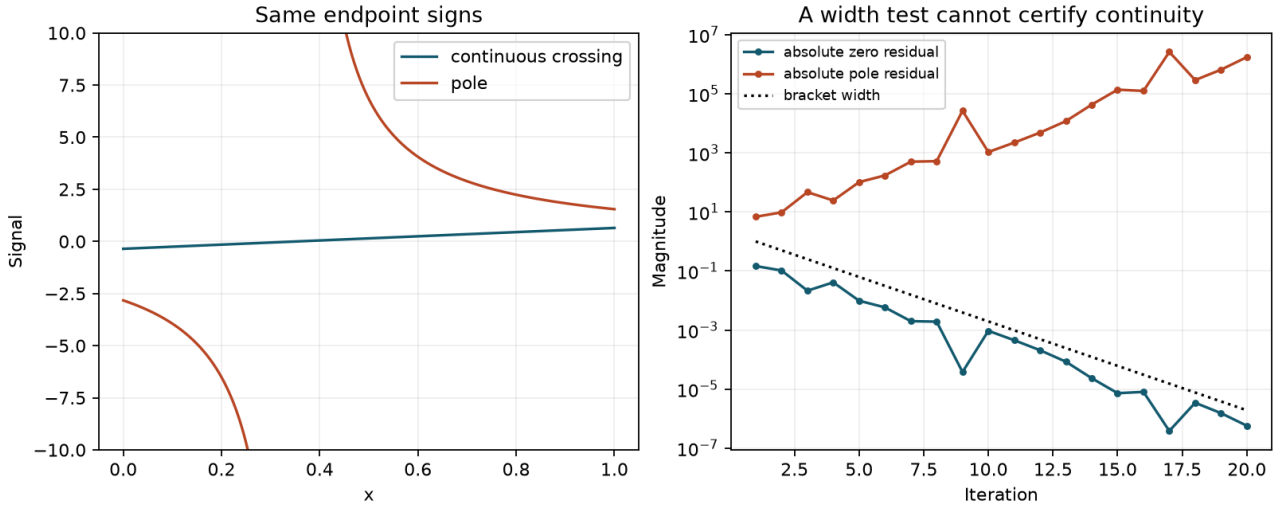
Start from this chapter. Use the previous temperature and the ambient temperature as candidate endpoints. Check the residual signs before comparing the numerical solution of the implicit stage.

OneSimulation laboratory · Locate the root of an implicit cooling step

A bracket can collapse onto a pole

A measuring circuit returns a signed response as a control coordinate x traverses the unit interval. Someone proposes logging only the sign and using bisection to locate the point at which the response vanishes. How much can those signs actually certify? Let $a = \sqrt{2}/4$. Compare two hypothetical responses $f(x) = x - a$ and $g(x) = 1/(x - a)$. They have the same sign at every sampled point except the singular coordinate itself. Only the first has a zero. The second is undefined at a and never vanishes.

Start with endpoints zero and one. On every trial the bisection coordinate, sign decision and resulting interval will be identical for these two responses. This follows by induction: identical current intervals have the same midpoint, the midpoint signs coincide, and the same endpoint is replaced. The shrinking width therefore tracks a crossing in the sign pattern. Continuity on the entire retained interval is the extra information the intermediate value theorem requires to guarantee a zero.



Both searches take the same path through parameter space. Their measured response magnitudes evolve in opposite ways. Individual midpoint residuals need not be monotone even though the bracket width always decreases.

Include the amplitude of the response in the notebook's log. At any midpoint m, excluding an exact tie, the product $|f(m)||g(m)|$ equals one. As the bracket contracts about a, the first magnitude tends to zero and the second grows without bound. If the first residual at a sample is 2^{-14}, the reciprocal response there is

$$2^{14} = 16384.$$

A software check that accepts only the final width would treat these two very different observations as equally persuasive. A width criterion plus a scaled residual supplies stronger evidence, but finite samples alone cannot establish continuity for a completely unknown device between measurements.

In a physical instrument saturation may hide the growing amplitude. Then inspect the forward model and its domain: a quotient can change sign when its denominator passes through zero. Record the numerator, denominator, clipping events and units separately where possible. Scaling a residual by a quantity that also diverges can disguise the singularity a second time. A well-designed stopping test makes its assumptions about observability explicit.

A discontinuity also has consequences for existence guarantees based on intervals. If a constitutive law supplies an excluded denominator interval or explicitly brackets an undefined point, split the physical domain there before asking a root finder to operate. This small piece of model knowledge is stronger than an arbitrarily long list of sign samples.

Suppose one accepts whenever the bracket width is below 10^{-5}. How many halvings suffice on a unit interval, and what follows for the reciprocal model at the midpoint of that final bracket?

▼ **Work through the solution**

Seventeen halvings make the width $2^{-17} \simeq 7.63 \cdot 10^{-6}$. Since a remains in the interval, its distance to the midpoint is at most half the remaining width. Thus the reciprocal response has magnitude at least 2^{18}, unless the midpoint exactly hits the undefined point. A tiny horizontal uncertainty has here located a singularity with high precision.

Change parameters · Explore this notebook example

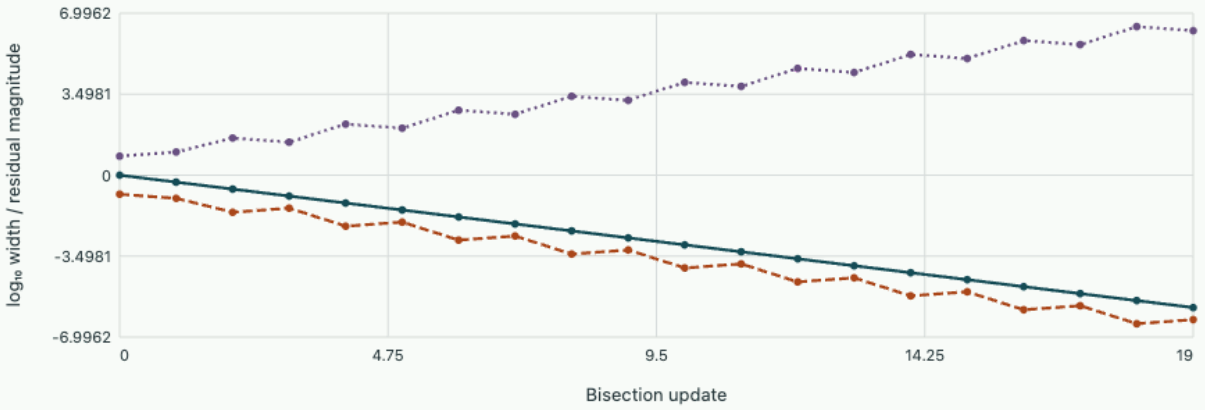
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Crossing location **0.35**

Bisection rounds **20**



The sign-based brackets are identical for x – a and 1/(x – a), but only the continuous function has a root. The pole residual increases while the bracket contracts.

Open editable notebook Download .ipynb

Establish the meaning of a root

A source supplies heat at a known rate, while a cooling surface loses heat at a rate that increases with temperature. A steady state equates the two rates. Writing generation minus loss as $F(T)$ turns a physical balance into one scalar equation. Before selecting a numerical method, specify an admissible temperature interval and state which assumptions make the relation continuous there. A temperature-dependent material table with a discontinuous switch may require separate treatment of each regime.

If F is continuous and $F(a)F(b) < 0$, the intermediate value theorem guarantees at least one root in (a, b). It does not count the roots. If F' has one strict sign throughout, the function is strictly monotonic and the root is unique. An even-multiplicity root can touch zero without changing sign, as

$F(x) = (x-1)^2$ does. Conversely, a sign change across the pole of $1/x$ guarantees nothing. An interval and its regularity assumptions are part of the mathematical answer.

Turn existence into a controlled search

Evaluate the midpoint $m = (a+b)/2$. If $F(m) = 0$ the machine has found a zero of its evaluated residual. Otherwise keep the half whose endpoint values have opposite signs. Continuity carries the existence argument through each subdivision. After n halvings the width is $2^{-n}(b-a)$, so the midpoint has distance at most $2^{-(n+1)}(b-a)$ from a root. A predetermined number of evaluations can therefore guarantee a target error in the independent variable even for an unknown local slope.

For $F(x) = x^3 + x - 1$ on $[0,1]$, the derivative $3x^2+1$ is positive, while the endpoints have opposite signs. Test 0.5, giving -0.375, and keep $[0.5, 1]$. Testing 0.75 gives 0.171875, narrowing the bracket to $[0.5, 0.75]$. The next midpoint 0.625 gives approximately -0.13086, hence $[0.625, 0.75]$. After ten halvings the midpoint is guaranteed within $1/2048$, independent of how the subsequent signs turn out. A function evaluation whose uncertainty covers zero cannot safely be classified as strictly positive or negative without additional information.

Translate a balance into a physical uncertainty

Let r be the root and suppose $|F'| \geq m > 0$ on a bracket containing both r and a candidate x. By the mean value theorem $F(x) - F(r) = F'(\xi)(x-r)$, so $|x-r| \leq |F(x)|/m$. For the cubic, $m = 1$ is valid globally, and within $[0.625, 0.75]$ the sharper bound $m = 1 + 3(0.625)^2$ applies. A balance mismatch can thus yield an informative location estimate when the slope is well separated from zero.

Suppose instead the residual of a thermal balance is expressed in watts and its slope is at least 2 W/K. A 0.1 W mismatch guarantees at most 0.05 K error under the exact model. Reporting a temperature tolerance of 0.1 solely from that residual would silently change units. Near a fold the slope can become arbitrarily small, so a tiny mismatch can coexist with substantial uncertainty about the state. In that case report the bracket, measurement uncertainty and nearby alternatives along with the small residual.

Calibrate the stop against a noisy measurement

Suppose temperatures 300 K and 302 K bracket a steady heat balance, and the residual slope is known to exceed 4 W/K. The net power sensor is accurate only to ±0.2 W. A reported mismatch of +0.1 W may have either sign, so it cannot safely choose a half-interval. Its magnitude plus the uncertainty is at most 0.3 W, however. The slope bound then implies distance at most 0.075 K from the exact balance under the model.

This gives two ways to use the same observation. If its error interval excludes zero, the sign can tighten a bracket. If it includes zero, a magnitude bound can still constrain the root location. A stopping rule that requires a residual below the instrument’s resolution might keep iterating without improving the physical estimate. Decide the temperature accuracy of interest first, then choose an appropriate balance test.

Background · What does a sign test mean at an instrument's resolution?

A calibrated instrument reports residual $m(x)$ with a guaranteed absolute error at most η. A positive report establishes a positive physical residual only if $m(x) - \eta > 0$. A certified bracket therefore uses intervals whose entire sign is known. If an interior interval contains zero, assigning its sign from the midpoint alone can discard the true root.

For a continuously differentiable monotone residual with $|f'| \geq s > 0$ between the candidate and a known root, the mean-value theorem also supplies a distance estimate:

$$|x - x_*| \leq \frac{|m(x)| + \eta}{s}.$$

This estimate explains why requesting more digits from bisection cannot settle a flat, noisy balance. Repeating observations may reduce random error; a guaranteed calibration offset needs independent information.

Preserve a bracket while using slope information

A secant through the two endpoints of a sign-changing bracket usually crosses the axis at a more informative location than the midpoint. For continuous f, opposite signs at a, b place the linear-interpolant root $c = [af(b) - bf(a)]/[f(b) - f(a)]$ strictly inside the bracket. Replace whichever endpoint has the same sign as $f(c)$. This false-position iteration preserves existence just as bisection does. It differs in one significant respect: without a further safeguard, the retained bracket width need not shrink by a prescribed factor.

Suppose f is strictly increasing and convex and has root r. Its chord lies above the graph. At the chord root c, therefore, $f(c) \leq 0$, which places $c \leq r$. Every false-position update may replace the lower endpoint while leaving the upper one fixed. The false-position estimates themselves can still approach the root. But a stopping rule requiring a narrow bracketing interval can then fail to terminate. The distinction separates convergence of a point estimate from convergence of both bounds in a certified interval.

Prevent a distant high-value endpoint from pinning the search

For $f(x) = x^{10} - 1$ on $[0, 2]$, the endpoint values are −1 and 1023. The first false-position estimate is $2/1024 = 1/512$, with function value indistinguishable from −1 to ordinary displayed precision. The second estimate lies near $1/512 + (2 - 1/512)/1024 \approx 0.003904$, although the true root is one. Both steps satisfy the strict interior condition; neither meaningfully resolves the enormous function variation near the right endpoint. Bisection happens to land exactly on the root here, making the consequences of this imbalance especially easy to measure.

One safeguard accepts a secant candidate only if it lies in the central half of the current bracket. Otherwise it takes the midpoint. With exact reliable signs, accepting any point between the one-quarter

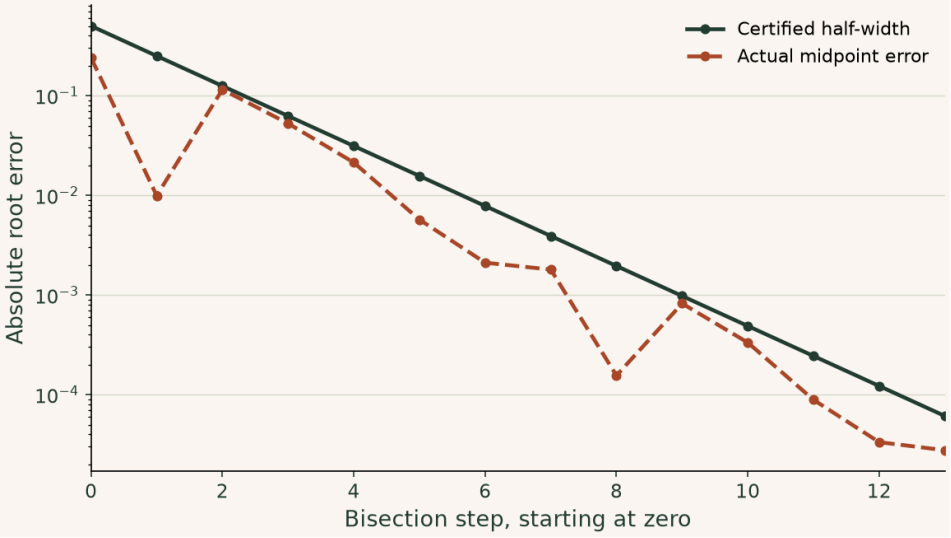
and three-quarter positions guarantees the retained width is at most three quarters of its old value. A midpoint fallback cuts it in half. Thus after k accepted or fallback steps the bracket width is bounded by $(3/4)^k(b-a)$, whether or not the slope model reflects the underlying function.

To test the policy away from this exact midpoint coincidence, shift the equation to $x^{10} - 1.1 = 0$ on the same interval. The initial chord estimate becomes $2(1.1)/1024$, below the central half, so the method takes one and obtains residual −0.1. The next bracket is $[1, 2]$, immediately discarding the whole range of tiny false-position candidates. A steeply varying function still calls for careful arithmetic in forming the chord, especially if endpoint values approach overflow. If field measurements cannot determine the sign at an accepted point, none of these purely geometric safeguards certifies which side to retain.

The shifted root is $1.1^{1/10} \approx 1.00958$, so the first midpoint is already close despite keeping a relatively wide bracket. This again separates observed point accuracy from the available certificate. Subsequent safeguarded steps can tighten the certificate without relying on advance knowledge of this exact value.

Compare a midpoint's actual error with its certificate

Bracket the positive root of $x^3 - 2$ on [1,2]. At each step record both the midpoint's distance to the analytic cube root and the half-width of the current sign-changing interval. Only the latter is known in a general application.



The half-width halves at every update. The actual error can rise after a lucky midpoint; selecting the smallest observed residual from a few iterates is a different stopping rule from certifying an interval.

▶ **Read values from the plot**

Change parameters · Compare a midpoint's actual error with its certificate

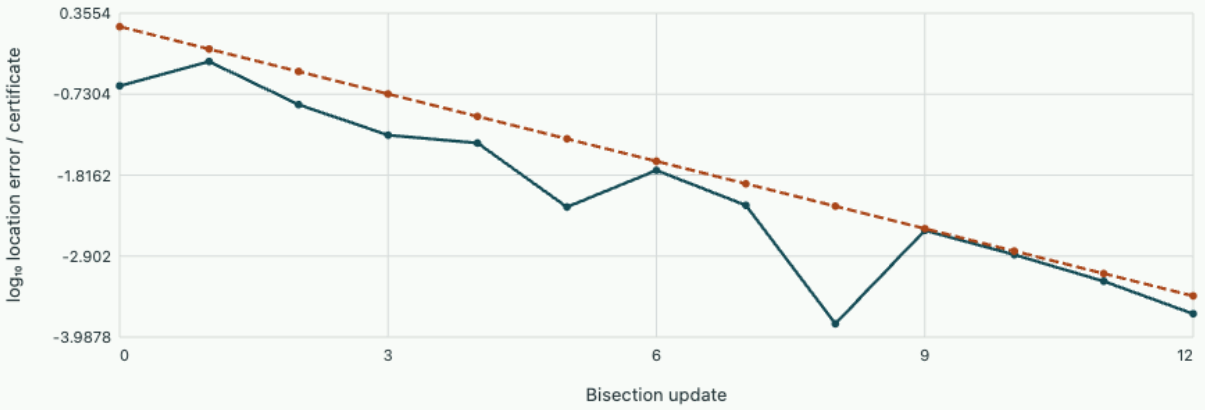
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Positive cubic target a **2**

Bisection updates **12**



Actual midpoint error Certified half-width

Root reference 1.2599. The bracket [0, 3] encloses every offered target; a lucky midpoint can have a much smaller error than the guaranteed half-width.

Chapter 2 · Fixed-point iteration and contraction

Select an iteration map for a nonlinear balance, prove convergence on an invariant interval and use an observed increment to bound the remaining error.

FROM FIRST PRINCIPLES

Bound fixed-point error by a geometric series

Rearrange an equation as $x = g(x)$ and repeatedly set $x_{k+1} = g(x_k)$. Suppose a closed interval contains the initial point, maps into itself under g, and has a uniform slope bound $|g'| \le q < 1$. The mean value theorem gives

$$|x_{k+1} - x_k| \le q|x_k - x_{k-1}|.$$

Successive increments therefore form a geometrically bounded series. Their sum converges, and the limit lies in the closed interval and is a fixed point by continuity. Two distinct fixed points would contradict the same contraction inequality. At step $k \ge 1$, summing the remaining increments gives the a posteriori estimate

$$|x_k - x_*| \le \sum_{j=1}^{\infty} q^j |x_k - x_{k-1}| = \frac{q}{1-q}|x_k - x_{k-1}|.$$

A small observed update is meaningful only in relation to the contraction factor. When q is close to one, many small updates may still remain.

Background · Contraction

$$\|G(x) - G(y)\| \le q\|x - y\|, \quad q < 1$$

The map must also keep the chosen complete region invariant to obtain the standard fixed-point guarantee.

DERIVATION AND APPLICATIONS

Choose a relaxation parameter from verified slope bounds

A fixed-point iteration can oscillate if the derivative of its update is negative, or move slowly if it stays near one. Suppose g is differentiable on an interval with $a \le g'(x) \le b < 1$, where a may be less than -1. Form a relaxed map $G_\omega(x) = x + \omega(g(x) - x)$ with $\omega > 0$. Every fixed point of one map is a fixed point of the other. The derivative transforms affinely,

$$G'_\omega(x) = 1 - \omega[1 - g'(x)], \qquad 1 - g'(x) \in [1 - b, 1 - a].$$

Both bounds inside the final interval are positive. The largest allowed magnitude of G'_ω is attained at one of its endpoints. Equalizing the signed endpoint derivatives with opposite signs gives the optimal common parameter and bound,

$$1 - \omega(1 - b) = -[1 - \omega(1 - a)],$$

$$\omega_* = \frac{2}{2 - a - b},$$

$$q_* = \frac{b - a}{2 - a - b}.$$

The denominator exceeds the numerator because $b < 1$. A usable global contraction certificate also requires the relaxed map to preserve the chosen domain. If a closed interval is centred at c with radius R, the slope bound gives a convenient sufficient condition:

$$|G_\omega(x) - c| \le q_* R + |G_\omega(c) - c| \le R.$$

This last condition can fail even when the derivative bound is excellent. In such a case the slope bounds outside the verified interval would be unknown.

WORKED EXAMPLE

Stabilize a map with an oscillatory steep branch

Suppose $g' \in [-3, 0.5]$ on a verified domain. The unrelaxed iteration can amplify error by a factor of three. Choose $\omega_* = 2/4.5 = 4/9$, giving endpoint derivatives $-7/9$ and $7/9$. If the interval radius is 2 and the relaxed centre moves by 0.20, the domain check gives

$$2(7/9) + 0.20 = 1.76 < 2.$$

Both invariance and contraction are then certified.

SECOND APPLICATION

Accelerate a map with uniformly positive slopes

In a second model $g' \in [0.6, 0.8]$. The optimal relaxed map uses $\omega_* = 2/0.6 = 10/3$, an extrapolation beyond the ordinary update. The slope bound falls to $q_* = 1/3$. Assuming the domain condition is independently verified, five steps reduce an initial location error of at most 0.10 to $0.10/3^5 = 0.000412$. With the unrelaxed slope bound the corresponding guarantee is only 0.0328.

An equation can have several iteration maps

A fixed point satisfies $x = g(x)$, but the rearrangement chosen for g is an algorithmic decision. A temperature balance might be written as a source update, a loss inversion or a relaxed combination of both. Every rearrangement can represent the same root and yet produce a different sequence from the same initial measurement. To judge the proposed sequence, examine how perturbations are propagated before investing in detailed model evaluations.

Let g map a closed interval I into itself, and suppose $|g'(x)| \le q < 1$ there. Then $|g(x) - g(y)| \le q|x - y|$ by the mean value theorem. Starting from $x_0 \in I$, all iterates stay in I. Successive changes shrink at least geometrically:

$$|x_{k+1} - x_k| \le q^k |x_1 - x_0|.$$

Summing the remaining geometric tail shows that the sequence is Cauchy. Its limit belongs to the closed interval, and continuity gives $r = g(r)$. Two distinct fixed points would violate the strict contraction inequality, establishing uniqueness.

Estimate what remains after an observed update

The tail estimate gives

$$|r - x_k| \le |x_{k+1} - x_k| / (1 - q).$$

If only the last completed increment is available,

$$|r - x_k| \le q |x_k - x_{k-1}| / (1 - q).$$

This is an a posteriori bound from the actual run and a proved Lipschitz constant. If q is very near one, small increments provide weak information: the remaining sum of many small changes can still be large. Invariance and contraction over a range should be distinguished from a local derivative estimate at one guessed solution.

For $x^2 + x - 1 = 0$, take $g(x) = 1/(1 + x)$ on $I = [0.5, 1]$. Its image $[0.5, 2/3]$ lies within I, and $|g'| \le 4/9$. From $x_0 = 1$, the sequence begins $0.5, 2/3, 3/5, 5/8$. The oscillation about the limit reflects the negative derivative. At $x_4 = 0.625$, the last increment is 0.025, giving error at most $(4/9)/(5/9) \times 0.025 = 0.020$. The exact root $(\sqrt{5} - 1)/2 \simeq 0.61803$ confirms the guarantee is conservative.

Try a relaxed update

The alternative arrangement $g_2(x) = 1 - x^2$ has slope $-2r \simeq -1.236$ at the same positive root. Arbitrarily small generic perturbations grow under this unrelaxed iteration. The equivalent equation alone cannot make it reliable. Form a relaxed map $g_\omega(x) = (1 - \omega)x + \omega g_2(x)$. Its local multiplier is $1 - \omega + \omega g_2'(r)$. Choosing $\omega = 1/2$ changes that multiplier to approximately -0.118.

In fact $g'_{1/2} = 0.5 - x$ on $[0.5, 1]$, so its magnitude is at most 0.5. The map sends the endpoints to 0.625 and 0.5, is decreasing on this interval, and remains inside it. This establishes global contraction on that range. Local cancellation could be made still stronger by tailoring ω to an estimated derivative, but an inaccurate estimate can move the iteration outside the verified interval. A relaxation factor is a numerical parameter; changing it should alter convergence behaviour while leaving a converged physical state unchanged.

Transfer the same reasoning to two coupled states

For a vector iteration, a matrix norm bound replaces the scalar derivative bound. Consider $g(x, y) = (0.2x + 0.3y + 1,\ 0.1x + 0.4y + 2)$. In the infinity norm the largest absolute row sum is 0.5. Subtraction of two evaluations therefore shows $\|g(z) - g(w)\|_\infty \le 0.5\|z - w\|_\infty$, so the map contracts on all of $\mathbb{R}^2$. Starting from zero the first two iterates are $(1, 2)$ and $(1.8, 2.9)$. The completed second increment has norm 0.9, bounding the remaining error by $0.5(0.9)/(1 - 0.5) = 0.9$.

Solving the two fixed-point equations gives $x = 8/3,\ y = 34/9$, so the actual infinity-norm error after the second update is about 0.878. The bound is fairly sharp here. If the two state variables have different units, scale them before combining them into one maximum norm. The scale choice affects the usefulness of a contraction certificate while a converged balance still satisfies the same physical equations.

Check the invariant domain

A map can have derivative smaller than one on a chosen interval yet send its iterates outside it. For example $g(x) = x/2 + 2$ has derivative one half on $[0, 1]$ but maps zero to two. The usual contraction proof on that interval cannot even form its second iterate there. The same map on the full real line is a contraction with fixed point four. Domain and norm are part of a convergence certificate.

Eliminate a geometric error using three iterates

For a scalar iteration near a linearly attracting fixed point, successive errors often behave approximately as $e_{k+1} = q e_k$. If the relation is exact with $q \neq 1$, three successive observations x_0, x_1, x_2 determine both the geometric factor and the limiting state. The difference ratio $(x_2 - x_1)/(x_1 - x_0)$ equals q. Summing the remaining geometric series, or eliminating q algebraically, gives the accelerated value $x_0 - (x_1 - x_0)^2/(x_2 - 2x_1 + x_0)$, provided the denominator is nonzero.

This expression is Aitken's delta-squared transformation. For a smooth nonlinear fixed-point map, the exact geometric relation is replaced by a first-order Taylor approximation about a simple fixed point. Taking $x_1 = g(x_0)$, $x_2 = g(x_1)$ and restarting from the accelerated value produces Steffensen iteration. Its local quadratic behaviour requires sufficient smoothness and $g'(r) \neq 1$. Indeed the denominator is asymptotically $(g'(r) - 1)^2(x_0 - r)$. If that coefficient is tiny or if function evaluations are inaccurate, two successive differences can suffer more cancellation than the fixed-point step itself.

Quantify both the gain and the sensitivity

A sensor calibration fixed point satisfies

$$x = 1 + 0.8x.$$

Starting at zero gives $x_1 = 1$ and $x_2 = 1.8$. The second difference is –0.2, so the accelerated value is $0 - 1^2/(-0.2) = 5$, exactly the fixed point. Ordinary iteration still has error $5 - 1.8 = 3.2$ at this same two-evaluation budget. Here the gain is exact because the iteration contains a single geometric mode with the same factor at every distance from the root.

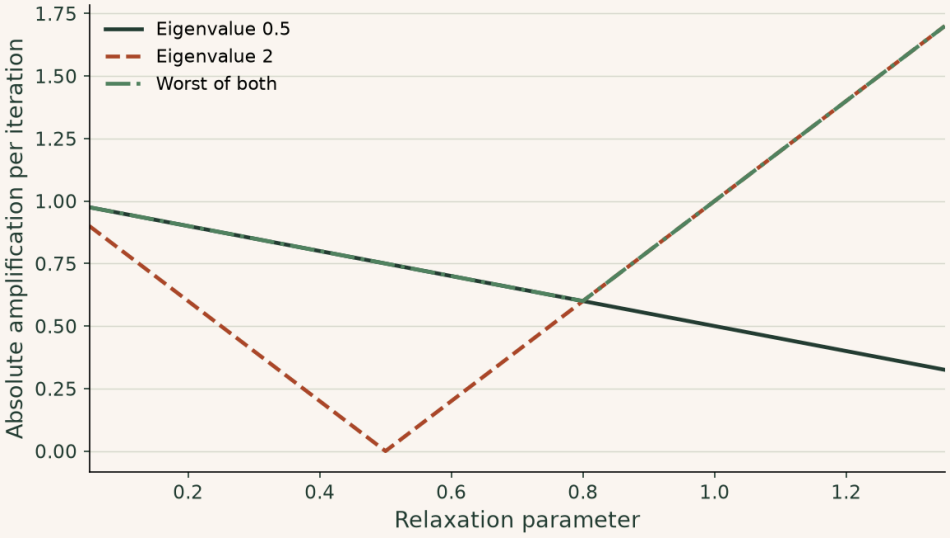
Perturb only the reported second iterate from 1.8 to 1.81. The inferred limit rises to $1/0.19 \approx 5.263$. A perturbation of 0.01 has moved the extrapolated estimate by 0.263. Differentiating the transformed estimate with respect to x_2, holding the first two values fixed, gives magnitude $(x_1 - x_0)^2/(x_2 - 2x_1 + x_0)^2 = 25$ at the unperturbed readings. The predicted change of 0.25 is close to the actual result. The apparent speedup has been purchased with sensitivity to slight deviations from a geometric trend.

A residual check distinguishes these two cases without knowing the exact solution. The exact extrapolated value satisfies $g(5) - 5 = 0$. At the perturbed estimate the residual is about –0.0526, much larger than machine precision. For a reliable contraction factor 0.8 this implies an absolute error bound about 0.263 by dividing by $1 - 0.8$. In a vector problem different modes generally have different convergence factors. Applying this scalar extrapolation separately to raw components may be unreliable when those components mix modes or when a denominator crosses zero in just one coordinate.

There is a second failure mode when the original map is nearly the identity. For $g(x) = 0.999x + 0.001$, starting from zero gives 0.001 and 0.001999. The second difference is only -10^{-6}. If observations are stored to three decimal places, the two reported values become 0.001 and 0.002 and the computed denominator vanishes. This behaviour is predictable from the accuracy of the acquisition and does not indicate that the underlying iteration lacks a fixed point.

Choose a relaxation that contracts both modes

Near a root let the Jacobian be symmetric with eigenvalues 0.5 and 2, and use $x^+ = x - \omega F(x)$. Its two independent error modes have factors $1 - \omega\lambda_i$. Minimise the larger magnitude, allowing a sign reversal in one mode.



At $\omega = 0.8$ the two factors are $+0.6$ and -0.6. Making the fastest mode vanish at $\omega = 0.5$ instead leaves the slow mode at 0.75. For ω greater than one the second mode grows even though the first still shrinks.

▶ **Read values from the plot**

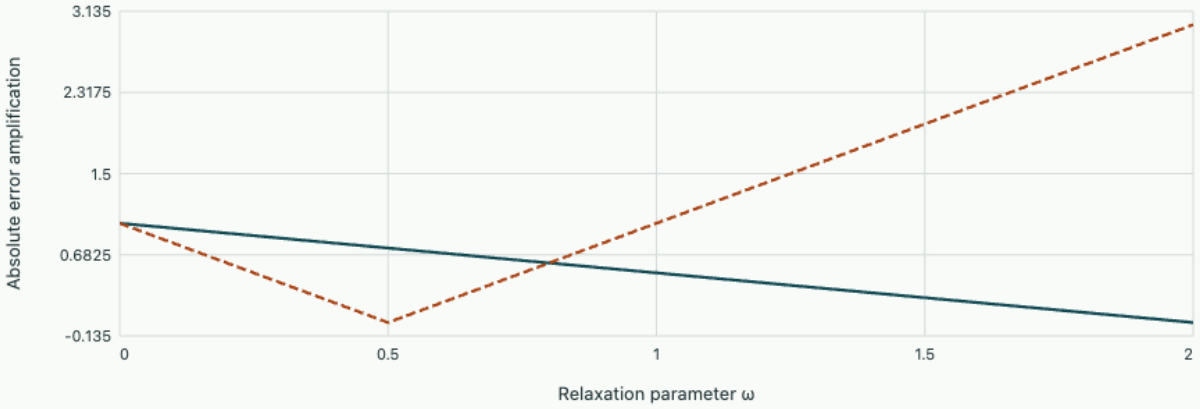
Change parameters · Choose a relaxation that contracts both modes

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

Open this calculation with adjustable inputs

Try the calculation

Small Jacobian eigenvalue	0.5
Large Jacobian eigenvalue	2
Chosen relaxation ω	0.8



At chosen ω the worst factor is 0.6. The minimax choice is 0.8 with factor 0.6.

Study the full derivation

Accelerating fixed points with differences of residuals · Nonlinear systems and iterative methods

A slowly convergent physical iteration contains information in the direction of successive errors. Use a scalar extrapolation and a small residual-minimisation problem to exploit that information, then test the extrapolated state against the original equation.

Problem sheet 1 · Isolating and iterating roots

Use a verified interval to certify one solution, recognise the effect of a transformation on convergence and quantify a stopping rule.

Find a root within a specified distance

Show that $F(x) = x + \ln x - 2$ has exactly one root in $[1, 2]$. Carry out two bisections. Determine how many halvings guarantee an absolute midpoint error below 0.001, and give a slope-based error bound for the trial value $x = 1.56$.

▼ **Worked solution**

At 1 the residual is −1, while $F(2) = \ln 2 > 0$. Its derivative $1 + 1/x$ is at least 1.5 throughout, giving strict monotonicity. At 1.5 the residual is approximately −0.0945 and at 1.75 it is +0.3096, so after two halvings the bracket is $[1.5, 1.75]$.

The initial width is one. After nine halvings the midpoint bound is $1/1024 < 0.001$. At $x = 1.56$, $F \simeq 0.004686$. The mean value bound gives an error at most $0.004686/1.5 = 0.00313$, assuming the exact logarithm on the stated bracket.

Explore the related calculation with adjustable inputs. Compare actual root error with the shrinking bracket certificate. The explorer uses a cubic test residual; the continuity and slope checks in this exercise must still be carried out for its stated residual.

Compare two rearrangements

For $x^2 - 3 = 0$ on $[1, 2]$, consider $g_1 = 3/x$ and $g_2 = (x + 3/x)/2$. Starting at $x_0 = 1.5$, compute two steps of each. Use the local derivatives to explain the different histories.

▼ **Worked solution**

The first map produces 2, then 1.5, because applying the reciprocal transformation twice returns the start. At $r = \sqrt{3}$ its derivative is $-3/r^2 = -1$. The second map first gives 1.75, then

$$(1.75 + 3/1.75)/2 = 1.7321429.$$

At the root $g_2' = (1 - 3/x^2)/2 = 0$, which cancels the linear error term. In particular, a fixed point for either formulation solves the original equation, but a repeated two-cycle of g_1 need not solve it.

Explore the related calculation with adjustable inputs. Explore how initial values and the residual slope affect Newton and secant convergence on the quadratic reference problem; retain the stated residual when checking the present hand calculation.

Interpret a small update

A contraction with known factor at most 0.95 has just changed its state by 0.002 K. Bound the current state error. How small must the next completed increment be to certify 0.01 K by the same test? Explain why stopping at a fixed increment of 0.01 K would be misleading.

▼ **Worked solution**

Summing the future geometric changes from the current iterate gives a bound $q|\Delta x|/(1-q) = 19(0.002) = 0.038$ K. To guarantee 0.01 K requires $|\Delta x| \le 0.01/19 \simeq 0.000526$ K.

For a slowly contracting process small updates can continue in the same direction over many iterations. A rule using the increment alone ignores that accumulated tail.

Explore the related calculation with adjustable inputs. Vary the contraction slope and compare the remaining error after a fixed evaluation budget.

Notebook practice · A bracket can collapse onto a pole

Chapter 3 · Newton and secant updates

Derive derivative-based and derivative-free updates, measure their convergence near a simple zero and explain why a repeated zero changes the convergence order.

FROM FIRST PRINCIPLES

Build one root correction from a local line

Suppose f is differentiable near a root and a trial point x_k has a nonzero derivative. Taylor expansion models the function a short displacement s away by $f(x_k) + f'(x_k)s$. Requiring this local line to vanish gives Newton's update,

$$s = -\frac{f(x_k)}{f'(x_k)},$$
$$x_{k+1} = x_k + s.$$

The true new residual is the nonlinear remainder. For twice continuously differentiable f and a segment where $|f''| \le M$, Taylor's theorem yields

$$|f(x_{k+1})| \le \tfrac{1}{2} M |s|^2.$$

Converting residual to a location error also requires a lower bound on $|f'|$ near the root. If derivatives are unavailable, estimate the local slope by $[f(x_k) - f(x_{k-1})]/[x_k - x_{k-1}]$ to obtain the secant method. Its denominator must be resolved numerically. Both methods need caution far from the root, where the line can extrapolate outside the region where it represents the function.

DERIVATION AND APPLICATIONS

Recover the multiplicity of a root from derivative measurements

At a multiple zero, the ordinary Newton step no longer removes the leading location error. Suppose $f(x) = (x - r)^m h(x)$, with integer $m \ge 1$, smooth h and $h(r) \ne 0$. Write $e = x - r$. Differentiating the product and keeping the leading powers gives

$$f' = m e^{m-1} h + e^m h',$$
$$\frac{f}{f'} = \frac{e}{m + e h'/h} = \frac{e}{m} + O(e^2).$$

It follows that ordinary Newton leaves the fraction $1 - 1/m$ of the error to first order. For known m, multiplying the Newton correction by m cancels the leading error. When multiplicity is unknown, consider the auxiliary function $u = f/f'$, which has a simple zero after continuous extension to r. Its derivative and Newton correction are

$$u' = 1 - \frac{ff''}{(f')^2},$$

$$x_{\text{new}} = x - \frac{u}{u'} = x - \frac{ff'}{(f')^2 - ff''}.$$

The leading powers in the twice differentiated product give a local diagnostic,

$$\frac{ff''}{(f')^2} \longrightarrow \frac{m-1}{m}, \qquad \frac{1}{1 - ff''/(f')^2} \longrightarrow m.$$

Far from the zero the inferred value need not be integral or even positive. Near a flat root both derivatives and the function value can become very small, so the ratio needs sufficiently accurate evaluations. The derivation gives a local method and does not establish that a distant initial guess will enter its convergence region.

WORKED EXAMPLE

Recover a triple zero from one set of values

For $f(x) = (x-2)^3$ at $x = 2.2$, $f = 0.008$, $f' = 0.12$, $f'' = 1.2$. The diagnostic gives

$$1/[1 - (0.008)(1.2)/0.12^2] = 3.$$

The ordinary Newton step lands at 2.13333, while multiplying its correction by three gives 2 exactly because h is constant.

SECOND APPLICATION

Check the effect of a changing smooth factor

For a different function $f(x) = x^2(1+x)$ at $x = 0.1$, the three values are $0.011, 0.23, 2.6$. The apparent multiplicity is

$$0.23^2/(0.23^2 - 0.011 \times 2.6) = 2.18.$$

Newton on u gives

$$x_{\text{new}} = 0.1 - (0.011)(0.23)/(0.0243) = -0.00412.$$

Crossing the zero is consistent with the curvature of u; the absolute location error nevertheless drops by a factor exceeding 24.

Start from this chapter. Differentiate the fourth-power loss in the implicit temperature residual. Predict which change alters the physical root: changing the time step or changing the Newton update.

OneSimulation laboratory · Locate the root of an implicit cooling step

Linearise the residual where you stand

Suppose a scalar balance $F(x) = 0$ is smooth, with a simple root r, meaning $F'(r) \neq 0$. Replacing the unknown function by its tangent at a current estimate x_k gives

$$0 \simeq F(x_k) + F'(x_k)(x_{k+1} - x_k).$$

Hence $x_{k+1} = x_k - F(x_k)/F'(x_k)$. The quotient has units of the unknown variable. When it is used in a coupled thermal model, for instance, a mismatch in watts divided by a derivative in W/K produces a temperature correction.

Assume both the current iterate and the root lie where $|F''| \leq M$ and $|F'| \geq m > 0$. Taylor expand $F(r)$ about x_k and insert the Newton update to obtain

$$|x_{k+1} - r| \leq M|x_k - r|^2/(2m).$$

For sufficiently small initial error the next iterate stays in the same neighbourhood and its error is of quadratic order. This is a local conclusion. A nearly horizontal tangent far away can give a huge update even when the desired root is simple.

Calculate the improvement explicitly

For $F(x) = x^2 - 2$, Newton's update is $(x + 2/x)/2$. From 1.5 it gives $17/12 = 1.4166667$, then

$$577/408 = 1.4142157.$$

The absolute errors fall from approximately 0.08579 to 0.002453 and then to 2.12×10^{-6}. An exact identity explains these figures:

$$x_{k+1} - \sqrt{2} = (x_k - \sqrt{2})^2/(2x_k).$$

The positive sequence stays on the high side after the first update and approaches the physically selected positive root.

When analytic derivatives are expensive, replace $F'(x_k)$ by the slope between x_k and x_{k-1}. The resulting secant update uses two initial estimates and only one new residual evaluation per subsequent step. For the same quadratic, the divided difference is exactly $x_k + x_{k-1}$; algebra gives

$$e_{k+1} = e_k e_{k-1}/(x_k + x_{k-1}),$$

where $e_j = x_j - \sqrt{2}$. Near the root the denominator is nearly constant. If e_{k+1} scales as e_k^p, consistency with the product relation gives

$$p^2 = p + 1.$$

The positive solution is about 1.618, superlinear but less than two.

Examine a root of higher multiplicity

For $F(x) = (x - r)^m$, the tangent correction is only $(x - r)/m$. Newton therefore leaves $e_{k+1} = (1 - 1/m)e_k$, a fixed linear reduction. With $m = 3$ and an initial error of 0.3, the first two errors are 0.2 and 0.13333. Meanwhile the residual is cubed, so even a small residual may mask appreciable positional error. Multiplying the Newton correction by m restores the exact root in one step for this pure power and restores quadratic convergence more generally when the multiplicity is known and the remaining factor is smooth and nonzero.

Numerically differentiating a residual introduces its own scale. For the forward difference $[F(x + s) - F(x)]/s$, Taylor truncation decreases as $O(s)$, while rounding or evaluation noise in the numerator is divided by s. Blindly shrinking s can therefore spoil the derivative. Recompute at a second scale, compare with an analytic derivative when available, and check the final state against the original balance. A derivative that is accurate near the starting point need not stay accurate along the entire sequence.

Background · How can the iterates suggest a multiplicity?

Suppose an isolated root has multiplicity $m > 1$, so $f(x) = (x - r)^m g(x)$ with smooth $g(r) \neq 0$. Dividing the function by its derivative shows that the leading Newton correction is $(x - r)/m$. Sufficiently near the root, successive corrections therefore obey

$$\frac{|x_{k+1} - x_k|}{|x_k - x_{k-1}|} \longrightarrow 1 - \frac{1}{m}.$$

A sustained ratio of about $2/3$ suggests a triple root. A single such ratio is weak evidence: a line search that repeatedly accepts one third of a full step can produce linear convergence even near a simple root. Check the actual step acceptance and the residual model before multiplying all subsequent corrections by three. Ratios of corrections also become unreliable once their sizes approach floating-point resolution.

Set stopping tests that survive a change of units

Multiplying a residual by a nonzero constant c leaves a Newton correction unchanged because both the residual and its derivative acquire that factor. Secant updates have the same invariance. An absolute stopping threshold on the residual alone, however, changes its meaning by c. For $F = x^2 - 2$ the trial $x = 1.414$ has $|F| \simeq 0.000604$. Reporting the equivalent balance as $1000F$ gives a mismatch about 0.604, with the identical prospective Newton step.

A useful report contains a scaled residual, the size of the accepted update in variable units and the conditions used to leave the iteration. Near a simple root a final undamped correction estimates the remaining positional error to first order. A small *damped* correction can equally be the consequence of a line search rejecting progress. Recheck the undamped balance or a bracket before treating a tiny change in the displayed state as convergence.

Prove a local quadratic bound with an explicit radius

Let r be a simple root of a twice differentiable scalar f. On a surrounding interval suppose $|f'| \geq m > 0$ and $|f''| \leq M$. Taylor expansion about a trial point x gives $0 = f(r) = f(x) + f'(x)(r - x) + f''(\xi)(r - x)^2/2$ for an intermediate ξ. Subtract the exact root from the Newton update $x_+ = x - f(x)/f'(x)$. The linear terms cancel and $|x_+ - r| \leq [M/(2m)]|x - r|^2$. This supplies both the mechanism and a coefficient for the familiar quadratic statement.

The constants must continue to hold at every iterate. If the closed ball of radius R about the root lies in the stated interval and $MR/(2m) < 1$, any starting point in that ball maps strictly closer to the root and remains in the ball. Repeating the bound is then justified by induction. Without such an invariant region, a single Taylor estimate says nothing about what happens after a step leaves the region of controlled derivatives. A relative update tolerance cannot substitute for a verified lower derivative bound near a flat or multiple root.

Certify a square root and assess an inexact division

For $f(x) = x^2 - 9$, use the interval $[2, 4]$, root three, $m = 4$, $M = 2$. The resulting quadratic error coefficient is $1/4$. From $x_0 = 2.4$ the true initial error is 0.6 and the first error is bounded by 0.09. The actual Newton update is $(2.4 + 9/2.4)/2 = 3.075$, leaving error 0.075. The next bound using the actual first error is

$$0.075^2/4 = 0.00140625.$$

Independently the exact update error obeys $(x - 3)^2/(2x)$, giving $0.075^2/(2 \cdot 3.075) \approx 0.0009146$. The guaranteed number is larger because it allows the smallest derivative on the whole interval.

Now suppose the computed Newton update has an additional absolute rounding uncertainty of 10^{-4} each time. The same reasoning gives

$$|e_+| \le |e|^2/4 + 10^{-4}.$$

Once $|e|$ falls below roughly 0.02, the allowed update uncertainty exceeds the quadratic contribution. Continuing iterations can still help, but the mathematical inequality no longer promises a doubling of correct digits. This transition is visible from the size of the actual input and evaluation errors and should not be fitted away as a mysterious change of convergence order.

A final distance certificate can avoid using the exact root in a real calculation. If a sign-changing bracket is known to contain the root and $|f'| \ge 4$ throughout it, then $|x - r| \le |f(x)|/4$ for a point x inside the bracket. The local Newton proof and the residual-based certificate therefore serve complementary roles: one predicts improvement per iteration, while the other assesses the current uncertainty from computable information.

In the same square-root example a residual of 0.0008 with the iterate inside $[2, 4]$ certifies error at most 0.0002, even without relying on observed convergence ratios. If one instead used the interval $[-4, 4]$, the derivative lower bound would be zero and the same residual could not tell which of the two roots was nearby. Certifying a branch is part of certifying a distance.

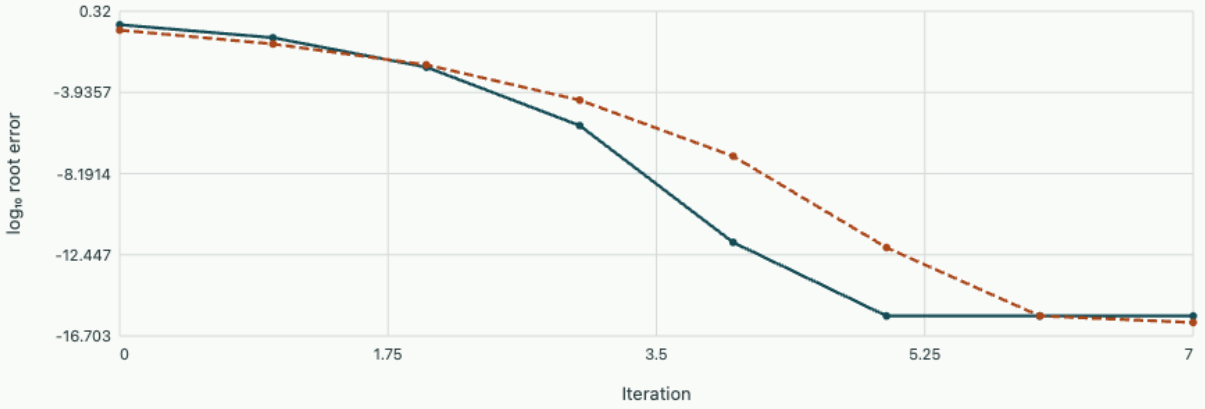
Change parameters · Newton and secant updates

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

Positive target a	2
Initial Newton value	1
Iterations	7



Positive reference root 1.4142. Secant starts at x_0 and $x_0 + 0.2$; both methods recompute their local slopes at each step.

Study the full derivation

Choosing a derivative step in a nonlinear solve · Nonlinear systems and iterative methods

A smaller finite-difference step eventually makes Newton iterations worse. Derive truncation and evaluation errors, compare a complex perturbation and a directional derivative, and test which numerical derivative is meaningful for the actual residual implementation.

Chapter 4 · Newton's method for coupled balances

Assemble a Jacobian from two coupled residuals, solve for a simultaneous correction and interpret how scaling changes the meaning of a norm.

FROM FIRST PRINCIPLES

Derive simultaneous corrections from all balance equations

Suppose two unknowns u, v must satisfy residual equations $F_1(u, v) = 0,\ F_2(u, v) = 0$. Changing either unknown can affect both balances. Expand each residual around the same trial state $x = (u, v)^T$, with a small proposed correction s. The partial derivatives form the Jacobian J,

$$F_i(x + s) = F_i(x) + \sum_{j=1}^{2} \frac{\partial F_i}{\partial x_j}(x) s_j + O(\|s\|^2).$$

If the derivatives are locally Lipschitz and the linear model is reliable, choose the correction that cancels both leading residuals together,

$$J(x)s = -F(x),$$
$$x_{\text{new}} = x + s.$$

A nonsingular Jacobian guarantees a unique local correction. Solving the system directly avoids constructing its inverse. A final small residual can only establish accurate state variables if the balances are sufficiently sensitive to changes in those variables. Scaling each quantity by a physically meaningful reference helps distinguish genuine insensitivity from a misleading comparison between different units.

Background · Newton correction

$$J(x_k)s_k = -F(x_k)$$

The correction solves a local linear model. Quadratic local convergence needs a suitable nonsingular derivative and a sufficiently close starting point.

DERIVATION AND APPLICATIONS

Eliminate a coupled balance without losing its derivative

Some models contain one equation that can be solved quickly for an auxiliary variable. Suppose $F(u, v) = 0,\ G(u, v) = 0$ and, locally, $G_v \neq 0$. If the auxiliary equation gives $v = g(u)$, the reduced residual is

$$H(u) = F(u, g(u)).$$

Differentiating $G(u, g(u)) = 0$ first accounts for the effect of a changed u on the auxiliary state,

$$g' = -G_u/G_v,$$
$$H' = F_u - F_v G_u/G_v.$$

This same combination arises even when the current auxiliary state is not perfectly equilibrated. The full two-row Newton system gives a formula for the auxiliary correction from the second row. Substituting it into the first row yields

$$\delta v = \frac{-G - G_u \delta u}{G_v}, \qquad \left(F_u - \frac{F_v G_u}{G_v}\right)\delta u = -F + \frac{F_v G}{G_v}.$$

At an equilibrated state $G = 0$, this is exactly Newton's scalar step for H. The second row's imbalance contributes to the right-hand side otherwise. The determinant identity

$$\det J = F_u G_v - F_v G_u = G_v H'$$

shows why an apparently harmless derivative F_u can hide a nearly singular coupled model. After a finite step, directly recomputing $g(u + \delta u)$ can differ from $v + \delta v$ by a second-order term when g is curved. Both choices share the same first-order model.

WORKED EXAMPLE

Calculate a simultaneous correction from an imbalanced state

Take $F = u + 2v - 5,\ G = v - u^2$, initially $(u, v) = (1, 0.8)$. Then

$$F = -2.4,\ G = -0.2,\ G_u = -2,\ G_v = 1.$$

The reduced derivative is 5 and the corrected right side is

$$2.4 - 0.4 = 2.0.$$

Thus

$$\delta u = 0.4,\ \delta v = 1.0.$$

The new state $(1.4, 1.8)$ satisfies the linear F balance exactly; the remaining $G = -0.16$ is the neglected quadratic term.

SECOND APPLICATION

Identify a cancellation in a reduced slope

In a second model $F = 2u - v - 1,\ G = v - 1.99u$. The individual slope $F_u = 2$, but eliminating v leaves $H(u) = 0.01u - 1$. Starting at zero the correct Newton correction is 100, giving $v = 199$, an exact solution of both linear balances. Using the partial slope of 2 while holding v fixed would instead propose $u = 0.5$ and miss the compensating response.

Linearise the whole model at one state

Two unknown physical quantities may influence both measured balances. Changing one to erase one residual can disrupt the other. The vector Taylor expansion $F(x + s) = F(x) + J(x)s + O(\|s\|^2)$ collects the simultaneous first-order response. Entry J_{ij} is the derivative of residual i with respect to unknown j while holding the other unknowns fixed. A Newton step solves the linear system $J(x)s = -F(x)$ and sets $x_{new} = x + s$. Factor or iteratively solve that system rather than explicitly computing an inverse.

Suppose J is Lipschitz with constant L near a root, and the inverses in that neighbourhood satisfy

$$\|J^{-1}\| \le K.$$

Integrate J along the line from the current point to the root. The error left by replacing J with its value at the start is bounded by $L\|x - r\|^2/2$. Multiplying by the inverse yields the local Newton bound $KL\|x - r\|^2/2$. This derivation makes explicit which objects must stay well behaved: derivative variation and amplification by the linear solve.

Intersect an energy constraint with a matching condition

Take the dimensionless equations $F_1 = x^2 + y^2 - 1,\ F_2 = x - y$, and choose the positive branch. The linearised system at $(0.8, 0.6)$ has $F = (0, 0.2)$ and rows $(1.6, 1.2), (1, -1)$. The point already lies on the unit circle but violates matching. Newton requires

$$s_y = s_x + 0.2;$$

substituting in the first row gives $2.8s_x = -0.24$. Thus $s_x = -0.0857143,\ s_y = 0.114286$ and the new state is $(5/7, 5/7)$.

The linear constraint is now exact. The circle residual is $2(5/7)^2 - 1 = 1/49$, a second-order consequence of moving along its tangent rather than along the circle itself. At the next iteration the two corrections are equal. Solving $(20/7)s = -1/49$ gives $s = -1/140$, hence

$$x = y = 99/140 = 0.707143.$$

The desired value is

$$1/\sqrt{2} = 0.707107.$$

A different initial state could lead to the negative intersection. Physical admissibility must therefore supplement the algebraic equations.

Quantify the cost of freezing a Jacobian

A new Jacobian factorisation at every iterate can dominate the work of solving a coupled equilibrium. Fix $B = J(x_0)$, factor it once, and take $x_{k+1} = x_k - B^{-1}F(x_k)$. Near a simple root r write $e = x - r$. Taylor expansion gives

$$e_+ = (I - B^{-1}J(r))e + O(\|e\|^2).$$

With a current exact Jacobian the first-order term vanishes. A frozen Jacobian normally retains it, so convergence is generally linear and even local convergence requires control of this iteration matrix. A sufficient condition on a neighbourhood is $\sup_x \|I - B^{-1}J(x)\| < 1$, together with an invariant neighbourhood.

Consider $F(x, y) = (x^2 + y - 3, x - y)^T$, starting at $(1, 1)$. The fixed matrix has rows $(2, 1)$ and $(1, -1)$. The second residual remains zero if both coordinates change equally. On that line the correction satisfies

$$3s = -(x^2 + x - 3).$$

The first increment is $1/3$. The new first residual is $1/9$. Reusing the factors gives the next increment $-1/27$ and the state $x = y = 35/27 \approx 1.29630$. Refactoring at $4/3$ would instead give increment $-1/33$, reaching $43/33 \approx 1.30303$. The exact positive root is $(\sqrt{13} - 1)/2 \approx 1.30278$.

The local error multiplier along this line is $1 - \sqrt{13}/3 \approx -0.202$, predicting alternating errors for the frozen calculation. Reusing a poor factorisation indefinitely can cost more residual evaluations than it saves. Compare actual reduction against the expected local factor and refresh the Jacobian when model changes make reuse ineffective.

Practise on sheet 2 · Decide when a reused slope is still informative. Further reading · ETH Zürich · Numerical Methods for Computational Science and Engineering · Remark 8.5.1.43, Simplified Newton method, and Code 8.5.1.44, p. 625.

Select scales and verify a derivative action

Changing units from metres to millimetres multiplies the displayed displacement by 1000 without changing the model. Yet an unscaled Euclidean norm of a state mixing displacement and temperature can then change radically. Choose characteristic scales in a positive diagonal matrix D_x and residual scales D_F. Form the dimensionless unknown $z = D_x^{-1}x$ and residual $G = D_F^{-1}F(D_x z)$. By the chain

rule $J_G = D_F^{-1} J_F D_x$. For an exact undamped Newton solve this produces the same physical correction, but stopping tests, trust regions and finite differences now acquire interpretable scales.

To check a programmed Jacobian, choose a direction d with nonzero entries in multiple unknowns and compare Jd with $[F(x + sd) - F(x)]/s$ over several s. For the circle constraint the discrepancy is exactly $s(d_x^2 + d_y^2)$, while for the linear matching condition it is zero in exact arithmetic. A one-direction check can miss a coding error orthogonal to that direction. Try additional directions emphasizing different couplings, keeping clear of nonsmooth switches. This test inspects the local derivative independently of the solver's final residual.

Separate state sensitivity from linearisation error

Suppose the matching constraint in the intersection example is perturbed to $x - y = \epsilon$, while the unit-circle measurement remains exact. About the positive root $x = y = 1/\sqrt{2}$, differentiating the circle implies $\delta x + \delta y = 0$. Differentiating the matching constraint gives

$$\delta x - \delta y = \epsilon.$$

Thus a measurement offset of 0.02 moves the coordinates by approximately $+0.01$ and -0.01, even if the perturbed nonlinear equations are solved to roundoff.

Geometrically, uncertainty becomes much larger when the two constraint curves meet almost tangentially. Their Jacobian rows then point in nearly the same direction, leaving a poorly resolved transverse displacement. Reducing the Newton tolerance improves the algebraic solution of the supplied measurements, but cannot increase their independence. Inspecting singular directions of the scaled Jacobian near the converged state helps distinguish this physical uncertainty from difficulties caused by a poor initial guess.

Analyse the basins of a coupled two-dimensional problem

Identify a real state (x, y) with $z = x + iy$. Solving $z^2 - 1 = 0$ is equivalent to solving $x^2 - y^2 - 1 = 0,\ 2xy = 0$ for two real unknowns. The real Jacobian has rows $(2x, -2y)$ and $(2y, 2x)$, with determinant $4(x^2 + y^2)$. It is nonsingular away from the origin. Complex division by $2z$ implements exactly the real Newton correction for this system. Therefore the full update is $N(z) = (z + 1/z)/2$, and both roots $+1$ and -1 have local quadratic convergence.

To determine which root is selected, set $w = (z - 1)/(z + 1)$, excluding the already solved point -1. Direct substitution shows $N(z) - 1 = (z - 1)^2/(2z)$ and $N(z) + 1 = (z + 1)^2/(2z)$. Dividing yields

$$w_{\text{new}} = w^2.$$

Hence $w_k = w_0^{2^k}$ as long as the updates remain defined. If $|w_0| < 1$, this tends to zero and $z = (1 + w)/(1 - w)$ tends to one. The condition means the starting point is closer to $+1$ than to -1, precisely the open right half-plane. The left half-plane similarly converges to the other root.

Predict a basin and trace a boundary failure

Start at $z_0 = 1 + i$, so

$$1/z_0 = (1 - i)/2.$$

The first update is

$$z_1 = 0.75 + 0.25i.$$

Its inverse equals $1.2 - 0.4i$, giving $z_2 = 0.975 - 0.075i$. The iterates moved from above to below the real axis, but both remained in the basin of $+1$. In transformed coordinates $w_0 = i/(2 + i) = (1 + 2i)/5$, of magnitude $1/\sqrt{5}$; after two steps the magnitude is $1/25$. This quantifies progress without estimating the distance from z_1 to either root by visual inspection.

The basin boundary is completely different. For $z_0 = i$, the first update is zero and the next Jacobian is singular. More generally a purely imaginary nonzero start remains purely imaginary for as long as the map is defined, because both z and $1/z$ have zero real part. It cannot converge to either real root while staying on this line. Infinitesimal perturbations toward opposite sides of the line select different roots. A solver reporting a converged root from a point numerically very close to this boundary should identify which initial or rounding perturbations could change that result.

This example makes global multiplicity observable in a system whose local steps are cheap and whose Jacobian is invertible almost everywhere. A small final residual confirms that the reported state solves the equations. It says little about whether the originally intended physical branch was selected. Additional physical conditions or a continuation path from a known state can supply information that the equations and a nominal starting vector alone do not specify.

Initial guesses only 10^{-6} to the right and left of the imaginary axis can therefore give final real parts differing by two. That discontinuity concerns the map from starting guess to selected solution, not a discontinuity of the original polynomial equations. It is one reason to report how physical branch selection was imposed when several roots correspond to substantially different states, even when every final residual is small.

Change parameters · Explore the worked calculation

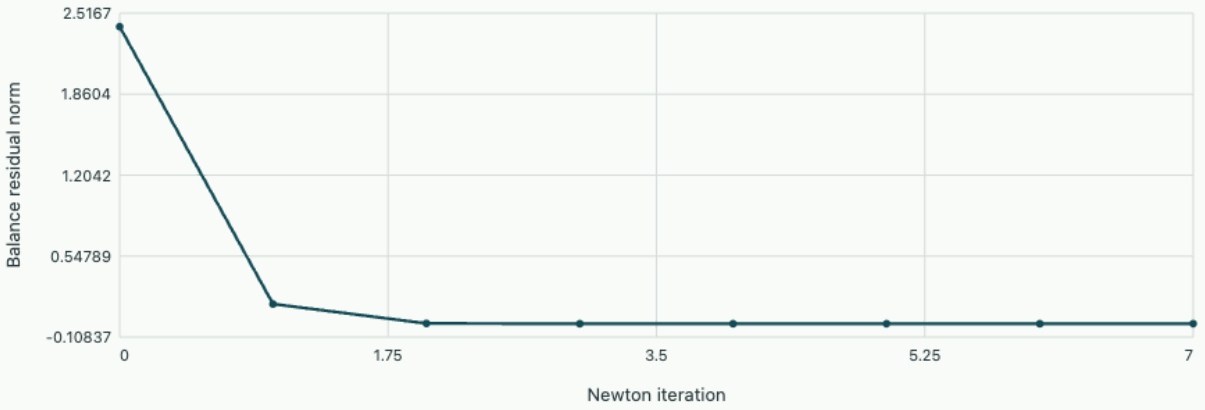
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Initial u **1**

Initial v **0.8**



Simultaneous two-balance Newton solve

Final state u = 1.3508, v = 1.8246. The cross derivatives Fv = 2 and Gu = −2u enter the same linear correction; treating v as fixed changes the problem.

Study the full derivation

Choosing a derivative step in a nonlinear solve · Nonlinear systems and iterative methods

A smaller finite-difference step eventually makes Newton iterations worse. Derive truncation and evaluation errors, compare a complex perturbation and a directional derivative, and test which numerical derivative is meaningful for the actual residual implementation.

Problem sheet 2 · Tangents and Jacobians

Compute scalar, secant and coupled Newton corrections, and test what an incorrectly assembled derivative changes.

Use the secant on a cubic

For $F(x) = x^3 - 3$, use the two starts $x_0 = 1$ and $x_1 = 2$. Find the next secant iterate, compare it with one Newton update from 2, and evaluate both residuals. Does one comparison prove an asymptotic convergence order?

▼ **Worked solution**

The secant slope is $(5 - (-2))/(2 - 1) = 7$, giving $x_2 = 2 - 5/7 = 9/7$ and residual $729/343 - 3 \simeq -0.875$. Newton uses slope 12, giving $19/12 \simeq 1.58333$ and residual $6859/1728 - 3 \simeq 0.970$.

The methods take corrections from different local models. One residual comparison on early iterates does not establish their asymptotic rates, which concern successive errors near the root.

Explore the related calculation with adjustable inputs. Explore how initial values and the residual slope affect Newton and secant convergence on the quadratic reference problem; retain the stated residual when checking the present hand calculation.

Locate two coupled unknowns

Take $F_1 = x^2 + y - 3,\ F_2 = x + y^2 - 3$. At $x = y = 1$, compute the Jacobian and full Newton update. Compare the residual after that update with the result of dropping both coupling terms from the Jacobian.

▼ **Worked solution**

The residual is $(-1, -1)$ and $J = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. Symmetry gives $s_x = s_y = 1/3$, hence

$$x = y = 4/3.$$

Both new residuals equal $1/9$.

Dropping the off-diagonal entries instead gives $s_x = s_y = 1/2$, landing at 1.5 in both coordinates with residual 0.75 in each row. The uncoupled derivative overestimates the correction because each unknown changes both balances.

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.

Roundoff in a derivative test

Estimate the derivative of $F(x) = x^2 - 2$ at 1.5 using a forward quotient with positive step s. Derive the exact truncation error and suppose arithmetic introduces absolute numerator error at most 10^{-14}. Choose s to minimise the sum of the two error bounds.

▼ **Worked solution**

Subtracting the function values and dividing by s gives $3 + s$, so the exact truncation error is s. Numerator noise contributes at most $10^{-14}/s$. The sum $s + 10^{-14}/s$ is minimised by $s^2 = 10^{-14}$, giving $s = 10^{-7}$ and a bound 2×10^{-7}.

A perturbation smaller than this model's optimum amplifies cancellation. The numerical optimum depends on evaluation scale and arithmetic, so use several step sizes when testing an unfamiliar Jacobian.

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.

Decide when a reused slope is still informative

Solve $x^2 = 9$ from $x_0 = 4$ using the initial slope for the first two steps. Compare the second result with refreshing the slope at the first iterate. For an arbitrary positive frozen slope B, find the range giving a strict local contraction at the positive root.

▼ **Worked solution**

The initial slope is 8. The first iterate is

$$4 - 7/8 = 25/8.$$

Its residual is $49/64$. Reuse gives

$$25/8 - 49/512 = 1551/512 \approx 3.02930.$$

Refreshing the slope to $25/4$ gives

$$25/8 - (49/64)(4/25) = 3.0025.$$

At the root the error multiplier is $1 - 6/B$. The strict inequality $|1 - 6/B| < 1$ requires $B > 3$. In particular $B = 8$ gives multiplier $1/4$: factor reuse converges but does not retain Newton's local quadratic rate.

Explore the related calculation with adjustable inputs. Explore how initial values and the residual slope affect Newton and secant convergence on the quadratic reference problem; retain the stated residual when checking the present hand calculation.

Revisit chapter 4 · Quantify the cost of freezing a Jacobian.

Distinguish root accuracy from temperature sensitivity

For one implicit cooling step take $T_n = 2$ K, $T_a = 1$ K, $h = 1$ s and $\beta = 0.0015$ K⁻³s⁻¹. Differentiate the stage equation with respect to the input temperature, treating h, β, T_a as fixed. Use the tabulated root 1.97851479 K to predict the effect of a 0.001 K increase in T_n. Compare its scale with an algebraic residual near 10^{-16} K.

▼ **Work through the solution**

Write $F = x - T_n + h\beta(x^4 - T_a^4)$. Implicit differentiation yields

$$(1 + 4h\beta x^3)\,\partial x/\partial T_n = 1.$$

The derivative at the tabulated root is approximately 0.956. A small 0.001 K input increase therefore shifts the physical stage root by about 0.000956 K. A tiny numerical residual verifies the equation for specified inputs, while the derivative measures uncertainty resulting from a change to those inputs.

Use the OneSimulation laboratory to check your answer

Chapter 5 · Line searches and trust regions

Diagnose a cycling Newton sequence, compute a merit-function decrease and explain how damping and trust regions change the step while preserving the root.

FROM FIRST PRINCIPLES

Choose a step using actual reduction in residual

A local Newton model can predict a zero far outside its useful region. To test the correction, use the merit function $\phi(x) = \|F(x)\|_2^2/2$. Along a candidate direction s, differentiate this scalar at step length zero. For a Newton direction solving $Js = -F$, the predicted slope is strictly negative whenever $F \neq 0$,

$$\frac{d}{dt}\phi(x+ts)\Big|_{t=0} = F^T J s = -\|F\|_2^2.$$

Choose a fraction $0 < c < 1$ of that initial decrease and accept a trial length $\alpha > 0$ when the actual residual satisfies

$$\phi(x+\alpha s) \le \phi(x) + c\alpha\,\nabla\phi(x)^T s.$$

Try full length first and shorten if necessary. For a differentiable merit function and a descent direction, sufficiently short feasible steps pass because the linear term dominates the remainder. This establishes a local acceptance mechanism. Bounds on variables, nondecreasing directions, or stationary points of ϕ with nonzero residual require additional interpretation.

DERIVATION AND APPLICATIONS

Bound the number of reductions in a backtracking line search

A sufficient decrease test can be guaranteed once the step is short enough. Assume a differentiable merit ϕ, a nonzero direction p with $g^T p < 0$ for $g = \nabla\phi(x)$, and an L-Lipschitz gradient on the whole feasible trial segment. The change in merit equals an integral of directional derivative. Replacing the gradient at a displaced point by its initial value introduces a bounded error,

$$\phi(x+\alpha p) - \phi(x) = \int_0^\alpha \nabla\phi(x+tp)^T p\,dt \le \alpha g^T p + \tfrac{1}{2}L\alpha^2\|p\|^2.$$

To make the right-hand side no larger than the Armijo requirement $c\alpha g^T p$, rearrange the inequality using $g^T p < 0$ and $\alpha > 0$. This yields a guaranteed acceptance range

$$0 < \alpha \le \alpha_{\text{safe}} = \frac{2(1-c)(-g^T p)}{L\|p\|^2}.$$

Starting at α_0 and reducing by a fixed factor $0 < \rho < 1$, the number of reductions before reaching this range is at most

$$j_{\max} = \max\left(0, \left\lceil \frac{\ln(\alpha_{\text{safe}}/\alpha_0)}{\ln \rho} \right\rceil\right).$$

Larger steps can already pass because the curvature bound is only sufficient. This estimate makes the roles of a nearly tangential descent direction and a poor scale for the search direction explicit. It assumes trial points lie inside the domain on which the function and its gradient bound are valid.

WORKED EXAMPLE

Predict a line search on an exact quadratic

Use $\phi(x) = 5x^2$, start at $x = 1$, and take $p = -10$. Here $L = 10,\ g^T p = -100,\ \|p\|^2 = 100$. With $c = 0.1$ the safe length is 0.18. Starting from 1 and halving reaches 0.125 after three reductions. The merit becomes $5(-0.25)^2 = 0.3125$, below the required

$$5 - 1.25 = 3.75.$$

The previous length 0.25 gives merit 11.25 and fails.

SECOND APPLICATION

Check an acceptance claim under curvature uncertainty

Another model has $-g^T p = 0.8,\ \|p\| = 2,\ L = 3,\ c = 0.25$. The sufficient length is

$$2(0.75)(0.8)/(3 \times 4) = 0.100.$$

An initial 0.80 reduced by a factor 0.4 produces $0.32,\ 0.128,\ 0.0512$; the third reduction is guaranteed to pass. Accepting 0.128 on the actual merit would also be consistent with the bound.

> **Start from this chapter.** Compare damped and undamped Newton on the same stage equation. Substitute each returned temperature into the residual independently to see whether they reach the same root.
>
> OneSimulation laboratory · Locate the root of an implicit cooling step

A locally excellent algorithm can cycle

Consider $F(x) = x^3 - 2x + 2$. Starting from zero gives $F(0) = 2, F'(0) = -2$, so Newton proposes $x_1 = 1$. At one, $F(1) = 1, F'(1) = 1$, and the next iterate returns to zero. The sequence will repeat indefinitely in exact arithmetic. A real root does exist between -2 and -1, where the function changes

sign. The trouble lies in the update chosen from this starting point. It does not refute the local convergence result near a simple root.

For coupled equations define the merit function $\phi(x) = \|F(x)\|_2^2/2$. Its directional derivative along an exact Newton step s is $\nabla\phi^T s = F^T J s = -\|F\|^2$, strictly negative when $F \neq 0$. Thus a sufficiently short step in that direction decreases the squared mismatch. This identity assumes $Js = -F$ is solved accurately. With a strongly truncated inner iteration, check the actual directional derivative before treating the step as a descent direction.

Compute an accepted step

A backtracking line search starts at $\alpha = 1$ and reduces it until $\phi(x + \alpha s) \leq \phi(x) + c\alpha\nabla\phi^T s$ with $0 < c < 1$. This sufficient-decrease condition compares actual progress with the local prediction. In the cycling example at $x = 1$ the Newton direction is $s = -1$, the initial merit is 0.5 and its slope along s is -1. With $c = 10^{-4}$, the full trial returns to zero and gives $\phi = 2$, which fails. A half-step gives $F(0.5) = 1.125$ and fails too. A quarter-step gives $F(0.75) = 0.921875$, so $\phi \simeq 0.425$, satisfying the condition.

A smaller mismatch at this point does not certify a nearby root. This example has a positive local minimum of ϕ near $x = \sqrt{2/3}$, where F' vanishes but F stays nonzero. From a poor starting point, globalisation may lead near such a stationary point. Check the residual itself, use a bracket when available and try physically motivated alternative starts. A long sequence of tiny accepted steps should prompt inspection of local derivatives or constraints rather than being counted as successful convergence.

Compare a regional model with the real residual

A trust-region method minimises the local model $m(s) = \|F + Js\|^2/2$ subject to $\|s\| \leq \Delta$. Its radius states how far the current linearisation is trusted in the chosen variable scales. After a trial, compute the ratio $\rho = [\phi(x) - \phi(x + s)]/[m(0) - m(s)]$ when predicted reduction is positive. A negative ratio means the true merit increased. A ratio near one indicates accurate local prediction. One can shrink the radius after a poor trial or enlarge it when the model predicts progress well near the boundary.

In the quarter-step above the linear model predicts residual $1 - 0.25 = 0.75$ and merit 0.28125, a reduction of 0.21875. The actual reduction is about 0.075, giving $\rho \approx 0.34$. The trial moved in the intended direction but the predicted gain was optimistic. Using such comparisons is especially valuable when model coefficients change sharply with state. On a simple monotone thermal stage, damped and full Newton can converge to the same root. The resulting residual and a comparison at identical time step test this expectation. Compare two nonlinear cooling solves.

Include physical bounds in a step decision

A variable representing a concentration cannot be negative. If a positive current value c receives a negative proposed correction s, then $c + \alpha s \geq 0$ imposes $\alpha \leq c/(-s)$. For a concentration of 0.04 and a direction of −0.10 in matching units, only factors up to 0.4 keep this component admissible. Starting a merit-function search at 0.36 leaves a small distance from the boundary. Across many components use the most restrictive bound before evaluating an expensive residual.

Admissibility alone cannot guarantee progress: a physically acceptable state may have a larger balance mismatch than the old one. Conversely, clipping a negative candidate to zero changes the search direction and can alter conservation. A genuinely constrained equilibrium may require a complementarity or inequality formulation rather than repeated clipping. Record which condition limited each step when diagnosing persistent stagnation near a boundary.

Derive the step along a trust-region boundary

For a smooth squared-residual objective, a local quadratic model has change $m(s) = g^T s + s^T B s/2$, often with $B = J^T J$. A trust region limits this model to $\|s\|_2 \leq \Delta$ in scaled coordinates. Suppose first that B is symmetric positive definite. If the unconstrained minimiser $-B^{-1}g$ lies inside the ball it solves the constrained problem. Otherwise the minimiser lies on the sphere. The stationarity condition there is $(B + \lambda I)s = -g$, with $\lambda > 0$ chosen so the length equals Δ.

Diagonalising $B = Q\,\mathrm{diag}(\beta_i)Q^T$ turns the length condition into $\sum_i (q_i^T g)^2/(\beta_i + \lambda)^2 = \Delta^2$. For a nonzero gradient this left-hand side decreases continuously to zero with λ, starting at the squared unconstrained-step length. Thus the boundary multiplier exists uniquely. A positive shift damps different curvature directions by different factors. The step is generally not a scalar fraction of the full unconstrained step. If B has negative eigenvalues, positive semidefiniteness of $B + \lambda I$ and possible singular boundary cases require additional treatment.

Measure what a restricted model predicted

Let $B = \mathrm{diag}(1, 4)$ and $g = (-2, -5)^T$. The full minimiser is $(2, 1.25)^T$, of norm about 2.36. Set the permitted radius to $\sqrt{2}$. At $\lambda = 1$ the shifted equations give

$$s = (2/2, 5/5)^T = (1, 1)^T,$$

with exactly the required length. The low-curvature component has been reduced to one half of its unconstrained value while the second retains four fifths. That orientation change follows from the different denominators; shortening the original step to the same norm would predict a different model decrease.

The linear term at the accepted candidate is -7. The quadratic term is $(1 + 4)/2 = 2.5$, so the predicted decrease is $7 - 2.5 = 4.5$. Suppose the actual objective is this local model plus $0.6 s_1^4$, up to an irrelevant constant. The actual decrease at $(1, 1)$ is 3.9 and the ratio of actual to predicted decrease

is $3.9/4.5 \approx 0.867$. The extra quartic term does not alter the gradient or Hessian at the base point, yet it changes the observed progress over a finite step.

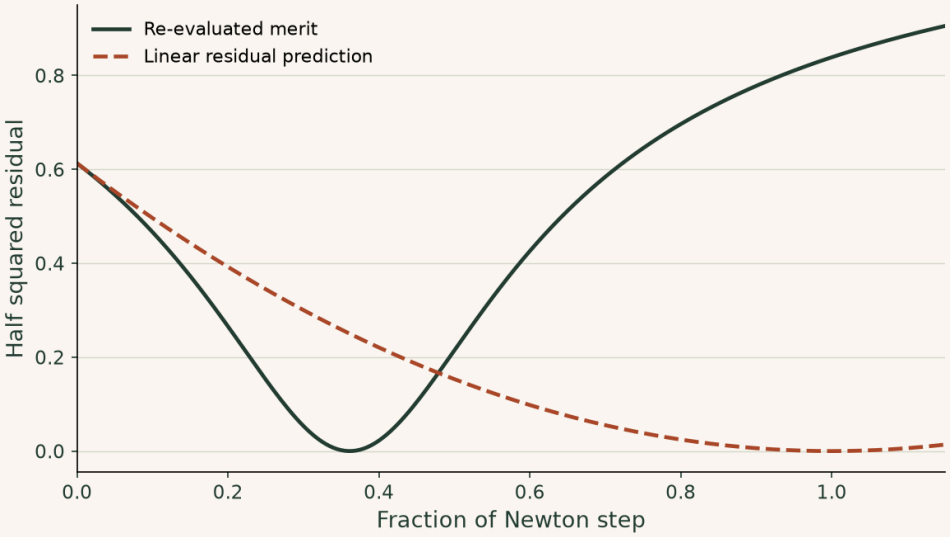
If that coefficient were 6 instead, the actual change would be an increase of 1.5 while the local quadratic still predicts a decrease of 4.5. Shrinking the radius diminishes the neglected term more rapidly than the leading linear improvement. Comparing achieved and predicted reductions supplies information about model accuracy that neither the multiplier nor the trust radius alone provides.

The choice of norm defining the ball also influences the calculation. If one coordinate is measured in millimetres and the other in metres, changing only the unit of the first coordinate alters an unscaled Euclidean ball. Instead specify physically meaningful scales and constrain the length of the dimensionless step. Transform the gradient and curvature consistently with that scaling; otherwise the shifted equations describe a different model than the one used to evaluate the actual decrease.

An equivalent specification is a weighted ellipsoid in the physical variables, with semiaxes equal to admissible changes in each parameter.

Measure how far a Newton model remains predictive

Apply Newton’s method to $\arctan x = 0$ from $x = 2$. The full step is $-5\arctan(2)$. Evaluate the true merit along that line and compare with the model that predicts zero residual at its end.



The true root lies at about 36.1% of the proposed step. At a full step the actual merit is larger than at the start while the model predicts zero. Backtracking needs fresh residual evaluations to distinguish these cases.

▶ **Read values from the plot**

Change parameters · Measure how far a Newton model remains predictive

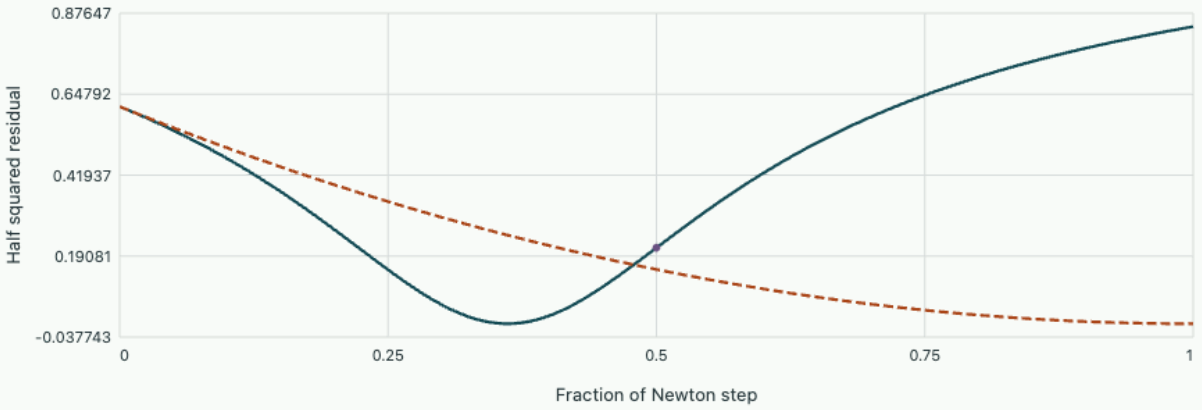
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Initial point **2**

Accepted step fraction **0.5**



Actual merit | Linear residual prediction | Chosen step

Full Newton step -5.5357. Chosen fraction changes merit by -0.39848. Negative change is improvement.

Chapter 6 · Continuation and folds

Track a family of equilibria through parameter changes, identify the failure of ordinary parameter continuation at a fold and build a predictor–corrector step.

FROM FIRST PRINCIPLES

Use the tangent of a solution curve to cross a fold

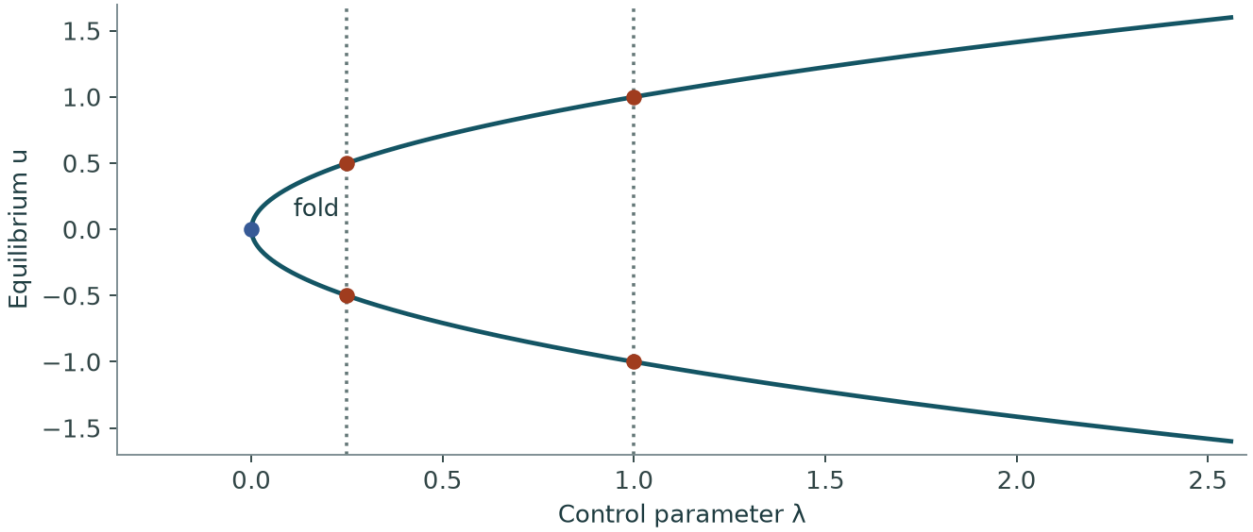
A parameter-dependent equation $F(x, \lambda) = 0$ defines a curve of equilibria when the gradient in (x, λ) is nonzero. Differentiating along any smooth parametrization by s shows that its tangent $(\dot{x}, \dot{\lambda})$ must be perpendicular to the gradient,

$$F_x \dot{x} + F_\lambda \dot{\lambda} = 0,$$
$$\dot{x}^2 + \dot{\lambda}^2 = 1.$$

The second equation fixes the length in chosen, consistently scaled coordinates. Choose the sign to match the preceding direction and predict a new state $z_p = z_0 + h t_0$. The predictor generally misses the curved solution set. To correct it, solve the original equation together with a plane condition fixing its position along the tangent,

$$F(z) = 0,$$
$$t_0^T (z - z_p) = 0.$$

Near a regular curve these two rows are independent. At a simple fold the parameter component of the tangent vanishes, so using λ alone to parameterize fails locally. The augmented system can remain nonsingular because the state component is still free to progress along the curve.



A turning point in the parameter is a regular point on the curve

For F(u, λ) = u² − λ, solving for u at prescribed λ gives du/dλ = 1/(2u), which diverges at zero. Taking u as the curve coordinate gives the regular parametrisation (λ, u) = (u², u). This picture explains why an arclength constraint can carry a branch follower through the turn.

DERIVATION AND APPLICATIONS

Predict the size and direction of a continuation correction

A tangent predictor from a known solution is first-order accurate along a smooth branch. Its leading transverse miss can be calculated before running a Newton corrector. Take a scalar equation $F(z) = 0$ for $z = (x, \lambda)$ in consistently scaled coordinates. At the starting point z_0 assume a nonzero gradient g, a unit tangent t with $g^T t = 0$, and sufficiently many bounded derivatives. Write H for its Hessian. Taylor expansion along the predictor $z_p = z_0 + ht$ yields

$$F(z_p) = \frac{h^2}{2} t^T H t + O(h^3).$$

The usual corrector plane is perpendicular to t. In two dimensions the permitted correction therefore lies along the initial unit normal $n = g/\|g\|$. Try $z = z_p + \delta n$. Since $\delta = O(h^2)$, the derivative at the initial point gives the leading normal correction,

$$0 = \tfrac{1}{2} h^2 t^T H t + \delta \|g\| + O(h^3),$$

$$\delta = -\frac{h^2 t^T H t}{2\|g\|} + O(h^3).$$

Thus a local curvature estimate $K = |t^T H t|/\|g\|$ suggests a predictor length for a desired transverse error ε,

$$\frac{1}{2} K h^2 \lesssim \varepsilon, \qquad h \lesssim \sqrt{2\varepsilon/K}.$$

The absolute curvature is invariant when F is multiplied by a fixed nonzero scalar. The estimate depends on the coordinate metric and becomes unreliable if the gradient is nearly zero or higher derivatives vary substantially over the step. A measured corrector displacement tests the prediction.

WORKED EXAMPLE

Predict the normal correction on a circle

On $F = x^2 + \lambda^2 - 4 = 0$, start at $(2, 0)$ with tangent $(0, 1)$ and choose $h = 0.20$. Then $t^T H t = 2$, $\|g\| = 4$, and the predicted normal displacement is -0.0100. At fixed $\lambda = 0.20$, the exact correction is

$$\sqrt{4 - 0.20^2} - 2 = -0.01003.$$

Its sign points inward, restoring the radius.

SECOND APPLICATION

Choose a step near a parabolic turning point

For $F = \lambda - 3x^2$ at the origin, take $t = (1, 0)$, $n = (0, 1)$. Here $t^T H t = -6,\ \|g\| = 1$, so the exact corrector at fixed $x = h$ is $+3h^2$ in the parameter. To keep it below 0.0012, choose $h \le 0.020$. A proposed $h = 0.05$ would require correction 0.0075, over six times the target.

A solution curve contains more than one isolated solve

A physical model often depends on a parameter λ, such as heat input, end displacement or applied load. Solving $F(x, \lambda) = 0$ independently at each parameter value discards information about how neighbouring states are related. On a smooth branch where F_x is invertible, differentiation gives

$$F_x \, dx/d\lambda = -F_\lambda.$$

This linear solve predicts how the state moves under a small parameter change. Re-evaluating the full residual at the prediction and correcting it by Newton accounts for the curvature omitted by the tangent.

For a scalar illustration take $F = x^2 + \lambda - 1$. The admissible solution set is a parabola in the (x, λ) plane, with branches $x = \pm\sqrt{1 - \lambda}$ for $\lambda < 1$. Its derivative $dx/d\lambda = -1/(2x)$ grows without bound as x approaches zero. Starting at $x = 0.5, \lambda = 0.75$, a change $\Delta\lambda = 0.1$ predicts $x = 0.4$. The positive exact value is

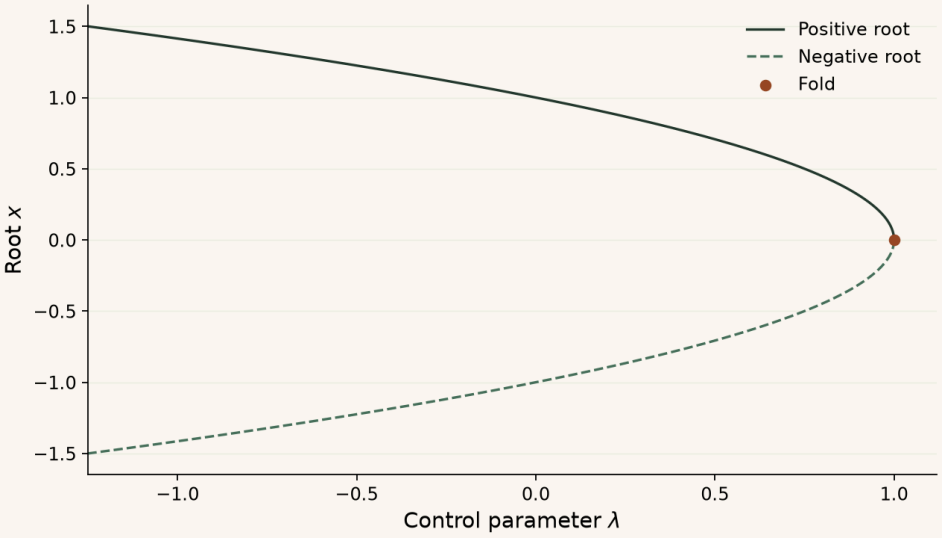
$$\sqrt{0.15} = 0.38730.$$

The predictor is close enough for a local correction here. Increasing the load beyond one would ask for a real equilibrium that this model cannot supply.

Change the parameter used to follow the curve

The curve itself is smooth at $(0, 1)$ although λ is a poor local coordinate there. Introduce an arclength-like parameter and a normalised tangent $t = (t_x, t_\lambda)$ satisfying $F_x t_x + F_\lambda t_\lambda = 0$. At a known point $z_k = (x_k, \lambda_k)$ predict $z_p = z_k + \Delta s \, t$. Correct by solving the physical equation together with a hyperplane constraint $t^T(z - z_p) = 0$. This pseudo-arclength condition selects a nearby point on the branch while allowing the load parameter itself to reverse direction.

At the fold of the example $F_x = 0,\ F_\lambda = 1$, and the tangent can be $t = (-1, 0)$. The augmented Jacobian has rows $(0, 1)$ and $(-1, 0)$, with determinant one. A Newton correction in the two unknowns is well posed even though the original scalar Jacobian for fixed λ was singular. The construction depends on a simple fold and the correct tangent. A more complicated branch intersection can require additional information to choose which path to follow.



The solution curve can be followed through the fold even though increasing the parameter beyond one would leave this particular equation with no real root.

OneSimulation · Follow an equilibrium through two load reversals

The dimensionless restoring law $F(x, \lambda) = x^3 - x + \lambda$ crosses two folds. At a prescribed load, Newton's scalar displacement update needs $1/(3x^2 - 1)$; the denominator vanishes at the turns. In the displacement–load plane the tangent t remains defined. Starting from $(1.2, -0.528)$, the code orients that tangent toward decreasing displacement, makes a prediction p, and requires the corrected point z to satisfy

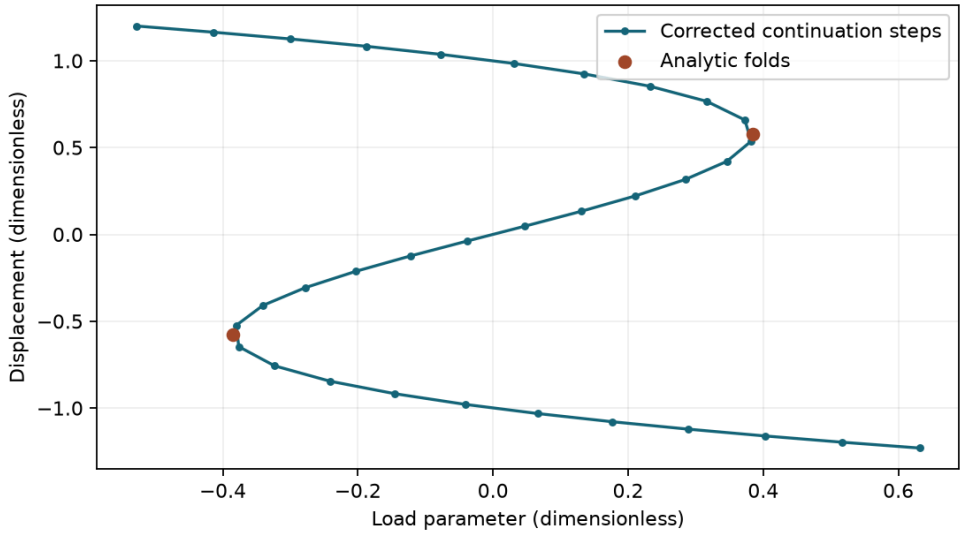
$$F(z) = 0,$$

$$t^T(z - p) = 0.$$

The OneSimulation linear solver computes each two-component Newton correction. At a fold the tangent points entirely in the displacement direction, so the second equation fixes the displacement component locally and the first equation determines the load component. Try halving the step to sample the turns more densely. Load parametrisation visits the same load at up to three equilibria; the path history identifies which one has been followed.

```
"""Follow both folds of a scalar equilibrium with a two-variable Newton correction."""
import matplotlib.pyplot as plt
import numpy as np
from onesim.numerics.linear.matrices.dense_linear_algebra import solve_dense

step = 0.12  # arclength in dimensionless displacement-load coordinates
point = np.array([1.2, 1.2 - 1.2**3])
path = [point.copy()]
for _ in range(100):
    x, load = point
    tangent = np.array([-1.0, 3*x*x - 1])
    tangent /= np.linalg.norm(tangent)
    predicted = point + step * tangent
    point = predicted.copy()
    for _ in range(10):
        x, load = point
```

```
        residual = np.array([x**3 - x + load, tangent @ (point - predicted)])
        if np.linalg.norm(residual, np.inf) < 1e-12:
            break
        jacobian = np.array([[3*x*x - 1, 1], tangent])
        point -= solve_dense(jacobian, residual)
    else:
        raise RuntimeError("Reduce the continuation step")
    path.append(point.copy())
    if point[0] < -1.2:
        break

x, load = np.array(path).T
fold_x = np.array([-1, 1]) / np.sqrt(3)
assert x[-1] < -1.2 and np.max(np.abs(x**3 - x + load)) < 1e-11
assert np.count_nonzero(np.diff(np.sign(np.diff(load)))) == 2
print("Fold loads from the derivative:", fold_x - fold_x**3)
fig, ax = plt.subplots(figsize=(6.8, 3.8), constrained_layout=True)
ax.plot(load, x, ".-", color="#176879", label="Corrected continuation steps")
ax.scatter(fold_x-fold_x**3, fold_x, color="#a4492b", label="Analytic folds", zorder=3)
ax.set(xlabel="Load parameter (dimensionless)", ylabel="Displacement (dimensionless)")
ax.grid(alpha=0.2)
ax.legend()
fig.savefig("nonlinear-equations.png", dpi=160)
```



The markers follow the middle branch through both folds. The isolated highlighted points use the exact criterion $3x^2 - 1 = 0$, giving load magnitudes $2/(3\sqrt{3})$.

Download the Python example

Take one step past a turning point

Start exactly at the fold, choose the negative-x tangent and take $\Delta s = 0.1$. The predictor gives $(-0.1, 1)$. Its physical residual is 0.01. The transverse plane now fixes $x = -0.1$. Correcting the physical balance therefore gives

$$\lambda = 1 - 0.1^2 = 0.99.$$

The parameter has decreased while the curve coordinate has advanced. This simple calculation illustrates why choosing fixed increments of λ alone cannot traverse both sides of the same load maximum.

A solver can report an accurate branch even if a physical system would not settle on that branch. To test stability one needs a dynamical law. For instance, if $\dot{x} = 1 - \lambda - x^2 = -F$, then perturbing an equilibrium gives

$$\dot{e} = -2x_\star e.$$

The positive branch is locally attracting and the negative branch repelling. The algebraic root residual vanishes on both. For a different specified dynamics the interpretation must be rederived. Keep separate records of equilibrium balance, tangent orientation, correction size and dynamical stability when interpreting a continuation diagram.

Check the prediction error against the step length

For the example, predict away from $(0.5, 0.75)$ with the slope $dx/d\lambda = -1$. A change of only 0.05 predicts $x = 0.45$, while the exact positive branch gives $\sqrt{0.20} = 0.447214$, an error about 0.00279. The larger change 0.10 gave error about 0.01270. Near this starting point halving the parameter step reduces the tangent prediction error by roughly a factor of four, consistent with the second-order remainder of a first-order predictor.

Approaching the fold makes that comparison deteriorate because the second derivative with respect to the load becomes unbounded. Monitoring the ratio of correction length to predictor length provides one way to shorten steps before Newton loses the branch. At a successful step, orient the new tangent to have positive dot product with its predecessor in the scaled coordinates. Without that convention, an arbitrary sign change in a computed tangent could send the next predictor back along the path already followed.

Distinguish a fold from a genuine branching point

The scalar equilibrium equation $F(x, \mu) = x^3 - \mu x = 0$ describes a symmetry-breaking pitchfork. There is a trivial branch $x = 0$ for every μ and two additional real branches $x = \pm\sqrt{\mu}$ when $\mu > 0$. Ordinary parameter continuation along the trivial branch encounters no difficulty evaluating the solution on either side of zero. Yet the collection of real equilibria changes there. An algorithm that accurately follows one branch can therefore entirely miss the existence of the other two.

A fold often has $F_x = 0$ while $F_\mu \neq 0$, so the full derivative with respect to state and parameter retains rank one. At the pitchfork origin both derivatives vanish:

$$F_x = 3x^2 - \mu = 0, \; F_\mu = -x = 0.$$

Adding one arclength condition to the original scalar equation cannot restore full rank there because the equilibrium row itself is zero. Arclength continuation avoids a poor choice of independent coordinate at a

regular fold; it does not automatically select a tangent at an actual branch crossing. Higher-order terms or an independent symmetry condition carry the missing information.

Follow the stable state through a loss of stability

To distinguish algebraic solutions from dynamically attracting equilibria, embed the example in $x' = \mu x - x^3 = -F(x, \mu)$. Linearising near an equilibrium gives perturbation growth rate $\mu - 3x^2$. At $x = 0, \mu = -0.04$, the rate is -0.04 and small perturbations decay. At $x = 0, \mu = +0.04$, it becomes $+0.04$ and perturbations grow. The additional equilibria are $x = \pm 0.2$, both with linear rate $0.04 - 3(0.04) = -0.08$. A static Newton solve initialised at zero on the positive side returns zero residual immediately, even though that solution is now dynamically unstable.

To switch branches deliberately, use the state amplitude as the small parameter. Cancelling the known factor x on a nontrivial branch gives

$$\mu = x^2.$$

At $x = 0.1$ the predicted parameter is 0.01; at $x = -0.1$ the same parameter gives the mirror branch. This cancellation is valid only because the branch sought has nonzero amplitude. Dividing the original equilibrium equation by x everywhere would have silently deleted the entire trivial branch.

The potential $V = x^4/4 - \mu x^2/2$ provides an independent physical check. At $\mu = 0.04$, the trivial state has energy zero, whereas either nontrivial state has energy $(0.2)^4/4 - 0.04(0.2)^2/2 = -0.0004$. Its lower energy and positive second derivative agree with the negative linear growth rate. Near a bifurcation, record branch existence, stability and which branch a continuation procedure actually followed as three separately testable facts.

If small external imperfections break the symmetry, replacing the equation by $x^3 - \mu x - \epsilon = 0$ eliminates the exact $x = 0$ branch whenever $\epsilon \neq 0$. At $\mu = 0$ the remaining equilibrium then has amplitude $\epsilon^{1/3}$. A forcing imperfection of 10^{-6} produces displacement 0.01, much larger than a linear response of the same size. This calculation identifies why branch diagrams near an ideal symmetry-breaking point are sensitive to tiny unmodelled loads.

A careful numerical continuation can map the resulting changes in the number of equilibria before interpreting any single sweep from one initial condition.

Locate the fold created by an imperfect pitchfork

An imperfect pitchfork can be analysed by solving the equilibrium and fold conditions together. With $F = x^3 - \mu x - \epsilon$, a fold requires $F_x = 3x^2 - \mu = 0$, hence $\mu = 3x^2$. Substitution into $F = 0$ gives $-2x^3 - \epsilon = 0$, therefore $x_f = -(\epsilon/2)^{1/3}$ and $\mu_f = 3(\epsilon/2)^{2/3}$ for positive imperfection. This is a nonanalytic dependence on the forcing perturbation. Halving the perturbation multiplies the fold displacement by $2^{-1/3}$, rather than by one half as a linear sensitivity would predict.

For $\epsilon = 0.002$, the fold occurs at $x_f = -0.1, \mu_f = 0.03$. Indeed $x^3 - 0.03x - 0.002 = (x + 0.1)^2(x - 0.2)$, showing the merging pair and the separate simple positive equilibrium. At this point $F_\mu = 0.1 \neq 0$ and $F_{xx} = -0.6 \neq 0$, restoring the usual regular-fold conditions away from the perfect pitchfork. At $\mu = 0$ there is only the positive cube root of 0.002. Above 0.03 the negative pair splits into two simple real roots, so the number of equilibria becomes three. This count follows also by inspecting the local maximum at negative x and local minimum at positive x. An observed delayed change in the number of equilibrium branches can therefore measure an imperfection that a symmetric idealisation omits.

Change parameters · Continuation and folds

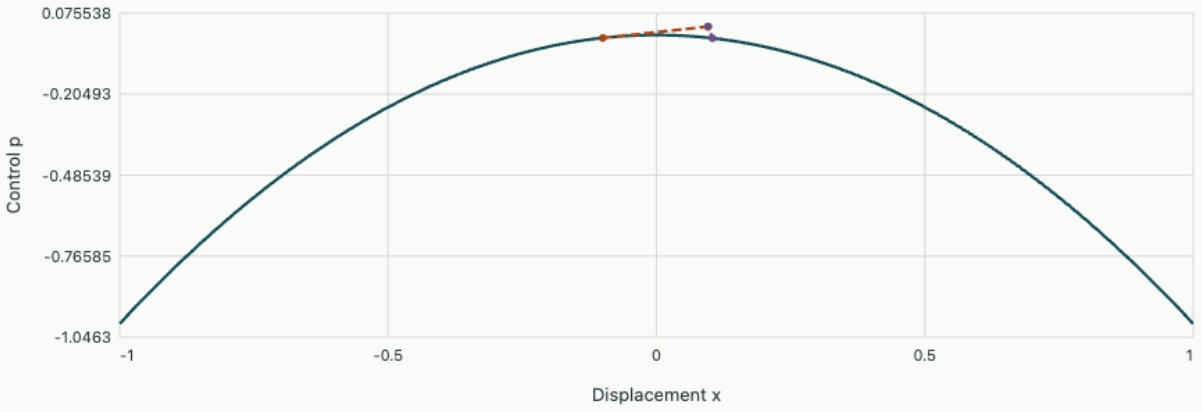
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Starting displacement **-0.1**

Tangent step length **0.2**



Predictor (0.096116, 0.029223). Nearby corrected point (0.10413, -0.010843).

Study the full derivation

Finding another equilibrium after one root is known · Nonlinear systems and iterative methods

Newton iteration repeatedly finds the same structural equilibrium. Derive how deflation changes its basins while preserving other solutions, use a rank-one Jacobian update, and distinguish a list of discovered equilibria from a completeness certificate.

Problem sheet 3 · Safeguards and continuation

Control steps within a scalar bracket, track a turning branch and distinguish equilibrium from stability under a specified evolution law.

Reject an inadmissible tangent

Solve $\ln x = 0$ on $[0.5, 4]$, starting Newton's method at 4. Compute the full trial, then two successively halved trials. Which is the first to reduce $\ln^2 x/2$? How could the bracket protect a failed update?

▼ **Worked solution**

The direction is $-4 \ln 4 \simeq -5.5452$, so the full trial is negative and the logarithm undefined. At half length the trial is 1.2274, with merit about 0.021 versus 0.961 initially, already a decrease. A quarter-step gives 2.6137, also decreasing merit but less strongly.

Reject an inadmissible trial before evaluating its logarithm. A fallback midpoint of 2.25 is admissible and cuts the sign-changing interval to $[0.5, 2.25]$; the interval maintains a root guarantee independent of tangent quality.

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

Choose the continuation coordinate

For $F(x, \lambda) = x^2 + \lambda - 4$, find the two solutions at $\lambda = 3$, their slopes $dx/d\lambda$, and the turning point. From that point take a pseudo-arclength predictor of length 0.2 in the positive-x direction and correct to the curve.

▼ **Worked solution**

At $\lambda = 3$ the roots are ± 1. Their slopes are $-1/2$ at +1 and $+1/2$ at –1. The branches meet at $(0, 4)$, where these slopes cease to be finite.

The tangent there is $(1, 0)$, so the predicted point is $(0.2, 4)$. The hyperplane constraint fixes $x = 0.2$, and the physical residual gives $\lambda = 3.96$. Its reduction in λ is compatible with forward arclength progress.

Explore the related calculation with adjustable inputs. Use the equilibrium parabola to inspect sensitivity and the tangent-coordinate continuation step. Translate the state coordinate if the fold is centred at a different value here.

Reverse the dynamics

Use the same F at $\lambda = 3$, first with $\dot{x} = -F$ and then with $\dot{x} = F$. Determine which equilibrium is stable in each model and the linearised time constant at its stable root. Can a static root-finding algorithm distinguish these dynamics from its roots alone?

▼ **Worked solution**

For $\dot{x} = -F$ the linearised rate at a root is $-2x_\star$, so +1 is stable with decay rate 2 and −1 is unstable. Reversing the evolution gives rate $+2x_\star$, hence −1 becomes stable with the same time constant $1/2$ in the time units of the equations.

Both static models have the same root set. Rate information from the specified dynamics is necessary to conclude whether a perturbation decays. Numerical convergence of Newton to a root is independent of that physical stability test.

Explore the related calculation with adjustable inputs. Use the equilibrium parabola to inspect sensitivity and the tangent-coordinate continuation step. Translate the state coordinate if the fold is centred at a different value here.

Chapter 7 · Polynomial roots and coefficient sensitivity

Evaluate a polynomial and its derivative in one sweep, then estimate how far a coefficient error can move a simple root.

FROM FIRST PRINCIPLES

Derive the change in a simple root from coefficient errors

Let $p(x) = \sum_{j=0}^{n} a_j x^j$ have a simple real or complex root r, meaning $p'(r) \neq 0$. Change the coefficients by small amounts δa_j, and suppose the root changes continuously to $r + \delta r$. Keeping first-order terms in the perturbed root equation gives

$$0 \simeq p(r) + p'(r)\delta r + \sum_{j=0}^{n} \delta a_j r^j,$$

$$\delta r \simeq -\frac{\sum_j \delta a_j r^j}{p'(r)}.$$

For a factored polynomial the derivative at a simple root has a direct interpretation,

$$p'(r_i) = a_n \prod_{j \neq i} (r_i - r_j).$$

Nearby roots can make this derivative small and the inferred root sensitive to modest coefficient uncertainty. The relation is a local estimate, with coefficient sizes and variable units affecting what counts as a small perturbation. At an exactly repeated root the derivative vanishes and the first-order expression cannot be used.

DERIVATION AND APPLICATIONS

Measure a computed polynomial root against its coefficient budget

A raw polynomial residual can be changed by multiplying every coefficient by the same constant. A more informative diagnostic asks how much the specified coefficients must change to make the reported root exact. Consider a polynomial $p(x) = \sum_{j=0}^{n} a_j x^j$, a trial real root z, and admissible coefficient changes $|\delta a_j| \leq \eta w_j$, with nonnegative weights w_j chosen from measurement tolerances. Requiring z to solve the perturbed equation gives

$$\sum_j \delta a_j z^j = -p(z), \qquad |p(z)| \leq \eta \sum_j w_j |z|^j.$$

For real data and real trial root, each term can be aligned in sign against the residual. If the denominator is positive, scaling those aligned perturbations simultaneously attains equality. The smallest admissible coefficient multiplier is therefore

$$\eta_* = \frac{|p(z)|}{\sum_j w_j |z|^j}.$$

A zero weight freezes the corresponding coefficient; taking $w_j = |a_j|$ tests componentwise relative backward error. To compare this diagnostic with a forward root error, linearize about the trial value provided $p'(z) \neq 0$ and higher terms are small,

$$r - z \simeq -\frac{p(z)}{p'(z)}.$$

A small coefficient backward error and large root displacement can coexist near a multiple root. Conversely, an accurately approximated location can appear to have a large raw residual when all coefficients carry a large common scale. This sign construction must be reconsidered if a complex trial root is tested using only real coefficient perturbations.

WORKED EXAMPLE

Compute a scale-invariant certificate for a trial root

For $p(x) = x^2 - 3x + 2$, take $z = 1.01$ and $w_j = |a_j|$. The residual is -0.0099; the denominator is

$$1.0201 + 3.03 + 2 = 6.0501.$$

The backward error is 0.00164, despite the 0.010 location error relative to the root at one. Multiplying the polynomial and all its weights by 1000 leaves the ratio unchanged.

SECOND APPLICATION

Test which coefficient is actually uncertain

For a different polynomial $p(x) = x^2 - 4$, suppose only the constant term can be adjusted, with tolerance 0.020. The trial $z = 2.01$ gives $p(z) = 0.0401$. Its required tolerance multiplier is $0.0401/0.020 = 2.005$, so it falls outside the specified data envelope. Indeed the constant would have to move from -4 to -4.0401, which directly checks the diagnostic.

A root belongs to a set of coefficients

Write $p(x) = a_n x^n + \cdots + a_0$ and suppose a simple real root r satisfies $p'(r) \neq 0$. Perturb the coefficients and write the displaced root as $r + \delta r$. Keeping first-order terms in the equation gives

$$p'(r)\delta r + \sum_{j=0}^{n} \delta a_j r^j = 0.$$

Consequently the effect of a given coefficient error depends on both the slope at the root and the size of the corresponding power. An evaluation routine can be backward accurate and still return an uncertain root when the input coefficients are insufficiently precise.

For $p = (x - 1)(x - 1.01) = x^2 - 2.01x + 1.01$, the derivative at 1 is -0.01. Increasing the constant coefficient by 10^{-6} moves this root upward by approximately 10^{-4}. The exact perturbation is easier to see by completing the square:

$$p = (x - 1.005)^2 - 0.000025.$$

The new lower root is $1.005 - \sqrt{0.000024} \approx 1.00010102$. The first-order estimate captures the direction and scale. A coefficient change of 0.000025 merges the roots; a larger upward change removes both real roots.

Carry derivative information through Horner evaluation

Collect the polynomial as a sequence of nested multiply-add operations starting with $b = a_n$. Appending the next coefficient replaces b by $xb + a_j$. Differentiating that operation shows that a derivative accumulator d must be replaced by $xd + b$, where b is its old value. Starting $d = 0$ and updating it before updating b yields $p(x)$ and $p'(x)$ without forming high powers. For $x^3 - 3x + 1$ at $x = 2$, the successive values of b are $1, 2, 1, 3$ and the derivative accumulator ends at 9. A Newton step gives $2 - 3/9 = 5/3$.

Operation count alone says little about cancellation: near a root, large terms can subtract to produce a tiny value. A scale for the accumulated terms is $s(x) = \sum |a_j||x|^j$, also computable by a Horner sweep. Reporting $|p(x)|/s(x)$ when the denominator is nonzero measures how small a relative componentwise coefficient change could explain that residual. Dividing by the slope provides an approximate forward correction for a simple isolated root.

The cost of removing a root

Dividing the polynomial by $x - \hat{r}$ after locating one root reduces subsequent work. Synthetic division leaves a remainder $p(\hat{r})$. Discarding it changes the equations for every remaining root. In the close-pair example, using $\hat{r} = 1.001$ produces quotient $x - 1.009$ and remainder -0.000009. The remaining estimate shifts by the same 0.001 magnitude. Recheck candidate roots against the original polynomial; where coefficients represent uncertain measurements, a range of compatible roots can communicate more information than many printed digits.

Choose a formula by the size of the root

For $x^2 - 10^8 x + 1 = 0$, the usual formula for the small root subtracts two numbers close to 10^8. Rounding in their last digits can dominate the small difference. Compute the large root instead by adding the positive square root and dividing by two. The product of the two roots is one, so the small root follows as its reciprocal, approximately 10^{-8}. This change removes a source of arithmetic cancellation. Whether the underlying coefficients determine the desired root accurately can then be assessed separately using the derivative-based sensitivity formula.

Count real roots before iterating on their locations

A sign change between two endpoints certifies an odd number of crossings for a polynomial when all roots in the interval are simple. It does not show whether the number is one or three. A Sturm sequence begins with the polynomial p_0 and its derivative p_1, then repeatedly sets p_{j+1} to the negative remainder from dividing p_{j-1} by p_j, stopping at a constant for a square-free polynomial. For a real test value x, count sign changes in the sequence of evaluated values, skipping zero entries. The decrease in this count between two nonroot endpoints equals the number of distinct roots between them.

The signs can change count only when some sequence member crosses zero. For an intermediate member, the remainder relation at its zero makes its two neighbours opposite in sign; the sign-change count across that triple is one on both sides. At a simple zero of the original polynomial, the next member is its nonzero derivative. Immediately to the left the polynomial has opposite sign to its derivative, and immediately to the right it has the same sign. The count therefore decreases by exactly one. These observations explain why the negative remainder sign and the assumption on multiple roots matter.

Isolate three roots of a cubic without choosing Newton starts

Take $p = x^3 - 3x + 1$ and $p' = 3x^2 - 3$. Dividing the cubic by the derivative leaves remainder $-2x + 1$, so use $p_2 = 2x - 1$. The remainder of p' on division by p_2 equals its value at $x = 1/2$, namely $-9/4$, producing final positive constant $9/4$. At $x = -2$ the signs are $(-, +, -, +)$, giving three variations. At $x = 2$ all entries are positive and the count is zero. The cubic therefore has exactly three roots in $(-2, 2)$, a more informative conclusion than the change of sign of p alone.

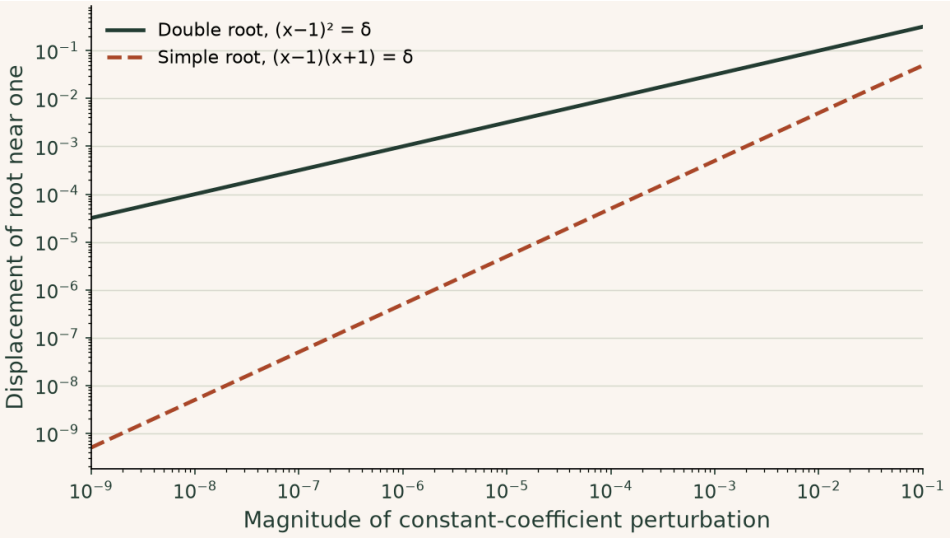
At $x = -1$, the sequence is positive, zero, negative, positive; ignoring the zero gives two variations. At zero its signs are $(+, -, -, +)$, also two. At $x = 1$ the signs are negative, zero, positive, positive, giving one variation. It follows that $(-2, -1)$, $(0, 1)$ and $(1, 2)$ each contain a single root, whereas $(-1, 0)$ contains none. Bisection or safeguarded Newton can now refine each isolating interval without ambiguity about missing another real root.

For a polynomial represented by rounded measured coefficients, this count addresses the specific polynomial actually supplied. Near a repeated root, small coefficient changes can alter the number of real roots. Exact rational arithmetic or certified sign enclosures may be needed if sequence values are close to zero; evaluating a delicate remainder sequence with ordinary rounding cannot by itself certify the sign variations of the intended polynomial.

As a check on isolation, the candidate intervals account for all three roots allowed by the degree. Further sign-preserving splits in the empty interval can never uncover a real root there. This accounting reduces the risk of reporting the same attractive Newton root twice while missing a separate crossing.

Compare a simple and a repeated root under coefficient error

Start with a polynomial whose target root is at one. In both examples subtract the same small positive constant δ; select the perturbed root above one. Use a rationalised expression for the simple-root displacement to avoid subtracting close floating-point numbers.



At a coefficient change of 10^{-6} the double-root displacement is 10^{-3}, while the simple-root displacement is about 5×10^{-7}. A tiny polynomial residual is particularly weak evidence about location near a multiple root.

▶ **Read values from the plot**

Change parameters · Compare a simple and a repeated root under coefficient error

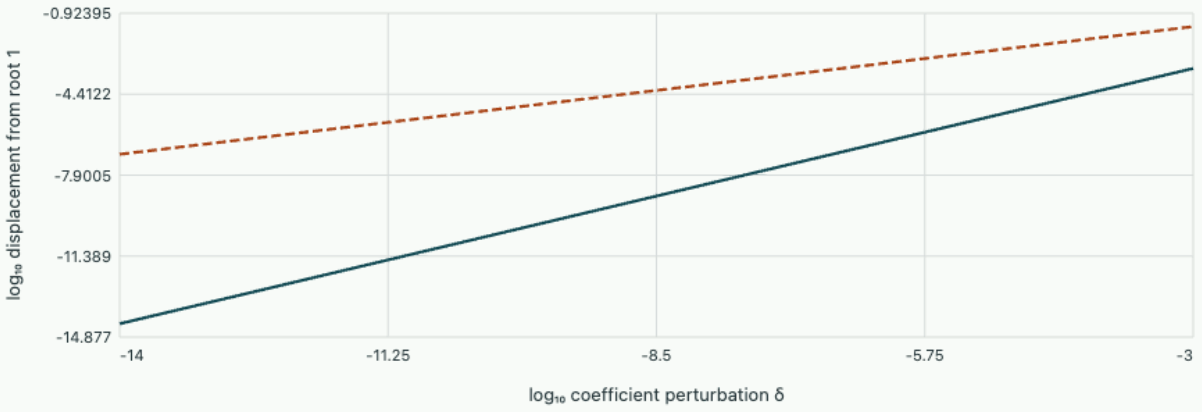
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Maximum $\log_{10}$ coefficient perturbation **-3**

Repeated-root multiplicity **2**



Simple root of $x^2 - 1 - \delta$ Root of $(x - 1)^m - \delta$

Multiplicity 2 gives displacement δ^(1/2). A polynomial residual and a root-location certificate answer different questions.

Study the full derivation

Count enclosed zeros with a contour before locating them · Nonlinear systems and iterative methods

A complex residual may have several nearby zeros whose Newton basins are hard to distinguish. Use the argument principle to count them, contour moments to estimate their locations and an independent perturbation test to certify a root cluster.

Chapter 8 · Root enclosures with interval Newton

Combine interval bounds and a single function evaluation to exclude roots or isolate one with a verifiable enclosure.

FROM FIRST PRINCIPLES

Extract a root enclosure from the mean value theorem

Suppose a differentiable scalar f has a root r in a current interval X, and choose a midpoint or other point $m \in X$. The mean value theorem relates $f(m)$ to the derivative at an unknown intermediate point. If an interval D contains every derivative on X and excludes zero, then

$$0 = f(m) + f'(\xi)(r - m), \qquad r \in m - \frac{f(m)}{D}.$$

Dividing by the entire derivative interval accounts for all possible intermediate slopes. Intersect with the original interval to retain both facts about the root,

$$X_{\text{new}} = X \cap \left(m - \frac{f(m)}{D} \right).$$

An empty intersection proves there was no root in X. A nonempty intersection alone does not establish existence; a verified sign bracket supplies a separate existence proof. Outward-rounded evaluations of the function and derivative bounds are necessary for a numerical certificate. If D contains zero, ordinary interval division cannot justify this simple enclosure.

DERIVATION AND APPLICATIONS

Include function-evaluation uncertainty in an interval Newton step

A derivative enclosure is insufficient for a reliable root bound when the value at the midpoint is itself uncertain. Suppose every root under consideration lies in a real interval X, and f' stays between two verified positive bounds $0 < d_- \le d_+$ throughout X. At $m \in X$ the numerical or measured value supplies only $f(m) \in [u, v]$. For any root r in the interval the mean value theorem now gives

$$r = m - \frac{f(m)}{f'(\xi)} \in m - \frac{[u, v]}{[d_-, d_+]}.$$

If $0 \le u \le v$, all ratios are nonnegative. The smallest ratio uses the smallest numerator and largest denominator; the largest uses the largest numerator and smallest denominator. Consequently

$$X_{\text{new}} = X \cap [m - v/d_-,\ m - u/d_+].$$

If $u < 0 < v$, both extremes instead use the smallest denominator. The new candidate interval becomes

$$X_{\text{new}} = X \cap [m - v/d_-,\ m - u/d_-].$$

These enclosures retain any root in X; existence still needs independent evidence such as verified opposite signs at the ends. A sign-uncertain midpoint can prevent further narrowing once evaluation error dominates. Tightening only the derivative interval need not improve the result materially in that regime. For machine-certified endpoints, rounded arithmetic must widen each interval as needed.

WORKED EXAMPLE

Enclose a square root with a rounded residual

For $f(x) = x^2 - 2$ on $[1.4, 1.5]$, use $m = 1.45$, verified derivative interval $[2.8, 3.0]$, and a coarse residual enclosure $[0.102, 0.103]$. Then $X_{\text{new}} = [1.45 - 0.103/2.8,\ 1.45 - 0.102/3.0]$, which is contained in the outward-rounded interval $[1.41321, 1.41600]$. The endpoint signs of the original interval establish existence, and the positive derivative makes the root unique.

SECOND APPLICATION

Recognize the resolution floor of a sensor

A separate monotone balance has derivative in $[4, 5]$. At $m = 3$, its instrument reads zero within ± 0.020. The sign-uncertain formula gives $r \in [2.995, 3.005]$ for any root in the working bracket. To certify a half-width 0.001 using the same lower derivative bound, the residual measurement must be resolved to ± 0.004 or better unless additional location information is available.

From a point step to an enclosure

Let f have a continuous derivative throughout a closed interval I. Choose $m \in I$ and compute an interval D containing every derivative value on I. If $0 \notin D$, any root $r \in I$ satisfies $0 = f(m) + f'(\xi)(r - m)$ by the mean value theorem. Therefore $r \in N = m - f(m)/D$. Intersecting N with I retains every possible root. An empty intersection certifies that the original interval contained none. Since the derivative has a fixed sign, there can be at most one.

Existence needs an additional argument. Opposite endpoint signs suffice by continuity. Alternatively if N lies strictly inside I, its extremal corrections force a sign change or a zero at m: for an increasing function and $f(m) > 0$, moving left by $f(m)/\min D$ decreases the function by at least $f(m)$. That point lies in I. The other signs follow by reversing the interval or taking $-f$. This argument explains why proper containment has stronger meaning than partial overlap.

Enclose a square root

For $f = x^2 - 2$ on $[1, 2]$, derivatives lie in $[2, 4]$. The midpoint gives $f(1.5) = 0.25$, so $N = [1.5 - 0.25/2, 1.5 - 0.25/4] = [1.375, 1.4375]$. Its width is one sixteenth of the starting width.

Ordinary Newton at 1.5 would give $1.4166\overline{6}$; an accurate single estimate carries a different type of evidence from an enclosure with guaranteed endpoints.

Choose $m = 1.40625$ in the new interval. Now $f(m) = -0.0224609375$ and $D = [2.75, 2.875]$. The next interval lies between $1.40625 + 0.0224609375/2.875$ and $1.40625 + 0.0224609375/2.75$, approximately $[1.414062, 1.414418]$ with endpoints rounded outward. For narrower intervals the variation in $1/f'$ shrinks proportionally to width when f' is Lipschitz and bounded away from zero. Because $f(m)$ is itself proportional to width near the root, the new width often decreases quadratically.

Keep the certificate valid

Each arithmetic operation needs outward rounding when performed in floating point. Rounding inward can exclude a root even when the algebra is exact. If a derivative interval crosses zero, ordinary interval division cannot give a single bounded correction. Subdivide and test again, or use an independent enclosure argument. For x^2 on $[-1, 1]$, repeated subdivision approaching zero encounters a true double root where the derivative vanishes. The usual simple-root contraction argument does not apply there. A sign-change test also misses this root because the function just touches the axis.

Use a stronger derivative enclosure when possible

For $f(x) = x + x^3$ on $[-1, 1]$, the derivative is $1 + 3x^2$. Multiplying two independent copies of the interval $[-1, 1]$ gives the loose product $[-1, 1]$, leading to the derivative enclosure $[-2, 4]$. Treating the expression as an actual square instead gives $[1, 4]$, excluding zero. The latter bound allows the interval Newton step at the midpoint to collapse directly onto the root zero. Both interval results enclose every derivative, but keeping algebraic dependence between two identical factors greatly improves the certificate.

Extend enclosure to a system using a preconditioned map

For a differentiable vector function F, choose an approximate inverse Jacobian C, a box X and a centre m inside it. Roots of F are fixed points of $G(x) = x - CF(x)$, and the converse holds when C is nonsingular. The integral mean-value identity expresses $G(x)$ as $m - CF(m)$ plus an averaged derivative $[I - CJ]$ acting on $x - m$. If every true Jacobian on the box is enclosed in an interval matrix, interval arithmetic produces an outer enclosure for the complete image $G(X)$. This is the basis of a Krawczyk-type inclusion test.

A simple sufficient certificate is to bound the induced infinity norm of $I - CJ(x)$ by $q < 1$ throughout the box and show that $G(X) \subseteq X$. The closed box is complete, so the contraction theorem then gives a unique root in it. In a cube of radius ρ around m, it suffices that $\|CF(m)\|_\infty + q\rho \le \rho$. If this test fails, no conclusion follows: the box may be too large, C may be a poor inverse, or the interval enclosures may have lost correlation between entries.

Certify a coupled nonlinear equilibrium with one matrix bound

Let $F_1 = x + 0.1y^2 - 1,\ F_2 = y + 0.1x^2 - 1$. Use centre $m = (0.90, 0.90)^T$, cube radius 0.05 and $C = I$. Each central residual is

$$0.90 + 0.081 - 1 = -0.019.$$

The derivative of $G(x, y) = (1 - 0.1y^2, 1 - 0.1x^2)$ has zero diagonal and off-diagonal entries $-0.2y, -0.2x$. On the cube both coordinate magnitudes are at most 0.95, so every absolute row sum is at most 0.19. The total displacement bound is

$$0.019 + 0.19(0.05) = 0.0285 < 0.05.$$

This verifies existence and uniqueness within the box without assuming that the answer has equal coordinates.

The first map evaluation gives $(0.919, 0.919)$. A tighter direct image uses the monotonic decrease of $1 - 0.1s^2$ for positive s: both coordinates land in $[1 - 0.1(0.95)^2, 1 - 0.1(0.85)^2] = [0.90975, 0.92775]$. Uniqueness together with symmetry under swapping coordinates forces the root in this box to satisfy $x = y = s$. The scalar equation $s + 0.1s^2 = 1$ then gives

$$s = -5 + \sqrt{35} \approx 0.91608,$$

which lies within the certified interval.

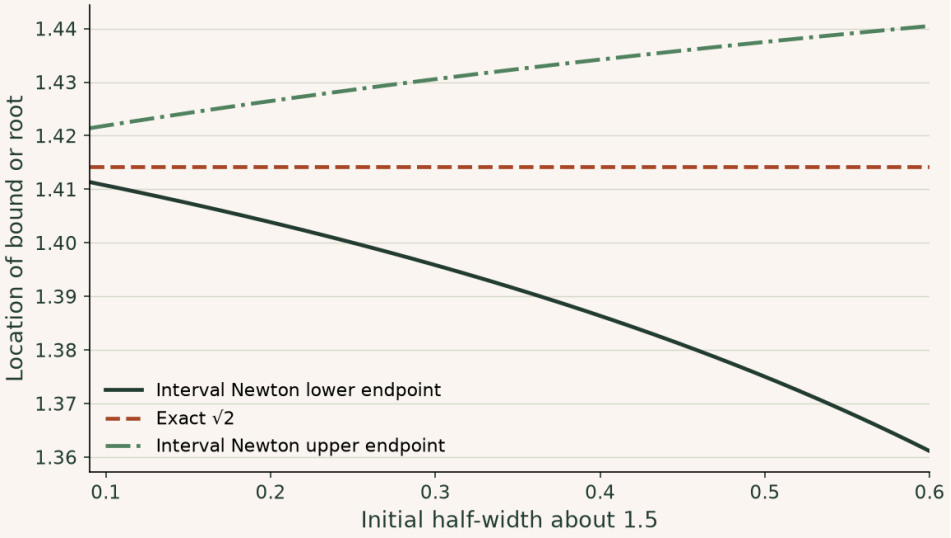
The other scalar root $-5 - \sqrt{35}$ lies far outside the box. The local certificate excludes no roots elsewhere and must not be reported as global uniqueness. In floating-point certification, round every bound outward, particularly evaluations that nearly touch a face. A convenient approximate Newton iterate can suggest the centre, while the independent box computation supplies the actual existence claim.

Consider expanding the initial certificate all the way to the cube $[-6, 6]^2$. The corresponding row-sum bound would be 1.2, violating the simple contraction criterion, and the scalar symmetry reduction exhibits two roots in that cube. Thus in this enlarged domain failure of the uniqueness test correctly reflects an actual multiplicity problem. On a moderately larger domain the very same bound might merely be too conservative. Either situation calls for subdivision or sharper derivative information, rather than interpreting a failed sufficient test as evidence of nonexistence.

Subdividing preserves the distinction because the certificate can be applied independently on each smaller closed box.

Enclose a root using a derivative interval

Start with $I = [1.5 - h, 1.5 + h]$ containing the positive root of $x^2 - 2$. By the mean value theorem the root lies in $1.5 - 0.25/[3 - 2h, 3 + 2h]$. Take care when dividing by a positive interval and subtracting it.



At h = 0.25 the new enclosure is [1.4, 1.42857…], much narrower than the original [1.25, 1.75]. These plotted endpoints use ordinary arithmetic for illustration. A certified implementation must round bounds outwards.

▶ **Read values from the plot**

Change parameters · Enclose a root using a derivative interval

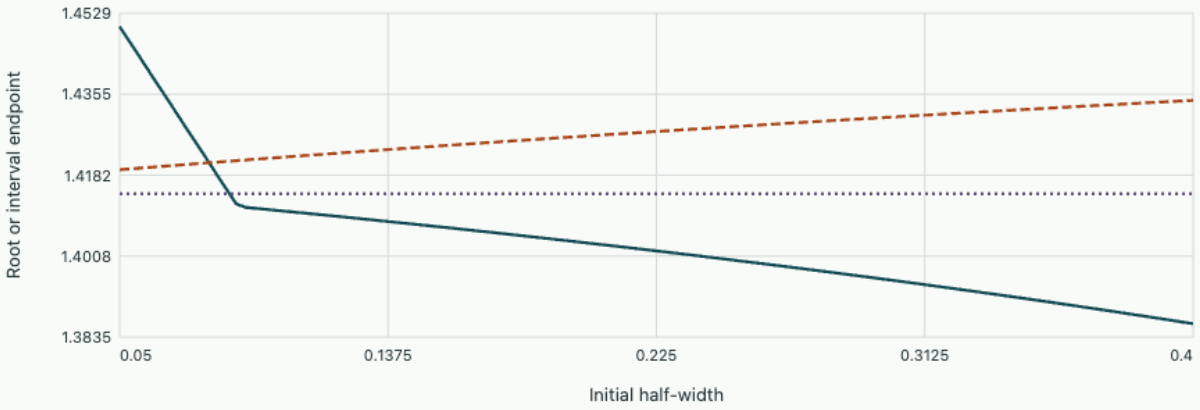
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Centre c **1.5**

Selected initial half-width **0.25**



Intersected lower endpoint Intersected upper endpoint Reference $\sqrt{2}$

Selected intersection [1.4, 1.4286]. Ordinary browser arithmetic illustrates the construction; outward rounding is required for a machine certificate.

Problem sheet 4 · Root sensitivity and certificates

Quantify the influence of coefficients and use derivative intervals to make verifiable statements about roots.

A multiple root splits

Take $p = (x-3)^2$. Add a constant perturbation δ and determine all real roots for $\delta = -0.0004, 0, +0.0004$. Explain why the usual derivative-based sensitivity formula fails.

▼ **Worked solution**

Solving $(x-3)^2 = -\delta$ gives 2.98 and 3.02 for the negative perturbation, a double root 3 for zero, and no real roots for positive perturbation.

The derivative $2(x-3)$ vanishes at 3, so dividing by this derivative is undefined. The displacement 0.02 is proportional to the square root of the coefficient error. It need not be small relative to the coefficient error itself.

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.

Evaluate before deflating

For $p = x^3 - 6x^2 + 11x - 6$, use Horner evaluation at 2.1 to obtain p, p', and the synthetic quotient upon division by $x - 2.1$. Check that the quotient and remainder reconstruct the original value at zero.

▼ **Worked solution**

The coefficient sweep gives $1, -3.9, 2.81, -0.099$. The derivative sweep uses the old coefficients and gives $0, 1, -1.8, -0.97$, so $p'(2.1) = -0.97$.

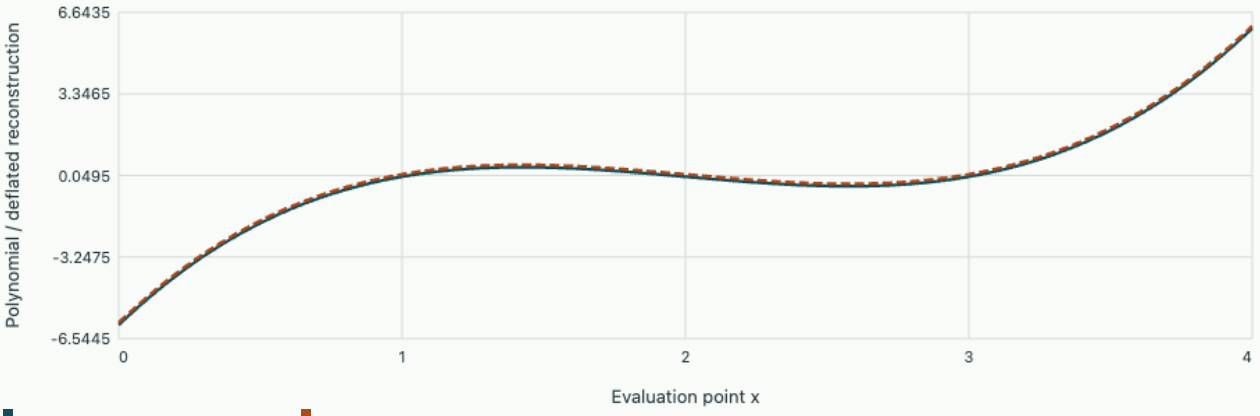
The quotient is $x^2 - 3.9x + 2.81$, with remainder -0.099. At zero this gives $(-2.1)(2.81) - 0.099 = -6$. Treating the remainder as zero would change the constant term of the reconstructed polynomial.

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

Trial divisor root a **2.1**



Original polynomial Discarded-remainder reconstruction

Quotient x² + (-3.9)x + 2.81, remainder -0.099. The derivative at a is -0.97. Dropping the remainder shifts the entire polynomial by a constant.

Contract or reject

Apply interval Newton at the midpoint to $f(x) = x^2 - 3$ on $[1.5, 2]$. Then repeat on $[2, 2.2]$. Compute both interval corrections and interpret their intersections.

▼ **Worked solution**

In the first interval $m = 1.75$, $f = 1/16$, and $D = [3, 4]$. Subtracting the quotient gives $[1.75 - 1/48, 1.75 - 1/64]$, approximately $[1.72916, 1.73438]$ outward. Opposite endpoint signs and a positive derivative give exactly one root.

In the second interval $m = 2.1$, $f = 1.41$ and $D = [4, 4.4]$. The enclosure before intersection is $[1.7475, 1.77955]$ rounded outward, entirely below 2. Its intersection is empty, excluding a root in the starting interval.

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

Chapter 9 · Secant information for systems

Derive Broyden's rank-one Jacobian update and decide whether the direction it learned is sufficient for the next step.

FROM FIRST PRINCIPLES

Update a Jacobian using what one step measured

Suppose a system has residual map F and the current Jacobian approximation is B. Moving by a nonzero displacement s yields an observed residual change $y = F(x + s) - F(x)$. A local linear map consistent with this observation must satisfy

$$B_{\text{new}} s = y.$$

Look for the smallest adjustment $D = B_{\text{new}} - B$ in Frobenius norm. For each row the minimum-length vector with prescribed dot product against s is parallel to s. Thus

$$Ds = y - Bs,$$
$$D = \frac{(y - Bs)s^T}{s^T s}.$$

Multiplying by s checks that the observation is reproduced exactly. For every vector v perpendicular to s, $Dv = 0$, so the action in such directions is preserved. The resulting rank-one update is

$$B_{\text{new}} = B + \frac{(y - Bs)s^T}{s^T s}.$$

One secant identifies only one directional response. Curvature and noise both affect y, making a tiny or poorly resolved step a weak basis for updating an entire Jacobian model.

DERIVATION AND APPLICATIONS

Make a secant update respect physical coordinate scales

The smallest change in the entries of a Jacobian depends on the units used for each input variable. To construct a scale-aware update, choose a positive definite state metric W, so that squared step length is $s^T W s$. Write scaled coordinates $z = W^{1/2} x$, leaving the output units fixed. If B approximates the derivative with respect to physical x, the corresponding scaled derivative and step are

$$\widetilde{B} = BW^{-1/2},$$
$$\tilde{s} = W^{1/2} s,$$
$$\widetilde{B}\tilde{s} = Bs.$$

Apply the ordinary smallest-Frobenius-change secant update in the scaled coordinates. Conversion back to x multiplies the resulting matrix on the right by $W^{1/2}$. Simplifying gives

$$B_{\text{new}} = B + \frac{(y - Bs)s^T W}{s^T W s}.$$

The secant equation holds for every positive definite metric. The action preserved outside the measured direction is now determined by weighted orthogonality,

$$B_{\text{new}} s = y, \qquad (B_{\text{new}} - B)v = 0 \quad \text{if } s^T W v = 0.$$

This construction assumes the selected scales describe relevant changes in each variable and that the step is observed accurately. It is invariant under a common multiplication of W by a positive scalar. Different relative scales legitimately produce different unmeasured directional predictions, since a single step cannot identify those predictions from data alone.

WORKED EXAMPLE

Compare scaled updates for an imbalanced step

Take one output, $B = (0, 0)$, step $s = (1, 10)^T$ and observed change $y = 2$. An unscaled update yields $(2, 20)/101$. If typical variable scales are 1 and 10, choose $W = \text{diag}(1, 0.01)$. Then $s^T W s = 2$, giving $B_{\text{new}} = (1, 0.1)$. Both estimates give dot product 2 on the observed step. Their predictions on $(1, 0)$ differ by a factor of 50.5.

SECOND APPLICATION

Check invariance after a unit conversion

Suppose $s = (2, 1)^T,\ W = I,\ B = 0,\ y = 5$. The new row is $(2, 1)$. Express the second step component in units one hundred times smaller:

$$s' = (2, 100)^T,\ W' = \text{diag}(1, 10^{-4}).$$

The scaled update is $(2, 0.01)$, which gives the same output change $4 + 1 = 5$. A future physical unit displacement along the second coordinate also gives the same response under either representation.

What one trajectory identifies

After a step $s = x_{k+1} - x_k$, two measured residuals provide $y = F(x_{k+1}) - F(x_k)$. Integrating the Jacobian along the connecting segment yields

$$y = \int_0^1 J(x_k + ts)s\, dt.$$

Thus y measures the action of an averaged Jacobian on just one direction. Rebuilding an unrestricted $n \times n$ Jacobian from these n numbers is underdetermined. Impose the secant equation $B_{\text{new}} s = y$ and prefer the smallest possible change to the current approximate Jacobian B in Frobenius norm.

Set $w = y - Bs$. Each row of the correction must have specified dot product with s. The shortest such row is a multiple of s^T; any added component orthogonal to s increases its squared length. Solving the scalar normalisation in each row gives

$$B_{\text{new}} = B + \frac{(y - Bs)s^T}{s^T s}.$$

The new matrix is exact on s for the observed change and leaves every vector perpendicular to s unchanged. An earlier secant equation can be disturbed if its direction has a component along the latest step. This update approximates the Jacobian; updating an inverse directly requires its own formula and conditions.

Observe two independent directions

Let $F(x) = Ax - b$, $A = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 3 \end{smallmatrix}\right)$ and $B_0 = I$. A displacement $s_1 = (1, 0)^T$ gives

$$y_1 = (2, 1)^T.$$

The first correction changes column one, giving $B_1 = \left(\begin{smallmatrix} 2 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. An independent orthogonal displacement $s_2 = (0, 1)^T$ gives $y_2 = (1, 3)^T$ and recovers the second column exactly. After these two measurements the whole linear operator is known. Such orthogonal probe steps illustrate identifiability; an actual solver selects directions from its approximate Newton equations, so its steps need not be orthogonal.

Instead repeat the first step scaled by two. The predicted and observed residual changes now agree on that direction and the update vanishes, even though B_1 is still wrong on e_2. Small secant corrections by themselves therefore do not demonstrate that all directions of the linearisation are reliable.

When to request a new Jacobian

Given B_k, solve $B_k s = -F(x_k)$, test the step and update with the displacement actually accepted. Reusing the proposed full displacement after a line search violates the secant equation for the measurements taken. If function data contain errors of size ϵ, then dividing a change in their difference by a tiny step can corrupt the learned derivative by order $\epsilon / \|s\|$. Compare achieved reduction with the predicted reduction, monitor the size of the residual itself and occasionally recompute a Jacobian if the model becomes unreliable. These diagnostics matter most when evaluating the underlying residual is substantially cheaper than assembling its derivatives.

Judge a secant across a bend

For the scalar residual $f = x^2 - 1$, moving from one to two supplies secant slope three, the average of endpoint slopes two and four. Reusing it near two underestimates the current slope by one. Over a step from 1.9 to two the secant slope is instead 3.9. This simple example isolates two types of Jacobian discrepancy: even perfectly accurate data constrain only the sampled direction, and nonlinear variation makes that directional measurement an average over a finite segment. Shorter accepted steps reduce the latter discrepancy until residual evaluation noise becomes important.

Update an inverse without rebuilding a factorisation

Suppose an invertible Jacobian approximation B has inverse H. A secant update of the form $B_+ = B + uv^T$ has inverse $H_+ = H - Huv^TH/(1 + v^THu)$ provided the scalar denominator is nonzero. Multiply this candidate by $B + uv^T$ to verify the formula: the two added rank-one terms from the individual products are cancelled by the cross term after factoring the common denominator. If the denominator vanishes, $B_+Hu = (1 + v^THu)u = 0$ and the updated matrix is singular whenever $u \neq 0$.

For Broyden's Jacobian update $u = y - Bs,\ v = s/(s^Ts)$, the denominator becomes $s^THy/(s^Ts)$. Substitution yields

$$H_+ = H + (s - Hy)s^TH/(s^THy).$$

This satisfies the inverse secant condition $H_+y = s$. It is specifically the inverse of the minimal-change Jacobian update. Directly minimising the Frobenius change in the inverse subject to this condition gives a different formula in general, because minimum changes in a matrix and its inverse are not equivalent objectives.

Distinguish harmless secant data from a singular update

Start with $B = H = I$, take $s = (1, 0)^T$ and observe $y = (2, 1)^T$. The new Jacobian replaces its first column with y and leaves its second unchanged, giving $B_+ = \left(\begin{smallmatrix} 2 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. Here $s^THy = 2$. The inverse update adds $(s - y)s^T/2$ to the identity, producing $H_+ = \left(\begin{smallmatrix} 1/2 & 0 \\ -1/2 & 1 \end{smallmatrix}\right)$. Multiplication confirms both $H_+B_+ = I$ and $H_+y = s$. Applying it to a new residual $(0, 1)^T$ leaves that residual unchanged before the sign change for the Newton correction.

Now observe $y = (0, 1)^T$ for the same step. The updated Jacobian has identical columns $(0, 1)^T$ and determinant zero. This need not mean the true Jacobian is singular. For example the perfectly invertible exchange matrix with columns $(0, 1)^T$ and $(1, 0)^T$ gives exactly that observation. The problem is that the unmeasured old second column has become incompatible with the newly learned first one. One directional observation cannot determine which of those explanations is correct.

If the observed first component were a small ϵ rather than exactly zero, the inverse update would amplify certain residuals by order $1/|\epsilon|$. Monitoring the update denominator relative to the norms of s and Hy,

and verifying the residual reduction after the step, can detect this failure mechanism. Rebuilding independent derivative information is more informative than applying repeated almost-singular rank-one corrections derived from essentially the same trajectory.

One diagnostic is to probe the unmeasured direction directly before committing to a new inverse. In the exchange-matrix example choosing an independent displacement $s_2 = e_2$ yields measured change e_1, while the singular provisional matrix predicts e_2. The disagreement localises the missing derivative information in a particular column. Returning to the original linear equations then gives an invertible Jacobian with determinant minus one. A diagonal perturbation of the singular provisional matrix would also make a solve possible, but it would not reveal which physical direction the model couples to the second coordinate. Additional measurements and algebraic regularisation address different uncertainties in the step.

In a larger system, probe selection can focus on directions where the model repeatedly gives inaccurate residual predictions.

Change parameters · Explore the worked calculation

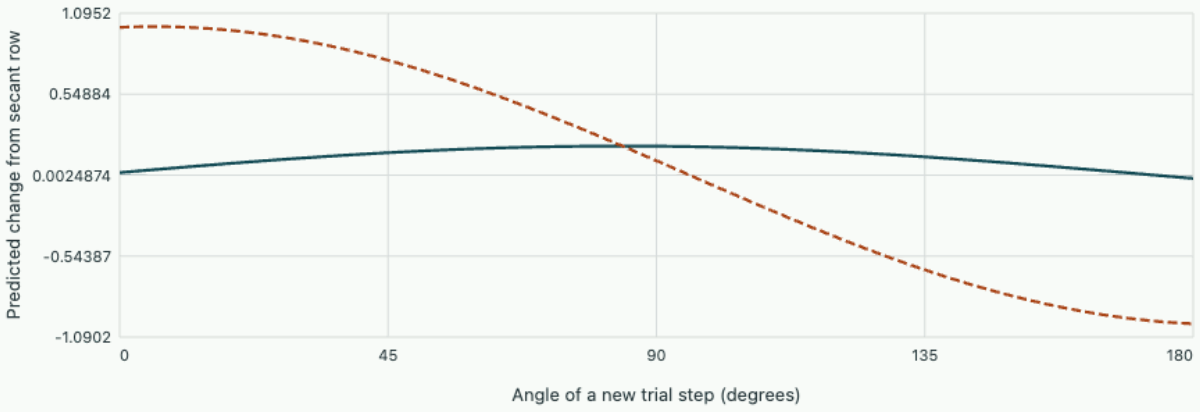
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Second step component **10**

Typical scale of second variable **10**



Both rows predict change 2 for the observed step (1, 10). They disagree in unobserved directions because the variable metric selects a different minimum-change extension.

Study the full derivation

Accelerating fixed points with differences of residuals · Nonlinear systems and iterative methods

A slowly convergent physical iteration contains information in the direction of successive errors. Use a scalar extrapolation and a small residual-minimisation problem to exploit that information, then test the extrapolated state against the original equation.

Chapter 10 · Solving Newton equations inexactly

Allocate work between the inner linear iteration and outer nonlinear iteration, and estimate the accuracy needed from a Jacobian action.

FROM FIRST PRINCIPLES

Track the residual left by an approximate Newton solve

A large nonlinear system may make an exact Newton correction too expensive. Suppose the computed step s_k solves its linearized equation only approximately. Define the linear defect $d_k = J(x_k)s_k + F(x_k)$ and require its norm to be a fraction η_k of the starting imbalance,

$$\|d_k\| \le \eta_k \|F(x_k)\|, \qquad 0 \le \eta_k < 1.$$

If the Jacobian is L-Lipschitz along the trial segment, integrating its change along the step bounds the nonlinear remainder. Substituting the observed defect into the Taylor expansion gives

$$\|F(x_k + s_k)\| \le \|d_k\| + \tfrac{1}{2} L \|s_k\|^2.$$

The first term is controllable by the inner solver; the second is inherent in using a local linear model. Solving the inner equation much more accurately than the nonlinear remainder demands can waste effort. Close to a regular root, progressively smaller forcing fractions permit faster outer convergence when derivative products and residual evaluations are also accurate enough.

DERIVATION AND APPLICATIONS

Derive a subtraction-free directional derivative for analytic residuals

A matrix-free Newton iteration needs products $J(x)v$. A complex perturbation can obtain such a product without subtracting two nearly equal real residuals. Assume each component of F is real for real arguments and has an analytic extension along $x + zv$ near $z = 0$. The implementation must propagate complex arithmetic through the same analytic formulas. Expanding a scalar component in the direction v gives

$$F(x + ihv) = F(x) + ihJv - \frac{h^2}{2} D_v^2 F - \frac{ih^3}{6} D_v^3 F + O(h^4).$$

Real coefficients make even powers real and odd powers imaginary. Extracting the imaginary part and dividing by h therefore produces a second-order directional approximation for sufficiently smooth analytic data,

$$\frac{\operatorname{Im} F(x+ihv)}{h} = Jv - \frac{h^2}{6}D_v^3 F + O(h^4).$$

The unperturbed real value never appears in a cancellation-prone subtraction. If a product approximation $\widetilde{J}v$ is used within an inexact Newton solve, the true linear defect still includes its directional error. For the proposed step s,

$$\|Js + F\| \le \|\widetilde{J}s + F\| + \|Js - \widetilde{J}s\|.$$

A small reported inner residual can conceal inaccurate Jacobian actions. An operation that discards imaginary parts, takes an absolute value, or branches differently under complex inputs violates the argument. Extreme steps can also underflow or interact with inaccurate library computations, so agreement across several steps is a useful check.

WORKED EXAMPLE

Check the truncation error on a cubic

For $F(x) = x^3,\ x = 2,\ v = 1,\ h = 0.1$, the true action is $3x^2 = 12$. The imaginary part of $(2 + 0.1i)^3$ is $1.2 - 0.001 = 1.199$, giving 11.99 after division. Halving h gives 11.9975, quartering the error exactly because all higher derivatives vanish.

SECOND APPLICATION

Expose a nonanalytic implementation

Suppose a real residual is implemented as $F(x) = |x|$. At $x = 2$, its real derivative is one. A direct complex call at $2 + ih$ returns the real modulus $\sqrt{4 + h^2}$; its imaginary quotient is zero for every h. For a proposed correction $s = -0.1$, the Jacobian action is thus wrong by 0.1, far above an inner defect budget of 0.001. Shrinking the complex perturbation cannot repair the implementation mismatch.

The linear defect survives into the next residual

For a differentiable system $F(x) = 0$, an exact Newton step cancels the first two terms in the Taylor expansion $F(x + s) = F(x) + J(x)s + R(s)$. If the Jacobian is Lipschitz continuous with constant L on the segment, integrating its variation gives

$$\|R(s)\| \le L\|s\|^2/2.$$

For a computed step define the linear defect $d = F(x) + J(x)s$. An inner iterative solver can stop once $\|d\| \le \eta_k\|F(x)\|$. Its dimensionless forcing factor η_k controls how much of the old residual may survive the linear solve.

Near a simple root assume $\|J^{-1}\| \le M$ and full steps. Then $\|s\| \le M(1 + \eta_k)\|F(x)\|$, because $Js = -F + d$. Substitution gives

$$\|F(x+s)\| \le \eta_k\|F(x)\| + \tfrac{1}{2}LM^2(1+\eta_k)^2\|F(x)\|^2.$$

A forcing factor bounded below one permits local linear convergence. A factor tending to zero gives superlinear convergence, and $\eta_k = O(\|F(x)\|)$ restores a quadratic local bound when units and scaling have been fixed. Solving early linear systems to near machine precision may waste work because model curvature dominates the outer error.

Observe the uncancelled term

For $f(x) = x^2 - 2$, start at $x = 1.5$, where $f = 0.25$ and $J = 3$. The exact step is $-1/12$; its new residual equals $s^2 = 1/144$. If the computed step is -0.075, the defect is $0.25 - 0.225 = 0.025$, a forcing factor of 0.1. The actual residual at 1.425 is $0.030625 = 0.025 + 0.005625$. Most of this residual now comes from the inner solve. Halving the step simply for caution would increase the linear defect again; a line search and an inner tolerance address different questions.

Apply a Jacobian without storing it

A Krylov solver can request a directional action through $Jv \approx [F(x + hv) - F(x)]/h$. With normalised v, a smooth residual and absolute evaluation uncertainty ϵ, the error behaves like $C_1 h + C_2 \epsilon / h$. Minimising that model gives a scale proportional to $\sqrt{\epsilon}$, with units and scales supplied by x, v and F. Automatic directional differentiation can avoid this differencing error when the residual implementation permits it.

An inner iteration using approximate products may report a small residual for its own approximate linear model. Verify a candidate full step by evaluating the actual nonlinear function, and compare the measured reduction with the forcing prediction. Once errors in evaluating F exceed the requested outer tolerance, neither decreasing the forcing factor nor making difference steps smaller can supply the claimed accuracy.

Produce a Jacobian action by forward differentiation

A matrix-free Newton solve asks for $J(x)v$. Finite differences subtract two complete residual evaluations. Forward-mode automatic differentiation instead propagates a directional derivative alongside each intermediate value. Represent an intermediate by $a + \dot{a}\varepsilon$, with $\varepsilon^2 = 0$. Addition combines the two components separately. Multiplication gives $ab + (a\dot{b} + \dot{a}b)\varepsilon$. A smooth scalar operation ϕ gives $\phi(a) + \phi'(a)\dot{a}\varepsilon$. Applying these rules to the operations of the original residual program is just repeated use of the chain rule.

Seed each independent input with its component of v. The final values give $F(x)$. The accompanying derivative components give $J(x)v$ in the same sweep. There is no differencing increment to select. Errors in ordinary floating-point arithmetic and the accuracy of the underlying residual program still matter. A branch switch or an absolute value at zero can leave the mathematical derivative undefined.

Differentiating the particular executed branch cannot supply the missing derivative. For a full dense Jacobian the method generally requires a separate direction for each independent column.

For example use the shared intermediate $s = x + y$ in $F = (se^y - 1, s^2 + x - 4)^T$. At $(2, 0)$ propagate direction $(1, -2)$. The pairs for x, y are $(2, 1)$ and $(0, -2)$. s becomes $(2, -1)$, and e^y becomes $(1, -2)$. The first output is therefore $(1, -5)$, the second $(2, -3)$. Thus $F = (1, 2)^T$ and the Krylov product is $(-5, -3)^T$. As a separate check the analytic Jacobian at this point has rows $(1, 3)$ and $(5, 4)$, whose dot products with the seed give the same two numbers.

Practise on sheet 5 · Differentiate a reciprocal with a shared input. Further reading · Massachusetts Institute of Technology · 18.S096 Matrix Calculus for Machine Learning and Beyond · Chapter 8, Forward and Reverse-Mode Automatic Differentiation, §8.1.3 Dual numbers.

Scale the forcing ratio for coupled balances

Consider two model balances expressed in newtons and millimetres. A raw Euclidean norm compares numerical values in these unrelated units. Changing the length unit alone can then make an inner solve appear more or less successful. Introduce a diagonal row scaling W based on meaningful force and length scales, and test $\|W(F + Js)\| \le \eta\|WF\|$. With characteristic magnitudes 1000 N and 0.1 mm, physical defects of 1 N and 0.001 mm become dimensionless defects 0.001 and 0.01. The second equation is now the larger unmet balance, even though its dimensional number is smaller.

Budget a matrix-free central derivative separately

When the linearised equation is solved through matrix-vector actions, the inner solver works with estimates of Jv. For a unit search direction v, set $\psi(t) = F(x + tv)$. The centred difference $[\psi(h) - \psi(-h)]/(2h)$ cancels even Taylor terms and has truncation error bounded by $Mh^2/6$ if the third directional derivative has norm at most M on the segment. If each residual evaluation carries absolute error bounded by ϵ, subtraction and division can contribute up to ϵ/h. This is a worst-case model; paired evaluations with strongly correlated errors can behave differently.

Minimising $Mh^2/6 + \epsilon/h$ gives $h = (3\epsilon/M)^{1/3}$ for positive M. This balance is specific to the centred, second-order formula and requires evaluating the residual twice for each direction. A derivative already available by an analytic action or automatic differentiation has a different accuracy and cost model. Moreover M depends on both the operating state and the direction; a single unscaled absolute step is rarely well calibrated for a system mixing quantities with strongly different units.

Compare the derivative error with the requested inner accuracy

For $f(x) = x^3 - 2$ at $x = 1$, the true derivative is three and the exact centred estimate at step h is $3 + h^2$, so $M = 6$. Let each function evaluation have absolute uncertainty at most 2×10^{-9}. The worst-case derivative error is bounded by $h^2 + 2 \times 10^{-9}/h$, minimised at $h = 0.001$. The truncation contribution is then 10^{-6} and the evaluation contribution is 2×10^{-6}, giving a combined bound of

3×10^{-6}. Reducing the step to 10^{-5} makes the evaluation contribution 2×10^{-4}, about sixty-seven times the previously balanced total error.

The exact Newton correction at this state is $s = 1/3$. If the inner solver uses the approximate scalar derivative $3 + e$, it can solve its own equation exactly by taking $s = 1/(3 + e)$. The true linear defect is then

$$-1 + 3/(3 + e) = -e/(3 + e).$$

For the balanced derivative bound this can be about 10^{-6} in size. An inner residual display near machine zero therefore need not establish a true forcing ratio of 10^{-8}. The derivative action, rather than the Krylov stopping tolerance, has become the limiting approximation.

A practical way to expose this discrepancy is to repeat a few directional actions over a geometric sequence of h values, looking first for the expected quadratic truncation trend and then for deterioration. Such comparisons estimate sensitivity to evaluation and discretisation errors; they are not a rigorous lower bound on what all directions achieve. Recompute the actual nonlinear residual after an accepted step to discover whether the reduced model accuracy still suffices for useful outer progress.

If the finite-difference increment represents a physical parameter change that crosses a material switch, the required third derivative need not exist on the segment. A tidy numerical plateau across several step sizes can then obscure a change of branch. Verify that both directional evaluations use the same smooth model region before interpreting the calculated optimum increment.

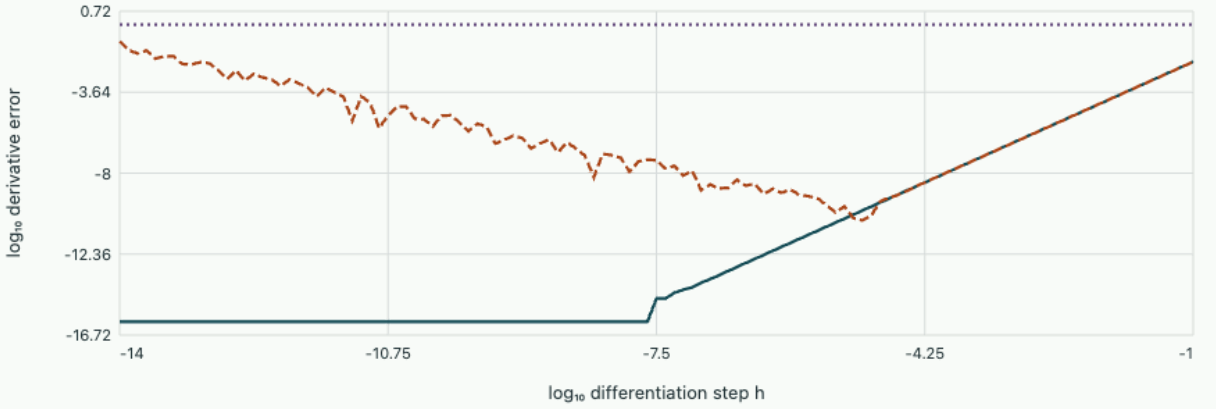
Change parameters · Explore the worked calculation

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

Evaluation point x **2**



Complex step on z^3 | Centred real difference on x^3 | Complex step applied to modulus $|z|$

Exact cubic derivative 12. For the nonanalytic modulus implementation the imaginary quotient is zero, so its error against d|x|/dx = 1 is exactly one at positive x.

Problem sheet 5 · What an approximate Jacobian knows

Construct a minimal secant update, diagnose the loss of an earlier secant equation, and budget an inexact Newton step.

A diagonal measurement direction

Given $B = I$, $s = (1, 1)^T$, $y = (3, 1)^T$, compute Broyden's updated matrix. Verify the new secant equation and its action on $(1, -1)^T$.

▼ **Worked solution**

The discrepancy is $(2, 0)^T$ and $s^T s = 2$. The correction is $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, giving $B_{\text{new}} = \left(\begin{smallmatrix} 2 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.

Multiplication by s yields $(3, 1)^T$. Multiplication by $(1, -1)^T$ yields $(1, -1)^T$, unchanged because that vector is perpendicular to the observed step.

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.

A new measurement can disturb an old one

Suppose $B = \operatorname{diag}(2, 1)$ already predicts the action of $A = \operatorname{diag}(2, 3)$ on e_1. Take $s = (1, 1)^T$, compute the updated matrix, and check both secant directions. Which subsequent probe would measure the second column directly?

▼ **Worked solution**

The new measured change is $(2, 3)^T$ while $Bs = (2, 1)^T$, so add $\left(\begin{smallmatrix} 0 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. The new matrix is $\left(\begin{smallmatrix} 2 & 0 \\ 1 & 2 \end{smallmatrix}\right)$.

Its action on $(1, 1)^T$ is exact but its action on e_1 is now $(2, 1)^T$ rather than $(2, 0)^T$. Taking $s = e_2$ would directly observe $Ae_2 = (0, 3)^T$, independently of the first column.

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.

Budget an inner defect

For $f = x^2 - 4$ at $x = 2.1$, an inner solver returns $s = -0.095$. Compute its linear defect, forcing ratio and true new residual. If the target forcing ratio were 0.02, would this step meet the target?

▼ **Worked solution**

$f = 0.41$, $f' = 4.2$, so the defect is $0.41 - 4.2(0.095) = 0.011$ and the ratio is $0.011/0.41 \approx 0.02683$.

At $x + s = 2.005$, the residual is $0.020025 = 0.011 + 0.095^2$. The permitted linear defect would be 0.0082, so this step misses the inner tolerance despite substantially reducing the outer residual.

Explore the related calculation with adjustable inputs. Explore how initial values and the residual slope affect Newton and secant convergence on the quadratic reference problem; retain the stated residual when checking the present hand calculation.

Differentiate a reciprocal with a shared input

Let $s = 1 + xy$ and $F = (s, 1/s + x)^T$. Evaluate both the residual and its Jacobian action at $(2, 1)$ in direction $(-1, 2)$ using dual arithmetic. Then propagate only the seed $(0, 1)$. State what must be checked before applying the same rules to a program that replaces s by $|s|$.

▼ **Worked solution**

For the first seed the product xy has derivative $(-1)1 + 2(2) = 3$, so $s = (3, 3)$ in value–derivative notation. The reciprocal has pair $(1/3, -3/9)$. Adding x gives

$$F = (3, 7/3)^T,$$

$$Jv = (3, -4/3)^T.$$

The second seed yields $\dot{s} = 2$ and second output derivative $-2/9$. This supplies exactly the second Jacobian column. At the stated positive s, the absolute-value operation has derivative one. At $s = 0$ the absolute value has a kink and the reciprocal is undefined, so no ordinary Jacobian action of this modified residual exists there.

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

Revisit chapter 10 · Produce a Jacobian action by forward differentiation.

Chapter 11 · Gauss–Newton and residual curvature

Derive Gauss–Newton from a local residual model and locate when omitted curvature or insufficient information changes the step.

FROM FIRST PRINCIPLES

Derive Gauss–Newton from a local model of measured residuals

Let $r(x)$ contain the differences between model predictions and measured data, expressed in comparable units or whitened by their known uncertainties. With more measurements than unknowns, exact agreement may be impossible. Minimize $\phi(x) = \|r(x)\|^2/2$. A first-order local model of each differentiable residual produces a linear least-squares problem for a step s,

$$r(x+s) \simeq r(x) + Js, \qquad \min_s \tfrac{1}{2}\|r + Js\|^2.$$

At the minimum the predicted residual is orthogonal to every column of J, yielding the Gauss–Newton equations

$$J^T J s = -J^T r.$$

If J has independent columns, the local model has a unique minimizer. In computation a QR or singular-value solve can be preferable to forming $J^T J$. Residual functions with strong curvature may make a full model step inaccurate, particularly when the best fit still has large residuals. Comparing actual and predicted decrease tests the local approximation.

DERIVATION AND APPLICATIONS

Predict local Gauss–Newton convergence when the best residual is nonzero

For a nonlinear least-squares problem a stationary point can retain a mismatch with the observations. That mismatch multiplies residual curvature in the exact Hessian and can control the iteration speed. To expose the effect, take one scalar parameter x, residuals $r_i(x)$ with two continuous derivatives, and objective $\phi = \frac{1}{2}\sum_i r_i^2$. Differentiate twice to separate two contributions,

$$g = \phi' = \sum_i r_i r_i',$$

$$\phi'' = H + S,$$

$$H = \sum_i (r_i')^2,$$

$$S = \sum_i r_i r_i''.$$

Gauss–Newton replaces the actual curvature by the always nonnegative sensitivity term H. Assume x_* has $g(x_*) = 0$ and $H(x_*) > 0$. Differentiate the update map $G(x) = x - g(x)/H(x)$. At the stationary point the term involving H' cancels exactly because $g = 0$, giving

$$G'(x_*) = 1 - \frac{\phi''(x_*)}{H(x_*)} = -\frac{S(x_*)}{H(x_*)}.$$

The local error thus behaves as

$$e_{k+1} = -\frac{S(x_*)}{H(x_*)} e_k + o(e_k),$$
$$e_k = x_k - x_*.$$

If all residuals vanish, the linear error term disappears. A strict local minimum of the true objective need not attract the full Gauss–Newton iteration: positive total curvature $H + S$ alone does not enforce $|S/H| < 1$. Damping changes the local factor by multiplying the modeled correction by a fraction.

WORKED EXAMPLE

Predict oscillation near a weakly curved mismatch

Take $r_1 = x,\ r_2 = 1 + 0.2x^2$. At $x_* = 0$, $H = 1$, $S = 0.4$, giving local factor -0.4. Starting at $x = 0.10$, $g = 0.14008$, $H = 1.0016$, and the full update lands at -0.03986, close to the linear prediction -0.040. The fitted residual of one in the second channel controls the alternating approach.

SECOND APPLICATION

Stabilize a minimum that repels full steps

For $r_1 = x,\ r_2 = 1 + x^2$, the origin is a strict minimum since $\phi''(0) = 3$. The full-step factor there is -2, so small errors grow. A half-step changes the factor to $1 - 0.5(3/1) = -0.5$, restoring local contraction. The residual value at the optimum is identical in both examples; its curvature determines the different behaviour.

More equations than unknowns

Repeated measurements generally leave a residual even at the best parameters. For $r(x) \in \mathbb{R}^m$, minimise $\Phi(x) = \|r(x)\|^2/2$. Differentiating gives

$$\nabla\Phi = J^T r.$$

At a stationary point the residual is perpendicular to every first-order change accessible by varying the parameters. It may remain large if the model cannot reproduce the observations. Different physical units require weights or a clear rescaling before squaring coordinates; otherwise a change of units silently changes what solution is optimal.

Approximate $r(x+s)$ by $r+Js$ and minimise its squared norm. The least-squares condition is

$$J^T J s = -J^T r.$$

In practice a QR or singular-value factorisation of J avoids squaring its condition number. The exact Hessian is $J^T J + \sum_i r_i \nabla^2 r_i$. Thus the neglected term vanishes at a zero-residual solution, but can matter for strongly curved residuals with a poor fit. A descent direction follows with full column rank whenever the gradient is nonzero: multiplying the normal equations by s^T gives

$$s^T \nabla \Phi = -\|Js\|^2 < 0.$$

Fit two incompatible observations

A parameter x predicts x for one sensor and x^2 for another. Suppose their dimensionless observations are 1 and 2. Then $r = (x-1, x^2-2)^T$. At $x=1$ the residual is $(0,-1)^T$ and $J=(1,2)^T$. The linear model chooses $s = 2/5 = 0.4$, moving to 1.4. The objective drops from 0.5 to $(0.4^2 + 0.04^2)/2 = 0.0808$. An exact fit is impossible because $x=1$ would require the other observation to be one.

At the original point the exact second derivative is $5-2=3$, so a full scalar Newton step on the objective would be $2/3$, quite different from the Gauss–Newton step. At 1.4 the gradient is

$$0.4 + 2.8(-0.04) = 0.288.$$

Another Gauss–Newton step is

$$-0.288/(1+2.8^2) \approx -0.03258.$$

The iteration reverses direction as the two sensor discrepancies begin to balance. Monitoring only the second sensor would favour a different parameter.

Regularise the step model

Minimising $\|r+Js\|^2 + \lambda\|s\|^2$ gives

$$(J^T J + \lambda I)s = -J^T r.$$

A positive λ controls overly large steps along weakly observed directions. With $\lambda = 5$ at the original point the step becomes 0.2 and the true objective at 1.2 is 0.1768. This is a smaller first improvement but may be more reliable elsewhere. Compare the actual decrease with that predicted by the local model to adjust damping. This temporary penalty on the step differs physically from assuming in advance that the final parameter itself must be small.

Changing one sensor's reliability changes the fit

Suppose the squared residual of the second sensor receives weight one quarter because its standard deviation is twice that of the first. In the worked example at $x = 1$, the scalar normal-equation numerator becomes $0 + (1/4)\,2(-1) = -0.5$; its denominator becomes

$$1 + (1/4)\,2^2 = 2.$$

The step is now 0.25. The model has moved less far from the exactly matched first sensor because the second sensor's discrepancy now carries less statistical weight. Repeated observations with correlated errors would require a full covariance model to interpret this weighting as a statistical fit.

Derive a bounded-influence residual fit

Ordinary least squares makes a single residual's contribution to the gradient proportional to its magnitude. If rare gross errors are plausible, choose a loss that is quadratic for small discrepancies and linear for large ones. For the Huber loss with transition $\delta > 0$, use $\rho(r) = r^2/2$ for $|r| \le \delta$ and $\delta|r| - \delta^2/2$ otherwise. Matching value and derivative at both transition points ensures a continuously differentiable convex objective in a location parameter. Its score is $\psi(r) = \max(-\delta, \min(r, \delta))$.

For differentiable model residuals $r_i(x)$ the stationarity equation becomes

$$J^T \psi(r) = 0.$$

Set $w_i = \psi(r_i)/r_i$ for nonzero residuals and one at zero. The score is now a weighted residual and solving a succession of weighted least-squares problems supplies an iteratively reweighted method. The weights must be recomputed from each current fit. Treating them as externally measured inverse variances would misinterpret their statistical role. Convexity of ρ in r does not by itself guarantee convexity with respect to nonlinear parameters.

Trace the effect of a contaminated reading

Four sensors estimate the same scalar location and report 0, 1, 2 and 20 in common standardised units. Least squares fits their mean $23/4 = 5.75$. Use $\delta = 1$ instead, with residuals $x - b_i$. In the interval $[1, 2]$ the first sensor contributes score $+1$ and the last contributes -1. The two central readings contribute $x - 1$ and $x - 2$. The sum is $2x - 3$, which vanishes at 1.5. Convexity makes this a global minimum. Increasing the last reading to 200 leaves its score at minus one throughout the same interval and does not change the fitted location.

At $x = 1.5$ the four weights are $2/3, 1, 1, 1/18.5$. A weighted-average check confirms stationarity more transparently through the signed weighted residuals: $+1, +0.5, -0.5, -1$ sum to zero. The final gross discrepancy is still 18.5 units. A robust fit limits how much that observation can move the fitted value; it does not explain why the instrument returned such a reading. Record it for investigation rather than replacing it silently by a reconstructed value.

For a physical model with differently scaled sensors, first divide discrepancies by appropriate instrument uncertainty scales or choose distinct physically defensible transition values. A single threshold measured in raw units can suppress every reading from one instrument merely because its unit is larger. Conversely an overly large threshold reduces to ordinary least squares over the whole observed residual range. Compare fits under plausible thresholds and investigate whether the prediction of interest changes, especially when a group of related sensors shares the same fault.

A gross error in an input coordinate is a further distinction. A calibration point at an extreme input can have large leverage through its Jacobian row even after its residual score has been capped. Thus the bounded score controls sensitivity to residual amplitude, while the geometry of the design still controls how much that bounded discrepancy can rotate or translate the final fit.

Change parameters · Explore the worked calculation

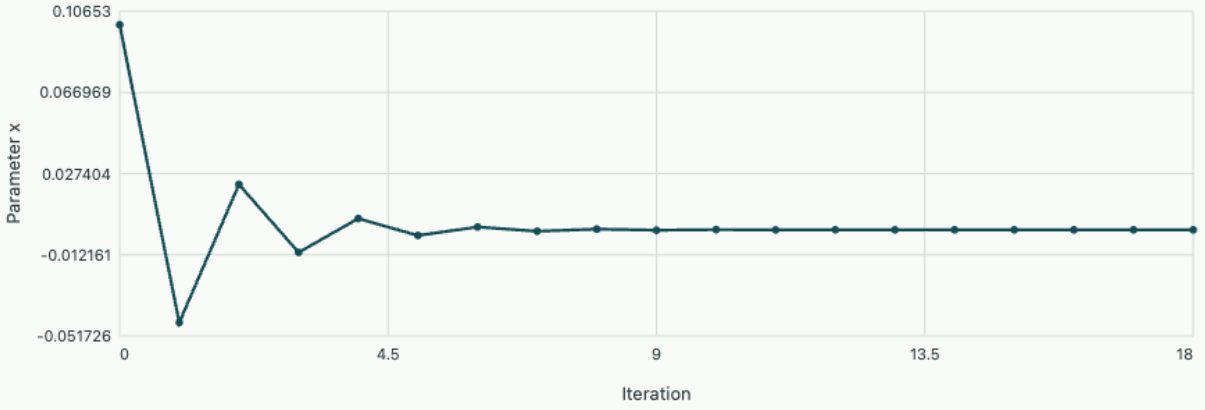
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Residual curvature c **1**

Damping fraction α **0.5**



Gauss–Newton for $(x, 1 + cx^2)$

The optimum x = 0 is a strict minimum. The local damped factor is 1 – α(1 + 2c) = -0.5. Its magnitude, not the existence of the minimum, predicts local attraction.

Chapter 12 · Inequalities and active constraints

Translate one-sided equilibrium into complementarity equations, identify an active set and verify both force and feasibility.

FROM FIRST PRINCIPLES

Express one-sided contact as a complementarity condition

Consider a point constrained to have gap $g \geq 0$ above a frictionless rigid stop. A compressive reaction $p \geq 0$ can prevent penetration, but a separated stop exerts no force. These requirements give the scalar conditions

$$g \geq 0, \qquad p \geq 0, \qquad gp = 0.$$

They allow either an open gap with zero force or closed contact with a nonnegative force. To connect this with optimization, suppose the physical coordinate is the gap itself and the unconstrained stored energy is differentiable and convex. At a constrained minimum on $g \geq 0$, every feasible first-order change has nonnegative energy change. Thus the equilibrium derivative obeys

$$E'(g) = p, \qquad p \geq 0,$$
$$gE'(g) = 0.$$

In the interior one may move in both directions, forcing zero derivative. At the boundary only increases in gap are feasible, allowing positive derivative. The balance equation for a specific spring or load supplies the remaining relationship needed to determine the gap and force.

Background · Complementarity

$$x \geq 0, \quad y \geq 0, \quad xy = 0$$

Both quantities are nonnegative and at least one vanishes. Active-set changes can make a smooth Newton model inappropriate across the switch.

DERIVATION AND APPLICATIONS

Quantify the gap introduced by smoothing a contact condition

Algebraic regularization can replace the sharp switch between an open and a closed contact with a differentiable approximation. First scale the gap and reaction into dimensionless variables g, p, using stated positive reference scales. One exact residual is $\sqrt{g^2 + p^2} - g - p$: setting it to zero implies $g + p \geq 0$, and squaring gives $gp = 0$, which together force nonnegativity. Introduce a smoothing parameter $\mu > 0$,

$$\phi_\mu(g,p) = \sqrt{g^2 + p^2 + 2\mu} - g - p.$$

At a zero, $g + p > 0$, and squaring gives $gp = \mu$. Thus both quantities are positive even where the physical contact would be exactly open or closed. For a one-degree-of-freedom linear balance $p = k(g - g_0)$ with positive scaled stiffness k and unloaded gap g_0, solve the positive quadratic root:

$$\begin{aligned} kg(g - g_0) &= \mu, \\ g_\mu &= \frac{g_0 + \sqrt{g_0^2 + 4\mu/k}}{2}, \\ p_\mu &= \frac{\mu}{g_\mu}. \end{aligned}$$

For an exactly grazing load $g_0 = 0$, the smoothing affects gap and force on square-root scales,

$$\begin{aligned} g_\mu &= \sqrt{\mu/k}, \\ p_\mu &= \sqrt{k\mu}. \end{aligned}$$

Far from this switching point the smaller of the two quantities decays linearly with μ. A small residual of the smoothed equation only proves that the regularized balance has been solved. The value of μ and the physical reference scales determine the discrepancy from exact contact.

WORKED EXAMPLE

Measure parasitic force across an open gap

Let $k = 4,\ g_0 = 0.50,\ \mu = 0.010$. The exact unsmoothed balance is

$$g = 0.50,\ p = 0.$$

The smoothed solution gives

$$g_\mu = (0.50 + \sqrt{0.26})/2 = 0.50495,$$

with $p_\mu = 0.01980$. Multiplication gives $g_\mu p_\mu = 0.0100$, an independent check on the source of this residual force.

SECOND APPLICATION

Choose smoothing at a grazing contact

With $k = 9$ and dimensionless admissible gap error 0.002, a grazing contact requires $\sqrt{\mu/9} \le 0.002$. Thus $\mu \le 3.60 \times 10^{-5}$. At this upper value the induced reaction is 0.0180. If the admissible reaction is

only 0.009, the stricter requirement is

> $$\mu \le 0.009^2/9 = 9.0 \times 10^{-6}.$$

A stop can push in only one direction

A constrained coordinate u must remain nonnegative and has an elastic energy $E(u) = ku^2/2 - fu$, $k > 0$. Minimising a differentiable energy on a half-line gives an interior stationary point when $u > 0$. At the endpoint zero, only nonnegative variations are possible, so the one-sided derivative must be nonnegative. Introduce the reaction $w = ku - f$. These conditions read $u \ge 0,\ w \ge 0,\ uw = 0$. They express that at least one of gap and reaction vanishes. A similar rule applies componentwise to a vector of gaps coupled through a stiffness matrix.

The scalar solution is

$$u = \max(f/k, 0),$$
$$w = \max(-f, 0).$$

Increasing f through zero changes the slope of the load–displacement graph. The solution stays continuous while its derivative jumps. Near that change an unconstrained differentiable root solver can produce a negative gap or an inadmissible attractive reaction unless its equations include the inequalities.

Couple two potential contact points

Let $K = \left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$, $f = (-1, 1)^T$, $w = Ku - f$ and $u_i \ge 0$. The unconstrained equilibrium gives $u = K^{-1}f = (-1/3, 1/3)^T$, violating the first gap. Test the set with $u_1 = 0$, $w_2 = 0$. The second equation becomes $2u_2 - 1 = 0$, hence $u_2 = 1/2$. Recover the first reaction $w_1 = -1/2 + 1 = 1/2$. Every gap and reaction is now nonnegative and every product vanishes. Positive definiteness of K makes the quadratic energy strictly convex, so a feasible stationary state satisfying these inequalities is its unique minimum.

For a changed load $f = (-0.4, 1)^T$, the same assumed active set yields $w_1 = -0.1$, which is inadmissible. Release the first coordinate and solve $Ku = f$. Using $K^{-1} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)/3$ gives $u = (1/15, 8/15)^T$, positive in both coordinates, with zero reactions. A small change of loading has switched the equations that determine the state.

A residual for switching equations

After compatible nondimensionalisation, $\min(u_i, w_i) = 0$ is equivalent to the three componentwise complementarity conditions. Where $u_i < w_i$, linearising the minimum selects the gap equation; where $w_i < u_i$, it selects the equilibrium equation. At equality the derivative is not unique, which explains the

need for a switching rule or a nonsmooth Newton formulation. A product alone is insufficient: a negative gap and zero reaction give a zero product. Report the largest feasibility violation and a complementarity residual together. At the scalar switching load $f = 0$, both gap and reaction vanish and the active designation is ambiguous even though the state itself is unique.

Read the reaction as the price of a constraint

Return to a single gap with stiffness 2 and load −4. Its minimum over $u \geq 0$ occurs at zero with reaction 4. Now move the lower stop to $u \geq d$ for a small positive d. The constrained minimum becomes $E(d) = d^2 + 4d$, whose derivative at zero is 4, precisely the original reaction. A small imposed displacement of the active stop requires work to first order. For a positive load producing a strictly open gap, the same small stop shift would initially have no effect and the reaction would be zero. These energy changes provide an independent check on the reaction sign.

Derive a discrete obstacle problem from virtual work

Consider a membrane represented by three interior displacements with fixed zero displacements at its ends. In scaled units the elastic energy is one half the sum of squared neighbouring differences, including the terms at both supports. Its stiffness matrix has diagonal two and neighbouring entries minus one. An obstacle at height g_j requires

$$u_j \geq g_j.$$

Minimising $E(u) = u^T K u/2 - f^T u$ over this convex feasible set gives a simple test involving every feasible variation: $(Ku - f)^T (v - u) \geq 0$ for every feasible v.

To see necessity, move a small fraction of the distance from u toward v and differentiate the energy at zero step. For sufficiency, the exact energy difference equals that linear term plus $(v - u)^T K (v - u)/2$, which is nonnegative and strictly positive for a nonzero displacement because both endpoints are fixed. At an interior free node both signs of variation are allowed, so its reaction vanishes. At an active obstacle only upward variations are feasible, so the signed reaction is nonnegative. These properties give complementarity between the gap $u - g$ and the reaction $Ku - f$.

Determine which part of an obstacle actually supplies force

Let $g = (0, 1, 0)^T$, $f = 0$. Setting all displacements equal to the obstacle would give

$$Ku = (-1, 2, -1)^T,$$

with negative endpoint reactions at the first and third interior nodes. Those two active guesses are inadmissible. Keep only the middle node in contact and solve free-node equilibrium: $2u_1 - u_2 = 0$ and $2u_3 - u_2 = 0$. With $u_2 = 1$ this yields $u = (0.5, 1, 0.5)^T$, whose reactions are $(0, 1, 0)^T$. The end gaps are both positive and the central gap is zero, satisfying the complete optimality test.

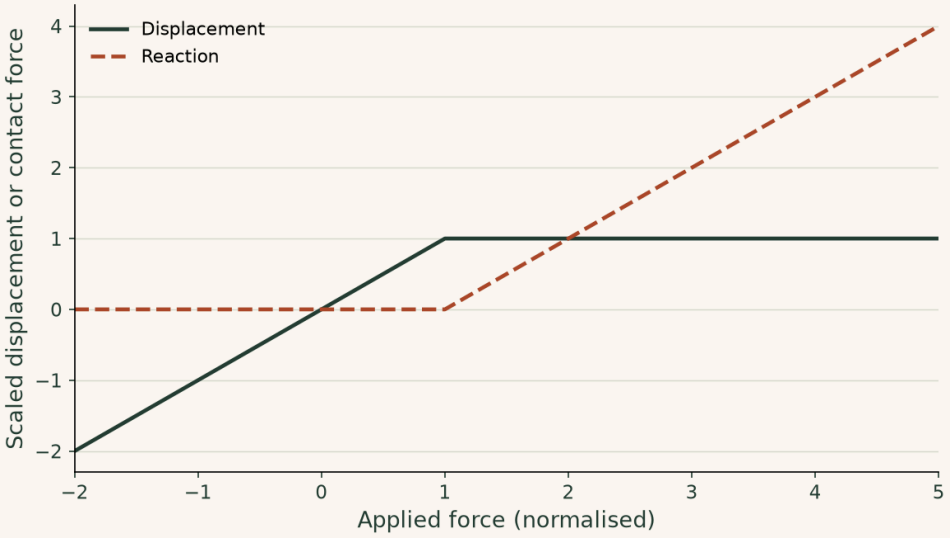
The energy is half the sum of four squared slopes $0.5, 0.5, -0.5, -0.5$, hence 0.5. The candidate that hugs the entire obstacle has slopes $0, 1, -1, 0$ and energy one. Lifting the two free nodes lets the membrane distribute curvature more economically while remaining above the obstacle. Independently the sum of the support reactions balances the unit central contact force, since each outer spring slope has magnitude 0.5 in these scaled units.

Apply an additional positive force $f_2 = t$ at the centre while keeping the obstacle fixed. The contact-state reaction becomes $1 - t$. Up to $t = 1$ the position stays fixed and the contact load decreases. For $t > 1$ the contact can no longer sustain a negative reaction; solving the unconstrained equations gives $u = (t/2, t, t/2)^T$, now above every obstacle node. This piecewise response exhibits an observable change of slope in displacement at the contact transition. Reusing the old active set without checking the sign of its reaction would incorrectly predict an unchanging displacement for arbitrarily large lifting forces.

A force test and a gap test remain necessary even when the computed energy looks sensible. For example $u = (0.4, 1, 0.4)^T$ is feasible and has energy only 0.52, close to the optimal 0.5. Yet its outer reactions are each −0.2 at strictly open gaps. The complementarity equations identify the reducible energy error even without knowing the exact energy minimum.

Locate the switch between free motion and contact

A unit-stiffness spring has free displacement equal to applied force but a rigid stop enforces $x \le 1$. With contact force $r \ge 0$ the force balance is $x + r = f$ and complementarity gives $r(1 - x) = 0$. Examine each active set.



At f = 0.5 the gap is 0.5 and contact force vanishes. At f = 2 the gap is zero and the contact supplies one force unit. The solution is continuous but its derivative jumps at engagement, a feature a smooth unconstrained Newton model misses.

▶ **Read values from the plot**

Change parameters · Locate the switch between free motion and contact

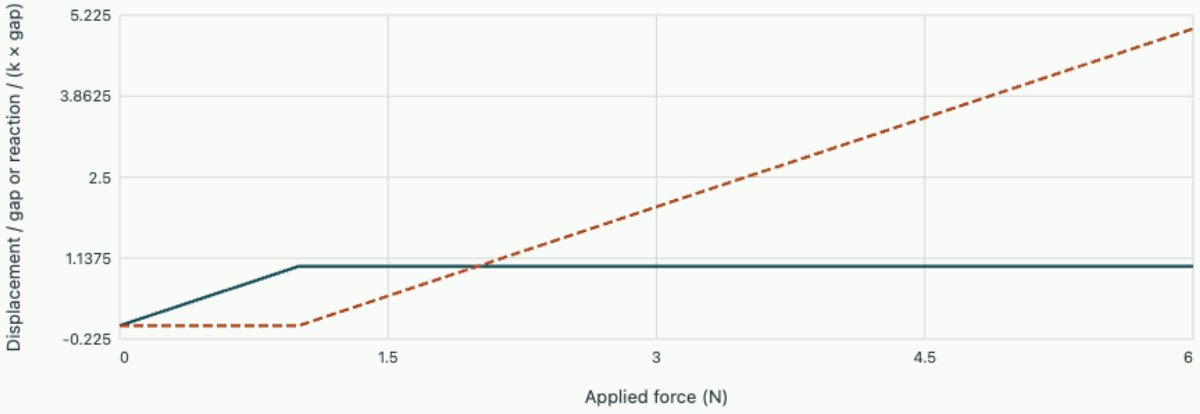
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Spring stiffness (N/m) 1

Stop displacement (m) 1



Normalised spring displacement Normalised contact reaction

Contact engages at applied force 1 N. Both displayed outputs are dimensionless: displacement divided by the stop gap, and reaction divided by k times that gap.

Problem sheet 6 · Misfit and one-sided equilibrium

Distinguish a stationary residual fit from an exact root and identify which unilateral constraints carry load.

Fit incompatible sensors

Two dimensionless residuals are

$$r_1 = x - 2,\ r_2 = 2x - 3.$$

Compute the exact least-squares state, both residuals and their dot product with the Jacobian. How many undamped Gauss–Newton steps suffice from zero?

▼ **Worked solution**

The gradient is $(x-2) + 2(2x-3) = 5x - 8$, giving $x = 1.6$. The residual is $(-0.4, 0.2)^T$ and the dot product with $(1, 2)^T$ is zero.

The residual is affine, so its linear model is exact everywhere. One step from zero produces 1.6 even though neither individual observation is exactly satisfied.

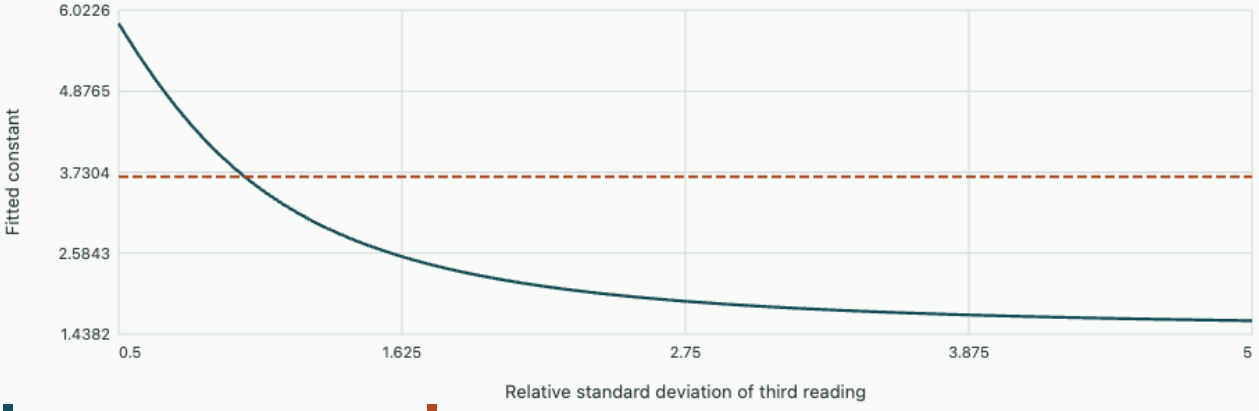
Try this calculation. Change one observation's uncertainty and recompute the weighted projection. More than one fitted coefficient requires the additional basis columns specified in this exercise.

Open this adjustable calculation

Try the calculation

Third reading **8**

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



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

A stationary point with a poor model

For the scalar residual $r = x^2 - 1$, evaluate the residual, gradient and curvature of $\Phi = r^2/2$ at zero. Analyse the undamped Gauss–Newton linear least-squares subproblem there, and compare it with a trial step to $x = 0.1$.

▼ **Worked solution**

At zero $r = -1$, $J = 0$ and $Jr = 0$. Exact curvature $J^2 + rr'' = -2$ identifies a strict local maximum.

The model $r + Js = -1$ has the same size for every step, so cannot select a descent direction. Directly testing $x = 0.1$ gives $\Phi = 0.99^2/2 = 0.49005$, below 0.5. A small gradient alone cannot certify a minimum.

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

Locate the switching load

Use $K = \left(\begin{smallmatrix} 3 & -1 \\ -1 & 2 \end{smallmatrix}\right)$, $f = (t, 2)^T$, $u, w \geq 0$, $w = Ku - f$, $u_i w_i = 0$. Assume the first gap is closed. Determine the range of t for which that assumption holds and solve the case $t = 0$.

▼ **Worked solution**

With $u_1 = 0$, the open second gap has $2u_2 = 2$, so $u_2 = 1$. The first reaction is $-1 - t$, nonnegative precisely for $t \leq -1$. At the threshold the closed coordinate has zero reaction.

At $t = 0$ release the first gap. Solving $3u_1 - u_2 = 0$, $-u_1 + 2u_2 = 2$ gives $(0.4, 1.2)^T$ with both reactions zero. The positive gaps verify the new active set.

Explore the related calculation with adjustable inputs. Inspect the active-set switch in the scalar contact example. The two-coordinate exercise still requires checking both inequalities with its stated stiffness matrix.

Locate and follow the two equilibria of a fold

Consider equilibria of the dimensionless evolution u′ = λ − u². For λ = 0.04 find both roots, classify their stability, and estimate the movement of the positive root when λ increases by 0.002. Then find out why stepping in λ fails near zero and construct a continuation step through the fold.

▼ **Worked solution**

The equilibria are u = ±0.2. A small perturbation obeys δu′ ≈ −2u δu, so +0.2 attracts while −0.2 repels. Both are perfectly valid zeros of the steady residual. A solver’s ability to converge to either root has no

direct implication for dynamical stability.

Differentiating $u^2 - \lambda = 0$ gives $du/d\lambda = 1/(2u)$. At the positive root the predicted change is +0.005. The exact change is $\sqrt{0.042} - 0.2 \approx 0.00494$. At the negative root the sign reverses. As the fold approaches, sensitivity diverges even though the equilibrium curve itself is the smooth parabola $\lambda = u^2$.

At $(u,\lambda) = (0,0)$ choose the tangent along increasing u. An arclength predictor of length 0.05 gives (0.05,0). Keep the correction normal to that tangent by imposing u = 0.05 together with the residual. The corrected point is (0.05,0.0025). The two-equation Jacobian remains invertible there and at the fold. A parameter sweep fixing λ below zero would have no real point to correct to.

Explore the related calculation with adjustable inputs. Use the equilibrium parabola to inspect sensitivity and the tangent-coordinate continuation step. Translate the state coordinate if the fold is centred at a different value here.

Chapter 13 · Why a flat zero slows Newton's method

Factor a residual near a multiple root, recover the missing factor in a Newton step and distinguish a small residual from a small location error.

Expand around the actual root

Write $f(x) = (x - r)^m g(x)$, with g smooth and $g(r) \neq 0$. Denote the current error by $e = x - r$. Differentiation gives

$$f' = e^{m-1}(mg + eg').$$

Away from the root the quotient stays finite even when both function and derivative become very small. Substitution into Newton's update gives

$$e_+ = e - \frac{eg}{mg + eg'} = (1 - 1/m)e + O(e^2).$$

If $m = 1$, the first-order error cancels. If $m > 1$, a fixed fraction survives. Multiplying the correction by the known multiplicity gives $e_+ = e^2 g'/(mg + eg')$, restoring local quadratic convergence. A wrong value of m can send steps across the root or make them diverge.

A contact measurement

Near tangency a gap can behave like $f = (x - 2)^2$ in dimensionless coordinates. Starting from 3 produces 2.5, 2.25 and 2.125. The gap residual falls by a factor of four per step. The location error only halves. A stopping rule $|f| < 10^{-8}$ certifies merely $|x - 2| < 10^{-4}$ in this exact model. The multiplicity-corrected step reaches 2 immediately.

Now take $f = e^2(1 + e)$ at $e = 0.1$. The residual and derivative are 0.011 and 0.23. An ordinary correction leaves error 0.05217. Twice the correction leaves 0.00435. The analytic denominator explains why this second update is no longer exact when g varies.

Compare errors over twelve iterations

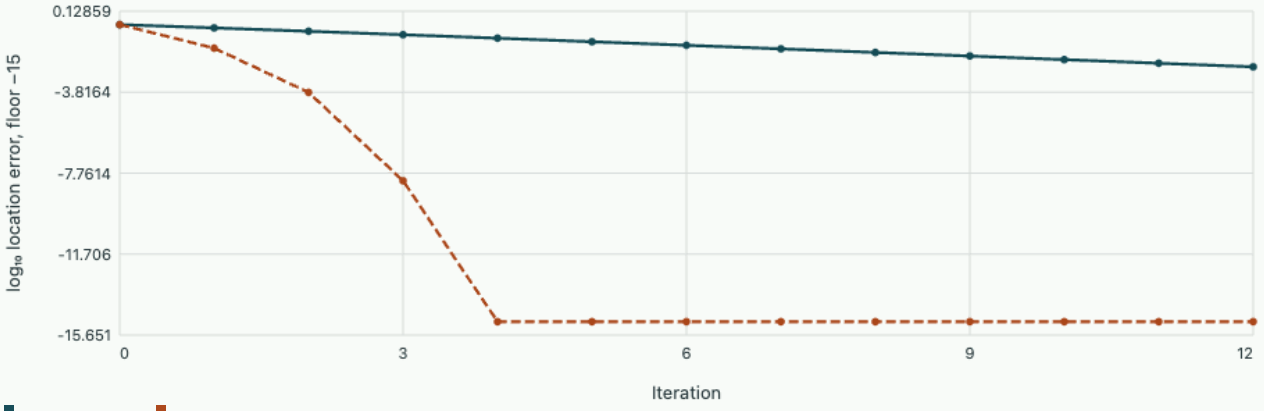
Use $f = e^m(1 + e)$ with positive starting error. Follow the ordinary and multiplicity-corrected ratios without forming underflowing powers. The logarithmic display stops at a floor of 10^{-15}. Check the slope of the ordinary iteration after the first few steps.

Open this experiment online

Try the calculation

Root multiplicity **3**

Initial distance to root **0.3**



Newton Multiplicity corrected

Asymptotic factor for ordinary Newton: 0.66667. For the corrected step the linear term vanishes.

Tutorial on error certificates

A zero of $f = e^4$ has residual 10^{-12}. How many decimal places of location accuracy follow from this observation alone?

▼ **Worked solution**

The magnitude of the location error is 10^{-3}. Four powers have compressed the residual. Any inference based on a nonzero lower bound for $|f'|$ fails across this root.

Why might a sign-changing bracket never isolate a double root?

▼ **Worked solution**

The leading sign is the sign of $g(r)$ on both sides. A sampler can step across the root without a sign change. Information about stationary points or a separate lower-bound argument is needed.

Read the university treatment

Compare the error analysis at a simple root with the extra multiplicity factor derived here. Driscoll and Braun · Solving nonlinear equations

Study the full derivation

Count enclosed zeros with a contour before locating them · Nonlinear systems and iterative methods

A complex residual may have several nearby zeros whose Newton basins are hard to distinguish. Use the argument principle to count them, contour moments to estimate their locations and an independent perturbation test to certify a root cluster.

Chapter 14 · Eliminating a fixed point's leading geometric error

Recover a limit from three consecutive iterates, derive Steffensen's update and find the cancellation that makes an apparently faster step unreliable.

Three readings of an unknown limit

Suppose $x_n = r + cq^n$ with unknown r, c, q. Consecutive differences are $c(q-1)q^n$ and $c(q-1)q^{n+1}$. Their ratio estimates q. Substituting that ratio into the sum of the geometric tail gives Aitken's delta-squared estimate

$$\hat{r} = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}.$$

For an exactly geometric sequence the denominator is $cq^n(q-1)^2$ and the cancellation is exact. For a smooth fixed-point map g with $|g'(r)| < 1$, the leading error is locally geometric, with a quadratic remainder. Replacing x_n, x_{n+1}, x_{n+2} by $x, g(x), g(g(x))$, then restarting at the estimate, gives Steffensen's iteration. Each outer step costs two map evaluations.

Two computed limits

For $g(x) = 1 + 0.8(x-1)$ and $x = 0$, the three numbers are 0, 0.2 and 0.36. The denominator is -0.04. The accelerated value is

$$0 - 0.04/(-0.04) = 1.$$

A slowly contracting map can supply a very accurate extrapolation when the data are exact.

To solve $x = \cos x$, begin at 0.5. The next two map evaluations are approximately 0.87758 and 0.63901. The accelerated value is about 0.73139. Its fixed-point residual is much smaller than the initial $\cos(0.5) - 0.5$, but plugging the result back into the original equation is essential: the geometric model is now only local.

Compare by number of map evaluations

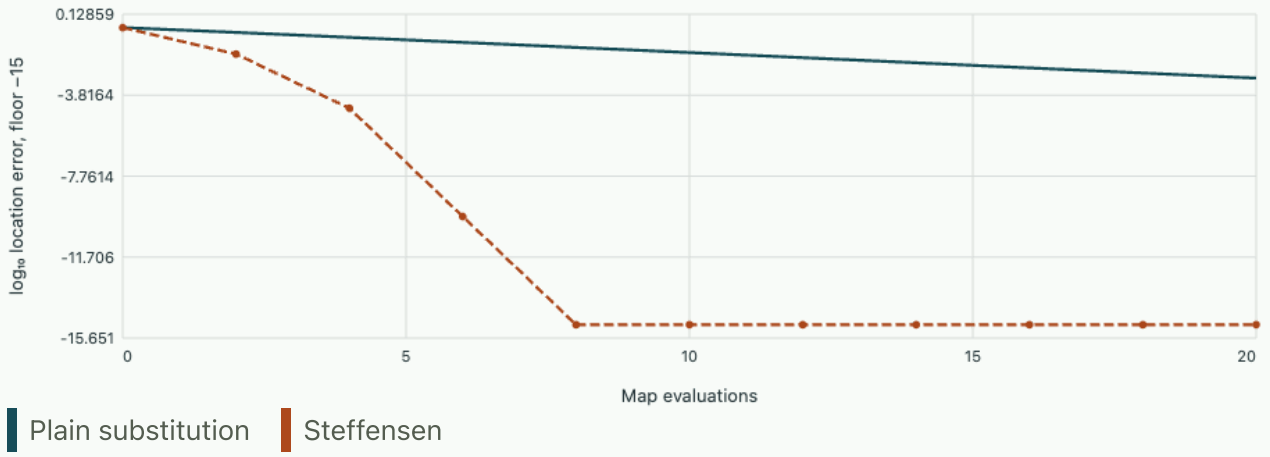
Use the smooth map $g(1+e) = 1 + qe + 0.05e^2$ close to its fixed point. Plain substitution gets one update per function call. The accelerated curve receives two calls per update. Read the horizontal axis before deciding which method made better use of the budget.

Open this experiment online

Try the calculation

Local slope q **0.75**

Initial error **0.3**



After twenty map evaluations, plain error 0.0010328 and accelerated error 0.

Tutorial on stability of the transformation

For an exact linear map with $q = 0.999$ and error one, what order of magnitude is the second difference?

▼ **Worked solution**

Its magnitude is $(1 - q)^2 = 10^{-6}$, while the three original readings may be of order one. A measurement perturbation comparable to this denominator can completely change the extrapolation.

What should an implementation do at an exact fixed point where all three readings agree?

▼ **Worked solution**

Return that fixed point after checking the residual. The delta-squared expression is then $0/0$. Away from a certified fixed point, a small denominator warrants retaining a verified plain step or a different local model.

Study the full derivation

Accelerating fixed points with differences of residuals · Nonlinear systems and iterative methods

A slowly convergent physical iteration contains information in the direction of successive errors. Use a scalar extrapolation and a small residual-minimisation problem to exploit that information, then test the extrapolated state against the original equation.

Chapter 15 · Deciding how far to trust a linearised residual

Derive an acceptance test from predicted versus actual decrease and see how it rescues a Newton direction that crosses the root by a large distance.

Choosing a step from a local model

For a scalar residual f, finding a zero also minimises $\phi = f^2/2$ if a zero is attainable. Near x, approximate $f(x+s)$ by $f(x) + f'(x)s$, yielding a quadratic model $m(s) = (f + f's)^2/2$. Minimise this model subject to $|s| \le \Delta$. In one dimension this clips the ordinary Newton step to the interval; when $f' = 0$, the model contains no preferred direction.

Measure both predicted and actual decrease. Provided the predicted decrease is positive, define

$$\rho = \frac{\phi(x) - \phi(x+s)}{m(0) - m(s)}.$$

A ratio near one supports the local model. A negative ratio means that a predicted gain produced an actual loss. Shrink the radius and repeat from the same point. A successful boundary step can justify enlarging the next radius. These decisions require an extra residual evaluation and do not require a globally accurate Taylor approximation.

Background · Trust radius

$$\|s\| \le \Delta$$

The local model is minimised on this region. Agreement between predicted and achieved reduction guides the next radius.

Two proposals for the same equation

For $f(x) = x^2 - 2$ at $x = 0.2$, $f = -1.96$, $f' = 0.4$. The full Newton proposal $s = 4.9$ lands at 5.1, where the residual is 24.01. The linear model predicted zero. The merit function rises catastrophically.

With radius one, $s = 1$ gives $x = 1.2$ and $f = -0.56$. Predicted merit decrease is

$$(1.96^2 - 1.56^2)/2 = 0.704.$$

Actual decrease is

$$(1.96^2 - 0.56^2)/2 = 1.764.$$

The ratio exceeds one because the curvature accelerated the improvement. A large ratio is possible and has a direct interpretation.

Look along the step before accepting it

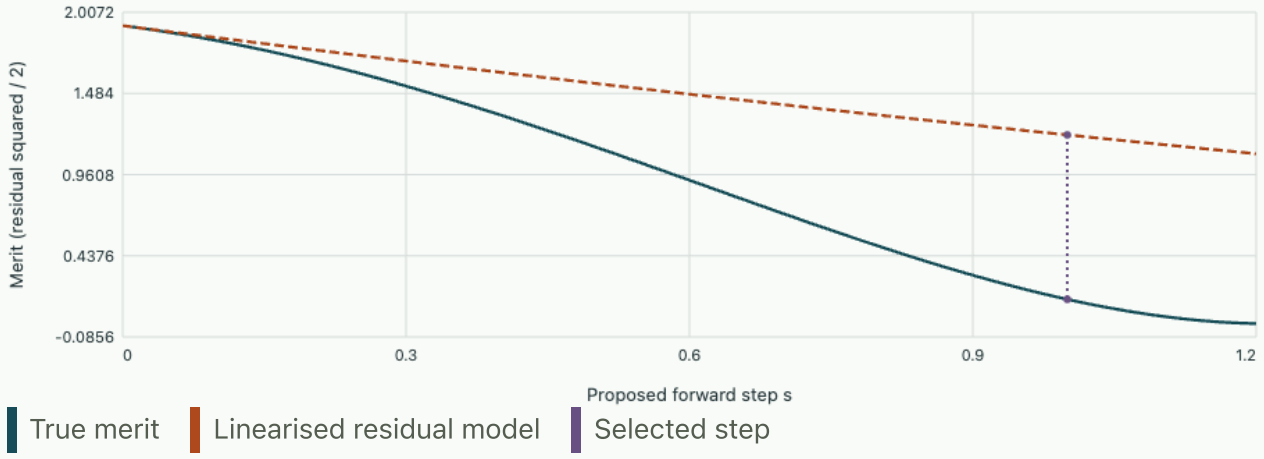
Compare the true merit function and the local model along positive steps from a starting value below the root. Move the trust radius and read the ratio for the constrained minimiser. Locate a region in which predicted and observed changes have opposite signs.

Open this experiment online

Try the calculation

Starting value **0.2**

Trust radius **1**



Proposed step 1. Actual / predicted decrease 2.5057.

Tutorial on choosing a merit function

Start at zero for the same equation. Can reducing the radius find a direction from this particular model?

▼ **Worked solution**

Here $f' = 0$ while $f = -2$. The linearised residual is constant, hence every step predicts zero decrease. The ratio is undefined. A curvature model, perturbation, or another starting point is needed.

Is a small gradient of $f^2/2$ sufficient to certify a root?

▼ **Worked solution**

Its gradient is ff'. It can vanish either at a root or at a stationary residual. Report the original residual and test feasibility to distinguish these outcomes.

Read the university treatment

Follow the nonlinear least-squares section to compare damping a proposed step with solving a constrained local subproblem. Driscoll and Braun · Nonlinear equations

Chapter 16 · A physically positive root across six decades

Reparametrise the unknown and rescale the residual in a nonlinear resistance model. Follow what each choice changes in a Newton step.

Start from a positive constitutive quantity

An ideal thermistor calibration can be written $R/R_* = e^z$. More generally a sensor fit may require a positive amplitude x satisfying its logarithmic reading $\log x = d$. A Newton correction on $f(x) = \log x - d$ gives

$$x_+ = x[1 + d - \log x].$$

Even with a unique perfectly regular root e^d, this formula can propose a negative amplitude when the initial residual exceeds one. Evaluating the next logarithm would then be impossible.

Choose $z = \log x$ instead. The equation in the new variable is $z = d$ and one step yields the answer. For an arbitrary residual $f(x)$, chain differentiation gives

$$df(e^z)/dz = xf'(x).$$

Newton moves z by $-f/(xf')$, so the update in x is multiplicative:

$$x_+ = x\exp[-f(x)/(xf'(x))].$$

Near a simple root, expanding the exponential gives the ordinary first-order correction; the difference is quadratic in the relative step. A logarithmic coordinate enforces positivity at every finite iterate. It cannot restore information missing from the measurement.

Measure failure and sensitivity separately

At $d = 0$, $x_0 = 10$ gives $x_1 = 10(1 - \log 10) \approx -13.03$ with additive Newton. Log coordinates give $x_1 = 1$. Start instead at 0.1. Additive Newton gives about 0.3303 and remains admissible. Positivity depends on the proposal relative to its scale.

If $d = \log(1000)$ with measurement uncertainty 0.01, the parameter is about 1000 with relative uncertainty 1%. The direct sensitivity follows by differentiating the equation: $dx/dd = x$. The same 0.01 measurement error near $x = 0.001$ means an absolute error near 10^{-5}, again about 1%. Changing variables improves interpretation of a relative uncertainty, without changing the information in the observation.

Inspect a Newton proposal without losing its invalid values

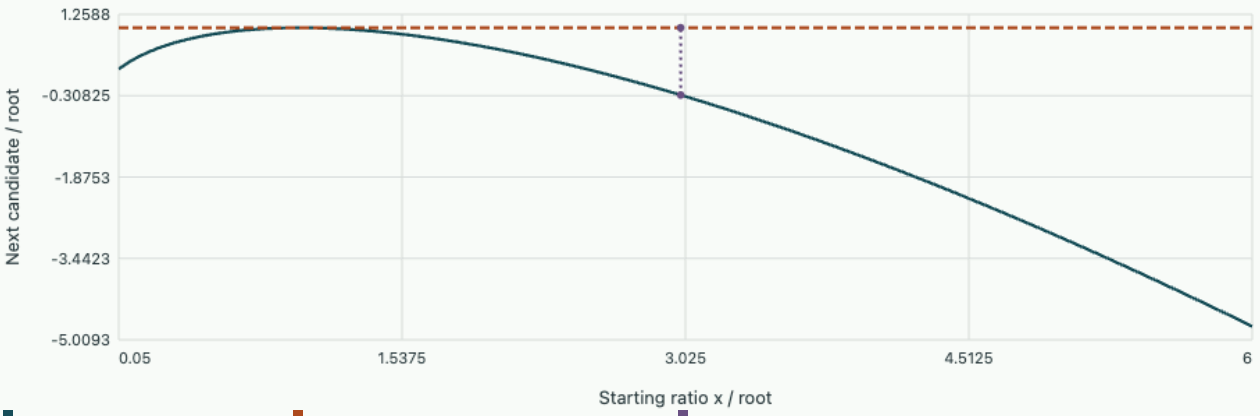
The ordinate shows the next candidate divided by the exact positive root. The horizontal axis is the starting ratio. The additive curve crosses zero at e; the logarithmic update reaches the root from every displayed start. Move the marker to read the actual candidates for a chosen target scale.

Open this experiment online

Try the calculation

Exact root (positive units) 1

Initial ratio to true root 3



Additive candidate | Log-coordinate candidate | Selected start

Starting from 3, additive Newton proposes -0.29584 while the log-coordinate step gives 1.

Tutorial on relative residuals

A model supplies $F(x) = e^{0.1x}$ and a positive observation y. How do residuals $F - y$ and $\log F - \log y$ weight small changes near a fit?

▼ **Worked solution**

Near $F = y$,

$$\log F - \log y \approx (F - y)/y.$$

A squared log residual weights errors approximately by relative error. Choosing it implies a different observation-noise model when fitting several readings.

If the true parameter can be exactly zero, what does the logarithmic coordinate miss?

▼ **Worked solution**

The zero boundary is at $z = -\infty$. A finite transformed iterate cannot represent it. One must allow a boundary solution explicitly when the physics admits it.

Chapter 17 · Finding a turning point without using the load as a coordinate

Use tangent geometry to continue across a fold, and distinguish branch sensitivity from an ill-posed choice of variable in an equilibrium computation.

Two equilibria share one control value

Use the scaled normal form $F(x, p) = x^2 + p = 0$. The equilibrium curve is $p = -x^2$. At $p = -0.01$ two displacements, $x = \pm 0.1$, give the same load. Direct continuation in p uses $dx/dp = -F_p/F_x = -1/(2x)$. The derivative diverges at the fold. Plotting p against x, however, produces a smooth curve there.

> **Background · Backward error**
>
> $$F(\hat{x}) + \delta F = 0$$
>
> Specify how perturbations in data or equations are measured. A tiny perturbation making the computed value exact can coexist with a large solution error near a fold.

Add a geometric condition

At a known point $y_0 = (x_0, p_0)$, let t be a unit tangent perpendicular to $(2x_0, 1)$. A predictor $y_* = y_0 + \Delta s\, t$ lies on the tangent. Correct it by solving $F(x, p) = 0$ together with $t \cdot (y - y_*) = 0$, which confines corrections to a line normal to the old tangent. At the fold the Jacobian of this pair is

$$J = \begin{pmatrix} 0 & 1 \\ t_x & t_p \end{pmatrix}, \qquad \det J = -t_x.$$

With the tangent aligned with the displacement axis, t_x is nonzero. The pair remains locally invertible where load control alone fails. This is a change in the computation's coordinate, not removal of the physical multiple equilibria at a fixed load.

A step that crosses the maximum

Start at $x_0 = -0.1, p_0 = -0.01$ with increasing x, so $t = (1, 0.2)/\sqrt{1.04}$. A step of length 0.2 predicts $x_* \simeq 0.0961, p_* \simeq 0.0292$, which is off the curve and has positive load. The correction line satisfies

$$x + 0.2p = x_* + 0.2p_*.$$

Substitute $p = -x^2$ to obtain $x - 0.2x^2 \simeq 0.1020$. The nearby root is about 0.1041 and the corrected load about −0.0108. Both endpoints have negative load despite crossing the maximum.

Test the predictor before trusting it

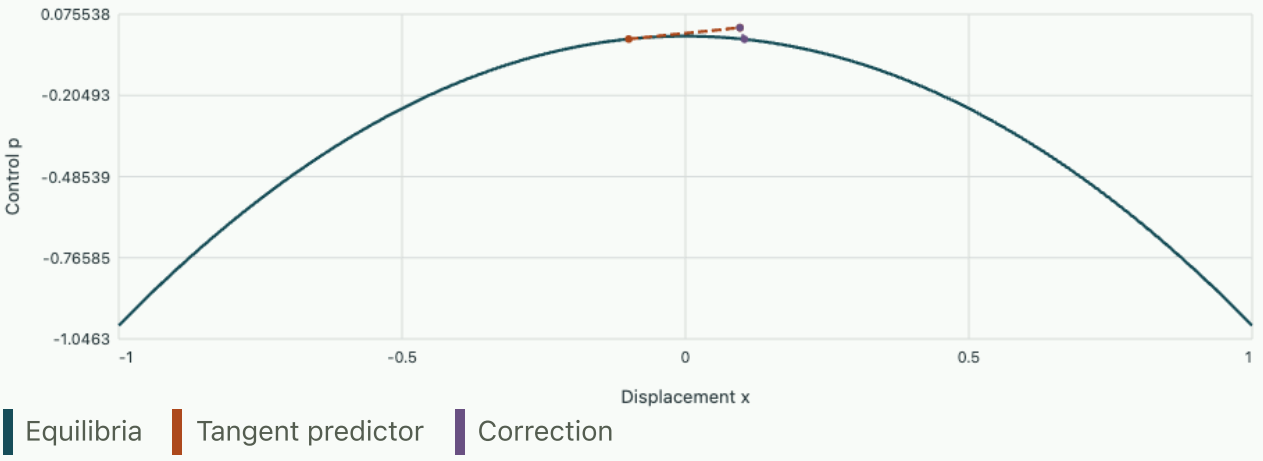
Choose the starting displacement and step length. Plot the tangent segment, the original curve and the selected nearby intersection of the correction line with the curve. A large step makes the tangent less faithful. The correction must stay on the intended local branch.

Open this experiment online

Try the calculation

Starting displacement **-0.1**

Tangent step length **0.2**



Predictor (0.096116, 0.029223). Nearby corrected point (0.10413, -0.010843).

Tutorial on the experimental meaning

At $p = -0.01$, a pressure measurement has uncertainty 0.001. Estimate its local displacement uncertainty and compare it near $p = -0.0001$.

▼ **Worked solution**

At $|x| = 0.1$, $|dx/dp| = 5$ and the linear estimate is 0.005. At $|x| = 0.01$, the formal factor is 50. The same pressure error now exceeds the distance to the fold, so the local linear approximation itself loses validity.

If the continuation algorithm traces both sides smoothly, are both equilibria dynamically stable?

▼ **Worked solution**

A static residual gives equilibrium locations. Stability requires the linearisation of a specified evolution law or a second variation of the appropriate energy. Smooth numerical traversal gives no such certificate.

Study the full derivation

Finding another equilibrium after one root is known · Nonlinear systems and iterative methods

Newton iteration repeatedly finds the same structural equilibrium. Derive how deflation changes its basins while preserving other solutions, use a rank-one Jacobian update, and distinguish a list of discovered equilibria from a completeness certificate.

Chapter 18 · Newton's method for an unknown covariance square root

A pair of independent unit-variance signals must be transformed to a prescribed covariance ellipse. One can solve a scalar square-root equation along each principal axis. A calibration that starts in a different frame exposes the genuinely matrix-valued derivative. Derive it, solve for its correction, and determine which residual will detect a wrong update.

Translate a covariance requirement into an equation

Suppose z is a zero-mean random vector with identity covariance. Let a deterministic device produce $y = Xz$. Multiplying out yy^T and averaging gives

$$\mathbb{E}(yy^T) = X\mathbb{E}(zz^T)X^T = XX^T.$$

A target covariance A is symmetric and positive semidefinite by construction. Restrict the exercise to positive definite targets. They correspond to genuine variation in every direction and allow us to discuss an invertible square root. The notebook uses two dimensionless readout channels, so mixing their coordinates has a defined physical meaning.

If the sole purpose were generating random samples, any factor satisfying $XX^T = A$ would do. Indeed, replacing X by XQ for an orthogonal Q leaves this output covariance alone. We make an extra modelling decision: the calibration transform should be symmetric and positive. This picks a distinguished factor that stretches along principal axes and introduces no orthogonal rotation. The equation becomes $X^2 = A$. A Cholesky factor produces the correct covariance too, but squaring that generally triangular factor is a different calculation.

To see existence, diagonalise $A = Q\Lambda Q^T$. The matrix $S = Q\sqrt{\Lambda}Q^T$ squares to A. For uniqueness on the positive branch, a putative root commutes with its own square and consequently preserves each eigenspace of A. Within an eigenspace of eigenvalue $a > 0$, a symmetric positive root has only positive eigenvalues whose squares equal a. It must act there as $\sqrt{a}$ times the identity. Repeated eigenvalues leave the choice of basis in that subspace free while the positive root itself stays fixed.

Differentiate before choosing a linear solver

For scalar data, expanding $(x + h)^2 = x^2 + 2xh + h^2$ supplies Newton's familiar correction. For matrices write every factor in its original order. The cross terms are XH and HX. They coincide only under a commutation condition. The first variation is therefore a linear map on a whole matrix, not multiplication by a single derivative matrix on the left.

$$F(X) = X^2 - A,$$

$$F(X + H) = F(X) + XH + HX + H^2.$$

Set the linearised residual to zero. The correction H solves the Sylvester equation $XH + HX = A - X^2$. If X and the right-hand side are symmetric, the transpose of any solution satisfies the same equation. Existence and uniqueness on the positive branch therefore guarantee a

symmetric correction. This argument is a practical implementation check: an update with substantial antisymmetric part could not have been obtained by solving this particular Newton equation accurately.

Diagonalise the current iterate $X = U\,\mathrm{diag}(d_i)U^T$. Write a tilde for the representation in the U basis. The ij entry of $XH + HX$ becomes $(d_i + d_j)\widetilde{H}_{ij}$. We can solve entry by entry and transform back. All denominators are positive if X is positive. The minimum denominator is $2d_{\min}$; as the smallest stretch approaches zero, a fixed residual can request a large correction in a weakly varying direction. This sensitivity also appears when estimating a covariance from data near a rank loss.

$$\widetilde{R} = U^T(A - X^2)U,$$

$$\widetilde{H}_{ij} = \frac{\widetilde{R}_{ij}}{d_i + d_j},$$

$$X_{\mathrm{new}} = X + U\widetilde{H}U^T.$$

A tilted measurement that defeats the scalar update

Choose the axes of the current iterate for a hand calculation and take $X = \mathrm{diag}(2, 1)$. Suppose the target covariance measured in these axes is $A = \left(\begin{smallmatrix} 5 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. It is positive because its determinant is nine and its first diagonal entry is positive. Then $R = A - X^2$ has all four entries equal to one. The diagonal corrections are $1/4$ and $1/2$. Both off-diagonal corrections equal $1/3$, so the update rotates the stretching axes as well as changing their magnitudes.

What if one used $2XH = R$? The two off-diagonal entries of that candidate would be $1/4$ and $1/2$ instead. It fails even the requirement that the correction be symmetric. Its upper off-diagonal Sylvester residual is

$$3/4 - 1 = -1/4.$$

A code that checks only whether $2XH - R$ vanishes would report success on exactly the wrong derivative. A well-chosen verification problem needs noncommuting matrices. If both the initial and target matrices are diagonal in the same basis, the extra terms disappear from this test.

With the correct correction, substitution into the expanded formula gives $F(X + H) = H^2$ in exact arithmetic. This provides a surprisingly inexpensive check: solve one linear system, form the new nonlinear residual, and compare it with the square of the correction. For the hand example the first diagonal entry of H^2 equals

$$1/16 + 1/9 = 25/144.$$

The initial value of F_{11} was -1. There is no reason every residual component should decrease at every remote initial condition, but this one-step identity holds regardless of the starting error.

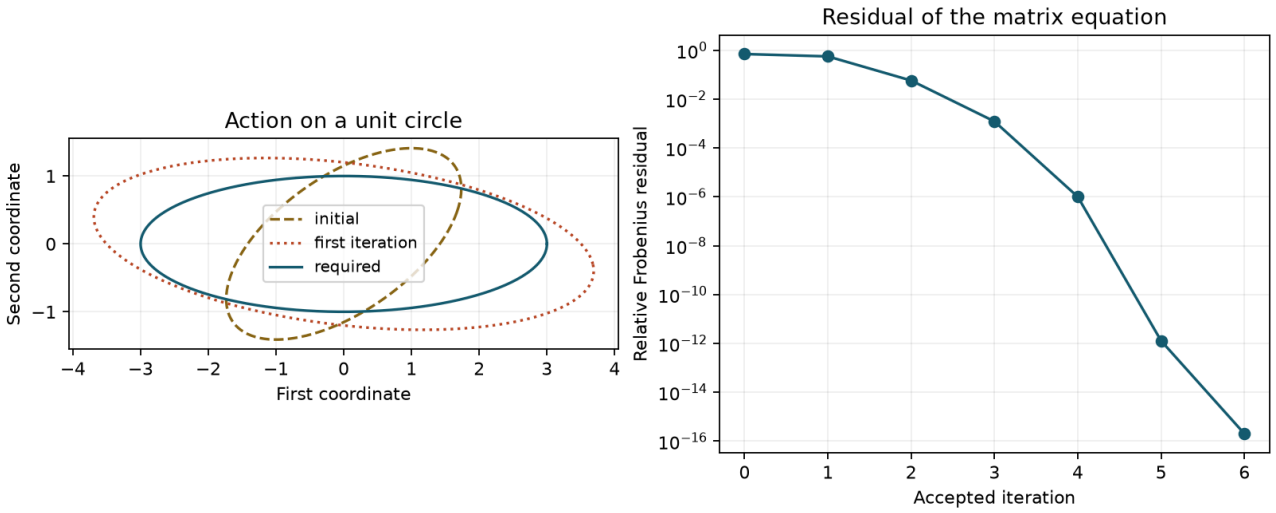
Which errors does quadratic convergence describe?

Near the desired root, the inverse Sylvester map is bounded if the smallest eigenvalue of the iterate stays away from zero. In Frobenius norm the correction obeys $\|H\|_F \le \|R\|_F/(2d_{\min})$. Together with submultiplicativity and $F(X+H) = H^2$, this yields a bound on the next residual proportional to the square of the preceding one. The proportionality constant depends on the current scale. This explains the near-doubling of correct digits late in the calculation and also why such a rule cannot be assumed for an arbitrary ill-conditioned target from the first step.

For this particular problem even preservation of the positive branch can be understood. Rearrange the equation for $Y = X + H$ to give

$$XY + YX = A + X^2.$$

The right side is positive. In an eigenbasis of positive X, the inverse operator admits the integral representation $Y = \int_0^\infty e^{-tX}(A + X^2)e^{-tX}\,dt$. Differentiate the integrand and integrate over the half line to verify the equation; the exponential vanishes at infinity. Each integrand is positive definite for finite t. Thus an exact step from positive X yields positive Y. This is specific to the equation and branch presently being solved.



A tilted initial stretch must change both eigenvalues and directions. The final residual is a matrix-equation diagnostic; the ellipse makes the action of each iterate geometrically observable.

Use a residual, a branch check and an independent solver

Run the notebook with target eigenvalues nine and one and a starting tilt of 35 degrees. The desired standard-deviation stretches are three and one. The plotted circle has radius one in input coordinates; the output contour has those semiaxes only for the target factor. A unit-radius contour in two dimensions is not a statement that 68 per cent of samples are enclosed. That probability requires a specified distribution and dimension. Read the ellipse here as a drawing of the linear transformation.

Set the initial angle to zero and repeat the comparison between the full and scalar-style derivatives. They now agree because X and A share eigenvectors. Next switch back to the tilted start and increase the

target anisotropy. Observe the size of the first correction and the number of accepted steps. At the final iterate compare its square with the target and check its smallest eigenvalue. A negative definite square root could have residual zero as well. The sign constraint is therefore informative even for a solver that seems to have converged perfectly.

The optional OneSimulation calculation in the notebook takes an independent route through the derivative. Stacking the columns of H converts the Sylvester map into $I \otimes X + X^T \otimes I$. A dense linear solve of the resulting four-by-four system should reproduce the rotated entrywise correction. Stacking rows while building the column-stacking operator would generally be a mistake. Large matrices would make the explicit Kronecker matrix expensive; the small example is valuable because it lets us compare two independently constructed maps directly.

Predict the effect of a small covariance error

The same derivative describes uncertainty in the result. If a measured target changes by a small symmetric matrix E and the root changes from S to $S + \Delta$, expand the defining equation to obtain

$$S\Delta + \Delta S = E + O(\|\Delta\|^2).$$

Errors in the target's small-eigenvalue direction have a comparatively large effect on the fitted stretch there. For target eigenvalues nine and one, a diagonal covariance error of 0.06 produces first-order stretch errors of 0.01 or 0.03, depending on its direction. The relative errors differ by an additional factor of three because the original stretches are three and one.

An off-diagonal perturbation has a different interpretation. In the target eigenbasis its leading root entry is divided by the sum of the two stretches, here four. It changes the orientation of the ellipse at first order while leaving the axis lengths unchanged at that order for distinct eigenvalues. The distinction explains how visually similar residual norms can imply rather different changes to a calibrated instrument. It also shows what to record when comparing two repeat measurements: the directions of their errors should accompany the overall covariance norm.

Open editable notebook Download .ipynb

Tutorial on branches, scaling and derivatives

1. Take $A = \operatorname{diag}(4, 9)$. Give two distinct symmetric roots and identify which one the stated calibration selects.

▼ **Work through the solution**

Both $\operatorname{diag}(2, 3)$ and $\operatorname{diag}(-2, 3)$ square to A. Only the first is positive definite. The residual cannot distinguish them; a test of the eigenvalues or a positive-branch construction is needed.

---

2. At $X = \operatorname{diag}(3, 0.2)$, compare corrections produced by equal residual entries of magnitude 0.01 on each diagonal.

▼ **Work through the solution**

The first diagonal correction has magnitude $0.01/6 \simeq 0.00167$. The second has magnitude $0.01/0.4 = 0.025$. Thus identical absolute covariance errors require corrections differing by a factor of fifteen in the two stretch directions.

---

3. In the tilted hand example, what is the predicted first off-diagonal entry of the updated nonlinear residual? How would one use it to distinguish a failed linear solve from an ordinary Newton truncation?

▼ **Work through the solution**

The correct H_{12}^2 equals

$$1/12 + 1/6 = 1/4.$$

With an inexact solve, $F(X + H) - H^2 = F(X) + XH + HX$ equals the residual of the linearised equation. Its norm tests solver accuracy separately from the nonlinear term intentionally left after a Newton step.

Further reading · Nicholas Higham, Newton's method for the matrix square root

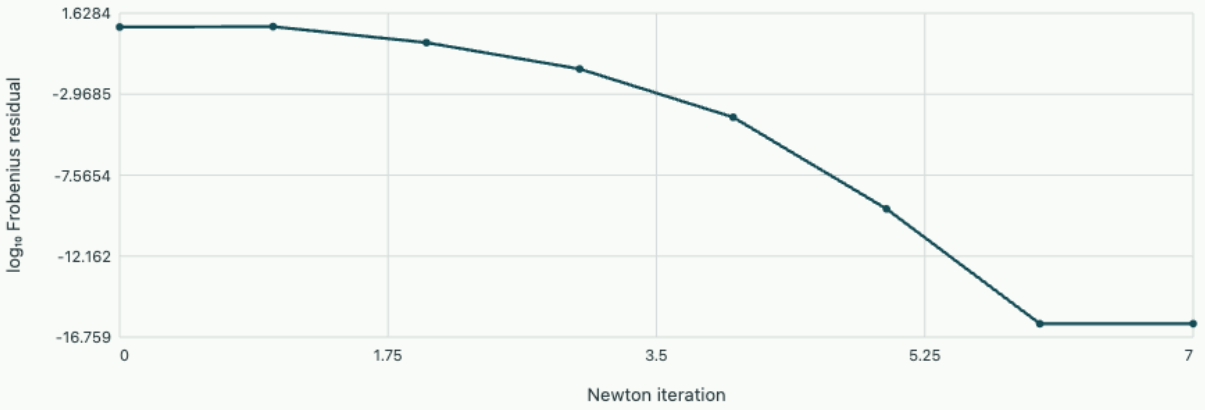
Change parameters · Newton's method for an unknown covariance square root

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

Open this calculation with adjustable inputs

Try the calculation

Target eigenvalue ratio	9
Starting orientation (degrees)	45



Symmetric Sylvester Newton correction

Target positive square root is diag(1, 3). Each step solves XH + HX = A − X² in the three symmetric matrix coordinates.

Problem sheet 7 · Project · Locate contact and calibrate a misleading residual

Two surfaces touch tangentially. A height sensor measures gap rather than signed distance. Decide what rootfinding, residual checks and a second sensor can each certify.

Recover the collision of two mathematical roots

Near the contact, model the gap as $g(x, p) = (x - 1)^2 - p$, with all variables scaled by their reference units. Sketch its zeros for p above and below zero. For $p = 0.01$ begin with $[0.95, 1.2]$ and test whether a sign-changing bracket isolates a root.

▼ **Work through the solution**

Real roots are $1 \pm \sqrt{p}$ for positive p, a double root at one for $p = 0$, and none for negative p. At 0.95 the residual is -0.0075, while at 1.2 it is 0.03. The bracket contains only the upper root at 1.1. A wider interval with both endpoints outside the two roots can have equal endpoint signs even when it contains both.

Explore the related calculation with adjustable inputs. Use the equilibrium parabola to inspect sensitivity and the tangent-coordinate continuation step. Translate the state coordinate if the fold is centred at a different value here.

Calibrate a stopping tolerance

With $p = 0$ begin at 1.2. Take three ordinary Newton steps and one multiplicity-corrected step. A calibrated sensor reading places the nonnegative gap in $[0, 10^{-6}]$. What location interval follows from that bound? Compare the slopes in the multiple-root experiment.

▼ **Work through the solution**

The successive locations are 1.1, 1.05 and 1.025. Multiplying the first Newton correction by two lands exactly on 1 for this pure square. A gap of 10^{-6} is compatible with a displacement of magnitude 10^{-3}. Supplying more digits in the root does not improve that data-based bound.

Audit an algorithm that always reports convergence

At $p = -0.001$ a least-squares routine stops at $x = 1$ and reports zero gradient. Find the residual and explain the report. Use the trust-region model to distinguish a decreasing merit function from a good final fit.

▼ **Work through the solution**

The residual is 0.001, and the merit gradient gg' vanishes because $g' = 0$. This is the best fit of this model to the infeasible contact target. The positive minimum is 1,000 times the 10^{-6} gap tolerance used in the preceding test. Report a nonzero closest separation and the uncertainty rather than a located contact.

Choose an additional measurement

You may measure one signed local slope g_x with uncertainty 0.002, or repeat the same absolute gap measurement. Which option resolves the side of a nearby contact? Express the position uncertainty before combining measurements.

▼ **Work through the solution**

$g_x = 2(x - 1)$ changes sign through contact. Its stated uncertainty yields an uncertainty of 0.001 in the inferred x. An interval for the signed slope that excludes zero resolves which side of contact the instrument occupies. If the measured magnitude is at most 0.002, both signs still fit within the uncertainty. Repeated absolute gaps can improve precision in distance magnitude while retaining the symmetry between left and right solutions.

Explore the related calculation with adjustable inputs. Use the equilibrium parabola to inspect sensitivity and the tangent-coordinate continuation step. Translate the state coordinate if the fold is centred at a different value here.

Problem sheet 8 · Find a periodic orbit when the starting point has no preferred phase

A nonlinear oscillator returns to its starting point after one period. Writing that fact as a root equation leaves a neutral direction and several unrelated solutions. A solvable amplitude oscillator lets you construct, stabilise and independently check a shooting method.

Count the unknowns in a periodic return

Consider the autonomous planar oscillator $\dot{x} = (\mu - r^2)x - \omega y,\ \dot{y} = (\mu - r^2)y + \omega x,\ r^2 = x^2 + y^2$, with $\omega > 0$. The example uses scaled amplitudes and seconds. Let $\Phi_T(u)$ denote the state after integrating for time T from $u = (x, y)$. A shooting residual is

$$F(u, T) = \Phi_T(u) - u.$$

Determine what could go wrong if a three-component unknown (x, y, T) is handed to a root solver with only these two residuals.

Recover the orbit and identify spurious roots

Use polar variables away from the origin to predict which residual zeros correspond to a single nontrivial revolution.

▼ **Work through the solution**

Taking $x\dot{x} + y\dot{y}$ gives

$$r\dot{r} = (\mu - r^2)r^2.$$

The angular numerator $x\dot{y} - y\dot{x} = \omega r^2$ gives

$$\dot{\theta} = \omega.$$

For $\mu > 0$, $r_* = \sqrt{\mu}$ is stationary and the period is

$$T_* = 2\pi/\omega.$$

Every phase on the circle gives the same periodic orbit, so there is one redundant physical direction. Furthermore $T = 0$ gives zero residual for every point; the origin gives zero residual for every T; and each integer multiple of T_* solves the return condition on the nonzero circle. A physically informative solve needs an initial estimate in the basin of the desired one-turn branch, positive period and a section choosing a phase. A tiny residual alone discriminates none of these cases.

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.

Differentiate through a complete return

Border the Newton equations with a transverse section

Take $g(u) = y = 0$ with $x > 0$. Derive a Newton matrix including period as an unknown. Why would fixing $x = \sqrt{\mu}$ be a bad phase condition at this same starting point?

▼ **Work through the solution**

Let $M = D_u\Phi_T(u)$ be the fundamental matrix propagated along the trajectory. Its columns solve $\dot{M} = Df(\Phi_t(u))M,\ M(0) = I$. The derivative with respect to final time is the vector field evaluated at the end of the integration. Thus the linearised return and section equations are

$$\begin{pmatrix} M - I & f(\Phi_T(u)) \\ (0,1) & 0 \end{pmatrix} \begin{pmatrix} \delta u \\ \delta T \end{pmatrix} = - \begin{pmatrix} \Phi_T(u) - u \\ y \end{pmatrix}.$$

At a nontrivial periodic point the vector $f(u)$ is an eigenvector of M with eigenvalue one, because starting a little later along the same orbit merely shifts the final point along it. This explains the singularity of $M - I$ before bordering. At $u = (\sqrt{\mu}, 0)$ the tangent points along positive y, so

$$\nabla g \cdot f = \omega\sqrt{\mu} \neq 0.$$

A section fixing $x = \sqrt{\mu}$ is tangent to the circle there; its gradient annihilates f. The new row then fails to remove phase degeneracy to first order even though it seems geometrically to locate the rightmost point.

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.

Check stability and conditioning with an exact return map

Compute the transverse multiplier without an eigensolver

Put $q = 1/r^2$ and solve its scalar initial-value problem for $r_0 > 0, \mu > 0$. Use the result to check one period of numerical propagation for $\mu = 0.1,\ \omega = 2\pi$.

▼ **Work through the solution**

Differentiating gives $\dot{q} = 2 - 2\mu q$, hence $q(T) = 1/\mu + (q_0 - 1/\mu)e^{-2\mu T}$. The return radius on the positive axis after T_* is therefore

$$P(r_0) = \left[\frac{1 - e^{-2\mu T_*}}{\mu} + \frac{e^{-2\mu T_*}}{r_0^2} \right]^{-1/2}.$$

At the fixed radius $P'(r_*) = e^{-2\mu T_*}$. This also follows by linearising $\dot{r} = \mu r - r^3$, whose derivative is -2μ at r_*. In the numerical case $T_* = 1.00$ s, $r_* = 0.31623$ and the transverse multiplier is 0.81873. The radial shooting residual has slope -0.18127. A radial residual error of 10^{-5} corresponds locally to a radius error of about 5.52×10^{-5}. Near onset, the multiplier approaches unity and the phase condition alone cannot cure the true loss of transverse conditioning. For $\mu \leq 0$ the positive finite-amplitude cycle ceases to exist, even though the origin remains an exact residual zero.

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.

Separate time-integration error from the nonlinear residual

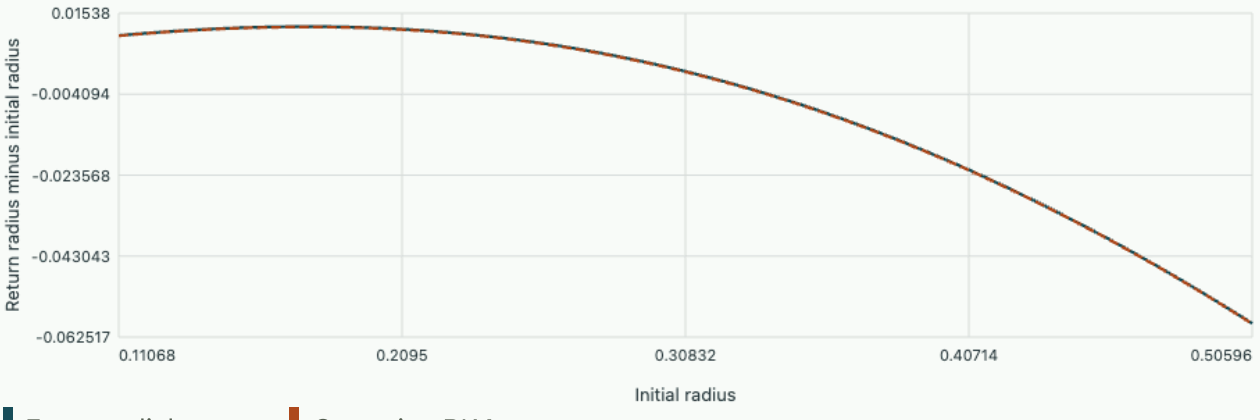
The experiment plots the signed one-revolution radial residual for many initial radii. One curve propagates the vector oscillator using classical fourth-order Runge–Kutta (RK4) in Cartesian coordinates. The other uses the analytical radial return. A sign change gives a candidate amplitude. The phase drift at that approximate root, reported separately, catches angular error that comparing final and initial radii alone would miss. Predict how both diagnostics should change when the time steps per turn double.

Open this experiment online

Try the calculation

Amplitude parameter μ **0.1**

RK4 steps per revolution **16**



Exact amplitude 0.31623, RK4 radial fixed point 0.31566, phase defect -6.75e-2°. Transverse multiplier 0.81873.

Choose a convergence test that excludes an equilibrium and a multiple turn

Describe a validation at $\mu = 0.1$ and design a check for a solver that accidentally reports $T \simeq 2$ seconds instead of one.

▼ **Work through the solution**

Begin on a positive transverse section with radius near 0.316 and period near one second. Impose a positive radius floor only when selecting the intended branch, and check the integrated speed or path length to exclude a stationary solution. Resolve the radial residual with several integration meshes, testing that both return radius and phase error converge. A Newton iteration can drive the residual associated with a coarse integrator nearly to zero while its orbit is still inaccurate. Make integration error appreciably smaller than the desired orbit error after amplification by the near-singular transverse slope.

To check multiplicity, unwrap the trajectory angle or count positive-direction section crossings over the proposed orbit. The desired orbit accumulates 2π and returns once; a two-turn solve accumulates 4π and its nontrivial multiplier is $e^{-4\mu T_*} = 0.67032$ in this example. Stability can still appear physically sensible for a two-turn return. Counting turns or measuring a shorter repeated return tests something the residual and stability sign cannot. For systems without an analytic polar reduction, an event-defined Poincaré map and independently propagated variational equations provide the corresponding diagnostics.

Explore the related calculation with adjustable inputs. Change the parameters in the experiment on this same project page and compare its output with this part of the derivation.

Further reading

- BifurcationKit · shooting equations, phase sections and Floquet multipliers

Notebook practice · Newton's method for an unknown covariance square root

Nonlinear systems and iterative methods

Begin with a residual and ask which solution is accessible. Study how curvature, scaling and conditioning affect a computed root.

1. ## Choosing a derivative step in a nonlinear solve

   A smaller finite-difference step eventually makes Newton iterations worse. Derive truncation and evaluation errors, compare a complex perturbation and a directional derivative, and test which numerical derivative is meaningful for the actual residual implementation.

2. ## Accelerating fixed points with differences of residuals

   A slowly convergent physical iteration contains information in the direction of successive errors. Use a scalar extrapolation and a small residual-minimisation problem to exploit that information, then test the extrapolated state against the original equation.

3. ## Finding another equilibrium after one root is known

   Newton iteration repeatedly finds the same structural equilibrium. Derive how deflation changes its basins while preserving other solutions, use a rank-one Jacobian update, and distinguish a list of discovered equilibria from a completeness certificate.

4. ## Count enclosed zeros with a contour before locating them

   A complex residual may have several nearby zeros whose Newton basins are hard to distinguish. Use the argument principle to count them, contour moments to estimate their locations and an independent perturbation test to certify a root cluster.

5. ## Where does a complex Newton step take its starting point?

   A single cubic has three equally simple roots. Derive the iteration, colour starting values by destination and use a separate convergence history to distinguish a bad first step from slow local convergence.

6. ## Certifying a Newton step before the root is known

   A small correction is meaningful only together with a curvature estimate. Construct a ball that contains a solution and use a frozen linear solve to make damping decisions independent of equation units.

7. ## Which correction does an experiment identify?

   Separate parameter combinations visible in a nonlinear fit, pass a regular fold by continuation and decide when to refine a boundary-value mesh before trusting its Newton residual.

Nonlinear systems and iterative methods · 1. Choosing a derivative step in a nonlinear solve

A smaller finite-difference step eventually makes Newton iterations worse. Derive truncation and evaluation errors, compare a complex perturbation and a directional derivative, and test which numerical derivative is meaningful for the actual residual implementation.

Derive competing contributions to error

Suppose f is smooth and you need its slope at x to form a Newton correction. A forward difference $D_h = [f(x+h) - f(x)]/h$ has Taylor expansion $f'(x) + hf''(x)/2 + O(h^2)$. A centred difference $[f(x+h) - f(x-h)]/(2h)$ instead cancels the even terms, leaving $f'(x) + h^2 f'''(x)/6 + O(h^4)$. Both limits are exact in real arithmetic as $h \to 0$. The program, however, returns approximate residual values, and a subtraction of nearby values can lose significant digits. If each absolute evaluation error is at most δ, the derivative error from those perturbations is at most $2\delta/h$ for forward difference and δ/h for the centred one.

Thus an illustrative forward error bound has shape $ah + b/h$; minimisation gives

$$h_* = \sqrt{b/a}.$$

The centred bound $ah^2 + b/h$ has minimiser $h_* = (b/(2a))^{1/3}$. Floating-point roundoff acting on $O(1)$ residual values often gives the familiar square-root and cube-root machine-precision scales, but the physical units of x, the scale of the derivative and noise from the residual solver must be included. A nonlinear function evaluated by an iterative PDE solve can have a much larger δ than roundoff. Repeating the inner solve with tighter tolerance can shift the optimal step considerably.

For $f(x) = e^x$ at zero, the forward leading error is $h/2$, centred error $h^2/6$. If $\delta = 10^{-10}$ as an absolute per-value uncertainty, a forward worst-case bound gives optimal $h = 2 \times 10^{-5}$, while the centred model gives

$$h = (3 \times 10^{-10})^{1/3} = 6.69 \times 10^{-4}.$$

The centred step can be much larger yet retain lower truncation error. This is a scale estimate using worst-case errors, not a guarantee about the phase or stochastic correlation of actual errors. If queries share common random numbers or a warm-started solver, errors can partly cancel or acquire a systematic derivative bias.

Obtain the slope from an imaginary perturbation

For a real-valued function with an analytic complex extension near the real argument, expand $f(x + ih) = f(x) + ihf'(x) - h^2 f''(x)/2 - ih^3 f'''(x)/6 + \cdots$. Taking its imaginary part and dividing by h gives second-order accuracy with no explicit subtraction of two nearly equal real numbers. For e^x at zero, the result is $\sin h/h$. At very small representable h the imaginary signal is itself proportional to h and may still be carried accurately by analytic elementary operations. This advantage

disappears if the computation discards imaginary parts, uses nonanalytic absolute values, branches on the perturbation, or applies a black-box routine using complex arithmetic for unrelated purposes. Automatic differentiation supplies another route through differentiable implemented operations, but it too needs the program's mathematical definition to match the derivative being requested.

Audit a Jacobian action before assembling a matrix

For $F : \mathbb{R}^n \to \mathbb{R}^m$, pick direction v and examine $R(h) = \|F(x + hv) - F(x) - hJv\|$. If a proposed Jacobian action is correct and F is twice differentiable, $R(h)$ initially scales like h^2. A wrong first derivative generally produces a leading $O(h)$ term. The test checks orientation, units, signs and chain rules without constructing a dense finite-difference matrix. Test several directions and scales; a wrong matrix can annihilate the one chosen direction. At very small h a roundoff or inner-solver floor eventually hides both ideal slopes. When working with constraints use perturbations remaining inside the physical domain or give derivatives in a local feasible parametrisation.

Find the useful step before a Newton update

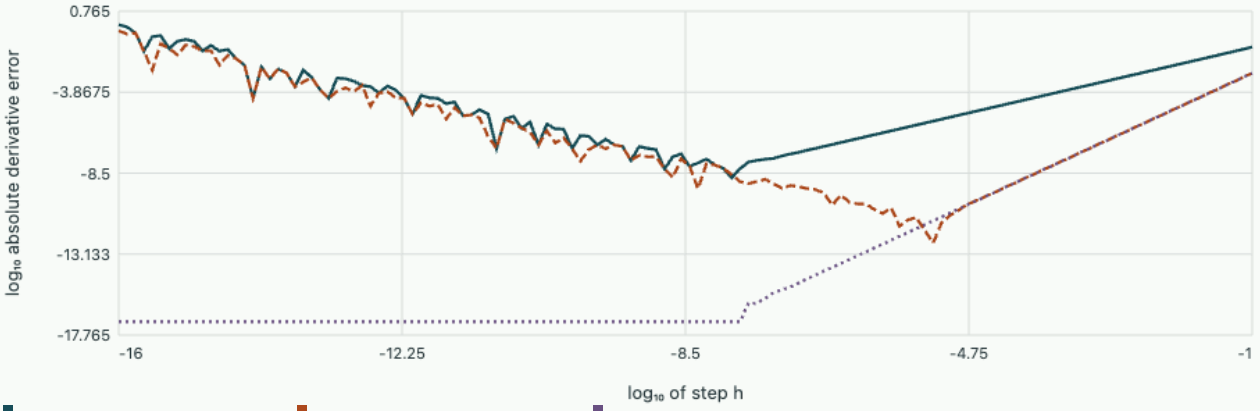
The experiment evaluates the forward, centred and complex-step slopes of a smooth elementary residual in floating-point arithmetic. It plots absolute errors on logarithmic scales and compares them with the exact derivative. Move the evaluation point and note that occasional accidental zeros are rounded at the plotting floor. Their exact location is not a robust guarantee for a different residual. Compare the range of steps with a reproducible error trend before interpreting one apparently perfect result.

Open this experiment online

Try the calculation

Evaluation point x **0**

Largest plotted $\log_{10}$ step **-1**



Forward difference Centred difference Complex-step imaginary part

Exact derivative exp(x) = 1.00000. Errors below 10^{-17} are drawn at the plotting floor. The complex exponential is evaluated analytically in real and imaginary parts.

Change the evaluation before changing the step

For the exponential the forward numerator has an exact factorisation $e^{x+h} - e^x = e^x(e^h - 1)$. A numerical library's `expm1(h)` evaluates the last factor accurately even for very small h, using a local series or a reduced-argument algorithm. Form the quotient as $e^x \operatorname{expm1}(h)/h$. Its mathematical truncation error is still $e^x h/2 + O(h^2)$. Its rounding behaviour differs because the two terms near e^x are never computed separately. At $x = 1$ and $h = 10^{-18}$, adding h to a binary64 value of one rounds back to one. The direct forward quotient then returns zero. The factored evaluation gives the derivative e to working precision.

This experiment diagnoses the source of an error curve. If the algebraic form is inaccessible, a finite-difference checker sees black-box function values and cannot apply the factorisation. If the entire residual is available, stable elementary functions can change the derivative accuracy of a Newton step without changing its formal order.

Tutorial · design tests for smoothness and implementation errors

Diagnose a forward-difference Newton iteration

For $f(x) = x^2 - 2$ at $x = 1.4$, obtain the exact dependence of a forward derivative estimate on h. Compare one Newton step using $h = 0.1$ with the exact Jacobian.

▼ **Work through the solution**

The difference quotient is

$$2x + h = 2.8 + h.$$

The residual is -0.04. The exact step gives

$$x_{\text{new}} = 1.4 + 0.04/2.8 = 1.414286;$$

using 0.1 gives 1.413793. Centred differences evaluate this polynomial's derivative exactly in real arithmetic for every nonzero h, but roundoff can spoil very small steps. The discrepancy at large h is a calculable systematic truncation, so merely tightening the outer Newton stopping tolerance cannot remove it.

Catch a direction missed by one test

A proposed Jacobian differs from the correct one by $E = \left(\begin{smallmatrix} 1 & -1 \\ 0 & 0 \end{smallmatrix}\right)$. Compare tests along $(1, 1)$ and $(1, 0)$ and show which observed slope would be misleading.

▼ **Work through the solution**

The first direction lies in the nullspace of E, so its Taylor remainder can show the expected h^2 slope despite an incorrect Jacobian. The second has $Ev = (1, 0)$, producing an $O(h)$ defect. Random directions augment a set of unit- or physics-informed probes, particularly when an error affects a single column or constraint block. One successful slope does not establish correctness of an entire matrix.

Determine whether complex step applies to a norm

A residual returns $f(x) = |x|$ using a complex modulus routine. At $x = 1$, what does a complex-step call produce? Is this evidence that the real function has zero derivative there?

▼ **Work through the solution**

The routine computes $|1 + ih| = \sqrt{1 + h^2}$, purely real, so the reported complex-step slope is zero. The real derivative at $x = 1$ equals one. The implementation used a nonanalytic map of a complex variable. At a known positive branch the analytic local extension $f(z) = z$ gives the correct derivative, while the real function at the kink $x = 0$ has no classical derivative. Check both the real regularity and what complex arithmetic actually executes before applying this technique.

Further reading

- Martins, Sturdza and Alonso · complex-step derivative approximation and implementation
- ETH Zürich · numerical methods for CSE, chapter 8 on nonlinear equations
- Course investigation · inexact Newton updates
- Solve an implicit cooling step with a changing residual

Nonlinear systems and iterative methods · 2. Accelerating fixed points with differences of residuals

A slowly convergent physical iteration contains information in the direction of successive errors. Use a scalar extrapolation and a small residual-minimisation problem to exploit that information, then test the extrapolated state against the original equation.

Eliminate an unknown leading error without knowing the root

Suppose a sequence obeys $x_j = x_* + cq^j$ for an unknown limit and mode multiplier $q \neq 1$. Subtract successive terms to find

$$\Delta x_{j+1} / \Delta x_j = q.$$

Eliminating q, c yields the Aitken extrapolation $\hat{x}_j = x_j - (x_{j+1} - x_j)^2 / (x_{j+2} - 2x_{j+1} + x_j)$. For a single exact geometric error it returns the limit after observing three values, even if the multiplier is negative and the iterates oscillate. If the denominator is near zero relative to numerical uncertainty, the extrapolation can amplify noise. A nearly constant first difference may indicate almost linear drift rather than a trustworthy asymptotic geometric error.

For $g(x) = 0.95x + 0.10$, ordinary iteration from zero gives $0, 0.10, 0.195$; the fixed point is 2. The second difference is -0.005 and the formula gives exactly 2. If the last reported value is rounded to 0.196, the denominator changes to -0.004, giving an extrapolated 2.50. Substituting this candidate into the original residual $g(x) - x$ gives -0.025: it is a worse fixed point than 2 but still has a modest residual because the fixed point itself is weakly conditioned. Always evaluate the actual residual and estimate its relation to state error if a meaningful physical tolerance is required.

Choose affine weights that minimise residual history

For a vector fixed-point map g, let $f_i = g(x_i) - x_i$. Seek a combination of recent residuals with small norm while preserving a fixed point under averaging. Anderson mixing chooses coefficients α_i minimising $\| \sum_i \alpha_i f_i \|$ subject to $\sum_i \alpha_i = 1$, then evaluates a candidate $x_{\text{new}} = \sum_i \alpha_i g(x_i)$ in a basic undamped formulation. The affine constraint allows negative coefficients, which is precisely what makes acceleration beyond ordinary averaging possible. For one-dimensional affine maps the combination can eliminate a single mode exactly.

For two residual vectors write their affine combination as $f_1 + \beta(f_0 - f_1)$. Its least-squares optimum is $\beta = -f_1^T (f_0 - f_1) / \|f_0 - f_1\|^2$ if the denominator is nonzero. With several vectors a small QR or SVD solve in terms of differences is preferable to building normal equations when differences are nearly collinear. Choose component units or a weighted norm consistent with the physical equation. Otherwise a residual expressed in large numerical units can dominate the fit and accelerate one variable while neglecting another important constraint.

For a strictly affine map $g(x) = Gx + c$,

$$f(\sum \alpha_i x_i) = \sum \alpha_i f_i.$$

The predicted residual at the affine combination of old states is exactly the linear combination. The candidate formed by applying g to it has residual $G\sum\alpha_i f_i$. For nonlinear g neither equality need hold globally. An Anderson residual minimisation is a local extrapolation hypothesis, not a proof that the next nonlinear state decreases residual or stays feasible. Evaluating $f(x_{\mathrm{new}})$, accepting a damped mixture when appropriate, and dropping near-dependent history are practical ways to test that hypothesis.

Background · What happens when two residual histories almost coincide?

Take two residuals from nearly consecutive states and minimise the norm of $r_1+\alpha(r_2-r_1)$. Write $d=r_2-r_1$. Differentiating the squared norm with respect to α gives

$$\alpha=-\frac{d^Tr_1}{d^Td}.$$

The denominator vanishes for duplicate residuals. A tiny denominator can also demand enormous positive and negative weights that cancel in the modelled residual. For scalar residuals 1 and 0.999 those weights are about -999 and 1000. Extrapolating the nonlinear states this far relies on local behaviour far outside the sampled interval. One practical response is to discard a nearly redundant column, then evaluate the true residual of the proposed state. A small residual of the affine model alone provides no acceptance test.

Separate fast and slow modes in a two-component relaxation

Use $g(x,y)=(0.9x+0.1,0.2y+0.8)$, whose fixed point is $(1,1)$. Start at $(0,0)$, then $x_1=(0.1,0.8)$. Residuals are $f_0=(0.1,0.8)$ and $f_1=(0.09,0.16)$; their difference is $(0.01,0.64)$. The two-history coefficient above is about -0.252. The predicted residual of the affine state is approximately $(0.0875,-0.0010)$, largely eliminating the fast second component while leaving the slower first. Three appropriately independent histories can eliminate two geometrically distinct error directions for this affine example if their residual vectors affinely span zero in exact arithmetic. Near-equal decay factors can make that calculation badly conditioned. A finite memory window trades acceleration against conditioning, storage and the cost of residual evaluation.

Follow a two-mode residual through a real mixing step

Set the slow multiplier to 0.90 and the second to 0.20. The first two ordinary states coincide in both runs because the accelerated iteration has no history at the start. Thereafter the calculator solves the two-vector minimisation at each iteration, forms the proposed mixed state and evaluates its actual residual. The second component initially dominates the residual norm even though it is the more rapidly damped mode. Notice how the mixed and plain curves separate as the slow component becomes limiting. The plotted logarithmic residual measures convergence to the stated map; the physical state error in mode j is its residual divided by the magnitude of one minus that mode multiplier.

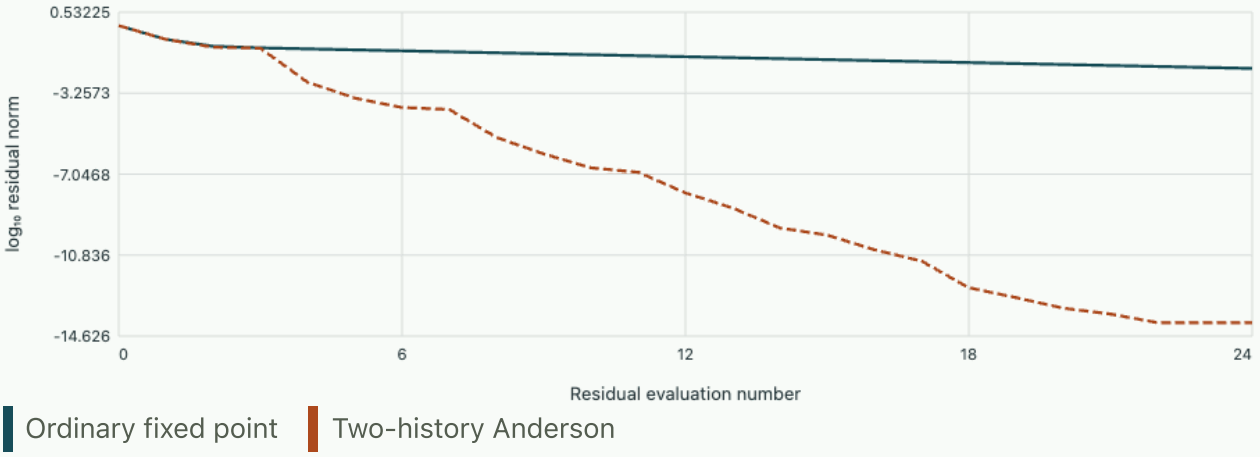
Move the two multipliers close together so the map becomes nearly a scalar relaxation, then choose a negative second multiplier so its errors alternate. Explain why affine mixing can nearly remove a single geometric mode, while nearly dependent histories make weights sensitive to evaluation noise. The implementation falls back to the ordinary map if the difference norm is already at the roundoff threshold. Very small residuals are marked at a finite plotting floor. Compare curves before that floor rather than fitting a convergence rate to the flat portion.

Open this experiment online

Try the calculation

Slow multiplier **0.9**

Second multiplier **0.2**



After six updates: ordinary residual 0.0531, mixed residual 0.000119. Plot floor 10^{-14}.

Tutorial · identify good and bad extrapolation data

Check an alternating scalar series

Three iterates are 1.50, 0.75 and 1.125. Infer the geometric multiplier and limit. How should the next value be checked?

▼ **Work through the solution**

The differences are -0.75 and $+0.375$, so $q = -0.5$. The Aitken prediction is

$$1.50 - 0.75^2/1.125 = 1.00.$$

If the geometric assumption persists, the next raw iterate would be $1 - 0.0625 = 0.9375$. Comparing both the new observed iterate and the actual residual at 1 tests the single-mode hypothesis. Three arbitrary numbers alone can always fit such a geometric extrapolation when the denominator is nonzero.

Show the connection to a secant root step

Use two scalar residuals $f_i = g(x_i) - x_i$. Derive the affine combination of x_0, x_1 whose predicted residual is zero and compare with a secant step for $f = 0$.

▼ **Work through the solution**

Solve $\alpha f_0 + (1 - \alpha) f_1 = 0$ to get

$$\alpha = -f_1/(f_0 - f_1).$$

The mixed state is $x_1 - f_1(x_1 - x_0)/(f_1 - f_0)$, the ordinary secant prediction. The combination of $g(x_i)$ is the same state because the weighted residual vanishes. For nearly equal residuals the denominator can be tiny, generating a large extrapolation. A bounded step or fallback to the base fixed-point map may be necessary if the candidate leaves the valid state domain.

Select a restart from nearly identical residual columns

Two stored vectors have residuals $(1, 0)$ and $(1, 10^{-9})$, measured with absolute component errors 10^{-7}. Can their difference support a reliable two-history fit?

▼ **Work through the solution**

The direction being used for the second independent sample is smaller than the evaluation uncertainty by two orders of magnitude. Its orientation and length can be dominated by numerical error, so a large coefficient in a least-squares solve carries little physical information. Drop the redundant direction or use a rank-revealing solve scaled to the expected residual error. Evaluate any retained candidate on the original model. A tiny predicted mixed residual from cancelling inaccurate large coefficients has no independent evidential value.

Further reading

- Walker and Ni · Anderson acceleration for fixed-point iterations
- Course investigation · contractive maps and error bounds
- Solve an implicit cooling step with a changing residual

Nonlinear systems and iterative methods · 3. Finding another equilibrium after one root is known

Newton iteration repeatedly finds the same structural equilibrium. Derive how deflation changes its basins while preserving other solutions, use a rank-one Jacobian update, and distinguish a list of discovered equilibria from a completeness certificate.

Prevent a known simple zero from attracting another iteration

Suppose a differentiable $F : \mathbb{R}^n \to \mathbb{R}^n$ has a known zero r, with nonsingular Jacobian $J(r)$. Near r, $F(r+h) = J(r)h + O(\|h\|^2)$, so its norm is bounded below by a positive constant times $\|h\|$ in a sufficiently small neighbourhood. Multiplying by a nonvanishing scalar multiplier away from r preserves all other zeros. Taking $M(x) = \|x - r\|^{-p} + \alpha$, $\alpha > 0$, yields deflated residual $G(x) = M(x)F(x)$ on $x \neq r$. Its norm behaves at least as a constant times $\|x - r\|^{1-p}$ near the simple root. For $p = 1$ it need not have a directional limit but cannot approach zero there; for $p > 1$ it diverges. The scalar shift α prevents the multiplier from vanishing at large distances.

If the root has multiplicity m in a scalar problem, $|F|$ can vanish like distance to the power m. An exponent $p < m$ can leave the deflated norm tending to zero, invalidating the reasoning for simple roots. In systems with singular Jacobians different directions can have different leading powers or a whole manifold of solutions can pass through r. One should identify symmetries and solve a properly constrained problem before treating every coordinate representation of a continuum of roots as a new equilibrium. Deflation changes a nonlinear solver's route; it does not make a singular physical problem isolated by decree.

Reuse a Newton correction with a rank-one change

Differentiate the scalar product to obtain

$$J_G = MJ_F + F(\nabla M)^T.$$

When J_F is invertible, let $d = -J_F^{-1}F$ be the ordinary Newton correction. Look for a deflated step $s = \beta d$. Applying J_G gives $\beta[-MF + F(\nabla M)^T d]$. Equating this to $-MF$ yields

$$s = \frac{d}{1 - (\nabla \log M)^T d}.$$

This is a convenient exact relation for a scalar deflation multiplier away from a vanishing denominator. The Jacobian of the modified problem can become singular at points where the original Newton step is perfectly well-defined. A denominator near zero can create an enormous proposed state change. A line search or trust region should verify the candidate against the chosen objective and the physically admissible variables. Once a different candidate root is located, evaluate the original F, its Jacobian conditioning and any invariant constraints independently.

See the change of basin in an analytical scalar model

Take $F(x) = x(x-1)(x+2)$ with roots $0, 1, -2$, and let $r = 0$ have already been found. On the positive half-line choose $M = 1/x$ for a transparent algebraic comparison, giving $G = (x-1)(x+2)$. This unshifted local example works on this domain; a global implementation often uses the positive shifted norm expression instead. From $x = 0.2$,

$$F = -0.352,\ F' = 3x^2 + 2x - 2 = -1.48.$$

Original Newton gives approximately -0.0378, entering the basin of the known zero. Deflated Newton has $G = -1.76,\ G' = 1.4$ and gives approximately 1.4571, entering a region directed toward the other positive root. Neither update used knowledge of the other roots to choose that direction.

For a polynomial with known exact factor, explicit algebraic division may be available. A numerical residual from a mechanics or fluid solver need not be a polynomial and generally offers no such division. Numerical deflation uses the known state and a norm to remove its attraction while continuing to evaluate the same physical governing equations. Use physical scaling in that norm. The distance between two states measured in metres and kelvins is arbitrary until component scales are chosen; an apparently huge separation in the deflation metric can represent tiny physical changes. If roots are known only approximately, multiplying a near-singular factor around a slightly wrong centre can leave the true nearby root or create severe numerical conditioning. Refine and verify each original root before including it in a list of deflation centres.

Change the Newton basin by dividing out a known root

Here the original residual is the cubic in the worked example. Its known zero at the origin is algebraically divided out, leaving a quadratic on the full line. This exact-factor experiment isolates the change in Newton geometry without introducing a singular numerical evaluation of an inverse distance. For both residuals, the calculator starts at the same chosen positive point and applies the same damping factor. At the default 0.20 the first ordinary step goes slightly negative and approaches zero; the deflated step passes one and converges back to it. The readout verifies the second candidate in the original cubic.

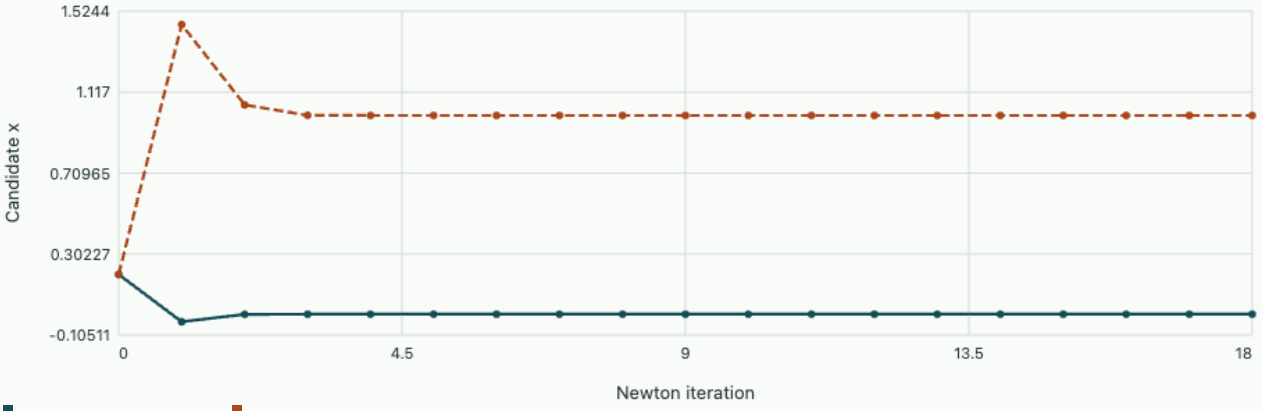
Try a smaller damping factor and compare the number of iterations needed before the two candidate sequences settle. Damping moderates each proposed correction but does not restore the removed root of the quadratic. A finite iteration count with heavy damping may end with a visible original residual; a current candidate is not an exact equilibrium merely because its curve has begun to flatten. This illustration relies on an exact polynomial factor, while deflation around a numerically inferred state requires an explicit centre tolerance and original-model verification.

Open this experiment online

Try the calculation

Initial x **0.2**

Newton damping **1**



Final candidates 0.000 and 1.000. Original cubic residual on the deflated branch 0.00e+0.

Tutorial · separate discovery, verification and completeness

Work through a simple shifted multiplier

For $F = x(x-1)$, $x > 0$, use $M = 1 + 1/x$ to deflate zero. Find the new residual, its value as $x \to 0^+$, and the deflated Newton step from $x = 0.5$.

▼ **Work through the solution**

Here $G = (x+1)(x-1) = x^2 - 1$, tending to -1 at the excluded zero. From $x = 0.5$, $G = -0.75,\ G' = 1$, giving $x_{\text{next}} = 1.25$. The original derivative $2x - 1$ vanishes exactly at the starting point even though the deflated derivative does not. The shortcut requiring J_F^{-1} is unavailable there; directly differentiating and solving for G remains valid. This example shows why a formula derived for nonsingular J_F should not become a universal implementation assumption.

Choose a multiplier for a double root

For $F(x) = (x-1)^2(x+1)$ deflate $r = 1$ with $M = |x-1|^{-p} + 1$. Compare $p = 1$ and $p = 2$. What limits a near-root termination test?

▼ **Work through the solution**

Near 1, $|F| \sim 2|x-1|^2$. With $p = 1$, $|G| \sim 2|x-1|$ still tends to zero; with $p = 2$, it tends to two. A test accepting just a small deflated residual can therefore return very close to the old double root when

using the inappropriate exponent. A small original residual also occurs over a relatively wide neighbourhood of a multiple root. Report a root-separation test and independent bounds or refinement where multiplicity affects the required state tolerance.

Construct a genuine completeness argument in one dimension

Several initial guesses and deflations recover the three roots of $x^3 + x^2 - 2x$. What additional reasoning certifies that the list is complete, and what can be reported if only a black-box smooth system is available?

▼ Work through the solution

The polynomial is nonzero of degree three, so it has at most three distinct complex roots. Substitution verifies three distinct real ones and proves completeness in this case. Multiple Newton solves by themselves would not establish that count. For a smooth black-box scalar problem over a compact interval one can partition it, enclose the function and its derivative, exclude root-free pieces and isolate all retained candidates when the bounds allow. In a higher-dimensional problem a list found by starts and deflation should be reported as verified discovered roots unless an independent argument covers the entire specified domain.

Further reading

- Farrell, Birkisson and Funke · deflation for distinct nonlinear solutions
- Course investigation · continuing an equilibrium branch
- Solve an implicit cooling step with a changing residual

Nonlinear systems and iterative methods · 4. Count enclosed zeros with a contour before locating them

A complex residual may have several nearby zeros whose Newton basins are hard to distinguish. Use the argument principle to count them, contour moments to estimate their locations and an independent perturbation test to certify a root cluster.

Turn winding of a residual into a root count

Let f be analytic inside and on a positively oriented simple closed contour Γ, with no zeros on it. Suppose z_j is an enclosed zero of multiplicity m_j. Locally $f(z) = (z - z_j)^{m_j} g(z)$ with $g(z_j) \neq 0$, so $f'/f = m_j/(z - z_j) + g'/g$. Cauchy's residue theorem gives

$$N = \frac{1}{2\pi i} \oint_\Gamma \frac{f'(z)}{f(z)}\, dz = \sum_j m_j.$$

A meromorphic f instead gives number of zeros minus poles, counted with multiplicity. On the image curve $f(\Gamma)$ the same integral equals its change of complex argument divided by 2π. This provides an alternative if only complex function evaluations are available. Both versions require that the contour avoid zeros and poles. Moving the contour through one changes its integer count discontinuously; a point extremely close to the boundary demands finer sampling or a different contour.

For a circle $z = c + Re^{i\theta}$, periodic trapezoidal quadrature evaluates the integral as the mean of $(z - c)f'(z)/f(z)$. The smoothness and distance to nearby singularities govern convergence. Simply rounding the result to the nearest integer can disguise severe undersampling. Compare independent contour radii and doubled angular grids, inspect the minimum $|f|$ on the boundary, and examine unwrapped phase increments. A grid with only endpoint samples of a rapidly rotating image can miss entire turns even when successive complex values look nearly identical. An upper bound on phase speed or adaptive interval enclosure is needed for a rigorous count from discrete samples.

Locate small groups after counting them

Multiply the integrand by z^k. Its residues become $m_j z_j^k$, giving power sums $s_k = (2\pi i)^{-1} \oint z^k f'/f\, dz$. For exactly one simple enclosed root, s_1 is its location. For two simple roots z_1, z_2, obtain their sum s_1, and their product $(s_1^2 - s_2)/2$ from the identity $(z_1 + z_2)^2 = z_1^2 + z_2^2 + 2z_1 z_2$. Solve the resulting quadratic for initial guesses, then polish on the original residual and recheck the contour count. For large clusters reconstructing a high-degree polynomial from power sums can be ill-conditioned; recursively subdivide or use a contour method designed for stable extraction.

Take $f(z) = (z - 0.2)(z + 0.4)(z - 2)$ on the unit circle. Its count is two; its moments are

$$s_1 = -0.2,\ s_2 = 0.20.$$

The reconstructed product is $(0.04 - 0.20)/2 = -0.08$, giving $z^2 + 0.2z - 0.08 = 0$ and roots $0.2, -0.4$. The exterior factor changes the residual values and Newton basins but contributes no interior pole of f'/f, hence no moment. A contour enclosing all three instead changes the power sums. Severe

cancellation can occur when several large root powers sum to a small number. Shift and scale coordinates about the contour centre before evaluating high-order moments.

Use a boundary inequality to protect a count under perturbation

If an analytically tractable g and target $f = g + h$ satisfy $|h(z)| < |g(z)|$ on Γ, then the continuous homotopy $g + th,\ 0 \le t \le 1$, has no boundary zero. Its winding number around zero cannot change continuously between distinct integers, so f and g enclose the same number of zeros. This is Rouché's theorem, useful when a perturbation is small on the boundary even if it strongly splits individual roots inside. The strict bound must cover the entire curve. Observations at a handful of sample points without a bound between them are numerical evidence, not a theorem.

For $f(z) = (z-1)^2 + \epsilon z,\ |\epsilon| = 0.001$, use a circle of radius 0.1 about one. Its unperturbed modulus is 0.01 and the perturbation has magnitude at most 0.0011. Both zeros stay inside this circle. Their individual shifts can be of order $\sqrt{|\epsilon|}$ because the reference zero is double. A small coefficient perturbation can therefore leave a reliably countable cluster while making its individual members comparatively sensitive. This distinction between a cluster as a whole and a single ill-conditioned root is important when a model residual depends on uncertain fitted coefficients.

Count one movable complex zero by discrete contour integration

The experiment integrates the logarithmic derivative of a polynomial with one zero strictly inside the unit circle. Its exact count is therefore one everywhere in the slider range. The real and imaginary parts of the quadrature estimate are plotted separately against the number of uniformly spaced angular samples. At radius 0.9 on the positive real axis, eight samples give about 1.756; rounding the real part would report two zeros. Moving the root farther from the contour makes convergence faster without changing the true count.

Rotate the root slightly while keeping its radius fixed. The absolute distance from the contour is unchanged, but its phase relative to the sampling nodes changes. A nonzero imaginary estimate and oscillatory errors may appear even though the exact winding number is real. The finite-grid formula for a root at complex a is one divided by one minus a raised to the number of samples. Use it to predict locations of large errors before reading the graph. An apparent agreement for one particular sample count can be accidental; compare successive refinements and the complex error.

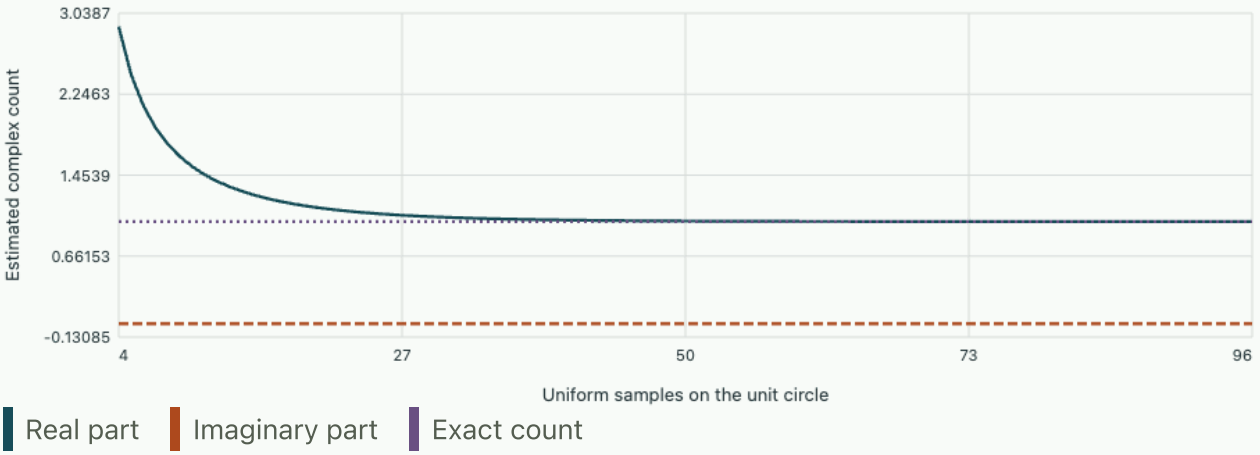
Open this experiment online

Try the calculation

Root radius **0.9**

Root angle (degrees) **0**

Inspected samples **8**



At 8 samples: 1.75583 + (-0.00000)i. True count 1.

Tutorial · audit a contour computation

Calculate the exact aliasing of a one-root count

For $f(z) = z - a$ on the unit circle with $|a| < 1$, use M equally spaced points. Derive the trapezoidal estimate of N and evaluate it for real $a = 0.9, M = 8$.

▼ Work through the solution

At each point $zf'/f = 1/(1 - a/z) = \sum_{k \geq 0} a^k z^{-k}$. Averaging roots of unity cancels all terms except powers divisible by M, giving $N_M = 1/(1 - a^M)$. At $M = 8$ it is about 1.756, which would round to the wrong integer two. Increasing to $M = 64$ gives about 1.00118. This explicit example illustrates why apparent smoothness of the original analytic function does not ensure adequate sampling of its logarithmic derivative.

Recover a repeated root from moments

A contour yields $N = 2, s_1 = 1.2, s_2 = 0.72$ to high numerical accuracy. Infer the candidate roots and explain the sensitivity to a small error in s_2.

▼ Work through the solution

The product is $(1.44 - 0.72)/2 = 0.36$; the polynomial $z^2 - 1.2z + 0.36$ has double root 0.6. A change δ in s_2 changes the product by $-\delta/2$, giving inferred splitting $\pm\sqrt{\delta/2}$ relative to 0.6. Tiny moment noise can therefore mimic two distinct roots. Verify a multiplicity claim against derivative information on the original function and clarify whether distinct roots can actually be resolved.

Explain a zero net winding that hides a zero

A rational residual has a simple zero at 0.1 and a simple pole at -0.2 inside the unit circle. What would the count report and how could one recover the desired zero count?

▼ Work through the solution

The logarithmic derivative has residues plus one and minus one, so net winding is zero even though a zero exists. If the pole and its order are known, add the pole count to the measured winding or multiply by a known denominator that removes it and then check the resulting analytic function. A near pole–zero cancellation can be difficult to resolve numerically. Treating a rational or discretely singular residual as a globally analytic polynomial would make the count's interpretation incorrect.

Further reading

- ETH Zürich · numerical methods for CSE, polynomial roots and nonlinear systems
- NIST Digital Library of Mathematical Functions · contour integration, residues and zeros
- Course investigation · polynomial roots and conditioning
- Solve an implicit cooling step with a changing residual

Nonlinear systems and iterative methods · 5. Where does a complex Newton step take its starting point?

A single cubic has three equally simple roots. Derive the iteration, colour starting values by destination and use a separate convergence history to distinguish a bad first step from slow local convergence.

Continue the tangent step into the complex plane

A holomorphic function has a complex linear approximation $f(z+s) \approx f(z) + f'(z)s$. Requiring this approximation to vanish gives $s = -f(z)/f'(z)$ wherever the derivative is nonzero. For $f(z) = z^3 - 1$, the resulting map is

$$N(z) = \frac{2z}{3} + \frac{1}{3z^2}.$$

The second term is a pole at zero. Starting at a small positive number sends the next iterate far to the right, after which large iterates return approximately by multiplication by 2/3. At a root r, write $z = r + e$ and substitute $r^3 = 1$. Exact cancellation of the linear error gives

$$N(r+e) - r = e^2(3r + 2e)/[3(r+e)^2].$$

Sufficiently near r, the new error is about e^2/r. This argument predicts the speed after entry into a local neighbourhood. It has no information about which root a distant initial point will reach.

Let $s = \exp(2\pi i/3)$. Since $s^3 = 1$, symmetry gives

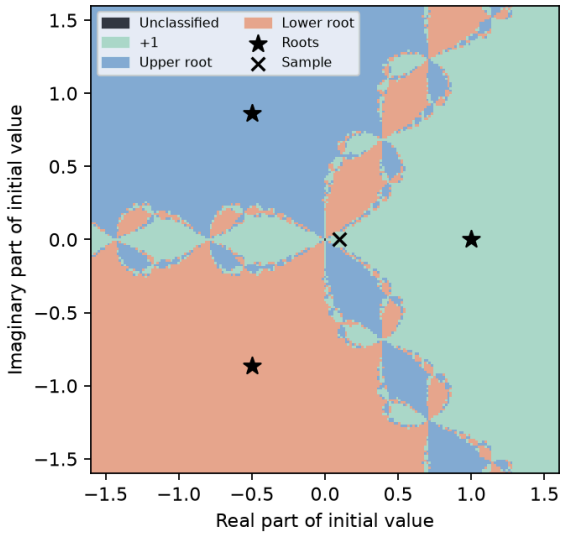
$$N(sz) = sN(z).$$

The entire sequence, including its root, rotates with the start. Conjugate starting points likewise produce conjugate histories because the coefficients are real. These are two independent checks for numerical basin maps. We classify a pixel only once its residual falls below a declared tolerance. A starting pole and a trajectory that exhausts its iteration allowance can both appear unclassified. Increasing the allowance separates these cases for many pixels.

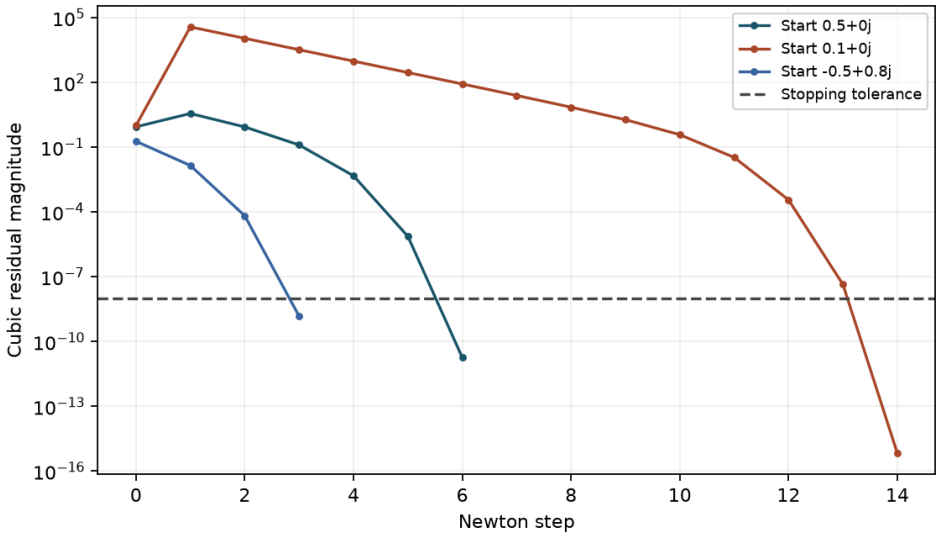
Follow a detour before quadratic convergence

From $z_0 = 1/2$ the first iterate is $1/3 + 4/3 = 5/3$. The next is $10/9 + 3/25 = 277/225 \approx 1.23111$. The error initially rises from 0.5 to 0.667. By contrast, starting at 0.1 gives $z_1 = 33.4$. The subsequent near-geometric return explains the long shoulder in the second figure. Both starts ultimately approach the real root.

Move the orange sample by writing a complex number such as $-0.5 + 0.85i$ in Python as `-0.5+0.85j`. The sample mark and the residual trace follow the same new starting point. The intricate boundaries in the image are computed from independent point trajectories. Refining the grid examines finer samples of those boundaries; the speed on the residual graph belongs to one particular start.



Each grid point starts an independent cubic iteration. Dark grey means the residual criterion was never met within the allowed steps, including the pole at zero.



The dashed line is the actual residual tolerance. A small initial magnitude can produce a large first residual followed by a long return.

Exercise · obtain two basin colours for free

Suppose $0.75 + 0.10i$ reaches the real root in five updates. What are the destinations and update counts after multiplying the start by either non-real cube root of one?

▼ Work through the solution

Write the original iterates as z_k. By induction $N^k(sz_0) = sz_k$ and $N^k(s^2 z_0) = s^2 z_k$. The destinations are s and s^2. Since $(sz_k)^3 - 1 = z_k^3 - 1$, every residual and hence the first crossing of the same threshold is unchanged. Differences in a finely sampled plot may arise because the rotated point falls between pixels.

Exercise · quantify the final two digits

For a real iterate $z = 1.01$, estimate the next error by the quadratic approximation and calculate its exact value using the error formula.

▼ **Work through the solution**

The leading term gives $0.01^2 = 10^{-4}$. The exact error is $10^{-4}(3.02)/(3 \times 1.01^2) \approx 9.8683 \times 10^{-5}$. The difference is of the next order in the initial error. The small residual after this update is about three times the error because $f'(1) = 3$.

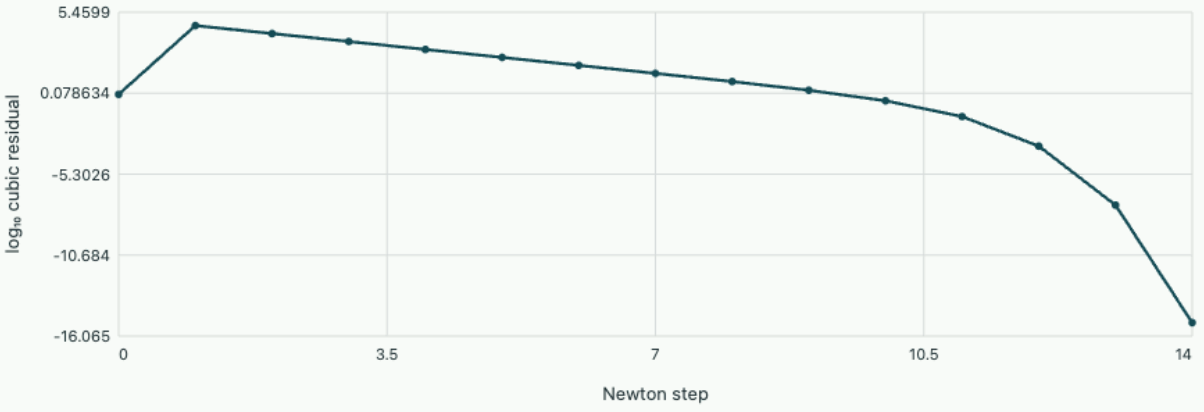
Change parameters · Explore this notebook model

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

Real initial value	0.1
Imaginary initial value	0
Maximum iterations	30



Chosen complex starting point

Converged to 1 + 0 i. Move either coordinate slightly near a basin boundary to test sensitivity.

Open editable notebook Download .ipynb

Nonlinear systems and iterative methods · 6. Certifying a Newton step before the root is known

A small correction is meaningful only together with a curvature estimate. Construct a ball that contains a solution and use a frozen linear solve to make damping decisions independent of equation units.

Bound all future corrections from initial data

Let F be differentiable on a convex neighbourhood, $J_0 = F'(x_0)$ invertible, and fix an induced norm. Suppose $\|J_0^{-1}F(x_0)\| \le b$ and $\|J_0^{-1}(F'(x) - F'(y))\| \le L\|x - y\|$. These estimates must hold on the entire proposed ball. Define the scalar majorant $g(t) = b - t + Lt^2/2$. If $2Lb < 1$, its smaller zero is $r = 2b/(1 + \sqrt{1 - 2Lb})$. Require the closed ball of that radius to lie in the domain.

Set $t_0 = 0$ and apply scalar Newton to g. Since g is decreasing and convex on $[0, r]$, the iterates increase to r. At a point no farther than t from x_0, the Neumann series for $J_0^{-1}F'(x) = I + E$, $\|E\| \le Lt < 1$, bounds $\|F'(x)^{-1}J_0\| \le 1/(1 - Lt)$. If s is an exact Newton correction, the constant and linear terms cancel. Integrating the Jacobian variation along that step gives

$$\|J_0^{-1}F(x + s)\| \le L\|s\|^2/2.$$

At the scalar Newton points exactly the same formula holds with equality for $g(t_{j+1}) = L(t_{j+1} - t_j)^2/2$. Induction now gives $\|x_{j+1} - x_j\| \le t_{j+1} - t_j$. The vector sequence is Cauchy and stays in the ball, its limit is a root and $\|x_* - x_j\| \le r - t_j$. Two roots in that ball would be joined by an averaged Jacobian differing from the identity, after scaling, by at most $Lr < 1$. It is invertible, so the roots coincide.

Use the linearised correction as the measuring stick

Premultiply the equations by any fixed nonsingular matrix S. The Newton correction and the majorant data stay the same since $(SJ_0)^{-1}S = J_0^{-1}$. The Euclidean norm of the raw residual generally changes. For a trial factor λ, form $s = -J^{-1}F(x)$ and $\bar{s} = -J^{-1}F(x + \lambda s)$ with the same factorisation. Linear prediction gives $\bar{s} = (1 - \lambda)s + e$, where local curvature bounded in the current scaled norm implies $\|e\| \le L_x\lambda^2\|s\|^2/2$. Backtrack until $\|e\| \le \lambda\|s\|/2$. Then $\|\bar{s}\| \le (1 - \lambda/2)\|s\|$. For smooth regular points sufficiently small steps pass. Rebuild the Jacobian after acceptance. Global convergence requires additional bounds keeping iterates within a regular level set.

For $F = (x_1^2 - 2, x_2^2 - 1)$ start at $(1.5, 0.9)$ in the maximum norm. Here $b = 0.19/1.8$ and $L = \max(1/1.5, 1/0.9)$. Both follow directly from the diagonal Jacobian. With $2Lb < 1$ the positive quadrant contains the certified ball.

OneSimulation experiment · Test the radius after a change of units

Replace a starting coordinate by a smaller positive value. Recheck the certificate before iterating.

Download the executable notebook · Python source

Adjust the experiment in the browser. Change the two starting coordinates for $F = (x_1^2 - 2, x_2^2 - 1)$. The plot recomputes the iterates, the maximum-norm correction β, the preconditioned derivative bound $L = \max(1/x_{1,0}, 1/x_{2,0})$, and the guaranteed radius whenever $\beta L \leq 1/2$. Move beyond the certified region and observe that failure of this sufficient test does not establish divergence. Rescale the first equation by up to six orders of magnitude: its Jacobian must receive the same scale. The browser uses the explicit diagonal solve; the notebook performs the corresponding OneSimulation dense solve.

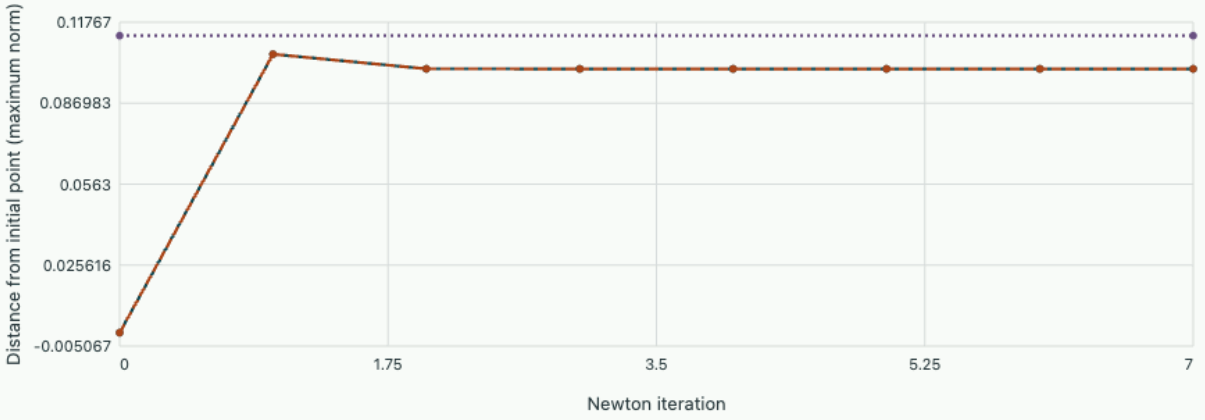
Open the live experiment

Try the calculation

Initial x_1 **1.5**

Initial x_2 **0.9**

$\log_{10}$ first-equation scale **4**



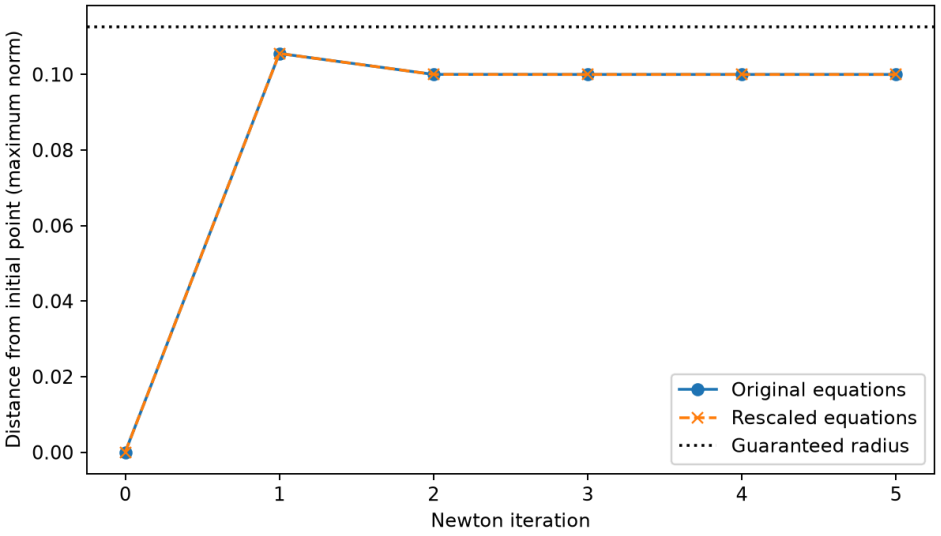
β = 0.10556, L = 1.1111, βL = 0.11728. The certified radius is 0.1126. All displayed iterates remain in that ball. Rescaling a whole equation and its Jacobian leaves the Newton step unchanged up to rounding.

```
"""Test a Newton certificate without knowing the solution in advance."""
import matplotlib.pyplot as plt
import numpy as np
from onesim.numerics.linear.matrices.dense_linear_algebra import solve_dense

start = np.array([1.5, 0.9])
target = np.array([2.0, 1.0])
beta = np.linalg.norm((start**2 - target) / (2 * start), np.inf)
lipschitz = max(1 / start)
radius = 2 * beta / (1 + np.sqrt(1 - 2 * beta * lipschitz))
print(f"beta={beta:.6f}, L={lipschitz:.6f}, certified radius={radius:.6f}")
fig, ax = plt.subplots(figsize=(7, 4), layout="constrained")
for label, scale in [("Original equations", [1., 1.]), ("Rescaled equations", [1e4, 0.01])]:
    x = start.copy()
    distances = [0.0]
    for _ in range(5):
        residual = np.asarray(scale) * (x*x - target)
```

```
        jacobian = np.diag(2 * np.asarray(scale) * x)
        x -= solve_dense(jacobian, residual)
        distances.append(np.linalg.norm(x-start, np.inf))
    assert max(distances) <= radius and np.allclose(x**2, target)
    ax.plot(distances, "o-" if scale[0] == 1 else "x--", label=label)
ax.axhline(radius, color="black", ls=":", label="Guaranteed radius")
ax.set(xlabel="Newton iteration", ylabel="Distance from initial point (maximum norm)")
ax.legend()
fig.savefig("numerics-certified-newton.png", dpi=170)
```

Output from the reference run

```
beta=0.105556, L=1.111111, certified radius=0.112599
```



Two Newton paths coincide inside the same a priori convergence ball.

Tutorial on certification and damping

Compute the first majorant step and use it as a bound

▼ Work through the solution

Scalar Newton at zero gives $t_1 = b$. The vector correction has size b. After that move the distance to the root is at most $r - b$, which can be evaluated without substituting a known root.

Change the starting point to a nearly singular configuration

▼ Work through the solution

For $(1.5, 0.1)$, the second correction is $0.99/0.2 = 4.95$ and $L = 10$. The majorant gives no conclusion. For that coordinate a full step reaches 5.05, whereas halving the step decreases its size. The failure of

the sufficient inequality leaves open whether an actual iteration converges.

Deuflhard · local and global Newton methods

Nonlinear systems and iterative methods · 7. Which correction does an experiment identify?

Separate parameter combinations visible in a nonlinear fit, pass a regular fold by continuation and decide when to refine a boundary-value mesh before trusting its Newton residual.

Solve only for directions constrained by data

Given covariance $C > 0$, whiten the residual $r = C^{-1/2}(f(x) - d)$ and its Jacobian J. Gauss–Newton minimises $\|r + Js\|^2/2$. Write $J = U\Sigma V^T$. Components orthogonal to its range cannot be changed to first order. Each reliably measured direction obeys $v_i^T s = -u_i^T r/\sigma_i$. A direction in the nullspace is free. Setting its correction to zero selects the minimum Euclidean norm in the chosen parameter units.

$$s_\tau = -\sum_{\sigma_i > \tau} (u_i^T r/\sigma_i) v_i, \qquad \|s_\tau\| \le \|r\| / \min_{\sigma_i > \tau} \sigma_i.$$

If the forward derivative carries error e_J, singular values below $\|e_J\|$ cannot be distinguished reliably from zero. Also compare predicted changes to measurement noise over physically allowed parameter displacements. Damping with a penalty gives factors $\sigma_i/(\sigma_i^2 + \lambda^2)$ and is continuous in the singular values. A penalty adds assumptions about scales.

Take two equal-noise readings whose model rows depend only on $a + b$, with sensitivities $(1, 1), (2, 2)$. At $a = b = 0$ readings $(3, 6)$ demand $a + b = 3$. Minimum norm assigns $a = b = 1.5$. Changing parameter units can change that individual assignment while retaining the identifiable sum. A new observation of $a - b$ adds an independent singular direction.

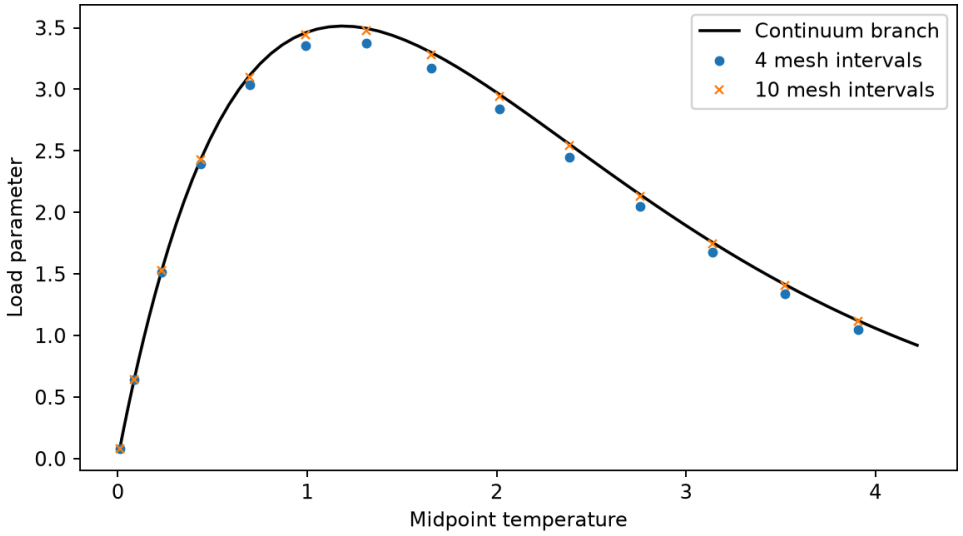
Exchange a failing parameter for a curve coordinate

Let $F(u, p) = 0$ be n equations in $n + 1$ unknowns and suppose $[F_u, F_p]$ has row rank n. Its nullspace supplies a unit tangent t oriented with the preceding tangent. Predict $z_p = z_k + ht$. The corrector solves $F(z) = 0,\ t^T(z - z_p) = 0$. At z_k the augmented derivative is invertible since its nullspace would lie both along t and perpendicular to it. This remains true locally for sufficiently small prediction steps, including a simple fold with singular F_u.

For a boundary-value application consider $u'' + \lambda e^u = 0,\ u(0) = u(1) = 0$. A centred grid gives $F_i = (u_{i-1} - 2u_i + u_{i+1})/\Delta x^2 + \lambda e^{u_i}$. The augmented rows have tridiagonal state derivative with diagonal $-2/\Delta x^2 + \lambda e^{u_i}$ and parameter column e^{u_i}. Use grid-weighted state inner products for tangents so refinement does not change the meaning of arclength. Halve h and repeat the prediction if corrector displacements are too large relative to h. If halving reduces the displacement by about four, the quadratic predictor defect is a geometrical effect.

How can a perfectly converged branch still predict a wrong maximum load?

Newton controls the residual of the chosen grid only. Correct the same branch point on a grid of half the spacing, matching arclength or a monotonic observed amplitude rather than load at the fold. In the regular second-order regime divide the fine–coarse output difference by three to estimate the fine-grid error. Refine the grid when this exceeds the output tolerance. Tighten the augmented linear and nonlinear solves enough that their estimated effect on that output is smaller.



A small nonlinear residual places a point on its discrete branch. Increasing the number of intervals changes the predicted turning load even after the nonlinear solves converge.

Tutorial on a fold with an exact reference

Find a mesh-independent check for the boundary-value branch

▼ **Work through the solution**

Set $u(x) = 2\log[\cosh(a/2)/\cosh(a(x-1/2))]$. Then $u'' = -2a^2\,\mathrm{sech}^2(a(x-1/2))$ and $e^u = \cosh^2(a/2)\,\mathrm{sech}^2(a(x-1/2))$. Hence $\lambda = 2a^2/\cosh^2(a/2)$ and the end values vanish. This parametric curve can test either side of the turning point.

Locate the fold and choose a comparison observable

▼ **Work through the solution**

Differentiating $\log\lambda$ gives $2/a - \tanh(a/2)$. Solve $a\tanh(a/2) = 2$, yielding a near 2.40 and λ near 3.51. The midpoint temperature $2\log\cosh(a/2)$ is strictly increasing for positive a, making it a well-defined matching observable across meshes.

Deuflhard · Gauss–Newton and continuation

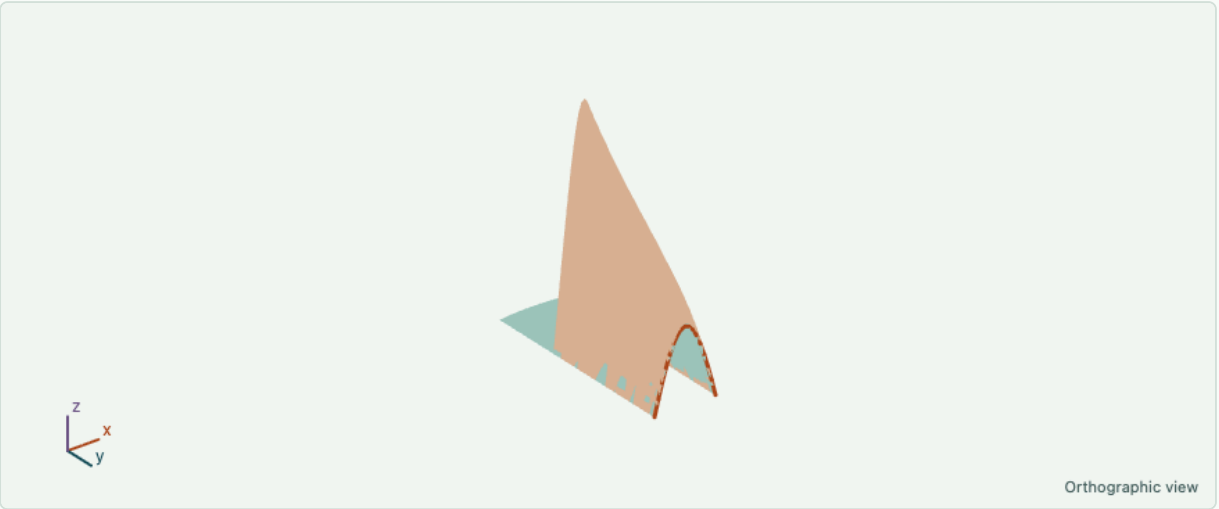
Change parameters · Which correction does an experiment identify?

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.

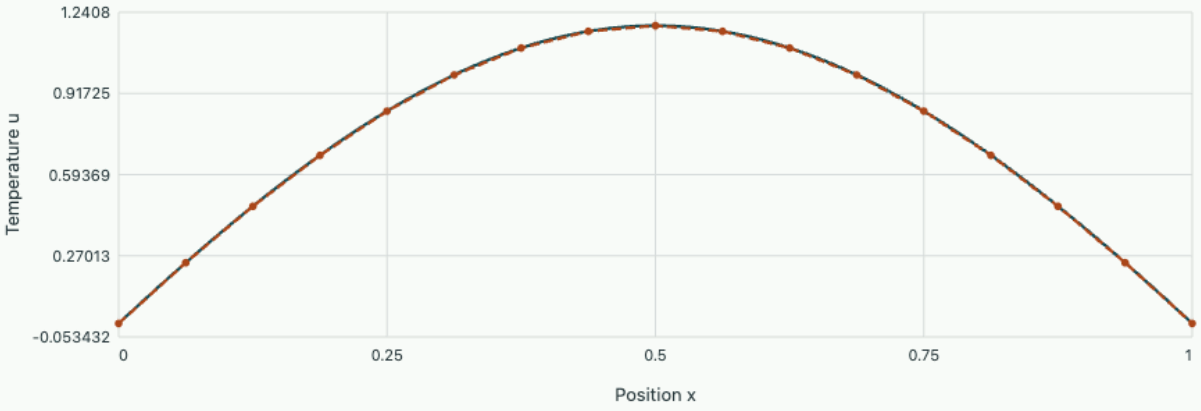
Open this calculation with adjustable inputs

Try the calculation

Branch parameter a	2.4
Mesh intervals	16



A three-dimensional graph of a family of one-dimensional boundary-value solutions. The folded sheet shows two temperature profiles at some loads; its vertical coordinate is temperature. The highlighted line is the computed finite-difference branch at the selected midpoint temperature. Axes: x = x; y = load λ; z = temperature u.



Analytical fixed-midpoint branch Augmented finite-difference solve

Exact load λ = 3.5138, discrete λ = 3.5067, augmented residual 5.4589e-13. Matching midpoint temperature allows comparison on both sides of the fold.

Download the executed OneSimulation notebook · Python source

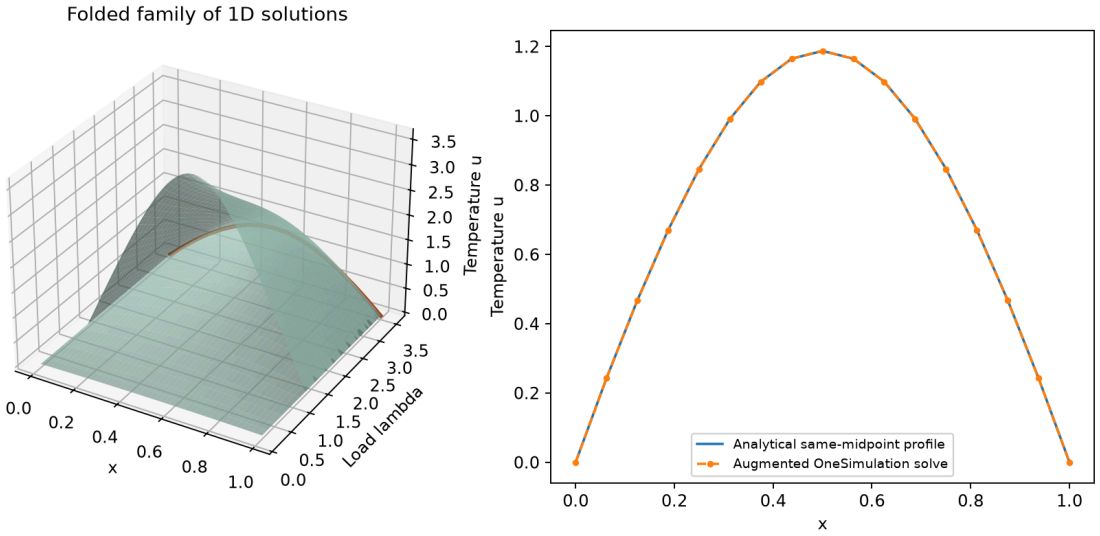
The notebook reproduces the reference case and exposes the mesh or geometry parameters. Its saved output below comes from the actual OneSimulation solver.

```
"""Follow both Bratu branches using OneSimulation augmented Newton solves."""
import argparse
import matplotlib.pyplot as plt
import numpy as np
```

```
from onesim.numerics.linear.matrices.dense_linear_algebra import solve_dense


def branch(parameter=2.4, intervals=16):
    if parameter <= 0 or intervals < 4 or intervals % 2:
        raise ValueError("Use a positive branch parameter and an even interval count >= 4")
    x = np.linspace(0, 1, intervals+1)
    exact = 2*np.log(np.cosh(parameter/2)/np.cosh(parameter*(x-.5)))
    reference = 2*parameter**2/np.cosh(parameter/2)**2
    state, load = exact[1:-1].copy(), reference
    h = 1/intervals
    for iteration in range(15):
        full = np.r_[0., state, 0.]
        residual = np.r_[(full[:-2]-2*state+full[2:])/h**2 + load*np.exp(state),
                         state[intervals//2-1]-exact[intervals//2]]
        if np.linalg.norm(residual) < 1e-10:
            break
        jacobian = np.zeros((intervals, intervals))
        jacobian[:-1, :-1] = (np.diag(-2/h**2 + load*np.exp(state))
                              + np.diag(np.ones(intervals-2)/h**2, 1)
                              + np.diag(np.ones(intervals-2)/h**2, -1))
        jacobian[:-1, -1] = np.exp(state)
        jacobian[-1, intervals//2-1] = 1
        correction = solve_dense(jacobian, -residual)
        state += correction[:-1]
        load += correction[-1]
    assert np.linalg.norm(residual) < 1e-8
    return x, np.r_[0., state, 0.], load, reference, np.linalg.norm(residual)


def experiment(parameter=2.4, intervals=16):
    x, state, load, reference, residual = branch(parameter, intervals)
    print(f"Analytical load: {reference:.9f}; discrete load: {load:.9f}")
    print(f"Augmented Newton residual: {residual:.3e}")
    refined = branch(parameter, 2*intervals)
    assert abs(refined[2]-reference) < abs(load-reference)
    print(f"Load error after doubling the mesh: {abs(refined[2]-reference):.3e}")
    fig = plt.figure(figsize=(10, 4.6), layout="constrained")
    view = fig.add_subplot(121, projection="3d")
    xx, aa = np.meshgrid(np.linspace(0, 1, 45), np.linspace(.3, 5, 45))
    lam = 2*aa**2/np.cosh(aa/2)**2
    field = 2*np.log(np.cosh(aa/2)/np.cosh(aa*(xx-.5)))
    view.plot_surface(xx, lam, field, color="#9dc6b9", alpha=.7, linewidth=0)
    view.plot(x, np.full_like(x, load), state, color="#ad491d", lw=2.5)
    view.set(xlabel="x", ylabel="Load lambda", zlabel="Temperature u", title="Folded family of 1D
solutions")
    ax = fig.add_subplot(122)
    ax.plot(x, 2*np.log(np.cosh(parameter/2)/np.cosh(parameter*(x-.5))), label="Analytical same-
midpoint profile")
    ax.plot(x, state, "o--", ms=3, label="Augmented OneSimulation solve")
    ax.set(xlabel="x", ylabel="Temperature u")
    ax.legend(fontsize=8)
    fig.savefig("complete-numerics-bratu.png", dpi=170)
    return load, residual


if __name__ == "__main__":
    parser = argparse.ArgumentParser(description=__doc__)
    parser.add_argument("--parameter", type=float, default=2.4)
    parser.add_argument("--intervals", type=int, default=16)
    args, _ = parser.parse_known_args()
```

```
    experiment(args.parameter, args.intervals)
    plt.show()
```

```
Analytical load: 3.513830356; discrete load: 3.506676842
Augmented Newton residual: 5.459e-13
Load error after doubling the mesh: 1.786e-03
```



A folded family of one-dimensional Bratu temperature profiles; the highlighted curve is computed with an augmented OneSimulation Newton solve.

Equilibria, branches and stability

Turn a physical stationarity condition into a nonlinear equation. Trace the solutions through loss of stability and decide which roots survive a perturbation.

1. Bifurcation of a rotating hoop
2. Why two coexisting phases share a tangent
3. Mean-field ordering, response and the critical point
4. When does a polar crystal change its preferred state?
5. When a suspended capacitor loses its stable equilibrium
6. Finite-size criticality and the evidence for a phase transition
7. Phase nucleation, metastability and waiting-time statistics

Bifurcation of a rotating hoop

Classical mechanics

Use an effective potential to locate competing equilibria and predict the frequency of small oscillations on either side of a stability change.

FROM FIRST PRINCIPLES

Find the instability threshold of the rotating hoop

Assume a frictionless bead constrained to a vertical circular hoop of radius R. The hoop rotates at prescribed constant angular speed Ω about its vertical diameter. With angle θ measured upward from the bottom, the bead moves along the hoop at speed $R\dot{\theta}$ and around the axis at speed $\Omega R\sin\theta$. These directions are perpendicular, and its gravitational height above the bottom is $R(1-\cos\theta)$. Thus its Lagrangian is

$$L = \tfrac{1}{2}mR^2(\dot{\theta}^2 + \Omega^2\sin^2\theta) - mgR(1-\cos\theta).$$

Apply the one-coordinate Lagrange equation. The rotation term tends to move the bead outward while gravity acts toward the bottom:

$$\ddot{\theta} = \sin\theta(\Omega^2\cos\theta - g/R).$$

Linearizing about the bottom gives

$$\ddot{\theta} \simeq (\Omega^2 - g/R)\theta.$$

For $\Omega^2 < g/R$ the coefficient restores a small displacement; above threshold the bottom has growing small disturbances. Setting the exact right side to zero gives the additional equilibrium and its range:

$$\Omega_c = \sqrt{g/R},$$

$$\cos\theta_* = g/(R\Omega^2) \quad (\Omega \geq \Omega_c).$$

The two nonzero angles are related by reflection. Their existence follows from the prescribed rotation, so a freely slowing hoop would require an additional dynamical coordinate.

DERIVATION AND APPLICATIONS

How high is the barrier between the two tilted equilibria?

In the rotating hoop, a new pair of stable angles appears where the bottom loses stability. Its curvature predicts a small-oscillation frequency; its depth determines whether a larger disturbance can carry the

bead across the bottom. Consider a bead of mass m on a circular hoop of radius R spinning at prescribed constant angular speed Ω about its vertical diameter. Ignore friction. Measure θ upward along the hoop from the bottom. The vertical height change is $R(1-\cos\theta)$ and the distance to the rotation axis is $R\sin\theta$.

The constrained bead has kinetic energy $mR^2(\dot\theta^2+\Omega^2\sin^2\theta)/2$. Collect the terms independent of $\dot\theta$ as the potential governing motion in the rotating coordinate:

$$U_{\text{eff}}(\theta) = mgR(1-\cos\theta) - \tfrac{1}{2}mR^2\Omega^2\sin^2\theta.$$

Write $s=\Omega^2R/g$ and differentiate. A tilted equilibrium exists only for $s>1$. At either stable angle θ_* substitute $\cos\theta_*=1/s$ into the potential and factor the result:

$$U'_{\text{eff}} = mgR\sin\theta(1-s\cos\theta),$$
$$U_{\text{eff}}(\theta_*) = -mgR\,\frac{(s-1)^2}{2s}.$$

The bottom has $U_{\text{eff}}(0)=0$. Subtracting the energies gives the barrier and the threshold angular speed for a bead launched from a tilted minimum toward the bottom:

$$\Delta U = mgR\,\frac{(s-1)^2}{2s},$$
$$|\dot\theta|_{\text{th}} = \sqrt{2\Delta U/(mR^2)}.$$

The velocity here is angular speed relative to the hoop. At threshold the trajectory approaches the unstable bottom asymptotically; crossing takes slightly more initial energy. Near $s=1$ the barrier collapses quadratically even though two distinct equilibria already exist. The externally maintained rotation can exchange ordinary laboratory energy with the bead, while the effective one-coordinate energy is conserved.

WORKED EXAMPLE

Estimate the speed needed to change branches

Change the inputs and investigate this derivation.

Take $m=0.0500$ kg, $R=0.200$ m and choose Ω so that $s=2.00$. The two minima then occur at $\pm60.0^\circ$ and the barrier is

$$mgR/4 = 0.0245\,\text{J}.$$

The tangential threshold speed is $\sqrt{2\Delta U/m}=0.990$ m/s, or 4.95 rad/s along the hoop. A launch at 0.800 m/s has only 0.0160 J and returns to its original branch.

SECOND APPLICATION

Use two measured angles to compare barrier depths

Change the inputs and investigate this derivation.

On the same hoop reduce the rotation until the two minima are at $\pm 36.9^\circ$. Their cosine is about 0.800, giving $s = 1.25$. The new barrier is $mgR(0.25)^2/2.50 = 0.00245$ J, a tenth of the previous value. The new threshold tangential speed is 0.313 m/s. A smaller change in angle than this tenfold energy reduction might suggest is a measurable feature of a shallow double well.

Locate the bifurcation by measuring curvature

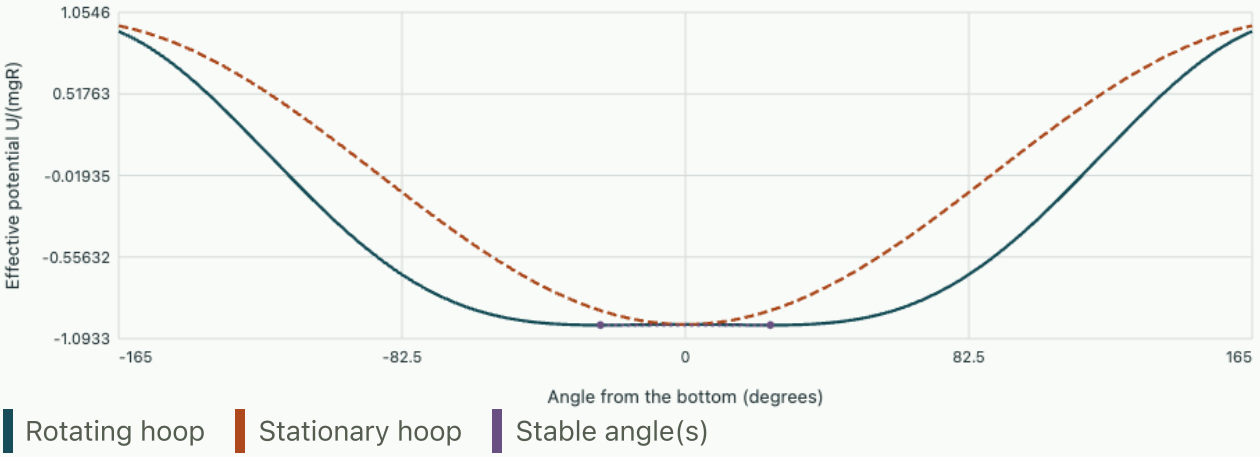
A vertical hoop rotates about its vertical diameter at a prescribed angular speed Ω. A frictionless bead can move along the hoop. Measure θ from the bottom. The laboratory kinetic energy is $mR^2(\dot{\theta}^2 + \Omega^2 \sin^2\theta)/2$. In the equation for the bead the effective potential per unit mgR is $-\cos\theta - (\Omega^2 R/2g)\sin^2\theta$. Its derivative factors as $\sin\theta[1 - (\Omega^2 R/g)\cos\theta]$.

Open this experiment online

Try the calculation

Hoop radius (m) **0.3**

Imposed rotation (rad/s) **6**



Critical speed 5.7184 rad/s. Bottom curvature in units of mgR is -0.10092. Stable angular magnitude 24.724°.

Plot the potential before and after crossing $\Omega_c = \sqrt{g/R}$. At a stable equilibrium the slope vanishes and the curvature is positive. Above the critical speed the bottom loses positive curvature and two stable angles appear. The motor holding Ω fixed exchanges angular momentum with the system. A freely rotating hoop with fixed total angular momentum describes a different experiment.

Reduce the constraint to one coordinate

A bead slides on a vertical circular hoop of radius R. An external motor rotates the hoop at prescribed constant angular velocity Ω about its vertical diameter. Let θ be the angle from the lowest point. The bead's height relative to that point is $R(1-\cos\theta)$ and its distance from the rotation axis is $R\sin\theta$. Assume negligible friction and a rigid constraint. The bead moves both along the hoop and azimuthally with it, even if θ is momentarily stationary.

In laboratory coordinates its speed squared is $R^2\dot\theta^2 + \Omega^2 R^2 \sin^2\theta$. Subtract gravitational potential energy from this kinetic energy to construct a one-coordinate Lagrangian. Lagrange's equation gives

$$\ddot\theta = \sin\theta(\Omega^2\cos\theta - g/R).$$

Although motion in the laboratory is constrained by a driven apparatus, the reduced equation is autonomous for constant Ω.

Track equilibria as the rotation rate changes

The torque vanishes at the bottom and the top because $\sin\theta = 0$. Off-axis equilibria also appear when $\cos\theta = g/(\Omega^2 R)$. They are possible only if the rotation exceeds $\Omega_c = \sqrt{g/R}$. Each positive-angle equilibrium then has a negative-angle partner. The two have equal energies because the apparatus is symmetric.

For small displacement from the bottom,

$$\ddot\theta \simeq -(g/R - \Omega^2)\theta.$$

Below threshold the bottom is stable with angular frequency $\sqrt{g/R - \Omega^2}$. As the threshold is approached this frequency tends to zero; the potential flattens. Above threshold the sign reverses and a small perturbation grows. Linearising about either displaced equilibrium instead gives a positive frequency squared $\Omega^2\sin^2\theta_{\mathrm{eq}}$. Measuring this small-amplitude response distinguishes two predictions: where the bead rests and how steeply it is restored when disturbed.

Interpret the effective energy

The reduced motion admits the effective potential $U_{\mathrm{eff}} = mgR(1-\cos\theta) - m\Omega^2R^2\sin^2\theta/2$. Its derivative predicts the reduced equation. Its negative centrifugal term differs in sign from the positive laboratory azimuthal kinetic energy. The motor supplies or absorbs work as the bead's distance from the rotation axis changes. Treating the bead's laboratory mechanical energy alone as conserved would therefore lead to the wrong equilibrium.

At exactly the critical rate, the quadratic terms in a small-angle expansion cancel and the first restoring term in U_{eff} is positive and quartic. A finite disturbance can still oscillate there, but its period depends on its amplitude. An arbitrarily small oscillation cannot be assigned a finite harmonic frequency at that point.

Estimate the transition in a real apparatus

With $R = 0.250$ m and $g = 9.81$ m/s² the critical rate is 6.26 rad/s, or about 59.8 rpm. At 5.0 rad/s the bead returns to the bottom with small-oscillation frequency $\sqrt{39.24 - 25} = 3.77$ rad/s. At 8.0 rad/s the two new equilibria satisfy $\cos\theta = 39.24/64 = 0.613$, or $\theta = \pm 52.2^\circ$. Small oscillations about either have angular frequency $8\sqrt{1 - 0.613^2} = 6.32$ rad/s.

When observing the transition, tilt of the axis or a small asymmetry of the hoop can favour one side. Hysteresis-like delays during a fast sweep can also arise from finite settling time in weak damping. Recording both equilibrium angles on slow sweeps and the independently measured oscillation frequency supplies a more discriminating test of the symmetric bifurcation than simply noting which side the bead occupies.

Break the symmetry with a weak bias

The two off-centre positions of the ideal rotating bead have equal energy. A small constant generalized torque τ_b favouring positive θ breaks that symmetry. Over a small working angular range it adds a term $-\tau_b\theta$ to the effective potential. Write the original potential relative to its value at the bottom as $mgR(1 - \cos\theta) - m\Omega^2 R^2 \sin^2\theta/2$. Expanding near the critical rate $\Omega_c^2 = g/R$ gives quadratic coefficient $mR^2(\Omega_c^2 - \Omega^2)/2$. The quartic coefficient evaluated at threshold is $mR^2\Omega_c^2/8$.

On the high-speed side define $a = \Omega^2 - \Omega_c^2 > 0$, $b = \Omega_c^2/2$, and $h = \tau_b/(mR^2)$. To leading order stationarity and local stability become

$$b\theta^3 - a\theta - h = 0, \qquad 3b\theta^2 - a > 0.$$

With zero bias there are two symmetric minima and an unstable central position. Increasing positive bias deepens the positive minimum and eventually destroys the negative one. At the point where the negative minimum meets the adjacent maximum, both the stationarity equation and its derivative in θ vanish. Thus

$$\theta_* = -\sqrt{a/(3b)},$$

$$h_* = (2a/3)\sqrt{a/(3b)}.$$

The critical bias scales as $a^{3/2}$. Near the threshold a surprisingly small imperfection can remove one of the two resting positions entirely.

Estimate the bias needed to select a branch

Use a bead of mass 0.0500 kg on a hoop of radius 0.250 m. Then $\Omega_c^2 = 39.24$ s⁻². Set $\Omega^2 = \Omega_c^2 + 0.600$ s⁻², making $b = 19.62$ s⁻². The unbiased displaced equilibrium angles are approximately $\pm\sqrt{0.600/19.62} = \pm 0.175$ rad. As positive bias is increased, the negative branch disappears near $\theta_* = -0.101$ rad. The corresponding generalized acceleration is $h_* = 0.0404$ s⁻², giving

$\tau_b = mR^2 h_* = 1.26 \times 10^{-4}$ N m. A tangential force with moment arm R would supply about 0.505 mN.

This estimate uses both small angles and operation close enough to the critical speed that quartic truncation is accurate. An exact calculation differentiates the full trigonometric potential twice and solves the two equations simultaneously. In a slow sweep with weak damping, a bead initially on the unfavoured branch can remain there until its minimum disappears. Vibration or thermal agitation may make it switch earlier by crossing the remaining barrier. Repeating the sweep in both directions and at several bias amplitudes helps distinguish an imperfect constraint from random switching in a nearly symmetric system.

The same model predicts increasing sensitivity before an actual switch. Differentiate the cubic stationarity condition with respect to bias while following a particular equilibrium branch:

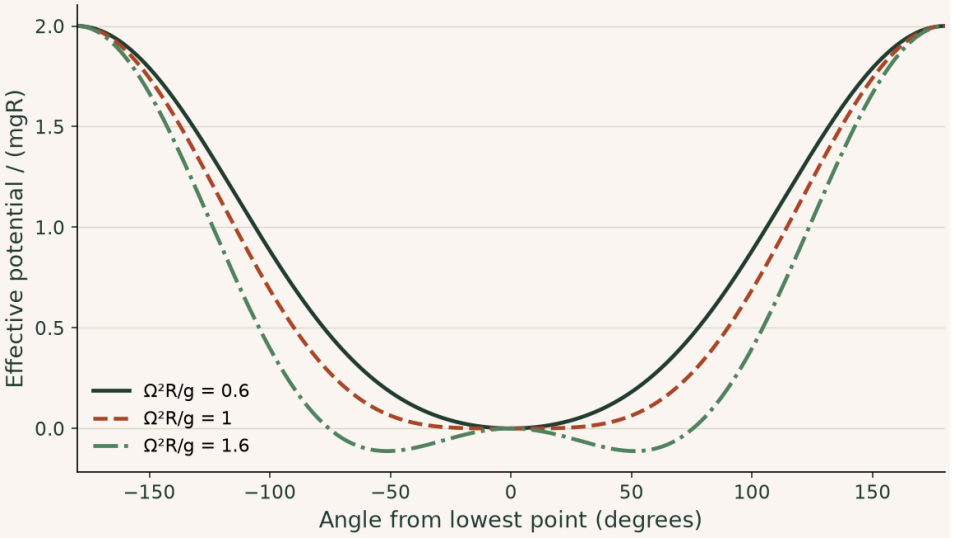
$$d\theta/dh = 1/(3b\theta^2 - a).$$

Its denominator is the local dimensionless restoring curvature. Thus a branch nearing disappearance both moves a large angular distance under a small bias increment and shows a very low small-oscillation frequency. Measuring those two effects on the same branch tests the local shape of the potential without requiring a precise measurement of the noisy switching moment. Away from a minimum the static slope can instead be negative, describing an unstable mathematical solution.

The sharp loss of a branch is best resolved with bias steps small compared with the measured interval between its creation and disappearance.

Where will a bead settle on a rotating hoop?

Choose zero potential at the bottom. Gravity contributes $mgR(1-\cos\theta)$; the rotating-frame centrifugal term contributes $-m\Omega^2R^2\sin^2\theta/2$. Inspect both the slope and the curvature at the bottom while increasing speed.



At $\Omega^2 R/g = 1.6$ the two minima satisfy $\cos\theta = 0.625$, at approximately $\pm 51.3°$. At the critical value the quadratic term vanishes; the bottom is much flatter than an ordinary harmonic minimum.

▶ **Read values from the plot**

Why two coexisting phases share a tangent

Thermodynamics and statistical physics

A material prepared with an intermediate composition can separate into two regions of different compositions. Its total composition is conserved. How can a free-energy graph predict both final compositions and the amount of each phase?

Minimise the free energy of a mixture

Work at fixed temperature and volume with a constant density of equivalent sites. Let $f(c)$ be the Helmholtz free energy per site of a spatially uniform state with fraction c of species B. If the two regions have compositions c_1, c_2 and the second contains fraction x of all sites, conservation requires

$$\bar{c} = (1-x)c_1 + xc_2.$$

Neglect interface energy per site in the macroscopic limit. The mixture free energy per site is then the weighted sum of the two f's, a point on the chord joining them.

Introduce a multiplier μ and minimise $(1-x)f(c_1) + xf(c_2) - \mu[(1-x)c_1 + xc_2 - \bar{c}]$. Varying c_1, c_2 when both regions are present gives

$$f'(c_1) = f'(c_2) = \mu.$$

This derivative measures the chemical work to replace an A particle by B. Varying x supplies the missing condition,

$$f(c_2) - f(c_1) = \mu(c_2 - c_1).$$

The common slope therefore equals the slope of the connecting chord. A straight line must touch the free energy at both endpoints and lie below competing homogeneous states. Equivalently $\int_{c_1}^{c_2} [f'(c) - \mu]\, dc = 0$. This equal-area condition is the composition version of Maxwell's construction. For liquids plotted as pressure versus specific volume, the conjugate slope acquires the sign from $p = -\partial F / \partial V$.

Find the coexistence and stability boundaries separately

Take $f(c) = A[(c-m)^2 - d^2]^2 + \mu_0 c$, with $A > 0$, $m = 0.5$ and $0 < d < 0.5$. Subtracting the linear term tilts all trial chords by the same amount. At $c = m \pm d$ the remaining term vanishes with zero derivative. Hence the common tangent is $\mu_0 c$, and the coexistence compositions are $m - d, m + d$. At $d = 0.30$ a sample of average composition 0.45 separates into 0.20 and 0.80 regions. Conservation gives $x = (0.45 - 0.20)/(0.80 - 0.20) = 5/12$ for the B-rich region.

Now perturb a uniform state into two equal parts at $c \pm \eta$. Their average energy differs from the uniform energy by $f''(c)\eta^2/2 + O(\eta^4)$. Here $f'' = 4A[3(c-m)^2 - d^2]$. Between $m - d/\sqrt{3}$ and $m + d/\sqrt{3}$ infinitesimal fluctuations reduce the energy. Between either of these limits and the nearest coexistence endpoint a uniform state resists small fluctuations but a sufficiently large rearrangement can still lower

the energy. This accounts for metastability and for the distinction between spinodal and coexistence boundaries. The uniform model does not predict a nucleation time because interfaces carry no energy in this construction.

Get a phase fraction from a graph

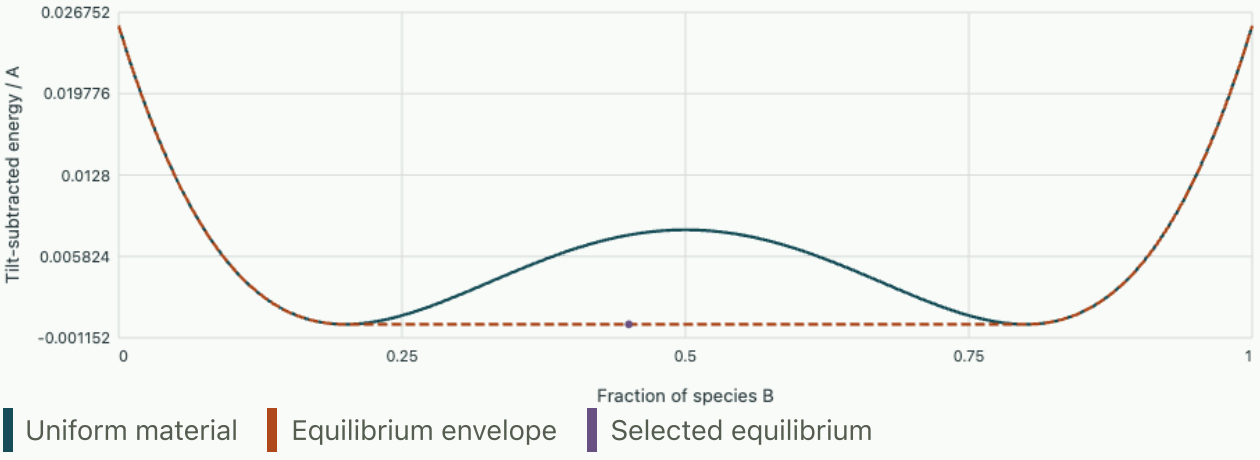
The display subtracts $\mu_0 c$ and measures energies in A. Move the mean composition across the whole interval. Outside the coexistence window the stable composition is homogeneous. Inside, the endpoints stay put while the rich-region fraction changes linearly. Compare $c = 0.45$ to $c = 0.55$ for the same gap. Reflection symmetry exchanges the two phase fractions and leaves the free-energy saving unchanged.

Open this experiment online

Try the calculation

Composition half-gap d **0.3**

Mean B fraction **0.45**



The B-rich phase occupies 41.67% of sites. Energy saving is 0.007656 A per site.

Tutorial on a composition quench

A homogeneous sample is prepared at $c = 0.26$ with $d = 0.30$. Does an infinitesimal composition fluctuation lower its energy? What fraction of the final mixture occupies the B-rich phase in equilibrium?

▼ **Work through the solution**

The lower spinodal lies at $0.5 - 0.3/\sqrt{3} = 0.327$, so the prepared state lies below it and has positive curvature. Nevertheless 0.26 lies between 0.20 and 0.80. Its equilibrium rich fraction is

$$0.06/0.60 = 0.10.$$

Local stability and the global free-energy minimum answer different experimental questions.

For $m = 0.50$, $d = 0.30$, compare free-energy gains of demixing at mean compositions 0.20 and 0.50 after subtracting the tilt. Report the gain in units of A.

▼ **Work through the solution**

At a coexistence endpoint the pure homogeneous material already lies on the common tangent and has zero gain. At the midpoint the homogeneous energy exceeds the tangent by $d^4 A = 0.0081 A$ per site. The equilibrium mixture uses equal amounts of the two compositions. The gain tends continuously to zero at either endpoint.

Would adding $b + mc$ to the free energy change the two selected compositions or spinodals? What would change in the tangent construction?

▼ **Work through the solution**

Conservation fixes the mean composition, so the new term contributes $b + m\bar{c}$ to every candidate mixture at that mean. The minimiser is unchanged. Second derivatives are unchanged, preserving spinodals. Both endpoint tangent slopes, and thus the chemical-potential difference assigned to coexistence, shift by m.

Further reading · Cambridge statistical physics, phase equilibrium and Maxwell construction.

Mean-field ordering, response and the critical point

Thermodynamics and statistical physics

Interactions can make a magnet favour a nonzero moment even when the applied field vanishes. Averaging the local exchange field gives a solvable model whose critical predictions can be compared with the exact one-dimensional chain.

Build the free energy from the count of spin states

Take N spins $s_i = \pm 1$, nearest-neighbour energy $-J\sum_{\langle ij\rangle} s_i s_j$ and Zeeman energy $-h\sum_i s_i$, where h has energy units. Denote mean spin by m. The fractions of positive and negative spins are $p_\pm = (1 \pm m)/2$. The binomial count and Stirling's approximation give entropy per spin $-k_B(p_+ \ln p_+ + p_- \ln p_-)$. Independent-neighbour mean field replaces each pair average by m^2. With z neighbours and each bond counted once, interaction energy per spin becomes $-zJm^2/2$.

$$f(m) = -\tfrac{1}{2}zJm^2 - hm + k_BT[p_+ \ln p_+ + p_- \ln p_-].$$

Differentiate at fixed T, h. The entropy contribution has derivative $k_BT\,\operatorname{atanh} m$, giving $m = \tanh[(zJm + h)/(k_BT)]$. This equation can have three real roots, so an iterative solver returning a root alone is insufficient. The equilibrium phase minimises f; the second derivative $k_BT/(1 - m^2) - zJ$ distinguishes local minima and maxima.

Expand about the loss of stability

Near zero magnetisation,

$$\operatorname{atanh} m = m + m^3/3 + \cdots.$$

Set $T_c = zJ/k_B$ and integrate the expanded derivative to obtain

$$f(m) - f(0) \simeq \tfrac{k_B}{2}(T - T_c)m^2 + \tfrac{k_BT}{12}m^4 - hm.$$

At zero field and $T > T_c$, $m = 0$ is stable. Below T_c, symmetry gives two minima whose leading magnitudes satisfy $m^2 \simeq 3(T_c - T)/T$. Thus the order parameter vanishes like $(T_c - T)^{1/2}$ close to T_c. At $T = T_c$ a small nonzero field instead gives

$$m \simeq [3h/(k_BT_c)]^{1/3}.$$

These powers describe this model's limiting behaviour. For a finite sample symmetry makes the unbiased ensemble average zero if it samples both minima. Spontaneous magnetisation refers to the thermodynamic limit with an infinitesimal field selecting a sign.

Compute a response, then compare with an exact counterexample

Above T_c the small-field stationary equation gives

$$\chi = \partial m/\partial h = 1/[k_B(T - T_c)].$$

Below T_c, differentiate the exact stationarity equation on the positive branch, obtaining $\chi = [k_B T/(1 - m^2) - zJ]^{-1}$. Expanding near T_c there gives $\chi \simeq [2k_B(T_c - T)]^{-1}$, half the above-transition amplitude at equal distance. At $T = 0.8T_c$, the full zero-field solution is about $m = 0.7104$; the small-m estimate $\sqrt{3(1 - 0.8)/0.8} = 0.866$ is already visibly inaccurate.

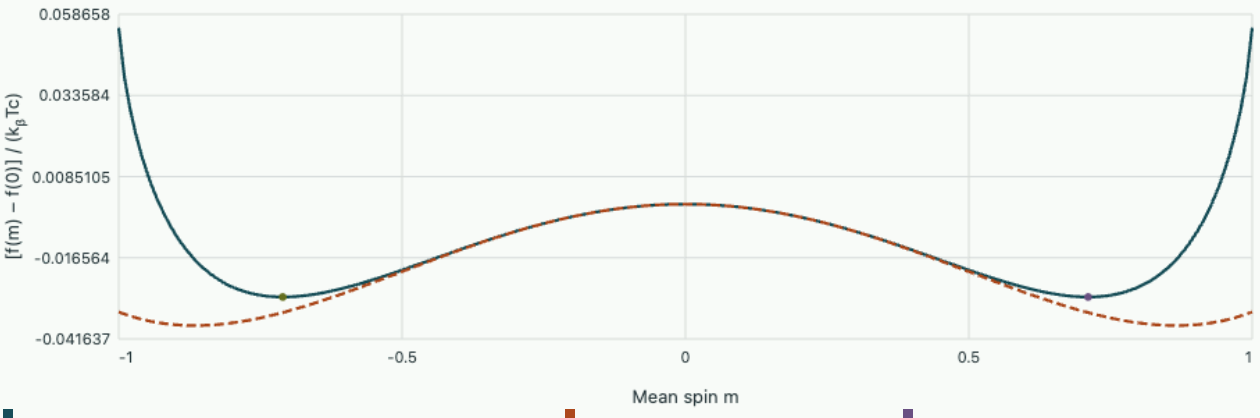
Mean field predicts an ordering temperature even for a one-dimensional chain with $z = 2$. The exact finite-temperature correlation length in the earlier transfer-matrix solution is finite at all $T > 0$. Spatial fluctuations and domain walls invalidate the independent-neighbour closure in one dimension. The experiment plots the full mean-field free energy and its quartic approximation for the same normalised T, h. Find the minima before inspecting the calculated order parameter. Compare how a weak bias changes their relative depths.

Open this experiment online

Try the calculation

Temperature T/Tc **0.8**

Field h/(k$_\beta$Tc) **0**



Full independent-neighbour free energy · Quartic approximation · Equilibrium minimum · Symmetry-related minimum

Equilibrium mean spin ±0.7104. Curvature at the minimum 0.6151 in units of k$_\beta$Tc.

Tutorial on critical response

At $T = 1.10T_c$, what is the linear-response mean spin expected at $h/(k_B T_c) = 0.002$? At the critical point use the cubic equation to estimate m for the same field.

▼ **Work through the solution**

Above the transition $m \simeq 0.002/0.10 = 0.020$. At criticality $m \simeq (0.006)^{1/3} = 0.182$. The larger response signals the loss of the quadratic restoring term; linear extrapolation through T_c would give a divergent susceptibility.

At $h = 0$ and $T = 0.95T_c$, is $m = 0$ a possible output from a fixed-point solver? How should an equilibrium calculation prevent a wrong interpretation?

▼ **Work through the solution**

Starting the symmetric iteration at exactly zero returns zero repeatedly, even below T_c. Its free-energy curvature is $k_B(T - T_c) < 0$; it is a local maximum. Compare the free energies or bracket the two symmetry-related nonzero stationary points before reporting an equilibrium moment.

Further reading · Cambridge statistical physics, mean-field Ising and Landau theory.

When does a polar crystal change its preferred state?

Fields and the response of matter

Derive stable and unstable dielectric responses from a free-energy landscape. Separate a reversible small-signal capacitance from a switching experiment.

Why one voltage can support several measured polarisations

Some insulating crystals develop a spontaneous electric polarisation as temperature is lowered. The high-temperature phase may be symmetric under inversion so that states with polarisation $+P$ and $-P$ are equivalent in zero field. Near a continuous transition, treat one component of polarisation as an order parameter. A local free-energy density compatible with the symmetry is $f(P) = a(T - T_c)P^2/2 + bP^4/4$ with $a, b > 0$. The omitted terms limit how far in temperature or field this truncated expansion is reliable. The positive quartic term keeps the energy bounded when the quadratic coefficient is negative.

For a device whose applied electric field E is controlled, the appropriate potential in this simple model is $g(P; E) = f(P) - EP$. The linear term records external work; changing what is held fixed changes the energy to minimise. Differentiation gives the constitutive equation $E = a(T - T_c)P + bP^3$. An algebraic root is an equilibrium candidate. A stable local response also requires positive second derivative $a(T - T_c) + 3bP^2$. A very accurate rootfinder can return an unstable stationary point if this second test is omitted.

Find the equilibrium branches and their response

Introduce scaled polarisation p, field e and temperature parameter s such that $g = p^4/4 + sp^2/2 - ep$. Then $e = p^3 + sp$. For $s > 0$, the slope $3p^2 + s$ stays positive. Each field has one stable solution. At $e = 0$, $p = 0$ and the small-signal susceptibility $dp/de = 1/s$ increases when approaching the ideal critical point from above.

For $s = -1,\ e = 0$, stationary values are $-1, 0, +1$. The middle root has curvature –1 and is a local maximum. Each outer root has curvature 2 and is a minimum. The susceptibility about a selected outer state is $1/2$ in the scaled units. The fact that the two domains can carry opposite zero-field polarisation does not make the harmonic response negative. The mere existence of a negative-slope segment in the cubic constitutive curve is insufficient to infer a stable isolated small-signal device with that differential response.

Derive the limit of a metastable homogeneous state

Imagine preparing the negative-polarisation state at $s = -1$, then increasing the applied field slowly. A positive field lowers the energy of the positive state. The negative state can persist locally behind a barrier even after it ceases to be the global minimum. A model of perfectly homogeneous evolution without early nucleation follows this metastable branch until the curvature vanishes. Solve $3p^2 - 1 = 0$, choose $p = -1/\sqrt{3}$, and insert it into $e = p^3 - p$. The positive limiting field is

$$e = 2/(3\sqrt{3}) \simeq 0.3849.$$

At the opposite spinodal the signs reverse.

This is a limit of the particular energy model. Real switching can occur earlier through domain walls, nucleation, defects or thermally activated barrier crossing. Observing a smaller coercive field in a specimen is not evidence that the cubic equation was solved inaccurately. To compare an actual hysteresis loop with a calculation, state the sweep rate, boundary circuit, sample preparation and a dynamical or nucleation model in addition to equilibrium free energy.

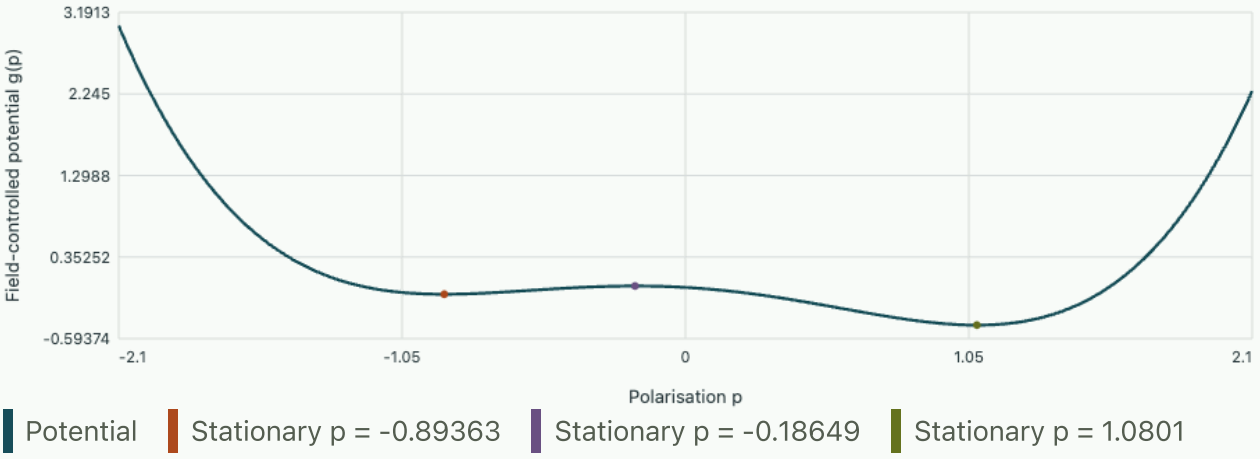
Read a landscape rather than tracing an unlabelled loop

Open the live experiment

Try the calculation

Temperature coefficient s **-1**

Applied field e **0.18**



Stationary polarisations: -0.89363 (minimum), -0.18649 (maximum), 1.0801 (minimum).

Set $s = -1$. At zero applied field, locate the two equal wells. Increase positive e gradually and watch the negative well become shallower. The browser locates real stationary points of the cubic, classifies them by curvature and reports their positions. Its plotted potential shows the energy barrier separating basins. Choose $s > 0$ and repeat to see why fitting the same S-shaped hysteresis loop above a continuous transition would be inappropriate.

A below-transition sample can also form spatially varying domains. Adding a positive gradient term $K(\nabla P)^2/2$ penalises interfaces, while depolarising fields can favour a mixture with reduced net surface charge. This competition sets new length scales absent from the homogeneous model. An electrode measuring total charge can return near-zero mean polarisation from oppositely oriented domains even

though each domain is strongly polarised. A local probe and a bulk capacitance measurement therefore constrain different parts of the model.

Turn a capacitance measurement into a consistency test

For a small harmonic voltage imposed about a prepared local minimum, the differential response is $\delta P/\delta E = 1/[a(T - T_c) + 3bP^2]$, provided the driving frequency is slow enough for this local variable to respond and the amplitude is small enough to stay in the same well. Add the vacuum and fast polarisation terms when converting to total displacement D. In a parallel-plate geometry capacitance per area then follows from $\delta D/\delta E$ divided by thickness.

A large-signal charge–voltage measurement crosses different branches and may dissipate work around a loop. The energy per volume over a closed polarisation cycle is $\oint E\,dP$ with the sign convention for electrical work. A single equilibrium graph of $f(P)$ returns to its initial state with no prescribed loop area. Measured hysteresis thus carries dynamical information that a polynomial fit to one reversible branch cannot uniquely reconstruct.

Tutorial

For $s = -4$ in the scaled model, find zero-field minima and the positive homogeneous spinodal field.

▼ **Work through the solution**

Zero-field minima lie at $p = \pm 2$. The negative spinodal is

$$p = -\sqrt{4/3} = -2/\sqrt{3}.$$

Substitution gives positive field $16/(3\sqrt{3}) \simeq 3.08$. The $|s|^{3/2}$ dependence follows from the cubic.

A Newton solver initialised at $p = 0$ reports exact convergence for $s = -1, e = 0$. What additional observation rejects that result as the prepared stable state?

▼ **Work through the solution**

Its residual is exactly zero but the curvature $3p^2 + s = -1$ is negative. A small perturbation lowers the potential. Check a measured weak-field slope or the relaxation after a perturbation, and compare the energy of the two minima.

Why can the homogeneous coercive field be much larger than the measured coercive field?

▼ **Work through the solution**

The calculation removes a local barrier only when a spatially uniform state ceases to exist. A real material can switch by creating a small favourable domain and moving a wall; electrodes and defects can lower the needed activation energy. The relevant material and time scales differ.

Compare material descriptions

Use MIT's ferroelectrics and polarisation material to distinguish a microscopic source of displacement from the macroscopic free-energy model used here.

When a suspended capacitor loses its stable equilibrium

Electromechanics, optics and measurement

An electrode suspended on a spring can remain motionless at a bias that makes the calculated electrostatic force grow as it moves. Static force balance alone misses the voltage at which that equilibrium disappears. Deriving the small-signal frequency gives an independent measurement of the same instability.

Start with the electrical and mechanical control variables

A movable conducting plate of area A faces a fixed electrode. Its unstressed gap is g. Define positive displacement x toward the fixed plate, with spring force $-kx$, mass m and viscous damping coefficient b. For $0 \le x < g$ assume ideal parallel plates, negligible tilt and fringing, constant ϵ, and a linear elastic spring. The capacitance is $C(x) = \epsilon A/(g - x)$. The model ends when the plates approach contact or when the suspending structure departs from its specified elastic law. It cannot by itself calculate adhesion, contact resistance or post-contact motion.

Before differentiating electrical energy specify whether the source holds voltage or charge. At fixed charge Q, electrical energy is $Q^2/[2C(x)]$, and its negative displacement derivative gives the force toward increasing x. At fixed voltage V, a source supplies $V\,dQ$ as the plates move. The appropriate potential for mechanical equilibrium is the spring energy plus the field energy minus QV, which here is $\mathcal{U}_V = kx^2/2 - C(x)V^2/2$. Differentiating gives the outward spring reaction and the attractive electrical force $F_e = V^2 C'(x)/2 = \epsilon A V^2/[2(g - x)^2]$. An independent derivation multiplies the conductor pressure $\epsilon E^2/2$ by area with $E = V/(g - x)$. Both give a positive force. Replacing the source potential with the stored energy at fixed voltage would reverse its sign.

Locate all equilibria before choosing a branch

Define $s = x/g$ and $\Lambda = \epsilon A V^2/(2kg^3)$. The static force equation reduces to

$$\Lambda = s(1 - s)^2.$$

At a specified voltage one can find s by scanning the curve $s \mapsto s(1 - s)^2$ over the physically admissible interval $0 \le s < 1$. Its derivative is $(1 - s)(1 - 3s)$, giving an interior maximum $4/27$ at $s = 1/3$. For $0 < \Lambda < 4/27$ there are two physical roots. The smaller is a local minimum of $\mathcal{U}_V$ while the larger is a local maximum. Their merger is a fold of the static equilibrium branch:

$$x_{\mathrm{pi}} = g/3,$$

$$V_{\mathrm{pi}} = \sqrt{\frac{8kg^3}{27\epsilon A}}.$$

The often-quoted one-third displacement is the result for a linear spring, ideal voltage drive and this capacitance geometry. It is not a universal travel limit determined by the gap alone. Additional mechanical restoring forces or a different electrical boundary condition change the fold. Neither positive net electrical force nor the mere existence of a real solution is evidence of stable operation. A root finder

started near the almost closed gap can converge to the unstable large-displacement root with an extremely small static residual.

The two branches can be distinguished experimentally without deliberately reaching contact. At static equilibrium $kx = F_e$. Linearise the total force about x_0 while maintaining voltage at its prescribed value. The incremental restoring stiffness and corresponding weak-damping natural frequency are

$$k_{\text{eff}} = k - \frac{\epsilon A V^2}{(g - x_0)^3} = k\,\frac{1 - 3s_0}{1 - s_0},$$

$$f_{\text{small}} = \frac{1}{2\pi}\sqrt{\frac{k_{\text{eff}}}{m}}.$$

A small push toward the fixed electrode is restored only when $k_{\text{eff}} > 0$. The familiar electrostatic softening is the reason the smaller equilibrium is stable. The small-signal frequency tends to zero as the fold is approached. With appreciable damping the observed amplitude resonance and ringdown oscillation have slightly different frequencies; fit the full mechanical response rather than identifying an arbitrary peak with the undamped f_{small}.

Background · Can a sudden step pull in a statically stable actuator?

An ideal voltage step applied to a plate initially at rest can push it past the unstable equilibrium. During the motion at fixed voltage the effective potential per kg^2 is $u(s) = s^2/2 - b/[2(1 - s)]$, where $s = x/g$ and $b = \epsilon A V^2/(kg^3)$. Neglect damping and begin at $s = 0$. At the dynamical threshold the initial energy equals the saddle energy and the slope vanishes at that saddle.

$$u(s) = u(0) \implies b = s(1 - s),$$

$$u'(s) = 0 \implies b = 2s(1 - s)^2.$$

The two equations meet at $s = 1/2$, $b = 1/4$. The quasistatic fold gives $b = 8/27$. Consequently the undamped step threshold is only $\sqrt{27/32} \simeq 0.919$ times the slow-ramp pull-in voltage. A device operated between the two voltages can survive a ramp but collapse after an abrupt step. Dissipation and the actual rise time change this prediction.

Reconstruct the prepared state from a frequency measurement

Take a 100 µm square electrode in vacuum, $g = 2.00$ µm, $k = 2.00$ N/m and $m = 5.00 \times 10^{-10}$ kg. The predicted pull-in voltage is 7.32 V, the zero-bias capacitance 44.3 fF and the zero-bias natural frequency 10.07 kHz. On the low branch $s_0 = 0.20$ gives $x_0 = 0.400$ µm and voltage 6.80 V. Then $k_{\text{eff}} = 1.00$ N/m and the small-signal frequency is 7.12 kHz. A measured frequency ratio f/f_0 can be inverted via $s = (1 - r)/(3 - r)$, where $r = (f/f_0)^2$. The result is independent of absolute optical displacement calibration if the mass and underlying spring stiffness have not changed during the bias sweep.

At the same voltage the other physical root is $s = 0.48769$. A numerical force balance at this root would give the same supply reading but a negative tangent stiffness. The ideal voltage-controlled potential in

units kg^2 is $u(s) = s^2/2 - \Lambda/(1-s)$. The barrier height $kg^2[u(0.48769) - u(0.20)]$ for this example is 7.26×10^{-14} J. A sufficiently energetic disturbance can carry the plate out of the stable basin even below static pull-in. This is different from loss of the static equilibrium itself. A quasistatic model predicts the existence and local stiffness of the equilibrium; a transient simulation must additionally specify initial displacement, velocity, drive ramp and damping.

Change the preparation to fixed charge

Disconnect a perfect isolated capacitor after placing a known charge Q on it, neglecting leakage and parasitic capacitance. Then $F_Q = Q^2/(2\epsilon A)$ is independent of the remaining gap, because $D = Q/A$ is fixed. Static equilibrium is $x_0 = Q^2/(2\epsilon A k)$, with incremental stiffness exactly k. At the same initial voltage and gap the instantaneous force agrees with the voltage-controlled value, since $Q = C(x_0)V$. The derivatives governing subsequent motion differ. At $x_0 = 0.400$ µm the required charge in the numerical example is 0.376 pC. After ideal electrical isolation the small-signal frequency returns to the underlying elastic value even though the loaded static position stays the same.

Ideal charge isolation is delicate in a small device. A fixed parasitic capacitance C_p from the isolated node to a reference means the conserved charge belongs to $C(x) + C_p$. Motion changes the fraction of total charge residing on the actuator. The electrical energy becomes $Q_{\text{total}}^2/[2(C + C_p)]$, so the force is $Q_{\text{total}}^2 C'/[2(C + C_p)^2]$, generally displacement dependent. If $C_p \gg C$, the node voltage changes little with a small displacement and behavior is closer to voltage control. Interpreting an observed softening after disconnecting the supply as evidence against electromagnetic force balance would overlook this boundary change.

A different practical preparation puts an intentional series capacitor C_s between a fixed voltage supply and the actuator. Its series combination is $C_{\text{ser}} = \epsilon A/[g - x + \epsilon A/C_s]$. For static mechanical work at fixed supply voltage use $F = V_s^2 C'_{\text{ser}}/2$. Defining an effective electrical spacing $g_e = g + \epsilon A/C_s$, the static fold occurs at $x = g_e/3$ if this lies within the open physical gap. If $\epsilon A/C_s \geq 2g$, contact would occur before any such static fold in the ideal model. This calculation shows a direct connection between passive circuit design and allowed mechanical travel, while making clear that the maximum actuator voltage can differ from the voltage printed on the supply.

Why a sudden voltage change can reach contact below static pull-in

For a rapid electrical step the plate has a position and velocity inherited from the previous preparation. Suppose it starts at rest at its zero-bias position, the supply steps to a constant V much faster than a mechanical period, and mechanical damping is initially negligible. Immediately after the step, energy in the mechanical potential $u(s) = s^2/2 - \Lambda/(1-s)$ is $u(0) = -\Lambda$ in units of kg^2. Motion under this constant voltage can reach the force barrier if the initial value exceeds the potential at the unstable stationary point. This question can be solved by comparing two energies without integrating an entire time history.

Let s_b be the unstable root of $\Lambda = s_b(1 - s_b)^2$. At threshold the initial state is on the separatrix, the trajectory separating oscillations in the well from excursions to the closing gap. Set $u(0) = u(s_b)$, subtract the two expressions and cancel s_b from the nontrivial solution. The two conditions become

$$\Lambda = \frac{s_b(1 - s_b)}{2},$$

$$\Lambda = s_b(1 - s_b)^2.$$

Thus $s_b = 1/2$, $\Lambda = 1/8$, and the ideal undamped step threshold is

$$V_{\mathrm{step}}/V_{\mathrm{pi}} = \sqrt{27/32} = 0.919.$$

In our example this is 6.72 V. A slow increase to 6.80 V leads to the observed stable 0.40 µm deflection, whereas a sudden ideal voltage step to the same value from rest at the origin contains enough energy to escape the well. Starting on the biased equilibrium before applying a small perturbation is a different experiment. The direction of the threshold shift makes physical sense: damping removes mechanical energy during the first excursion, and a slow ramp allows the equilibrium to move with the plate. The undamped value specifies a particular ideal initial-value experiment, rather than a voltage rating for an unspecified drive waveform.

One can check a simulated step by monitoring $m\dot{x}^2/2 + \mathcal{U}_V$ after the supply settles. With viscous damping its derivative is $-b\dot{x}^2$, so any sustained increase during a constant-voltage segment points to a numerical or modelling error. Before the step ends there is an explicit drive term $-CV\dot{V}$ in this energy balance. The mechanical potential already includes the source work. Adding $V\dot{Q}$ to the same balance a second time gives an artificial energy imbalance. A transient solver should stop at a physically specified minimum gap instead of marching across the singularity; a statement about contact time depends on what counts as contact.

A finite source resistance changes what a ringdown measures

A dc supply with series resistance R can maintain the static voltage accurately while failing to hold the actuator voltage fixed over a mechanical cycle. Let its steady node voltage be V, with equilibrium capacitance $C = C(x_0)$. A small sinusoidal displacement δx causes current $i\omega(C\,\delta v + VC'\delta x)$ into the actuator. The small-signal supply voltage is zero, so $\delta v = -Ri$. Eliminating current gives

$$\delta v = -\frac{VC'}{C}\,\frac{i\omega\tau}{1 + i\omega\tau}\delta x,$$

$$\tau = RC.$$

Linearise $F_e = V^2C'/2$ allowing both voltage and position to vary. The direct derivative is $K_{e0} = V^2C''/2$, while the change from node voltage is $VC'\,\delta v$. For an ideal parallel plate $C''/2 = (C')^2/C$, so the two terms combine to the simple complex relation $\delta F_e = K_{e0}\delta x/(1 + i\omega\tau)$. At low frequency the electrical force follows displacement with its fixed-voltage slope. At high frequency

the slope vanishes, as it must for an isolated ideal plate carrying fixed charge over one oscillation. At intermediate frequency a motion-induced current dissipates heat in the resistor.

Bring δF_e to the left side of the mechanical equation. Its real part subtracts $K_{e0}/[1+(\omega\tau)^2]$ from the mechanical stiffness; its imaginary part adds $K_{e0}\tau/[1+(\omega\tau)^2]$ to the viscous damping coefficient. This damping is positive for a passive resistor. At $s = 0.20$ the static electrical stiffness is 1.00 N/m, the operating capacitance is 55.3 fF and $R = 400$ MΩ gives

$$\tau = 22.1\,\mu\text{s}.$$

A 7.2 kHz interrogation has $\omega\tau \simeq 1$. Only about half the static electrical slope survives and the added damping is roughly $11\ \mu\text{N}\,\text{s}/\text{m}$. Treat these as frequency-dependent coefficients in the driven susceptibility. Substituting the low-frequency stiffness into a ringdown fit at this frequency would give the wrong displacement when the fitted frequency is inverted. Fitting the complex response at several frequencies avoids that ambiguity.

An experiment can separate the circuit effect from mechanical nonlinearity. Keep the same static bias and displacement, insert two known series resistances, then measure both in-phase and quadrature response to a small sinusoidal force. The displaced equilibrium should remain essentially unchanged at dc if leakage stays small. A stiffness dispersion and dissipative peak that shift in proportion to $1/R$ support an electrical cause. Then repeat at a smaller oscillation amplitude. Amplitude-dependent frequency shift that survives the change in resistance can arise from nonlinear spring or force curvature, which the first-order susceptibility leaves out. A simple change to the external circuit has turned the slope of an energy curve into a prediction for a measured phase lag.

Compare force intersections for voltage and charge control

The initial gap is 2.00 µm, the area is a 100 µm square and the mass is 0.50 µg. Use the voltage ratio to move the force curve through the fold. At 93% of pull-in the low-displacement fixed-voltage intersection has a positive restoring slope. The high-displacement intersection has a negative slope. The charge-controlled line is prepared at the selected voltage at the original open gap. It gives the same initial force but a different subsequent curve. The readout reports both open-gap roots even if the upper root lies beyond the plotted interval. At the baseline stiffness, static pull-in is 7.32 V.

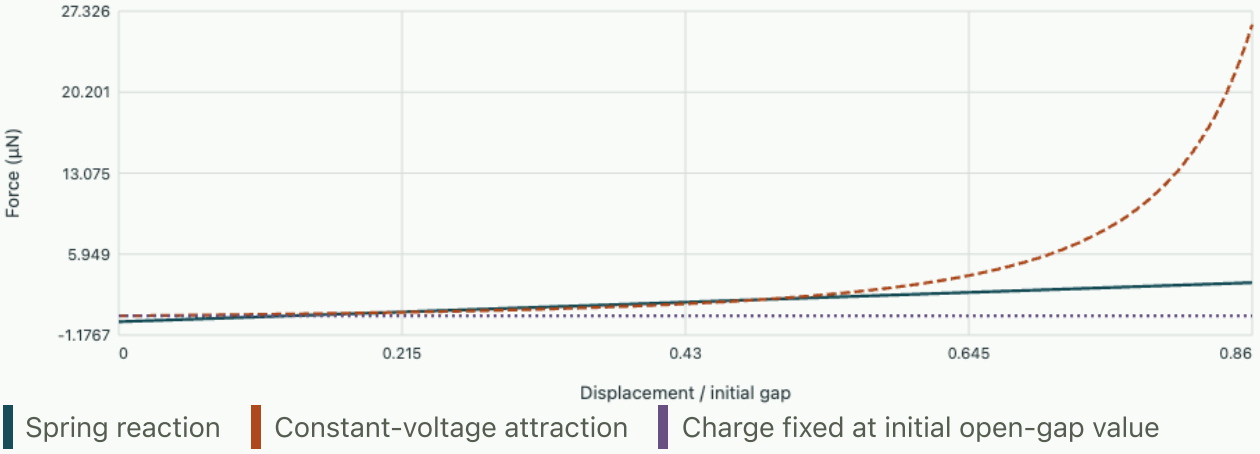
Change the spring stiffness while holding the voltage ratio fixed. The dimensionless locations of the roots stay in place, but their dimensional force and small-signal frequency change. This checks the scaling that eliminated the physical dimensions in the fold equation. Then set the ratio between the dynamic step threshold and unity. Use the force intersections to locate the potential maximum and sketch the area between attraction and spring reaction from the initial position. Its signed integral supplies the work which distinguishes a bounded excursion from crossing the maximum. The displayed curves predict initial-value behaviour only when combined with a choice of initial kinetic energy and damping.

Open this experiment online

Try the calculation

Bias as fraction of static pull-in **0.93**

Spring stiffness (N/m) **2**



Bias 6.805 V; static pull-in 7.317 V. Stable x/g = 0.200, unstable x/g = 0.487, small-signal frequency 7.11 kHz.

Tutorial · distinguish stability from force balance

Infer a gap fraction and a bias from a resonance

An otherwise identical actuator shows $f/f_0 = 0.80$ at a stable fixed voltage. Find the deflection as a fraction of gap, the voltage relative to pull-in, and the independent capacitance change. What could spoil this comparison?

▼ Work through the solution

Here $r = 0.64$ and $s = (1 - 0.64)/(3 - 0.64) = 0.1525$. The bias ratio is $\sqrt{(27/4)s(1-s)^2} = 0.860$ approximately. The parallel-plate capacitance should rise by $1/(1-s) - 1 = 18.0\%$. Measure the change relative to the active capacitance after removing any constant wiring capacitance. Disagreement can come from changing mode shape, a nonlinear spring, a tilted plate, a measurement frequency above the electrical equilibration bandwidth, or an incorrect extraction of mechanical stiffness from a strongly damped resonance. With a measured gap and mode shape these explanations yield different quantitative corrections.

Choose a series capacitor to move the static fold

The original 44.3 fF zero-bias actuator should remain on its stable slope as far as $x = 0.60g$, ignoring contact and other limits. Find the series capacitance whose fold falls exactly at this displacement. How much supply voltage appears across the actuator at that point?

▼ **Work through the solution**

Set $g_e/3 = 0.60g$, giving $\epsilon A/C_s = 0.80g$ and $C_s = C(0)/0.80 = 55.4$ fF. At $0.60g$ the actuator capacitance is $C(0)/0.40 = 110.7$ fF. Series charge is identical in both capacitors, so

$$V_{\text{act}}/V_{\text{supply}} = C_s/(C_s + C_{\text{act}}) = 1/3.$$

Choosing exactly this value places operation at vanishing ideal stiffness, so a strictly positive measured restoring slope at that displacement requires a smaller series capacitance or a different margin. The result is sensitive to additional capacitance connected between the series node and ground.

Further reading

- Derive capacitor pressure from field stress and work

Finite-size criticality and the evidence for a phase transition

Statistical mechanics from microscopic counting to material response

A finite magnet in zero field can reverse between two equally magnetised directions. Its exact mean magnetisation is then zero even below the temperature where a bulk material orders. Learn how to extract order, correlation length and critical exponents from finite measurements, and which apparent crossings or peaks would survive the thermodynamic limit.

What a finite magnet can and cannot spontaneously break

Consider a ferromagnetic Ising model with spins $s_i = \pm 1$, $N = L^d$ sites, short-range coupling $J > 0$, and uniform field h in energy units. Its Hamiltonian is $-J \sum_{\langle ij \rangle} s_i s_j - h \sum_i s_i$. A finite partition sum contains finitely many positive analytic terms at positive temperature. For finite N it cannot develop a true nonanalyticity on the real positive-temperature axis. Yet derivatives can become large and measured distributions strongly bimodal as size grows.

Define total magnetisation $M = \sum_i s_i$ and intensive order $m = M/N$. At zero field, spin inversion pairs configurations of M and $-M$, making the exact finite equilibrium mean zero. To define spontaneous order in the positive direction, take the thermodynamic limit at positive h, then let $h \to 0^+$. Reversing that order gives zero. A finite record that stays on one side because reversals are rare has not performed the full finite-system equilibrium average. It can approximate a symmetry-selected state only with that interpretation stated.

The absolute magnetisation or its root-mean-square value remains positive in finite zero-field calculations. Even independent disordered spins produce $\langle |m| \rangle$ of order $N^{-1/2}$. Therefore a nonzero magnitude in one simulation cannot establish bulk order. Repeat over sizes and inspect its scaling. A weak applied field, surfaces or slow field drift can also select one sign in a laboratory, modifying the finite-size comparison.

Derive a response from connected fluctuations

Differentiate the finite partition function at fixed temperature. The susceptibility per site to energy field h is

$$\chi_L = \frac{\partial \langle m \rangle}{\partial h} = \frac{\beta}{N}(\langle M^2 \rangle - \langle M \rangle^2) = \beta \sum_{\mathbf{r}} G_c(\mathbf{r}),$$

where the last equality assumes a translationally invariant system and $G_c(\mathbf{r}) = \langle s_0 s_{\mathbf{r}} \rangle - \langle s_0 \rangle \langle s_{\mathbf{r}} \rangle$. Long correlations amplify the response because many sites contribute together when one responds to a field. For an experiment using magnetic induction instead of the energy field, include magnetic moments and unit-system factors.

In a symmetric finite zero-field sample below the ordering temperature, transitions between the two bulk-like peaks can give a large contribution proportional to N to the full fluctuation susceptibility. To estimate within-phase fluctuations one sometimes forms $\beta N(\langle m^2 \rangle - \langle |m| \rangle^2)$. This is a different observable and not exactly the derivative of the finite symmetric mean with respect to h. Report which

definition is plotted. Mixing those definitions in a size collapse can change both amplitudes and the apparent location of extrema.

A correlation length may be estimated from the large-distance connected correlation decay away from criticality, or from the ratio of small-wavevector structure factors. In lattice units, an often used periodic second-moment length is

$$\xi_L = \frac{1}{2\sin(\pi/L)}\sqrt{\frac{S(0)}{S(k_{\min})} - 1},$$

$$k_{\min} = 2\pi/L.$$

This relation uses the small-k expansion of the correlation structure factor and a lattice wavevector convention. Subtract means consistently, especially if estimating connected rather than symmetric correlations. If a small noisy $S(0)/S(k_{\min}) - 1$ is negative, replacing it by zero without describing the noise can bias a length. Larger samples or more suitable observables may be preferable.

Derive the finite-size variable from competing lengths

Write reduced temperature $t = (T - T_c)/T_c$. Suppose the bulk correlation length diverges as $\xi_\infty = \xi_0 |t|^{-\nu}$. A finite periodic sample cannot support correlations extending arbitrarily beyond L. The competition is governed by L/ξ_∞, equivalently $tL^{1/\nu}$ with its sign and a metric factor. If the bulk magnetisation, susceptibility and heat capacity have exponents β_m, γ, α, a standard finite-size ansatz below the appropriate upper critical dimension and without extra dangerous variables is

$$\langle |m| \rangle = L^{-\beta_m/\nu} F_m(tL^{1/\nu}),$$

$$\chi_L = L^{\gamma/\nu} F_\chi(tL^{1/\nu}) + \chi_{\mathrm{reg}}.$$

Here β_m is an order-parameter exponent, distinct from reciprocal thermal energy β. The functional forms match the bulk powers when the scaled argument becomes large at fixed nonzero t. A rounded finite-size peak of χ occurs at an approximately fixed scaled argument, so its temperature shift and width scale roughly as $L^{-1/\nu}$, allowing boundary and subleading corrections. Its height has an additional smooth background that can be significant for modest sizes.

For a two-dimensional short-range Ising universality class the known values $\nu = 1$, $\gamma = 7/4$, $\beta_m = 1/8$ give useful benchmarks. Doubling L at criticality should raise leading susceptibility by $2^{7/4} = 3.364$ and reduce order magnitude by only $2^{-1/8} = 0.917$. The very slow decrease of order is easy to mistake for nonzero bulk order at the critical temperature. The heat-capacity singularity in this example is logarithmic rather than a simple positive power; fitting a few peak heights to a power law can give a misleading small apparent exponent.

Compare a mean-field onset and susceptibility with the finite-size scaling argument

A dimensionless crossing and its systematic drift

A convenient shape diagnostic is $U_L = 1 - \langle m^4 \rangle / [3 \langle m^2 \rangle^2]$. A zero-mean Gaussian distribution has fourth moment three times variance squared and gives zero. Two infinitely narrow equal-weight peaks at $\pm m_0$ give $2/3$. At criticality, rescaling m by $L^{-\beta_m/\nu}$ cancels between numerator and denominator, so curves for different sizes can approach an approximately common crossing. The limiting value depends on geometry, boundary conditions and convention, despite the exponent universality.

Include a leading correction $U_L = U_* + a\,tL^{1/\nu} + bL^{-\omega} + \cdots$. Equating the expansions for L and sL shows that the pair crossing drifts as $t_\times \propto L^{-(1/\nu+\omega)}$ when the coefficients are nonzero. An exactly common crossing is therefore not guaranteed in finite data. Sample temperature resolution, uncertainty of higher moments and correlations between neighbouring reweighted temperatures can also shift the visual crossing.

For a data set with sizes 16, 32 and 64, suppose adjacent pair crossings occur at reduced-temperature-like coordinates 2.31 and 2.28 when plotted as $k_B T/J$. Without a calibrated correction exponent, taking their simple mean gives no justified bulk critical temperature. If a model independently predicts a crossing displacement decreasing by four per doubling, solve $T_{16,32} = T_c + A$, $T_{32,64} = T_c + A/4$, giving $T_c = 2.27 J/k_B$. Vary the correction assumption and use an additional larger size to test it. Extrapolated digits considerably beyond the scatter of these alternatives are not warranted.

Separate thermodynamic rounding from sampling error

Near criticality long spatial correlations often coincide with long relaxation times. A sequential Monte Carlo update can require many sweeps to produce independent configurations. If an observable has normalised autocorrelation $\rho(k)$, its effective independent sample count for a long record is approximately $N_{\mathrm{samp}} / [1 + 2 \sum_{k>0} \rho(k)]$. Blocking time records and recomputing Binder ratios from blocks is more reliable than treating individual sweeps as independent.

If a susceptibility peak is represented by ten nearby temperatures generated from the same energy histogram, their errors are correlated. Temperature reweighting uses factors $e^{-(\beta' - \beta)E}$ and is dependable only where sampled energy distributions overlap. A smooth reweighted curve extending outside sampled energies can look precise while depending on a few rare configurations. Use independent anchor simulations or directly measured temperatures to verify both tails.

Inspect distributions as well as means. A first-order transition can also show a large susceptibility peak and bimodality, with leading peak scaling proportional to volume in many finite systems. A continuous transition's order distribution may itself be broad or even possess two lobes. Distinguish using latent heat or its size dependence, interface free-energy barriers, scaling of peak width and appropriate thermal observables. One finite-size double peak in magnetisation is not a universal transition-order test.

Check how a long but finite correlation length can arise without a positive-temperature transition in one dimension

An exact finite chain tests what a rapidly growing length means

A one-dimensional nearest-neighbour Ising chain in zero field has no positive-temperature bulk ordering transition, but at low temperature its correlation length grows extremely rapidly. This makes it a valuable counterexample when interpreting a finite magnet. The transfer matrix has entries $e^{Kss'}$ with $K = J/(k_B T)$. Its two eigenvalues are

$$\lambda_+ = 2\cosh K,$$
$$\lambda_- = 2\sinh K.$$

For a periodic chain of L spins,

$$Z_L = \lambda_+^L + \lambda_-^L.$$

Keeping only the larger eigenvalue is permissible in the thermodynamic free-energy density at fixed positive temperature but need not accurately represent a short cold chain.

In the infinite chain the zero-field correlation is $\langle s_0 s_r \rangle = (\tanh K)^r$, yielding $\xi = -1/\log(\tanh K)$ in lattice spacings. At low temperature $\xi \simeq e^{2K}/2$. A chain of length 100 can therefore be strongly correlated across its entire length around $K \simeq 2.65$, while the infinite chain still has zero spontaneous order at every finite K. Setting $\xi \sim L$ gives an apparent ordering scale $k_B T_L / J \simeq 2/\log(2L)$, which moves slowly toward zero as the sample is enlarged.

For a periodic finite chain there are two paths connecting site zero and site r. The exact correlation is $[q^r + q^{L-r}]/[1 + q^L]$, $q = \tanh K$. Fitting it to a single exponential out to separations near half the box would overestimate the length because the second periodic image contributes. Use the periodic expression explicitly or restrict fits to a window that is both larger than microscopic distances and smaller than half the sample. A high-quality apparent fit over a narrow middle interval can still mask the wrong boundary geometry.

Numerically simulate chains of lengths 32, 64 and 128 at a sequence of K. Compare the measured absolute magnetisation, full connected correlation and transfer-matrix predictions. If a simulation remains stuck in one sign sector its short-distance correlation can agree with theory while the time average of signed magnetisation does not. Increasing the number of Monte Carlo sweeps and increasing the physical chain length are separate limits. Following only one of them cannot establish a nonzero thermodynamic order parameter.

A response derivative offers an exponent without partition-function normalisation

Let U_L be a dimensionless shape ratio such as the Binder cumulant. If $U_L = \mathcal{U}(tL^{1/\nu}) + \cdots$, its slope with respect to reduced temperature at the critical point scales as $L^{1/\nu}$. A logarithmic slope ratio between two sizes can estimate the correlation-length exponent independently of the magnetisation or susceptibility exponents. Corrections to scaling still matter, especially if the sampled point is not near a fixed critical argument for both sizes.

An equilibrium derivative can also be computed without a finite temperature difference. For any state observable A with no explicit temperature dependence, differentiating its canonical mean gives

$$\partial_\beta \langle A \rangle = -\mathrm{cov}(A, E).$$

Apply this to the second and fourth moments entering a Binder ratio, then use the chain rule. The resulting slope uses energy–magnetisation correlations from one run. Taking two numerically close independent temperatures instead can amplify noise by division by their small separation. A covariance estimator avoids that particular subtraction but requires accurate sampling of joint tails.

Suppose measured absolute crossing slopes on sizes 24, 48 and 96 are 5.7, 10.9 and 21.5 per unit reduced temperature. Adjacent estimates are $1/\nu = \log(10.9/5.7)/\log 2 = 0.935$ and 0.980. They approach the expected value one in this benchmark. A single fit through all sizes would give a smaller formal statistical uncertainty but might obscure a systematic drift. Report adjacent effective exponents and test sensitivity to omitting the smallest size.

The numerical derivative must use the same temperature variable for each size. A slope against inverse temperature differs by $d\beta/dT$, a known factor near a common critical temperature; plotting against independently fitted centre temperatures can introduce an additional calibration dependence. A nonlinear temperature sensor or a slowly drifting coupling also changes the physical conversion from laboratory temperature to the simulated dimensionless control. Use an independently determined parameter scale if the goal is to compare experiment with an absolute transition temperature rather than just a scaling exponent.

Estimate an interface cost from a probability trough

At a first-order transition, a finite periodic sample may switch between two phases separated by an interfacial configuration. A histogram of an appropriate extensive observable can show peaks corresponding to nearly pure phases and a trough representing intermediate states. At equal bulk phase weight, a probability ratio gives a free-energy difference $\Delta F_L = -k_B T \log(P_{\mathrm{trough}}/P_{\mathrm{peak}})$, provided the same bin measure or a properly defined density is used. An exponentially small trough amplifies sampling errors, so the measured minimum is often only an upper bound on its probability.

In a slab geometry under periodic boundaries, an intermediate configuration may contain two approximately flat interfaces. If each interface costs γL^{d-1}, the leading barrier is $2\gamma L^{d-1}$, with subleading capillary fluctuations, geometry corrections and possible droplet-to-slab crossovers. A cube with open boundaries can instead permit interfaces of different number or shape. Counting two interfaces automatically for every order-parameter histogram would misestimate the inferred tension.

For an illustrative two-dimensional periodic square, suppose the measured barrier grows from $8k_BT$ at $L = 40a$ to $15k_BT$ at $L = 80a$, where a is lattice spacing. From the difference of the two barriers and the increase in total interface length $2(40a)$, an effective line tension is $0.0875\,k_BT/a$. Using the absolute smaller-size barrier divided by $2L$ gives $0.100\,k_BT/a$, revealing a finite-size offset. More sizes, an observed slab geometry and adequate trough sampling are needed before assigning a bulk interface tension.

This analysis should not be transplanted unchanged to a continuous critical distribution with two shoulders in finite volume. A critical finite-size order distribution can be non-Gaussian without an interfacial barrier growing as L^{d-1} at its critical point. Plot the barrier's size dependence and inspect actual spatial configurations. A peak-height ratio alone contains no direct information about how many interfaces or what type of correlations generated the intermediate states.

Estimate a ratio of moments without losing its covariance

The second and fourth magnetisation moments in a Binder cumulant usually come from the same record. Treating their errors as independent can grossly misestimate the uncertainty because the fourth moment tends to be large precisely in configurations contributing a large squared moment. Divide the record into blocks longer than its integrated autocorrelation time, recompute the ratio after deleting each block and estimate uncertainty using a block jackknife or use independent replicate simulations. Recomputing the nonlinear statistic inside each resample retains numerator–denominator covariance.

Histogram reweighting across nearby temperatures should likewise be recomputed within each block when extracting a crossing. Fitting smooth curves to the full sample and then adding independent random errors to individual temperature points ignores strong correlations induced by the shared configuration weights. Repeat the root-finding step for each independent or block-resampled pair of systems to obtain an uncertainty on the crossing temperature itself.

Check block-size convergence. If estimated errors continue growing as block duration is increased, slow modes may not be resolved or the record may have drifted. A huge number of recorded spin configurations cannot compensate for fewer than a handful of independent reversals when the statistic depends on those reversals. This issue is measurable: compare direct covariance estimates, block estimates and independent starts, and extend the run where those diagnostics disagree.

A block-average variance exposes a correlation length without fitting a tail

Divide a large zero-field spin image into square blocks of side b and record the average spin in each block. Its variance is the sum of all connected correlations within the block divided by the squared number of sites. For independent spins of variance one, it is exactly $1/b^2$ in two dimensions. If neighbouring spins are correlated over a length ξ much smaller than the block, the variance still falls approximately with inverse block area, but with a larger prefactor proportional to the integrated correlation. Below block sizes comparable to ξ, this averaging becomes much less effective.

At a critical point the connected correlation decays algebraically rather than exponentially over the accessible range. If in dimension d it behaves as $G(r) \propto r^{-(d-2+\eta)}$, integrating pair separations to the block scale gives $\operatorname{var}(m_b) \propto b^{-(d-2+\eta)}$ in the scaling regime. For the two-dimensional short-range Ising benchmark with $\eta = 1/4$, doubling the block side reduces the variance by only $2^{-1/4} = 0.841$, compared with 0.25 for independent spins. This is an observable distinction even when a full bulk thermodynamic limit is unavailable.

Block analysis of one image does not generate independent thermodynamic systems. Neighbouring blocks share correlated surroundings, overlapping blocks are statistically dependent, and every block inherits global constraints or drift in the original preparation. Increasing block size also reduces the available count sharply. Repeat across independent configurations, compare several whole-image sizes and study block-size windows well above a lattice spacing and well below the full image. An apparently accurate critical exponent from the final two largest blocks of one image is particularly unreliable.

A real magnetic image may measure a blurred field projection rather than the local spin itself. Spatial point spread creates detector-induced correlations at small separations. Determine it with a reference pattern and omit or model blocks comparable to the blur. Conversely spatially varying illumination adds long-range covariance unrelated to the sample. Flat-field calibration and sample rotations can separate these artefacts from material correlations. A second-moment structure-factor estimate and a block-variance estimate respond differently to these measurement effects, providing a useful cross-check.

This method also links microscopic and macroscopic noise. The scaled block variance in the large-block regime approaches the same integrated connected correlation controlling uniform susceptibility, with boundary terms that shrink as block dimensions exceed ξ. Comparing a weak-field response with block statistics therefore tests equilibrium fluctuation–response and spatial calibration simultaneously. If the discrepancy grows only as blocks approach the full system size, a fixed total magnetisation constraint or incomplete sampling of global sign reversals deserves examination.

For a concrete noise budget, consider a camera averaging adjacent pairs of independent sites before reporting them as displayed pixels. Those displayed pixels can have a positive nearest-neighbour correlation if their averaging footprints overlap, even when the original spins are completely uncorrelated. A measurement of the apparent short-range correlation length would then reflect the instrument kernel. Blocking on sizes much larger than the overlap can recover the correct large-scale inverse-area trend, while direct comparison of single-pixel variance with independent-spin theory fails. Calibrating only the detector's global gain cannot fix this spatial response. A measured point-spread map, an uncorrelated reference preparation and tests of blocks at several image resolutions can separate the resolution scale from the material's growing correlation length as temperature is varied.

Compare a correlation length with a finite specimen

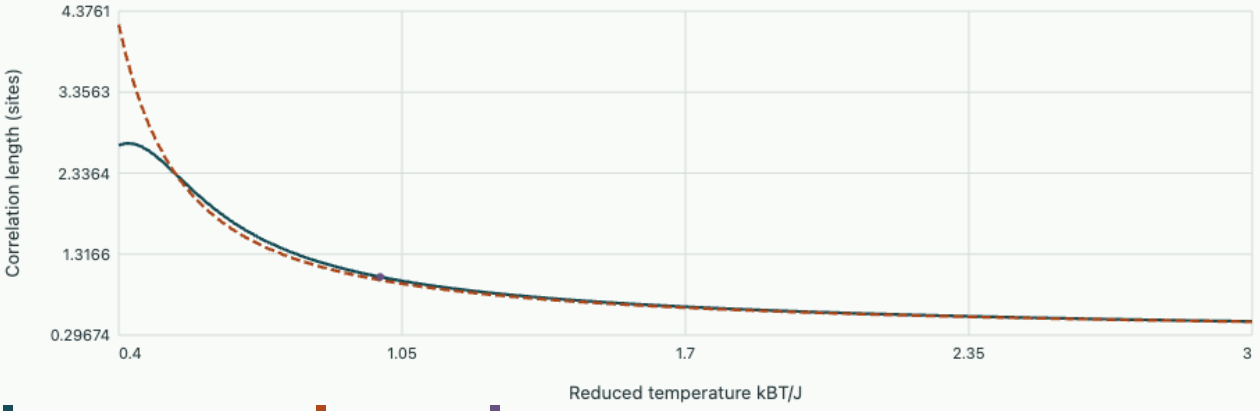
Determine how the subleading transfer-matrix eigenvalue controls the length seen in a colour measurement. A finite chain much shorter than this length can look ordered.

Open this experiment online

Try the calculation

Reduced temperature kBT/J **1**

Reduced colour field H/J **0.15**



Selected colour bias Zero bias Selected temperature

Leading eigenvalue 4.9881, second 1.8884, longest length 1.0295 sites.

Download the OneSimulation model

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

```
python spectra-and-orbits.py chain --temperature 1 --bias 0.15
```

Tutorial · Extract two exponent ratios from a size doubling

At a candidate critical temperature susceptibilities for sizes 32 and 64 are 110 and 360 in common units. Absolute magnetisations are 0.310 and 0.282. Estimate the corresponding exponent ratios and compare with the two-dimensional Ising benchmarks.

▼ **Use logarithmic ratios and inspect the small differences**

The susceptibility ratio gives

$$\gamma/\nu = \log(360/110)/\log 2 = 1.71.$$

For magnetisation $\beta_m/\nu = -\log(0.282/0.310)/\log 2 = 0.137$. These are near 1.75 and 0.125, but no error bars or corrections have been supplied. An additive susceptibility background or a slight critical-temperature offset can bias both effective exponents. Use several sizes and a dimensionless crossing diagnostic before assigning a universality class.

Tutorial · Determine whether a finite sample really switched phases

A zero-field record contains 20,000 sweeps of positive magnetisation near $+0.6$, followed by one reversal and 20,000 sweeps near -0.6. It yields a near-zero mean and large variance. Why does the nominal record length not establish a precise symmetric susceptibility? What extra records or controls would help?

▼ **Identify the slow event determining the uncertainty**

Within each sign sector there may be many locally decorrelated fluctuations, yet the relative weight of the two peaks is estimated from essentially one reversal. Treating 40,000 sweeps as independent underestimates the error of the peak weights and of statistics sensitive to global sign. Starting multiple independent runs with both signs can improve coverage, but deliberately balancing starting signs is not a measurement of true equilibrium weights if a weak residual field exists.

Record switching-time distributions over much longer runs or use an equilibrium-preserving algorithm that efficiently samples both sectors, while distinguishing computational updates from physical kinetics. Apply small positive and negative fields to test susceptibility and residual bias. For a within-phase susceptibility condition on a clearly defined phase and report the changed observable.

Tutorial · Explain a crossing that drifts with the boundary

Two groups study the same lattice Hamiltonian. One uses square periodic samples; the other uses twice-as-long rectangular samples with open ends. Their Binder curves cross near similar temperatures but at different cumulant values. Is this evidence against universal critical behaviour?

▼ **Separate universal exponents from finite-geometry functions**

Changing aspect ratio and boundary conditions changes the finite-size scaling function and hence the crossing value of a dimensionless ratio. Universal bulk exponent comparisons do not require identical Binder values across different finite geometries. Match boundary conditions, shape, operator and cumulant convention before comparing purported universal amplitude ratios.

Each group can still extrapolate its own crossing temperatures, examine moment slopes proportional to $L^{1/\nu}$ and test bulk exponent ratios with geometry fixed across its own size series. If one dimension is held fixed while the other grows, however, the family approaches a strip rather than a bulk two-dimensional limit. The thermodynamic sequence itself is an experimental variable, not just a plotting preference.

Cambridge Statistical Physics · phase transitions, criticality and scaling

Phase nucleation, metastability and waiting-time statistics

Statistical mechanics from microscopic counting to material response

A cooled liquid can remain liquid below its equilibrium freezing temperature and then transform abruptly. How can a bulk phase be thermodynamically favoured while a small embryo of it tends to shrink? Balance interfacial cost against bulk gain, convert a barrier into a probabilistic waiting-time prediction and test which measurements can distinguish surface nucleation from a bulk mechanism.

Choose a driving free-energy density at fixed controls

At specified temperature and pressure, phases in equilibrium have equal molar Gibbs free energies. Below a melting temperature T_m the stable solid has lower chemical potential than the liquid. For modest undercooling, use $d\mu = -s\,dT + v\,dp$. At fixed pressure the chemical-potential difference per molecule is approximately $\mu_l - \mu_s \simeq \Delta h_f (T_m - T)/T_m$ if latent enthalpy of fusion per molecule varies little. Multiplying by the number density of the new phase gives a positive driving free-energy density Δg. State whether each quoted latent heat is per particle, per mole or per unit volume; converting it twice gives enormous radius errors.

The underlying free-energy landscape has stable and metastable minima separated by a barrier. In a phase-field picture a local bulk density penalises intermediate order within the metastable region, and a gradient penalty charges a finite energy for an interface. A finite amount of stable phase can therefore cost free energy before its bulk advantage wins. A coexistence construction determines which bulk phases minimise free energy; it does not provide the rate at which an existing metastable preparation can reach them.

Compare a bulk coexistence construction with the cost of an interface

Derive the critical radius and barrier

In the simplest capillarity model an embryo is a sphere of radius r, with isotropic positive interface energy γ per area and a bulk gain Δg per volume. Treat its interface as thin and its bulk properties as those of the macroscopic phases. Neglect elastic strain, long-range fields and depletion of the surrounding reservoir, then the reversible work to create it is

$$\Delta G(r) = 4\pi r^2 \gamma - \frac{4\pi}{3} r^3 \Delta g,$$

$$r_* = \frac{2\gamma}{\Delta g},$$

$$\Delta G_* = \frac{16\pi\gamma^3}{3\Delta g^2}.$$

The derivative vanishes at r_* and the second derivative there is $-8\pi\gamma$, so this is a maximum along the size coordinate. A slightly smaller embryo tends to shrink and a slightly larger one tends to grow, subject to noise and kinetic constraints. The radius where the total formation free energy first drops below its initial zero is

$$3\gamma/\Delta g = 1.5 r_*.$$

A supercritical nucleus can grow downhill even while its cumulative creation free energy is still positive. Confusing these two radii yields a fifty percent size error.

Take illustrative values $\gamma = 0.010$ J per square metre and $\Delta g = 1.0 \times 10^{7}$ J per cubic metre. Then $r_* = 2.0$ nm and $\Delta G_* = 1.68 \times 10^{-19}$ J, about 40.5 thermal energies at 300 K. Doubling driving density halves critical radius and quarters the barrier if γ is unchanged. At a two-nanometre radius the embryo may be just a few molecular spacings across. This calculated barrier is a sensitivity estimate; the assumed bulk interface tension, sharp interface and spherical shape become increasingly questionable there.

An alternative derivation uses the pressure jump $2\gamma/r$ across a curved interface together with a bulk chemical-potential difference. Near coexistence the condition of zero growth recovers the same critical curvature. In a vapour close to ideal behaviour the driving chemical-potential difference involves $k_B T \log S$ with supersaturation $S = p/p_{\mathrm{sat}}$, not simply $S - 1$ at arbitrary supersaturation. Near $S = 1$ the two are close. A quantitative cloud-chamber inference must also account for the evolving supersaturation and evaporation or growth after nucleation.

Background · Why does a faster temperature ramp move the apparent freezing point?

Assume independent homogeneous nucleation with rate density $J(T)$ in volume V, and prompt detection of the first nucleus. The chance that the untransformed sample survives the next interval dt is $1 - V J(T)\, dt$. Multiply these conditional probabilities along a cooling ramp with constant positive speed $v = -dT/dt$. Starting at T_0 gives

$$S(T) = \exp\!\left[-\frac{V}{v}\int_{T}^{T_0} J(T')\, dT'\right].$$

At fixed undercooling the faster ramp has accumulated less nucleation probability. The median observed freezing temperature satisfies an integrated hazard of $\ln 2$. Doubling the sample volume and the ramp speed together leaves the predicted distribution unchanged. Failure of that scaling can signal surface-dominated nucleation, thermal lag or a growth time long enough to affect detection.

Derive the barrier reduction for a spherical cap

A nucleus forming on a flat inert wall can replace one existing interface with another. Let it be a spherical cap of curvature radius r and contact angle θ through the new phase. The liquid-facing area is $2\pi r^2(1 - \cos\theta)$, contact area $\pi r^2 \sin^2\theta$ and cap volume $\pi r^3(2 - 3\cos\theta + \cos^3\theta)/3$. Young's contact-angle condition relates the wall energy difference to minus $\gamma\cos\theta$ per replaced contact area. Combining all terms gives the homogeneous spherical result multiplied by

$$f(\theta) = \frac{(2+\cos\theta)(1-\cos\theta)^2}{4}.$$

Both surface and bulk coefficients receive the same factor. Critical curvature radius therefore remains $2\gamma/\Delta g$ in this simple model, while the barrier becomes $f(\theta)\Delta G_*$. At 180°, $f = 1$ as the full sphere barely touches the wall. At 90°, $f = 1/2$ for a hemisphere. At 60°, $f = 0.15625$, lowering the illustrative 40.5-thermal-energy barrier to 6.33 thermal energies. Even a relatively small number of active wall sites can dominate observed events.

Complete wetting, a chemically patterned surface, a very small defect or appreciable energy per contact-line length can invalidate the simple contact geometry. Contact angle measured on a macroscopic equilibrium droplet need not equal the local microscopic value for a tiny forming embryo. A fitted effective angle that varies systematically with undercooling can signal a missing mechanism rather than a temperature-dependent property of a clean ideal wall.

Distinguish an attempt rate from an equilibrium barrier

A classical rate expression is $J = J_0 e^{-\Delta G_*/(k_B T)}$ per unit volume per unit time for homogeneous events. The prefactor contains the supply of available configurations and the kinetics for traversing the critical region. A barrier alone is insufficient to predict a waiting time. Lowering temperature can raise thermodynamic driving yet also reduce mobility and attachment, giving a nonmonotone total rate as the two effects compete.

If a well-mixed sample of volume V has constant independent hazard JV, the survival probability of the metastable state follows from $P_0(t+dt) = P_0(t)(1 - JV dt)$: $P_0 = e^{-JVt}$. Mean wait is $1/(JV)$, median $\log 2/(JV)$. A 1.0 microlitre sample with $J = 2.0 \times 10^8$ per cubic metre per second has hazard 0.20 per second and median 3.47 seconds. After five seconds 63.2 percent have transformed at least once, if the detector registers nucleation immediately.

If surfaces contribute independent rate per area J_s, the combined first-event hazard is $J_v V + J_s A$. Geometrically similar droplets have $A \propto V^{2/3}$, so changing sample size can distinguish these contributions. Vary active surface chemistry at fixed geometry as a second probe. If most samples contain a single rare defect not scaling smoothly with area, survival can be a mixture of exponentials rather than one exponential. A straight-line fit to mean waits alone obscures such heterogeneity.

A cooling ramp integrates many different hazards

Under a deterministic temperature history $T(t)$, survival is $P_0(t) = \exp[-\int_0^t J(T(s))V ds]$ for homogeneous, conditionally independent events. In a constant cooling ramp $q = -dT/dt > 0$, the exponent becomes $-(V/q)\int_T^{T_0} J(T')dT'$. The observed freezing-temperature histogram is weighted by survival as well as the instantaneous rate. The most probable freezing temperature need not minimise any free energy, and different ramp speeds can shift it even when all material properties are unchanged.

For a ramp calculation on gridded temperatures, update the remaining fraction by $e^{-\Delta H}$ where ΔH is the integrated hazard in that interval. Using a transition probability $JV\Delta t$ even when it exceeds one violates probability bounds. Subdivide the interval and check convergence of median and distribution width. At steep temperature dependence, a midpoint estimate over a coarse interval can move a reconstructed barrier enough to dominate the apparent contact-angle uncertainty.

The detected event may be a threshold in light scattering, latent heat or visual crystal size. A nucleus must grow before producing that signal. At fixed growth speed u_g, reaching detection radius r_d adds an approximate lag $(r_d - r_*)/u_g$; under cooling, both growth and hazard vary. Compare multiple detection thresholds or resolve the early growth trajectory and extrapolate a starting time. A shift in waiting time due entirely to the imaging threshold could otherwise be misinterpreted as altered microscopic nucleation.

Test how changing undercooling alters a barrier in the course's phase-equilibrium example

A finite reservoir can stabilise a droplet that would otherwise grow

The standard critical-radius derivation supposes that the driving free-energy density remains fixed while the nucleus grows. In a small sealed droplet or isolated simulation box, formation of the new phase can consume a significant fraction of the supersaturation. The surrounding phase then approaches coexistence and the bulk driving force falls. A single critical radius computed from initial supersaturation may describe early stability but cannot describe the entire late-time growth or the final phase fraction.

As a transparent model let conversion of new-phase volume v reduce the driving density linearly to $\Delta g(v) = \Delta g_0 - a v$, with $a > 0$ incorporating finite-reservoir compressibility or depletion. The bulk free-energy gain is the integral of the driving density, not its final value times volume. For a spherical interface $A(v) = (36\pi)^{1/3} v^{2/3}$, the total model is

$$\Delta G(v) = \gamma (36\pi)^{1/3} v^{2/3} - \Delta g_0 v + \frac{a}{2} v^2 .$$

The initial derivative is positive because a very small sphere has large area per volume. At large v the positive depletion derivative can make further growth unfavourable. If the derivative turns negative at intermediate volume there can be both a barrier top and a larger stable or metastable finite-droplet minimum. If it never turns negative the finite reservoir is too small or too weakly driven for a stable droplet in this model. This explains why a finite system can show a rare nucleus followed by arrested growth without contradicting the initial local barrier idea.

For a dimensionless illustration choose $\Delta G/(k_B T) = 4 v^{2/3} - 4 v + 0.25 v^2$, with a specified arbitrary reference volume absorbed into v. Its derivative is $(8/3) v^{-1/3} - 4 + 0.5 v$. At $v = 0.10$ it is about 1.80, at $v = 1$ about minus 0.83 and at $v = 10$ about 2.24. At least two stationary sizes therefore lie along the growth path. A sign test on neighbouring sizes distinguishes the early maximum from the later minimum without assuming every stationary root is a critical nucleus. Numerical root finding should bracket these branches separately instead of passing a single initial guess to a solver.

An experiment can test depletion by varying reservoir size at the same initial supersaturation and surface treatment. If early appearance times are comparable after correcting for trial volume but final crystal sizes grow with available material, finite feeding is plausible. Limited thermal transport, impurities blocking the interface or elastic stress can also arrest growth, so monitor the remaining parent-phase composition and temperature. A small system is precisely where the thermodynamic ensemble and the available amount of material need explicit definition.

A detection time can be dominated by diffusion after nucleation

After crossing the barrier a supercritical spherical domain may grow by diffusion of a dilute species. In a quasi-steady infinite exterior with surface concentration c_s and far concentration $c_\infty > c_s$, the concentration solves the radial Laplace equation. Its solution is

$$c(r) = c_\infty - (c_\infty - c_s) R / r.$$

The total inward solute flux is $4\pi D R (c_\infty - c_s)$. Equate this to solute incorporated into an advancing sphere, $4\pi R^2 (c_{\mathrm{new}} - c_s)\dot{R}$, to obtain

$$R^2(t) - R^2(0) = 2D \frac{c_\infty - c_s}{c_{\mathrm{new}} - c_s} t.$$

This square-root growth law assumes quasi-steady diffusion, a spherical shape, nearly constant far conditions and a surface near local equilibrium. If interface attachment instead limits growth, a roughly constant $\dot{R}$ may be a better first model over a short range. Measuring time to several optical thresholds distinguishes R_d^2 scaling from R_d scaling. Threshold variation therefore separates growth and nucleation more effectively than one image definition held fixed.

Take $D = 1.0 \times 10^{-9}$ square metres per second and concentration ratio 0.020. From a negligibly small initial radius, reaching 5.0 micrometres takes $25 \times 10^{-12} / (4.0 \times 10^{-11}) = 0.625$ s; reaching ten micrometres takes 2.50 s. If true median nucleation time is 1.0 s, the detected median at the larger threshold can be 3.5 s even when the microscopic barrier and initial supersaturation are unchanged. Calling its inverse a nucleation rate would seriously underestimate the true first-event hazard.

At very small radius a Gibbs–Thomson curvature shift increases the equilibrium surface concentration. The same curvature effect that opposes a small embryo can therefore slow its initial excess growth even after it has crossed the most probable barrier. At late times the bath may not be infinite and growing domains compete for material. These effects change the simple square-root law in different regions. Image a radius trajectory for each event instead of assigning the same unverified growth lag to all samples.

A thermal ramp with a tractable hazard shape

Over a restricted undercooling interval an experimentally measured hazard can rise approximately exponentially with undercooling $\Delta T = T_0 - T$: take $h(\Delta T) = h_0 e^{a \Delta T}$. This is a local fit to the full

barrier and prefactor, not a universal replacement for their pronounced temperature dependence. At constant cooling rate $q > 0$, direct integration gives cumulative hazard $H = h_0(e^{a\Delta T} - 1)/(aq)$ and survival e^{-H}. The median solves

$$\Delta T_{1/2} = \frac{1}{a}\log\left[1 + \frac{aq\log 2}{h_0}\right].$$

For $h_0 = 0.010$ per second, $a = 1.0$ per kelvin and $q = 0.10$ K/s the median undercooling is 2.07 K. Doubling cooling rate raises it to 2.70 K. The shift is smaller than one kelvin but large relative to many thermometer precisions. It arises even with exactly the same thermodynamic coexistence temperature and material interfaces. Counting a median freezing temperature at only one ramp rate can therefore mix kinetics with the equilibrium phase boundary.

The probability density per unit undercooling is $[h(\Delta T)/q]e^{-H}$, where the first factor measures the rate conditional on surviving to that temperature. Taking a log derivative shows that an interior density maximum occurs when $h = aq$, provided this exceeds the initial hazard. The median and the most probable temperature are distinct. At slower ramps the maximum can even lie at the beginning of the observation range, depending on preparation and chosen T_0. A histogram peak should be interpreted together with the surviving population, not treated as a direct plot of microscopic rate.

To reconstruct hazard from a measured ramp, estimate the fraction still untransformed at each temperature and inspect minus the log of that fraction. Differentiate smooth cumulative hazards cautiously; differentiation amplifies noise and small late survivor counts. A joint fit of raw first-event and censored records over several ramp rates generally preserves more information than differentiating a single binned histogram. Varying sample volume as well helps separate a distribution of bulk rates from one associated with surfaces.

A committor test examines whether radius is a good reaction coordinate

The capillarity free-energy maximum uses one coordinate, radius. A practical microscopic test begins from candidate configurations near the proposed critical size, randomises thermal momenta or bath noise consistently with the preparation and launches many short independent trajectories. For each starting configuration estimate the fraction reaching a specified stable new-phase basin before returning to the parent phase. This fraction is the forward committor. A good transition-surface coordinate should identify configurations with committor near one half under the specified dynamics.

If configurations at the same radius show committors ranging from almost zero to almost one, that radius alone poorly locates the dynamical transition state. Shape, local order, strain or neighbour concentration may supply missing coordinates. Binomial uncertainty from finite trials also matters: 25 independent launches near one half give a standard error around 0.10 for an individual configuration. Increasing launches distinguishes true variation between configurations from sampling noise.

The committor depends on physically chosen basins and dynamics. An artificially accelerated Monte Carlo kinetic rule may preserve equilibrium free energies while changing which paths reach the new phase first. Use the same dynamics when comparing committor tests with measured waiting times.

Equilibrium barrier heights, kinetic attempt rates and the adequacy of a proposed reaction coordinate are connected questions but require different data.

Distinguish first appearance from transformed volume fraction

A first-event detector stops caring about a sample once the earliest nucleus is registered. A calorimeter or diffraction measurement instead often follows the total fraction of material transformed. Even if nucleation is homogeneous at a constant rate and crystals grow at constant speed, these two signals have different time dependences. Treat an infinite initially untransformed sample, with nuclei placed randomly at rate J per volume and spherical domains growing radially at speed u. To ask whether a particular point remains untransformed at time t, count possible earlier nucleation events whose growth sphere would have reached it.

A nucleus born at time s reaches points within radius $u(t-s)$, corresponding to volume $4\pi u^3(t-s)^3/3$. Integrate the attempt rate over that backwards influence volume from zero to t. For an ideal Poisson space-time nucleation model with this growth rule the probability of no such event is $\exp[-\pi J u^3 t^4/3]$, giving transformed fraction $X = 1 - \exp[-\pi J u^3 t^4/3]$. The result uses the standard extended-volume construction with random nucleation and isotropic constant-rate growth. It is not the simple first-event law $1 - e^{-JVt}$ for a finite observation sample.

For sites all pre-existing at time zero with random density n_0, the same backwards-volume reasoning gives

$$X = 1 - \exp[-(4\pi/3) n_0 u^3 t^3].$$

A log plot of minus the log remaining fraction against log time therefore has ideal slope four for continual three-dimensional nucleation and three for instantaneous site saturation. In two dimensions the powers would both be reduced by one. A measured slope between these values is not automatically evidence for a fractional geometric dimension. Changing growth law, active-site exhaustion, temperature ramps or overlapping different mechanisms can alter the apparent slope.

If measured isothermal transformed fraction is one half at 100 s in the continuous model, the product Ju^3 equals $3\log 2/(\pi\, 100^4) = 6.62 \times 10^{-9}$ per second to the fourth. This identifies the product, not the nucleation rate and growth speed separately. Independent imaging of isolated domain radii versus time, or a set of first-event measurements in small known volumes, can fix the remaining scale. A fit to the entire same ideal fraction curve cannot identify two quantities that enter it only through their product.

Actual calorimetric power is approximately latent enthalpy times dX/dt, convolved with thermal response. Its maximum occurs when transformation rate peaks rather than when the first nucleus forms or exactly when half the volume has transformed. Compare integrated calorimetric energy with the final transformed fraction and deconvolve the calibrated instrument response before interpreting any schedule of peaks as a nucleation-time distribution.

The influence-volume argument also predicts a useful geometry check in a thin film. If domains reach both film surfaces early and then grow as expanding disks in the plane, a three-dimensional spherical

growth model will eventually overestimate newly transformed volume. The apparent exponent can cross from an initially three-dimensional value to a lower late-time value without any change of microscopic nucleation pathway. Compare films of several thicknesses, identify the time at which individual radius reaches the film thickness and then test the predicted crossover in transformed fraction. Altering film thickness usually changes wall-to-volume ratio as well, so site density and growth geometry can vary together. Independent first-event measurements and direct images of individual domains remain necessary if the objective is to separate those two effects.

Locate a critical cluster and its barrier

Vary surface energy and bulk driving force independently. Find the stationary radius before reading the height of the barrier.

Open this experiment online

Try the calculation

Interfacial energy (mJ/m²) 35

Bulk driving density (MJ/m³) 25



Critical radius 2.8 nm; barrier 7.174 eV. The free-energy curve crosses zero again at 4.2 nm.

Tutorial · Extract a rate from records with censored observations

Twelve equal 2.0 microlitre samples are held at constant temperature for at most 20 s. Eight show first events at times whose sum is 64 s; four remain untransformed at the end. Estimate the homogeneous rate and explain the bias from averaging only the observed event times.

▼ **Count events and sum the full time at risk**

Likelihood for independent exponential waits with right censoring is proportional to $h^{N_e} e^{-hT_{\text{exposure}}}$. The exposure is $64 + 4(20) = 144$ s, so $\hat{h} = 8/144 = 0.0556$ per second. Dividing by 2.0×10^{-9} cubic metres gives $J = 2.78 \times 10^{7}$ per cubic metre per second. The observed-only mean of 8 s would instead imply 0.125 per second and bias the rate upward by discarding samples known to survive a long time. With eight events, statistical uncertainty is substantial despite the precise arithmetic.

Tutorial · Distinguish area from volume nucleation

Two cubic samples have side lengths L and $2L$. Their measured first-event hazards are 0.060 and 0.360 per second. Assume homogeneous bulk and wall contributions $b = J_v L^3$, $s = J_s 6L^2$ in the smaller cube. Infer their relative importance and predict the hazard for a cube of side $3L$.

▼ **Use different geometric powers to separate the contributions**

The measurements give $b + s = 0.060$ and $8b + 4s = 0.360$. Subtracting four times the first equation gives $4b = 0.120$, hence $b = s = 0.030$ per second. For side $3L$, hazard is $27b + 9s = 1.08$ per second. This prediction requires the same temperature, wall state and sample history, and negligible detection lag. An unmodelled population of corner-active sites adds a term independent of L, so measurements at a third size test the two-term assumption rather than merely improving its fit.

Tutorial · Propagate interfacial uncertainty into a rate

Suppose the barrier is estimated at $35 k_B T$, with five-percent relative uncertainty in interfacial tension while driving force, temperature and prefactor are treated as known. Estimate its effect on log rate and explain why the rate uncertainty is strongly asymmetric when expressed in ordinary units.

▼ **Differentiate the barrier first and exponentiate last**

The barrier scales as γ^3, so its first-order relative uncertainty is 15 percent. The uncertainty in $\log J$ is therefore about $35(0.15) = 5.25$, corresponding to multiplicative factors of roughly $e^{5.25} = 191$. A symmetric error in interfacial tension need not produce a symmetric additive error in rate. Computing the exact cubed upper and lower tensions produces further asymmetry.

Repeat the inference with a driving-force uncertainty, whose logarithmic coefficient is minus two in the barrier. Tension and driving parameters estimated from common coexistence data may be correlated. A direct comparison of predicted log rates across several undercoolings is often more stable than comparing raw rates spanning many orders of magnitude.

Cambridge Statistical Physics · metastability, phases and transitions

Nonlinear inference and implicit equations

Investigate what a residual and a local derivative measure in inverse fitting, discretised boundary equations and an implicitly advanced stiff mode.

1. Nonlinear fitting and local models
2. Predictions, parameter profiles and nonlinear uncertainty
3. When a single uncertainty interval hides two explanations
4. Absolute stability and stiffness
5. Shooting for two-point boundary values
6. Differential–algebraic equations and consistency
7. Why a boundary value problem can be easier than its shooting equation
8. Anisotropic and nonlinear dielectrics from local measurements to global fields
9. From mobile ions to screening lengths and differential double-layer capacitance
10. Measure inertia and gravity with two suspension axes

Nonlinear fitting and local models

Inverse problems and least squares

Derive a Gauss–Newton step, recognise misleading local curvature and use step acceptance to protect a nonlinear fit.

FROM FIRST PRINCIPLES

Build a local inverse from sensitivities

Let a differentiable forward model $f(p)$ predict several measurements from parameters p. Whiten its prediction errors by the measured noise covariance, and denote the resulting residual by $r(p)$. To infer a parameter change s from the current mismatch, expand about the current state,

$$r(p+s) = r(p) + J(p)s + O(\|s\|^2).$$

Here column j of J records the response to changing parameter j, with the other parameters fixed at the current values. Minimizing the squared norm of the linearized prediction supplies the Gauss–Newton step,

$$s = \operatorname*{argmin}_d \|r + Jd\|_2^2,$$

$$J^T J s = -J^T r.$$

Independent columns make the local step unique. Nearly parallel columns indicate parameter combinations that can compensate for each other in the data. Because J describes only the present neighbourhood, compare the actual residual at the proposed new state with the linear prediction. A new trial point must also respect any physical parameter bounds assumed by the forward model.

DERIVATION AND APPLICATIONS

Compare finite updates in a parameter and its logarithm

A positive coefficient such as a diffusivity or decay rate can be fitted directly or via a logarithm that enforces positivity automatically. Consider $p > 0$ and use a reference p_0 of the same units so that $q = \ln(p/p_0)$ is dimensionless. Let $r(p)$ be a whitened residual vector and J_p its derivative column at the current parameter. The chain rule and a linearized solve give

$$J_q = pJ_p,$$

$$\delta q = -\frac{J_q^T r}{J_q^T J_q} = \frac{\delta p}{p},$$

$$\delta p = -\frac{J_p^T r}{J_p^T J_p}.$$

Thus direct and logarithmic coordinates describe the same first-order change when the column is nonzero. Their undamped finite proposals differ. The direct step adds the change; the logarithmic step exponentiates it. Expanding the exponential gives the difference explicitly,

$$p_{\text{direct}} = p + \delta p,$$
$$p_{\log} = p\, e^{\delta p / p} = p + \delta p + \frac{\delta p^2}{2p} + O(\delta p^3 / p^2).$$

The local variance from the same linear model transforms consistently, assuming unit whitened noise and no added prior:

$$\operatorname{var}(\hat{q}) \simeq \frac{1}{p^2 J_p^T J_p},$$
$$p^2 \operatorname{var}(\hat{q}) \simeq \operatorname{var}(\hat{p}).$$

The positivity of a proposal does not ensure it reduces the nonlinear residual. A line search or other acceptance test can still be needed. Likewise, a prior density specified in one coordinate must be transformed with its Jacobian when a Bayesian analysis changes coordinates.

WORKED EXAMPLE

Compare two proposals against an exact one-channel fit

A direct sensor predicts $y = p$, with reading 1 and initial $p = 4$. The direct correction is -3, reaching the exact fit immediately. The log-coordinate correction is $-3/4$, giving $4e^{-0.75} = 1.889$. Both reduce the original error of 3, but the two linearizations have different finite remainders. The logarithmic model is nonlinear even for this linear physical sensor.

SECOND APPLICATION

Translate a local interval into physical units

A second inversion gives $p = 2.00\,\mathrm{mm}^2/\mathrm{s}$ and local standard deviation $s_q = 0.10$ in log coordinates. First-order propagation gives $s_p = 0.20\,\mathrm{mm}^2/\mathrm{s}$. Mapping the symmetric log interval $q \pm 0.20$ exactly gives $p \in [2e^{-0.2}, 2e^{0.2}] = [1.637, 2.443]\,\mathrm{mm}^2/\mathrm{s}$. The asymmetric physical interval retains the bounds specified in logarithmic units.

Expand the residual around a trial parameter

Let $r_i(p)$ be predicted minus observed data after whitening. Near p_k,

$$r(p_k + d) = r(p_k) + Jd + O(\|d\|^2),$$

where $J_{ij} = \partial r_i / \partial p_j$. Minimise the squared norm of this linear expression to obtain the Gauss–Newton step. The exact Hessian of $\frac{1}{2}\|r\|^2$ is $J^{\mathsf{T}}J + \sum_i r_i \nabla^2 r_i$. Gauss–Newton omits the second term. It is

especially accurate near a good fit with small residuals. With substantial residuals or severe model curvature it can predict the wrong improvement even if the Jacobian is computed perfectly.

Compute the step from a QR or SVD solve $Jd \simeq -r$. Verify a derivative independently using a directional difference $[r(p + hv) - r(p - hv)]/(2h)$ and compare with Jv across a few h. Errors that shrink and then grow as h decreases are consistent with truncation followed by roundoff. A disagreement stable across several step sizes often signals inconsistent units, residual signs or parameter ordering.

Work one update by hand

For $g(t; p) = e^{-pt}$, observe $y(1) = 0.50,\ y(2) = 0.25$ with equal standard deviations. Start at $p = 0$. The residuals are 0.50 and 0.75 and the Jacobian column is $(-1, -2)^{\mathsf{T}}$. The least-squares step is

$$d = -J^{\mathsf{T}} r/(J^{\mathsf{T}} J) = 0.4.$$

At $p = 0.4$ the two predictions are about 0.670 and 0.449. The residual shrinks, but a single linear step falls short of the exact parameter $\ln 2$. Recompute the derivative at the new point before the next update. Fitting the logarithms instead would change the implied measurement-noise model unless the weights are transformed appropriately.

Limit poor steps and enforce physical scales

A damped step minimises $\|r + Jd\|^2 + \mu\|Dd\|^2$. The diagonal scale D makes a step of one natural parameter unit comparable across components. Large μ restricts the step strongly; small μ approaches the ordinary Gauss–Newton solve. Evaluate the true objective at the trial point and compare its actual decrease with the decrease predicted by the linear residual model. A step with negative actual decrease should be rejected or shortened. Repeatedly tiny steps with a large gradient are evidence of a stalled solve, not convergence.

For a strictly positive rate, one can write $k = e^{\theta}$; then $\partial g/\partial\theta = k\,\partial g/\partial k$. This transformation enforces positivity and represents relative changes naturally. It also changes what a penalty on parameter distance means. Distinguish damping of an individual iteration from regularisation in the final objective: damping may decrease toward zero as a local method converges even when no penalty on the final parameters is intended.

The limits of a local uncertainty ellipse

At a well-determined full-rank fit with correctly scaled independent errors, $(J^{\mathsf{T}} J)^{-1}$ approximates parameter covariance in a local linear model. For $g(p) = p^2$ and observed positive y, either root has nonzero slope and a narrow local interval can arise around each. That interval misses the other root entirely. Bounds, multiple starts, profile fits and independent predictions give different checks on global uniqueness and nonlinearity. Distinct parameter values with similar residuals are a result to report; selecting the one reached by the first start can hide a scientifically important ambiguity.

Eliminate linear coefficients from a nonlinear fit

A decay amplitude may be unknown alongside its time constant. For fixed positive rate k the vector of predictions is $a v(k)$, with $v_i(k) = e^{-k t_i}$ and an unknown scalar amplitude a. Assume the data have already been whitened, applying the same transformation also to v. Minimising the squared residual at fixed k is an ordinary one-dimensional projection. Differentiating with respect to a gives $v^T(a v - y) = 0$, so $a(k) = v^T y / (v^T v)$ whenever the column is nonzero. Substitution leaves a scalar objective $F(k) = \|a(k) v(k) - y\|^2 / 2$.

To compute its derivative, differentiate the complete residual $r = a v - y$. The chain rule gives

$$F' = r^T [a' v + a v'].$$

But the optimal amplitude has already made $r^T v = 0$, so the first term vanishes exactly. Thus the profile derivative is $a r^T v'$, even though $a(k)$ changes as the rate changes. This cancellation requires an accurately solved inner projection. If an approximate amplitude leaves a visible residual along the column, ignoring its derivative can yield the wrong outer step. Several linear columns can be handled by the same projection argument while their rank remains fixed.

Diagnose the effect on three observations

Take times 0, 1 and 2 seconds, observed amplitudes $y = (4, 2, 1.5)^T$, and trial rate $k = \log 2$ per second. Then $v = (1, 1/2, 1/4)^T$, $v^T v = 21/16$, and $v^T y = 43/8$. Hence $a = 86/21$. Subtract the observations from the predictions to obtain

$$r = (2, 1, -10)^T / 21.$$

Its product with the design column is proportional to $2 + 1/2 - 10/4 = 0$, checking that no amplitude adjustment alone improves the fit. The value of the objective is

$$105 / (2 \cdot 21^2) = 5/42.$$

Now $v' = (0, -1/2, -1/2)^T$ in seconds, so

$$r^T v' = 3/14.$$

Multiplying by the fitted amplitude gives $F' = 43/49$ per unit rate in the chosen amplitude scaling. The positive derivative says that slightly reducing k lowers the local discrepancy. This sign has a physical explanation: the late observation lies above the fitted curve and a smaller decay rate raises the relative late amplitude. The amplitude will readjust simultaneously to keep the projection optimal.

What if all three acquisition times are the same? Every $v(k)$ is a nonzero multiple of the same constant column. Changing k rescales that column and the best amplitude compensates perfectly. The profiled objective is then flat although the inner projection is perfectly well defined. Eliminating a linear nuisance

coefficient improves the structure of a search but cannot cure missing information about decay rates when the timing protocol transmits only one measurement direction.

One can diagnose that limiting design before taking data by projecting the derivative column v' perpendicular to v. Only this perpendicular component can change the fitted curve after amplitude compensation. At identical times it is identically zero because $v' = -tv$. With distinct times it usually survives. Under approximately constant additive noise, inspecting its magnitude across candidate sampling schedules gives a quantitative way to choose the times that distinguish a changed shape from a changed amplitude.

The projected derivative has units of amplitude times time, so compare candidate schedules using the same time and amplitude scales.

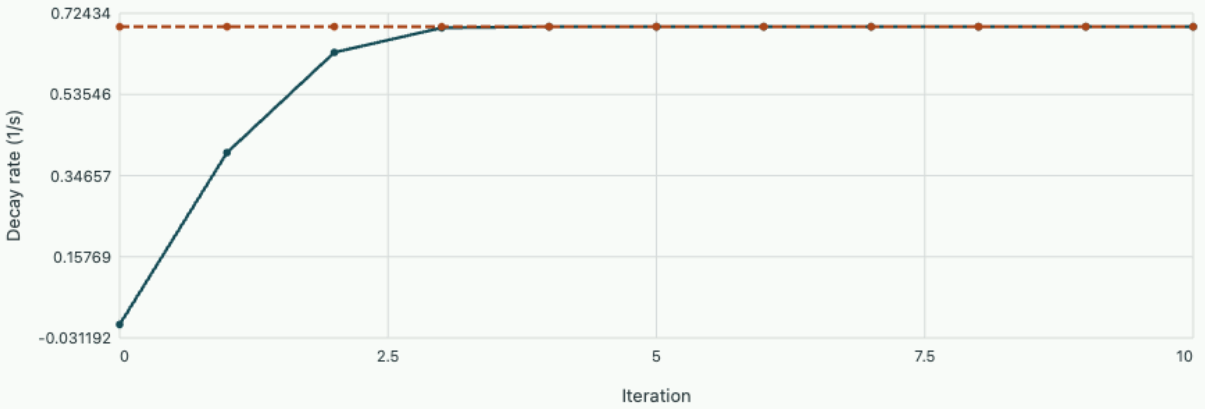
Change parameters · Nonlinear fitting and local models

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

Open this calculation with adjustable inputs

Try the calculation

Initial rate (1/s)	0
Second observation offset	0
Damping μ	0



Final rate 0.69315. Changing the second observation makes the two data inconsistent with a single exact decay; the solver then seeks a least-squares compromise.

Predictions, parameter profiles and nonlinear uncertainty

Inverse problems and least squares

Calculate intervals for a quantity of interest, then find where a linear covariance approximation loses the geometry of a nonlinear fit.

FROM FIRST PRINCIPLES

Propagate fitted uncertainty to a reported quantity

Suppose a fit gives a parameter estimate m with covariance C, and the quantity to be reported is $z = h(p)$. Near m, expand a differentiable h to first order. Writing $g = \nabla h(m)$, each small parameter change contributes through its directional sensitivity,

$$h(m + \delta p) - h(m) \simeq g^T \delta p,$$
$$\operatorname{var}(z) \simeq g^T C g.$$

Correlations can either reduce or enlarge the propagated variance. For a genuinely nonlinear quantity or a large parameter region, a local line may miss the full shape. One alternative holds the target value fixed and finds the best fit compatible with it. With an objective Φ, define

$$\Phi_{\text{prof}}(z) = \min_{p:\, h(p)=z} \Phi(p).$$

Comparing this constrained objective with its unconstrained minimum tracks compensating changes in the nuisance parameters. A confidence level for its increase requires an explicit statistical model; tracing an objective curve alone supplies no probability calibration.

DERIVATION AND APPLICATIONS

Derive a profile interval using the covariance of a linear fit

Suppose a Gaussian linear experiment gives unconstrained estimate m and strictly positive covariance C, with correctly known noise scale. Its negative log likelihood relative to the optimum is the quadratic distance $\frac{1}{2}(p - m)^T C^{-1}(p - m)$. To profile a reported scalar $z = c^T p$, require the candidate parameters to satisfy that equality. Introduce a multiplier and differentiate with respect to the vector,

$$C^{-1}(p - m) + \lambda c = 0,$$
$$p = m - \lambda C c.$$

The constraint determines $\lambda = (c^T m - z)/(c^T C c)$. Substitution gives both the parameter combination selected at each target value and the profile height,

$$p_z = m + Cc\frac{z - c^T m}{c^T C c},$$

$$2[\Phi_{\text{prof}}(z) - \Phi(m)] = \frac{(z - c^T m)^2}{c^T C c}.$$

This reproduces the usual propagated Gaussian variance while showing exactly how the other fitted coordinates move along the profile. For a monotone function $q = g(z)$, imposing q fixes the corresponding $z = g^{-1}(q)$. The profile then transforms without approximation,

$$2\Delta\Phi(q) = \frac{[g^{-1}(q) - c^T m]^2}{c^T C c}.$$

An interval selected by a likelihood-ratio threshold maps by applying g to its two endpoints. It can be asymmetric even when the underlying likelihood in p is exactly quadratic. If g is not one-to-one, every feasible branch must be considered and the smallest objective chosen.

WORKED EXAMPLE

Track the nuisance displacement along a profile

Take $m = (2, 4)^T$, $C = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$, and target $z = p_1 + p_2$. The target mean is 6 and variance 3. At $z = 9$, $Cc = (1.5, 1.5)^T$, giving constrained state $(3.5, 5.5)^T$. The twice-log-likelihood increase is $3^2/3 = 3$. Holding the second parameter fixed would impose a different and more restrictive path.

SECOND APPLICATION

Transform an interval for a positive output

A second model estimates $z = \ln(q/q_0)$ with mean 0.4 and deviation 0.3; let $q_0 = 10$. At threshold $2\Delta\Phi = 4$, the interval in z is $[-0.2, 1.0]$. Mapping its endpoints gives $q \in [8.19, 27.18]$, around the maximum-likelihood value $10e^{0.4} = 14.92$. A symmetric interval about that value would lose the exact profile geometry.

Push uncertainty through the prediction

Often a user needs tomorrow's temperature or a time to threshold, rather than all fitted coefficients. For a linear target $h(x) = c^{\mathsf{T}}x$ and Gaussian posterior, h is Gaussian with mean $c^{\mathsf{T}}m$ and variance $c^{\mathsf{T}}Cc$. Include covariance terms: positively correlated errors raise uncertainty in a sum and reduce it in a difference. For a new noisy measurement add its own independent measurement variance. Reporting this predictive interval as the uncertainty of the noise-free underlying response would mix two questions.

For a smooth nonlinear h, expand about the mean m. The first-order variance is $\nabla h^{\mathsf{T}} C \nabla h$; the leading correction to the mean is $\frac{1}{2}\mathrm{tr}[H_h(m)C]$. If the gradient vanishes, first-order variance propagation can return zero despite substantial uncertainty. For example, if x is centred normal with variance s^2 and

$h = x^2$, then $\mathbb{E}h = s^2$ and $\operatorname{Var} h = 2s^4$, using the Gaussian fourth moment. The linear approximation at zero would miss both.

A positive rate and a nonlinear lifetime

Suppose $\theta = \ln(k/k_{\text{unit}})$ is normal with mean m and standard deviation s. A decay lifetime is

$$t_* = 1/k = e^{-\theta}/k_{\text{unit}}.$$

Its median is e^{-m}/k_{unit}; its mean is larger by $e^{s^2/2}$. A central 95% interval is $[e^{-m-1.96s}, e^{-m+1.96s}]/k_{\text{unit}}$. These endpoints need not be equally far from the median. For median 10 s and $s = 0.5$, the interval is approximately [3.75, 26.6] s and mean 11.33 s. Transform samples or quantiles if strong nonlinearity makes symmetric error bars misleading.

A profile follows compensating changes

Let the fit have parameters (p, q). To profile p, fix it temporarily at each trial value and minimise the objective over q. The resulting curve reflects how much the other parameter can compensate. Holding q fixed instead gives a slice that is generally steeper and gives deceptively narrow intervals. For the sum-of-squares surface $F = (p + q - 2)^2 + 0.04(q - 1)^2$, minimising gives

$$q = (2.04 - p)/1.04.$$

The resulting profile is $0.04(p - 1)^2/1.04$. A slice at the joint optimum's $q = 1$ is just $(p - 1)^2$. Both have minimum at one but radically different curvature.

Read the level as well as the curve

A profile likelihood is a maximisation over nuisance parameters. A Bayesian marginal instead integrates their probability. Even when their modes agree the two objects can differ in shape in nonlinear models. A likelihood-ratio threshold yields approximate confidence sets under regularity assumptions and enough informative data. Boundaries, multiple modes and poorly identified nuisance parameters may invalidate a familiar chi-square cutoff. Report the objective convention and how the threshold was calibrated, e.g. by simulated repetitions of the full experiment. A plot of the raw profile and representative predicted curves along it remains interpretable even when an asymptotic confidence level is questionable.

Invert a test to put an interval on a ratio

A scientifically relevant quantity may be $q = \mu_y/\mu_x$, for example the response per unit excitation. Estimates $\hat{x}, \hat{y}$ can each have nearly Gaussian uncertainty even when their ratio is highly asymmetric. Suppose for now that they are independent normals with known variances s_x^2, s_y^2, and that the true

denominator is nonzero. To test a proposed ratio q, consider the contrast $\hat{y} - q\hat{x}$. If the proposed ratio is correct, that contrast has mean zero and variance $s_y^2 + q^2 s_x^2$. The source of difficulty, the uncertain denominator, now enters the reference variance explicitly.

Collect all values whose absolute standardised contrast is at most a normal quantile z. Squaring and expanding gives the quadratic inequality

$$(\hat{x}^2 - z^2 s_x^2) q^2 - 2\hat{x}\hat{y} q + (\hat{y}^2 - z^2 s_y^2) \le 0.$$

This inversion gives a confidence set because it contains the true ratio exactly when the correctly standardised contrast falls in the specified acceptance region. Its shape need not resemble a symmetric error bar. If the leading coefficient changes sign, the accepted set may be unbounded. A covariance between $\hat{x}$ and $\hat{y}$ changes the variance to $s_y^2 + q^2 s_x^2 - 2q\,\mathrm{Cov}(\hat{x}, \hat{y})$, so the preceding coefficients require modification.

A precise reading divided by an uncertain excitation

Use $z = 2$, corresponding to about 95.45 percent coverage for the stated normal model. Set $\hat{x} = 2, \hat{y} = 4,\ s_x = 0.5,\ s_y = 1$. The quadratic coefficients are 3, −16 and 12. Its roots are $(16 \pm \sqrt{112})/6$, about 0.903 and 4.431. The accepted set is the closed interval between them because the polynomial opens upward. The point estimate is two. The interval extends considerably farther upward than downward because a smaller plausible excitation can produce a much larger ratio. A symmetric linearised interval loses this asymmetry.

Now keep the observations and numerator uncertainty fixed but raise s_x to 1.5. The resulting inequality is $-5q^2 - 16q + 12 \le 0$, whose roots are $(-16 \pm \sqrt{496})/10$, about −3.827 and 0.627. The accepted set consists of the two exterior rays. It includes very large positive and negative values. For such values the standardised contrast approaches $|\hat{x}|/s_x = 4/3$, inside the threshold two, precisely because the data cannot rule out an excitation arbitrarily close to zero.

An investigator with independent physical information that the excitation is strictly positive may intersect the set with the resulting permitted parameter space. A numerical routine that simply discards negative draws or clips ratios at a convenient maximum is instead changing the uncertainty analysis. The unbounded set is a measurable statement about what an experiment failed to localise, and identifies the denominator measurement as an appropriate target for redesign.

When $\hat{x}^2 = z^2 s_x^2$ exactly, the quadratic term disappears. The confidence set is then ordinarily a half-line determined by a linear inequality. This is the transition between bounded and exterior solutions. A robust implementation must handle it directly instead of dividing by a leading coefficient that may be zero or extremely small.

Change parameters · Explore the worked calculation

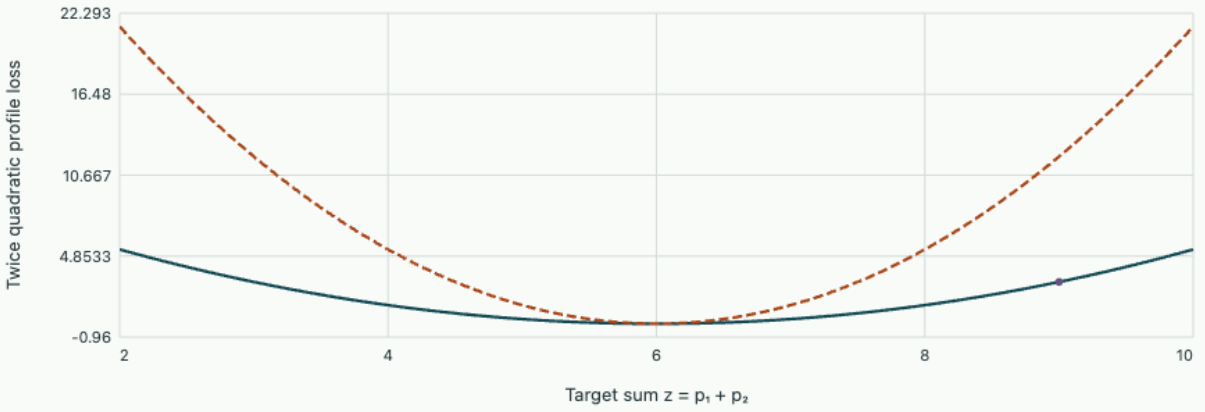
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Prior parameter correlation **0.5**

Selected sum target z **9**



Both parameters refitted | Second parameter held at its mean | Selected profiled target

Selected profiled state (3.5, 5.5). Sum variance 3. Profiling moves nuisance parameters; fixing them gives a different curvature.

When a single uncertainty interval hides two explanations

Inverse problems and least squares

Start from a squared-amplitude measurement, derive the full posterior and design a second observation capable of identifying the sign.

Symmetry is an identifiability statement

A sensor measures $d = x^2 + \eta$ with Gaussian noise of standard deviation s. A symmetric Gaussian prior of variance τ^2 puts equal weight on opposite signs. Bayes' rule gives the density

$$p(x \mid d) \propto \exp\left[-\frac{(d - x^2)^2}{2s^2} - \frac{x^2}{2\tau^2}\right].$$

The exponent is even, so the posterior mean is exactly zero whenever the integral exists. That mean may make an extremely poor prediction of the observed quantity. To locate modes, differentiate the negative log density to get $x[2(x^2 - d)/s^2 + 1/\tau^2]$. Two off-centre modes appear precisely when $d > s^2/(2\tau^2)$. At smaller d, the prior and noise merge the explanations into one central mode.

Numbers and competing summaries

Set $d = 1$, $s = 0.2$, and $\tau = 2$. The modes are at $\pm\sqrt{0.995}$, close to ± 0.9975. The density ratio of $x = 0$ to $x = 1$ is $\exp[-1/(2s^2) + 1/(2\tau^2)]$, about 4.2×10^{-6}. The reported mean of zero sits in a very low-density region.

For $d = -0.05$, there is no real parameter that exactly fits the measured amplitude squared. A noisy sensor can still produce this value. The distribution has a single mode at zero, and there is no justification for taking the square root of the negative datum or discarding the observation.

Buy a measurement that identifies the sign

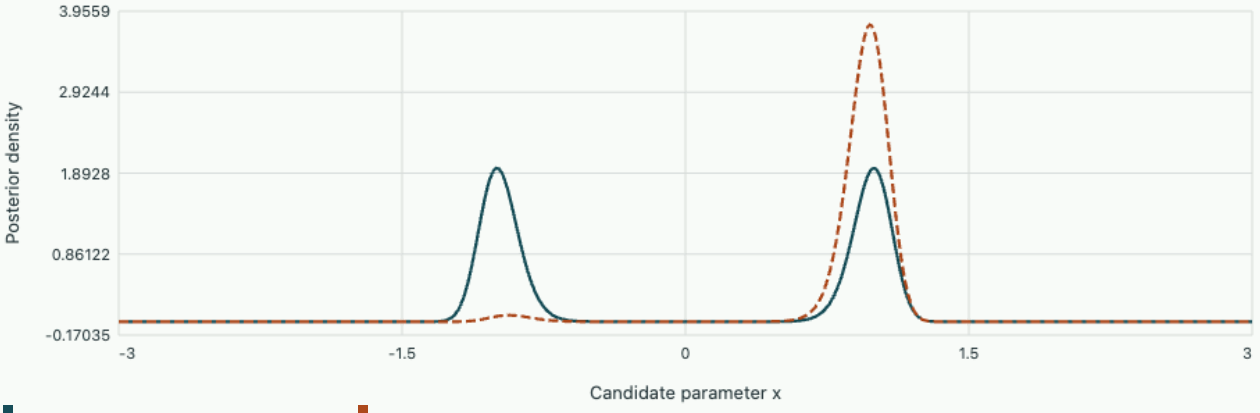
Assume an independent second sensor reports $b = x + \epsilon$ with known standard deviation 0.5. Add $-(b - x)^2/(2 \cdot 0.5^2)$ to the log density. The original two signs now receive different weights. Both curves in the plot are independently normalised by quadrature on the displayed interval; the sign probability is conditional on that interval. Compare the observation $b = 0$ with $b = 1$.

Open this experiment online

Try the calculation

Squared-amplitude datum d **1**

Signed sensor reading b **0.5**



Squared amplitude only With signed sensor

Probability of positive x on [−3,3]: 50% before the signed measurement, 97.698% afterwards.

Tutorial on designing another experiment

Repeat the squared-amplitude measurement a hundred times with independent noise. What can become precise?

▼ **Worked solution**

The magnitude x^2 can become very precise. Every likelihood factor remains even in x, so the sign probabilities stay equal with this prior. More data of the same kind cannot break that symmetry.

If the two candidate states were exactly $x = \pm 1$, what likelihood ratio does $b = 0.5$ give in favour of the positive state?

▼ **Worked solution**

The exponent difference is

$$((0.5+1)^2 - (0.5-1)^2)/(2 \cdot 0.5^2) = 4.$$

The likelihood ratio is e^4, about 54.6. A continuous posterior also averages over the spread around each candidate.

Absolute stability and stiffness

Numerical ordinary differential equations

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:

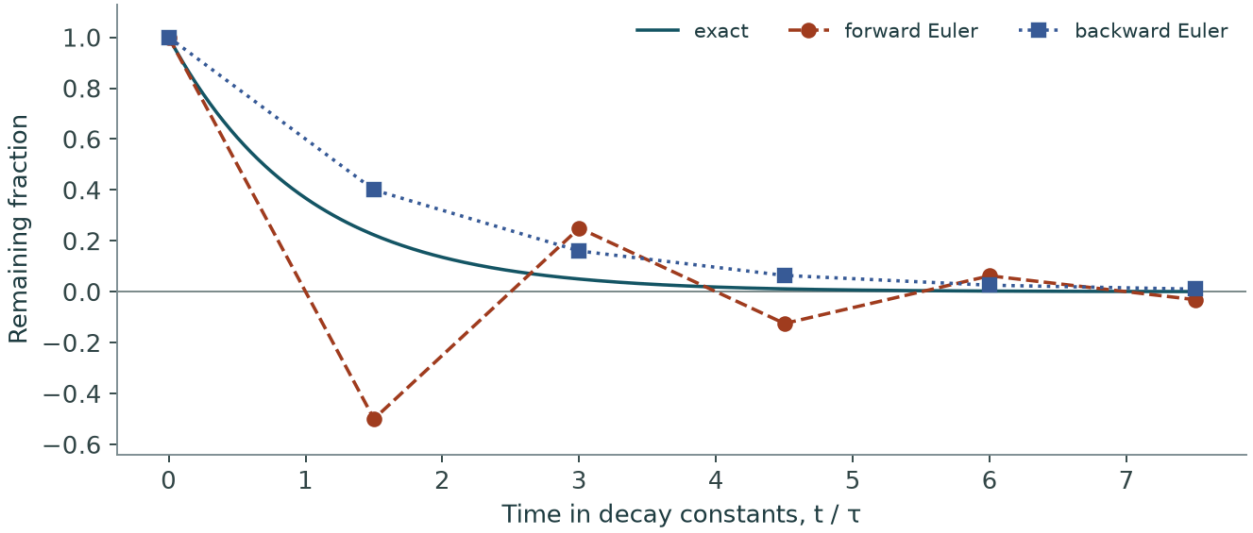
$$y_n = R(z)^n y_0,$$
$$R_{\text{forward}} = 1 + z,$$
$$R_{\text{backward}} = (1 - z)^{-1}.$$

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.

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.

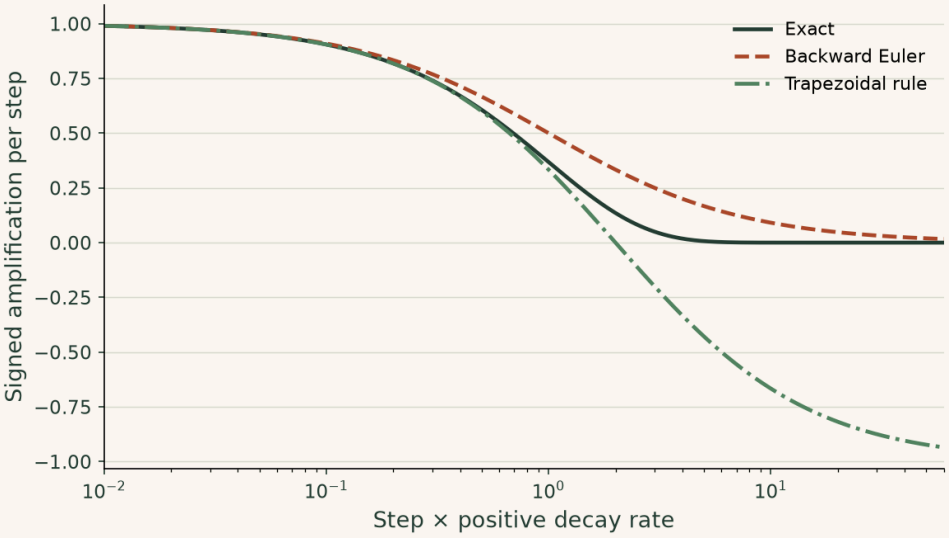
For comparison, trapezoidal integration of the slow-path derivative integrates the linear function $g' = 2t$ exactly over each interval. Substituting the exact nodal values $g(t_n)$ therefore satisfies its full discrete equations, so an exactly initialized trajectory stays on this particular path. However,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

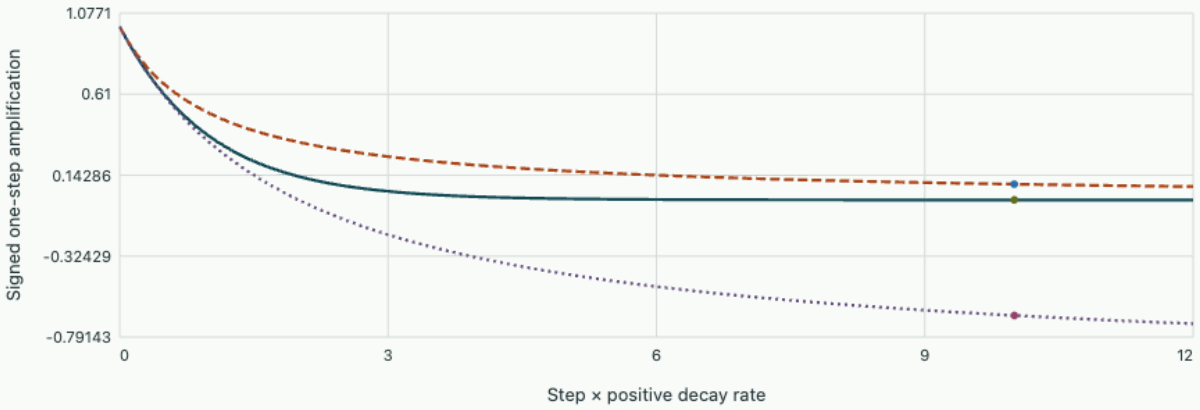
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Maximum step × rate **12**

Selected step × rate **10**



Exact exponential · Backward Euler · Trapezoidal · Selected exact amplification · Selected backward Euler · Selected trapezoidal

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

Shooting for two-point boundary values

Numerical ordinary differential equations

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 $\lambda_{\text{fold}} \simeq 3.5138$. 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}.

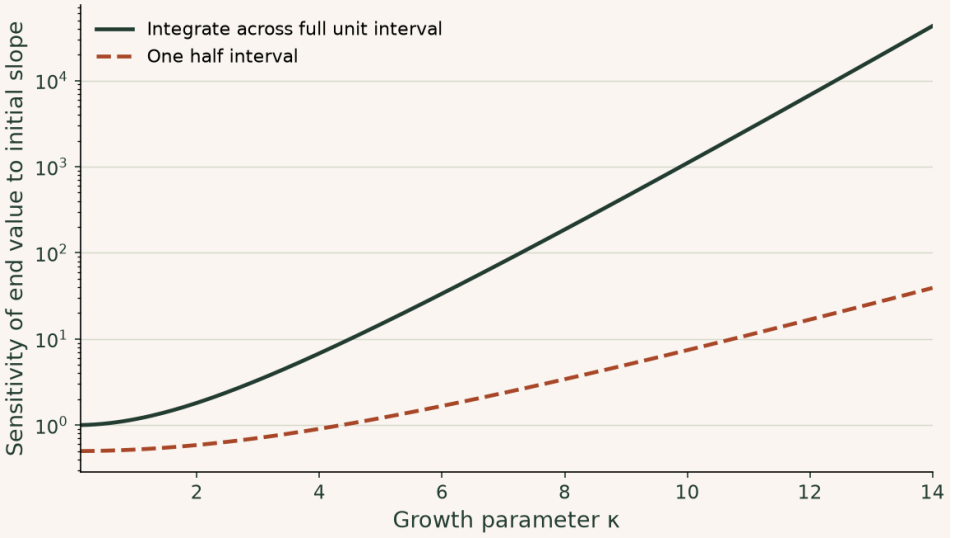
A residual search on $k \in [2, 7]$ could instead overlook both $k = \pi$ and 2π, since the residuals at the endpoints have the same positive sign. Counting interior nodes provides an independent mode label. With normalisation $u'(0) = 1$, the first eigenfunction has no internal zero and the second has 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

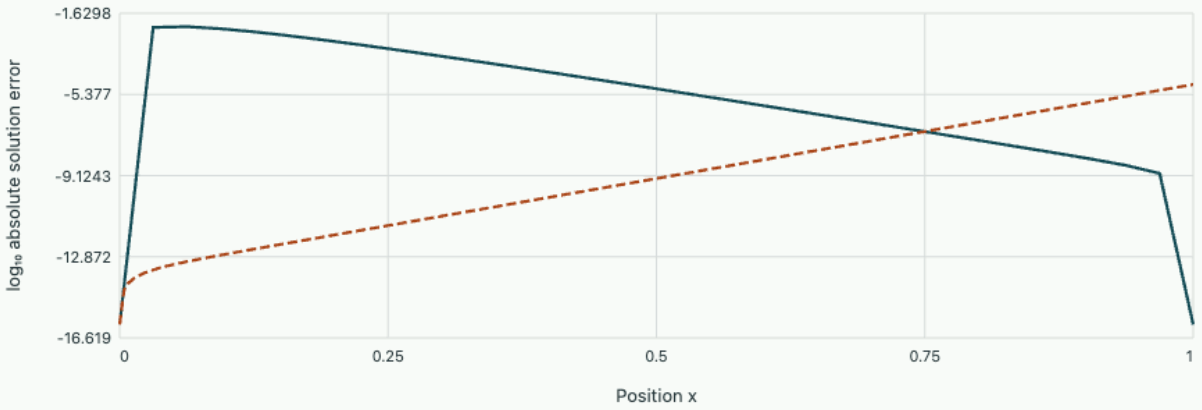
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Layer parameter λ **20**

Number of mesh intervals **32**



Finite-difference collocation Propagated slope error

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

Differential–algebraic equations and consistency

Numerical ordinary differential equations

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.

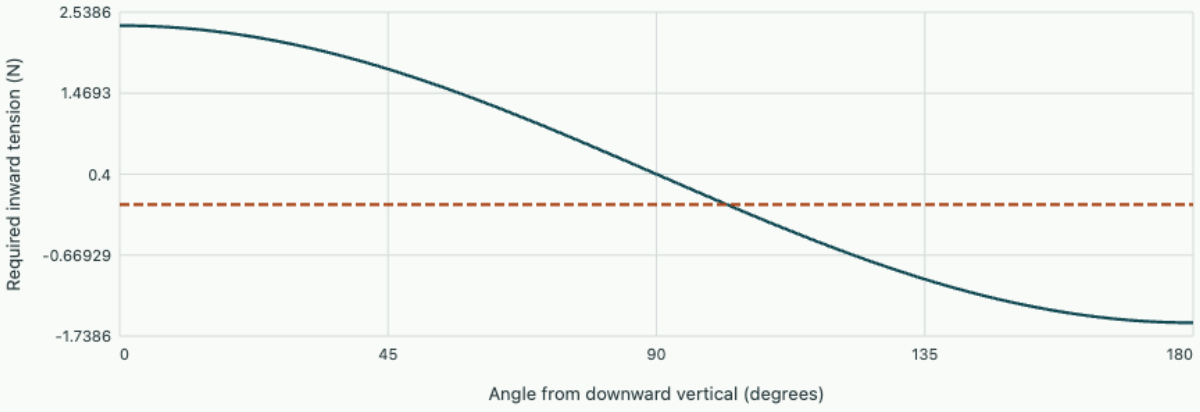
Change parameters · Explore the worked calculation

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.

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

Numerical ordinary differential equations

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

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

$$j = 1, \ldots, N-1.$$

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

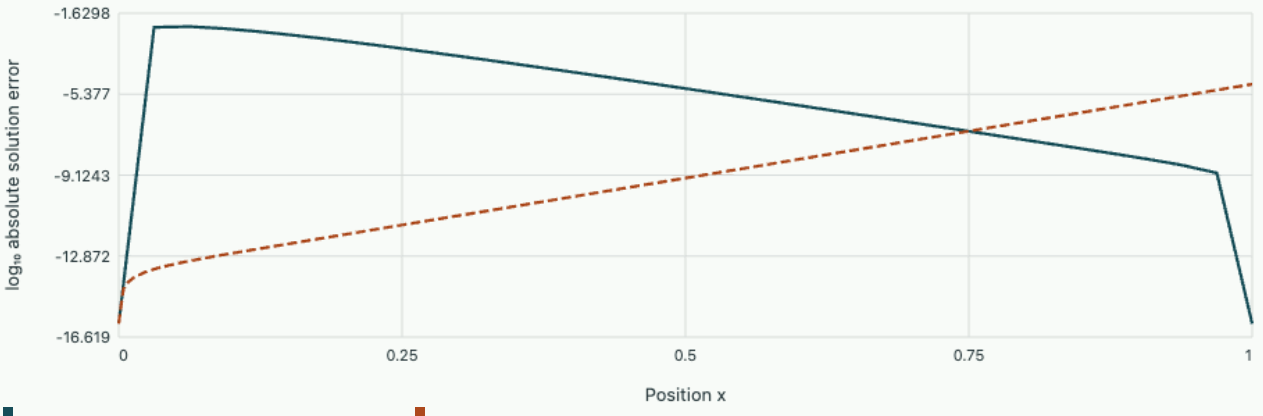
The graph plots logarithmic absolute errors against position. The shooting curve isolates propagation of a controlled absolute initial slope perturbation of 10^{-12}, using the analytical propagation factor to avoid cancellation in the demonstration. The collocation curve solves the tridiagonal system 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.

Anisotropic and nonlinear dielectrics from local measurements to global fields

Electrostatic field theory and measurement

A single dielectric constant cannot describe a crystal whose displacement tilts away from its electric field, or a reversible material whose incremental response changes with bias. Starting with the constitutive measurement clarifies which parts of ordinary electrostatics survive.

Measure a tensor rather than assuming collinearity

In a homogeneous linear anisotropic material,

$$\mathbf{D} = \boldsymbol{\epsilon}\mathbf{E}.$$

The entry ϵ_{ij} measures displacement component i in response to electric-field component j. A reversible reciprocal dielectric has a symmetric positive-definite tensor. Its principal axes are orthogonal and its positive eigenvalues are the principal permittivities. Field-energy density $u = \mathbf{E}\cdot\boldsymbol{\epsilon}\mathbf{E}/2$ is then positive for a nonzero field. Positive definiteness concerns all directions; having three positive diagonal entries in one basis alone is insufficient.

Choose principal relative permittivities $2, 4, 6$ along x, y, z. With a field $(1000, 0, 1000)$ V/m the displacement is $\epsilon_0(2000, 0, 6000)$. The field makes 45° to z, the displacement only 18.4°; their directions differ. Energy density is 3.54×10^{-5} J/m³. A flat free-charge sensor normal to z responds to D_z; an electrostatic force probe constrained along x measures E_x. Their readings cannot be interpreted as two component measurements of one vector differing by a single scalar.

Derive the field of a point charge in an anisotropic bulk

For a constant symmetric tensor write $\boldsymbol{\epsilon} = \epsilon_0 K$, with positive dimensionless matrix K. Poisson's equation is

$$-\epsilon_0 \nabla_{\mathbf{x}} \cdot K \nabla_{\mathbf{x}} V = q\delta^3(\mathbf{x}).$$

Change coordinates to $\mathbf{y} = K^{-1/2}\mathbf{x}$. The gradient transforms by the same symmetric matrix and the operator becomes the ordinary isotropic Laplacian in $\mathbf{y}$. The source transforms too: physical volume is $d^3x = \sqrt{\det K}\, d^3y$, so

$$\delta^3(\mathbf{x}) = \delta^3(\mathbf{y})/\sqrt{\det K}.$$

The decaying solution is

$$V(\mathbf{x}) = \frac{q}{4\pi\epsilon_0\sqrt{\det K}}\,\frac{1}{\sqrt{\mathbf{x}^T K^{-1}\mathbf{x}}},$$

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0\sqrt{\det K}}\,\frac{K^{-1}\mathbf{x}}{(\mathbf{x}^T K^{-1}\mathbf{x})^{3/2}}.$$

A check is the scalar case $K = kI$. The determinant contributes $k^{3/2}$ and the distance contributes $1/\sqrt{k}$, returning $q/(4\pi\epsilon_0 kr)$. The equipotentials in laboratory coordinates are ellipsoids, while $\mathbf{D} = \epsilon_0 K\mathbf{E}$ happens to point along $\mathbf{x}$ for this infinite homogeneous point-source problem. Its magnitude varies with direction on a physical sphere. Gauss's law still gives total free flux q; equating that flux to $4\pi r^2$ times a constant magnitude would be invalid. Conducting or material boundaries also transform under this coordinate map and generally cease to be simple physical spheres.

Along any principal axis i, at distance r, the potential is proportional to $\sqrt{K_i}/\sqrt{\det K}$. With principal values $2, 4, 6$, two probes equally far away on the z, x axes measure a voltage ratio $\sqrt{6/2} = \sqrt{3}$. One might have expected the largest permittivity to reduce the voltage most strongly along that axis. The observed result reflects how the full anisotropic medium distributes the displacement of a point source; it cannot be derived by inserting a direction-specific permittivity into the isotropic radial formula.

Which permittivity does a slab capacitance measure?

Place broad metal plates normal to a unit vector $\mathbf{n}$ and neglect edge effects. Constant electrode potentials force a purely normal electric field $V\mathbf{n}/d$, independent of the off-diagonal dielectric entries. The displacement generally has a tangential component, but the metal's free-charge reading is D_n. Thus the measured normal capacitance per area is

$$C/A = (\mathbf{n}^T \boldsymbol{\epsilon} \mathbf{n})/d.$$

For axes with relative values 2 and 6, cut the normal at 45° between them. The relative plate permittivity is 4.

A different preparation constrains displacement to be purely normal, for instance in an effectively uniform bulk state with known flux vector. Then $\mathbf{E} = \boldsymbol{\epsilon}^{-1} D_n \mathbf{n}$, so

$$D_n/E_n = 1/(\mathbf{n}^T \boldsymbol{\epsilon}^{-1} \mathbf{n}).$$

For the same cut the result is $1/[(1/2 + 1/6)/2] = 3$ in relative units. Both are correct for their own field constraints. Flat equipotential plates do not generally enforce tangential displacement equal to zero. A boundary-value calculation using that extra condition could report the wrong dielectric constant from an otherwise perfect experiment.

With an interface between two tensors, integrate Faraday's law around a small static rectangular loop to enforce continuous tangential E and Gauss's law over a pillbox to enforce the specified normal D jump. Write the first material's measured tangential field into the second's constitutive law and solve for its normal electric field from D_n. No scalar Snell-like rule involving only a ratio of two relative permittivities can represent arbitrary crystal orientations. Some field components may vanish by symmetry in one experiment and reappear after rotating the sample.

Reversible response distinguishes stored energy and coenergy

For a slowly and reversibly assembled homogeneous dielectric, increment of stored electrical energy per volume is $du = \mathbf{E} \cdot d\mathbf{D}$. A complementary density $w(\mathbf{E}) = \mathbf{E} \cdot \mathbf{D} - u$ has $dw = \mathbf{D} \cdot d\mathbf{E}$. In a scalar nonlinear law $D(E) = \epsilon E + \alpha E^3$ with $\epsilon, \alpha > 0$, integration gives

$$w(E) = \tfrac{1}{2}\epsilon E^2 + \tfrac{1}{4}\alpha E^4,$$

$$u(E) = ED(E) - w(E) = \tfrac{1}{2}\epsilon E^2 + \tfrac{3}{4}\alpha E^4.$$

The familiar $ED/2$ holds only in the linear case. For a small perturbation superposed on bias E_b, $dD/dE = \epsilon + 3\alpha E_b^2$ is the differential permittivity. A dc charge-to-voltage ratio measures the secant value $\epsilon + \alpha E_b^2$. At $\alpha E_b^2 = 0.20\epsilon$ these are 1.60ϵ and 1.20ϵ, a substantial difference even though the constitutive curve remains smooth.

The field equation becomes $\nabla \cdot \mathbf{D}(-\nabla V) = \rho_f$ and superposition generally fails. For a differentiable energy-based vector law, $\partial D_i/\partial E_j = \partial^2 w/\partial E_i \partial E_j$ is symmetric. A positive-definite differential tensor gives local stability and helps preserve ellipticity of the electrostatic boundary problem. A hysteretic material can depend on its prior history and need extra state variables; an area in a closed constitutive cycle then measures dissipated work, not reversible recoverable energy.

A nonlinear field calculation that needs only a scalar root

Place a sphere with charge q in an infinite isotropic nonlinear medium of the preceding type. Spherical symmetry and Gauss's law still determine $D_r = q/(4\pi r^2)$ outside. Invert $\epsilon E_r + \alpha E_r^3 = D_r$ point by point. For $\alpha > 0$ the derivative is positive everywhere, ensuring one physical real root for each D_r. A bisection bracket for $q > 0$ is $0 \le E_r \le D_r/\epsilon$; evaluating the residual at the upper endpoint gives positive cubic surplus. Compute voltage relative to infinity by integrating this field, rather than replacing D/E by an arbitrary constant taken at the surface.

For weak nonlinearity set $E = D/\epsilon + \delta$ and keep the leading cubic correction. One obtains

$$E \simeq D/\epsilon - \alpha D^3/\epsilon^4.$$

At large radii a $1/r^2$ displacement therefore acquires a negative $1/r^6$ correction in the electric field, giving a $1/r^5$ potential correction. A probe can misinterpret that extra radial scale as a change of source shape if material nonlinearity is not independently controlled. Double the source charge and compare the field after division by q. In a linear medium those curves coincide; the perturbative deviation here scales as q^2.

Resolve an oblique crystal interface component by component

A lateral detail or an oblique electrode can produce a field with both normal and tangential components at an interface. Consider a locally planar interface with tangent $\hat{\mathbf{x}}$ and normal $\hat{\mathbf{z}}$ from isotropic medium 1

into a crystal. In the xz plane let the crystal's relative tensor be $\begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}$, with entries in the order x, z. Medium 1 has relative permittivity 2 and measured field $E_{1x} = 120$ V/m, $E_{1z} = 240$ V/m. There is no free sheet at the interface. Neither input field is assumed normal to the boundary.

From a very short rectangular loop crossing the interface, the tangential electrostatic field is continuous, so $E_{2x} = 120$ V/m. The normal displacement on side 1 is $D_{1z}/\epsilon_0 = 2(240) = 480$ V/m. Equating it to $E_{2x} + 3E_{2z}$ yields $E_{2z} = 120$ V/m. The tangential displacement can jump. In the crystal it is $D_{2x}/\epsilon_0 = 4(120) + 120 = 600$ V/m, compared with $D_{1x}/\epsilon_0 = 240$ V/m. Neither tangent D nor normal E is required to be continuous. A field solver that enforces continuity of all components can silently erase the very anisotropy the experiment was intended to measure.

Find the bound surface charge using two routes. The true charge jump is $\epsilon_0(E_{2z} - E_{1z}) = -120\epsilon_0 = -1.06$ nC/m². With no free sheet this must be bound. Directly, $P_{1z}/\epsilon_0 = D_{1z}/\epsilon_0 - E_{1z} = 240$ V/m and $P_{2z}/\epsilon_0 = 480 - 120 = 360$ V/m. The net bound charge from the two facing polarisation surfaces is $P_{1z} - P_{2z} = -120\epsilon_0$, giving the same result. At a rotated interface it can be safer to perform these component checks than to infer charge from one visual change in field-line angle.

A controlled reversal can expose the off-diagonal tensor entry. Keep E_{1z} fixed and reverse E_{1x} using independent side electrodes. Then $E_{2x} = -120$ V/m and $E_{2z} = (480 + 120)/3 = 200$ V/m. The normal response therefore changes by 80 V/m even though the imposed normal displacement is unchanged. A scalar model, or a crystal whose tensor had zero xz entry in this coordinate system, would predict no such change. In practice the side electrodes must be far enough from the chosen interface patch for their intended approximately uniform fields to be realised locally.

Derive pressure for a reversible nonlinear capacitor under both electrical controls

Take a parallel-plate gap d, area A, completely filled by a reversible material with specified constitutive law $D(E)$. For this idealised calculation assume that changing d changes only the volume occupied by otherwise identical material at fixed thermodynamic properties. Exclude any additional mechanical elastic stress, electrostriction or chemical work needed to preserve that preparation. Keep the field homogeneous and neglect fringing. Under charge control the displacement $D = Q/A$ is fixed, so the stored electric energy is $U = Ad\,u(D)$ with $u = \int_0^D E(D')\,dD'$. The force tending to increase the gap is

$$F = -\partial_d U|_Q = -Au(D).$$

Thus the attractive field contribution has magnitude Au.

Now hold voltage fixed. The electric field is $E = V/d$; decreasing the gap also changes the charge the source must supply. Legendre transforming the energy by subtracting VQ gives the appropriate potential $\mathcal{G} = U - VQ = -Ad\,w(E)$. Its derivative at fixed V is $-Aw(E) + AE\,dw/dE = -Aw + AED = Au$. Hence $F = -\partial_d \mathcal{G}|_V = -Au$, agreeing for the same

instantaneous state. Differentiating U at fixed voltage without the source term would produce a different answer. The calculation identifies the source work excluded by that incorrect derivative.

For $D = \epsilon E + \alpha E^3$ with $\alpha E_b^2 = 0.20\epsilon$,

$$u(E_b) = 0.65\epsilon E_b^2.$$

The often-used linear formula $E_b D_b/2$ would give $0.60\epsilon E_b^2$, underestimating the stated ideal field contribution. With $\epsilon = 4\epsilon_0,\ E_b = 1.0 \times 10^5$ V/m, the predicted magnitude per area is 0.230 Pa. This number follows from a particular reversible constitutive model and mechanical variation. A real force between electrodes compressing a solid sample also contains material mechanical stresses, and the permittivity itself can change with density. The correct experiment must specify which matter moves when the gap changes.

An independent numerical test computes $U(d \pm \delta d, Q)$ at fixed charge or $\mathcal{G}(d \pm \delta d, V)$ at fixed voltage, forms centred derivatives and chooses the charges or voltages that describe the same midpoint state. If the two forces converge to opposite signs, check the mechanical displacement convention and the source-work term. Agreement of the pressures at a single bias does not prove the constitutive approximation is valid at other fields. A separate differential capacitance scan can verify the slope required by the energy model across the displacement range.

Solve a layered nonlinear capacitor by displacement continuation

A useful intermediate problem between a uniform nonlinear slab and a full finite-element solve is a pair of different nonlinear layers stacked normal to large plates. With no free interface charge, the same normal displacement D passes through both. Each has a reversible monotone law $D_i(E_i) = \epsilon_i E_i + \alpha_i E_i^3$, $\epsilon_i > 0, \alpha_i \geq 0$. At fixed terminal charge $D = Q/A$ is known. Invert each material separately and add the voltages,

$$V(D) = d_1 E_1(D) + d_2 E_2(D).$$

At fixed voltage solve this last scalar equation for D. Its derivative is positive:

$$\frac{dV}{dD} = \frac{d_1}{\epsilon_1 + 3\alpha_1 E_1^2} + \frac{d_2}{\epsilon_2 + 3\alpha_2 E_2^2},$$

$$\frac{C_{\mathrm{diff}}}{A} = \left(\frac{dV}{dD}\right)^{-1}.$$

Thus it is the differential permittivities, evaluated at the two actual and generally different local fields, that combine as a series small-signal capacitance. Using a differential response measured at the overall average field can give an inconsistent result even when the series formula itself is correct. The secant ratio D/V has another interpretation and should be compared only with a dc charge reading.

Choose a field scale $E_* = 1000$ V/m and $D_* = \epsilon_0 E_*$. In the first material let $D/D_* = 2e_1 + e_1^3$, $e_1 = E_1/E_*$, while the second is linear $D/D_* = 6e_2$. Give both layers thickness 1 mm. At $D = 3D_*$,

the first equation is solved by $e_1 = 1$; the second gives $e_2 = 0.50$. The total voltage is 1.50 V, and $D = 26.56$ nC/m². The dc capacitance per area is 17.71 nF/m². At this same operating point the differential relative permittivities are 5 and 6, giving $C_{\mathrm{diff}}/A = \epsilon_0/[0.001/5 + 0.001/6] = 24.15$ nF/m². A small voltage ripple measures a capacitance about 36% larger than the secant value.

Check the bias state from energy without using terminal current. The stored energy per volume in layer 1 is $\epsilon_0 E_*^2[e_1^2 + 3e_1^4/4]$ and in layer 2 is $3\epsilon_0 E_*^2 e_2^2$. Multiplying by the respective thicknesses at the chosen state gives total energy per area $0.00250\epsilon_0 E_*^2 = 22.14$ nJ/m². Integrating the voltage as a function of displacement from zero to $3D_*$ must give the same result, since $d(U/A) = V\,dD$. The shortcut $VD/2$ would instead give 19.92 nJ/m². It lies between neither two guaranteed bounds nor two voltage controls; it is simply unjustified for this nonlinear path.

For voltage-controlled computation use a continuation in applied voltage. Start at low bias with the known linear two-layer solution and reuse its displacement as a first guess for the next monotone scalar solve. A bracketing method with increasing positive endpoints always finds the root for positive applied voltage in these specific laws. Once converged, evaluate both layers' displacement residuals, their voltage sum and the independent energy integral. A lower iteration error alone cannot correct a model that accidentally multiplied the cubic coefficient by field rather than field cubed, or used a common E across both materials. Direct field checks at a selected bias make the numerical experiment physically interpretable.

Rotate the field relative to a microstructure

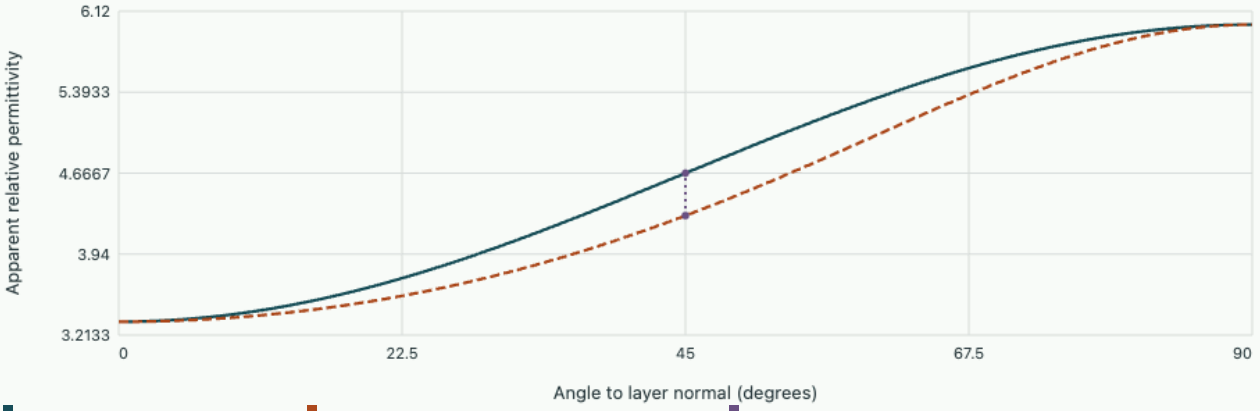
Study a layered material with calculable normal and tangential responses. Change its composition before deciding whether one scalar permittivity can describe both measurements.

Open this experiment online

Try the calculation

Fraction of low-permittivity phase **0.5**

Measurement angle (degrees) **45**



Fixed field direction | Fixed displacement direction | Selected angle

Principal relative permittivities: normal 3.3333, tangential 6. At 45° the two responses are 4.6667 and 4.2857.

Recover dielectric alignment torque from electrode work

Will positive or negative rotation increase capacitance?

A slab completely filling an ideal parallel-plate cell can rotate its uniaxial permittivity tensor reversibly without changing plate geometry. The angle θ between plate normal and the larger principal-permittivity axis is adjustable. Derive the torque on this ideal material coordinate, compare voltage and charge control, and calculate the maximum for a stated geometry.

▼ **Derive and check the result**

With area A, separation d, principal absolute permittivities $\epsilon_{\parallel} > \epsilon_{\perp}$, the normal capacitance is $C(\theta) = (A/d)(\epsilon_{\parallel}\cos^2\theta + \epsilon_{\perp}\sin^2\theta)$. For fixed voltage electrical coenergy is $CV^2/2$. The torque doing work in the direction of increasing θ is its derivative,

$$T = (A/d)(\epsilon_{\perp} - \epsilon_{\parallel})V^2\sin(2\theta)/2.$$

For $0 < \theta < \pi/2$ it drives the larger-permittivity principal axis towards the plate normal, increasing capacitance. The configurations at 0° and 90° both have zero torque but only the former minimises the fixed-voltage mechanical potential $-CV^2/2$.

Under fixed charge the stored energy is $Q^2/(2C)$, so the torque is $-\partial_\theta U = Q^2C'/(2C^2)$. At the same instantaneous voltage Q/C it agrees with the voltage-controlled result. Following two entire rotation curves with different preparations generally yields different magnitudes because the isolated cell voltage

varies with angle. This equality at the same state is an independent sign check and distinguishes the field torque from the external holding torque of opposite sign.

Take $A = 1.0$ cm², $d = 1.0$ mm, principal relative permittivities 6 and 2, and $V = 20$ V. At 45°, the magnitude is $0.5(A/d)4\epsilon_0 V^2 = 7.08 \times 10^{-10}$ N m. A small-angle torque measurement close to the aligned state would also reveal a positive angular restoring stiffness. In a real sample the crystal lattice, liquid-crystal elasticity or mount supplies additional energy and boundary conditions for rotation. The electrostatic contribution just calculated can be isolated by varying V^2 and checking the $\sin 2\theta$ angular dependence at sufficiently slow drive.

Tutorial · diagnose a tensor and a bias dependence

Recover a rotated tensor from three plate measurements

In a material plane, cuts with normals $x, y, (x+y)/\sqrt{2}$ yield relative plate permittivities $3, 5, 5$, respectively. Reconstruct the 2×2 tensor, check positivity and predict the result for the other bisector.

▼ **Work through the solution**

The first two values are

$$K_{xx} = 3, K_{yy} = 5.$$

The third is $(3+5)/2 + K_{xy} = 5$, so $K_{xy} = 1$. The symmetric matrix has determinant 14 and positive first diagonal entry, hence is positive definite. Its principal values are $4 \pm \sqrt{2}$. The other bisector gives $4 - 1 = 3$. This extra reading can test the assumed orientation and symmetry independently. All four inferences assume the plate experiment imposes electric field normal to each cut and fringing is negligible.

Distinguish charge measurements before integrating electrical work

At a bias E_b, a sample following the cubic constitutive law gives secant permittivity $1.30\epsilon_*$ and differential permittivity $1.70\epsilon_*$, where ϵ_* is a calibration permittivity. Infer its linear coefficient, cubic contribution at bias and true stored energy density.

▼ **Work through the solution**

Subtract the two measurements. The difference is $2\alpha E_b^2 = 0.40\epsilon_*$, so $\alpha E_b^2 = 0.20\epsilon_*$,

$$\epsilon = 1.10\epsilon_*.$$

The energy at bias is

$$[0.55 + 0.15]\epsilon_* E_b^2 = 0.70\epsilon_* E_b^2.$$

Using $ED/2$ would give $0.65\epsilon_* E_b^2$. To test the law rather than just fit two parameters, obtain a second set of secant and differential readings at a different bias and check the predicted quadratic change with bias field.

Further reading

- Inspect normal displacement across a dielectric interface
- Follow the optical consequences of a permittivity tensor
- Griffiths, 5th ed., §§4.3–4.4 · constitutive laws and dielectric work

From mobile ions to screening lengths and differential double-layer capacitance

Electrostatic field theory and measurement

A charged planar wall in an electrolyte can be studied without assuming a capacitance in advance. Diffusion and electrostatic drift first establish an ion cloud. The charge–voltage slope of that cloud is an experimental output of the model.

An electric field competes with entropy of mixing

Take a bulk reservoir containing mobile ions with charges $+e$, $-e$, each at number density n_b. The solvent is a uniform continuum of permittivity ϵ, the bath holds absolute temperature T, and no reactions or Faradaic currents occur at the wall. At thermal equilibrium the electrochemical potential of each species is spatially constant. For an ideal dilute gas of particles it contains $k_B T \ln(n_\pm / n_b)$ from mixing plus electrostatic energy $\pm eV$. Set the bulk potential to zero. Equating the local and bulk electrochemical potentials yields

$$n_\pm = n_b e^{\mp eV/k_B T}.$$

At positive V the negative counterions concentrate and the positive coions are depleted. Total mobile charge is

$$\rho = e(n_+ - n_-) = -2en_b \sinh(eV/k_B T).$$

Substituting into Poisson's equation and defining dimensionless $\psi = eV/(k_B T)$ gives

$$\nabla^2 \psi = \kappa^2 \sinh \psi,$$

$$\kappa^2 = \frac{2e^2 n_b}{\epsilon k_B T}.$$

The parameter n_b is a density of one species, not both together. For a 1:1 salt specified in mol/litre convert to formula units per cubic metre using $1000 N_A$ before substitution. For a mixture the linear screening coefficient generalises to $e^2 \sum_i z_i^2 n_i / (\epsilon k_B T)$, whereas the full nonlinear right-hand side remains a sum of exponentials. An externally maintained reservoir fixes chemical potentials; a sealed very small cell can instead require separate number conservation for every species.

What the small-potential limit predicts and what it neglects

For $|\psi| \ll 1$, replace $\sinh \psi$ by ψ. Above a planar wall at $z = 0$ with prescribed potential V_s and $V \to 0$ far away, the bounded solution is

$$V(z) = V_s e^{-\kappa z}.$$

With the normal into liquid $+\hat{\mathbf{z}}$, the wall displacement is

$$D_z(0^+) = -\epsilon V'(0) = \epsilon\kappa V_s.$$

For a conducting wall with no interior field this equals its free surface charge. Therefore $C_{\mathrm{diff}}/A = d\sigma/dV_s = \epsilon\kappa$, dimensionally a dielectric layer of thickness κ^{-1}. The ionic concentration actually varies continuously; it is not a plate of opposite charge located at that distance.

Integrating charge density $-\epsilon\kappa^2 V(z)$ over $z > 0$ gives

$$-\epsilon\kappa V_s = -\sigma.$$

The screening cloud and wall together are neutral per area. The fraction of compensating charge within distance h is $1 - e^{-\kappa h}$. At one screening length the potential has fallen to $1/e$ and 63.2% of the countercharge has accumulated. For a spherical bead curvature changes this simple relationship between potential ratio and enclosed charge. The ionic constitutive relation is the same, but the spatial Laplacian and collecting area differ.

With $\epsilon_r = 78.5,\ T = 298$ K and 1.00 mM of a fully dissociated symmetric monovalent salt, $n_b \simeq 6.02 \times 10^{23}$ m⁻³ and $\kappa^{-1} \simeq 9.62$ nm. The linear diffuse capacitance is about 0.072 F/m². Increasing concentration fourfold halves the screening length and doubles this capacitance at fixed solvent response. Close to molecular length scales the continuum assumptions and concentration-independent permittivity become progressively questionable, so the square-root law should be tested in the intended concentration range.

Background · How small is a small electrostatic potential?

In a symmetric electrolyte the mean charge term is proportional to $\sinh u$, with $u = ze\psi/(k_B T)$ measured relative to the neutral reservoir. A potential being small in volts gives no absolute criterion: the valence and temperature set the relevant scale. Taylor expansion yields the fractional error in replacing the ionic source by its linearised value,

$$\frac{\sinh u - u}{u} = \frac{u^2}{6} + O(u^4).$$

For roughly 1% local source accuracy this estimate asks for $|u| \lesssim 0.245$. At 298 K it gives about 6.3 mV for monovalent ions and 3.1 mV for divalent ions. This is a criterion on the local differential equation. An integrated charge or differential capacitance can have a different relative error and merits its own check.

Solve the nonlinear planar boundary problem

With planar symmetry $\psi'' = \kappa^2 \sinh\psi$. Multiply both sides by ψ' and integrate from the remote bulk where $\psi, \psi' \to 0$. This gives

$$\tfrac{1}{2}(\psi')^2 = \kappa^2(\cosh\psi - 1) = 2\kappa^2 \sinh^2(\psi/2).$$

Choose the branch $\psi' = -2\kappa \sinh(\psi/2)$, which decays to zero for either sign of surface potential. The identity $d \ln |\tanh(\psi/4)| / d\psi = [2 \sinh(\psi/2)]^{-1}$ turns the separated equation into

$$\tanh \frac{\psi(z)}{4} = \tanh \frac{\psi_s}{4} \, e^{-\kappa z}.$$

The physical potential itself is not a simple exponential at finite bias. Its distant tail has the same decay length but extrapolates back to a different amplitude $4 \tanh(\psi_s/4)$ in dimensionless units. A far-field slope can therefore recover κ, while a linear extrapolation of its intercept underestimates large surface potentials. The nonlinear solution reduces continuously to the linear one because $\tanh x \sim x$.

At $z = 0$, using $D_z = -\epsilon k_B T \psi' / e$ gives

$$\sigma(V_s) = \frac{2 \epsilon \kappa k_B T}{e} \sinh\left(\frac{e V_s}{2 k_B T} \right),$$

$$\frac{C_{\text{diff}}}{A} = \epsilon \kappa \cosh\left(\frac{e V_s}{2 k_B T} \right).$$

The first relation is odd in surface voltage and the differential capacitance is even for this symmetric salt model. Increasing voltage concentrates counterions and allows more charge per further voltage increment. For $V_s = 50$ mV at 298 K, $V_s / (2 k_B T / e) \simeq 0.973$. The differential capacitance is about 1.51 times its zero-bias value, while the secant ratio σ / V_s is only about 1.17 times that value. A small AC perturbation superposed slowly about a DC bias measures the differential slope if ionic transport remains equilibrated.

An independent mechanical check of the nonlinear field

The ideal mobile ions have an osmotic pressure excess relative to the reservoir $\Delta p = k_B T (n_+ + n_- - 2 n_b) = 2 n_b k_B T (\cosh \psi - 1)$. From the first integral the electrical normal stress magnitude $\epsilon E^2 / 2$ equals this excess. Thus $\Delta p - \epsilon E^2 / 2$ is spatially constant and vanishes in the bulk of the isolated-wall problem. One can derive the same result by differentiating the pressure: $d \Delta p / dz = \rho E$ at equilibrium. The cancellation provides a sensitive joint test of the ion distribution and the potential gradient.

Confine electrolyte between two equal positively biased parallel walls instead. By symmetry $E = 0$ at the midplane but ψ_m generally remains positive. The constant excess normal pressure is then $2 n_b k_B T (\cosh \psi_m - 1) > 0$. Overlap of the diffuse layers produces a repulsive osmotic force per area in this mean-field model. With large separation the midplane potential tends to zero and so does that force. A solution built by naively superposing two high-voltage nonlinear wall profiles will generally violate the constant-pressure test.

Separate a near-wall compact layer from the diffuse response

A finite exclusion distance for ion centres near a real electrode can be modelled at first by a compact dielectric layer of capacitance per area c_s. Its voltage drop is σ/c_s; the diffuse layer starts at potential V_d. For applied bias measured relative to the bulk and a chosen neutral reference,

$$V_{\mathrm{app}} = V_d + \sigma(V_d)/c_s.$$

Differentiate with respect to surface charge:

$$\frac{1}{c_{\mathrm{meas}}} = \frac{1}{c_s} + \frac{1}{c_d(V_d)}.$$

The diffuse constitutive relation must use V_d, not the entire applied voltage. For $c_s = 0.20$ F/m² and low-bias $c_d = 0.072$ F/m², the measured capacitance is about 0.053 F/m². Even as c_d grows, this simple series model cannot exceed c_s. It offers a physically distinct explanation for a measured saturation, alongside finite ion size, adsorption, dielectric variation and other model extensions. A static capacitance measurement alone cannot uniquely distinguish all of them.

Use layer overlap to infer a mechanical force independently of capacitance

Consider identical parallel electrodes at $z = \pm H$, held at the same small positive potential V_s relative to a salt reservoir in chemical equilibrium with the gap. Neither electrode is an ordinary reference for the other: a voltage meter connected only between them reads zero while each diffuse layer experiences a nonzero potential relative to the reservoir. Under the linear screening approximation the potential obeys $V'' = \kappa^2 V$. Reflection symmetry enforces $V'(0) = 0$. Matching both wall potentials gives $V(z) = V_s \cosh(\kappa z)/\cosh(\kappa H)$, so the midplane voltage is

$$V_m = V_s/\cosh(\kappa H).$$

The magnitude of free charge per area on each inner electrode face is $\epsilon\kappa V_s \tanh(\kappa H)$. At large H each plate recovers its isolated-wall charge. At small H the whole gap is nearly at V_s, so very little normal field is required to maintain it and each facing surface carries less charge. This behaviour differs from two oppositely biased plates, whose nearly uniform internal field grows as their separation decreases. The reservoir reference and voltage symmetry decide which experiment is being performed.

To leading nonzero order in V_s, the osmotic excess force per area between these identical walls can be found at the field-free midplane. There it is $2n_b k_B T(\cosh\psi_m - 1) \simeq n_b k_B T \psi_m^2 = \epsilon\kappa^2 V_s^2/[2\cosh^2(\kappa H)]$. Why use a term quadratic in V_s when the potential was found linearly? The pressure's zeroth-order excess vanishes and the first-order ionic density changes cancel between opposite species. Its first measurable correction is quadratic. The cubic error in the linear potential changes the computed pressure only at fourth order in V_s, making the mixed expansion consistent.

For $\kappa^{-1} = 10$ nm, $\epsilon = 78.5\epsilon_0$ and $V_s = 5.0$ mV, at half-gap $H = 10$ nm the predicted repulsive pressure is about 36.5 Pa. At $H = 20$ nm it drops to roughly 6.14 Pa. The potential rose by only a few millivolts yet its ionic osmotic consequence extends across the entire overlap region. As H becomes large, $\cosh^{-2}(\kappa H) \simeq 4e^{-2\kappa H}$, giving exponential decay with the full separation $2H$. This is a force measurement of the same screening scale measured independently by a voltage probe. Surface roughness, salt exclusion and adsorption can influence a real nanometre-scale force curve and should be distinguished from deviations in the bulk screening length.

If the electrode charges rather than their potentials are fixed, the wall derivative is prescribed. The even solution then has $V_m = \sigma/[\epsilon\kappa \sinh(\kappa H)]$ in the linear regime, and pressure $\sigma^2/[2\epsilon \sinh^2(\kappa H)]$. As H shrinks the linear approximation will eventually fail for fixed nonzero σ, because V_m grows. Comparing the constant-potential and constant-charge formulas without adjusting their boundary conditions could suggest contradictory predictions where each applies to a different experiment.

Estimate how quickly the ion cloud can follow a capacitance measurement

Static Poisson–Boltzmann theory assumes ions are in equilibrium at each measured voltage. A time-dependent measurement can test that assumption. For dilute symmetric monovalent ions with equal diffusivity D_i, small density perturbations obey the linearised number fluxes $\mathbf{j}_\pm = -D_i[\nabla \delta n_\pm \pm e n_b \nabla V/(k_B T)]$, combining diffusion with electric drift. Assume no bulk reactions. Subtract the two continuity equations and multiply by e, yielding a charge-current density $\mathbf{J} = -D_i[\nabla \rho + 2e^2 n_b \nabla V/(k_B T)]$. Use $\nabla^2 V = -\rho/\epsilon$ in the bulk to obtain

$$\partial_t \rho = D_i(\nabla^2 \rho - \kappa^2 \rho).$$

A spatial harmonic of wavenumber k decays at rate $D_i(k^2 + \kappa^2)$ in the unbounded linear bulk model. The long-wavelength Debye charge-relaxation time is $\tau_D = 1/(D_i \kappa^2)$, also recovered by dividing permittivity by the dilute ionic conductivity $\sigma_{\mathrm{ion}} = 2e^2 n_b D_i/(k_B T)$. The total salt-density perturbation instead obeys ordinary diffusion because the leading positive and negative drift terms cancel. Fast local charge relaxation and slower redistribution of neutral salt are different observables.

For $D_i = 1.0 \times 10^{-9}$ m²/s and $\kappa^{-1} = 10$ nm, this local time is 100 ns, corresponding to an angular rollover scale $1/\tau_D = 10^7$ s⁻¹ or ordinary frequency about 1.6 MHz. Blocking electrodes in a cell of thickness $L \gg \kappa^{-1}$ can display a slower charging time. Estimate its series resistance $R = L/(\sigma_{\mathrm{ion}} A)$ and diffuse capacitance $C \sim \epsilon\kappa A$ for one interface. Their product is of order $L/(D_i\kappa)$. For $L = 100$ µm with the preceding parameters this is about 1 ms, orders of magnitude above the local bulk time. Two identical interfaces in series supply an extra numerical factor; finite reservoirs and unequal mobilities can introduce further scales.

A low-frequency impedance experiment interpreted as the equilibrium $d\sigma/dV$ must therefore allow the actual circuit and cell to charge, not merely wait several local Debye times. Varying cell spacing helps distinguish a geometry-dependent charging time from an intrinsic screening-length change. This diffusion calculation and the static pressure calculation share the same κ, giving independent checks on a physical fit to voltage, force and time-resolved measurements.

Check reservoir and continuum assumptions using actual particle numbers

The grand-canonical ion profiles implicitly allow the diffuse layers to draw particles from a large reservoir without changing the reservoir's chemical potentials. Whether this is reasonable can be checked with a dimensioned inventory. For a 1 mM monovalent salt, each species has about 6.02×10^{23} particles per cubic metre. A cell of area 1 mm² and depth 100 µm therefore starts with 6.02×10^{13} ions of each species. A surface charge density of 1.0 mC/m² on one plate amounts to 6.24×10^{9} elementary charges on that area. Even a conservative estimate in which each net countercharge requires one ion from the bulk uses only about 10^{-4} of the available counterions.

Keep the same area and salt concentration but reduce the gap to 20 nm. Now only 1.20×10^{10} ions of each species occupy the cell initially. The same surface charge is comparable with half the inventory of one species. The remote-neutral-reservoir approximation is no longer automatically credible, especially if both walls attract counterions of the same sign. One must specify how ions and compensating charge can enter or leave, what the electrodes' chemical or electrical preparations enforce, and how much fluid actually belongs to the reservoir. An applied electrostatic boundary voltage fixes no ionic chemical potential by itself.

For a sealed cell write the equilibrium densities instead as $n_\pm(z) = A_\pm \exp[\mp eV(z)/(k_B T)]$. The constants $A_\pm$ must be chosen so that integrals of $n_\pm$ over the physical volume equal the known numbers of each species. Substitution into Poisson's equation gives a coupled boundary problem with two additional global constraints. Adding a constant to V and appropriately rescaling $A_\pm$ leaves the densities unchanged; electrical reference and chemical normalisation must be chosen consistently. A calculation holding both fixed bulk densities and exact sealed inventories can be overconstrained if diffuse layers hold appreciable salt.

The continuum approximation also makes a statement about scale. A cube of side one screening length near 10 nm at 1 mM holds roughly 0.60 ions of either species on average. This does not by itself invalidate a mean density for a large planar electrode, which averages over many lateral regions and thermal configurations. It does warn against interpreting the smooth one-dimensional density as a snapshot of a single nanometre-sized volume with a deterministic fractional particle. Force or voltage measurements with different lateral averaging areas can show quite different fluctuations even when their mean profiles follow the same equation.

As another self-check compare the predicted contact concentration with available molecular packing. The ideal theory gives

$$n_{\mathrm{counter}}(0) = n_b e^{|\psi_s|}.$$

Starting from 1 mM, a surface potential near 0.25 V at 298 K would multiply concentration by roughly 1.7×10^{4}, corresponding to tens of molar even before allowing room for solvent. A molecular model or a measured finite-size constitutive relation is then more appropriate for quantitative contact density. The failure is in extrapolating the dilute ideal entropy, not in Gauss's law or the electrostatic boundary condition. Source, solvent, reservoir and detector scales determine where the otherwise elegant analytic solution becomes an approximation to test.

Resolve a screening length

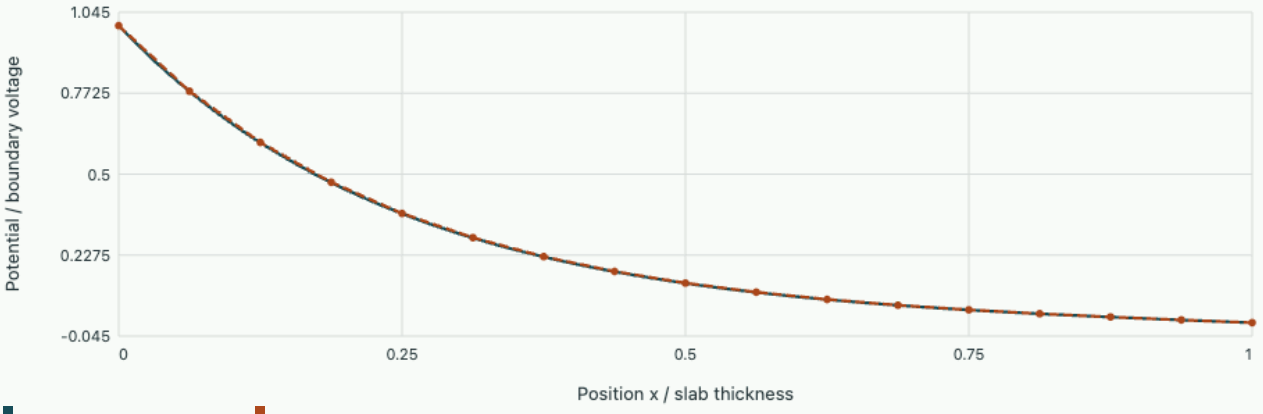
Compare the linear screened profile with a grid solution as the ratio of thickness to screening length grows. Separate failure to resolve the layer from failure of the linear material law.

Open this experiment online

Try the calculation

Cells across slab **16**

Gap / screening length **4**



Continuum slab Solved grid values

Midpoint 0.13357, exact 0.1329. Discrete decay constant × thickness = 3.9897.

Download the OneSimulation model

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

```
python field-laboratory.py screen --cells 16 --contrast 4
```

How many excess ions, not just how much net charge?

Count the salt adsorbed in a planar diffuse layer

The total countercharge per area equals the negative surface charge. Separately calculate the excess number of ions of both signs over the neutral bulk value for a symmetric electrolyte at prescribed surface potential. Which quantity changes sign if the wall voltage reverses?

▼ Derive and check the result

Define $\Gamma = \int_0^\infty [n_+(z) + n_-(z) - 2n_b]\,dz = 2n_b \int_0^\infty (\cosh\psi - 1)dz$. The integrand is nonnegative for either sign of wall voltage. For $\psi_s > 0$, substitute $dz = -d\psi/[2\kappa \sinh(\psi/2)]$ from the first integral. Then $\cosh\psi - 1 = 2\sinh^2(\psi/2)$ makes the resulting integral elementary. The result, also valid at

negative bias by symmetry, is $\Gamma = (4n_b/\kappa)[\cosh(\psi_s/2) - 1]$. It has units particles per area. The signed excess charge divided by elementary charge, in contrast, is

$$-\sigma/e = -(4n_b/\kappa)\sinh(\psi_s/2).$$

For positive surface charge the negative counterion excess and positive coion depletion follow by taking half the sum and difference of these two measurements. At small ψ_s, the net charge excess is linear in voltage while Γ is quadratic: the two ion species redistribute in opposite directions before appreciable neutral salt adsorption appears. Reversing the bias swaps which species is enriched but leaves the total ion-number excess unchanged. Thus a mass-sensitive and a charge-sensitive measurement give complementary responses even in the same equilibrium model.

At $\psi_s = 2$,

$$\Gamma/[|\sigma|/e] = [\cosh 1 - 1]/\sinh 1 = \tanh(1/2) = 0.462.$$

The total excess count can be smaller than the magnitude of countercharge measured in elementary charges because some compensation occurs through removing coions rather than adding counterions. An optical concentration measurement that tracks only total dissolved ion count cannot be converted into net countercharge by multiplying by e. Sampling both quantities, or calibrating the signal against the appropriate species, is needed to test the adsorption prediction. The derivation assumes the same dilute reservoir as the planar Poisson–Boltzmann solution and should not be applied unmodified to a cell with fixed small ion inventories.

Tutorial · infer a screening length and check the model

Use two potential measurements at finite bias

In a 298 K symmetric electrolyte, $\tanh[eV(z)/(4k_BT)]$ decreases from 0.30 to 0.15 over 8.0 nm. Infer the screening length, and reconstruct the two physical voltages. Why is taking a raw voltage ratio slightly different?

▼ Work through the solution

The transformed profile decays exponentially, so

$$\kappa^{-1} = 8.0/\ln 2 = 11.54\,\text{nm}.$$

Since $4k_BT/e \simeq 102.7$ mV, the voltages are $102.7\,\text{artanh}(0.30) = 31.8$ mV and $102.7\,\text{artanh}(0.15) = 15.5$ mV. Their physical ratio is about 0.488 rather than 0.500. Inferring a linear-log decay using only these points gives a biased length. Repeating at smaller voltages would test convergence towards the same result without assuming the nonlinear transformation.

Which capacitance was measured?

At low bias a diffuse model independently predicts 0.080 F/m² but a low-frequency measurement gives 0.050 F/m². Infer the compact capacitance of the ideal series model. What would make this inversion impossible with positive capacitances?

▼ **Work through the solution**

From the reciprocal difference $1/c_s = 20 - 12.5 = 7.5$ m²/F we obtain $c_s = 0.133$ F/m². A measured capacitance larger than the predicted diffuse one would imply a negative reciprocal compact capacitance in this two-positive-series model. Check the normalisation area, frequency response, adsorption or additional parallel charging processes before assigning a negative parameter. To distinguish the two capacitances experimentally, vary reservoir concentration at fixed temperature and compare the predicted $\sqrt{n_b}$ trend of the diffuse contribution with the remaining offset in reciprocal capacitance.

Further reading

- Compare planar screening to a spherical bead
- Cambridge Electromagnetism, chapter 6 · screening and electrostatics of mobile charge

Measure inertia and gravity with two suspension axes

Mechanics in the laboratory

A rigid pendulum can oscillate from either of two knife edges. If the periods agree at distinct distances from the centre of mass, the spacing of the edges determines gravity. Derive the cancellation, identify a second way equal periods can arise and find why a poorly chosen geometry magnifies timing errors.

Torque about an offset pivot

Take a rigid body of mass m with parallel suspension axes perpendicular to its plane of oscillation. Its moment of inertia about the centre of mass is $I_C = m\rho^2$; ρ is the radius of gyration. For a pivot a distance $a > 0$ from the centre, the parallel-axis theorem gives

$$I_a = m(\rho^2 + a^2).$$

Gravity exerts torque $-mga\sin\theta$. In the small-angle limit,

$$T_a^2 = \frac{4\pi^2}{g}\,\frac{\rho^2 + a^2}{a},$$

$$L_a = a + \frac{\rho^2}{a}.$$

Here L_a is the length of a point pendulum with the same period. It can lie beyond the physical centre of mass. A tap applied at this distance from a pivot gives the centre of percussion when the ideal instantaneous pivot reaction is to vanish. The same combination of mass distribution and geometry appears in an impulsive and in an oscillatory measurement because both depend on angular momentum about that pivot.

Match the periods and keep both algebraic branches

Put the second knife edge on the opposite side of the centre at distance b. Equal periods require

$$a + \rho^2/a = b + \rho^2/b.$$

Rearrange before dividing by $a - b$:

$$(a - b)\left(1 - \frac{\rho^2}{ab}\right) = 0.$$

Two equal distances always give equal periods. That branch provides no information about ρ. If the distances are distinct, equality implies

$$\rho^2 = ab.$$

Both equivalent lengths then equal $a + b$, the directly measurable edge separation. For $a = 0.200\,\mathrm{m},\ b = 0.450\,\mathrm{m},\ \rho = 0.300\,\mathrm{m}$, the equivalent length is 0.650 m and both periods at standard laboratory gravity are about 1.62 s. A small value of $a - b$ is an experimental warning sign, even if the observed equality is exceptionally clean.

Background · Can unequal release amplitudes imitate a mismatch between pivots?

A rigid pendulum about either knife edge obeys the same sine nonlinearity as a simple pendulum. To leading order in its release angle θ in radians, the measured period exceeds the small-amplitude period by the fraction $\theta^2/16$. At matched small-amplitude periods, measurements with amplitudes θ_a, θ_b therefore give

$$\frac{T_a - T_b}{T_0} \simeq \frac{\theta_a^2 - \theta_b^2}{16}.$$

Releasing at 6° about the first edge and 2° about the other creates a fractional mismatch of about 6.1×10^{-4}. For a two-second period this is 1.2 ms. Record the two amplitudes along with the timing data and reduce both periods to the same limit before moving the trim mass. Otherwise the trim can compensate a preparation error.

Extract gravity when the two measured periods differ

Multiply the first period equation by a and the second by b. Subtract to eliminate ρ^2. The exact small-angle result for unequal distances is

$$g = 4\pi^2 \frac{a^2 - b^2}{aT_a^2 - bT_b^2}.$$

At matched periods this reduces to the simple expression in terms of the edge spacing. The general formula involves a difference in its denominator. If a and b nearly coincide, the denominator is small and two independent timing errors subtract with poor relative precision. This is a property of the inverse measurement, even if computing each individual period is perfectly well conditioned.

Once g is known, use $\rho^2 = gaT_a^2/(4\pi^2) - a^2$ to recover the body inertia per unit mass. A negative value may indicate a wrong distance measured to the centre or a timing observable that is actually a half period. Verify the result using the second pivot and recover a separately predicted period after shifting a suspension point. A single successful cancellation is weaker than reproducing an additional measurement.

Move one knife edge in the experiment

Hold $a = 0.200\,\mathrm{m}$ fixed and vary both the radius of gyration and the second distance. The graph plots the period against the possible second distance. Before inspecting intersections, predict where the

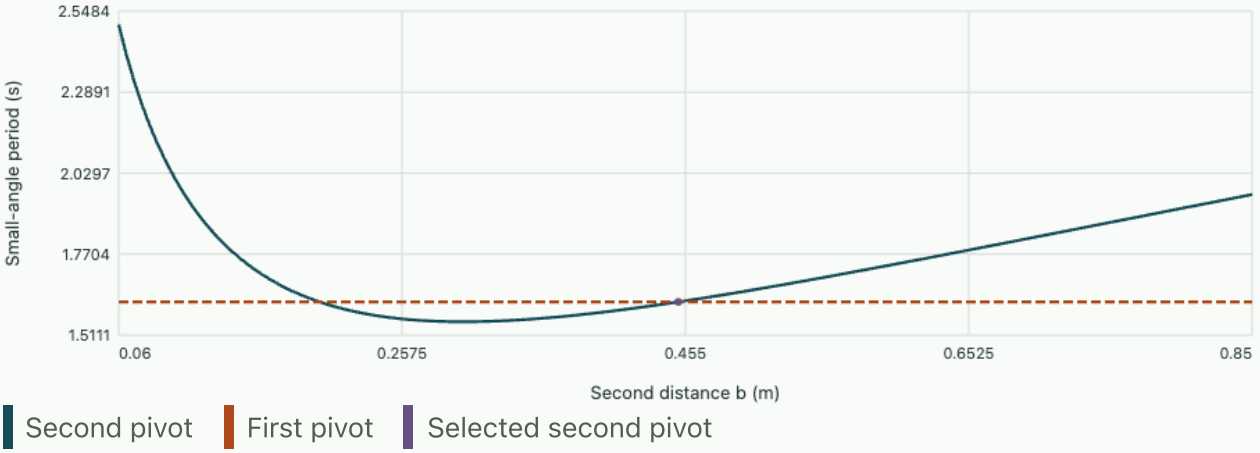
second match should lie as ρ grows. The dependence on distance has a minimum at $b = \rho$ because effective length comprises a growing and a decreasing term.

Open this experiment online

Try the calculation

Radius of gyration (m) **0.3**

Second pivot distance (m) **0.45**



First period 1.6173 s; second 1.6173 s. The other matching distance is 0.45 m.

With a finite oscillation amplitude, the physical pendulum has the same elliptic correction as a point pendulum of length L_a, provided the body is rigid and the frictionless pivot stays fixed. Matching different amplitudes can bias the cancellation. Measuring both at the same small angle removes the leading difference, and extrapolating both to zero amplitude provides a check. Rotational compliance of the knife edge and a shifting centre of mass require a different model.

Tutorial

For $a = 0.240$ m and $\rho = 0.300$ m, where should the second distinct match occur and what length should be used in $g = 4\pi^2 L/T^2$?

▼ **Work through the solution**

The second branch gives

$$b = 0.0900/0.240 = 0.375\,\mathrm{m}.$$

Because the pivots straddle the centre of mass the equivalent length is 0.615 m.

Both pivots lie on the same side of the centre, at the same distances a, b as in the previous question. What changes in the cancellation?

▼ **Work through the solution**

The periods depend on the magnitudes of the offsets, so equal periods still imply $\rho^2 = ab$ on the distinct branch. The sum $a + b$ still gives the equivalent length, but the physical ruler distance between pivots is $|a - b|$. The direct measurement shortcut relies on the chosen geometry.

Why does setting both axes exactly at the same distance make even noise-free measurements insufficient to determine both g and ρ?

▼ **Work through the solution**

The two equations are then identical. For every positive choice of inertia radius the first period determines a corresponding value of gravity. Improving timing precision cannot provide the missing independent equation.

Connect the offset-pivot calculation to the rigid-body inertia chapter.

Applied study

Bracket a root and decide when a Newton step is reliable.

Why solve an equation inside a time step?

A thermal mass cools by radiation. Its temperature obeys $\dot{T} = -\beta(T^4 - T_a^4)$. For $T > T_a$ the cooling rate decreases as the body cools. Backward Euler samples this rate at the *end* of the step. Calling the unknown end temperature x, write the discrete energy balance as a root problem

$$F(x) = x - T_n + h\beta(x^4 - T_a^4) = 0.$$

Assume a positive time step h, coefficient $\beta > 0$, and $T_n > T_a > 0$. The physical interval is $[T_a, T_n]$. Since $F(T_a) = T_a - T_n < 0$, $F(T_n) > 0$, and F is continuous, a root exists. Its derivative $1 + 4h\beta x^3$ is positive throughout this interval, so the root is unique. These facts suggest an algorithm even before choosing a starting guess.

Trade guaranteed progress for local speed

Bisection evaluates F at the midpoint and keeps the half containing a sign change. Each iteration halves the interval width. The midpoint error is at most half the remaining width, even when derivatives are unavailable.

Newton's method draws the tangent at x_j, $F(x_j) + F'(x_j)(x - x_j)$, and takes its zero $x_{j+1} = x_j - F(x_j)/F'(x_j)$. Near a simple root r, Taylor expansion about x_j gives

$$x_{j+1} - r = [F''(\xi)/(2F'(x_j))](x_j - r)^2.$$

If the derivative stays away from zero and the second derivative is bounded, this explains local quadratic convergence. Farther from the root, a practical safeguard is to replace a tangent step that leaves the bracket with bisection.

Residual and temperature error have a particularly simple relation here. For an approximation x in the physical bracket, the mean value theorem gives

$$|F(x)| = F'(\eta)|x - r| \geq |x - r|.$$

An absolute residual of 10^{-6} K certifies a root error no greater than 10^{-6} K for this scaled equation.

Follow two tangent steps

Set $T_n = 2$ K, $T_a = 1$ K, $h = 1$ s and $\beta = 0.0015$ K⁻³s⁻¹. This deliberately scaled thermal example keeps the arithmetic visible. Starting at $x_0 = 2$, the residual is $0.0015(16 - 1) = 0.0225$ K and the slope is

$$1 + 4(0.0015)2^3 = 1.048.$$

The first update gives $x_1 = 1.97853053$ K and $F(x_1) \simeq 1.65 \times 10^{-5}$ K.

Re-evaluating the slope at x_1 gives $x_2 = 1.97851479$ K, with residual about 8.7×10^{-12} K before rounding the iterate for display. This is the root for one backward Euler step. It still differs from the continuous-time temperature at one second because the rate varied during that second.

Problem class

Budget the bisections

Start from $[1, 2]$ K. After how many bracket halvings is the midpoint guaranteed to be within 10^{-5} K of the root without evaluating a derivative?

▼ Show the calculation

After n halvings, the width is 2^{-n} K. Midpoint error is bounded by $2^{-(n+1)}$ K. Thus

$$n \geq \lceil \log_2(1/(2 \times 10^{-5})) \rceil = 16.$$

Fifteen halvings still allow an error greater than the target.

Test a fixed-point iteration

Consider $g(x) = 2 - h\beta(x^4 - 1)$. For $h = 1$ s and $h = 25$ s, does the maximum slope magnitude on $[1, 2]$ certify contraction? What can a failed bound tell you?

▼ Show the calculation

The maximum $|g'|$ is $4h\beta\, 2^3 = 0.048h$ with h in seconds. Both maps take the bracket into itself: at one second the image is $[1.9775, 2]$, at 25 seconds it is $[1.4375, 2]$. At one second the maximum slope is 0.048, which certifies convergence from anywhere in the bracket. At 25 seconds the maximum is 1.2. The global contraction test is then inconclusive. Local convergence depends on the slope near the root.

ONESIMULATION LABORATORY

Locate the root of an implicit cooling step

Compare recorded OneSimulation runs on this page. To test your own inputs, download the Python model and run the model in an activated OneSimulation environment.

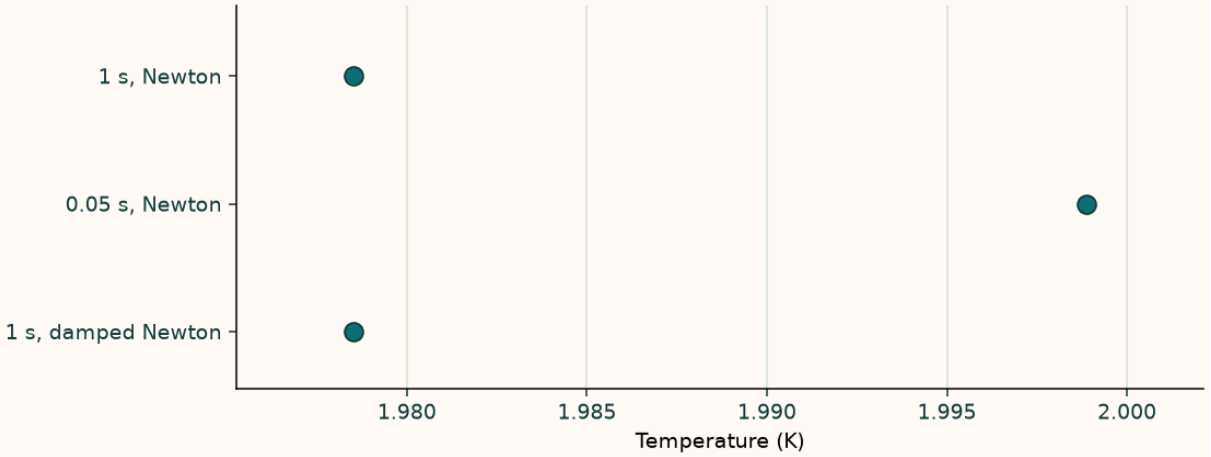
The downloadable model takes one backward Euler step with $C = 2$ J/K, $A = 0.3$ m², emissivity 0.5 and a test radiation constant of 0.02 W/(m²·K⁴), giving $\beta = 0.0015$ K⁻³s⁻¹. Its output substitutes the returned

temperature into the residual independently. Predict which edit changes the root. Download the Python experiment · OneSimulation implicit cooling comparison.

```
python nonlinear-equations.py
python nonlinear-equations.py --step 0.05
python nonlinear-equations.py --solver damped
```

End temperature and discrete residual from a single step.

Step, solver	Temperature (K)	|F| (K)
1 s, Newton	1.9785147906	5.6×10^{-17}
0.05 s, Newton	1.9988776913	3.7×10^{-17}
1 s, damped Newton	1.9785147906	5.6×10^{-17}



Damping the Newton iteration gives the same converged end temperature at a one-second step. Shortening the step changes the equation and gives a warmer endpoint.

A different step changes both the equation and the observation time. A converged line-search Newton method on the original step recovers the same physical root. To measure time discretisation error, compare solutions on successively finer grids at a common final time.

Change parameters · Change the laboratory inputs

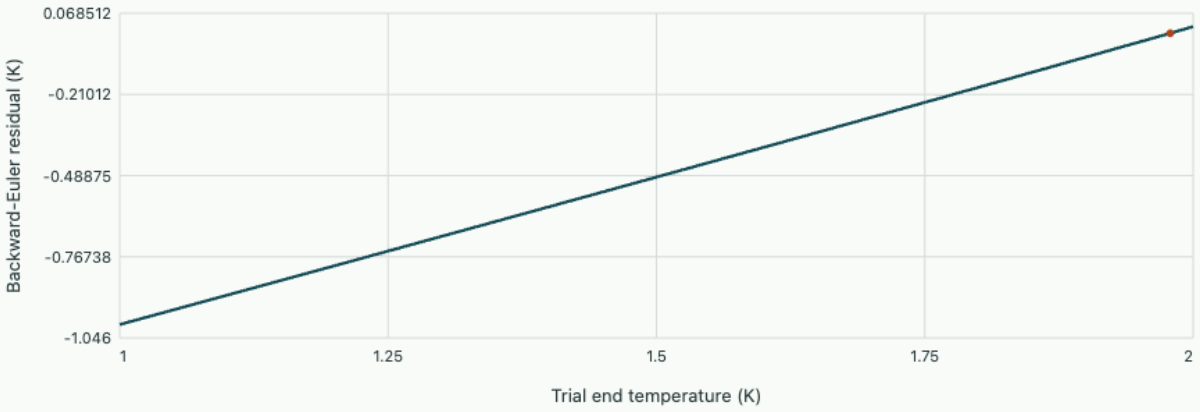
Change one input at a time and compare the recalculated curves with the argument above. The browser evaluates this small model locally; the accompanying numerical values state its assumptions and reference checks. Reset restores the reference inputs, and the data can be downloaded as CSV.

Open this calculation with adjustable inputs

Try the calculation

Implicit step duration (s) 1

Initial temperature (K) 2



Discrete cooling residual | Computed implicit root

End temperature 1.9785 K, residual -5.5511e-17 K. Coefficient 0.0015 combines the same capacity, area and radiation constants as the linked OneSimulation lumped model.

Problem sheet 2 · Distinguish root accuracy from temperature sensitivity

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.

Backward error

$$F(\hat{x}) + \delta F = 0$$

Specify how perturbations in data or equations are measured. A tiny perturbation making the computed value exact can coexist with a large solution error near a fold.

Derivation · Finding a turning point without using the load as a coordinate

Bracket

$$f(a)f(b) < 0$$

Continuity is needed to infer a zero between endpoints. A root of even multiplicity may exist without a sign change.

Derivation · Existence, brackets and error certificates

Complementarity

$$x \geq 0, \quad y \geq 0, \quad xy = 0$$

Both quantities are nonnegative and at least one vanishes. Active-set changes can make a smooth Newton model inappropriate across the switch.

Derivation · Inequalities and active constraints

Contraction

$$\|G(x) - G(y)\| \leq q\|x - y\|, \quad q < 1$$

The map must also keep the chosen complete region invariant to obtain the standard fixed-point guarantee.

Derivation · Fixed-point iteration and contraction

Newton correction

$$J(x_k)s_k = -F(x_k)$$

The correction solves a local linear model. Quadratic local convergence needs a suitable nonsingular derivative and a sufficiently close starting point.

Derivation · Newton’s method for coupled balances

Trust radius

$$\|s\| \le \Delta$$

The local model is minimised on this region. Agreement between predicted and achieved reduction guides the next radius.

Derivation · Deciding how far to trust a linearised residual

Further reading

- Chapter 4 develops scalar and vector rootfinding. Compare Newton and secant iterates at a simple root, then repeat at a multiple root to expose the change in convergence order. Driscoll and Braun · Fundamentals of Numerical Computation

Compare specific topics with the university sources