PHI TECHNICAL MONOGRAPHS

Signals, Systems and Communication Theory

Analytic models, implementation constraints and physical optimization



PHI research collective

Abstract

This volume develops signal representations, transform identities, filter synthesis, finite-precision models, random channels and constrained communication optimization. Its 110 bibliographically identified case studies connect exact mathematics to realizations and numerical experiments. Subsequent chapters examine passivity, conformal field transformations, parameter uncertainty and implementation-aware guarantees.

Each physical calculation specifies its normalization, units, boundary or initial conditions, admissible signals and power constraints. Frequency-response samples, whole-band polynomial certificates, state-machine properties and physical measurements have distinct evidential roles. Four source-defined legacy development-environment laboratories are supplied as procedures and mathematical acceptance models; the original proprietary target sessions remain unperformed.

The rate simulations are recorded computational experiments under the stated Gaussian models. They do not constitute measurements of an unspecified device. Whole-band FIR certificates establish the delivered coefficient arrays' inequalities; minimum FIR order is not asserted. Public illustrations are authored constructions. Bibliographic references identify the original exercise sources.

Reproduction material. The public source package contains the LaTeX manuscript and authored vector assets required to rebuild this volume. Its executable supplement covers selected numerical studies and certificate calculations. External textbook page reproductions and original externally acquired datasets are not distributed. References identify the scholarly works needed for independent source comparison.

Evidence convention. Established mathematical results retain their scholarly attribution. The exposition states the assumptions of each calculation and comparison. Recorded numerical results retain their original experimental scope. Formal specifications describe prospective proof developments unless an accepted proof object is explicitly identified. This public edition does not establish new proof-assistant kernel acceptance.

Contents

Abstract **i**

1 An Introduction to Digital Signal Processing: A Focus on Implementation **1**

1.1 SC-001 | Sampling a three-tone signal 1
1.2 SC-002 | Periods of sampled harmonics 2
1.3 SC-003 | Exact finite sinusoidal energy 2
1.4 SC-004 | Three Fourier invariants without a transform 3
1.5 SC-005 | A half-length frequency shift 4
1.6 SC-006 | DFT invariants of an eight-sample sequence 5
1.7 SC-007 | A complete four-point circular convolution 6
1.8 SC-008 | A circular shift in time 6
1.9 SC-009 | Four-point decimation-in-time FFT 7
1.10 SC-010 | Inverse DFT using a forward transform 8
1.11 SC-011 | A transform program for arbitrary lengths 9
1.12 SC-012 | Three discrete-time Fourier transforms 10
1.13 SC-013 | Time reversal of a real sequence 11
1.14 SC-014 | Z transforms with shifted one-sided support 11
1.15 SC-015 | A one-sided exponentially weighted cosine 12
1.16 SC-016 | An empty region of convergence 13
1.17 SC-017 | Inverse Z transform by residues 14
1.18 SC-018 | Inverse Z transform by partial fractions 14
1.19 SC-019 | A repeated pole and an impossible stated region 15
1.20 SC-020 | The first five impulse samples by long division 16
1.21 SC-021 | Transfer function of a four-tap impulse response 16
1.22 SC-022 | Moving-average impulse and transfer functions 17
1.23 SC-023 | Inverse transforms of real and imaginary spectral parts 17
1.24 SC-024 | Conjugate symmetry for a real DTFT 18
1.25 SC-025 | Real-sequence DFT symmetry 18
1.26 SC-026 | An inconsistent real-signal spectrum 19
1.27 SC-027 | Linear phase of a symmetric five-tap sequence 20
1.28 SC-028 | A linear-phase assertion contradicted by the printed 20
1.29 SC-029 | Four Fourier identities for complex sequences 21
1.30 SC-030 | Summability and energy of a one-sided exponential 22

1.31 SC-031 | Moving-average system properties 23
1.32 SC-032 | Backward-difference differentiator 24
1.33 SC-033 | Necessary and sufficient BIBO stability 24
1.34 SC-034 | A three-state canonical cascade with proved range scaling 25
1.35 SC-035 | Numerical integration as a difference equation 26
1.36 SC-036 | Bilinear transform of a rational second-order filter 27
1.37 SC-037 | Second-order digital Butterworth lowpass 28
1.38 SC-038 | Second-order digital Butterworth highpass 29
1.39 SC-039 | Sampling-parametric digital elliptic highpass 31
1.40 SC-040 | Sampling-parametric digital Butterworth bandpass 33
1.41 SC-041 | Sampling-parametric digital Chebyshev II bandstop 36
1.42 SC-042 | Impulse invariance of a causal exponential 39
1.43 SC-043 | Trapezoidal integration and the bilinear mapping 40
1.44 SC-044 | Backward-difference stability and its nonunit-circle frequency curve . . . 40
1.45 SC-045 | Highpass impulse invariance: an obstruction in the requested comparison 41
1.46 SC-046 | Butterworth transition, discrimination and degree 42
1.47 SC-047 | A second-order analog Butterworth prototype in physical units 44
1.48 SC-048 | Seventh-order Chebyshev I lowpass filter 45
1.49 SC-049 | Fifth-order elliptic lowpass filter 47
1.50 SC-050 | Highpass elliptic synthesis and the printed music-frequency units 50
1.51 SC-051 | A 455 kHz intermediate-frequency bandpass 52
1.52 SC-052 | A narrow analog power-line rejection filter 54
1.53 SC-053 | A continuously certified 101-tap FIR highpass 56
1.54 SC-054 | A continuously certified 101-tap FIR bandpass 59
1.55 SC-055 | Binary fractions in three signed representations 62
1.56 SC-056 | Subtraction by signed binary addition 63
1.57 SC-057 | Fixed-point multiplication with repeating operands 64
1.58 SC-058 | Finite dynamic range in IEEE binary formats 65
1.59 SC-059 | Aligned floating-point addition and guard bits 66
1.60 SC-060 | Four views of a factored IIR response 66
1.61 SC-061 | DFTs of conjugate and real/imaginary components 70
1.62 SC-062 | Four Z-transform operations with regions 70
1.63 SC-063 | Pole-zero analysis and a nonlinear printed recurrence 71
1.64 SC-064 | Direct and cascade realizations of a cubic FIR section 72
1.65 SC-065 | Canonical direct and cascade IIR realizations 73
1.66 SC-066 | Parallel first- and second-order realization 74
1.67 SC-067 | A 40-to-16-bit negative fixed-point truncation 75
1.68 SC-068 | Mantissa truncation from 32 to 24 bits 76
1.69 SC-069 | Pole locations under four-bit coefficient truncation 76
1.70 SC-070 | CCS introductory laboratory: a reproducible acceptance specification . 79
1.71 SC-071 | CCS project, debugger, visualization and profiling laboratory 79
1.72 SC-072 | DSP/BIOS scheduling and I/O laboratory 80

1.73 SC-073 | Visual Linker and memory placement 81

2 Fundamentals of Wireless Communication 83

2.1 SC-074 | The MMSE output as a sufficient statistic 83
2.2 SC-075 | Orthogonal observations of white Gaussian noise 84
2.3 SC-076 | Complex covariance and pseudo-covariance 85
2.4 SC-077 | Circular symmetry of a Gaussian vector 85
2.5 SC-078 | Real and imaginary parts of a circular Gaussian scalar 86
2.6 SC-079 | Which unitary matrices preserve an improper Gaussian law? 87
2.7 SC-080 | Sufficiency of a complex matched-filter statistic 87
2.8 SC-081 | Maximum discrete entropy and uniqueness 88
2.9 SC-082 | Jensen inequality and conditional entropy 89
2.10 SC-083 | When mutual information is additive 90
2.11 SC-084 | Maximum differential entropy at fixed second moment 90
2.12 SC-085 | Maximum entropy on a nonnegative half-line 91
2.13 SC-086 | Entropy maximization with several moment constraints 92
2.14 SC-087 | Real-vector and complex-scalar maximum entropy 92
2.15 SC-088 | Maximum entropy with fixed complex covariance 93
2.16 SC-089 | Optimal MISO beamforming under total power 94
2.17 SC-090 | Mutual information under an isotropic MISO input 95
2.18 SC-091 | Joint time–frequency waterfilling 95
2.19 SC-092 | Reciprocity of MIMO capacity 96
2.20 SC-093 | Gaussian ellipsoid concentration and volume 97
2.21 SC-094 | Uncorrelated fading does not guarantee diagonal optimal covariance . . 98
2.22 SC-095 | Orthogonal access versus simultaneous two-user transmission 99
2.23 SC-096 | User count, sum capacity and symmetric capacity 100
2.24 SC-097 | Low-SNR comparison with orthogonal access 106
2.25 SC-098 | Grouped SDMA between fully orthogonal and fully simultaneous access 106
2.26 SC-099 | Coupled waterfilling from a rank-one resolvent 108
2.27 SC-100 | Both corner points of a two-user Gaussian MIMO MAC 109
2.28 SC-101 | Isotropic covariance maximizes every fast-fading MAC constraint 110
2.29 SC-102 | Finite-memory Gaussian channels and the full Toeplitz limit 111
2.30 SC-103 | Fixed-association power control and its minimal feasible vector 113
2.31 SC-104 | Perron–Frobenius terminology and a feasibility equivalence 114
2.32 SC-105 | Joint association and componentwise power minima 115
2.33 SC-106 | Convergence of synchronous minimum-interference iteration 116
2.34 SC-107 | Totally asynchronous power updates 116
2.35 SC-108 | Two-user examples, trajectories, and association counterexamples 117
2.36 SC-109 | The chi-square density and its recurrence 119
2.37 SC-110 | Sphere hardening with a controlled shell approximation 120

3 Analytic response, passivity and branch selection **122**
3.1 All-pass filters: a phase budget from a proper map . . . 122
3.2 From upper-half-plane maps to positive-real impedances . . . 123
3.3 A stability certificate with a simpler independent baseline . . . 124
3.4 Operating branches: a fold is not a numerical convention . . . 125
3.5 Transmission lines and label-free propagation . . . 126
3.6 Energy, cascades and implemented state evolution . . . 127

4 Explicit circuits and finite implementations **129**
4.1 EU1: an LC calculation with dimensions intact . . . 129
4.2 EU2: a filter with constant magnitude and nonconstant delay . . . 130
4.3 EU3: a cubic stability boundary derived from an imaginary root . . . 130
4.4 EU4: a square root can simplify when zeros have even order . . . 131
4.5 EU5: sampling loses a logarithmic label . . . 131
4.6 EU6: a matrix response can avoid a mode label . . . 132
4.7 EU7: rounded implementation and a measurable error floor . . . 132
4.8 A six-part laboratory record . . . 133
4.9 Undergraduate comparison laboratory: a model is not its implementation . . . 133

5 Energy, cascades and switched models **136**
5.1 EG1: a lossless cascade has an additive storage function . . . 136
5.2 EG2: a paired-pole rational map gives a Foster-type network . . . 136
5.3 EG3: robust cubic stability over a coefficient box . . . 137
5.4 EG4: a transmission matrix without square-root cancellation . . . 137
5.5 EG5: annular energy under a degree-d cover . . . 138
5.6 EG6: a periodic system is governed by its product . . . 138
5.7 EG7: a two-layer certificate for quantized state evolution . . . 139
5.8 Graduate submission standard . . . 139
5.9 Graduate comparison: realization, whole-domain bounds and complexity . . . 139

6 Robust response and implementation guarantees **144**
6.1 EP1: uniform delay bounds over an all-pass parameter set . . . 144
6.2 EP2: finite data define a feasible analytic class . . . 145
6.3 EP3: the uncertainty disk at an unmeasured point . . . 145
6.4 EP4: passive square-root continuation under uncertainty . . . 146
6.5 EP5: robust stability is about the declared uncertainty set . . . 146
6.6 EP6: finite-state refinement of an analytic certificate . . . 147
6.7 EP7: translate mathematical error to the requested observable . . . 147
6.8 Research acceptance criteria . . . 148
6.9 Research comparison: uncertainty sets instead of fitted certainty . . . 148

7 Conformal fields, geometric uncertainty and response compression 152
7.1 E01. Conformal transformation of electrostatic potential 152
7.2 E02. Energy invariance and capacitance per unit length 153
7.3 E03. Boundary certification and interior field errors 154
7.4 E04. Clearance, conformal radius, and geometric uncertainty 154
7.5 E05. Bounded analytic response functions and sensitivity 155
7.6 E06. Analytic response compression and noise 156
7.7 E07. Model validation at singular and discretized limits 156

8 Field scales and explicit energy 158
8.1 EU01. Voltage coordinates and physical length 158
8.2 EU02. Energy of a coaxial cross-section 159
8.3 EU03. Impedance and reflection coordinates 159
8.4 EU04. Certifying unsampled boundary angles 160
8.5 EU05. A clearance estimate from a conformal radius 161
8.6 EU06. Exponential weighting of a causal response 162
8.7 EU07. A field singularity with finite energy 162
8.8 EU08. Boundary conditions select the physical problem 163

9 Wave coordinates and spectral constraints 165
9.1 EG01. Complex potential error and harmonic gradient error 165
9.2 EG02. The transformed conductivity equation 166
9.3 EG03. Disk geometry and response sensitivity 166
9.4 EG04. Two measurements and an analytic feasibility test 167
9.5 EG05. Geometric approximation error in a field certificate 168
9.6 EG06. Green functions as normalized geometric observables 169
9.7 EG07. Finite-frequency data and global response ambiguity 170
9.8 EG08. Conformal transformation of a wave equation 170

10 Uncertainty, matrix response and model restrictions 172
10.1 EP01. Energy forms and admissible coordinate transformations 172
10.2 EP02. Geometric uncertainty in an evaluation functional 173
10.3 EP03. Approximation rank and realization constraints 174
10.4 EP04. A local sampling certificate with uncertain node positions 175
10.5 EP05. Collar geometry and field-error allocation 175
10.6 EP06. Corner rounding and observable convergence 176
10.7 EP07. Multiport analytic models and finite-dimensional compactness 177
10.8 EP08. A physical certificate and its semantic premises 178

References 179

Chapter 1

An Introduction to Digital Signal Processing: A Focus on Implementation

Stanley H. Mneney. Reference problems and derived analyses.

1.1 SC-001 | Sampling a three-tone signal

Reference: [1], exercise 1.7.1.

Source task and conventions

Determine an admissible sampling rate, the sampled sequence, and the aliases when the angular sampling frequency is $10\Omega_0$.

Conventional solution and proof

Write $\Omega_s = 2\pi f_s$ and $T = 2\pi/\Omega_s$. The largest angular frequency is $11\Omega_0$, so an alias-free uniform sampling design for an arbitrary signal in this band requires

$$\boxed{\Omega_s > 22\Omega_0, \qquad f_s > 11\Omega_0/\pi.}$$

Strict inequality avoids the endpoint ambiguity at Nyquist. The sampled sequence is

$$x[n] = 3\sin(7\Omega_0 T n) + 5\sin(5\Omega_0 T n) + 7\sin(11\Omega_0 T n).$$

At $\Omega_s = 10\Omega_0$, the three digital angular frequencies are $7\pi/5, \pi, 11\pi/5$. For every integer n, $\sin(\pi n) = 0$, $\sin(7\pi n/5) = -\sin(3\pi n/5)$ and $\sin(11\pi n/5) = \sin(\pi n/5)$. Thus

$$\boxed{x[n] = -3\sin(3\pi n/5) + 7\sin(\pi n/5).}$$

The second component is lost and the first and third are aliased. If the symbol “sampling frequency” denotes cycles per second, the value $10\Omega_0$ must first be converted; angular and cyclic units cannot be equated.

Structural or analytic derivation

The sampling map sends an analog exponential $e^{i\Omega t}$ to the sequence $(e^{i\Omega T})^n$. Two frequencies are indistinguishable exactly when their difference is an integer multiple of $2\pi/T$: equality at $n = 1$ proves necessity and integer powers prove sufficiency. Reducing the three frequencies modulo this lattice proves every alias above without drawing overlapping spectra. For a band-limited function whose spectrum is supported in $[-B, B]$, sampling with $2\pi/T > 2B$ places its translated spectral copies on disjoint intervals, which is the frequency-domain injectivity condition behind the recommended rate.

Comparative cost and preferred method

Both a careful spectral-folding baseline and the exponential reduction use three frequency reductions. For m specified tones the work is $O(m)$, matching the need to inspect all frequencies. The analytic representation supplies an exact alias identity; no runtime advantage over a competent conventional calculation is asserted.

1.2 SC-002 | Periods of sampled harmonics

Reference: [1], exercise 1.7.2.

Source task and conventions

Find the periods of $2\sin(\omega_0 n)$, $3\sin(3\omega_0 n)$ and $4\cos(4\omega_0 n)$ and their sum.

Conventional solution and proof

A nonzero sampled exponential has period N precisely when $\omega N \in 2\pi\mathbb{Z}$. Suppose $\omega_0/(2\pi) = p/q$ in lowest terms, with $q \geq 1$. The individual exponential periods are q, $q/\gcd(q, 3)$ and $q/\gcd(q, 4)$. A sampled sine with reduced denominator one or two is identically zero and is assigned fundamental period one; all other sine periods equal their exponential periods. The cosine has period $q/\gcd(q, 4)$, including the constant case. The sum has the least common multiple of the nonzero-component periods. Here the unequal sine coefficients cannot cancel an aliased sine pair, and a real cosine coefficient cannot cancel an imaginary sine coefficient in the exponential expansion. Consequently its fundamental period is q for $q > 2$, and one for $q = 1, 2$. If $\omega_0/(2\pi)$ is irrational, all three components and their sum are aperiodic. Thus numerical periods cannot be given without ω_0.

Structural or analytic derivation

Expand each sine and cosine into the finitely many exponentials λ^n. Distinct exponentials are linearly independent: evaluating a linear relation at $n = 0, \ldots, r - 1$ gives a Vandermonde matrix with determinant $\prod_{j<k}(\lambda_k - \lambda_j) \neq 0$. Periodicity of the sum therefore requires $\lambda^N = 1$ for each nonzero combined coefficient. This establishes the least-common-multiple statement, including aliases and degenerate zero sequences.

Comparative cost and preferred method

For rational input p/q, Euclid's algorithm and least common multiples replace an unbounded search over candidate periods. The bit-operation cost is polynomial in $\log q$; a direct search through $1, \ldots, q$ has q candidates. The usual rational-frequency criterion already achieves the efficient method. Irrationality is a mathematical input property, not something certifiable from a finite floating-point decimal.

1.3 SC-003 | Exact finite sinusoidal energy

Reference: [1], exercise 1.7.11.

Source task and conventions

Compute $E_N = \sum_{n=0}^{N-1} \sin^2(\omega_0 n)$.

Conventional solution and proof

Using $\sin^2 u = (1 - \cos 2u)/2$,

$$E_N = \frac{N}{2} - \frac{1}{2}\Re \sum_{n=0}^{N-1} e^{2i\omega_0 n}.$$

If $\sin \omega_0 \neq 0$, multiplying the geometric sum by $1 - e^{2i\omega_0}$ and simplifying yields

$$\boxed{E_N = \frac{N}{2} - \frac{\sin(N\omega_0)\cos((N-1)\omega_0)}{2\sin\omega_0}.}$$

If $\omega_0 \in \pi\mathbb{Z}$, every sample vanishes and $E_N = 0$. The empty sequence $N = 0$ also has energy zero. For an integral number of nondegenerate half-periods, $\sin(N\omega_0) = 0$ and $E_N = N/2$.

Structural or analytic derivation

The identity is coefficient orthogonality on the unit circle. Writing the finite sum as $e^{i(N-1)\omega_0} \sin(N\omega_0)/\sin\omega_0$ explains both the energy and its exceptional denominator. Near a zero of $\sin\omega_0$, a scaled sinc evaluation or a series expansion avoids cancellation; the limiting formula alone is not a stable floating-point algorithm.

Comparative cost and preferred method

Direct summation uses N sine evaluations and N squares. The closed expression uses a fixed number of elementary-function evaluations for parameterized input. When the samples themselves are the input, reading them still costs $\Omega(N)$; no sublinear claim applies to arbitrary data.

1.4 SC-004 | Three Fourier invariants without a transform

Reference: [1], exercise 2.6.4.

Source task and conventions

For $x[-1], \ldots, x[4] = (1, 2, 1, 0, 1, 3)$, evaluate the DC value, the integral of the DTFT, and its squared-magnitude integral.

Conventional solution and proof

At zero frequency, $X(1) = \sum_n x[n] = 8$. For every integer m, $\int_{-\pi}^{\pi} e^{im\omega}\, d\omega$ is 2π if $m = 0$ and zero otherwise, as direct integration shows. Hence

$$\int_{-\pi}^{\pi} X(e^{i\omega})\, d\omega = 2\pi x[0] = 4\pi.$$

Expanding $|X|^2$ as a finite double sum and integrating kills all terms with unequal indices, giving

$$\boxed{\int_{-\pi}^{\pi} |X(e^{i\omega})|^2\, d\omega = 2\pi(1 + 4 + 1 + 0 + 1 + 9) = 32\pi.}$$

This is a complete finite proof of Parseval for this sequence.

Structural or analytic derivation

The three requested outputs are respectively evaluation at one, the constant Laurent coefficient, and the coefficient inner product. The index origin is essential: the constant coefficient is two, not the first listed value one.

Comparative cost and preferred method

The values require one six-term sum, one indexed lookup and one six-term sum of squares. Constructing a frequency grid and applying quadrature adds an arbitrary grid-size cost and approximation error. The exact coefficient method is both cheaper and exact for the requested invariants.

1.5 SC-005 | A half-length frequency shift

Reference: [1], exercise 3.6.5.

Source task and conventions

Recover $v[n]$ from $V[k] = W[(k-4)_8]$ and $w = (2, 0, 1, 3, 2, 1, 4, 1)$.

Conventional solution and proof

By the inverse DFT and the substitution $r = (k-4)_8$,

$$v[n] = \frac{1}{8}\sum_r W[r]e^{2\pi i(r+4)n/8} = e^{i\pi n}w[n] = (-1)^n w[n].$$

Thus

$$\boxed{v = (2, 0, 1, -3, 2, -1, 4, -1).}$$

Every index, including the wrap at $k = 0$, is covered by the periodic exponential identity.

Structural or analytic derivation

The shift of frequency coordinates is the character $n \mapsto (-1)^n$ acting on the original samples. It requires sign changes only; no root calculation or transform evaluation remains.

Comparative cost and preferred method

Exactly the odd-index entries change sign. For a general even length N, a half-spectrum shift takes at most $N/2$ negations, versus two transforms for a direct frequency-domain reconstruction. This achieves linear output complexity.

1.6 SC-006 | DFT invariants of an eight-sample sequence

Reference: [1], exercise 3.6.6.

Source task and conventions

Evaluate $X[0]$, $X[4]$, $\sum_k X[k]$ and $\sum_k |X[k]|^2$ for the given w, treating X as its DFT.

Conventional solution and proof

Directly from the definition,

$$X[0] = 2 + 0 + 1 + 3 + 2 + 1 + 4 + 1 = 14,$$
$$X[4] = 2 - 0 + 1 - 3 + 2 - 1 + 4 - 1 = 4.$$

Summing the DFT over k and using roots-of-unity orthogonality gives $\sum_{k=0}^{7} X[k] = 8w[0] = 16$. Expanding the squared magnitude and using the same identity gives

$$\boxed{\sum_{k=0}^{7} |X[k]|^2 = 8\sum_n |w[n]|^2 = 8(36) = 288.}$$

The source changes the name W to X and prints a mismatched summation index; the intended eight frequency values are specified here.

Structural or analytic derivation

The outputs are evaluation at the trivial and alternating characters, coefficient extraction at zero, and the unitary-norm identity. All are invariants of the DFT matrix and can be proved without computing its other six entries.

Comparative cost and preferred method

Two eight-term linear sums, one lookup and eight squares suffice. Computing a whole DFT produces outputs not requested. Both the conventional property-based solution and the analytic character method have $O(N)$ input work and exact equality.

1.7 SC-007 | A complete four-point circular convolution

Reference: [1], exercise 3.6.3.

Source task and conventions

Compute both DFTs, their product, its inverse and the circulant-matrix convolution for $x = (1, 2, 0, 1)$ and $h = (2, 1, 0, 1)$.

Conventional solution and proof

With $e^{-2\pi i/4} = -i$,

$$X = (4, 1 - i, -2, 1 + i), \quad H = (4, 2, 0, 2), \quad Y = (16, 2 - 2i, 0, 2 + 2i).$$

The inverse formula $y[n] = \frac{1}{4} \sum_{k=0}^{3} Y[k] e^{2\pi ikn/4}$ gives

$$\boxed{y = (5, 5, 3, 3).}$$

The circular convolution matrix is

$$C_h = \begin{pmatrix} 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{pmatrix}, \qquad C_h x = (5, 5, 3, 3)^T.$$

Each row implements $\sum_m h[(n - m)_4] x[m]$, verifying part (iv) independently.

Structural or analytic derivation

For any finite length N, expand the product of the two DFT sums and apply $N^{-1} \sum_k e^{2\pi ik(n-r)/N} = 1$ if $n \equiv r$ and zero otherwise. This proves that inverse-transform multiplication is circular convolution. For these short sparse vectors, the circulant product is the simplest check. Linear convolution would require padding to at least seven and is a different output.

Comparative cost and preferred method

Dense circular convolution uses N^2 scalar products; two forward FFTs, N products and one inverse FFT use $O(N \log N)$. At $N = 4$ the constants and zeros favour hand calculation. The exact arrays above, not an asymptotic claim, certify output equivalence.

1.8 SC-008 | A circular shift in time

Reference: [1], exercise 3.6.4.

Source task and conventions

Find the five-point DFT of $g[n] = h[(n - 2)_5]$ from the given spectrum.

Conventional solution and proof

Put $m = (n - 2)_5$. Periodicity of the exponential gives

$$G[k] = \sum_n h[(n - 2)_5] e^{-2\pi ikn/5} = e^{-4\pi ik/5} H[k].$$

Consequently the complete answer is

$$\boxed{G = (1+2i,\ (\tfrac{1}{2}-i)e^{-4\pi i/5},\ 3e^{-8\pi i/5},\ (\tfrac{1}{2}+i)e^{-12\pi i/5},\ (1-2i)e^{-16\pi i/5}).}$$

No reality assumption on h is needed, and the supplied spectrum does not satisfy the usual real-sequence symmetry at DC.

Structural or analytic derivation

Multiplication by a monomial z^{-2} in the circular time algebra becomes multiplication by a diagonal unit-modulus matrix in frequency coordinates. Thus magnitudes are unchanged and phase shifts linearly by $-4\pi k/5$.

Comparative cost and preferred method

Five scalar phase factors replace an inverse DFT followed by a shift and another DFT. For arbitrary N, this is $O(N)$ instead of $O(N \log N)$ transform work, and the output-size lower bound is $\Omega(N)$.

1.9 SC-009 | Four-point decimation-in-time FFT

Reference: [1], exercise 3.6.8.

Source task and conventions

Derive the length-four FFT flow and apply it to $(1, 0, 1, 0)$.

Conventional solution and proof

Split the defining sum into even and odd time indices. Define

$$a = x_0 + x_2, \quad b = x_0 - x_2, \quad c = x_1 + x_3, \quad d = x_1 - x_3.$$

The four outputs are

$$\boxed{X_0 = a + c, \quad X_2 = a - c, \quad X_1 = b - id, \quad X_3 = b + id.}$$

Substituting the input gives $(a, b, c, d) = (2, 0, 0, 0)$ and $X = (2, 0, 2, 0)$. The diagram records the two stages and the quarter-turn multiplication. There are eight complex additions/subtractions, no general complex multiplications, and one multiplication by $-i$, implemented by exchanging real/imaginary components and changing a sign. In the required two-stage radix-two butterfly architecture, each of four butterflies needs its sum and difference, so eight additions are necessary within that architecture. No lower bound for arbitrary unrestricted linear circuits is asserted.

Structural or analytic derivation

The algebra is the factorization of the four-point Fourier matrix into two block-diagonal two-point Fourier transforms, a diagonal twiddle and a permutation. The identity $(-i)^{2k} = (-1)^k$ proves the even/odd split directly and handles all four output indices.

Comparative cost and preferred method

A literal dense DFT has sixteen coefficient products and twelve additions before simplifying trivial roots. The butterfly has eight additions and only a sign/swap twiddle. A competent hand DFT may simplify to the same circuit, so the comparison is against the declared dense evaluation, not every conventional implementation.

Graphical reference. The shared construction is displayed in Volume II.

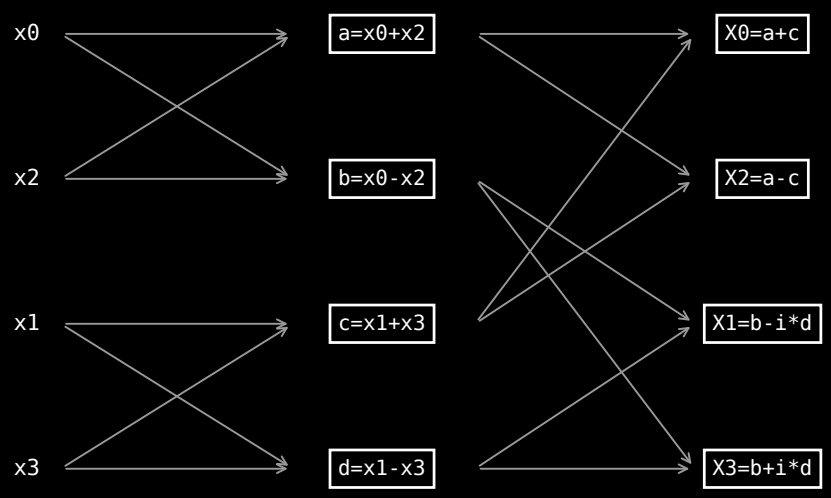


Figure 1.1: Four-point decimation in time: each labelled intermediate records its exact sign convention. Eight additions/subtractions and a quarter-turn implement the stated butterfly.

1.10 SC-010 | Inverse DFT using a forward transform

Reference: [1], exercise 3.6.9.

Source task and conventions

Give a block construction that uses an existing forward DFT routine to compute an inverse DFT.

Conventional solution and proof

For the negative-exponent forward transform F, complex conjugation of its defining sum yields

$$\boxed{F^{-1}Y = \frac{1}{N}\,\overline{F(\overline{Y})}.}$$

Indeed the right-hand component is $N^{-1}\sum_k Y_k e^{2\pi ikn/N}$, exactly the inverse definition. The block chain is input $Y \to$ conjugate $\to F \to$ conjugate $\to$ scale by $1/N$. Alternatively $F^{-1}Y = N^{-1}R(FY)$, where R is circular reversal; its proof is the same exponent substitution. The scale is essential.

Structural or analytic derivation

The identity $F^2 = NR$ follows from roots-of-unity orthogonality. Since $R^2 = I$, it provides an algebraic inverse and a verification test independent of any stored inverse-transform implementation.

Comparative cost and preferred method

The method requires one forward transform plus $O(N)$ conjugations/scales or a reversal. It has the same asymptotic cost as a dedicated inverse FFT and reduces the number of independently maintained transform kernels from two to one. That maintenance reduction is not a measured runtime speedup.

1.11 SC-011 | A transform program for arbitrary lengths

Reference: [1], exercise 3.6.10.

Source task and conventions

Write a MATLAB routine computing and plotting the DFT magnitude and phase of any nonempty finite sequence, including non-power-of-two lengths.

Conventional solution and proof

The delivered file `code/matlab/dftplots.m` validates a numeric finite vector, sets $N = \mathrm{numel}(x)$, computes $X = \mathrm{fft}(x)$ and uses the grid $\omega_k = 2\pi k/N$. Magnitude is $|X_k|$ and phase is $\arg X_k$; phase at numerically zero magnitude is marked undefined. The routine preserves the requested length: it does not silently zero-pad to a power of two. For $N = 1$, the single output is x_0. A direct defining-sum implementation is included as a small-input reference. The accompanying Python translation is executed against that reference on composite and prime lengths.

Structural or analytic derivation

For an arbitrary N, $kn = [k^2 + n^2 - (k-n)^2]/2$ gives

$$X_k = e^{-\pi i k^2/N} \sum_{n=0}^{N-1} \left(x_n e^{-\pi i n^2/N}\right) e^{\pi i (k-n)^2/N}.$$

The sum is a linear convolution after including the negative-index chirp and padding to at least $2N - 1$. This proves that a power-of-two convolution backend can compute the exact length-N DFT. A robust FFT library already incorporates efficient non-power-of-two algorithms; the delivered default uses it.

Comparative cost and preferred method

The direct defining sum uses N^2 products and $N(N-1)$ additions. Chirp convolution uses $O(N \log N)$ arithmetic with $O(N)$ workspace. The production library is the strong control; the explicit chirp implementation is explanatory and is not claimed faster. MATLAB execution is not asserted in this environment; the supplied routine and the executed Python equivalent are distinguished.

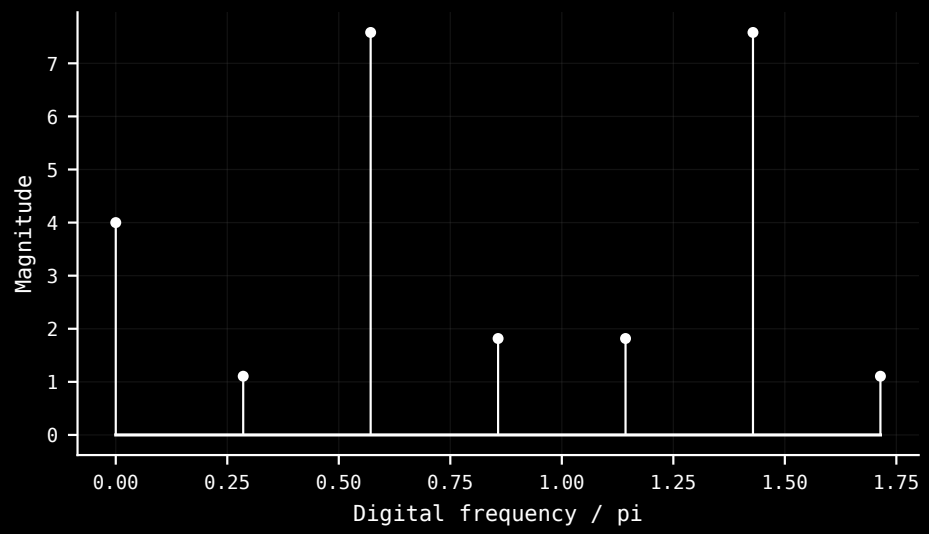


Figure 1.2: Seven-point example for the general source routine; the output length is seven, with no change to the transform by zero-padding.

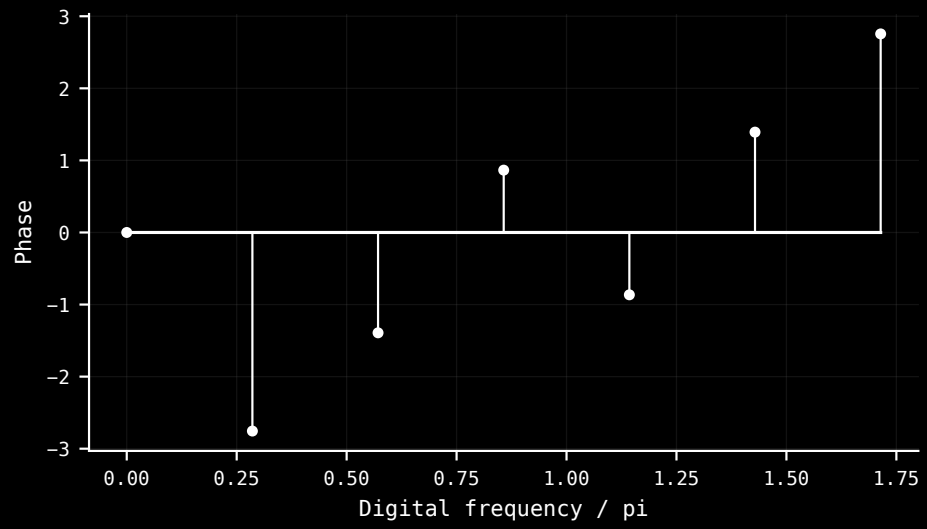


Figure 1.3: Seven-point example for the general source routine; the output length is seven, with no change to the transform by zero-padding.

1.12 SC-012 | Three discrete-time Fourier transforms

Reference: [1], exercise 2.6.1.

Source task and conventions

Transform the three exponential sequences in Problem 2.6.1, with $u[n] = 1$ for $n \geq 0$.

Conventional solution and proof

For (i), with $|\beta| < 1$,

$$W(e^{i\omega}) = \sum_{n=2}^{\infty} (\beta e^{-i\omega})^n = \frac{\beta^2 e^{-2i\omega}}{1 - \beta e^{-i\omega}}.$$

For (ii), with integer $L \geq 0$,

$$X(e^{i\omega}) = 1 + \sum_{n=1}^{L} \beta^n (e^{-in\omega} + e^{in\omega}) = 1 + \frac{r_-(1 - r_-^L)}{1 - r_-} + \frac{r_+(1 - r_+^L)}{1 - r_+}, \quad r_\pm = \beta e^{\pm i\omega}.$$

When a denominator vanishes, its finite-sum term is L; the polynomial expression is valid for every finite β. For (iii), put $m = -n \geq 4$ and use $\beta > 1$:

$$Y(e^{i\omega}) = \sum_{m=4}^{\infty} (\beta^{-1} e^{i\omega})^m = \frac{\beta^{-4} e^{4i\omega}}{1 - \beta^{-1} e^{i\omega}}.$$

The geometric sum follows by multiplying its partial sum by $1 - r$ and passing to the limit only when $|r| < 1$.

Structural or analytic derivation

All three calculations are evaluations of Laurent polynomials or Laurent series on $|z| = 1$. The right-sided sequence needs $|z| > |\beta|$, and the left-sided sequence needs $|z| < \beta$. Their Fourier transforms exist because the unit circle lies in the respective convergence region. The finite symmetric sequence combines two analytic geometric polynomials, so apparent poles at $r_{\pm} = 1$ are removable.

Comparative cost and preferred method

The infinite sums have fixed-size exact expressions with explicit convergence conditions. The finite direct sum needs $O(L)$ terms per frequency; geometric evaluation takes $O(\log L)$ multiplications for integer powers, or fixed transcendental calls in a floating-point model. The usual geometric-series method is equally efficient. The source's $\beta < 1$ in (i) is qualified to $|\beta| < 1$.

1.13 SC-013 | Time reversal of a real sequence

Reference: [1], exercise 2.6.2.

Source task and conventions

Express the DTFT of $x[-n]$ in terms of the DTFT of a real sequence $x[n]$.

Conventional solution and proof

Changing variable $m = -n$ gives $\sum_n x[-n]e^{-i\omega n} = \sum_m x[m]e^{i\omega m} = X(e^{-i\omega})$. Since $x[m]$ is real, conjugating the defining sum gives $X(e^{-i\omega}) = \overline{X(e^{i\omega})}$. Thus the answer is $\boxed{X(e^{-i\omega}) = X^*(e^{i\omega})}$. The argument holds pointwise for absolutely summable sequences and in the L^2 sense for square-summable sequences.

Structural or analytic derivation

Reflection on the unit circle sends z to $\bar{z} = z^{-1}$. Real coefficient symmetry identifies this reflection with complex conjugation. The proof contains the relevant convergence convention and does not assume the transform itself is real.

Comparative cost and preferred method

A stored spectrum needs only conjugation, $O(N)$ entries, instead of a new transform costing $O(N \log N)$ by FFT on a finite grid. If only one frequency is requested and its value is known, one conjugation suffices. The standard Fourier symmetry is already the preferable method.

1.14 SC-014 | Z transforms with shifted one-sided support

Reference: [1], exercise 4.8.1.

Source task and conventions

Find all three Z transforms and their exact regions of convergence.

Conventional solution and proof

For (i), summing from $n = 2$ gives

$$X_1(z) = \frac{(1/2)^2 z^{-2}}{1 - (1/2)z^{-1}}, \quad |z| > 1/2.$$

For a left-sided exponential $a^n u[-n-3]$, let $m = -n \geq 3$. Then

$$\sum_{n=-\infty}^{-3} a^n z^{-n} = \sum_{m=3}^{\infty} (z/a)^m = \frac{(z/a)^3}{1 - z/a}, \quad |z| < |a|.$$

Hence (ii) is $-\frac{(z/(2/5))^3}{1-z/(2/5)}$ with $|z| < 2/5$, and (iii) is $\frac{(z/(-2/5))^3}{1-z/(-2/5)}$ with $|z| < 2/5$. The origin is included for these strictly left-supported polynomial-at-zero expressions.

Structural or analytic derivation

The three rational expressions are accompanied by distinct Laurent expansion domains. The sign of a moves the pole but does not change the radius of convergence. Finite shifts change polynomial factors and the behaviour at zero or infinity, while the nonzero finite boundary radius remains $|a|$.

Comparative cost and preferred method

Geometric summation reduces each infinite series to fixed rational arithmetic. A transform-table baseline gives the same efficiency only if it also retains support and region of convergence. Omitting the region loses the information that distinguishes left and right sequences.

1.15 SC-015 | A one-sided exponentially weighted cosine

Reference: [1], exercise 4.8.2.

Source task and conventions

Derive the Z transform of $r^n \cos(\omega_0 n)u[n]$ and specialize to $2^n \cos(\pi n/2)u[n]$.

Conventional solution and proof

Writing the cosine as two exponentials and summing gives

$$Y(z) = \frac{1}{2}\left(\frac{1}{1 - re^{i\omega_0} z^{-1}} + \frac{1}{1 - re^{-i\omega_0} z^{-1}}\right) = \boxed{\frac{1 - r\cos\omega_0 z^{-1}}{1 - 2r\cos\omega_0 z^{-1} + r^2 z^{-2}}}, \quad |z| > |r|.$$

For $r = 2$ and $\omega_0 = \pi/2$,

$$\boxed{W(z) = \frac{1}{1 + 4z^{-2}}, \quad |z| > 2.}$$

The sequence is $1, 0, -4, 0, 16, 0, -64, \ldots$, confirming its expansion. For $r = 0$, the sequence is just $\delta[n]$ under the convention $0^0 = 1$ and the apparent region restriction is removable.

Structural or analytic derivation

The poles are $re^{\pm i\omega_0}$. This is the elementary spectral decomposition of a second-order recurrence; conjugate poles ensure a real sequence. Combining the two fractions before evaluating can reduce arithmetic but does not authorize use of the formula on the unit circle when $|r| \geq 1$.

Comparative cost and preferred method

Two geometric series give a fixed-size rational result. Computing a long prefix by the second-order recurrence is $O(N)$; evaluating the rational transform at one admissible point has fixed arithmetic cost. The standard exponential decomposition is already efficient.

1.16 SC-016 | An empty region of convergence

Reference: [1], exercise 4.8.3.

Source task and conventions

Decide whether $2^n u[-n-1] + 3^n u[n+1]$ has a bilateral Z transform.

Conventional solution and proof

The first summand is supported on $n \leq -1$ and requires $|z| < 2$. The second begins at $n = -1$ and has an infinite positive tail requiring $|z| > 3$. Their intersection is empty. More directly, on the positive tail the terms $(3/z)^n$ fail to tend to zero unless $|z| > 3$; on the negative tail the terms $(z/2)^m$ fail unless $|z| < 2$. These conditions cannot both hold. The overlap at $n = -1$ changes one finite coefficient and cannot cancel either tail. Therefore no nonempty annulus of bilateral convergence exists.

Structural or analytic derivation

Algebraically adding the two separately continued rational formulas gives a meromorphic expression, but it is not the sum of the original bilateral series on any common domain. Analytic continuation does not create a convergence region for a different specified Laurent series.

Comparative cost and preferred method

Two radius comparisons decide nonexistence; truncating a bilateral sequence cannot establish convergence. This is a constant-size exact certificate for the symbolic input. No numerical Z-transform evaluation is admissible under the requested convergence contract.

1.17 SC-017 | Inverse Z transform by residues

Reference: [1], exercise 4.8.4.

Source task and conventions

Invert the stated rational function in each of its three regions, using residues.

Conventional solution and proof

On an admissible counterclockwise circle, $h[n] = (2\pi i)^{-1} \oint H(z) z^{n-1}\, dz$. The residues at two and three are -2^n and $2\,3^n$. For $n \geq 0$ there is no pole at zero, so the enclosed nonzero poles give respectively both, only the first, or neither contribution. For $n \leq -1$, the integrand is $O(z^{-2})$ at infinity; its residues at zero, two and three sum to zero. Thus the origin residue equals $2^n - 2\,3^n$. Adding whichever nonzero poles lie inside the circle gives zero in the outer region, $-2\,3^n$ in the annulus, and $2^n - 2\,3^n$ in the inner region. This proves the same table by residues for every integer n.

Structural or analytic derivation

For $H(z) = z(z-1)/[(z-2)(z-3)]$, the partial fractions are $-z/(z-2) + 2z/(z-3)$. The complete inverse sequences are

$$\begin{array}{c|l} |z| > 3 & [-2^n + 2\,3^n]u[n] \\ 2 < |z| < 3 & -2^n u[n] - 2\,3^n u[-n-1] \\ |z| < 2 & [2^n - 2\,3^n]u[-n-1]. \end{array}$$

The identities $z/(z-a) = \sum_{n \geq 0} a^n z^{-n}$ for $|z| > |a|$ and $z/(z-a) = -\sum_{n \leq -1} a^n z^{-n}$ for $|z| < |a|$ follow by geometric summation, proving the table and every support endpoint.

Comparative cost and preferred method

There are two nonzero simple poles. Their residues require fixed arithmetic for any chosen index, with powers computed in $O(\log |n|)$ multiplications. A full prefix uses an $O(N)$ recurrence. Partial fractions are a strong conventional control and yield the same outputs; contour quadrature would add unnecessary numerical error here.

1.18 SC-018 | Inverse Z transform by partial fractions

Reference: [1], exercise 4.8.5.

Source task and conventions

Invert the same function in all three regions using partial fractions.

Conventional solution and proof

The partial-fraction decomposition and the inverse sequences for all three admissible regions are listed in SC-017. They apply to this identical rational transform and supply the requested inverse.

The identities $z/(z-a) = \sum_{n\geq 0} a^n z^{-n}$ for $|z| > |a|$ and $z/(z-a) = -\sum_{n\leq -1} a^n z^{-n}$ for $|z| < |a|$ follow by geometric summation, proving the table and every support endpoint.

Structural or analytic derivation

The complete residue calculation, including the origin contribution for negative indices and all three annular regions, is given in SC-017. It verifies the same inverse table independently of the partial-fraction method requested here.

Comparative cost and preferred method

The source explicitly asks for partial fractions, so the two-coefficient decomposition is retained as the preferred calculation. Residues provide an independent coefficient proof. Both have fixed symbolic setup and identical evaluation complexity; no speedup is attributed merely to a change of terminology.

1.19 SC-019 | A repeated pole and an impossible stated region

Reference: [1], exercise 4.8.6.

Source task and conventions

Invert $H(z) = (1 - z^{-1}/2)/[(1 - z^{-1}/3)(1 - 2z^{-1})^2]$ and examine the printed condition $|z| > 1/3$.

Conventional solution and proof

The exact decomposition is

$$H(z) = -\frac{1/50}{1 - z^{-1}/3} + \frac{3/25}{1 - 2z^{-1}} + \frac{9/10}{(1 - 2z^{-1})^2}.$$

It has a noncancelled pole at $z = 2$, so the entire set $|z| > 1/3$ cannot be a region of analyticity or convergence. Two valid regions satisfying its lower-radius intention give different inverses. For $|z| > 2$,

$$h[n] = \left[-\frac{1}{50}3^{-n} + \left(\frac{3}{25} + \frac{9}{10}(n+1)\right)2^n\right] u[n].$$

For $1/3 < |z| < 2$,

$$h[n] = -\frac{1}{50}3^{-n}u[n] + \left[-\frac{3}{25} + \frac{9}{10}(-n-1)\right]2^n u[-n-1].$$

In the second expression the repeated-pole contribution is zero at $n = -1$, as required. There is no unique inverse under the printed region alone.

Structural or analytic derivation

For a right-sided double pole, differentiating the geometric series yields $(1-r)^{-2} = \sum_{n\geq 0}(n+1)r^n$. For a left-sided expansion, $(1-2/z)^{-2} = (z/2)^2(1-z/2)^{-2}$, whose coefficient at $n \leq -2$ is $(-n-1)2^n$. These two expansions prove both inverses and explain why pole multiplicity produces a linear factor in the index.

Comparative cost and preferred method

Checking pole membership disproves the stated region before any inverse algorithm runs. The corrected alternatives each need three elementary partial-fraction terms. A method that reports only a causal inverse while accepting the printed region would omit a source ambiguity; the rigorous improvement is early feasibility checking.

Source qualification

Physical source page 127 visibly prints $|z| > 1/3$; it does not exclude the double pole at two.

1.20 SC-020 | The first five impulse samples by long division

Reference: [1], exercise 4.8.7.

Source task and conventions

Find the first five causal impulse samples for $H(z) = z/(3z^2 - 4z + 1)$.

Conventional solution and proof

Put $q = z^{-1}$ and write $H(q) = q/(3-4q+q^2) = \sum_{n\geq 0} h_n q^n$. Equating coefficients gives $3h_0 = 0$, $3h_1 - 4h_0 = 1$ and $3h_n - 4h_{n-1} + h_{n-2} = 0$ for $n \geq 2$. Thus

$$\boxed{h_0 = 0,\ h_1 = \tfrac{1}{3},\ h_2 = \tfrac{4}{9},\ h_3 = \tfrac{13}{27},\ h_4 = \tfrac{40}{81}.}$$

This coefficient recurrence is long division with exact fractions. It also proves uniqueness of the causal expansion.

Structural or analytic derivation

Factoring the denominator gives $H = \frac{1}{2}z/(z-1) - \frac{1}{2}z/(z-1/3)$, so $h_n = \frac{1}{2}(1 - 3^{-n})u[n]$. Substitution reproduces all five coefficients. The causal region is $|z| > 1$; the impulse response tends to $1/2$, so the system is not BIBO stable.

Comparative cost and preferred method

For five outputs, recurrence/long division is the preferred $O(N)$ method. A closed form helps remote indices but does not reduce the output lower bound for a complete prefix. The pole at one gives a stability conclusion without simulating a long response.

1.21 SC-021 | Transfer function of a four-tap impulse response

Reference: [1], exercise 4.8.8.

Source task and conventions

Obtain the transfer function for $h[n] = \sum_{k=0}^{3} \alpha_k \delta[n-k]$.

Conventional solution and proof

The defining Z transform contains only one nonzero contribution from each delta:

$$\boxed{H(z) = \alpha_0 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \alpha_3 z^{-3}.}$$

The input-output relation is $y[n] = \sum_{k=0}^{3} \alpha_k x[n-k]$. The nonzero finite plane is a convergence domain; zero is excluded when a positive-delay coefficient is nonzero, and any formal poles created by multiplying numerator and denominator are not extra physical modes.

Structural or analytic derivation

This is a polynomial in the delay operator $Sx[n] = x[n-1]$: $T = \sum \alpha_k S^k$. Its operator norm on bounded sequences equals $\sum |\alpha_k|$, by the finite phase-alignment proof of the convolution norm. The representation exactly separates delay from coefficient arithmetic.

Comparative cost and preferred method

Four coefficient reads and a four-term expression are sufficient and necessary for arbitrary coefficients. Direct and structural derivations coincide. No transform sampling is required.

1.22 SC-022 | Moving-average impulse and transfer functions

Reference: [1], exercise 4.8.9.

Source task and conventions

Derive the impulse response and transfer function of the causal M-point moving average.

Conventional solution and proof

Apply the system to $x[n] = \delta[n]$. This gives $h[n] = 1/M$ for $0 \le n < M$ and zero otherwise. Therefore

$$\boxed{H(z) = \frac{1}{M} \sum_{k=0}^{M-1} z^{-k} = \frac{1 - z^{-M}}{M(1 - z^{-1})}.}$$

The rational form has a removable value at $z = 1$, namely one; the true filter is FIR. On the unit circle,

$$H(e^{i\omega}) = \frac{1}{M} e^{-i(M-1)\omega/2} \frac{\sin(M\omega/2)}{\sin(\omega/2)},$$

with its continuous limiting values at multiples of 2π. Multiplication of the finite sum by $1 - z^{-1}$ proves both forms.

Structural or analytic derivation

The zeros are the nontrivial Mth roots of unity. The formal integrator pole in the rational expression cancels exactly, explaining why the rolling-sum realization can be BIBO stable as an input-output filter while its internal accumulator needs finite-precision care.

Comparative cost and preferred method

Direct filtering needs M samples per output; the exact rolling identity from Problem 1.7.7 needs constant arithmetic and a ring buffer. Both still store M delayed inputs for arbitrary streaming data. The symbolic transfer-function derivation itself already has constant-size geometric form.

1.23 SC-023 | Inverse transforms of real and imaginary spectral parts

Reference: [1], exercise 2.6.3.

Source task and conventions

Find the inverse transform of $\Re X(e^{i\omega})$ and $\Im X(e^{i\omega})$ for real $x[n]$.

Conventional solution and proof

Conjugate symmetry gives $X^*(e^{i\omega}) = X(e^{-i\omega})$, whose inverse is $x[-n]$. Linearity then gives

$$\boxed{\mathcal{F}^{-1}(\Re X)[n] = \tfrac{1}{2}(x[n] + x[-n]), \quad \mathcal{F}^{-1}(\Im X)[n] = \frac{x[n] - x[-n]}{2i}.}$$

The second inverse is generally imaginary; the inverse of $i\Im X$ is the real odd part. This follows directly from $\Re X = (X + X^*)/2$ and $\Im X = (X - X^*)/(2i)$. For a finite DFT use circular reversal $[-n]_N$.

Structural or analytic derivation

The two projections $P_e = (I + R)/2$ and $P_o = (I - R)/2$, where $Rx[n] = x[-n]$, satisfy $P_e^2 = P_e$, $P_o^2 = P_o$, $P_e P_o = 0$. They decompose the time sequence into even and odd components. The factor $1/i$ is required by the chosen transform convention.

Comparative cost and preferred method

Pairwise sums and differences cost $O(N)$ and use no inverse FFT. The saving against two fresh transforms is asymptotically $O(N \log N)$ to $O(N)$ for finite arrays, with the same outputs. The source mixes DTFT and DFT terminology; both conventions are supplied explicitly.

1.24 SC-024 | Conjugate symmetry for a real DTFT

Reference: [1], exercise 2.6.8.

Source task and conventions

Prove $X(\omega) = X^*(-\omega)$ and the even/odd symmetries of its components.

Conventional solution and proof

For real $x[n]$, conjugating $\sum_n x[n]e^{i\omega n}$ gives $\sum_n x[n]e^{-i\omega n} = X(\omega)$. Therefore $X(-\omega) = \overline{X(\omega)}$. Taking real parts proves $\Re X(-\omega) = \Re X(\omega)$; taking imaginary parts proves $\Im X(-\omega) = -\Im X(\omega)$. This supplies all three assertions directly, including zeros and endpoints. The same absolute-summability or L^2 qualification as in the preceding identities applies.

Structural or analytic derivation

In a Laurent representation, reality of the coefficients is precisely invariance under $X(z) \mapsto \overline{X(\bar{z})}$. On the unit circle the reflection changes the frequency sign. No symmetry of the original time sequence is required.

Comparative cost and preferred method

Only one half of a symmetric frequency grid needs independent evaluation. Completing the other half costs conjugations; the number of independent values is reduced by approximately two, with DC and Nyquist values real. This is a standard transform property, not a separate novel acceleration.

1.25 SC-025 | Real-sequence DFT symmetry

Reference: [1], exercise 3.6.2.

Source task and conventions

Prove the three circular symmetry relations for the DFT of a real sequence.

Conventional solution and proof

Since $x[n]$ is real,

$$X^*[(-k)_N] = \sum_{n=0}^{N-1} x[n] e^{2\pi i (-k)_N n/N} = \sum_n x[n] e^{-2\pi i k n/N} = X[k].$$

Writing $X[k] = a_k + ib_k$ then gives $a_{(-k)_N} = a_k$ and $b_{(-k)_N} = -b_k$. In particular $X[0]$ is real, and if N is even then $X[N/2]$ is real. These self-paired cases are part of the proof, not extra assumptions.

Structural or analytic derivation

The DFT sends real vectors to the fixed-point subspace of circular conjugate reversal. This subspace has exactly N real degrees of freedom, so storing N unrestricted complex numbers is redundant.

Comparative cost and preferred method

The symmetry reconstructs a full spectrum from $\lfloor N/2 \rfloor + 1$ complex-grid values with the endpoint reality constraints. Completion is linear in output size. A real FFT exploits this redundancy internally; comparison with a strong real-FFT library does not yield another factor-of-two saving.

1.26 SC-026 | An inconsistent real-signal spectrum

Reference: [1], exercise 3.6.7.

Source task and conventions

Determine whether the six printed values can be the first six entries of a twelve-point DFT of a real sequence.

Conventional solution and proof

For any real sequence, $V[0] = \sum_n v[n]$ is real. The source gives $V[0] = 1 + i$, so no real sequence has the stated spectrum. There is also no supplied value for the self-paired bin $V[6]$; even after replacing $V[0]$ by a real number it would remain a free real parameter. Under that explicitly modified premise, conjugate symmetry would give

$$V[7] = 1,\ V[8] = 3 - i,\ V[9] = 2,\ V[10] = 7 + i,\ V[11] = 3, \qquad V[6] \in \mathbb{R}.$$

The contradiction at DC resolves the literal question; a unique numerical completion is impossible.

Structural or analytic derivation

The fixed-point condition $V[k] = V^*[(-k)_{12}]$ gives both the immediate inconsistency and the unpaired real degree of freedom. It is stronger than attempting an inverse transform and inspecting its imaginary part.

Comparative cost and preferred method

One reality check disproves feasibility. With consistent endpoints, completing five paired entries is $O(N)$ work. No invented sixth value or shifted indexing is used to manufacture a unique answer.

Source qualification

The twelve-point length and $1 + i$ first bin were checked on source page 101.

1.27 SC-027 | Linear phase of a symmetric five-tap sequence

Reference: [1], exercise 2.6.5.

Source task and conventions

Prove linear phase for $x = (1, 2, 3, 2, 1)$ supported on $0, \dots, 4$.

Conventional solution and proof

Pair terms symmetrically around two:

$$X(e^{i\omega}) = e^{-2i\omega}[3 + 4\cos\omega + 2\cos 2\omega] = e^{-2i\omega}(1 + 2\cos\omega)^2.$$

The bracket is real and nonnegative. Therefore the phase is -2ω modulo 2π wherever $X \neq 0$, and the group delay is two samples. At $\omega = \pm 2\pi/3$ modulo 2π, the response is zero and phase is undefined. The identity also follows by recognizing $(1 + z^{-1} + z^{-2})^2$.

Structural or analytic derivation

Self-reciprocity $z^{-4}X(z^{-1}) = X(z)$ forces the centered response to be real on the circle. The explicit square additionally removes sign-induced π phase jumps. This is an exact polynomial factorization, not a visual inference from a phase plot.

Comparative cost and preferred method

Symmetric evaluation uses two cosines and three real coefficients instead of five complex exponential terms per frequency. The factorized formula uses one cosine, one affine expression and one square. For implementation, coefficient symmetry pairs input samples and reduces distinct nontrivial multipliers; the exact gain depends on treatment of the integer coefficients.

1.28 SC-028 | A linear-phase assertion contradicted by the printed

Reference: [1], exercise 2.6.6.

Source task and conventions

Test the claimed linear phase of $h = (1, -2, 3, 2, -1)$.

Conventional solution and proof

The displayed coefficients give

$$H(e^{i\omega}) = e^{-2i\omega}\{3 + i[2\sin 2\omega - 4\sin\omega]\}.$$

Set $b(\omega) = 2\sin 2\omega - 4\sin\omega$. Since the real part of the bracket is three, it never vanishes and its continuous argument is $\arctan(b/3)$. Thus

$$\phi'(\omega) = -2 + \frac{3b'(\omega)}{9 + b(\omega)^2}.$$

At zero, $b = 0, b' = 0$, so $\phi'(0) = -2$. At π, $b = 0, b' = 8$, so $\phi'(\pi) = 2/3$. These different derivatives prove that the phase is not affine. Consequently the requested assertion is false for the printed filter.

Structural or analytic derivation

For a real FIR polynomial with nonzero endpoints, exact generalized linear phase requires symmetric or antisymmetric coefficients. This sequence has neither: $h_0 = -h_4$, $h_1 = -h_3$, but the middle coefficient is nonzero, which violates antisymmetry; the endpoints violate symmetry. The direct derivative calculation supplies an independent counterexample without relying solely on that characterization.

Comparative cost and preferred method

The symmetry check costs $O(M)$ comparisons; two exact phase-derivative evaluations finish the disproof here. A dense phase plot cannot prove linearity and could conceal variation. No alternative coefficients are silently substituted to make the requested claim true.

Source qualification

The signs and central coefficient were checked visually on physical source page 68. This entry resolves a false source assertion by a complete disproof.

1.29 SC-029 | Four Fourier identities for complex sequences

Reference: [1], exercise 2.6.7.

Source task and conventions

Prove or correct all four identities in Problem 2.6.7.

Conventional solution and proof

Let $X(\omega) = \sum_n x[n]e^{-i\omega n}$. Reindexing gives $\mathcal{F}\{x[-n]\} = X(-\omega)$. Reindexing and conjugating gives

$$\mathcal{F}\{x^*[-n]\} = \sum_m x^*[m]e^{i\omega m} = X^*(\omega).$$

The source's second identity omits the conjugation on the result. For $x[n] = i\delta[n]$, its two sides would be $-i$ and i, an explicit contradiction. Also $\mathcal{F}\{x^*[n]\} = X^*(-\omega)$. Expanding real and imaginary parts now proves

$$\mathcal{F}\{\Re x\} = \tfrac{1}{2}[X(\omega) + X^*(-\omega)], \quad \mathcal{F}\{i\Im x\} = \tfrac{1}{2}[X(\omega) - X^*(-\omega)].$$

All manipulations are justified by absolute summability, or by density and continuity for ℓ^2 transforms.

Structural or analytic derivation

Conjugation C, reversal R and their product are commuting involutions. Their projections recover the real/imaginary and even/odd subspaces. Tracking both operations prevents the missing conjugation in part (ii).

Comparative cost and preferred method

For a length-N spectrum, all four transformed outputs can be obtained with reversal, conjugation, additions and scales in $O(N)$ work, avoiding four new FFTs. The source's false part is replaced only after it has been explicitly disproved.

Source qualification

The missing conjugation is present in the visually inspected printed statement, not introduced by this transcription.

1.30 SC-030 | Summability and energy of a one-sided exponential

Reference: [1], exercise 1.7.5.

Source task and conventions

Determine absolute summability and energy of $x[n] = A\alpha^n u[n]$.

Conventional solution and proof

Assume A is finite. If $A = 0$, both sums vanish. Otherwise

$$\sum_{n\in\mathbb{Z}} |x[n]| = |A| \sum_{n=0}^{\infty} |\alpha|^n, \qquad E = \sum_n |x[n]|^2 = |A|^2 \sum_{n=0}^{\infty} |\alpha|^{2n}.$$

For $|\alpha| < 1$, multiplying each geometric series by $1 - |\alpha|$ or $1 - |\alpha|^2$ gives

$$\boxed{\|x\|_1 = \frac{|A|}{1-|\alpha|}, \qquad E = \frac{|A|^2}{1-|\alpha|^2}.}$$

For $|\alpha| \geq 1$, the nonzero summands fail to tend to zero, so neither sum converges. The printed assumption $\alpha < 1$ alone allows $\alpha = -2$ and is insufficient.

Structural or analytic derivation

The transfer generating function $X(z) = A/(1 - \alpha z^{-1})$ converges for $|z| > |\alpha|$. Its pole is inside the unit disk exactly in the summable case. Parseval gives the same energy from the unit-circle integral; the geometric proof above already establishes the needed convergence and is shorter for one pole.

Comparative cost and preferred method

A truncation at N requires N terms and leaves exact energy tail $|A|^2|\alpha|^{2N}/(1 - |\alpha|^2)$. The closed form uses fixed arithmetic and gives the entire infinite sum. The conventional geometric-series solution is optimal for this symbolic input; complex analysis adds a pole interpretation.

Source qualification

The convergence condition is stated in modulus; this is an explicit qualification of the source's inequality.

1.31 SC-031 | Moving-average system properties

Reference: [1], exercise 1.7.7.

Source task and conventions

Classify the source's causal M-sample moving average, $Tx[n] = M^{-1} \sum_{k=0}^{M-1} x[n-k]$, for integer $M \geq 1$.

Conventional solution and proof

For scalars a, b, distributing the finite sum proves $T(ax + bv) = aTx + bTv$. If $|x[n]| \leq B$, then $|Tx[n]| \leq M^{-1} \sum_{k=0}^{M-1} B = B$, proving BIBO stability. Only indices $n - k \leq n$ occur, proving causality. For a shift $S_r x[n] = x[n - r]$, substitution gives $T(S_r x)[n] = M^{-1} \sum_k x[n - r - k] = (S_r Tx)[n]$, proving time invariance. The induced ℓ^∞ norm equals one: the upper bound is attained by $x[n] \equiv 1$.

Structural or analytic derivation

Its impulse response is $h[k] = 1/M$ on $0 \leq k < M$ and zero elsewhere. The finite support and sum $\|h\|_1 = 1$ encode causality and the exact gain. A rolling sum satisfies

$$y[n] = y[n-1] + \frac{x[n] - x[n-M]}{}.$$

Subtracting the two overlapping window sums proves the identity. It retains the same input-output map in exact arithmetic.

Comparative cost and preferred method

Direct evaluation needs $M-1$ additions per output and one scale. A rolling implementation needs one subtraction, one addition and one scale after $O(M)$ initialization, with an M-sample ring buffer. For a long stream this is $\Theta(N)$ versus $\Theta(NM)$ arithmetic. Rounding error in a recursively updated sum can accumulate, so periodic exact-window refresh or compensated accumulation is a separate numerical choice.

1.32 SC-032 | Backward-difference differentiator

Reference: [1], exercise 1.7.8.

Source task and conventions

Classify the source's discrete differentiator $Dx[n] = (x[n] - x[n-1])/T$, where $T > 0$.

Conventional solution and proof

Distributing the two terms proves linearity. The output uses the present and previous input, so it is causal. Replacing $x[n]$ by $x[n-r]$ shifts the output by the same r, proving time invariance. For bounded input $|x[n]| \le B$, $|Dx[n]| \le 2B/T$, so the discrete operator is BIBO stable. The alternating input $x[n] = B(-1)^n$ attains the bound; the exact induced norm is $2/T$.

Structural or analytic derivation

The impulse response is $T^{-1}(\delta[n] - \delta[n-1])$, and

$$H(e^{i\omega}) = T^{-1}(1 - e^{-i\omega}) = \frac{2i}{T} e^{-i\omega/2} \sin(\omega/2).$$

Its maximum magnitude is $2/T$. Taylor's integral identity gives $[x(t) - x(t-T)]/T = x'(t) - T^{-1} \int_{t-T}^{t} (s - t + T) x''(s)\, ds$; hence the derivative error is at most $T \sup |x''|/2$. This distinguishes exact discrete stability from the approximation to continuous differentiation.

Comparative cost and preferred method

One subtraction and one scale already realize the exact operator. The frequency representation gives no cheaper sample implementation; it provides the sharp gain and an accuracy condition. As $T \downarrow 0$, the norm grows, so stability for each fixed sample period does not imply a uniform noise bound.

1.33 SC-033 | Necessary and sufficient BIBO stability

Reference: [1], exercise 1.7.9.

Source task and conventions

Prove that a discrete convolution operator is BIBO stable exactly when its impulse response is absolutely summable.

Conventional solution and proof

If $h \in \ell^1(\mathbb{Z})$ and $\|x\|_\infty \le B$, absolute convergence and the triangle inequality give $|(h*x)[n]| \le B\sum_k |h[k]|$ for every n. Conversely, assume the convolution operator obeys a uniform BIBO bound $\|Tx\|_\infty \le C\|x\|_\infty$. For each finite set F, choose a bounded input supported at $-F$ with $x[-k] = \overline{h[k]}/|h[k]|$ for $h[k] \neq 0$ and zero otherwise. Then $(Tx)[0] = \sum_{k\in F} |h[k]| \le C$. Increasing F through finite intervals proves $\sum_k |h[k]| \le C$. This also proves

$$\boxed{\|T\|_{\ell^\infty \to \ell^\infty} = \|h\|_1.}$$

If stability is phrased only as every bounded input producing bounded output, the single input $x[-k] = \overline{h[k]}/|h[k]|$ makes the output at zero diverge whenever $\sum |h[k]| = \infty$, so the convolution is not defined for all bounded inputs.

Structural or analytic derivation

This is the exact functional-analytic operator norm. For a rational causal filter, poles strictly inside the unit disk imply exponentially decaying polynomial-weighted impulse coefficients, hence ℓ^1 summability. A surviving pole on or outside the disk prevents summability. Pole cancellation must be performed on the actual input-output transfer function; hidden internal-state stability is a separate question.

Comparative cost and preferred method

Checking finitely many sample responses cannot prove the universal quantifier over bounded inputs. The ℓ^1 criterion proves it once and supplies the best bound. For a finite impulse response of length M, exactly M absolute values and $M-1$ additions evaluate that bound; inspecting all coefficients is necessary in the unconstrained input model.

1.34 SC-034 | A three-state canonical cascade with proved range scaling

Reference: [1], exercise 7.7.9.

Source task and conventions

Give a first-order/second-order cascade and scale its internal states under a declared bounded input.

Conventional solution and complete derivation

With $q = z^{-1}$, exact factorization gives

$$H(q) = \frac{1}{1-(2/5)q}\,\frac{11/25 + (9/25)q + (1/50)q^2}{1+(4/5)q+(1/2)q^2}.$$

An unscaled direct-form-II cascade is $v[n] = x[n] + (2/5)v[n-1]$, $w[n] = v[n] - (4/5)w[n-1] - (1/2)w[n-2]$, and $y[n] = (11/25)w[n] + (9/25)w[n-1] + (1/50)w[n-2]$. There are exactly three delay states, matching denominator degree. Taking z transforms proves the original transfer; the displayed block diagram labels all gains and delays. Absolute overflow protection requires an input amplitude, state initialization and accumulator range. Suppose $|x[n]| \le 1$, zero initial state and normalized storage range $[-1, 1]$. One conservative fully specified realization is

$$v[n] = \tfrac{3}{5}x[n] + \tfrac{2}{5}v[n-1], \quad w[n] = \tfrac{1}{24}v[n] - \tfrac{4}{5}w[n-1] - \tfrac{1}{2}w[n-2],$$

$$\widetilde{y}[n] = \tfrac{11}{20}w[n] + \tfrac{9}{20}w[n-1] + \tfrac{1}{40}w[n-2].$$

Its transfer is $H/32$, so $y = 32\widetilde{y}$ restores the physical output outside the normalized storage path. The first recurrence gives $\|v\|_\infty \le 1$. The two poles of the second all-pole section have modulus $1/\sqrt{2}$. Factoring it into two complex first-order inverses and applying the convolution triangle inequality gives impulse-response one-norm at most $(1 - 1/\sqrt{2})^{-2} = 6 + 4\sqrt{2} < 12$. Consequently $\|w\|_\infty \le (6 + 4\sqrt{2})/24 < 1/2$. Even the absolute sum of products in its accumulator is at most $[1 + (13/10)(6 + 4\sqrt{2})]/24 < 1$, while the output absolute sum is at most $(41/40)(6 + 4\sqrt{2})/24 < 1$. Thus states and these specified accumulation orders are protected under exact arithmetic. All listed multipliers have magnitude below one. Finite-precision roundoff requires a further margin; the strict inequalities leave a calculable allowance, but the first stage needs input headroom below one for a fixed-point endpoint excluding $+1$. For input bound X, first divide by X and interpret the restored output gain as $32X$. Without an input bound, no nonzero linear filter can be made overflow-free by fixed internal scaling while preserving an unbounded physical output in a bounded register. The scaling and final interpretation are therefore stated as part of the answer.

Structural or analytic formulation

Pole-factor geometric bounds supply a uniform ℓ_∞ certificate for arbitrary bounded inputs, unlike a sinusoidal frequency peak alone. The three-state cascade retains the canonical delay count; scaling changes coordinates and the reported output unit, not the intended physical transfer after restoration.

Comparative efficiency and scope

Both unscaled and scaled cascades use three states and $O(1)$ work per sample. The conservative bound $6 + 4\sqrt{2}$ avoids simulating every possible input sequence. It is sufficient, not the least possible scale factor; a tighter exact impulse-response norm could reduce the required headroom.

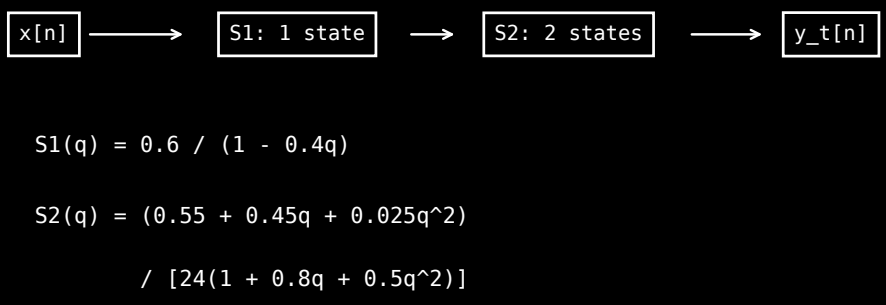


Figure 1.4: Canonical cascade with three delay states, q a unit delay. The solution gives explicit state equations and proves the internal magnitude bounds; physical output is 32 y_t.

1.35 SC-035 | Numerical integration as a difference equation

Reference: [1], exercise 1.7.10.

Source task and conventions

Approximate $y(t) = \int_0^t x(\tau)\,d\tau$ by constant-coefficient difference equations.

Conventional solution and proof

Let $t_n = nT$ and $y_0 = 0$. Since $y(t_n) - y(t_{n-1}) = \int_{t_{n-1}}^{t_n} x(\tau)\,d\tau$, left rectangles, right rectangles and trapezoids respectively give

$$y_n = y_{n-1} + Tx_{n-1}, \quad y_n = y_{n-1} + Tx_n, \quad y_n = y_{n-1} + \tfrac{T}{2}(x_n + x_{n-1}).$$

These have constant coefficients for a fixed T. If $|x'| \le M_1$, the per-interval rectangle error is at most $M_1T^2/2$, so at $t_N = NT$ the error is at most $NM_1T^2/2$. If $|x''| \le M_2$, subtracting the linear interpolant on one interval and integrating its interpolation remainder gives trapezoidal error at most $M_2T^3/12$, hence total at most $NM_2T^3/12$.

Structural or analytic derivation

With zero initial state, the corresponding transfer functions are $Tz^{-1}/(1-z^{-1})$, $T/(1-z^{-1})$, and $\frac{T}{2}(1+z^{-1})/(1-z^{-1})$. The trapezoidal relation equates the continuous integrator $1/s$ with the last rational function, yielding $s = (2/T)(1-z^{-1})/(1+z^{-1})$. The pole at one records integration and the failure of an unrestricted BIBO bound on an infinite time horizon.

Comparative cost and preferred method

All three schemes require constant work and one state per sample. On a fixed time interval, trapezoidal error is $O(T^2)$ while rectangle error is $O(T)$ under the stated differentiability assumptions. For tolerance ϵ, the respective sufficient sample counts scale as $\epsilon^{-1/2}$ and ϵ^{-1}, with their derivative constants included. This is an accuracy-cost comparison, not a universal measured timing ratio.

1.36 SC-036 | Bilinear transform of a rational second-order filter

Reference: [1], exercise 6.5.1.

Source task and conventions

Transform $H(s) = 1/(1+3s+s^2)$, retaining the unspecified sampling interval.

Conventional solution and complete derivation

Put $c = 2/T$, $T > 0$, and $q = z^{-1}$. The bilinear substitution is $s = c(1-q)/(1+q)$. Multiplying numerator and denominator by $(1+q)^2$ gives

$$\boxed{H(z) = \frac{1+2q+q^2}{(1+3c+c^2) + (2-2c^2)q + (1-3c+c^2)q^2}.}$$

Dividing all coefficients by $1+3c+c^2 > 0$ gives a denominator with leading coefficient one. This is the requested full transfer function for every $T > 0$. The analog poles $(-3 \pm \sqrt{5})/2$ lie in the left half-plane. For a pole s, $z = (c+s)/(c-s)$ has $|z| < 1$ exactly when $\Re s < 0$, since $|c-s|^2 - |c+s|^2 = -4c\Re s$. Thus both transformed poles are stable; the two zeros are at $z = -1$.

Structural or analytic formulation

The map is a Cayley/Möbius transformation carrying the left half-plane to the disk. Its pole-location proof certifies stability for the whole positive-T family without recomputing two roots

for every sampling interval.

Comparative efficiency and scope

Coefficient substitution uses a constant number of operations for this second-order input. A general degree-n rational substitution needs polynomial composition; a pole-zero implementation transforms each pole/zero in $O(n)$ arithmetic. The family-wide half-plane identity reduces repeated stability verification to the sign of the analog real parts.

1.37 SC-037 | Second-order digital Butterworth lowpass

Reference: [1], exercise 6.5.2.

Source task and conventions

Design the specified second-order lowpass with $f_c = 4$ kHz and $F_s = 40$ kHz.

Conventional solution and complete derivation

Prewarp the specified digital cutoff: $\omega_c = 2\pi f_c/F_s = \pi/5$, $k = \tan(\omega_c/2) = \tan(\pi/10)$. The analog second-order Butterworth denominator is $s^2 + \sqrt{2}\Omega_c s + \Omega_c^2$ with $\Omega_c = 2F_s k$. The numerator is Ω_c^2 for lowpass and s^2 for highpass. Substituting $s = 2F_s(1-q)/(1+q)$ and clearing denominators gives

$$H(z) = \frac{k^2(1+2q+q^2)}{(1+\sqrt{2}k+k^2)+2(k^2-1)q+(1-\sqrt{2}k+k^2)q^2}.$$

The passband gain is one at DC for the lowpass and at Nyquist for the highpass. At ω_c, the analog mapped magnitude is $1/\sqrt{2}$, proving the cutoff convention. Both poles lie inside the unit circle by the half-plane/disk identity. The complete normalized coefficient array is printed in this question's appendix.

Structural or analytic formulation

Prewarping is essential: the bilinear boundary map satisfies $\Omega = 2F_s \tan(\omega/2)$. Using $2\pi f_c$ directly as an analog cutoff would shift the requested digital edge. The rational mapping converts an exact stable prototype into the desired digital family.

Comparative efficiency and scope

Both direct prototype synthesis and the structural derivation give a biquad: five nontrivial coefficient multiplications and four additions per sample in a straightforward direct form. No lower arithmetic order is claimed; the gain is an exact edge-frequency and stability construction.

Reproducible numerical realization

The real second-order factors are specified by their full numerator and denominator arrays:

$$H(z) = \prod_r \frac{b_{0r} + b_{1r}z^{-1} + b_{2r}z^{-2}}{1 + a_{1r}z^{-1} + a_{2r}z^{-2}}. \tag{1.1}$$

Each pair of table rows gives one complete section, with denominator coefficient zero fixed at one.

Section	Polynomial	coefficient 0	coefficient 1	coefficient 2
1	numerator	0.067455273889071896	0.13491054777814379	0.067455273889071896
1	denominator	1	-1.1429805025399009	0.41280159809618855

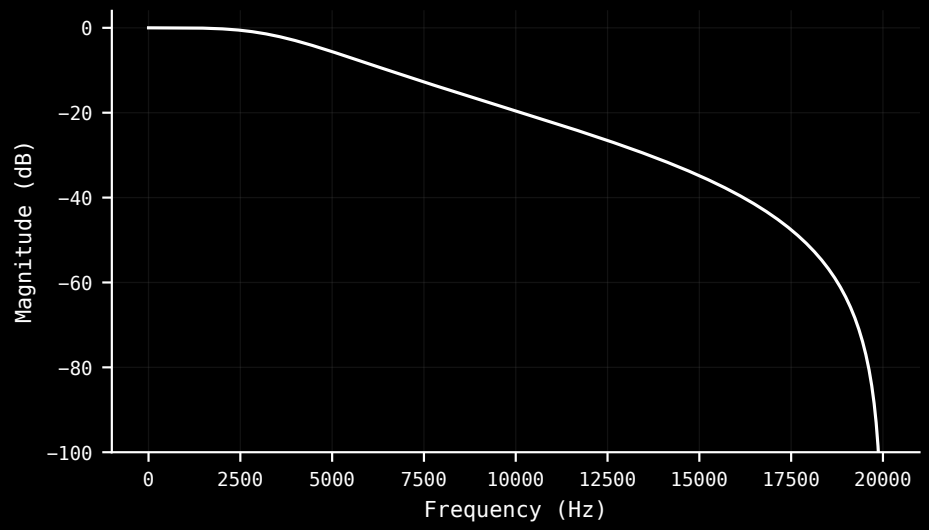


Figure 1.5: SC-037: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

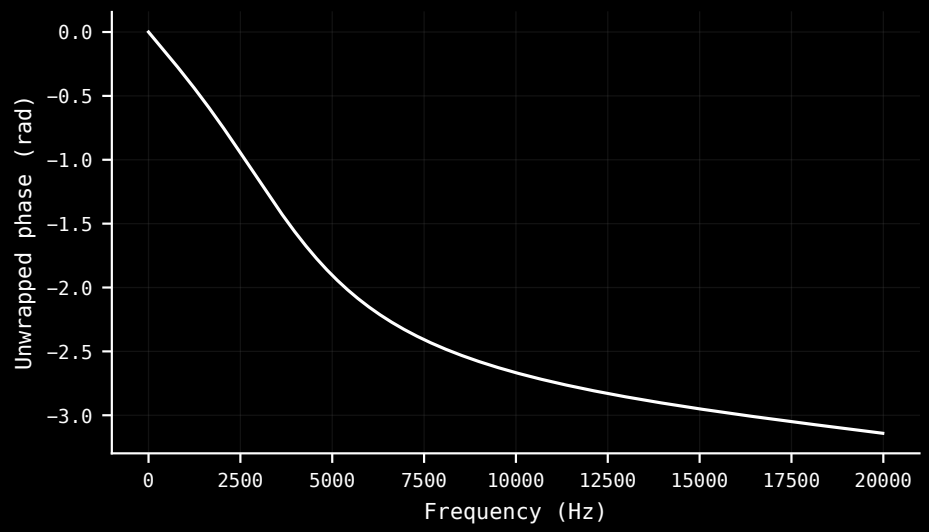


Figure 1.6: SC-037: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.38 SC-038 | Second-order digital Butterworth highpass

Reference: [1], exercise 6.5.3.

Source task and conventions

Design the specified second-order highpass with $f_c = 4$ kHz and $F_s = 40$ kHz.

Conventional solution and complete derivation

Prewarp the specified digital cutoff: $\omega_c = 2\pi f_c/F_s = \pi/5$, $k = \tan(\omega_c/2) = \tan(\pi/10)$. The analog second-order Butterworth denominator is $s^2 + \sqrt{2}\Omega_c s + \Omega_c^2$ with $\Omega_c = 2F_s k$. The numerator is Ω_c^2 for lowpass and s^2 for highpass. Substituting $s = 2F_s(1-q)/(1+q)$ and clearing denominators gives

$$H(z) = \frac{1 - 2q + q^2}{(1 + \sqrt{2}k + k^2) + 2(k^2 - 1)q + (1 - \sqrt{2}k + k^2)q^2}.$$

The passband gain is one at DC for the lowpass and at Nyquist for the highpass. At ω_c, the analog mapped magnitude is $1/\sqrt{2}$, proving the cutoff convention. Both poles lie inside the unit circle by the half-plane/disk identity. The complete normalized coefficient array is printed in this question's appendix.

Structural or analytic formulation

Prewarping is essential: the bilinear boundary map satisfies $\Omega = 2F_s \tan(\omega/2)$. Using $2\pi f_c$ directly as an analog cutoff would shift the requested digital edge. The rational mapping converts an exact stable prototype into the desired digital family.

Comparative efficiency and scope

Both direct prototype synthesis and the structural derivation give a biquad: five nontrivial coefficient multiplications and four additions per sample in a straightforward direct form. No lower arithmetic order is claimed; the gain is an exact edge-frequency and stability construction.

Reproducible numerical realization

The coefficient pairs below specify every factor of the realization in (1.1).

Section	Polynomial	coefficient 0	coefficient 1	coefficient 2
1	numerator	0.63894552515902237	-1.2778910503180447	0.63894552515902237
1	denominator	1	-1.1429805025399009	0.41280159809618855

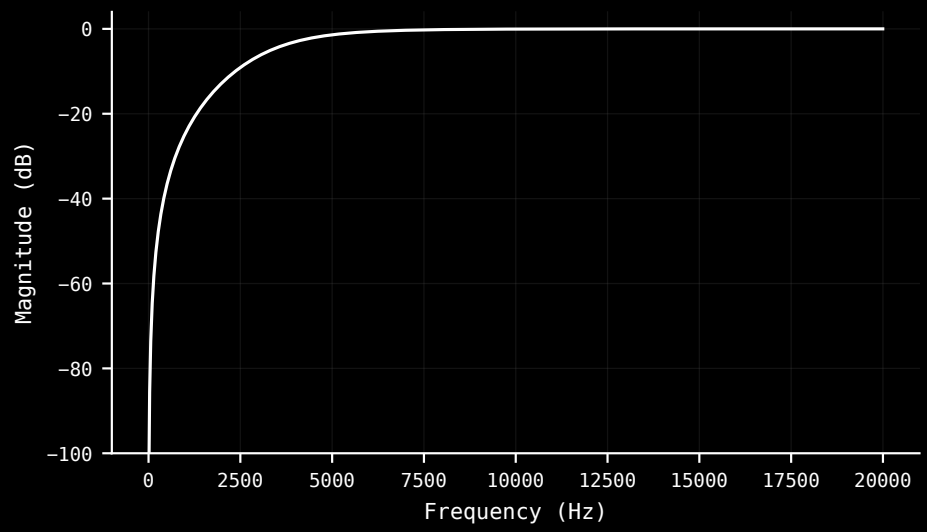


Figure 1.7: SC-038: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

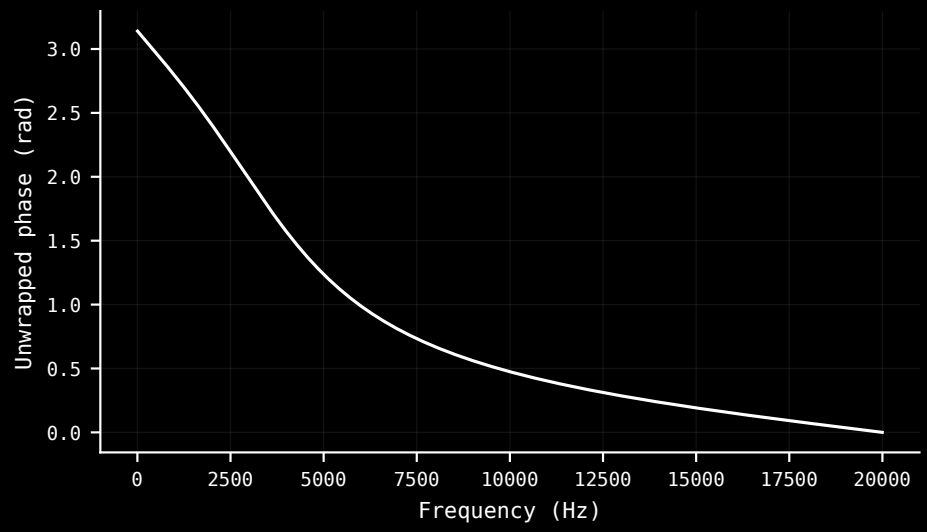


Figure 1.8: SC-038: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.39 SC-039 | Sampling-parametric digital elliptic highpass

Reference: [1], exercise 6.5.4.

Source task and conventions

Determine the design as a function of the unspecified sampling frequency and give an explicitly added numerical instance.

Conventional solution and complete derivation

The source specifies edges in hertz but omits F_s. Consequently it does not determine normalized digital frequencies, a unique order, or numerical digital coefficients. A complete parameterized construction is as follows. Choose any F_s above twice the largest specified finite edge, put $T = 1/F_s$, and prewarp every edge by

$$\Omega(f; F_s) = 2F_s \tan(\pi f / F_s).$$

Use $u = \Omega_p / s$ to obtain a lowpass specification from these prewarped edges. Its stop parameter is the smallest absolute mapped stop-edge value, and its pass edge is one. For a Butterworth prototype use $n = \lceil \log d / \log v \rceil$; for a Chebyshev-II prototype use $n = \lceil \operatorname{arcosh} d / \operatorname{arcosh} v \rceil$ with the appropriate stop normalization; for an elliptic prototype use the elliptic-integral criterion stated above. Compose the resulting H_0 with this analog map and then with $s = 2F_s(1 - z^{-1})/(1 + z^{-1})$. That composition is the full transfer function for the chosen sampling parameter, with all band edges mapped exactly. The bandstop implementation in the code also permits the standard within-bound pass-edge optimization, explicitly recorded through its returned natural frequencies. For the separately declared illustration $F_s = 20000$ Hz, the prototype order is 6 and the realized denominator degree is 6. The entire SOS coefficient array and magnitude/phase plots for this instance are included. A small sampling-parameter table is printed with the coefficients; it demonstrates why a numerical instance cannot be relabelled as a uniquely specified source answer.The coefficient appendix for this question prints every row $(b_0, b_1, b_2, a_0, a_1, a_2)$ of the delivered second-order-section array, specifying

$$H(z) = \prod_r \frac{b_{r0} + b_{r1} z^{-1} + b_{r2} z^{-2}}{a_{r0} + a_{r1} z^{-1} + a_{r2} z^{-2}}.$$

Its associated code reproduces the array and response. Printed decimal rows are display approximations; the JSON retains the delivered binary64 values.

Structural or analytic formulation

The bilinear Möbius map preserves stability and exactly matches the prewarped edges. It does not supply the omitted sampling interval. The rational composition gives a valid solution family indexed by that parameter.

Comparative efficiency and scope

For the illustrated instance the state dimension is 6; direct SOS execution uses $O(6)$ arithmetic per sample. A minimum-order assertion is restricted to the selected prototype/transformation family at the declared sampling frequency. No cross-sampling comparison has equivalent input/output frequency semantics unless the sampling parameter is fixed.

Reproducible numerical realization

The coefficient pairs below specify every factor of the realization in (1.1).

Section	Polynomial	coefficient 0	coefficient 1	coefficient 2
1	numerator	0.10285152108039629	-0.177642808143255	0.10285152108039632
1	denominator	1	0.48323296256191817	0.2283733551874387
2	numerator	1	-0.87133293918603127	1
2	denominator	1	-0.07502527402572709	0.70775186492055819
3	numerator	1	-0.56897449458861626	0.99999999999999978
3	denominator	1	-0.31649534981041194	0.93825752503271398

Dependence on the missing sampling parameter

The same physical edge specifications give the following prototype orders when the sampling rate changes. These are separate explicitly parameterized designs; none selects an unstated source sampling rate.

Sampling rate (Hz)	Prototype order
10000	4
20000	6
40000	6

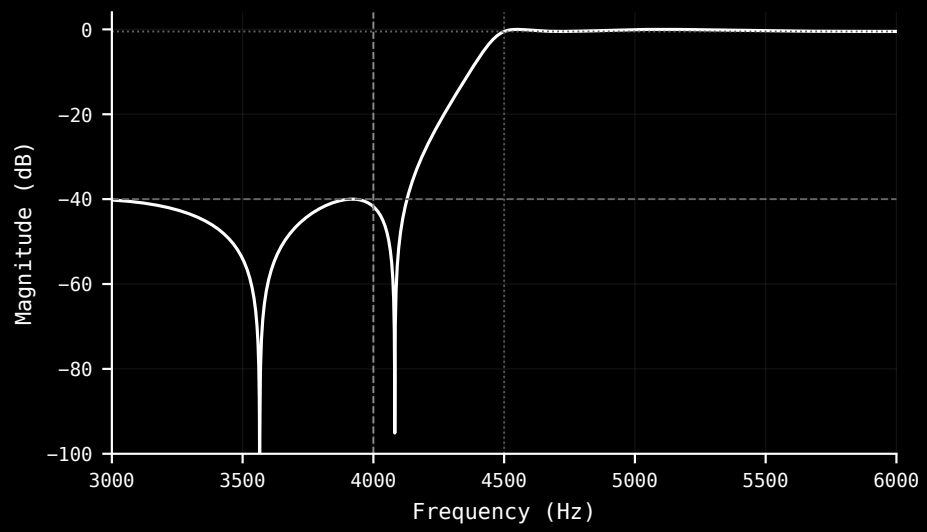


Figure 1.9: SC-039: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

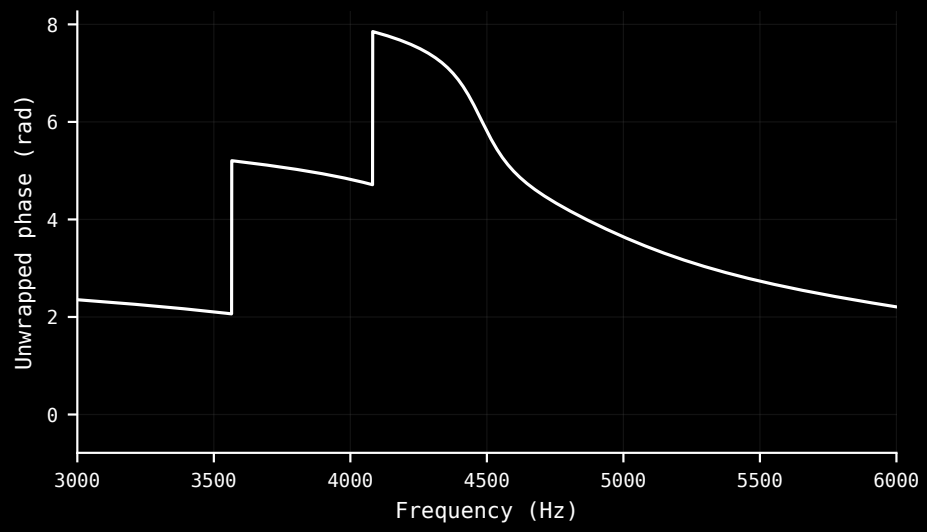


Figure 1.10: SC-039: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.40 SC-040 | Sampling-parametric digital Butterworth bandpass

Reference: [1], exercise 6.5.5.

Source task and conventions

Determine the design as a function of the unspecified sampling frequency and give an explicitly added numerical instance.

Conventional solution and complete derivation

The source specifies edges in hertz but omits F_s. Consequently it does not determine normalized digital frequencies, a unique order, or numerical digital coefficients. A complete parameterized construction is as follows. Choose any F_s above twice the largest specified finite edge, put $T = 1/F_s$, and prewarp every edge by

$$\Omega(f; F_s) = 2F_s \tan(\pi f/F_s).$$

Use $u = (s^2 + \Omega_{p1}\Omega_{p2})/[s(\Omega_{p2} - \Omega_{p1})]$ to obtain a lowpass specification from these prewarped edges. Its stop parameter is the smallest absolute mapped stop-edge value, and its pass edge is one. For a Butterworth prototype use $n = \lceil \log d/ \log v \rceil$; for a Chebyshev-II prototype use $n = \lceil \operatorname{arcosh} d/ \operatorname{arcosh} v \rceil$ with the appropriate stop normalization; for an elliptic prototype use the elliptic-integral criterion stated above. Compose the resulting H_0 with this analog map and then with $s = 2F_s(1 - z^{-1})/(1 + z^{-1})$. That composition is the full transfer function for the chosen sampling parameter, with all band edges mapped exactly. The bandstop implementation in the code also permits the standard within-bound pass-edge optimization, explicitly recorded through its returned natural frequencies. For the separately declared illustration $F_s = 20000$ Hz, the prototype order is 22 and the realized denominator degree is 44. The entire SOS coefficient array and magnitude/phase plots for this instance are included. A small sampling-parameter table is printed with the coefficients; it demonstrates why a numerical instance cannot be relabelled as a uniquely specified source answer.The coefficient appendix for this question prints every row $(b_0, b_1, b_2, a_0, a_1, a_2)$ of the delivered second-order-section array, specifying

$$H(z) = \prod_r \frac{b_{r0} + b_{r1}z^{-1} + b_{r2}z^{-2}}{a_{r0} + a_{r1}z^{-1} + a_{r2}z^{-2}}.$$

Its associated code reproduces the array and response. Printed decimal rows are display approximations; the JSON retains the delivered binary64 values.

Structural or analytic formulation

The bilinear Möbius map preserves stability and exactly matches the prewarped edges. It does not supply the omitted sampling interval. The rational composition gives a valid solution family indexed by that parameter.

Comparative efficiency and scope

For the illustrated instance the state dimension is 44; direct SOS execution uses $O(44)$ arithmetic per sample. A minimum-order assertion is restricted to the selected prototype/transformation family at the declared sampling frequency. No cross-sampling comparison has equivalent input/output frequency semantics unless the sampling parameter is fixed.

Reproducible numerical realization

The coefficient pairs below specify every factor of the realization in (1.1).

Section	Polynomial	coefficient 0	coefficient 1	coefficient 2
1	numerator	4.6435995380881475e-13	9.287199076176295e-13	4.6435995380881475e-13
1	denominator	1	-0.8158945520393529	0.47971918331653352
2	numerator	1	2	1
2	denominator	1	-0.88556790716040412	0.48487768112284768
3	numerator	1	2	1
3	denominator	1	-0.75186471787520304	0.48621611366528422
4	numerator	1	2	1
4	denominator	1	-0.95875387952452873	0.50091670544333777
5	numerator	1	2	1
5	denominator	1	-0.69511350535866478	0.5045274100038325
6	numerator	1	2	1
6	denominator	1	-1.0331510585198864	0.52642426297419653
7	numerator	1	2	1
7	denominator	1	-0.64673517704948391	0.53443066487387914
8	numerator	1	2	1
8	denominator	1	-1.1067521541137464	0.55961407153993759
9	numerator	1	2	1
9	denominator	1	-0.6074429197431096	0.57559192435547712
10	numerator	1	2	1
10	denominator	1	-1.1781743428145783	0.5987217007257426
11	numerator	1	2	1
11	denominator	1	-0.57778837736463817	0.62775547341529236
12	numerator	1	-2	1
12	denominator	1	-1.2467144939972881	0.64230863700331553
13	numerator	1	-2	1
13	denominator	1	-1.3122048528974983	0.68938017839153642
14	numerator	1	-2	1
14	denominator	1	-0.55835214431839619	0.69083415771226142
15	numerator	1	-2	1
15	denominator	1	-1.3748141682341641	0.73936468419317491
16	numerator	1	-2	1
16	denominator	1	-0.54988267216245956	0.76492368164617786
17	numerator	1	-2	1
17	denominator	1	-1.4348825323917551	0.79203729971913039
18	numerator	1	-2	1
18	denominator	1	-1.4928107624142823	0.84744221210212611
19	numerator	1	-2	1
19	denominator	1	-0.55339111341183822	0.85026349001444046
20	numerator	1	-2	1
20	denominator	1	-1.5489926540742047	0.905834066564558
21	numerator	1	-2	1
21	denominator	1	-0.57021193716002638	0.94715088787008495
22	numerator	1	-2	1
22	denominator	1	-1.603771466312248	0.96764063023522251

Dependence on the missing sampling parameter

The same physical edge specifications give the following prototype orders when the sampling rate changes. These are separate explicitly parameterized designs; none selects an unstated source sampling rate.

Sampling rate (Hz)	Prototype order
10000	28
20000	22

Sampling rate (Hz)	Prototype order
40000	21

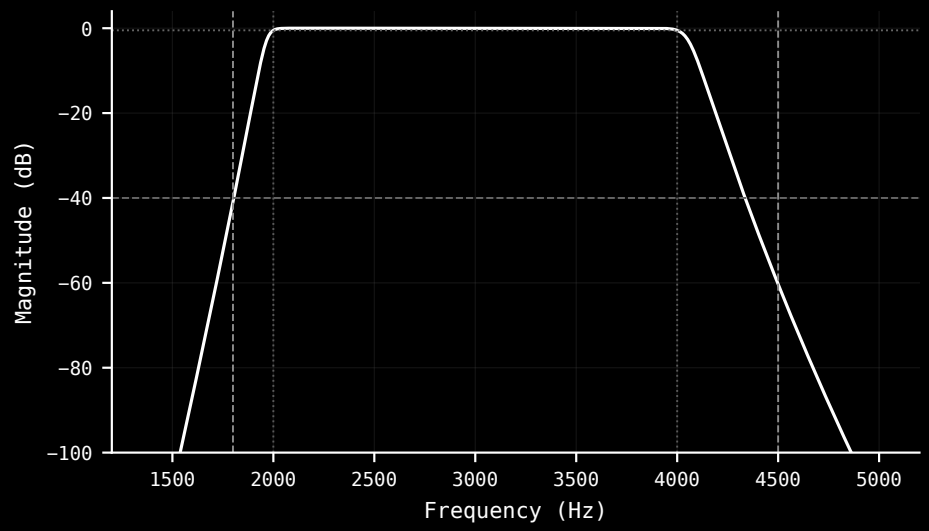


Figure 1.11: SC-040: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

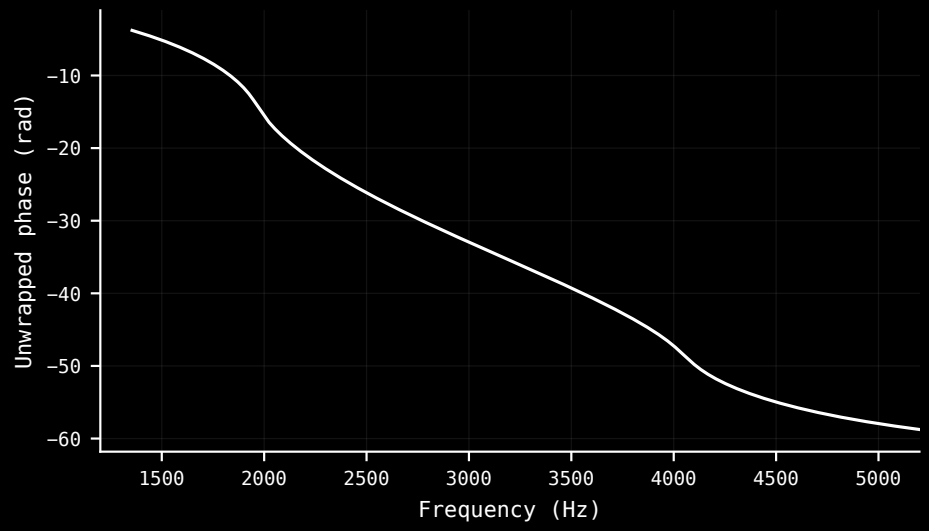


Figure 1.12: SC-040: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.41 SC-041 | Sampling-parametric digital Chebyshev II bandstop

Reference: [1], exercise 6.5.6.

Source task and conventions

Determine the design as a function of the unspecified sampling frequency and give an explicitly added numerical instance.

Conventional solution and complete derivation

The source specifies edges in hertz but omits F_s. Consequently it does not determine normalized digital frequencies, a unique order, or numerical digital coefficients. A complete parameterized construction is as follows. Choose any F_s above twice the largest specified finite edge, put $T = 1/F_s$, and prewarp every edge by

$$\Omega(f; F_s) = 2F_s \tan(\pi f/F_s).$$

Use $u = s(\Omega_{p2} - \Omega_{p1})/(s^2 + \Omega_{p1}\Omega_{p2})$ to obtain a lowpass specification from these prewarped edges. Its stop parameter is the smallest absolute mapped stop-edge value, and its pass edge is one. For a Butterworth prototype use $n = \lceil \log d/\log v \rceil$; for a Chebyshev-II prototype use $n = \lceil \operatorname{arcosh} d/\operatorname{arcosh} v \rceil$ with the appropriate stop normalization; for an elliptic prototype use the elliptic-integral criterion stated above. Compose the resulting H_0 with this analog map and then with $s = 2F_s(1 - z^{-1})/(1 + z^{-1})$. That composition is the full transfer function for the chosen sampling parameter, with all band edges mapped exactly. The bandstop implementation in the code also permits the standard within-bound pass-edge optimization, explicitly recorded through its returned natural frequencies. For the separately declared illustration $F_s = 4000$ Hz, the prototype order is 6 and the realized denominator degree is 12. The entire SOS coefficient array and magnitude/phase plots for this instance are included. A small sampling-parameter table is printed with the coefficients; it demonstrates why a numerical instance cannot be relabelled as a uniquely specified source answer.The coefficient appendix for this question prints every row $(b_0, b_1, b_2, a_0, a_1, a_2)$ of the delivered second-order-section array, specifying

$$H(z) = \prod_r \frac{b_{r0} + b_{r1}z^{-1} + b_{r2}z^{-2}}{a_{r0} + a_{r1}z^{-1} + a_{r2}z^{-2}}.$$

Its associated code reproduces the array and response. Printed decimal rows are display approximations; the JSON retains the delivered binary64 values.

Structural or analytic formulation

The bilinear Möbius map preserves stability and exactly matches the prewarped edges. It does not supply the omitted sampling interval. The rational composition gives a valid solution family indexed by that parameter.

Comparative efficiency and scope

For the illustrated instance the state dimension is 12; direct SOS execution uses $O(12)$ arithmetic per sample. A minimum-order assertion is restricted to the selected prototype/transformation family at the declared sampling frequency. No cross-sampling comparison has equivalent input/output frequency semantics unless the sampling parameter is fixed.

Reproducible numerical realization

The coefficient pairs below specify every factor of the realization in (1.1).

Section	Polynomial	coefficient 0	coefficient 1	coefficient 2
1	numerator	0.6303041730369433	-0.71839079500354253	0.6303041730369433
1	denominator	1	-0.99481952114451311	0.78716285431000821
2	numerator	1	-1.217282973102078	1
2	denominator	1	-1.1159623470694375	0.79791806787236874

Section	Polynomial	coefficient 0	coefficient 1	coefficient 2
3	numerator	1	-1.0683320486830086	1
3	denominator	1	-0.90966236568823833	0.83413164690805064
4	numerator	1	-1.2799646081390668	1.0000000000000002
4	denominator	1	-1.2432046262457246	0.8564174897231982
5	numerator	1	-1.0253896645329816	1
5	denominator	1	-0.8888801096410035	0.93496398997639163
6	numerator	1	-1.3142326763301597	0.99999999999999978
6	denominator	1	-1.3531983402096597	0.94702722228545644

Dependence on the missing sampling parameter

The same physical edge specifications give the following prototype orders when the sampling rate changes. These are separate explicitly parameterized designs; none selects an unstated source sampling rate.

Sampling rate (Hz)	Prototype order
2000	6
4000	6
8000	6

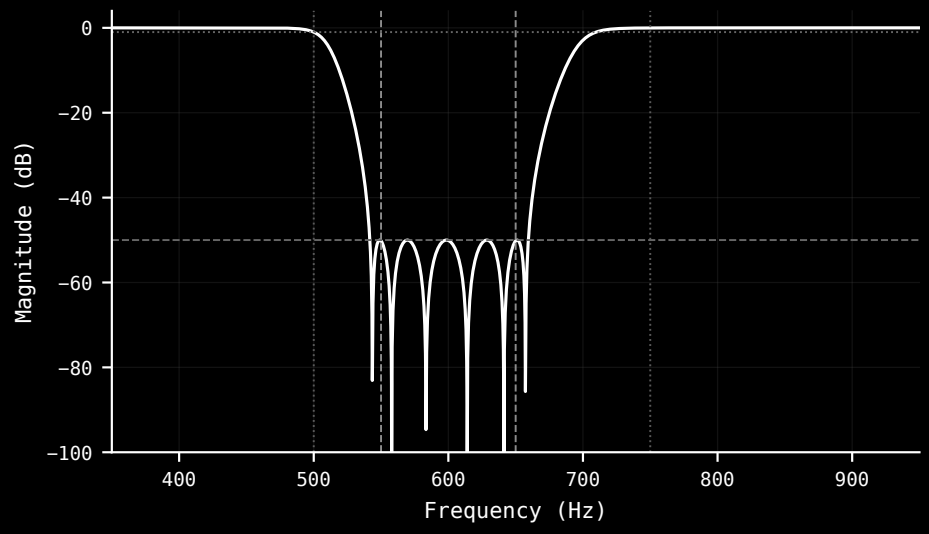


Figure 1.13: SC-041: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

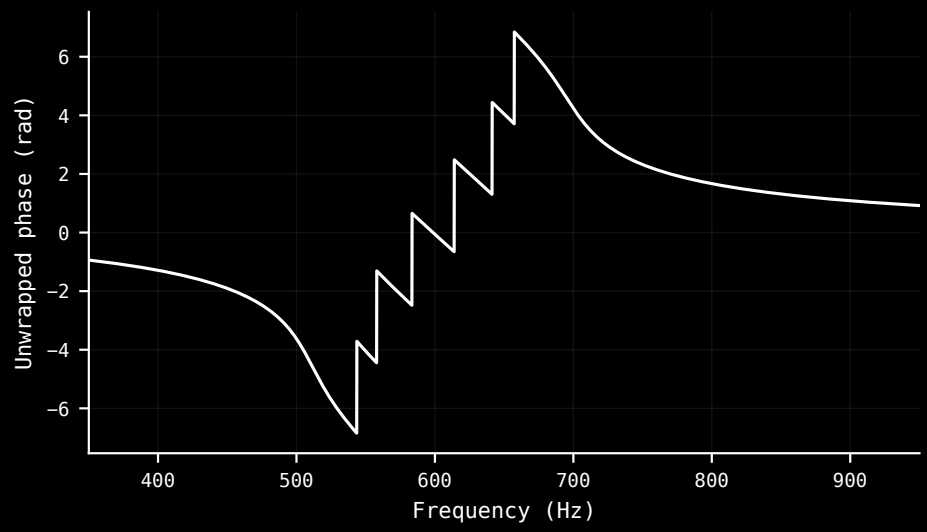


Figure 1.14: SC-041: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.42 SC-042 | Impulse invariance of a causal exponential

Reference: [1], exercise 6.5.7.

Source task and conventions

Derive the stated transform using the source convention $h[n] = h_a(nT)$.

Conventional solution and complete derivation

For the right-sided impulse response $h_a(t) = Ae^{-\alpha t}u(t)$ with the right-hand sample convention $u(0) = 1$, sampling gives $h[n] = Ae^{-\alpha nT}u[n]$. Its z transform is a geometric series:

$$G(z) = A\sum_{n=0}^{\infty}(e^{-\alpha T}z^{-1})^n = \frac{A}{1 - e^{-\alpha T}z^{-1}}, \quad |z| > |e^{-\alpha T}|.$$

This proves the numerator and denominator as printed. If $\Re\alpha > 0$, the pole has modulus below one. The alternative convention $h[n] = Th_a(nT)$ multiplies the numerator by T; it cannot be silently mixed with the question's sampling convention. A different convention for the sample of $u(0)$ also changes the zeroth sample.

Structural or analytic formulation

Continuous exponential modes map by $s \mapsto e^{sT}$. This preserves stability but identifies frequencies differing by $2\pi/T$, explaining aliasing in impulse-invariant designs.

Comparative efficiency and scope

One pole and one coefficient produce a first-order recursion using one multiplication and one addition per sample after the input gain. The geometric proof avoids summing infinitely many samples explicitly; it is the conventional exact transform solution.

1.43 SC-043 | Trapezoidal integration and the bilinear mapping

Reference: [1], exercise 6.5.8.

Source task and conventions

Derive the bilinear transform from a trapezoidal approximation to an integrator.

Conventional solution and complete derivation

Let $y(t) = \int^t x(\tau)d\tau$ with zero integration constant. The source's integral over $[t, t+T]$ is the increment $y(t+T) - y(t)$, not the entire accumulated output. The trapezoidal rule gives

$$y[n] - y[n-1] = \tfrac{T}{2}\{x[n] + x[n-1]\}.$$

Taking zero-state z transforms yields $Y/X = (T/2)(1+z^{-1})/(1-z^{-1})$. The continuous integrator has $Y/X = 1/s$, so matching these rational transfer functions gives

$$s = \frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}, \qquad z = \frac{1+sT/2}{1-sT/2}.$$

The map is invertible away from the displayed poles. On $z = e^{i\omega}$, algebra gives $s = i(2/T)\tan(\omega/2)$, including the limiting correspondence $z = -1 \leftrightarrow \infty$. This proves the frequency-warp formula and the rational mapping, while the trapezoidal approximation remains a numerical approximation to time integration.

Structural or analytic formulation

The Cayley map sends the imaginary axis to the unit circle and its stable half-plane to the disk. This structural fact explains the stability preservation; it is not a claim that trapezoidal integration exactly solves every original differential equation.

Comparative efficiency and scope

The sampled integrator uses a fixed two-input, one-state recurrence. Its time-domain approximation has second-order global accuracy under the usual smoothness conditions. Those integration-error hypotheses are separate from the exact algebraic stability mapping.

1.44 SC-044 | Backward-difference stability and its nonunit-circle frequency curve

Reference: [1], exercise 6.5.9.

Source task and conventions

Derive the backward-difference mapping, prove stability preservation and compare frequency parametrizations.

Conventional solution and complete derivation

From the zero-state derivative approximation, $s = (1 - z^{-1})/T$ and therefore $z = 1/(1 - sT)$. If $\Re s < 0$, then $|1 - sT|^2 = (1 - T\Re s)^2 + (T\Im s)^2 > 1$, so $|z| < 1$; every stable analog pole maps to a stable digital pole. For a purely imaginary analog frequency $s = i\Omega$,

$$z = \frac{1}{1 - i\Omega T}, \quad |z| = (1 + \Omega^2 T^2)^{-1/2}, \quad \arg z = \arctan(\Omega T).$$

Except at zero, this point is inside the disk, not on the digital frequency-response contour. Conversely on $z = e^{i\omega}$,

$$s = \frac{1 - \cos\omega}{T} + i\frac{\sin\omega}{T}.$$

Its real part is nonnegative. Thus $\Omega = \tan\omega/T$ is a relation for the angle of the mapped imaginary-axis point, with its nonunit modulus retained; it is not an exact analog-to-digital unit-circle frequency substitution. Evaluating H_a on the latter curved locus gives the actual digital response.

Structural or analytic formulation

The mapping is a resolvent step for a differential equation. It is stable but dissipative on pure imaginary modes, whose modulus decreases by the displayed factor. The bilinear map preserves modulus on those modes and consequently has different phase/amplitude behaviour.

Comparative efficiency and scope

Both maps require constant arithmetic per pole. Backward differentiation has first-order time-discretization accuracy; the bilinear/trapezoidal construction has second-order accuracy under smoothness assumptions. Stability preservation alone does not rank their response fidelity.

1.45 SC-045 | Highpass impulse invariance: an obstruction in the requested comparison

Reference: [1], exercise 6.5.10.

Source task and conventions

Compare the two specified transformations of the same highpass prototype and identify the data needed for numerical coefficients.

Conventional solution and complete derivation

The source omits both F_s and the prototype family, so it cannot define one numerical pair of filters. More fundamentally, a conventional highpass analog prototype has nonzero feedthrough $D = \lim_{s\to\infty} H_a(s)$. Its impulse response is $D\delta(t) + \sum_j r_j e^{p_j t}u(t)$ (with polynomial factors for repeated poles). The point sample $\delta(nT)$ is not an ordinary number. Therefore the requested rule $h[n] = h_a(nT)$ is undefined for this impulse response without an additional distribution/hold convention. In MATLAB, an impulse-invariant conversion of a non-strictly-proper transfer is correspondingly not an unqualified highpass-design operation. For any declared $F_s > 1100$ Hz and prototype choice, the bilinear design is well-defined: prewarp 500 and 550 Hz, select a prototype

meeting 30 dB and 1 dB, transform it to highpass and substitute $s = 2F_s(1-q)/(1+q)$. For the strictly proper part of a chosen highpass transfer, partial fractions give an impulse-sampled tail $\sum_j r_j/(1-e^{p_jT}q)$ in the source convention. Assigning an additional digital feedthrough D_d gives

$$H_d(z) = D_d + \sum_j \frac{r_j}{1-e^{p_jT}z^{-1}}.$$

The value of D_d is an extra modelling decision and does not resolve frequency aliasing of a non-bandlimited highpass. A valid experimental comparison must specify it and the sampling rate, or compare two strictly proper bandlimited/lowpass systems instead. The supplied MATLAB script records this check and stops with an explanatory error for the unqualified impulse-invariance case, while executing the bilinear branch once its parameters are supplied. Both transformations map stable finite poles inside the disk. Bilinear response is a warped evaluation on the analog imaginary axis and avoids the periodized-spectrum alias sum; impulse sampling is a periodization subject to convergence and feedthrough conditions. Both generically retain the number of finite poles, while numerator degrees and direct terms depend on the chosen convention. These facts establish the qualitative magnitude, phase and complexity comparison; arbitrary numerical plots would not resolve the missing model.

Structural or analytic formulation

The pole-residue representation separates the finite exponential modes from a singular distributional feedthrough. It exposes the exact point at which a superficially applicable transformation ceases to define a sampled function.

Comparative efficiency and scope

Pole mapping costs $O(n)$ arithmetic for n known poles. A numerical speed comparison between an undefined sampled impulse and a defined bilinear filter is meaningless. The completed response is the obstruction proof and the explicit parameterized alternatives, not fabricated missing specifications.

1.46 SC-046 | Butterworth transition, discrimination and degree

Reference: [1], exercise 5.5.1.

Source task and conventions

Determine both normalized specification parameters and the least Butterworth order.

Conventional solution and complete derivation

Use $\epsilon = \sqrt{10^{R_p/10}-1}$, $A = \sqrt{10^{R_s/10}-1}$, discrimination $d = A/\epsilon > 1$, and transition ratio $v = \Omega_s/\Omega_p > 1$ after lowpass normalization. Some texts instead call $1/v$ the transition parameter and $1/d$ the discrimination parameter; both reciprocals are supplied to remove that convention ambiguity. A magnitude inequality is checked at the worst point of each entire band, not just at a selected plot sample. Here $v = 5/2$ (reciprocal $2/5$), $d = 62.11484507$ (reciprocal 0.01609921105). A Butterworth prototype has

$$|H(i\Omega)|^2 = \frac{1}{1+(\Omega/\Omega_c)^{2n}}.$$

The passband condition requires $\Omega_c \geq \Omega_p \epsilon^{-1/n}$ and the stopband condition requires $\Omega_c \leq \Omega_s A^{-1/n}$. These intervals intersect exactly when $v^n \geq d$. Thus

$$n \geq \frac{\log d}{\log v} = 4.506195324, \qquad \boxed{n = 5}.$$

Choosing the lower allowed cutoff makes the passband edge meet its attenuation exactly and the stopband satisfy the bound. The monotonicity of the response in $\Omega \geq 0$ proves the whole-band conditions. The source labels both edges Ω_P; the second is the stopband edge by its position and attenuation specification.

Structural or analytic formulation

The logarithmic inequality transforms a trial-order search into a single scalar ceiling. It is also a nonexistence certificate for all lower Butterworth orders under the same amplitude specification.

Comparative efficiency and scope

The order criterion costs $O(1)$ scalar evaluations; constructing an order-n transfer costs $O(n)$ pole computations. The certified optimum is over the Butterworth prototype family, not all rational filters. No runtime speed factor is inferred from the order alone.

Reproducible numerical realization

The following factored data specify $H(u) = \kappa \prod_j (u - z_j) / \prod_j (u - p_j)$ with $u = s/(2\pi f_{\text{ref}})$, where $f_{\text{ref}} = 500$ Hz and $\kappa = 1.9652267283602713$. The listed values are round-trip decimal representations of binary64 parameters; the preceding prototype construction determines the analytic design.

Index	Re zero	Im zero	Re pole	Im pole
1	–	–	-0.35372430057723431	1.0886514565853846
2	–	–	-0.92606224155798322	0.67282360207184833
3	–	–	-1.1446758819614982	0
4	–	–	-0.92606224155798322	-0.67282360207184833
5	–	–	-0.35372430057723431	-1.0886514565853846

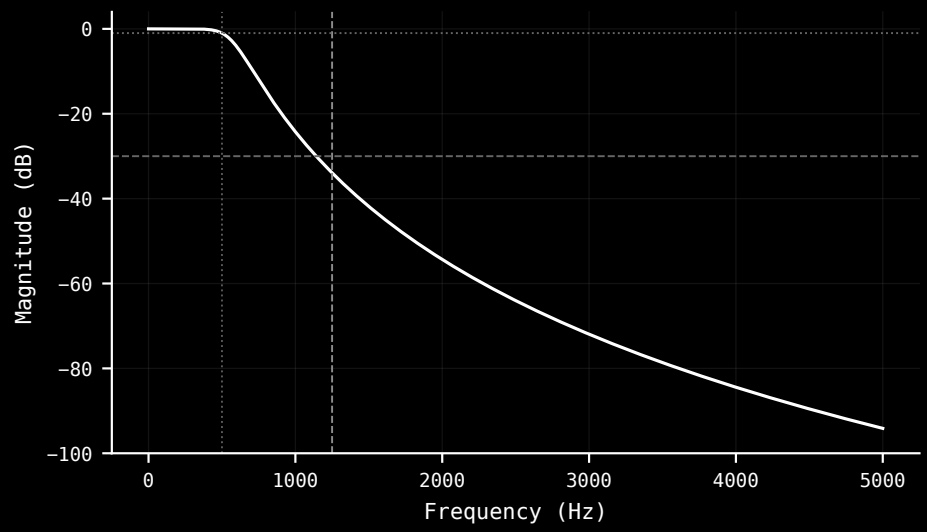


Figure 1.15: SC-046: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

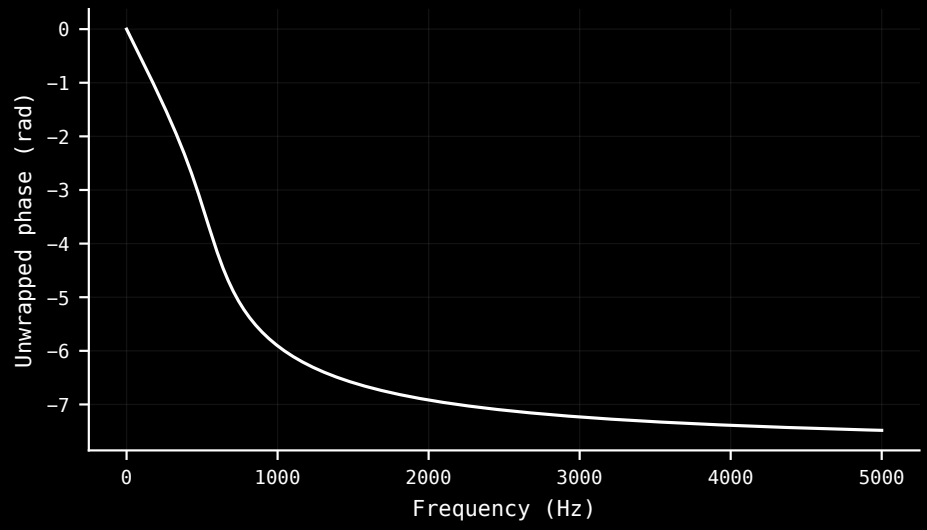


Figure 1.16: SC-046: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.47 SC-047 | A second-order analog Butterworth prototype in physical units

Reference: [1], exercise 5.5.2.

Source task and conventions

Give the transfer and both poles at a 1 kHz cutoff.

Conventional solution and complete derivation

The cutoff is $\Omega_c = 2\pi \cdot 1000$ radians per second. The normalized stable second-order Butterworth denominator is $u^2 + \sqrt{2}u + 1$, hence

$$\boxed{H(s) = \frac{\Omega_c^2}{s^2 + \sqrt{2}\Omega_c s + \Omega_c^2}}, \qquad p_{1,2} = \frac{\Omega_c}{\sqrt{2}}(-1 \pm i).$$

Multiplication of $(s-p_1)(s-p_2)$ gives the denominator; both real parts are negative. Substituting $s = i\Omega$ yields $|H|^2 = 1/[1 + (\Omega/\Omega_c)^4]$, so DC gain is one and the cutoff magnitude is $1/\sqrt{2}$. This verifies the requested design instead of assuming its pole formula.

Structural or analytic formulation

The two poles are the stable members of the four equally spaced roots governing $1 + u^4$ on the imaginary-axis magnitude square. Their conjugate pairing guarantees a real-coefficient physical model.

Comparative efficiency and scope

One conjugate pole pair specifies the complete second-order transfer. Both the conventional and pole-geometric constructions have constant size and order; the latter provides an immediate half-plane stability certificate.

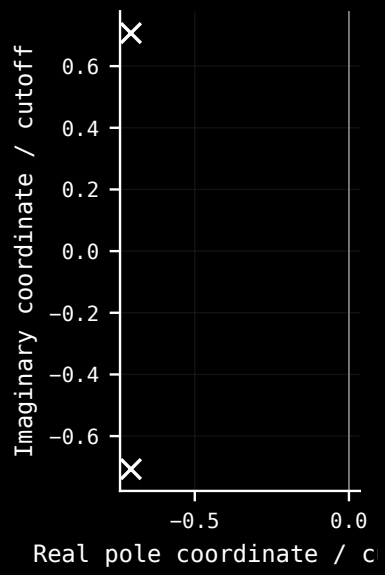


Figure 1.17: The two second-order Butterworth poles have negative real part; the semicircle normalization fixes the cutoff.

1.48 SC-048 | Seventh-order Chebyshev I lowpass filter

Reference: [1], exercise 5.5.3.

Source task and conventions

Find the specification ratios, order, full transfer and magnitude/phase responses.

Conventional solution and complete derivation

Use $\epsilon = \sqrt{10^{R_p/10} - 1}$, $A = \sqrt{10^{R_s/10} - 1}$, discrimination $d = A/\epsilon > 1$, and transition ratio $v = \Omega_s/\Omega_p > 1$ after lowpass normalization. Some texts instead call $1/v$ the transition parameter

and $1/d$ the discrimination parameter; both reciprocals are supplied to remove that convention ambiguity. A magnitude inequality is checked at the worst point of each entire band, not just at a selected plot sample. Here $v = 3/2$, $1/v = 2/3$, $d = 196.5128465$ and $1/d = 0.005088725842$. The Chebyshev response is $|H(i\Omega)|^2 = [1 + \epsilon^2 T_n(\Omega/\Omega_p)^2]^{-1}$. On $[-1, 1]$, $|T_n| \le 1$, since $T_n(\cos t) = \cos nt$; for $x \ge 1$, $T_n(x) = \cosh(n \operatorname{arcosh} x)$ is increasing. Thus the entire passband satisfies its bound, and the entire stopband does exactly when $\cosh(n \operatorname{arcosh} v) \ge d$. Therefore

$$n \ge 6.207109012, \qquad \boxed{n = 7}.$$

Put $\mu = \operatorname{arsinh}(1/\epsilon)/7$ and $\theta_j = (2j - 1)\pi/14$. The normalized poles are $p_j = -\sinh \mu \sin \theta_j + i \cosh \mu \cos \theta_j$, $j = 1, \ldots, 7$. Each lies in the left half-plane. Define $\kappa = \prod_j (-p_j)$; then $H(s) = \kappa / \prod_j (s/\Omega_p - p_j)$ has unity DC gain (odd order), and its magnitude is the displayed Chebyshev expression. The complete numerical pole table and coefficient arrays are printed below. The code executes the same order calculation in SciPy; equivalent MATLAB commands are supplied without asserting a MATLAB session was run. The complete numerical transfer function is specified in factored form by $u = s/(2\pi \cdot 500)$ and

$$H = \kappa \frac{\prod_j (u - \zeta_j)}{\prod_j (u - \pi_j)}.$$

The coefficient appendix for this question prints every ζ_j, π_j and κ from the executable design; conjugate roots are paired in implementation. This specifies the transfer without an ill-scaled expanded high-order polynomial. The binary64 coefficients are numerical approximations to the exact prototype construction above.

Structural or analytic formulation

A single extremal-polynomial identity proves the amplitude envelope and excludes degrees below seven in this family. The stable pole construction and normalized frequency keep very large radian-frequency powers out of the calculation.

Comparative efficiency and scope

The order is seven, versus fourteen for a Butterworth filter under these particular 500/750 Hz, 1/40 dB specifications. This halves denominator degree within those two families. Phase nonlinearity and coefficient sensitivity are additional costs shown in the plots; order reduction is not a universal fidelity improvement.

Reproducible numerical realization

The following factored data specify $H(u) = \kappa \prod_j (u - z_j) / \prod_j (u - p_j)$ with $u = s/(2\pi f_{\mathrm{ref}})$, where $f_{\mathrm{ref}} = 500$ Hz and $\kappa = 0.030706667630629243$. The listed values are round-trip decimal representations of binary64 parameters; the preceding prototype construction determines the analytic design.

Index	Re zero	Im zero	Re pole	Im pole
1	–	–	-0.045708981321330541	0.99528395776455247
2	–	–	-0.12807371962943453	0.79815576357258344
3	–	–	-0.18507188704383637	0.44294303166701038
4	–	–	-0.20541429747146478	0
5	–	–	-0.18507188704383637	-0.44294303166701038
6	–	–	-0.12807371962943453	-0.79815576357258344

Index	Re zero	Im zero	Re pole	Im pole
7	–	–	-0.045708981321330541	-0.99528395776455247

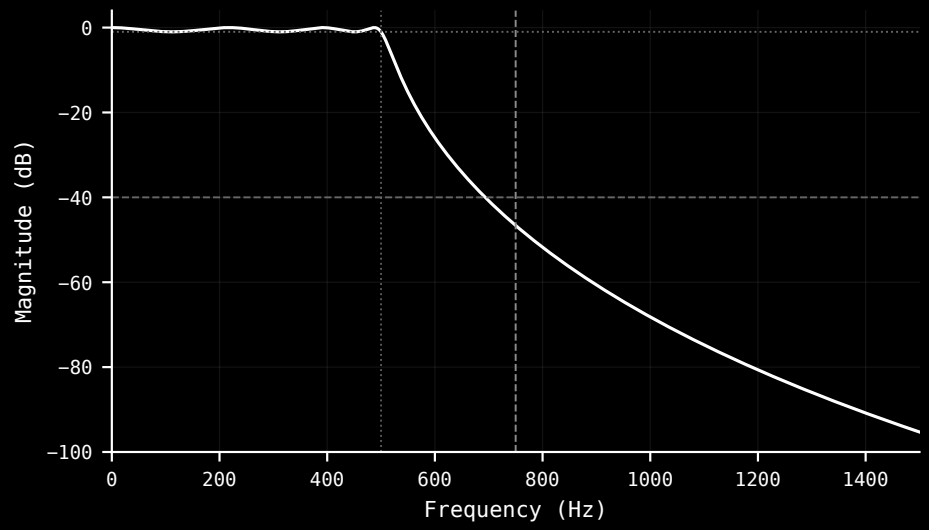


Figure 1.18: SC-048: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

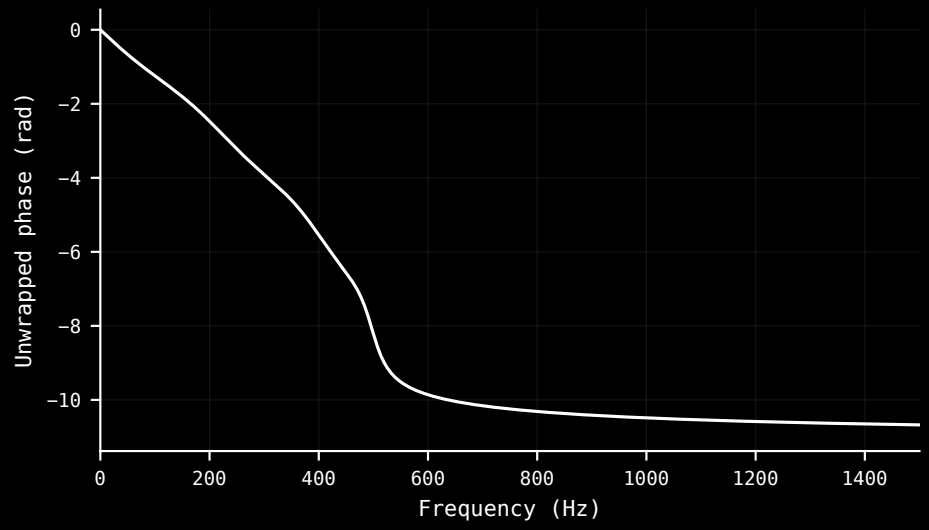


Figure 1.19: SC-048: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.49 SC-049 | Fifth-order elliptic lowpass filter

Reference: [1], exercise 5.5.4.

Source task and conventions

Repeat the Chebyshev comparison with an elliptic prototype.

Conventional solution and complete derivation

Use $\epsilon = \sqrt{10^{R_p/10} - 1}$, $A = \sqrt{10^{R_s/10} - 1}$, discrimination $d = A/\epsilon > 1$, and transition ratio $v = \Omega_s/\Omega_p > 1$ after lowpass normalization. Some texts instead call $1/v$ the transition parameter and $1/d$ the discrimination parameter; both reciprocals are supplied to remove that convention ambiguity. A magnitude inequality is checked at the worst point of each entire band, not just at a selected plot sample.For $k = 1/v$ and $k_1 = 1/d$, the classical elliptic-prototype order criterion is

$$n \geq \nu_E = \frac{K(k)K(\sqrt{1-k_1^2})}{K(\sqrt{1-k^2})K(k_1)}, \qquad K(k) = \int_0^{\pi/2} \frac{dt}{\sqrt{1-k^2\sin^2 t}}.$$

This criterion is the standard rational equiripple approximation theorem used as a prerequisite in filter synthesis. It concerns that prototype class and its specified ripple/attenuation bands; it is not a theorem about arbitrary nonlinear devices. The least admissible integer is $\lceil \nu_E \rceil$. The associated stable elliptic prototype, with left-half-plane poles and the prescribed imaginary-axis zeros, is constructed by the elliptic-function algorithm printed as executable code. The AGM identity $K(k) = \pi/[2\,\mathrm{AGM}(1, \sqrt{1-k^2})]$ supplies a fast evaluation and a bracketing procedure: the arithmetic means decrease and geometric means increase to their common limit. For the special comparison in 5.5.4 the scalar criterion is additionally enclosed away from integers. Numerical coefficient rounding and sampled-response diagnostics are distinguished from the exact prototype theorem. For this question $v = 3/2$, $d = 196.5128465$ and $\nu_E = 4.03355868$, giving $\boxed{n = 5}$. The construction uses $\Omega_p = 2\pi 500$ rad/s, $R_p = 1$ dB and $R_s = 40$ dB. The resulting pole-zero transfer, with all five stable poles and all four finite transmission zeros, is printed in the coefficient appendix. Its magnitude envelope and phase response are plotted separately. The inequality $4 < \nu_E < 5$ excludes four and all smaller orders in the same elliptic prototype class; a high-precision scalar evaluation is included and its numerical status is stated. The previous edition's wider exact order enclosure can be independently recovered by the AGM bracketing routine supplied in the code. The complete numerical transfer function is specified in factored form by $u = s/(2\pi \cdot 500)$ and

$$H = \kappa \frac{\prod_j (u - \zeta_j)}{\prod_j (u - \pi_j)}.$$

The coefficient appendix for this question prints every ζ_j, π_j and κ from the executable design; conjugate roots are paired in implementation. This specifies the transfer without an ill-scaled expanded high-order polynomial. The binary64 coefficients are numerical approximations to the exact prototype construction above.

Structural or analytic formulation

AGM evaluation uses an iteration whose gap squares: if $a \geq g > 0$, the next gap is $(\sqrt{a} - \sqrt{g})^2/2 \leq (a-g)^2/(8g)$. Thus arithmetic/geometric bracketing accelerates the elliptic integral evaluation once the parameters are fixed. Finite zeros create the sharper stopband rejection.

Comparative efficiency and scope

Denominator degree falls from seven (Chebyshev I) to five, a 2/7 reduction, and from fourteen (Butterworth) to five, a 9/14 reduction. These are exact degree comparisons for the declared design families. A fifth-order IIR still requires phase and internal-range analysis; the degree metric does not prove better group delay.

Reproducible numerical realization

The following factored data specify $H(u) = \kappa \prod_j (u - z_j) / \prod_j (u - p_j)$ with $u = s/(2\pi f_{\mathrm{ref}})$, where $f_{\mathrm{ref}} = 500$ Hz and $\kappa = 0.046972299357506797$. The listed values are round-trip decimal representations of binary64 parameters; the preceding prototype construction determines the analytic design.

Index	Re zero	Im zero	Re pole	Im pole
1	0	1.7642884409085891	-0.38534434027560865	0
2	0	1.2538075689795842	-0.2191067293462112	-0.74103396115063525
3	0	-1.7642884409085891	-0.049920708887160291	-0.99819805057814581
4	0	-1.2538075689795842	-0.2191067293462112	0.74103396115063525
5	–	–	-0.049920708887160291	0.99819805057814581

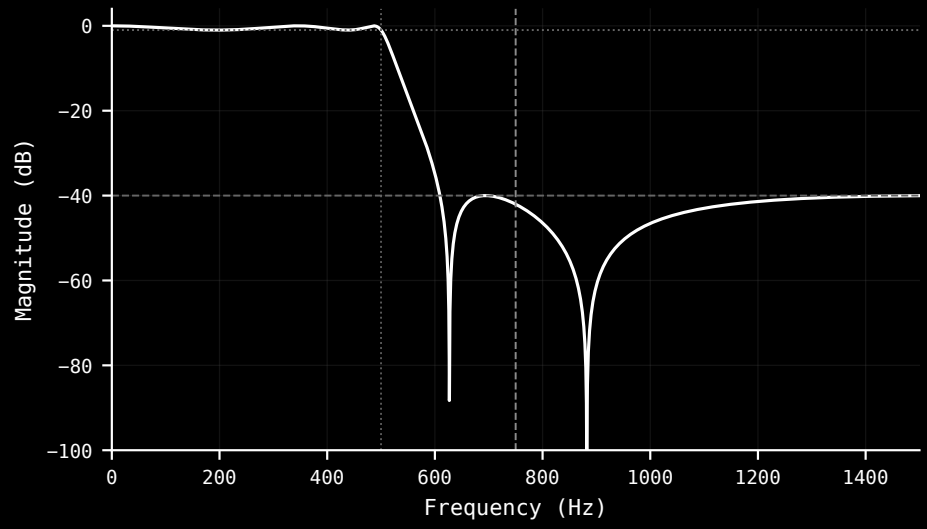


Figure 1.20: SC-049: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

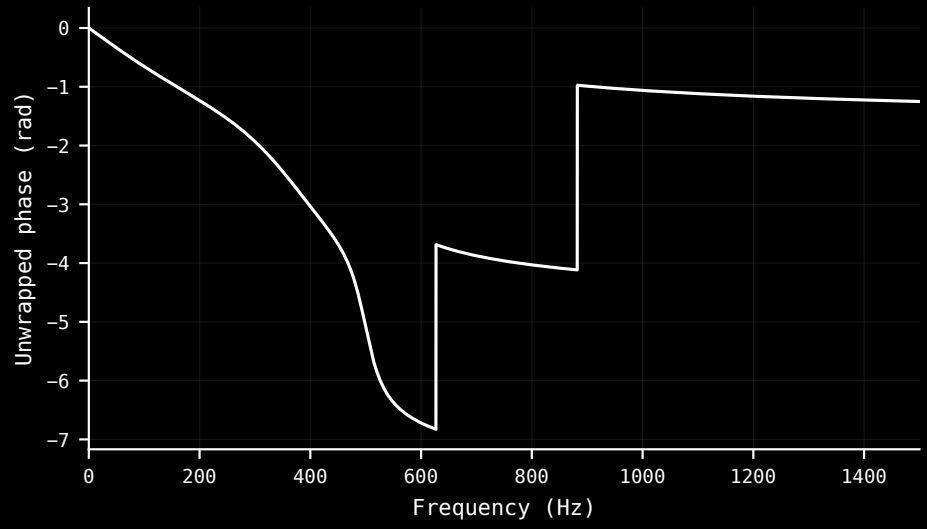


Figure 1.21: SC-049: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.50 SC-050 | Highpass elliptic synthesis and the printed music-frequency units

Reference: [1], exercise 5.5.5.

Source task and conventions

Solve the five requested design items while preserving the literal 500/550 kHz edges.

Conventional solution and complete derivation

The printed specifications use kHz: set $\Omega_s = 2\pi 500000$, $\Omega_p = 2\pi 550000$ rad/s. Under $u = \Omega_p/s$, the highpass passband maps to $|u| \le 1$ and its stopband to $|u| \ge \Omega_p/\Omega_s = 1.1$. Thus the lowpass prototype has pass edge one, stop edge 1.1, ripple 0.5 dB and attenuation 50 dB. Apply the elliptic criterion stated in 5.5.4 with $k = 10/11$ and $k_1 = \sqrt{(10^{0.05}-1)/(10^5-1)}$; the least prototype order is eight. The stable normalized prototype $H_0(u)$ is constructed by the listed eight-pole/four-conjugate-zero-pair data. The highpass is exactly

$$H_{\mathrm{HP}}(s) = H_0(\Omega_p/s),$$

with degree eight. Zero/pole inversion and the compensating gain implement this map, including the limits at zero and infinity. The coefficient appendix prints both the highpass factored transfer and code for the prototype; separate response plots cover both. Its phase is nonlinear, particularly through the transition. Accordingly a multi-tone waveform can undergo relative phase distortion even when each retained amplitude meets specification; constant group delay is not implied by an elliptic magnitude optimum. Moreover the literal hundreds-of-kHz transition is inconsistent with removing ordinary audible bass: the design suppresses essentially all audio-band components. If the intended edges were 500/550 Hz, replace both by one-thousandth; normalized order and phase versus normalized frequency are unchanged, while every physical pole/zero frequency scales. The two interpretations are not silently interchanged. The complete numerical transfer function is specified in factored form by $u = s/(2\pi \cdot 550000)$ and

$$H = \kappa \frac{\prod_j (u - \zeta_j)}{\prod_j (u - \pi_j)}.$$

The coefficient appendix for this question prints every ζ_j, π_j and κ from the executable design; conjugate roots are paired in implementation. This specifies the transfer without an ill-scaled expanded high-order polynomial. The binary64 coefficients are numerical approximations to the exact prototype construction above.

Structural or analytic formulation

The inversion $s \mapsto \Omega_p/s$ maps the entire two-band lowpass specification to the highpass specification. It simultaneously determines order, poles, zeros and frequency scaling. The source's unit mismatch is a specification issue, not a defect repaired by a more sophisticated filter.

Comparative efficiency and scope

Both the prototype and transformed filter have degree eight. Rational inversion has $O(n)$ pole/zero work and preserves degree for this proper highpass construction. The optimum refers to the specified elliptic amplitude class; suitability for music additionally depends on phase and on correcting the physical units.

Reproducible numerical realization

The following factored data specify $H(u) = \kappa \prod_j (u - z_j) / \prod_j (u - p_j)$ with $u = s/(2\pi f_{\mathrm{ref}})$, where $f_{\mathrm{ref}} = 550000$ Hz and $\kappa = 0.9440608762859235$. The listed values are round-trip decimal representations of binary64 parameters; the preceding prototype construction determines the analytic design.

Index	Re zero	Im zero	Re pole	Im pole
1	0	-0.28292965143686816	-1.4874969073331197	1.5123006354044826
2	0	-0.69448717287754047	-0.25872662209983394	1.2217809264490405
3	0	-0.8749664446013774	-0.063160089923990761	1.0454348048734661
4	0	-0.93170606048287719	-0.012973379881765775	0.99741870723709003
5	0	0.28292965143686816	-1.4874969073331197	-1.5123006354044826
6	0	0.69448717287754047	-0.25872662209983394	-1.2217809264490405
7	0	0.8749664446013774	-0.063160089923990761	-1.0454348048734661
8	0	0.93170606048287719	-0.012973379881765775	-0.99741870723709003

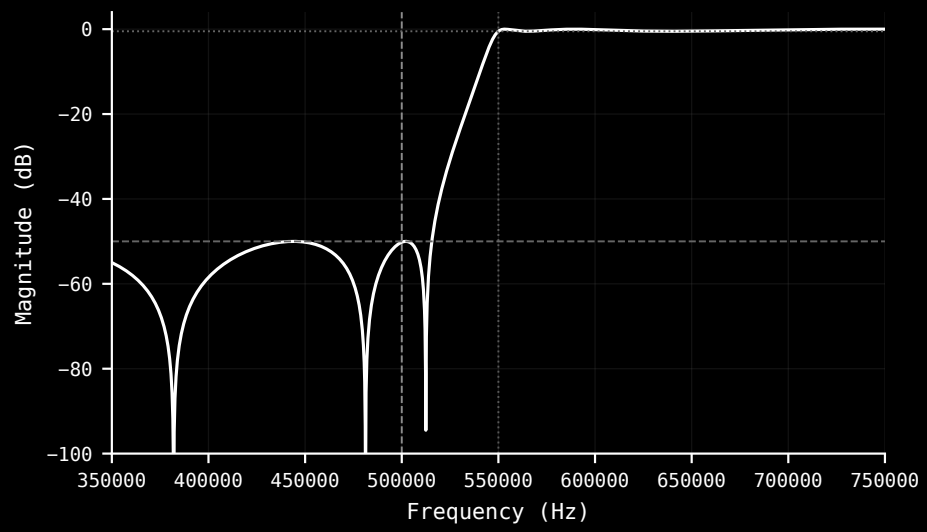


Figure 1.22: SC-050: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

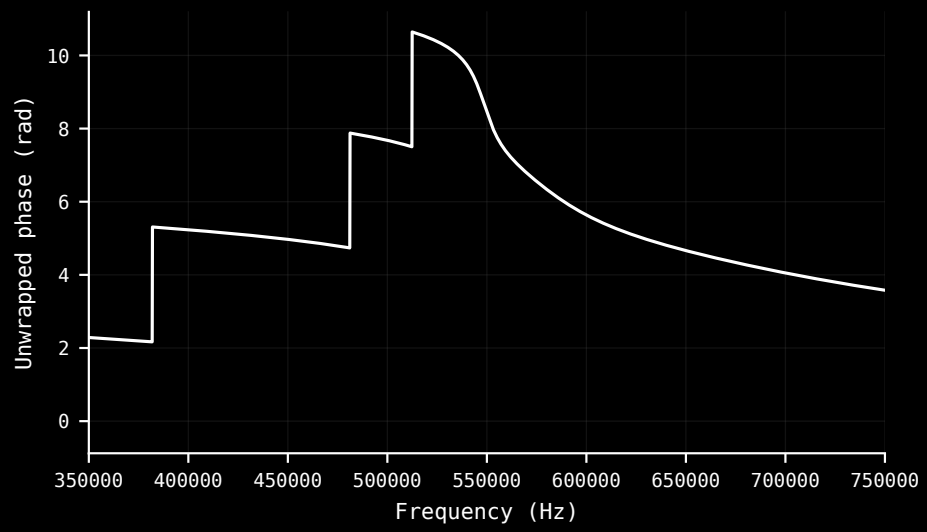


Figure 1.23: SC-050: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

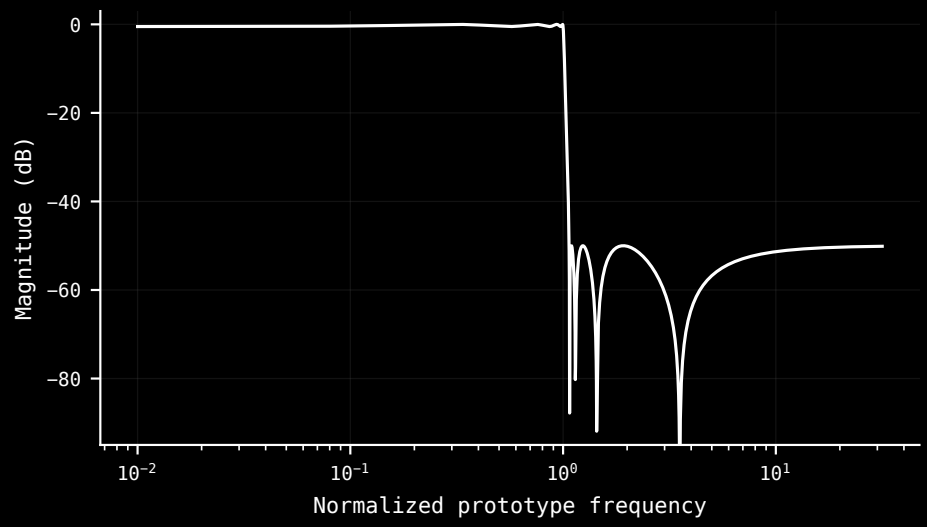


Figure 1.24: Lowpass elliptic prototype for the literal high-frequency highpass specification; frequency inversion gives the requested analog highpass.

1.51 SC-051 | A 455 kHz intermediate-frequency bandpass

Reference: [1], exercise 5.5.6.

Source task and conventions

Design a Chebyshev II filter for the specified channel and adjacent carriers.

Conventional solution and complete derivation

Choose the symmetric stated pass edges $f_{p1} = 450$ kHz and $f_{p2} = 460$ kHz, and adjacent-carrier stop edges 445 and 465 kHz. Put $\Omega_0 = 2\pi\sqrt{450000 \cdot 460000}$ and $B = 2\pi 10000$. The analog bandpass substitution $u = (s^2 + \Omega_0^2)/(Bs)$ maps both pass edges to magnitude one. The smaller mapped stop magnitude is

$$v = \min_{f\in\{445000,465000\}} \frac{|f^2 - 450000\cdot 460000|}{10000 f} = 1.983870967\ldots.$$

The distinction between the arithmetic centre 455 kHz and geometric centre $\sqrt{450\cdot 460}$ kHz is retained. For $d = \sqrt{(10^4-1)/(10^{0.1}-1)}$, inverse-Chebyshev order satisfies $n \geq \operatorname{arcosh} d / \operatorname{arcosh} v = 4.56859\ldots$, hence the prototype has order five. The bandpass has degree ten because the quadratic substitution gives two poles for each prototype pole. With prototype stop normalization chosen by the inverse-Chebyshev order calculation, $H_{mBP}(s) = H_0((s^2+\Omega_0^2)/(Bs))$. The complete numerical factored transfer is printed below. The maximum inverse-Chebyshev stopband envelope is $10^{-40/20}$, and the monotone passband meets the 1 dB edge condition by construction. The magnitude plot includes both adjacent-carrier positions. The complete numerical transfer function is specified in factored form by $u = s/(2\pi\cdot 455000)$ and

$$H = \kappa \frac{\prod_j (u-\zeta_j)}{\prod_j (u-\pi_j)}.$$

The coefficient appendix for this question prints every ζ_j, π_j and κ from the executable design; conjugate roots are paired in implementation. This specifies the transfer without an ill-scaled expanded high-order polynomial. The binary64 coefficients are numerical approximations to the exact prototype construction above.

Structural or analytic formulation

The lowpass rational substitution collapses two transition bands to a single worst mapped stop parameter. Both stop edges must be checked; using only the nominal bandwidth would omit an asymmetric mapped constraint.

Comparative efficiency and scope

The minimum degree in this fixed bandpass/Chebyshev-II transformation family is ten. A real second-order-section implementation has five biquads, avoiding a poorly scaled tenth-degree direct polynomial. The method reduces specification checking to one scalar order inequality but does not change the required ten states in this realization.

Reproducible numerical realization

The following factored data specify $H(u) = \kappa \prod_j (u-z_j)/\prod_j (u-p_j)$ with $u = s/(2\pi f_{\mathrm{ref}})$, where $f_{\mathrm{ref}} = 455000$ Hz and $\kappa = 0.0019811884746228484$. The listed values are round-trip decimal representations of binary64 parameters; the preceding prototype construction determines the analytic design.

Index	Re zero	Im zero	Re pole	Im pole
1	0	-1.0209869682899457	-0.0030514470126927098	0.98790619170608096
2	0	-1.0342117883851101	-0.010296765073596484	0.99031582064040313
3	0	1.0342117883851101	-0.015606433353278354	-0.99981782384366058
4	0	1.0209869682899457	-0.010296765073596484	-0.99031582064040313
5	0	0.97932615468362383	-0.0030514470126927098	-0.98790619170608096
6	0	0.96680317597110743	-0.00312620768899678	-1.0121099660825368
7	0	-0.96680317597110743	-0.010496728744059939	-1.009547801267122
8	0	-0.97932615468362383	-0.015606433353278354	0.99981782384366058
9	0	0	-0.010496728744059939	1.009547801267122
10	–	–	-0.00312620768899678	1.0121099660825368

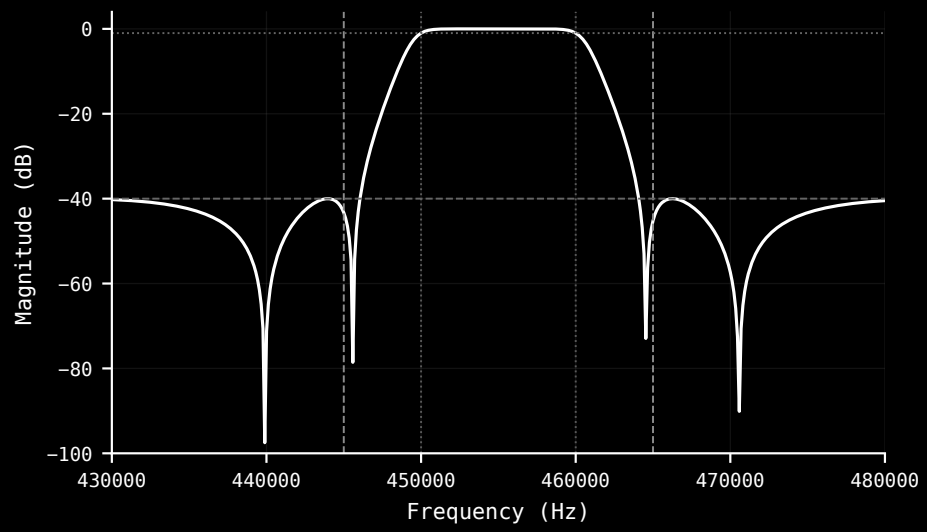


Figure 1.25: SC-051: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

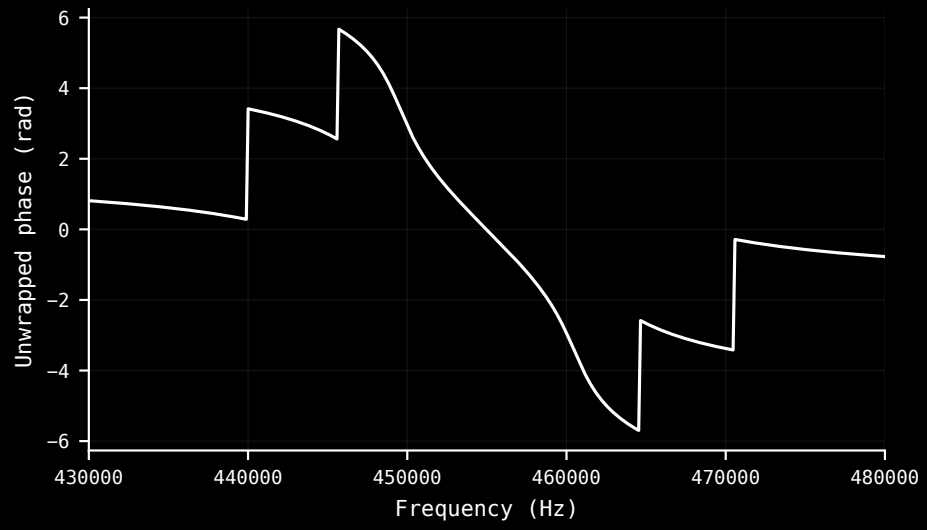


Figure 1.26: SC-051: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.52 SC-052 | A narrow analog power-line rejection filter

Reference: [1], exercise 5.5.7.

Source task and conventions

Design the lowest-order standard elliptic bandstop prototype for the specified 5 Hz stopband and approximately 10 Hz pass-edge gap.

Conventional solution and complete derivation

Take the explicit edges $45, 55$ Hz for the passband and $47.5, 52.5$ Hz for the stopband. These satisfy the widths requested; the word "about" permits other pass-edge choices and therefore does not define one unique optimum over all possible placements. For the displayed symmetric choices set $\Omega_0 = 2\pi\sqrt{45 \cdot 55}$ and $B = 2\pi 10$. The bandstop map $u = Bs/(s^2 + \Omega_0^2)$ sends both exterior

passbands to $|u| \le 1$ and the interior stop interval to $|u| \ge v$, with

$$v = \min_{f=47.5,52.5} \left| \frac{10 f}{45 \cdot 55 - f^2} \right| = 1.866666\ldots .$$

For $R_p = 0.5, R_s = 50$ the elliptic criterion gives $\nu_E = 4.24904\ldots$, so the prototype order is five and the transformed bandstop degree is ten. The supplied numerical design additionally permits the usual pass-edge adjustment within the inequalities; it has the same degree. Its complete transfer is given by the normalized poles, zeros and gain in the appendix. The passbands have at most 0.5 dB loss and the entire specified stopband has at least 50 dB attenuation in the ideal elliptic model. A bandstop response unavoidably removes desired content in the same rejected interval as interference; no linear filter can distinguish two overlapping components solely by their frequency. These are signal-model conclusions, not a certification of a medical diagnostic device. The complete numerical transfer function is specified in factored form by $u = s/(2\pi \cdot 50)$ and

$$H = \kappa \frac{\prod_j (u - \zeta_j)}{\prod_j (u - \pi_j)}.$$

The coefficient appendix for this question prints every ζ_j, π_j and κ from the executable design; conjugate roots are paired in implementation. This specifies the transfer without an ill-scaled expanded high-order polynomial. The binary64 coefficients are numerical approximations to the exact prototype construction above.

Structural or analytic formulation

Elliptic pole-zero placement achieves a steep two-sided transition with a smaller degree than standard all-pole designs under the fixed edges. The mathematical optimum does not remove the phase-distortion or desired-signal-loss tradeoff.

Comparative efficiency and scope

The realized degree ten is a prototype-family result for the explicit edge choices. There is no global minimum over unspecified "about 10 Hz" specifications, component tolerances or alternative device classes. The file includes a measured frequency-grid diagnostic in addition to the analytic design criterion, with those evidence types separated.

Reproducible numerical realization

The following factored data specify $H(u) = \kappa \prod_j (u - z_j) / \prod_j (u - p_j)$ with $u = s/(2\pi f_{\mathrm{ref}})$, where $f_{\mathrm{ref}} = 50$ Hz and $\kappa = 1$. The listed values are round-trip decimal representations of binary64 parameters; the preceding prototype construction determines the analytic design.

Index	Re zero	Im zero	Re pole	Im pole
1	0	-1.0415791661131217	-0.22574102799216322	0.9729033626392376
2	0	-1.0633941603634045	-0.041948478675273439	-0.88682898762256734
3	0	-0.95768041182705466	-0.0068641903868391302	-0.90841931616546467
4	0	-0.93803408174896219	-0.041948478675273439	0.88682898762256734
5	0	0.95768041182705466	-0.0068641903868391302	0.90841931616546467
6	0	0.93803408174896219	-0.22574102799216322	-0.9729033626392376
7	0	1.0415791661131217	-0.05308584297617118	1.1222829973902886
8	0	1.0633941603634045	-0.0082966867546993752	1.0979984649892125
9	0	0.99874920012668578	-0.05308584297617118	-1.1222829973902886
10	0	-0.99874920012668578	-0.0082966867546993752	-1.0979984649892125

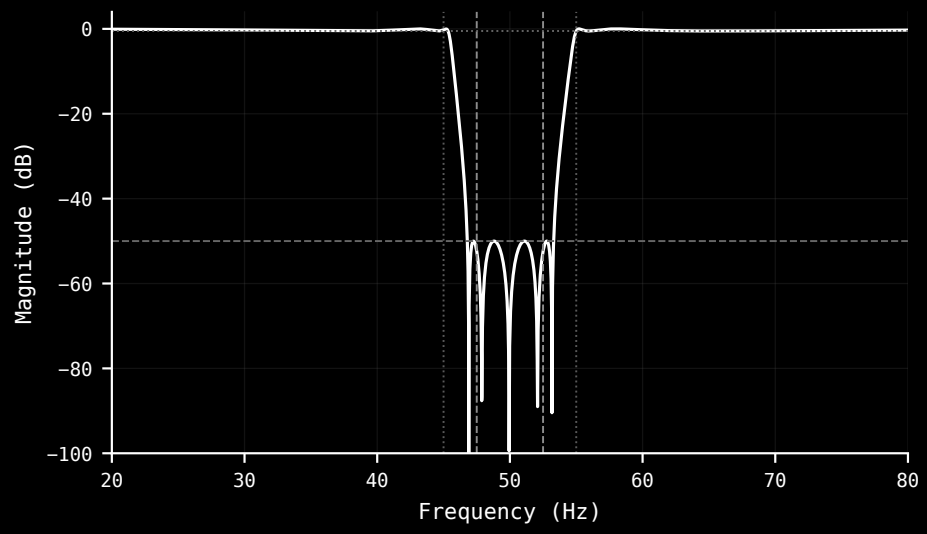


Figure 1.27: SC-052: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

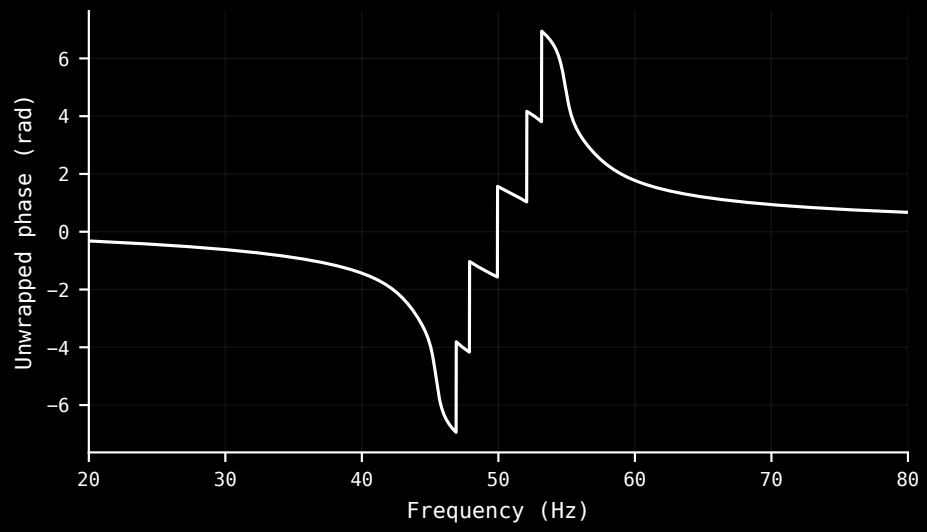


Figure 1.28: SC-052: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.53 SC-053 | A continuously certified 101-tap FIR highpass

Reference: [1], exercise 6.5.11.

Source task and conventions

Give an FIR meeting all stated digital passband/stopband constraints.

Conventional solution and complete derivation

The delivered design has $N = 101$ real symmetric taps $h_j = h_{100-j}$ and transfer $H(z) = \sum_{j=0}^{100} h_j z^{-j}$. Every independent coefficient is printed in the coefficient appendix; the remaining coefficients are fixed by symmetry. The design is obtained by an equiripple exchange calculation followed by a stated gain factor 0.999. This is a specific feasible design; minimum tap count is not asserted. Symmetry proves

$$H(e^{i\omega}) = e^{-50i\omega}A(\cos\omega), \quad A(x) = h_{50} + 2\sum_{k=1}^{50} h_{50+k}T_k(x).$$

Thus the passband group delay is exactly 50 samples, or 5 ms at 10 kHz, away from zeros and phase jumps. Finite support proves BIBO stability, since $|y[n]| \le \|h\|_1\|x\|_\infty$. The required pass magnitude lower bound is $10^{-0.1/20} < 0.99$; the stop magnitude upper bound is $10^{-50/20} > 0.003$. The delivered exact rational certificate establishes $0.99 \le A(x) \le 1.001$ on every pass interval and $|A(x)| \le 0.003$ on every stop interval. Therefore it meets the stated 0.1 dB maximum loss and 50 dB minimum rejection throughout the bands. This is a continuous-band result for the exact stored binary64 taps, not an inference from the plotted grid. To verify the certificate, write $x = a+(b-a)t$ on each certified subinterval, expand A in the Bernstein basis $\sum_{j=0}^{50} b_j \binom{50}{j} t^j(1-t)^{50-j}$, and check all b_j against the claimed bounds. For $0 \le t \le 1$, the basis functions are nonnegative and sum to one by the binomial theorem, proving the range bound. The checker computes every coefficient in rational arithmetic and subdivides by the exact de Casteljau identity. Endpoints corresponding to cosines of multiples of $\pi/10$ are enclosed by nested radical bounds obtained from integer square roots, so the certified intervals cover the true bands. The printed certificate summary and full JSON provide the subintervals and extreme coefficients. This proves all inequalities needed for the delivered design.

Structural or analytic formulation

Conjugate symmetry converts a complex frequency response into a real degree-50 polynomial. Exact convex-hull bounds then replace an arbitrarily dense frequency search with a finite replayable certificate. The original complex response remains available for the phase plot.

Comparative efficiency and scope

A direct 101-tap implementation uses 101 multiplications per sample. Pairing symmetric taps uses 50 pre-additions, 51 multiplications and 50 accumulation additions, a reduction of 50 multiplications for the same exact transfer. The certificate has finitely many accepted intervals, recorded in the metric table; it verifies infinitely many frequencies. This feasibility proof does not certify globally minimal FIR degree.

Complete FIR coefficients and band certificate

The table supplies $h_0, \ldots, h_{50}$; exact symmetry $h_{100-j} = h_j$ supplies every remaining tap. Seventeen significant decimal digits identify the delivered binary64 values on round-trip conversion. The certificate treats those binary values as exact dyadic rationals.

j	h_j
0	-5.3617638453317448e-05
1	-7.5117966638612217e-06
2	8.5641113205097371e-05
3	5.5746201565015741e-05
4	-0.00012270304912228528
5	-0.00013866916687642675
6	0.00014988416039945517
7	0.00028259508418279698
8	-0.00011840117499181697
9	-0.00046855292309119285
10	1.3155241404571084e-06
11	0.00068573422046004661
12	0.00025692126769600481
13	-0.00087024859204962027

j	h_j
14	-0.00067568219770246013
15	0.00095344778777590631
16	0.0012695500503778036
17	-0.00082280462105008766
18	-0.0019894661813683085
19	0.00037642904669524208
20	0.0027480387607900134
21	0.00049089789864490883
22	-0.0033791622332611845
23	-0.0018261486972158127
24	0.0036748592097979307
25	0.003617046087431772
26	-0.003375475507929005
27	-0.0057374548120722449
28	0.0022270231573141655
29	0.0079532058703577768
30	1.1038124925088001e-06
31	-0.0098937976201426171
32	-0.0034494410126740655
33	0.011081122923954185
34	0.0081469457793073435
35	-0.010928013351772984
36	-0.013963662898605312
37	0.0087688540632374166
38	0.020617796956720821
39	-0.00382036447909256
40	-0.027674096435850861
41	-0.0049400064323156315
42	0.034596927895137505
43	0.019272932613425209
44	-0.040795333325586089
45	-0.043445238559746358
46	0.045707522880301221
47	0.093280193593834618
48	-0.04886137590209097
49	-0.31364387232226693
50	0.54944977397990313

Band	Hz interval	Exact subintervals	Lower bound	Upper bound
Stop	0–2000	3	-0.000352558	0.000336285
Pass	2500–5000	3	0.997399886	1.000113530

The tabulated decimal bounds are rounded outward from the exact rational Bernstein extrema. The complete rational subintervals and coefficient extrema are in `results/fir_rational_certificates.json`. The response plot is a sampled illustration; the certificate covers the whole declared bands.

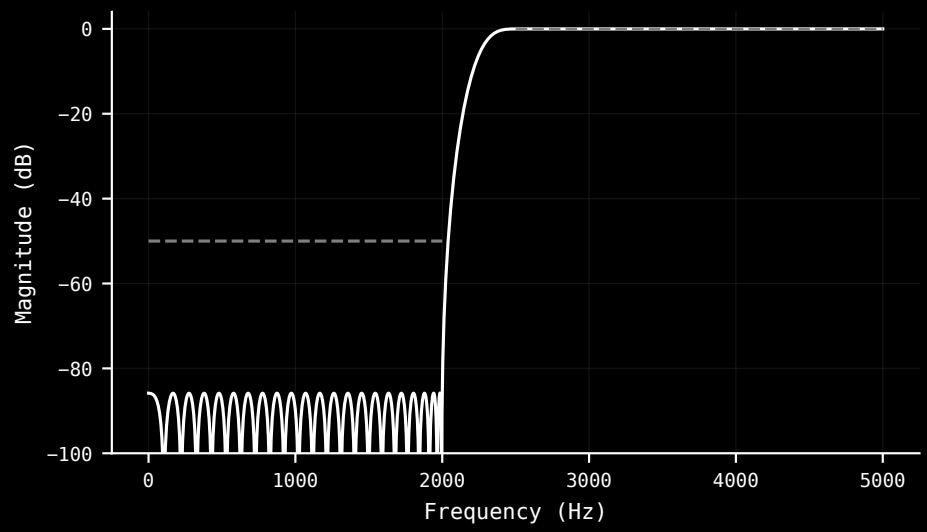


Figure 1.29: SC-053: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

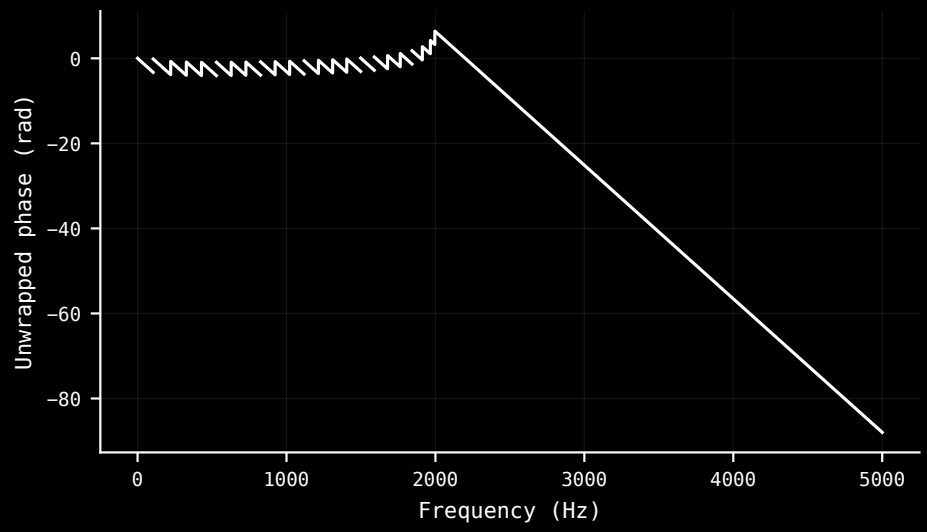


Figure 1.30: SC-053: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.54 SC-054 | A continuously certified 101-tap FIR bandpass

Reference: [1], exercise 6.5.12.

Source task and conventions

Give an FIR meeting all stated digital passband/stopband constraints.

Conventional solution and complete derivation

Use the symmetric 101-tap representation and the equiripple construction specified in SC-053. The present bandpass coefficients and its three band intervals differ from that highpass instance. The independent coefficient array printed below determines this transfer function completely, and the band-specific Bernstein certificate establishes its stated inequalities.

Structural or analytic formulation

Conjugate symmetry converts a complex frequency response into a real degree-50 polynomial. Exact convex-hull bounds then replace an arbitrarily dense frequency search with a finite replayable certificate. The original complex response remains available for the phase plot.

Comparative efficiency and scope

The exact 101-to-51 multiplication reduction is proved for the symmetric representation in SC-053; the bandpass uses the same tap count and therefore the same operation count. Its certificate uses its own three frequency intervals and their recorded subdivisions. Feasibility of these coefficients does not establish minimum bandpass order.

Complete FIR coefficients and band certificate

The table supplies $h_0, \ldots, h_{50}$; exact symmetry $h_{100-j} = h_j$ supplies every remaining tap. Seventeen significant decimal digits identify the delivered binary64 values on round-trip conversion. The certificate treats those binary values as exact dyadic rationals.

j	h_j
0	9.5010988210112373e-05
1	-3.8505119033563352e-05
2	4.4608787095339599e-06
3	-0.00014327787239157866
4	-8.3945618697839566e-05
5	0.00028372972171881169
6	1.0061868369111158e-06
7	0.00036633193918332944
8	-0.00066440390106210334
9	-0.00018249761331659667
10	-8.3863074803256322e-06
11	0.00034143686082428574
12	0.0013183139787594932
13	-0.0011510434436992538
14	1.0094007010548949e-05
15	-0.0019322944652802054
16	0.001051497619791515
17	0.0020417612002685591
18	1.5514528616280659e-05
19	0.0009957366173739767
20	-0.005525893513633787
21	0.0014538880363467541
22	-2.5900959488749729e-05
23	0.0046957531340546671
24	0.0026747807309914725
25	-0.0072317982500589104
26	-4.5027489122667981e-06
27	-0.0073392608005695666
28	0.011798208383887487
29	0.0033796749806706745
30	4.6484187251982519e-05
31	-0.0045375856682687002
32	-0.017957373177822571
33	0.01431906932481047
34	-2.180801497759776e-05
35	0.021920163800316056
36	-0.010927317504096029
37	-0.021967066203598468
38	-4.6044556351568447e-05
39	-0.010505722251923878
40	0.05537311062607006

j	h_j
41	-0.013994460978068491
42	5.6712570314671948e-05
43	-0.048745168113470294
44	-0.030871400281890529
45	0.086850418647988634
46	2.0357946919502091e-05
47	0.11968942224709686
48	-0.25604286264384779
49	-0.13854059322153275
50	0.49942857318978345

Band	Hz interval	Exact subintervals	Lower bound	Upper bound
Stop	0–1500	2	-0.000809709	0.001019648
Pass	2000–4000	2	0.996751329	1.000991379
Stop	4500–5000	1	-0.000090186	0.000089188

The tabulated decimal bounds are rounded outward from the exact rational Bernstein extrema. The complete rational subintervals and coefficient extrema are in `results/fir_rational_certificates.json`. The response plot is a sampled illustration; the certificate covers the whole declared bands.

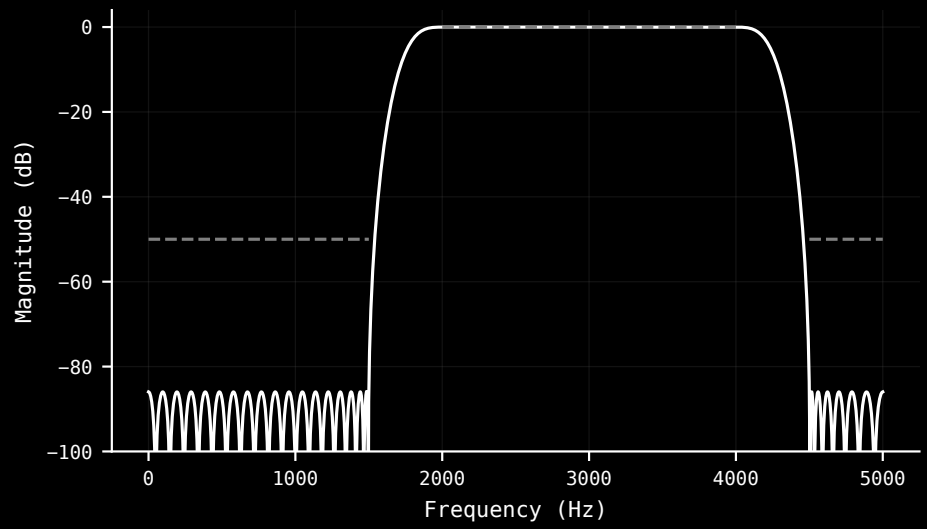


Figure 1.31: SC-054: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

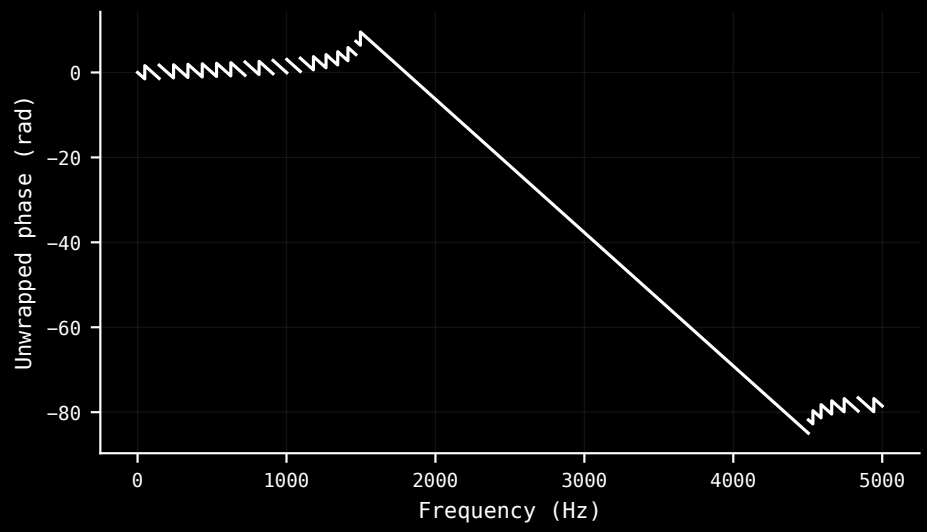


Figure 1.32: SC-054: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

1.55 SC-055 | Binary fractions in three signed representations

Reference: [1], exercise 7.7.1.

Source task and conventions

Represent 0.65625, −0.65625 and −0.05 and state the word-length limitation.

Conventional solution and complete derivation

Use one sign bit and b fractional bits, with the binary point immediately after the sign. Since $0.65625 = 21/32 = 0.10101_2$, all three positive representations are 0 10101 for $b = 5$. The negative representations are sign–magnitude 1 10101, one's complement 1 01010, and two's complement 1 01011. Complementing the whole positive six-bit word gives the one's-complement word; adding one unit in the least position gives the two's-complement word. Decoding gives exactly $-21/32$ in each case. Additional precision extends positive and sign–magnitude fractional words by zeros, one's-complement negatives by ones, and two's-complement values by the appropriate sign-preserving scaled integer encoding. The number $0.05 = 1/20$ has binary expansion $0.00\overline{0011}_2 = 0.00001100110011\ldots$. To verify it, sum the repeated block as $3 \cdot 2^{-6} \sum_{j \geq 0} 2^{-4j} = 1/20$. It has no finite binary expansion: a finite b-bit fraction has denominator dividing 2^b, whereas the reduced denominator 20 contains a factor five. Hence no unique exact finite word is determined for the third item. For a declared b, set $M = \lfloor 2^b/20 \rfloor$. Magnitude truncation gives $-M2^{-b}$, encoded as 1 followed by the b-bit representation of M for sign–magnitude; the one's-complement encoding is the bitwise complement of $0M$; its two's-complement encoding is $2^{b+1} - M$ (zero separately). For example at $b = 8$, $M = 12$, so the three words are 1 00001100, 1 11110011, 1 11110100, each representing $-3/64 = -0.046875$. Their error relative to -0.05 is $1/320$. If arithmetic truncation in two's complement means dropping low bits of an already negative word, it rounds toward minus infinity instead; that distinct convention is treated in 7.7.6.

Structural or analytic formulation

The denominator criterion proves the representation obstruction exactly; a longer decimal display cannot make $1/20$ a finite dyadic fraction. Explicit integer encodings make the sign and rounding

convention machine-checkable.

Comparative efficiency and scope

Converting to b fractional bits requires $O(b)$ output operations. A nearest rounding rule reduces worst-case absolute error from less than 2^{-b} to at most 2^{-b-1}, at the cost of inspecting guard/sticky information; it does not provide an exact finite representation of $1/20$.

1.56 SC-056 | Subtraction by signed binary addition

Reference: [1], exercise 7.7.2.

Source task and conventions

Evaluate $7/8 - 1/4$, $1/5 - 3/5$, and $3/5 - 1/10 - 1/8$ using signed addition.

Conventional solution and complete derivation

The exact rational answers are $5/8$, $-2/5$, and $3/8$. For the first, in one-sign/three-fraction-bit two's complement,

$$0\,111 + 1\,110 = 1\,0\,101 \longmapsto 0\,101 = 5/8,$$

where the leftmost carry is discarded. The negative quarter is encoded as the complement of $0\,010$ plus one. There is no signed overflow because the operands have opposite signs and the result is representable. The fractions with denominators five or ten have infinite repeating binary expansions. Thus their exact operations are operations on those infinite expansions, or equivalently on their exact rational values; a finite machine operation necessarily specifies precision. Here is a complete finite illustration with eight fractional bits and magnitude-truncated inputs. For the second, $1/5$ gives integer 51 and $-3/5$ gives the nine-bit two's-complement integer $512 - 153 = 359$:

$$0\,00110011 + 1\,01100111 = 1\,10011010 = -102/256.$$

The exact target is $-2/5$, so this finite result has error $1/640$. For the third, the encoded input integers are 153, $512 - 25 = 487$, and $512 - 32 = 480$. Adding modulo 512 gives

$$(153 + 487 + 480) \bmod 512 = 96 = 0\,01100000,$$

which decodes to $96/256 = 3/8$, exactly matching the rational result. Here the two initial quantization errors cancel; this is not a general property of subtraction. In general precision $b \geq 3$, $\lfloor (3/5)2^b \rfloor - \lfloor (1/10)2^b \rfloor = 2^{b-1}$ since the two arguments differ by that integer, so the third cancellation is exact under this convention. A different rounding convention must be analysed separately.

Structural or analytic formulation

Two's-complement addition is integer addition modulo 2^{b+1} followed by signed decoding and multiplication by 2^{-b}. Proving the integer congruence verifies every displayed bit operation; the rational residual separately verifies the approximation to the intended fractions.

Comparative efficiency and scope

A fixed-width add is constant time in the word-RAM model, or $O(b)$ bit operations. Replacing subtraction by addition of an encoded negative uses the same adder. The exactness of an output does not imply that all its quantized inputs were exact.

1.57 SC-057 | Fixed-point multiplication with repeating operands

Reference: [1], exercise 7.7.3.

Source task and conventions

Show the partial products for the three signed fractional multiplications.

Conventional solution and complete derivation

For (i), take magnitudes $3/8 = 0.011_2$ and $5/8 = 0.101_2$. The multiplier has nonzero bits at positions one and three. The partial products are $0.001100_2 = 3/16$ and $0.000011_2 = 3/64$, with a zero row for position two. Their sum is $0.001111_2 = 15/64$. One negative operand gives result $-15/64$; in one-sign/six-fraction-bit two's complement its word is 1 110001 because $128 - 15 = 113$. Keeping all six fractional product bits makes this exact. For (ii), the magnitudes are $5/16 = 0.0101_2$ and $2/3 = 0.\overline{10}_2$. Nonzero multiplier bits are at positions $1, 3, 5, \ldots$, giving successive partial products 5/32, 5/128, 5/512, and generally $5/2^{2j+5}$ for $j \geq 0$. Their sum is

$$\frac{5}{32}\sum_{j=0}^{\infty} 4^{-j} = \frac{5}{24} = 0.001\overline{10}_2.$$

The two negative signs cancel. In (iii) both operands are positive and the same magnitude computation gives 5/24. Thus (ii) and (iii) have equal positive results. After m nonzero partial products the exact value is $(5/24)(1 - 4^{-m})$, and the remaining error is $(5/24)4^{-m}$. This gives every intermediate accumulation and its exact error. A finite binary multiplier cannot represent 2/3 exactly; a finite-width answer must state m or a word length. The supplied program prints any requested number of these rows and checks their sum using exact fractions.

Structural or analytic formulation

The infinite shift-add pattern is a geometric series. Its closed form proves both the exact product and a finite-tail error bound, avoiding an unqualified termination claim for the non-dyadic operand.

Comparative efficiency and scope

Naive multiplication of two b-bit integers uses $O(b^2)$ bit operations, while these particular sparse/repeating operands admit m nonzero shift-add rows and an exact constant-size rational answer. That special structure does not establish a universal constant-time multiplier.

Nonzero partial rows	Exact sum	Exact remaining tail
1	5/32	5/96
2	25/128	5/384
3	105/512	5/1536

Nonzero partial rows	Exact sum	Exact remaining tail
4	425/2048	5/6144
5	1705/8192	5/24576
6	6825/32768	5/98304
7	27305/131072	5/393216
8	109225/524288	5/1572864

1.58 SC-058 | Finite dynamic range in IEEE binary formats

Reference: [1], exercise 7.7.4.

Source task and conventions

Determine dynamic range with explicit normal/subnormal and format conventions.

Conventional solution and complete derivation

For a binary format with t explicit fraction bits, normal exponents $e_{\min}, \ldots, e_{\max}$ and an implicit leading one, the largest positive finite value is $x_{\max} = (2 - 2^{-t})2^{e_{\max}}$. The smallest positive normal is $2^{e_{\min}}$, and the smallest positive subnormal is $2^{e_{\min}-t}$. These follow by setting the significand and exponent fields to their extreme finite patterns. Thus the normal dynamic-range ratio is $(2 - 2^{-t})2^{e_{\max}-e_{\min}}$ and the subnormal-inclusive ratio is $(2 - 2^{-t})2^{e_{\max}-e_{\min}+t}$. Zero is excluded from the denominator; infinities and NaNs are excluded from finite dynamic range. For binary32, $(t, e_{\min}, e_{\max}) = (23, -126, 127)$, giving $x_{\max} = (2 - 2^{-23})2^{127}$, $x_{\min,\text{normal}} = 2^{-126}$ and $x_{\min,\text{subnormal}} = 2^{-149}$. The two ratios are $(2 - 2^{-23})2^{253}$ and $(2 - 2^{-23})2^{276}$. For binary64, the corresponding tuple is $(52, -1022, 1023)$; the ratios are $(2 - 2^{-52})2^{2045}$ and $(2 - 2^{-52})2^{2097}$. Amplitude dynamic range in decibels is $20 \log_{10}$ of the ratio; a power-ratio convention uses $10 \log_{10}$. The source says only "IEEE format", so both common formats and the general formula are required to avoid an unstated word length. Flush-to-zero hardware uses the normal lower endpoint in practice.

Structural or analytic formulation

Separating the exponent range from the significand range gives exact symbolic bounds that do not overflow while being computed. Logarithms turn the enormous ratios into finite sums of exponent and significand contributions.

Comparative efficiency and scope

The symbolic formulas use constant arithmetic in the format parameters. Directly dividing binary64 extrema overflows a binary64 result; computing the log-ratio avoids this numerical failure. It does not enlarge the representable dynamic range of the actual hardware.

1.59 SC-059 | Aligned floating-point addition and guard bits

Reference: [1], exercise 7.7.5.

Source task and conventions

Add both pairs of the source's eight-fraction-bit mantissas and binary exponents.

Conventional solution and complete derivation

Interpret $0\Delta11010101$ as positive mantissa 213/256, $0\Delta01101010$ as 106/256, and the exponent strings as positive binary integers. For (i), the exponents are both $001_2 = 1$, so add mantissas directly:

$$.11010101_2 + .01101010_2 = 1.00111111_2.$$

Normalization gives $.100111111_2\, 2^{010_2}$, with nine fractional mantissa bits. Its exact value is $319/128 = 2.4921875$. Keeping eight fraction bits by truncation gives $.10011111_2\, 2^2 = 2.484375$, error $-1/128$. Round-to-nearest, ties-to-even instead gives $.10100000_2\, 2^2 = 2.5$, error $+1/128$; the discarded ninth bit is an exact tie. For (ii), align the smaller exponent 001 to the larger 011 by shifting its mantissa right two positions. The sum is

$$(213/256 + 213/1024)2^3 = (1065/1024)2^3.$$

After normalization it is $.100001010001_2\, 2^{100_2}$, exact value $1065/128 = 8.3203125$. Eight-fraction-bit truncation gives $.10000101_2\, 2^4 = 8.3125$, error $-1/128$. Nearest rounding gives the same eight-bit result because the discarded tail is less than half a retained unit. An exponent field too short or a different signed exponent convention would require an overflow/encoding rule; the displayed three-bit nonnegative exponents represent the required values one, two, three and four.

Structural or analytic formulation

The exact fractions make exponent alignment and renormalization transparent. Guard and sticky bits determine whether truncation or rounding is justified; discarding them before the addition can change the result.

Comparative efficiency and scope

Alignment, one mantissa add and normalization are constant operations at a fixed word size. Nearest rounding halves the worst-case quantization bound relative to truncation but needs guard/sticky information. The exact examples demonstrate the tradeoff without asserting zero error.

1.60 SC-060 | Four views of a factored IIR response

Reference: [1], exercise 2.6.9.

Source task and conventions

Plot magnitude, phase, real part and imaginary part of the supplied three-section transfer.

Conventional solution and complete derivation

Define $q = e^{-i\omega}$ and retain the printed factors:

$$H = 0.0124 \frac{1+q}{1-.6386q} \frac{1+1.5975q+q^2}{1-.9621q+.5708q^2} \frac{1+1.12724q+q^2}{1-.5811q+.8547q^2}.$$

For every sampled $\omega \in [0, \pi]$, evaluate the three numerators and denominators by Horner multiplication and multiply their ratios. The four requested ordinates are $|H|$, an unwrapped argument on each nonzero branch, $\Re H$, and $\Im H$. The delivered four plots and complete MATLAB/Python routines implement exactly these definitions. At a zero, the phase is undefined; the phase graph does not assign a meaningful continuous phase through that point. The denominator has a real pole .6386 and two conjugate pairs with moduli $\sqrt{.5708}$ and $\sqrt{.8547}$; their discriminants are negative and both moduli are below one. Thus the unit circle contains no pole and the response is finite throughout. The independent identity $(\Re H)^2 + (\Im H)^2 = |H|^2$ is checked on the same grid. The extra textbook prose accidentally included after the exercise in the extracted record is not part of the problem.

Structural or analytic formulation

The section product preserves conjugate factors and avoids expanding into a potentially poorly conditioned fifth-order polynomial. The analytic pole information establishes regularity of the plotted contour; numerical phase unwrapping remains a presentation convention.

Comparative efficiency and scope

At M frequencies, three constant-degree sections require $O(M)$ arithmetic. The factored and expanded degree-five forms have the same order, but section storage and local scaling are preferable for implementation sensitivity. No measured asymptotic speedup is claimed for this fixed filter.

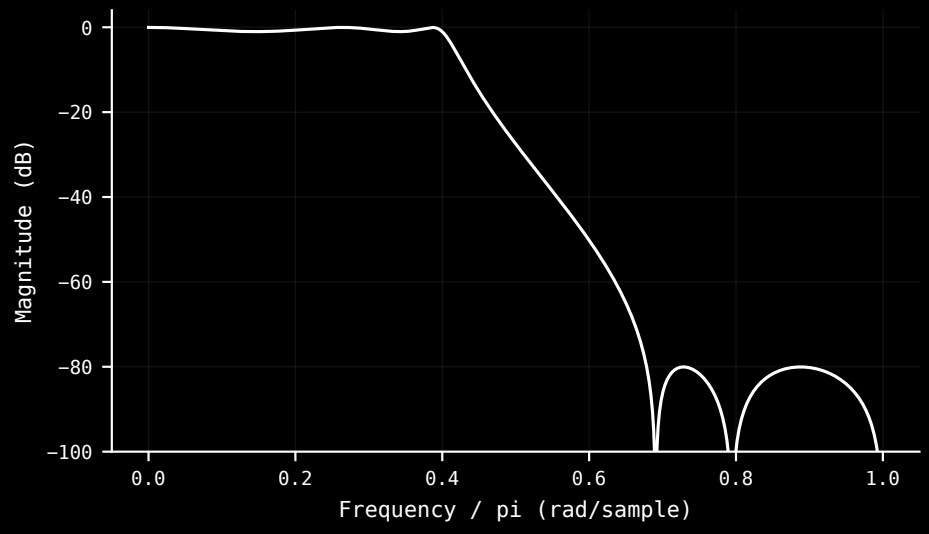


Figure 1.33: SC-060: magnitude of the specified factored transfer function or delivered tap array. Dashed constraints and grid samples do not replace the accompanying analytic or rational certificate.

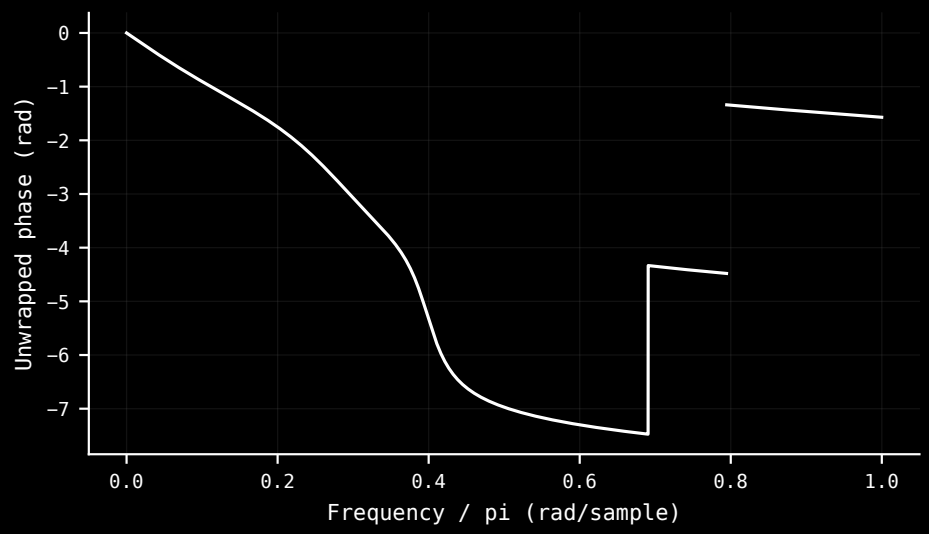


Figure 1.34: SC-060: unwrapped phase on the stored frequency grid. Samples with magnitude below 10^-7 are excluded because phase is undefined at a zero.

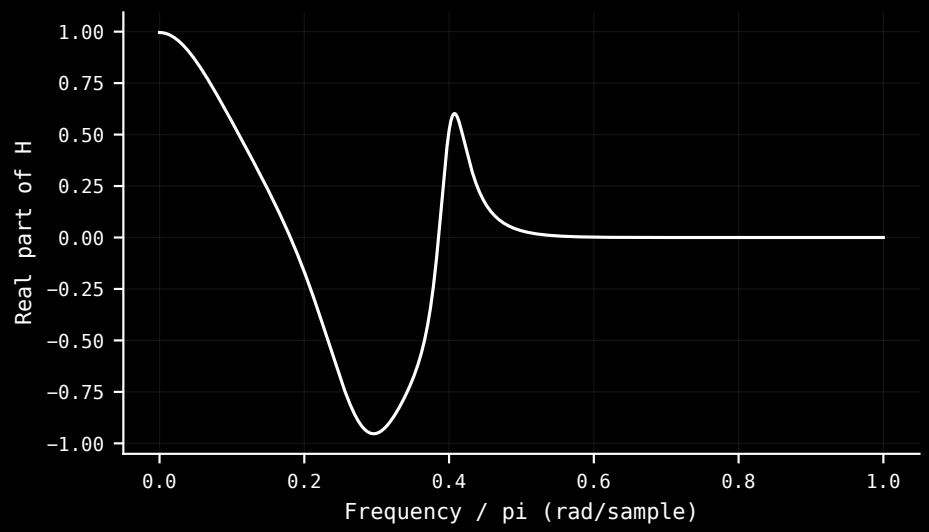


Figure 1.35: Real part computed from the three supplied sections without intermediate rounded polynomial expansion.

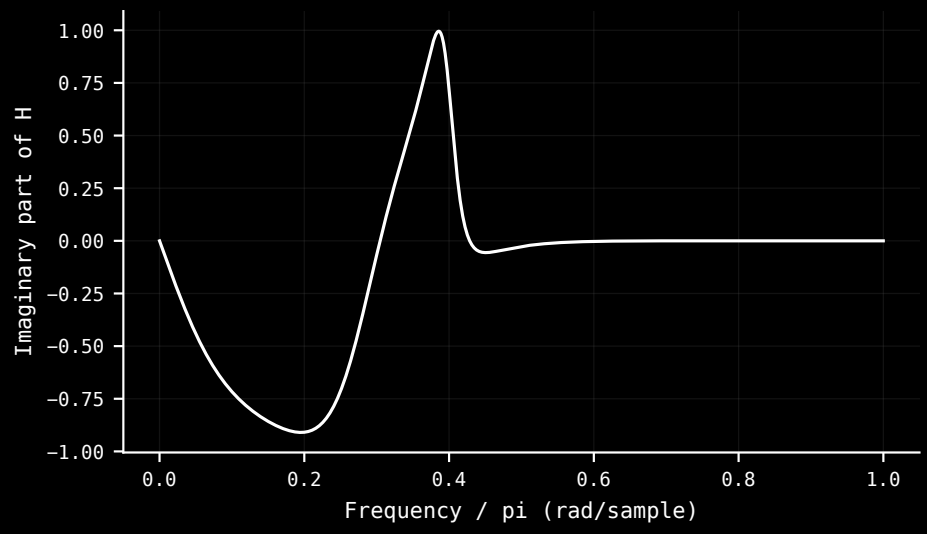


Figure 1.36: Imaginary part computed using the same section representation.

1.61 SC-061 | DFTs of conjugate and real/imaginary components

Reference: [1], exercise 3.6.1.

Source task and conventions

Express the three requested length-N DFTs in terms of $X[k]$, using the negative-exponent forward convention.

Conventional solution and proof

Let $\bar{k} = (-k) \bmod N$. Direct conjugation gives

$$\mathcal{D}\{x^*\}[k] = X^*[\bar{k}].$$

Since $\Re x = (x + x^*)/2$ and $i\Im x = (x - x^*)/2$, linearity gives

$$\boxed{\mathcal{D}\{\Re x\}[k] = \tfrac{1}{2}(X[k] + X^*[\bar{k}]), \quad \mathcal{D}\{i\Im x\}[k] = \tfrac{1}{2}(X[k] - X^*[\bar{k}]).}$$

The identity $e^{2\pi ikn/N} = e^{-2\pi i\bar{k}n/N}$ proves that the reversal is circular, including $k = 0$.

Structural or analytic derivation

These are projections induced by the conjugation-reversal involution on coefficient space. Applying either projection twice leaves it unchanged, as follows by replacing k with $\bar{k}$ and conjugating. This gives a second algebraic consistency check.

Comparative cost and preferred method

All outputs require $O(N)$ additions and conjugations once X is known, compared with $O(N \log N)$ for independent FFTs. Reading or writing N outputs supplies the matching $\Omega(N)$ output lower bound.

1.62 SC-062 | Four Z-transform operations with regions

Reference: [1], exercise 4.8.10.

Source task and conventions

Starting from $y[n] = ca^n u[n]$, $a = e^{-\alpha}$, transform $ny[n]$, $\beta^n y[n]$, $y[-n]$ and $y[n - n_0]$.

Conventional solution and proof

For $a \neq 0$ and finite c, $Y(z) = c/(1 - az^{-1})$, $|z| > |a|$. Termwise differentiation on compact subannuli gives

$$\mathcal{Z}\{ny[n]\} = -zY'(z) = \frac{caz^{-1}}{(1 - az^{-1})^2}, \quad |z| > |a|.$$

Scaling the coefficient base gives $\mathcal{Z}\{\beta^n y[n]\} = c/(1 - a\beta z^{-1})$, $|z| > |a\beta|$; if $\beta = 0$, the sequence is $c\delta[n]$. Reversal gives $Y(z^{-1}) = c/(1 - az)$ with $|z| < 1/|a|$. For integer n_0, substitution $m = n - n_0$ gives $z^{-n_0}Y(z)$, with the same finite nonzero radial boundary; zero/infinity inclusion is determined by the resulting finite shift. The identically zero case $c = 0$ has no genuine poles.

Structural or analytic derivation

The four operations are the Euler differential operator, a dilation of the argument, inversion on the Riemann sphere and a Laurent monomial. Their convergence domains follow from the transformed series, not solely from the rational denominator.

Comparative cost and preferred method

Each symbolic transformation has fixed cost once Y is known. Re-summing an infinite sequence is unnecessary. The standard table operations already realize this reduction; the analytic contribution is explicit domain tracking and removable-degeneracy handling.

1.63 SC-063 | Pole-zero analysis and a nonlinear printed recurrence

Reference: [1], exercise 4.8.11.

Source task and conventions

Analyze the rational filter in (i) and the displayed input-dependent recurrence in (ii), retaining its literal multiplication.

Conventional solution and proof

For (i), clearing negative powers gives

$$H(z) = \frac{z(z - 1)}{(z^2 - z + 1)(z + 4/5)}.$$

Zeros are $0, 1$ and poles are $e^{\pm i\pi/3}, -4/5$. For the causal interpretation the unit-circle poles do not cancel, so the impulse response is not absolutely summable and the filter is not BIBO stable. More generally no ordinary stable bilateral realization can have a noncancelled pole on its convergence circle.

Part (ii) literally reads $3y[n] = 3.7y[n - 1] - 0.7y[n - 2]x[n - 1]$. Its state transition depends on the input, so it is not an LTI input-output recurrence and has no single transfer function whose poles can be plotted. With zero initial state it produces the identically zero output for every input. With a nonzero state and bounded input $x \equiv 0$, it includes the growing mode $(37/30)^n$, so internal stability fails. An additive forcing term or an intended plus sign is not supplied by the source.

Structural or analytic derivation

For a fixed constant input $x[n] = u$, the state matrix is

$$A(u) = \begin{pmatrix} 37/30 & -7u/30 \\ 1 & 0 \end{pmatrix}.$$

Its eigenvalues depend on u. This parameterized linear state analysis is distinct from a transfer function of the original input-dependent system. The delivered plot shows the well-defined pole-zero diagram for (i); the MATLAB script reports (ii)'s modelling obstruction.

Comparative cost and preferred method

Factoring the quadratic gives an exact unit-circle certificate with fixed arithmetic. A numerical root plot is illustrative, not the stability proof. Detecting the nonlinear product prevents a formally incorrect Z-transform reduction. There is no equivalent LTI baseline to accelerate for part (ii).

Source qualification

The product $y(n-2)x(n-1)$ was checked visually on physical source page 128.

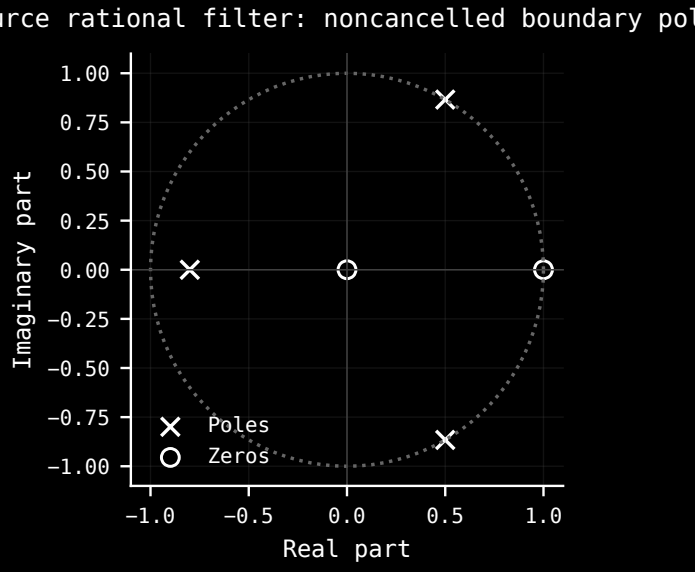


Figure 1.37: Exact pole-zero diagram for part (i). The literal nonlinear recurrence in part (ii) has no fixed transfer function.

1.64 SC-064 | Direct and cascade realizations of a cubic FIR section

Reference: [1], exercise 4.8.12.

Source task and conventions

Realize $H(z) = (1 - \alpha z^{-1} - \beta z^{-2})^3$ directly and as three second-order sections.

Conventional solution and proof

Put $q = z^{-1}$. Polynomial multiplication gives

$$H = 1 - 3\alpha q + (3\alpha^2 - 3\beta)q^2 + (6\alpha\beta - \alpha^3)q^3 + (3\beta^2 - 3\alpha^2\beta)q^4 - 3\alpha\beta^2 q^5 - \beta^3 q^6.$$

The direct realization is a six-delay tapped line with these seven coefficients, $y[n] = \sum_{k=0}^{6} b_k x[n-k]$. The cascade is $v_0 = x$ and, for $j = 1, 2, 3$,

$$v_j[n] = v_{j-1}[n] - \alpha v_{j-1}[n-1] - \beta v_{j-1}[n-2], \qquad y = v_3.$$

Each stage has two delays. Taking Z transforms with zero states gives $V_j = (1 - \alpha q - \beta q^2)V_{j-1}$; multiplication proves exact realization equality.

Structural or analytic derivation

The cascade retains the repeated polynomial factor and avoids forming powers of coefficients at setup. Both realizations have order at most six; special parameter cancellations can lower the effective order. The diagram gives each second-order signal equation and its delay states.

Comparative cost and preferred method

Ignoring unity multipliers, the cascade uses six coefficient multiplications and six additions per sample; the generic expanded FIR has six nonunity coefficients and six additions. Both use six delays. There is no generic arithmetic reduction, although coefficient storage in the identical-section cascade can be two shared parameters and numerical scaling can be arranged per stage.

1.65 SC-065 | Canonical direct and cascade IIR realizations

Reference: [1], exercise 4.8.13.

Source task and conventions

Give direct-form II, first/second-order cascade, and two-biquad cascade realizations of the stated third-order transfer function.

Conventional solution and proof

With $q = z^{-1}$ the expanded numerator and denominator are

$$B(q) = 1 + \tfrac{4}{5}q + \tfrac{17}{100}q^2 + \tfrac{1}{100}q^3, \quad A(q) = 1 + \tfrac{17}{50}q + \tfrac{53}{250}q^2 + \tfrac{3}{50}q^3.$$

A direct-form II realization is

$$w[n] = x[n] - \tfrac{17}{50}w[n-1] - \tfrac{53}{250}w[n-2] - \tfrac{3}{50}w[n-3],$$

$$y[n] = w[n] + \tfrac{4}{5}w[n-1] + \tfrac{17}{100}w[n-2] + \tfrac{1}{100}w[n-3].$$

It uses three shared delays. A first/second-order cascade uses $H_1 = (1 + q/2)/(1 + 3q/10)$ and $H_2 = (1+3q/10+q^2/50)/(1+q/25+q^2/5)$. For any section $(b_0+b_1q+b_2q^2)/(1+a_1q+a_2q^2)$, use $w = x-a_1w_{-1}-a_2w_{-2}$ and $y = b_0w + b_1w_{-1} + b_2w_{-2}$; the first-order case sets $a_2 = b_2 = 0$ and removes its unused delay. Two second-order slots can implement these same sections with a zero second-order coefficient in the first slot. Their unused delay must be omitted to retain canonical three-state storage.

Structural or analytic derivation

To check the degree, factor $1 + 0.3q + 0.02q^2 = (1 + 0.1q)(1 + 0.2q)$, which has no factor in common with the denominator; thus the displayed three-state order is genuine for these factors. The biquad state equations prove each realization by multiplication, and zero padding a coefficient array does not increase the McMillan degree. Reordering sections preserves the exact transfer but changes intermediate signal magnitudes.

Comparative cost and preferred method

Generic direct-form II uses three feedback and three nonunity feedforward multiplications, six additions, and three delays. The first/second-order cascade has the same generic counts. Two allocated full biquads would waste one state and may waste zero multiplications; removing those yields the canonical realization. No universal floating-point superiority is inferred from algebraic equivalence.

1.66 SC-066 | Parallel first- and second-order realization

Reference: [1], exercise 4.8.14.

Source task and conventions

Implement $H(z) = \gamma_0 + 0.3/(1 + 0.5z^{-1}) + (1 + 0.4z^{-1} + 0.33z^{-2})/(1 - 0.62z^{-1} + 0.43z^{-2})$.

Conventional solution and proof

Feed the same input to a direct branch, one first-order branch and one second-order branch:

$$v[n] = x[n] - 0.5v[n-1], \qquad y_1[n] = 0.3v[n],$$

$$w[n] = x[n] + 0.62w[n-1] - 0.43w[n-2], \quad y_2[n] = w[n] + 0.4w[n-1] + 0.33w[n-2],$$

$$\boxed{y[n] = \gamma_0 x[n] + y_1[n] + y_2[n].}$$

With zero initial states, the first pair gives $Y_1/X = 0.3/(1 + 0.5q)$, the second gives the stated biquad, and summation gives H. There are three dynamic states; γ_0 remains the source's free direct-feedthrough parameter.

Structural or analytic derivation

This is the decomposition of the rational operator into pole blocks. The first pole is -0.5 and the conjugate pair has modulus $\sqrt{0.43} < 1$, so each causal branch is stable and their finite sum is stable for finite γ_0. The result follows from the quadratic root product and negative discriminant, with no long simulation.

Comparative cost and preferred method

Parallel branches can be evaluated independently; total arithmetic remains constant per sample and order three. The source already supplies partial fractions, so recombining into a cubic and refactoring would add setup with no benefit. Parallelism is a scheduling option, not an asserted reduction in total arithmetic.

1.67 SC-067 | A 40-to-16-bit negative fixed-point truncation

Reference: [1], exercise 7.7.6.

Source task and conventions

Compute the maximum error for sign–magnitude, one's complement and two's complement.

Conventional solution and complete derivation

Fix the binary point after the sign bit: the original word has 39 fractional bits and the retained word 15. Let $\Delta = 2^{-15}$, $\delta = 2^{-39}$ and define error $e = Q(x) - x$. The discarded 24-bit tail is a nonnegative integer times δ, from zero through $\Delta - \delta$. For a negative sign–magnitude value, deleting low magnitude bits moves the result toward zero, so $0 \le e \le \Delta - \delta$. The same holds for one's complement: complementing the retained high bits gives the retained high bits of the positive magnitude, so low-bit deletion again truncates that magnitude. For two's complement, arithmetic deletion of low bits is signed floor division by 2^{24}; the represented value rounds toward minus infinity. Hence $-(\Delta - \delta) \le e \le 0$. The common maximum absolute error is therefore

$$\boxed{2^{-15} - 2^{-39}}.$$

It is attained when all discarded positive-magnitude bits are one in the first two conventions, or the signed two's-complement integer has the appropriate maximal nonzero remainder. If the word length excludes the sign bit, replace 39 and 15 by the actual fractional-bit counts; if the binary point is elsewhere, multiply by the corresponding power of two. Word length alone does not determine an absolute error without a binary-point convention.

Structural or analytic formulation

Integer quotient and remainder give an exact error interval for every possible discarded bit pattern. This is a uniform certificate over 2^{24} tails, with the direction of bias depending on representation.

Comparative efficiency and scope

Checking all discarded patterns would require 2^{24} cases. The remainder bound is constant-size algebra and proves the maximum for all of them. Round-to-nearest would reduce the maximum

absolute error to at most $\Delta/2$, given a representable result and the required guard information.

1.68 SC-068 | Mantissa truncation from 32 to 24 bits

Reference: [1], exercise 7.7.7.

Source task and conventions

Give absolute and relative error bounds for all signed conventions and exponent scalings.

Conventional solution and complete derivation

Write the represented value as $x = m2^E$. First suppose the specified mantissa lengths include one sign bit, so the original has 31 fractional bits and the result 23. With fixed exponent, low-bit deletion has maximum absolute value error

$$\boxed{|e|_{\max} = (2^{-23} - 2^{-31})2^E}.$$

For negative sign–magnitude and one's-complement mantissas, $e = Q(x) - x$ is nonnegative; for arithmetic two's-complement deletion it is nonpositive. The proof is the same quotient/remainder bound as in the preceding problem, now with eight discarded bits. All three maximum magnitudes are equal. For positive values truncation is downward. If "32 to 24" means fractional bits excluding the sign, the corresponding bound is $(2^{-24} - 2^{-32})2^E$. If it means significand precision including an implicit leading one in a $[1, 2)$ normalization, the 31-to-23 fractional-bit formula applies but the exponent convention is that normalization's exponent. For a normalized magnitude $|m| \geq 1/2$, the relative error in the first convention is at most $2(2^{-23} - 2^{-31})$; for $|m| \geq 1$ it is at most $2^{-23} - 2^{-31}$. Subnormal values do not have this uniform normalized relative bound. Without an exponent bound no finite global absolute maximum exists over all possible real scalings; a specific floating-point format supplies that bound.

Structural or analytic formulation

The error is an interval scaled by 2^E. Explicit exponent and normalization conventions prevent comparing mantissa ulps with physical-value error.

Comparative efficiency and scope

One remainder inequality proves the bound over all 256 discarded tails. Rounding uses additional bits and reduces the half-ulp bound, while retaining the same final storage. This is a proved precision comparison, not a wall-clock benchmark.

1.69 SC-069 | Pole locations under four-bit coefficient truncation

Reference: [1], exercise 7.7.8.

Source task and conventions

Plot exact-decimal and truncated-coefficient responses, compute all poles and explain the effect of truncation.

Conventional solution and complete derivation

The source transfer is

$$H(z) = \frac{.44 + .36z^{-1} + .02z^{-2}}{1 + .4z^{-1} + .18z^{-2} - .2z^{-3}}.$$

Treat the printed decimals as exact rationals. With four *fractional* bits and magnitude truncation toward zero, keep the leading denominator coefficient exactly one and obtain

$$H_q(z) = \frac{7/16 + (5/16)z^{-1}}{1 + (6/16)z^{-1} + (2/16)z^{-2} - (3/16)z^{-3}}.$$

The original denominator factors as $(1 - .4z^{-1})(1 + .8z^{-1} + .5z^{-2})$, so its poles are .4 and $-.4 \pm i\sqrt{.34}$; their moduli are $.4, 1/\sqrt{2}, 1/\sqrt{2}$. The quantized poles are the three roots of $z^3 + (3/8)z^2 + (1/8)z - 3/16$; all their computed coordinates are printed in the coefficient table, alongside the exact original values. Stability can be proved without trusting numerical roots: on $|z| = 1$, the sum of absolute nonleading coefficients is $.78 < 1$ originally and $11/16 < 1$ after quantization. Rouché's theorem gives exactly three roots inside the disk in each case. The original boundary margin is at least $1 - .78 = .22$; every coefficient perturbation with total absolute size below .22 retains this root count. The original DC gain is $.82/1.38 = 41/69$, while the quantized gain is $(12/16)/(21/16) = 4/7$. The numerator's small second-delay term disappears, and both poles and zeros move; the plotted magnitude differences are therefore not a simple constant-gain change. If "four bits" means one sign and three fractional bits instead, magnitude truncation gives numerator $(3/8, 1/4, 0)$ and denominator $(1, 3/8, 1/8, -1/8)$; this is a different filter. Its nonleading sum $5/8 < 1$ also proves stability. Both interpretations are recorded, with the primary plots explicitly labelled four fractional bits.

Structural or analytic formulation

A boundary polynomial inequality certifies an entire coefficient region. Quantized numerical poles illustrate the change but are not the sole stability evidence. The phase and amplitude still need evaluation after truncation; stable does not mean accurate.

Comparative efficiency and scope

A root solver has nontrivial per-instance cost and conditioning near multiple roots. The sufficient Rouché test here is three absolute values and two additions, proving all roots lie in the disk. It is not a necessary general stability test; failure of this bound would not prove instability.

Original and truncated realization

Index	b original	b truncated	a original	a truncated
0	0.44	0.4375	1	1
1	0.35999999999999999	0.3125	0.40000000000000002	0.375
2	0.02	0	0.17999999999999999	0.125
3	0	0	-0.20000000000000001	-0.1875

Index	Re original pole	Im original pole	Re truncated pole	Im truncated pole
1	-0.4	0.583095189485	-0.394772760438	0.544478674351
2	-0.4	-0.583095189485	-0.394772760438	-0.544478674351

Index	Re original pole	Im original pole	Re truncated pole	Im truncated pole
3	0.4	0	0.414545520876	0

The displayed pole coordinates are numerical approximations. The factorization and the strict boundary inequalities above provide the stability proof independently of these coordinates. Leading denominator coefficient one is fixed by the realization convention.

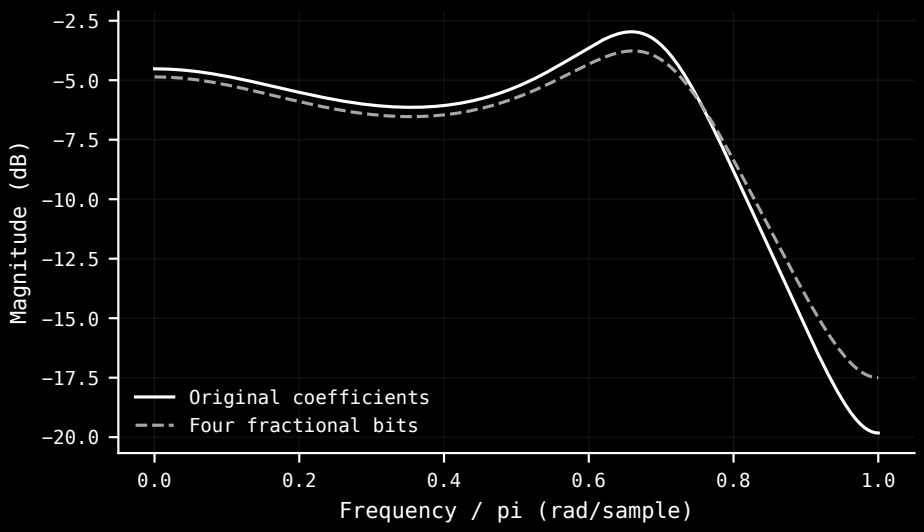


Figure 1.38: Original decimal transfer function and the explicitly specified four-fractional-bit truncation. The pole certificate in the solution covers a full perturbation region.

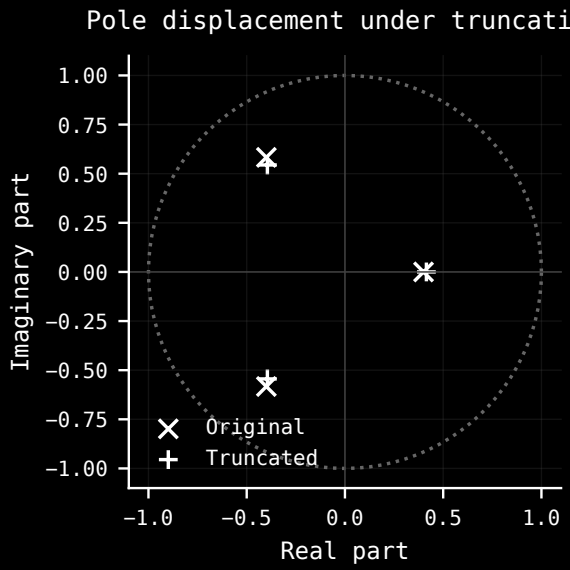


Figure 1.39: Original and quantized poles remain strictly inside the unit circle; exact stability is proved separately.

1.70 SC-070 | CCS introductory laboratory: a reproducible acceptance specification

Reference: [1], exercise 8.9.1.

Source task and conventions

Complete the written tutorial workflow for loading, running, modifying and configuring a program in the source CCS/DSP-BIOS environment.

Conventional solution and complete derivation

This exercise assigns a tutorial activity, not a uniquely determined calculation. A complete deliverable has a recorded CCS version, board/processor model, emulator, compiler, linker command file and DSP/BIOS version; the source's legacy interfaces must not be silently replaced by a contemporary incompatible tool. Use the shipped target-matched example and first build without modification. Preserve the build log and map. Connect the debugger, load the executable, run to its documented stop point, and inspect a known output value. Add a volatile scalar or a ten-element integer vector, modify the loop to double each entry, rebuild/reload, and verify every output against its specified integer value. In the configuration tool, create the required startup/task object, generate the configuration files and confirm that the generated objects appear in the map. Test a deliberate breakpoint and a single step, then restore the baseline. The supplied host reference computes the same deterministic vector and verifies it, providing an independent output oracle. Acceptance requires the original and modified binaries, source diff, successful load/run trace, captured outputs, configuration file and generated-map evidence. Merely supplying C source does not prove that it was loaded on a DSP. The host reference and analytical checks were executed in the recorded study; the requested proprietary target session has not been executed because that target/tool installation is unavailable. This execution boundary is intrinsic to an activity-based question and is retained in the completion ledger.

Structural or analytic formulation

Use a small invariant: after loop iteration j, exactly the first j elements equal twice their input and the remaining elements retain their initialization. Initialization at $j = 0$, preservation by the update and termination at ten prove the expected array. This independent oracle detects target/configuration mistakes without conflating them with algorithm design.

Comparative efficiency and scope

A ten-element reference needs ten multiplications or additions. Its linear invariant proves all intermediate states with one induction. There is no justified complex-analytic reduction of debugger loading, configuration or target execution time.

1.71 SC-071 | CCS project, debugger, visualization and profiling laboratory

Reference: [1], exercise 8.9.2.

Source task and conventions

Provide the complete written sequence and acceptance criteria for the advanced IDE tutorial topics.

Conventional solution and complete derivation

Create a version-controlled project containing the deterministic three-tap filter in the supplied C reference. Document target selection, include paths, compiler optimization, linker command and library versions. Build a debug configuration, inspect symbols, set breakpoints before and after the state update, and verify the first impulse outputs $(1, 2, 1, 0, \ldots)$. Use the memory/data window to inspect the delay line; graph a bounded multi-tone test input and output using a declared sampling interval. Profile an uninstrumented block of B samples, subtract empty-loop overhead where justified, and report total cycles, cycles/sample and run-to-run spread under a fixed cache/interrupt policy. Compare optimized and debug builds only after verifying identical outputs under the intended arithmetic. For GEL/configuration, use the target vendor's supplied initialization and record the exact script; do not write guessed PLL, memory-controller or reset registers. Introduce one controlled source error, diagnose it with the debugger, record its cause and restore the correct version. The deliverable includes project files, map, compiler flags, test input, output trace, graph export, profiler trace and the target initialization script. The reference program and host output checks accompany this monograph, while the source-required week of target-IDE practice is a user activity and is not falsely reported as performed here.

Structural or analytic formulation

The filter oracle follows $y[n] = x[n] + 2x[n-1] + x[n-2]$ with zero initial delay state. Its transfer $(1 + z^{-1})^2$ predicts the impulse and frequency responses. This gives an analytic cross-check for visualization; it does not verify board initialization or a debugging interface.

Comparative efficiency and scope

The reference requires two additions plus one doubling per sample and two delay values. Profiling must measure the actual target implementation; operation counts do not establish DSP cycle counts in the presence of pipelines, memory stalls or instrumentation.

1.72 SC-072 | DSP/BIOS scheduling and I/O laboratory

Reference: [1], exercise 8.9.3.

Source task and conventions

Provide a reproducible scheduling, profiling and I/O experiment with explicit real-time acceptance conditions.

Conventional solution and complete derivation

Use the target-matched DSP/BIOS tutorial project, record all versions and configure a periodic producer, a filter task and a consumer with bounded buffers. A useful declared reference schedule has periods $T_1 = 1$ ms and $T_2 = 4$ ms and measured worst-case execution requirements to be inserted as C_1, C_2. Fix priorities, interrupt masking, preemption rules and the buffer capacity.

Instrument release, start and finish timestamps and distinguish blocked time from running time. Test empty/full buffers, delayed consumer, task overrun and an interrupted transfer. Confirm sample order, no out-of-range writes and the filter's deterministic reference output. For a uniprocessor fixed-priority model with task one higher priority and no blocking, its response is $R_1 = C_1$ and task two's response is the least fixed point

$$R_2 = C_2 + \lceil R_2/T_1 \rceil C_1.$$

Starting at C_2 and iterating counts all higher-priority jobs released during the candidate window; monotonicity proves the least fixed point. If it is at most T_2 and $C_1 \leq T_1$, the two-task deadlines are certified in this declared model. Add blocking/interrupt terms before applying the result to a real system. The host scheduling oracle tests the model with $C_1 = .1$ ms,$C_2 = .3$ ms, giving $R_1 = .1$ ms and $R_2 = .4$ ms. Deliver the configuration, source, timing traces, buffer logs, I/O waveform and deadline analysis. The host model is executed; no proprietary DSP/BIOS scheduling trace or physical I/O capture is claimed. The source's extended hands-on duration cannot be certified by a generated document.

Structural or analytic formulation

A response-time fixed point is a more informative timing certificate than average CPU utilization. It handles the specified finite schedule and its explicit preemption assumptions. A complex transfer function can verify signal processing but cannot certify task deadlines.

Comparative efficiency and scope

A three-tap filter has constant work per sample; the scheduling recurrence has a finite iteration bound when its response remains in a bounded integer-time horizon. Actual worst-case C_i and blocking bounds must come from the target. The illustrative host values are not measurements of the user's board.

1.73 SC-073 | Visual Linker and memory placement

Reference: [1], exercise 8.9.4.

Source task and conventions

Give the complete written linker-recipe and memory-management procedure with a checkable placement example.

Conventional solution and complete derivation

Start from the vendor's target-matched linker example, open its recipe, and record memory regions, origins, lengths, access permissions and section alignments. Inspect placement of executable text, initialized data, uninitialized data, stack, heap and generated runtime/configuration sections. Produce a map before and after one controlled section-placement change; verify symbols, entry point and load/run addresses. In a declared host model, let a RAM region be $[0x80000000, 0x80001000)$, text require 1536 bytes aligned to 256, data 512 bytes aligned to 64, and stack 1024 bytes aligned to 16. The placement text $[0x80000000, 0x80000600)$, data $[0x80000600, 0x80000800)$, stack $[0x80000800, 0x80000C00)$ meets all alignments and leaves 1024 bytes. Sorting these half-open intervals and checking each end against the next start proves absence of overlap; comparing endpoints with the RAM bounds proves containment. A test increasing the stack to 3072 bytes

must be rejected because the final end exceeds the region. This numerical example is not the actual memory map of any unspecified DSP board. The deliverable is the recipe, linker command/map files, before/after placement diagram, alignment/overlap checks, load/run verification and a rejected over-allocation case. Device registers, reserved memory and external-memory initialization come from the actual board documentation. The host interval checker is executed; opening a proprietary Visual Linker project and programming hardware remain unexecuted target activities.

Structural or analytic formulation

Memory placement is a finite interval/alignment constraint problem. Exact integer endpoints provide a complete overlap certificate. A conformal or spectral method does not replace linker semantics or target reservations.

Comparative efficiency and scope

For s sections, sorting intervals costs $O(s \log s)$ and the overlap scan is $O(s)$. A supplied sorted map can be verified in linear time. These are exact host-model costs, not a prediction of a graphical linker's elapsed time.

Chapter 2

Fundamentals of Wireless Communication

David Tse, Pramod Viswanath. Reference problems and derived analyses.

2.1 SC-074 | The MMSE output as a sufficient statistic

Reference: [2], exercise 8.19.

Source task and conventions

Prove sufficiency in the coloured proper-Gaussian vector channel and recover its information expression.

Conventional solution and complete derivation

Use $y = hx + z$, $z \sim \mathcal{CN}(0, K_z)$ independent of x, $K_z \succ 0$, and scalar input variance $P > 0$. The conditional density has exponent

$$-y^* K_z^{-1} y + 2\Re\{x^* h^* K_z^{-1} y\} - |x|^2 h^* K_z^{-1} h.$$

Its dependence on y relevant to x is only $t = h^* K_z^{-1} y$, proving sufficiency for any input prior. Put $a = h^* K_z^{-1} h$. The linear MMSE vector is

$$v = (PK_h + K_z)^{-1} hP = \frac{P}{1 + Pa} K_z^{-1} h, \qquad K_h = hh^*.$$

Multiplying $K_z^{-1} h$ by $K_z + Phh^*$ verifies the identity directly, so $v^* y$ is a nonzero scalar multiple of t when $h \neq 0$. For Gaussian input, $t = ax + \xi$ with $\xi \sim \mathcal{CN}(0, a)$. Thus its scalar SNR is Pa, and

$$I(x; y) = I(x; t) = \log(1 + Pa), \qquad \text{mmse} = \frac{P}{1 + Pa}.$$

The information equality follows from the chain rule and $I(x; y \mid t) = 0$ supplied by sufficiency; the scalar entropy difference follows from the Gaussian density. The error formula follows by substituting v into $E|x - v^* y|^2$. If $h = 0$, both mutual informations and the output statistic are zero and the error is P.

Structural or analytic formulation

The rank-one resolvent identity separates the signal direction from coloured noise. It simultaneously proves sufficiency, the MMSE weight and the information–SNR relation. The formula $\log(1 + Pa)$ uses Gaussian input; sufficiency itself does not.

Comparative efficiency and scope

One factorization of K_z costs $O(n^3)$ in a dense model, followed by $O(n^2)$ solves per new h and $O(n)$ per observation. After this setup, the entire observation reduces to one complex scalar without information loss for the specified model. Repeatedly inverting $K_z + Phh^*$ for each P adds unnecessary cubic work.

2.2 SC-075 | Orthogonal observations of white Gaussian noise

Reference: [2], exercise A.2.

Source task and conventions

Find the joint law of the M noise coefficients.

Conventional solution and complete derivation

Interpret ideal white noise as a centred Gaussian generalized process: for real square-integrable waveforms f, g, $E[W(f)W(g)] = (N_0/2)\langle f, g\rangle$. This definition avoids treating white noise as an ordinary pointwise finite-variance function. Every linear combination $\sum_i a_i z_i = W(\sum_i a_i s_i)$ is Gaussian with variance $(N_0/2)\sum_i a_i^2$, since the waveforms are orthonormal. Its characteristic function is therefore $\exp[-N_0\|a\|^2/4]$, which factors into the characteristic functions of M independent $N(0, N_0/2)$ variables. Thus

$$z \sim N(0, (N_0/2)I_M), \qquad f_z(u) = (\pi N_0)^{-M/2} \exp(-\|u\|^2/N_0).$$

Zero covariance implies independence here because the complete vector is jointly Gaussian. Without Gaussianity orthogonality would establish only zero correlation. Complex waveforms require the sesquilinear covariance and the corresponding complex-noise convention.

Structural or analytic formulation

Projection onto an orthonormal waveform family turns the noise covariance operator into a scalar multiple of the identity. Unitary/orthogonal basis changes preserve its law because they preserve the norm in its characteristic function.

Comparative efficiency and scope

The joint density requires M squared coordinates and one exponential, using $O(M)$ arithmetic. A generic dense Gaussian covariance representation uses $O(M^2)$ storage and $O(M^3)$ factorization; orthonormality removes that setup. Computing sampled waveform projections still requires their own input-dependent cost.

2.3 SC-076 | Complex covariance and pseudo-covariance

Reference: [2], exercise A.3.

Source task and conventions

Reconstruct the full real covariance and specialize to a circularly symmetric vector.

Conventional solution and complete derivation

Write the centred vector as $x = u + iv$ and set $A = Euu^T$, $B = Evv^T$, $C = Euv^T$. The complex covariance $K = E[xx^*]$ and pseudo-covariance $J = E[xx^T]$ satisfy

$$K = A + B + i(C^T - C), \qquad J = A - B + i(C + C^T).$$

Solving these linear equations gives the full real covariance

$$\Sigma_{\mathbb{R}} = \frac{1}{2}\begin{pmatrix} \Re(K+J) & \Im(J-K) \\ \Im(J+K) & \Re(K-J) \end{pmatrix}.$$

This explicitly recovers all second-order information. If x is circularly symmetric, x and $e^{i\theta}x$ have the same distribution. Their finite mean obeys $\mu = e^{i\theta}\mu$, hence $\mu = 0$. Their pseudo-covariance obeys $J = e^{2i\theta}J$, hence $J = 0$. Consequently

$$\Sigma_{\mathbb{R}} = \frac{1}{2}\begin{pmatrix} \Re K & -\Im K \\ \Im K & \Re K \end{pmatrix}.$$

These identities do not assume Gaussianity. The implication from symmetry to vanishing pseudo-covariance has been proved, not its converse for arbitrary distributions.

Structural or analytic formulation

The pair (K, J) separates Hermitian energy from noncircular second-order structure. The displayed invertible real-linear transformation gives a direct implementation and an entry-by-entry verification rule.

Comparative efficiency and scope

For a general n-vector the conversion is $O(n^2)$ arithmetic, matching the number of matrix entries. Under circular symmetry J is known to be zero and need not be stored. This is a parameter/storage reduction, not a universal halving of every downstream computation.

2.4 SC-077 | Circular symmetry of a Gaussian vector

Reference: [2], exercise A.4.

Source task and conventions

Prove the mean/pseudo-covariance criterion and assess whether the covariance identity alone suffices.

Conventional solution and complete derivation

Necessity follows from the mean and pseudo-covariance transformation laws just proved. For sufficiency, suppose $\mu = 0, J = 0$. Under $x \mapsto e^{i\theta}x$, the mean stays zero, K stays unchanged, and J stays zero. The preceding reconstruction formula therefore gives identical real means and covariance matrices before and after rotation. A real Gaussian vector is determined by these quantities, since its characteristic function is $\exp(it^T m - t^T \Sigma t/2)$. The two vectors have the same distribution for every θ. The covariance relation in part (2) establishes $J = 0$, but says nothing about the mean. A counterexample is $x = 1 + z$ with $z \sim \mathcal{CN}(0, 1)$. Its centred real covariance is $I_2/2$ and its pseudo-covariance is zero, yet its mean one is changed by rotation. Thus the answer is no unless zero mean is added. If "covariance" were used for an uncentred second moment, that alternative convention would require separate treatment; the source defines centred covariance.

Structural or analytic formulation

Gaussian identifiability reduces equality of distributions to equality of first and second moments. For a non-Gaussian vector the same reduction fails: a variable uniform on $\{1, -1, i, -i\}$ has zero mean and pseudo-covariance but is not invariant under every angle.

Comparative efficiency and scope

Testing the Gaussian criterion uses one mean vector and one $n \times n$ pseudo-covariance, $O(n^2)$ stored quantities. Checking a few simulated angles cannot certify invariance under a continuum of rotations. The criterion supplies an exact distributional test under the Gaussian model.

2.5 SC-078 | Real and imaginary parts of a circular Gaussian scalar

Reference: [2], exercise A.5.

Source task and conventions

Prove independence and identical distributions.

Conventional solution and complete derivation

Let $x = u + iv$ be circular Gaussian with finite variance κ. Its mean is zero. Its pseudo-variance is

$$Ex^2 = E(u^2 - v^2) + 2iEuv = 0.$$

Hence $Eu^2 = Ev^2 = \kappa/2$ and $Euv = 0$. The jointly real Gaussian vector (u, v) therefore has characteristic function $\exp[-\kappa(t_1^2 + t_2^2)/4]$, a product of two equal univariate characteristic functions. Thus u, v are independent and identically distributed as $N(0, \kappa/2)$. The degenerate case $\kappa = 0$ gives two deterministic zeros and satisfies the same statement.

Structural or analytic formulation

Multiplication by i exchanges the coordinate directions up to a sign; invariance forces equal variances and zero cross-covariance. Joint Gaussianity then converts this covariance condition to independence.

Comparative efficiency and scope

Two variance values and one covariance suffice, so verification of the declared scalar model has constant algebraic cost. Independence is a theorem of the model; no finite empirical sample proves it.

2.6 SC-079 | Which unitary matrices preserve an improper Gaussian law?

Reference: [2], exercise A.6.

Source task and conventions

Give the condition for a fixed unitary matrix and the stronger condition for all unitary matrices.

Conventional solution and complete derivation

Write the common real/imaginary covariance as $K_x = \left(\begin{smallmatrix} a & c \\ c & b \end{smallmatrix}\right) \succeq 0$, with mean zero. Independence across entries gives complex covariance $(a+b)I_n$ and pseudo-covariance jI_n, where $j = a - b + 2ic$. For Ux, these become $(a+b)I_n$ and jUU^T. The Gaussian mean and real covariance reconstruction prove the exact criterion

$$Ux \overset{d}{=} x \quad \Longleftrightarrow \quad j(UU^T - I_n) = 0.$$

Thus either $a = b, c = 0$ (the proper/circular case), or the specified matrix satisfies $UU^T = I_n$. A matrix that is both unitary and complex orthogonal has $U^* = U^T$, hence has real entries and is real orthogonal. Consequently every real orthogonal U preserves any such i.i.d. Gaussian law. Invariance for every unitary U requires $j = 0$, as can be seen by choosing $U = iI_n$. The result includes degenerate covariance matrices.

Structural or analytic formulation

The pseudo-covariance transforms by congruence $J \mapsto UJU^T$, whereas covariance transforms by $K \mapsto UKU^*$. Confusing these two transformations incorrectly declares every improper Gaussian unitary-invariant.

Comparative efficiency and scope

For the distributional condition under all unitary transformations, checking $a = b, c = 0$ has constant cost. For a supplied dense U, checking $UU^T = I$ takes $O(n^3)$ classical arithmetic unless its structure is already certified. A numerical residual supplies a tolerance check, not exact equality.

2.7 SC-080 | Sufficiency of a complex matched-filter statistic

Reference: [2], exercise A.7.

Source task and conventions

Characterize when h^*y retains all binary detection information under possibly improper Gaussian noise.

Conventional solution and complete derivation

Write $h = u + iv$, $r = (\Re y, \Im y)^T$, $s = (u, v)^T$, and

$$\Omega = \begin{pmatrix} aI & cI \\ cI & bI \end{pmatrix}, \quad M = \begin{pmatrix} u^T & v^T \\ -v^T & u^T \end{pmatrix}.$$

Then $r = xs + \eta$ and $Mr = (\Re h^*y, \Im h^*y)^T$. First assume $K_x \succ 0$. Expanding the two Gaussian quadratic forms gives log-likelihood ratio $2s^T\Omega^{-1}r$. The statistic Mr is sufficient exactly when this linear functional is constant on every fibre of M, namely

$$\Omega^{-1}s \in \operatorname{range} M^T.$$

Sufficiency follows by factoring the likelihood through Mr. Necessity follows because a vector in $\ker M$ with nonzero score changes the likelihood while leaving the statistic fixed; full Gaussian support allows all such perturbations. If u, v are linearly independent, writing $K_x^{-1} = \left(\begin{smallmatrix} a' & c' \\ c' & b' \end{smallmatrix}\right)$ and comparing coefficients in $\Omega^{-1}s = \alpha s + \beta(-v, u)^T$ gives $a' = b'$ and $c' = 0$. Thus the necessary and sufficient condition is $K_x = \sigma^2 I_2$. If u, v are linearly dependent, h is a common complex phase times a real vector; $\Omega^{-1}s$ always belongs to that two-dimensional real span, so every positive-definite K_x works. For $h = 0$ every statistic is sufficient because the hypotheses have identical distributions. In particular the scalar-output case $n = 1, h \neq 0$ is always sufficient, since multiplication by h^* is invertible. For completeness, if Ω is singular and $s \in \operatorname{range}\Omega$, the criterion becomes $s \in \operatorname{range}(\Omega M^T)$, obtained by restricting the score $\Omega^\dagger s$ to the common Gaussian support. If $s \notin \operatorname{range}\Omega$, the two supports are disjoint and full observation detects perfectly; sufficiency then holds exactly when $Ms \notin \operatorname{range}(M\Omega M^T)$, so the projected supports also remain disjoint. These support criteria cover noiseless degeneracies.

Structural or analytic formulation

Whitening identifies the sufficient direction $\Omega^{-1}s$. Ordinary matched filtering is optimal only when its two real output coordinates span that direction. Circularity is sufficient for every channel; a fixed channel can admit weaker conditions.

Comparative efficiency and scope

For independent identically distributed noise pairs, applying K_x^{-1} costs $O(n)$ and the final score is one dot product. A generic $2n \times 2n$ whitening factorization would cost $O(n^3)$. The channel-specific criterion prevents attributing an information-preserving reduction to an invalid noise assumption.

2.8 SC-081 | Maximum discrete entropy and uniqueness

Reference: [2], exercise B.1.

Source task and conventions

Maximize entropy over all probability vectors on K symbols.

Conventional solution and complete derivation

Let $u_i = 1/K$ and use $0 \log 0 = 0$. The elementary inequality $\log t \le t - 1$ for $t > 0$ follows because $t - 1 - \log t$ has minimum zero at one. Applying its equivalent relative-entropy form gives $D(p\|u) = \sum_{i:p_i>0} p_i \log(p_i/u_i) \ge 0$, with equality only when $p = u$; one proof applies $-\log(u_i/p_i) \ge 1 - u_i/p_i$ and sums, with any missing positive u_i giving strict inequality. Since

$$H(p) = \log K - D(p\|u),$$

the maximum is $\log K$, attained uniquely at the uniform distribution. For $K = 1$ the sole distribution has entropy zero.

Structural or analytic formulation

The uniform law is the entropy maximizer because the entropy deficit is exactly its Kullback–Leibler divergence. This identity proves the bound and the equality condition together.

Comparative efficiency and scope

Computing the entropy of a supplied K-entry vector requires $O(K)$ terms. Determining the maximizing symbolic distribution requires no numerical constrained optimizer. The direct analytic answer uses K identical output entries, which is output-optimal if all probabilities are written.

2.9 SC-082 | Jensen inequality and conditional entropy

Reference: [2], exercise B.2.

Source task and conventions

Give a geometric definition, prove Jensen and characterize entropy equality including a conditional counterexample.

Conventional solution and complete derivation

A function on a convex domain is concave when $f((1-t)a + tb) \ge (1-t)f(a) + tf(b)$ for every a, b and $0 \le t \le 1$; its graph lies above each chord. Iterating this inequality proves Jensen for every finitely supported distribution by induction on its support size. For an integrable real variable with mean m in the interior of the interval domain, concavity provides a supporting slope c with $f(x) \le f(m) + c(x - m)$. To see existence, every left chord slope at m is at least every right chord slope, so choose c between the one-sided slopes. Taking expectations proves $Ef(X) \le f(EX)$ whenever the expectations are defined. Boundary means force X to equal that boundary almost surely for a one-sided domain, or follow by limits under the usual integrability hypotheses. For discrete variables with finite entropies,

$$H(X) - H(X \mid Y) = \sum_y P(y) D(P_{X|y}\|P_X) \ge 0.$$

The divergence equality condition proves equality iff the conditional laws equal P_X for every positive-probability y, exactly independence. As a counterexample to a pointwise conditional bound, let $P(Y = 1) = 1/10$, take $X = 0$ when $Y = 0$ and make X a fair bit when $Y = 1$. Then $P(X = 1) = 1/20$, so $H(X) = h_2(1/20) < \log 2 = H(X \mid Y = 1)$, while $H(X \mid Y) = (1/10) \log 2 \le H(X)$. Conditioning reduces the average entropy, not necessarily each conditional entropy.

Structural or analytic formulation

The entropy difference is an average of nonnegative divergences. This establishes all equality conditions without solving a family of Lagrange equations. Infinite entropies require care because subtracting $\infty - \infty$ is undefined; the mutual-information divergence remains meaningful.

Comparative efficiency and scope

For an explicit finite joint table of $r \times s$ entries, the entropy comparison costs $O(rs)$ arithmetic and reads the input once. The geometric proof covers all distributions satisfying its integrability hypotheses; finite plots cannot substitute for that universal statement.

2.10 SC-083 | When mutual information is additive

Reference: [2], exercise B.3.

Source task and conventions

Characterize $I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y)$.

Conventional solution and complete derivation

Assume the mutual informations involved are finite. Two applications of the chain rule give

$$I(X_1, X_2; Y) - I(X_1; Y) - I(X_2; Y) = I(X_1; X_2 \mid Y) - I(X_1; X_2).$$

For example this follows by replacing each information with entropy differences and cancelling $H(X_1)$, $H(X_2)$ and their conditional counterparts. Therefore the necessary and sufficient condition is equality between unconditional and conditional dependence: $I(X_1; X_2 \mid Y) = I(X_1; X_2)$. A useful sufficient condition is independence of X_1, X_2 both unconditionally and conditionally on Y. Unconditional independence alone is not sufficient: for independent fair bits and $Y = X_1 \oplus X_2$, both individual informations vanish but the pair reveals one bit. Conditional independence alone is also insufficient: let $X_1 = X_2 = Y$ be a fair bit; the left side is one bit and the right side is two.

Structural or analytic formulation

This identity isolates the interaction-information term. It distinguishes complementary information from redundant information and prevents unjustified summation of channel rates.

Comparative efficiency and scope

The equality is an exact scalar criterion derived from the same joint law. Computing it from a dense finite table still costs at least the table size; no sublinear algorithm is claimed. The two binary counterexamples each require only four or fewer nonzero atoms.

2.11 SC-084 | Maximum differential entropy at fixed second moment

Reference: [2], exercise B.4.

Source task and conventions

Prove that the centred Gaussian uniquely maximizes entropy with $EX^2 = P$.

Conventional solution and complete derivation

Take $P > 0$ and $g(x) = (2\pi P)^{-1/2} \exp[-x^2/(2P)]$. For any admissible density f with the required second moment and well-defined entropy,

$$D(f\|g) = -h(f) + \tfrac{1}{2}\log(2\pi P) + \frac{EX^2}{2P} = \tfrac{1}{2}\log(2\pi eP) - h(f).$$

The integral inequality $D(f\|g) \geq 0$ follows from $t\log t - t + 1 \geq 0$ integrated against g, with equality iff $f = g$ almost everywhere. Consequently $h(f) \leq \frac{1}{2}\log(2\pi eP)$, uniquely attained by $N(0, P)$. This treats the source's uncentred second-moment constraint: a nonzero mean spends part of P and cannot attain the maximum. If $P = 0$ the distribution is a point mass and does not have the stipulated everywhere-positive Lebesgue density.

Structural or analytic formulation

The Lagrangian stationary equation $-1 - \log f + \lambda_0 + \lambda_2 x^2 = 0$ suggests an exponential quadratic. Normalization and the second moment select $\lambda_2 = -1/(2P)$. The divergence calculation supplies global optimality and uniqueness beyond this stationary calculation.

Comparative efficiency and scope

The infinite-dimensional variational problem is solved by one scalar parameter P. A discretized density optimizer is unnecessary for this model and would require discretization error control. This is a proved entropy optimum, not a runtime benchmark.

2.12 SC-085 | Maximum entropy on a nonnegative half-line

Reference: [2], exercise B.5.

Source task and conventions

Maximize differential entropy for a nonnegative variable of fixed mean m.

Conventional solution and complete derivation

For $m > 0$, let $g(x) = m^{-1}e^{-x/m}$ for $x \geq 0$. For every density on the same half-line with $EX = m$,

$$D(f\|g) = -h(f) + \log m + EX/m = 1 + \log m - h(f) \geq 0.$$

Thus the exponential density uniquely maximizes entropy with value $1 + \log m$. Equality requires $f = g$ almost everywhere. The assumption $m = 0$ would force a point mass at zero, outside the given positive-density class.

Structural or analytic formulation

The exponential family is determined by the affine constraint in x. The nonnegative divergence is the global certificate; a stationary Euler–Lagrange equation alone would not prove maximality without this concavity or divergence argument.

Comparative efficiency and scope

One scalar mean specifies the maximizing density. The comparison with a general discretized density search is a representation reduction under exactly the same support and mean constraint. Changing the support or imposing a second moment produces a different optimization problem.

2.13 SC-086 | Entropy maximization with several moment constraints

Reference: [2], exercise B.6.

Source task and conventions

Prove global optimality of the proposed exponential density when its multipliers exist.

Conventional solution and complete derivation

Suppose $g(u) = \exp(\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(u))$ on S is normalized and has the stated moments A_i, and assume the required integrals are finite. For any other feasible density f,

$$D(f\|g) = -h(f) - \lambda_0 + 1 - \sum_i \lambda_i A_i.$$

The same expression at $f = g$ has divergence zero. Hence $h(g) - h(f) = D(f\|g) \geq 0$, with equality exactly when the densities agree almost everywhere. This proves the assertion without an existence claim for multipliers in every possible moment problem. Some constraint sets are empty; some have no entropy maximizer or attain one on a smaller support. The conditional statement in the exercise explicitly supplies multipliers producing a feasible density.

Structural or analytic formulation

The log-density belongs to the linear span of the constraint functions. Every feasible perturbation has the same expected log-density, so all entropy loss is relative entropy. This gives a common proof for the Gaussian and exponential cases without repeating separate variational optimizations.

Comparative efficiency and scope

After solving for the $m + 1$ multipliers, evaluating the density requires m feature evaluations and one exponential. Multiplier existence and numerical solution costs depend on the functions r_i; no universal complexity bound is asserted.

2.14 SC-087 | Real-vector and complex-scalar maximum entropy

Reference: [2], exercise B.7.

Source task and conventions

Prove both covariance-constrained entropy bounds.

Conventional solution and complete derivation

For part (1), assume $K \succ 0$ and the source's actual constraint $E[xx^T] = K$. The centred Gaussian density has logarithm $-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det K - \frac{1}{2}x^T K^{-1}x$. Since $E[x^T K^{-1}x] = \operatorname{tr}(K^{-1}K) = n$, the same divergence calculation gives

$$h(x) \le \tfrac{1}{2}\log((2\pi e)^n \det K),$$

with equality only for $N(0, K)$. For a fixed centred covariance with a prescribed nonzero mean the translated Gaussian instead maximizes entropy; the displayed uncentred constraint selects mean zero. For part (2), write $x = u + iv$ and $S = E[(u, v)^T(u, v)]$. Its nonnegative eigenvalues s_1, s_2 obey $s_1 + s_2 \le P$. The real-vector bound and $(s_1 - s_2)^2 \ge 0$ give $\det S = s_1 s_2 \le (P/2)^2$. Therefore $h(x) \le \log(\pi e P)$, attained by independent $u, v \sim N(0, P/2)$, i.e. $\mathcal{CN}(0, P)$. Equality forces zero mean, a Gaussian law, full use of the power budget and equal component variances. Singular covariance has no full-dimensional density; it is treated on its lower-dimensional support and cannot attain the positive-P full-dimensional bound.

Structural or analytic formulation

The two-dimensional determinant is maximized by equal eigenvalues at fixed trace. Thus circular symmetry is a consequence of entropy optimization, not an extra assumption imposed on competitors.

Comparative efficiency and scope

For a fixed covariance, the exact entropy bound is one log-determinant, costing $O(n^3)$ for a dense Cholesky factorization. Its optimizer is explicit. In the scalar-complex case the trace–determinant inequality is constant algebraic work.

2.15 SC-088 | Maximum entropy with fixed complex covariance

Reference: [2], exercise B.8.

Source task and conventions

Prove $h(x) \le \log\det(\pi e K)$ and its equality condition.

Conventional solution and complete derivation

For $K \succ 0$, the centred proper complex Gaussian has density $g(z) = (\pi^n \det K)^{-1}\exp(-z^* K^{-1} z)$ on $\mathbb{R}^{2n}$. This density is normalized by diagonalizing K unitarily and integrating n scalar complex Gaussians. For a zero-mean complex vector with $E[xx^*] = K$,

$$E[x^* K^{-1} x] = \operatorname{tr}(K^{-1}K) = n.$$

Thus $D(f\|g) = \log(\pi^n \det K) + n - h(f)$, and nonnegativity proves the bound. Equality holds iff $f = g$ almost everywhere; in particular an improper Gaussian cannot also attain equality. If K is singular, restrict to its complex range; the full $2n$-dimensional entropy formula is not a finite Lebesgue-density statement. On rank-r support the analogue uses $\log((\pi e)^r \operatorname{pdet} K)$.

Structural or analytic formulation

The proof does not need a separate optimization over pseudo-covariances. Relative entropy to the proper Gaussian simultaneously controls every competitor with the same Hermitian covariance.

Comparative efficiency and scope

One Cholesky factorization provides the exact covariance-dependent bound and Gaussian normalization. This is a scope-limited global optimum over all admissible densities, with no claim of an algorithm for arbitrary entropy estimation from samples.

2.16 SC-089 | Optimal MISO beamforming under total power

Reference: [2], exercise B.9.

Source task and conventions

Prove the covariance and input law maximizing the MISO mutual information.

Conventional solution and complete derivation

Use $y = h^* x + w$ with $w \sim \mathcal{CN}(0, N_0)$ independent, and $\operatorname{tr} Q \le P$ for $Q = E[xx^*]$. The entropy bound yields

$$I(x; y) = h(y) - h(w) \le \log(1 + h^* Q h / N_0).$$

For $h \neq 0$, choose an orthonormal basis whose first vector is $h/\|h\|$. In that basis $h^* Q h = \|h\|^2 Q_{11} \le \|h\|^2 \operatorname{tr} Q \le P\|h\|^2$, since $Q \succeq 0$. Equality is attained by

$$Q_* = P \frac{hh^*}{\|h\|^2}, \quad x = \sqrt{P} \frac{h}{\|h\|} s, \quad s \sim \mathcal{CN}(0, 1),$$

so capacity is $\log(1 + P\|h\|^2/N_0)$. If the source writes the row channel as h^T, conjugate the beam vector accordingly. For $h = 0$ the capacity is zero and every feasible input ties. The conclusion assumes known fixed channel, independent Gaussian noise and a total-power constraint; per-antenna or constant-envelope constraints alter the feasible set.

Structural or analytic formulation

Cauchy–Schwarz in the channel direction turns a matrix-valued covariance optimization into a scalar power bound. The rank-one construction proves attainability.

Comparative efficiency and scope

Forming the beam costs $O(L)$ arithmetic and stores L complex values, compared with storing an arbitrary $L \times L$ covariance. Against isotropic allocation, received SNR improves by factor L for any nonzero fixed h, or total power falls by factor L for a fixed SNR. This is a proved physical-model optimum, not a universal RF implementation guarantee.

2.17 SC-090 | Mutual information under an isotropic MISO input

Reference: [2], exercise B.11.

Source task and conventions

Compute the achieved rate and compare it with optimal coherent transmission.

Conventional solution and complete derivation

For $x \sim \mathcal{CN}(0, (P/L)I_L)$, h^*x is proper complex Gaussian with variance $(P/L)\|h\|^2$. Therefore y has variance $N_0 + (P/L)\|h\|^2$, and direct Gaussian entropy subtraction gives

$$I(x;y) = \log\left(1 + \frac{P\|h\|^2}{LN_0}\right).$$

The result in bits uses $\log_2$. The optimum with transmitter channel knowledge replaces P/L by P, as proved in B.9. Both rates vanish for $h = 0$ or $P = 0$.

Structural or analytic formulation

Only the one-dimensional image of the isotropic covariance through h^* affects the observation. The other transmit directions carry no information to this single receiver.

Comparative efficiency and scope

Both rate formulas require an $O(L)$ norm calculation. The improvement of coherent beamforming is an SNR/power factor L, not an L-fold rate improvement: the exact rate ratio is the ratio of the two logarithms.

2.18 SC-091 | Joint time–frequency waterfilling

Reference: [2], exercise 5.30.

Source task and conventions

Optimize within a coherence block, across finitely many blocks, and in a stationary ergodic limit.

Conventional solution and complete derivation

Use a unitary length-T_c DFT. The eigenvalues of circular convolution are the unnormalized transform $\widetilde{h}_n = \sum_{\ell=0}^{L_i-1} h_\ell[i] e^{-2\pi i n\ell/T_c}$, equivalently $\sqrt{T_c}$ times the unitary DFT of the zero-padded tap vector. Under the source's underspread/edge approximation, orthogonal transformed noises remain independent $\mathcal{CN}(0, N_0)$ and the rate for block i is

$$\frac{1}{T_c} \max_{p_{in}\geq 0,\, \sum_n p_{in} \leq T_c P[i]} \sum_n \log(1 + p_{in} g_{in}/N_0), \quad g_{in} = |\widetilde{h}_n[i]|^2.$$

The derivative is $g/(N_0 + pg)$. A common multiplier on the sum constraint and concavity prove the unique active allocations $p_{in} = (\nu_i - N_0/g_{in})_+$, with zero for zero gain and ν_i chosen to meet the total power. This proves (1), including the normalization. For (2), the intermediate block powers can be eliminated: the combined constraint is $\sum_{i,n} p_{in} \le MT_cP$. The same KKT argument yields one common water level ν_M across all MT_c time–frequency cells, not separately optimized levels with arbitrary preassigned $P[i]$. Then $P[i] = T_c^{-1} \sum_n p_{in}$. Sorting the noise-to-gain ratios and checking active sets gives the exact finite allocation. For (3), assume fixed T_c, a stationary ergodic sequence of block tap vectors, a nonzero gain with positive probability, and sufficient log-integrability, for example $E \log(1 + g) < \infty$. The empirical power function $F_M(\nu) = (MT_c)^{-1} \sum_{i,n} (\nu - N_0/g_{in})_+$ converges at every fixed ν by the ergodic theorem to $F(\nu) = E[T_c^{-1} \sum_n (\nu - N_0/g_n)_+]$. These functions are continuous, nondecreasing and 1-Lipschitz. The positive-power equation $F(\nu) = P$ has a unique solution on its increasing range; positivity of gain on a set of positive probability ensures eventual increase without bound. Bracketing at $\nu \pm \epsilon$ proves $\nu_M \to \nu$. Log-integrability and these bounded water levels justify the corresponding convergence of empirical rates, giving ergodic capacity $E[T_c^{-1} \sum_n \log(\max\{1, \nu g_n/N_0\})]$ under the adopted block model. Exact finite convolution has edge terms; their neglect is the explicit operational approximation stipulated by the source, not an identity for every T_c, L_i.

Structural or analytic formulation

All allocations are a concave optimization in parallel scalar eigenchannels. The unitary transform supplies independence and power conservation; KKT sufficiency establishes global optimality.

Comparative efficiency and scope

For $D = MT_c$ gain cells, sorting and active-set scanning cost $O(D \log D)$ and $O(D)$ respectively; evaluating the final allocation costs $O(D)$. An arbitrary search over all power tuples is unnecessary. The limiting theorem is conditional on the declared ergodicity and integrability, and contains no unverified iid assumption.

2.19 SC-092 | Reciprocity of MIMO capacity

Reference: [2], exercise 8.1.

Source task and conventions

Compare a channel H with its adjoint under equal total power and noise normalization.

Conventional solution and complete derivation

Let $H = U\Sigma V^*$ be a singular-value decomposition and let $\lambda_1, \ldots, \lambda_r > 0$ be its nonzero squared singular values. Unitary preprocessing preserves the isotropic Gaussian noise and total input power. The capacity optimization is therefore

$$C(H) = \max_{p_i \ge 0, \sum p_i \le P} \sum_{i=1}^{r} \log(1 + p_i \lambda_i / N_0).$$

To justify diagonal covariance, in the singular-vector coordinates Hadamard's determinant inequality bounds the positive-definite output determinant by the product of its diagonal entries; retaining the covariance diagonal attains this bound at the same trace. Hadamard follows by

Gram–Schmidt: the squared volume of column vectors is the product of squared orthogonal residual lengths, each no greater than its original squared length. The adjoint decomposition $H^* = V\Sigma^* U^*$ has exactly the same nonzero singular values, so the two optimizations are identical. Extra nullspace dimensions contribute zero and receive no power in a positive-capacity optimum. For $H = 0$ both capacities vanish.

Structural or analytic formulation

The spectra of H^*H and HH^* coincide off zero: if $H^*Hv = \lambda v$ with $\lambda > 0$, then $Hv \neq 0$ and $HH^*(Hv) = \lambda Hv$; the reverse map uses H^*/λ. This gives an alternative proof independent of a second SVD.

Comparative efficiency and scope

One SVD and one waterfilling calculation serve both directions, avoiding two independent decompositions. The dense SVD has $O(n_r n_t \min(n_r, n_t))$ arithmetic in a classical model. Equality requires the same total-power and noise assumptions; equal per-antenna constraints need not be reciprocal.

2.20 SC-093 | Gaussian ellipsoid concentration and volume

Reference: [2], exercise 8.2.

Source task and conventions

Prove the three high-probability and volume assertions for a length-N MIMO block.

Conventional solution and complete derivation

Assume the source's capacity-achieving architecture: independent time samples $x[m] \sim \mathcal{CN}(0, K_x)$ and $w[m] \sim \mathcal{CN}(0, N_0 I_{n_r})$. Put $S = HK_xH^* + N_0 I \succ 0$ and order the block so its covariance is $S \otimes I_N$ (a different ordering is a coordinate permutation). Whitening gives $z^N \sim \mathcal{CN}(0, I_{n_r N})$. Each squared complex coordinate is exponential with mean and variance one, so the quadratic form $Q = (y^N)^*(S \otimes I_N)^{-1} y^N$ has mean $n_r N$ and variance $n_r N$. For every $\varepsilon > 0$,

$$P(Q > N(n_r + \varepsilon)) \leq \frac{n_r}{N\varepsilon^2} \to 0,$$

proving (1). The mean-radius ellipsoid $E(0)$ itself need not have probability tending to one; the positive slack in the assertion is essential. For (2), map the complex ball of radius $\sqrt{n_r N}$ by $(S^{1/2} \otimes I_N)$. The real Jacobian determinant of a complex linear map A is $|\det_{\mathbb{C}} A|^2$, as follows first for a complex scalar and then from unitary diagonalization/singular values. Thus this map scales volume by $\det(S)^N$, yielding the required ellipsoid volume. For (3), $\|w^N\|^2/N_0$ has the same mean and variance as Q, giving the identical Chebyshev bound. The mean-radius noise ball has radius $\sqrt{N_0 n_r N}$. Its volume is $N_0^{n_r N}$ times the unit-normalized reference ball. Consequently the received/noise volume ratio is

$$\left[\det\left(I + HK_xH^*/N_0\right)\right]^N.$$

The logarithm divided by N is the Gaussian mutual information. Volumes at positive slack have the same limiting logarithmic rate as the slack is subsequently taken to zero. This geometric count explains the formula; the coding theorem giving operational achievability remains the channel-theoretic prerequisite.

Structural or analytic formulation

Whitening reduces the high-dimensional ellipsoid to a gamma radius and isolates all channel dependence in one determinant. The moment proof provides an explicit concentration bound instead of relying on a drawing of the ellipsoid.

Comparative efficiency and scope

Whitening setup can use a single $n_r \times n_r$ factorization, reused across N time samples, instead of factorizing the $n_r N$ block matrix. Costs become $O(n_r^3 + N n_r^2)$ instead of a generic $O((n_r N)^3)$ factorization. This exploits the exact Kronecker covariance.

2.21 SC-094 | Uncorrelated fading does not guarantee diagonal optimal covariance

Reference: [2], exercise 8.3.

Source task and conventions

Validate the unitary change of coordinates and assess the proposed diagonal restriction.

Conventional solution and complete derivation

Part (1) is valid: substitute $H = U_r H_a U_t^*$ into $\log\det(I + HK_xH^*/N_0)$, then conjugate by U_r^*; the determinant is unchanged. The variable $Q = U_t^* K_x U_t$ remains positive semidefinite with the same trace. Part (2), as stated using only jointly uncorrelated zero-mean entries, is false. Consider one receive antenna, two transmit antennas, $P = N_0 = 1$. Let the angular row H_a equal $\pm(1,1)$, each with probability $1/3$, and $\pm(2,-1)$, each with probability $1/6$. The mean is zero and $E[\overline{h}_1 h_2] = (2/3)\cdot 1 + (1/3)\cdot(-2) = 0$. For diagonal $Q = \mathrm{diag}(t, 1-t)$, $0 \le t \le 1$,

$$F(Q) = \tfrac{2}{3}\log 2 + \tfrac{1}{3}\log(2+3t),$$

which is maximized at $t = 1$. Now take

$$Q_\delta = \begin{pmatrix} 1-\delta & \sqrt{\delta(1-\delta)} \\ \sqrt{\delta(1-\delta)} & \delta \end{pmatrix}, \quad 0 < \delta < 1.$$

This covariance is positive semidefinite (rank one) with trace one. Its objective is

$$\tfrac{2}{3}\log(2+2c) + \tfrac{1}{3}\log(5-3\delta-4c), \quad c = \sqrt{\delta(1-\delta)}.$$

Relative to the best diagonal covariance, its leading change as $\delta \downarrow 0$ is $(2/3 - 4/15)\sqrt{\delta} + O(\delta) = (2/5)\sqrt{\delta} + O(\delta) > 0$. The limit of the change divided by $\sqrt{\delta}$ proves strict improvement for sufficiently small δ; $\delta = 1/100$ is also checked numerically in the supplied calculations. Thus a diagonal covariance need not be optimal. The proposed intermediate matrix $I + H_a K_{\mathrm{off}} H_a^*/N_0$ need not be positive definite, so its log-determinant need not even be defined. Moreover $E\det X = \det EX$ is false in general: for $X = \mathrm{diag}(Z, Z)$ with $Z = 1, 2$ equally likely, the two sides are $5/2$ and $9/4$. Uncorrelated entries do not justify the printed determinant manipulation. A correct sufficient hypothesis is invariance of the entire law under independent column sign changes (in particular, independent proper Gaussian entries suffice). For every diagonal sign matrix D, invariance gives $F(DQD^*) = F(Q)$. Concavity of $\log\det$ implies

$$F(\operatorname{diag} Q) = F(E_D[DQD^*]) \geq E_D F(DQD^*) = F(Q).$$

Averaging all 2^{n_t} signs eliminates each off-diagonal entry and keeps the trace. This proves the corrected diagonal-restriction theorem under the stated stronger distributional hypothesis.

Structural or analytic formulation

The required object is the symmetry of the full fading law, not merely its covariance. Group averaging gives the valid optimization proof when that symmetry is present; the finite four-atom distribution proves the original weaker assertion cannot be completed.

Comparative efficiency and scope

Under the corrected symmetry condition, a covariance search with n_t^2 real degrees of freedom reduces to n_t nonnegative diagonal powers. Under the original assumptions that reduction can decrease the attainable objective, so no efficiency claim is valid. The counterexample is an exact obstruction, not an unfinished proof.

2.22 SC-095 | Orthogonal access versus simultaneous two-user transmission

Reference: [2], exercise 10.4.

Source task and conventions

Optimize the orthogonal fraction and prove the exact and high-SNR gaps.

Conventional solution and complete derivation

Put $a = P_1\|h_1\|^2/N_0$, $b = P_2\|h_2\|^2/N_0$. The orthogonal sum rate is

$$R(\alpha) = \alpha \log(1 + a/\alpha) + (1-\alpha)\log(1 + b/(1-\alpha)).$$

Concavity of log bounds this by $\log(1+a+b)$; for $a, b > 0$, equality requires $a/\alpha = b/(1-\alpha)$, so $\alpha_* = a/(a+b)$. If one is zero, allocate all useful resources to the other; if both are zero every fraction ties. The determinant lemma, or expansion on the span of h_1, h_2, gives simultaneous sum capacity

$$C_\Sigma = \log\left(1 + a + b + \frac{P_1 P_2}{N_0^2} D\right), \quad D = \|h_1\|^2\|h_2\|^2 - |h_1^* h_2|^2 \geq 0.$$

Cauchy–Schwarz proves the inequality, with equality iff the vectors are linearly dependent. Consequently, for positive powers, the exact gap is

$$\Delta = \log\left(1 + \frac{P_1 P_2 D/N_0^2}{1 + a + b}\right),$$

zero precisely for linearly dependent channel vectors (including a zero vector). The wording $h_1 = ch_2$ needs this symmetric interpretation if $h_2 = 0, h_1 \neq 0$. With $P_1 = P_2 = P$, $\rho = P/N_0$ and $D > 0$,

$$\Delta = \log\rho + \log\frac{D}{\|h_1\|^2 + \|h_2\|^2} + o(1).$$

The gap diverges with slope one nat per unit $\log\rho$, or one bit per unit $\log_2\rho$.

Structural or analytic formulation

The Gram determinant D is the squared area of the two receive directions. It is the single invariant deciding whether a second spatial degree of freedom exists.

Comparative efficiency and scope

A one-variable concave maximization is solved exactly by the energy fraction α_*. The two large receive-matrix determinants reduce to norms and one inner product, costing $O(n_r)$ instead of generic $O(n_r^3)$ determinant arithmetic. The physical rate improvement is the exact logarithmic gap, not a computation-speed factor.

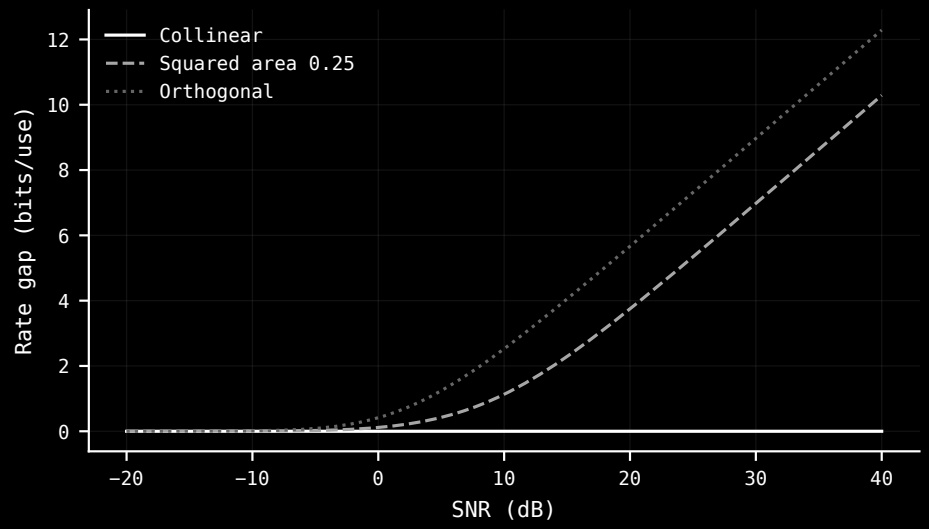


Figure 2.1: Unit-norm signatures: the exact Gram-area invariant determines the gain of simultaneous reception.

2.23 SC-096 | User count, sum capacity and symmetric capacity

Reference: [2], exercise 10.5.

Source task and conventions

Prove the four limiting assertions and calculate the requested Rayleigh-channel curves.

Conventional solution and complete derivation

Let $W_K = \sum_{k=1}^{K} h_k h_k^*$, where $h_k \sim \mathcal{CN}(0, I_r)$ independently, and define $F_K(\rho) = E \log\det(I_r + \rho W_K)$ in nats. Adding a user and using the determinant lemma gives

$$F_{K+1} - F_K = E \log\{1 + \rho h_{K+1}^* (I + \rho W_K)^{-1} h_{K+1}\} > 0 \quad (\rho > 0).$$

This proves (1). For (2), concavity of log-determinant and $EW_K = KI_r$ give $F_K \le r \log(1 + \rho K)$, hence $F_K/K \to 0$. The nonnegative lower bound completes the squeeze argument. For (3), $0 \le \rho\lambda - \log(1 + \rho\lambda) \le \rho^2\lambda^2/2$ for $\lambda \ge 0$, by integrating $t/(1 + t) \le t$. Thus

$$0 \le \rho K r - F_K \le \tfrac{1}{2}\rho^2 K r (K + r),$$

because $E \operatorname{tr} W_K = Kr$ and $E \operatorname{tr} W_K^2 = Kr(K+r)$. For the latter identity, the K same-column terms have mean $E\|h\|^4 = r(r+1)$, and the $K(K-1)$ distinct-column terms have mean $E|h_i^* h_j|^2 = r$. Hence, for fixed K, r, symmetric capacity is $\rho r + O(\rho^2 r(K+r))$. This low-SNR expansion is not uniform in unbounded K. For (4), at $K \geq r$ the Wishart matrix has full rank almost surely, and

$$F_K(\rho) = r \log \rho + E \log \det W_K + o(1) \quad (\rho \to \infty).$$

Gram–Schmidt applied to the r independent Gaussian rows shows $\det W_K$ is the product of independent gamma variables with shapes $K, K-1, \ldots, K-r+1$. Their expected logarithms are $\psi(K-j+1)$, defined by differentiating the gamma integral. Consequently $E \log \det W_K = \sum_{j=1}^{r} \psi(K-j+1) = r \log K + O(r^2/K)$ for fixed r and large K; the final estimate follows from the harmonic-number representation $\psi(m) = H_{m-1} - \gamma$ and integral bounds on harmonic sums. Thus adding users beyond r changes the power/array gain logarithmically, while it no longer increases the spatial pre-log beyond r. For (5), the supplied script generates common Gaussian realizations for $K = 1, \ldots, 40$, $r = 3, 4, 5, 6$ and SNR $0, 5, \ldots, 30$ dB. It records both sample means and standard errors and plots $F_K / \log 2$ and $F_K/(K \log 2)$. The transition near $K = r$ is sharpest at high SNR: before it, adding a user adds a spatial mode; afterwards, it mainly raises the eigenvalue scale. At low SNR the symmetric rate remains nearly flat over the range where $\rho(K+r) \ll 1$. Thus $K = r$ is an engineering operating heuristic, not a universal optimum of an unspecified objective. The numerical table is appended with the actual execution status.

Structural or analytic formulation

Exchangeability and diminishing rank-one gains also prove that F_m/m decreases with m: inverse covariance decreases as columns are added, so the expected increments form a decreasing sequence. Therefore the full-user subset is the tight symmetric constraint in this i.i.d. model, justifying $C_{\text{sym}} = F_K/K$.

Comparative efficiency and scope

Each simulated realization can reuse prefix updates $W_{K+1} = W_K + hh^*$ for all K, instead of rebuilding every matrix independently. A Cholesky log-determinant or eigenvalue calculation remains $O(r^3)$ per requested curve point. The analytic bounds supply monotonicity and asymptotic proofs; simulation measures finite-parameter values and uncertainty.

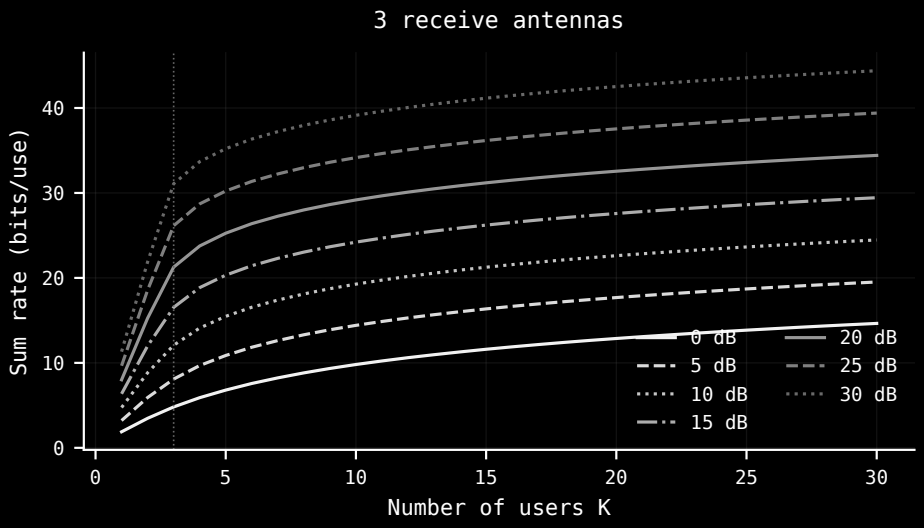


Figure 2.2: Rayleigh MAC, r=3: 2,048 shared realizations per curve; standard errors for every point are in the experiment register. The line K=r is a heuristic operating reference, not an exact threshold.

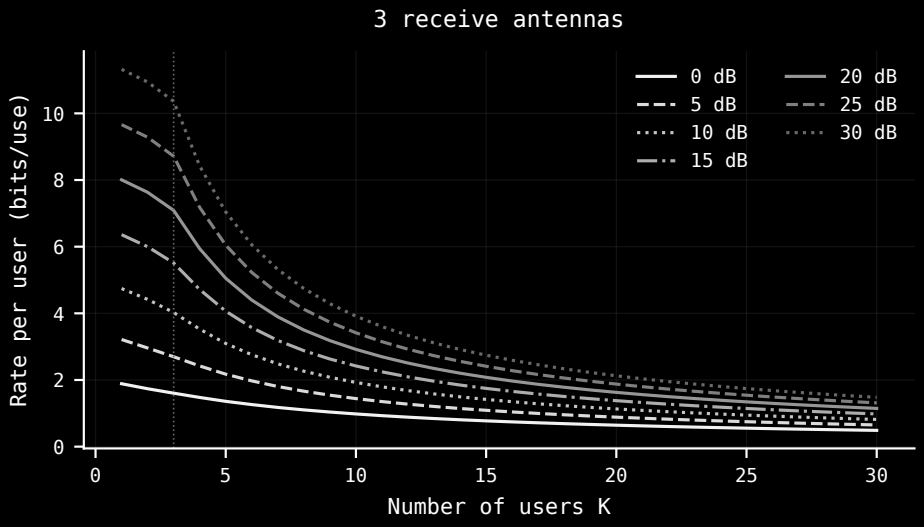


Figure 2.3: Rayleigh MAC, r=3: 2,048 shared realizations per curve; standard errors for every point are in the experiment register. The line K=r is a heuristic operating reference, not an exact threshold.

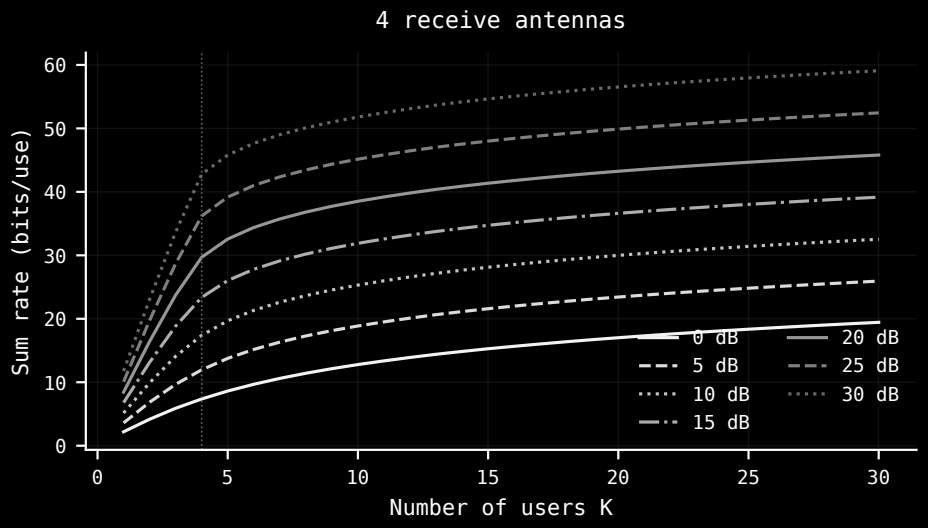


Figure 2.4: Rayleigh MAC, r=4: 2,048 shared realizations per curve; standard errors for every point are in the experiment register. The line K=r is a heuristic operating reference, not an exact threshold.

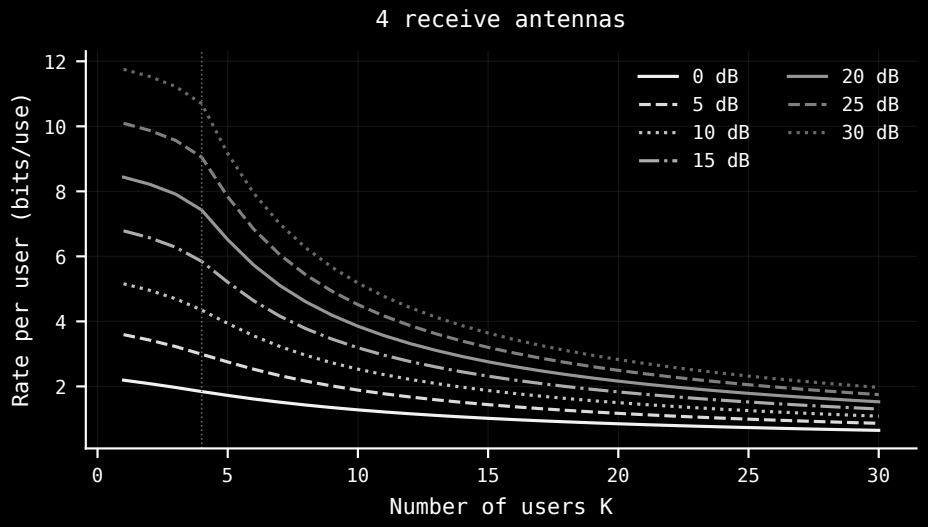


Figure 2.5: Rayleigh MAC, r=4: 2,048 shared realizations per curve; standard errors for every point are in the experiment register. The line K=r is a heuristic operating reference, not an exact threshold.

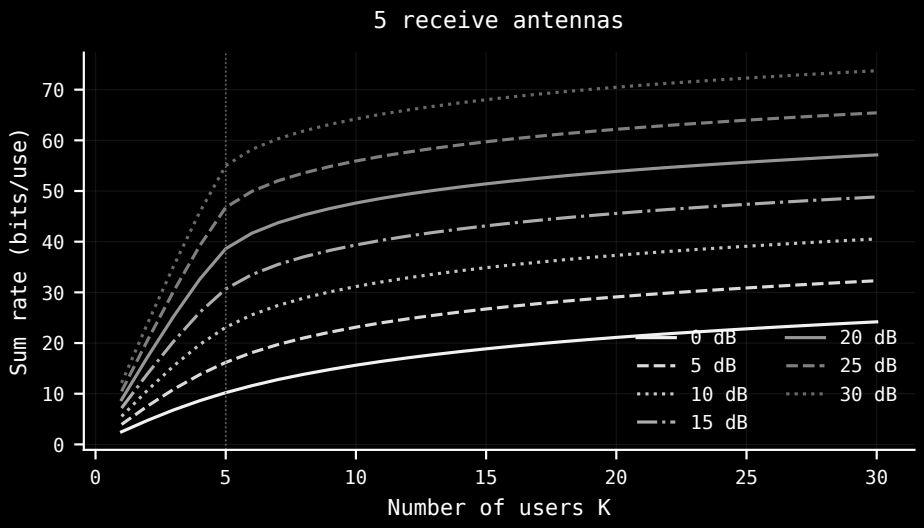


Figure 2.6: Rayleigh MAC, r=5: 2,048 shared realizations per curve; standard errors for every point are in the experiment register. The line K=r is a heuristic operating reference, not an exact threshold.

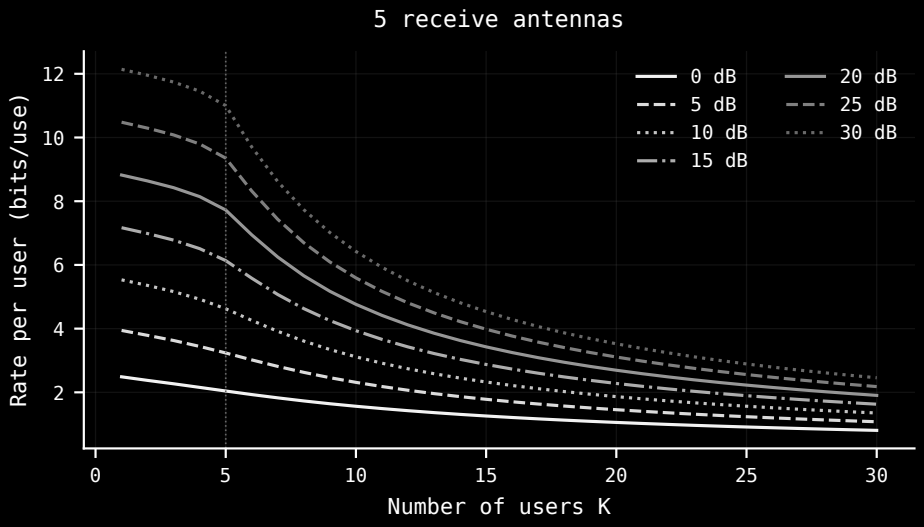


Figure 2.7: Rayleigh MAC, r=5: 2,048 shared realizations per curve; standard errors for every point are in the experiment register. The line K=r is a heuristic operating reference, not an exact threshold.

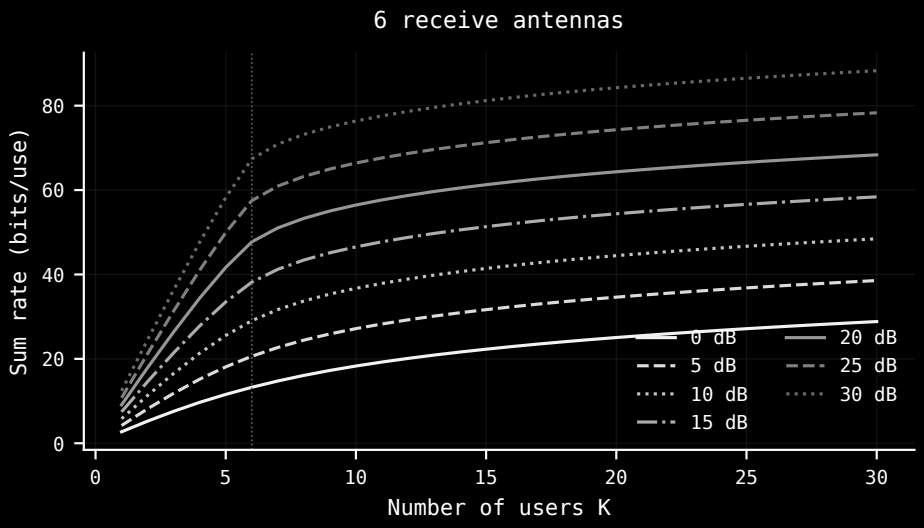


Figure 2.8: Rayleigh MAC, r=6: 2,048 shared realizations per curve; standard errors for every point are in the experiment register. The line K=r is a heuristic operating reference, not an exact threshold.

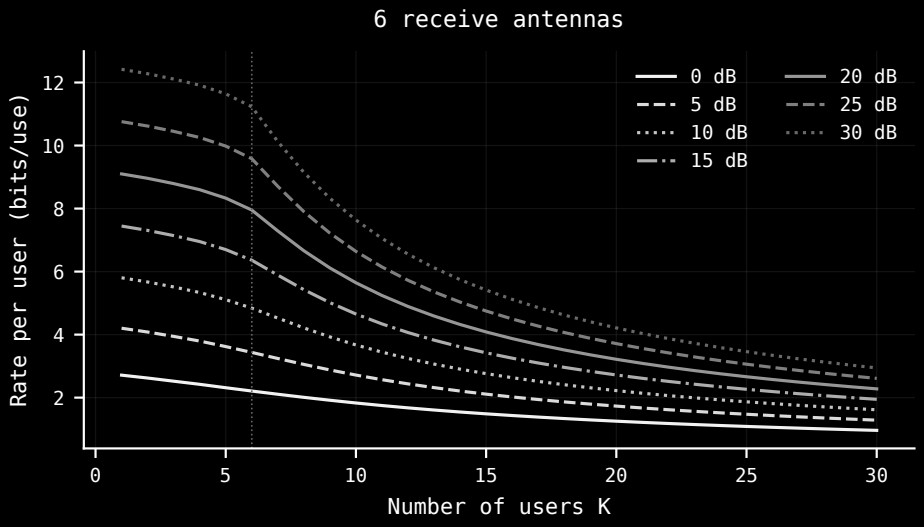


Figure 2.9: Rayleigh MAC, r=6: 2,048 shared realizations per curve; standard errors for every point are in the experiment register. The line K=r is a heuristic operating reference, not an exact threshold.

2.24 SC-097 | Low-SNR comparison with orthogonal access

Reference: [2], exercise 10.6.

Source task and conventions

Compare equal-rate orthogonal access and simultaneous transmission at low SNR.

Conventional solution and complete derivation

With K equal-power users, orthogonal access gives each a fraction $1/K$ of the resource and active power KP. Its symmetric rate in nats is

$$R_{\mathrm{OMA}} = \frac{1}{K} E \log(1 + K\rho\|h\|^2).$$

Using $E\|h\|^2 = r$ and $E\|h\|^4 = r(r+1)$ yields

$$0 \le \rho r - R_{\mathrm{OMA}} \le \tfrac{1}{2} K\rho^2 r(r+1).$$

For SDMA, the preceding proof gives $0 \le \rho r - C_{\mathrm{sym}} \le \rho^2 r(K+r)/2$. Both therefore have the same first derivative r at zero SNR, and their ratio tends to one for fixed K, r as $\rho \downarrow 0$. Converting to bits divides all terms by $\log 2$. The statement assumes average power can be concentrated into the user's active fraction; a peak-power cap can change the comparison.

Structural or analytic formulation

The first-order log-determinant term is the trace, so inter-user spatial correlations affect only higher orders. The common low-SNR performance is an energy/receive-array gain, independent of the multiple-access allocation to first order.

Comparative efficiency and scope

The two explicit remainder bounds quantify "comparable" and specify its range. A fixed low SNR does not remain in that range as K grows without bound. Evaluating the scalar OMA expression is cheaper than a matrix log-determinant, but this computational saving is separate from the proved rate agreement.

2.25 SC-098 | Grouped SDMA between fully orthogonal and fully simultaneous access

Reference: [2], exercise 10.8.

Source task and conventions

Derive the symmetric rates, low- and high-SNR limits, and the $K = 30, r = 5$ comparison.

Conventional solution and complete derivation

For $K = gs$ with groups of equal size $s \le r$, symmetry makes equal group fractions $1/g$ optimal among orthogonal group schedules. A group has active user SNR $g\rho$, so

$$R_{\text{group}} = \frac{1}{K} F_s(g\rho), \quad R_{\text{SDMA}} = \frac{1}{K} F_K(\rho), \quad R_{\text{OMA}} = \frac{1}{K} E \log(1 + K\rho \|h\|^2).$$

The equal-fraction assertion follows by concavity of the perspective $\tau F_s(\rho/\tau)$: averaging a schedule over permutations of identical groups keeps feasibility and cannot lower its common rate. For unequal group sizes s_j, the exact optimal common rate is the solution of

$$\max_{\tau_j \ge 0, \sum \tau_j = 1} \min_j \frac{\tau_j}{s_j} F_{s_j}(\rho/\tau_j).$$

This formula specifies the remainder-group case without silently assuming equal groups. At fixed K, r, the trace expansion gives $R_{\text{group}} = \rho r + O(\rho^2)$, the same as both other schemes. When $s = r$ and $\rho \to \infty$, grouped SDMA and full SDMA both have leading term $(r/K) \log \rho$; OMA has leading term $(1/K) \log \rho$. Their additive constants generally differ, so high-SNR comparability means matching pre-log, not identical rates. With an unequal last group the best leading symmetric allocation uses equal group fractions and pre-log $1/\lceil K/r \rceil$, which is comparable but not exactly r/K. The $K = 30, r = 5$ case has $g = 6$: the three expressions are $F_{30}(\rho)/30$, $F_5(6\rho)/30$, and $E \log(1 + 30\rho\|h\|^2)/30$. The experiment evaluates these on 0–30 dB using common Gaussian samples and records errors of the Monte Carlo estimates. Full SDMA is at least as good, since the grouped and orthogonal schedules are admissible strategies in the full multiple-access capacity region. Grouping retains five spatial modes within each active group and is consequently much closer in high-SNR scaling than OMA. The finite-SNR gap is reported numerically, without asserting it vanishes.

Structural or analytic formulation

A spatial-mode decomposition exposes the design tradeoff: per-group decoding involves five users instead of thirty, while the high-SNR symmetric pre-log is preserved in the divisible case. Decoding complexity is implementation-dependent and is not numerically equated to group size alone.

Comparative efficiency and scope

The exact resource fractions and power concentration are included in the rate comparison. A claim based on unboosted active power would compare different average-power budgets. The per-user rate curves and their standard errors provide the finite-parameter metric; the pre-log statements are analytically proved.

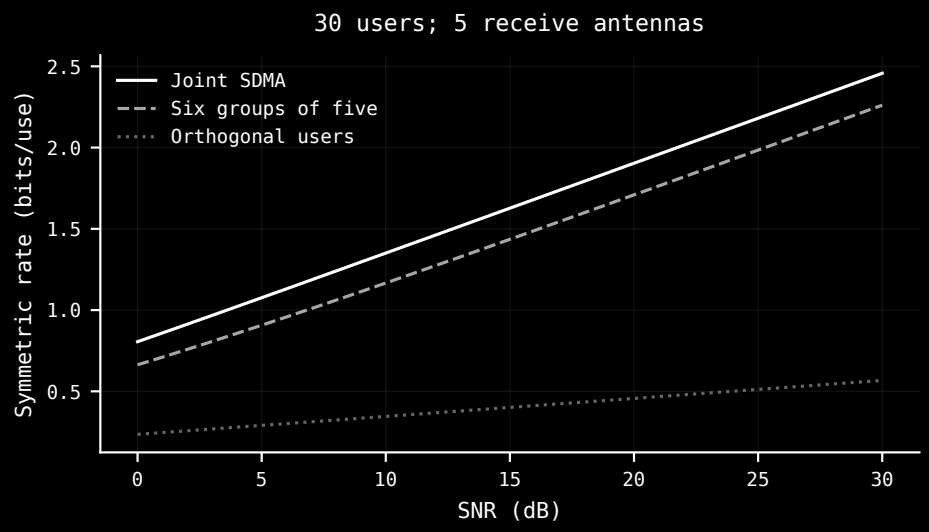


Figure 2.10: Equal per-user average power, including the appropriate instantaneous-power scaling during orthogonal resource fractions.

2.26 SC-099 | Coupled waterfilling from a rank-one resolvent

Reference: [2], exercise 10.10.

Source task and conventions

Prove the inverse identity and derive the KKT power rule and effective interference.

Conventional solution and complete derivation

Use unit noise covariance for the notation in the question and set $B_k = I + \sum_{j \neq k} P_j h_j h_j^* \succ 0$ and $a_k = h_k^* B_k^{-1} h_k$. Multiplying out verifies

$$(B_k + P_k h_k h_k^*)^{-1} = B_k^{-1} - \frac{P_k B_k^{-1} h_k h_k^* B_k^{-1}}{1 + P_k a_k}.$$

Taking the quadratic form gives $h_k^*(B_k + P_k h_k h_k^*)^{-1} h_k = a_k/(1 + P_k a_k)$, proving part (1). For part (2), the derivative of the concave instantaneous objective $\log\det(I + \sum_j P_j h_j h_j^*)$ with respect to P_k is that quadratic form. A multiplier λ for the average-power constraint gives $a_k/(1 + P_k a_k) = \lambda$ on an active coordinate, and at a zero coordinate $a_k \leq \lambda$. Hence

$$P_k^*(H) = \left(\frac{1}{\lambda} - \frac{1}{a_k}\right)_+.$$

These are necessary and sufficient KKT conditions for the concave power optimization when the multipliers enforcing the power constraints exist. They are coupled because B_k contains the other powers; writing the rule does not independently solve each coordinate. For part (3), the MMSE-output SINR is $P_k a_k$. For an active nonzero channel, the effective interference is $I_0 = \|h_k\|^2/a_k$, so the rule equals $(1/\lambda - I_0/\|h_k\|^2)_+$. The cancellation defines the continuous effective-noise value even if the chosen $P_k = 0$. If $h_k = 0$, allocate zero power and avoid division by zero. The subsequent large-system constancy claim requires a normalized random-matrix model and additional assumptions on powers and their dependence on channel columns. Finite-dimensional i.i.d. second moments alone do not prove it. Under an independently specified deterministic-equivalent approximation it can be used as an approximate scalar waterfilling rule;

the exact result proved here is the finite-dimensional identity and coupled KKT system requested in (1)–(3). The source itself cites a separate theorem for that asymptotic step.

Structural or analytic formulation

The rank-one inverse formula replaces a fresh inverse for each user by one update and separates a scalar sufficient-statistic gain. The log-determinant's second directional derivative is $-\operatorname{tr}(A^{-1}BA^{-1}B) \le 0$, after writing it as the negative trace of a Hermitian square, which proves the concavity used in the optimization.

Comparative efficiency and scope

Given one inverse/factorization, a rank-one update and a quadratic form cost $O(r^2)$ instead of $O(r^3)$ reinversion. Solving the coupled optimal powers still needs a convergence-controlled optimizer. No speed or exactness claim is attached to an unverified large-system approximation.

2.27 SC-100 | Both corner points of a two-user Gaussian MIMO MAC

Reference: [2], exercise 10.12.

Source task and conventions

Give dimensionally correct rate pairs with arbitrary positive noise power.

Conventional solution and complete derivation

Put $S_i = H_i K_i H_i^* \succeq 0$ and $R_i = N_0 I + S_i$. If user one is decoded first and user two last, Gaussian entropy differences give

$$R_1^{(A)} = \log\det(N_0 I + S_1 + S_2) - \log\det(N_0 I + S_2),$$

$$R_2^{(A)} = \log\det(I + S_2/N_0).$$

Equivalently $R_1^{(A)} = \log\det(I + R_2^{-1/2} S_1 R_2^{-1/2})$. After user one's decoded codeword is subtracted, user two sees only the background noise, proving the successive-cancellation interpretation. Reversing the order gives

$$R_1^{(B)} = \log\det(I + S_1/N_0), \quad R_2^{(B)} = \log\det(N_0 I + S_1 + S_2) - \log\det(N_0 I + S_1).$$

Both sums equal $\log\det(I + (S_1 + S_2)/N_0)$. If the drawing's labels A, B use the opposite order, interchange the labels; the decoding order specifies the rates unambiguously. The source's first corner formula omits noise normalization unless $N_0 = 1$, and its other displayed product can have incompatible matrix dimensions. The whitened expression above is valid for arbitrary transmit/receive dimensions.

Structural or analytic formulation

The mutual information with coloured residual noise is a difference of log-determinants or a determinant in whitened coordinates. Sylvester's identity $\det(I + AB) = \det(I + BA)$ allows a smaller transmit-side determinant when its dimensions are advantageous; it follows by eliminating either block in $\left(\begin{smallmatrix} I & A \\ -B & I \end{smallmatrix}\right)$.

Comparative efficiency and scope

Compute three positive-definite log-determinants and reuse them for both corners; a separate calculation for each rate is redundant. For dense receive dimension r the setup is $O(r^3)$. A rank-d transmit covariance permits d-dimensional determinant calculations after solving with the residual-noise factor.

2.28 SC-101 | Isotropic covariance maximizes every fast-fading MAC constraint

Reference: [2], exercise 10.15.

Source task and conventions

Prove simultaneous optimization of the single-user and sum constraints, hence of every linear rate objective.

Conventional solution and complete derivation

Assume independent i.i.d. proper Rayleigh channel matrices, receiver CSI only, fixed input covariances $Q_i \succeq 0$, and $\operatorname{tr} Q_i \le P_i$. The fading law is invariant under separate unitary rotations of each transmitter's columns. Each single-user constraint $E \log\det(I + H_i Q_i H_i^*/N_0)$ and the sum constraint $E \log\det(I + \sum_i H_i Q_i H_i^*/N_0)$ is concave in every covariance. Average Q_i over diagonal sign changes and then over coordinate permutations. The average is $(\operatorname{tr} Q_i/n_i)I$, the distributional invariance keeps the objective equal before averaging, and concavity makes it no smaller after averaging. Increasing the scalar diagonal power to P_i/n_i cannot decrease a constraint, because its added output covariance is positive semidefinite and each positive eigenvalue of the whitened increment increases the determinant. Thus the same pair $Q_i = (P_i/n_i)I$ simultaneously maximizes both individual bounds and the sum bound. Every feasible rate pair under any other fixed covariance lies within the region described by these three maximized constraints. Conversely the Gaussian MAC coding theorem attains that region under the isotropic pair, with time sharing on its boundary. Hence its support function maximizes every $a_1R_1 + a_2R_2$; negative coefficients select axes or the origin and are also covered by the containment argument. The resulting nonnegative region is the usual pentagon, with possible degeneracies at zero powers or zero channel dimensions.

Structural or analytic formulation

Finite group averaging is sufficient; no calculation of a full Haar integral is required. The argument uses symmetry of the entire Rayleigh distribution, a stronger hypothesis than uncorrelated entries alone, as the counterexample to Exercise 8.3 demonstrates.

Comparative efficiency and scope

The transmitter covariance search reduces from $n_1^2 + n_2^2$ real matrix parameters to the fixed scalar allocations prescribed by the total powers. Receiver expectation evaluation is still required for numerical rates. This is a proved optimizer of the stated no-transmitter-CSI model, not of arbitrary adaptive fading transmission.

2.29 SC-102 | Finite-memory Gaussian channels and the full Toeplitz limit

Reference: [2], exercise B.12.

Source task and conventions

Prove Gaussian optimality, spectral waterfilling, the Toeplitz structure, and the limiting frequency-domain capacity for a fixed finite impulse response.

Conventional solution and complete derivation

Let $h = (h_0, \dots, h_{L-1})$ be fixed, G_n the $(n + L - 1) \times n$ full linear-convolution matrix, and $Y = G_n X + Z$ with independent circular Gaussian noise $N_0 I$, $N_0 > 0$. The total input constraint is $\operatorname{tr} Q \le nP$. For any input covariance Q, maximum entropy at fixed covariance gives

$$I(X;Y) \le \log\det(I + G_n Q G_n^*/N_0),$$

with equality for a zero-mean circular Gaussian input. The Gaussian entropy bound follows from nonnegativity of relative entropy against the Gaussian with the same covariance, proved in the entropy questions; conditioning on X leaves exactly the noise entropy. Thus this class attains the optimum. Diagonalize $G_n^* G_n = V \operatorname{diag}(\lambda_1, \dots, \lambda_n) V^*$. Sylvester's determinant identity and Hadamard's inequality bound the objective by $\sum_j \log(1 + \lambda_j (V^* Q V)_{jj}/N_0)$, with equality for diagonal $V^* Q V$. Trace is unchanged. Strict concavity on positive-gain modes and the tangent inequality for a concave logarithm prove optimality of

$$p_j = (\nu_n - N_0/\lambda_j)_+, \qquad n^{-1} \sum_j p_j = P.$$

Set $p_j = 0$ at $\lambda_j = 0$. The multiplier ν_n is unique for a nonzero channel and $P > 0$; complementary slackness makes the tangent upper bound equal at this allocation. This proves finite-block waterfilling, not merely stationarity. For all valid indices, $(G_n^* G_n)_{ij} = \sum_\ell \overline{h}_\ell h_{\ell+i-j}$, extending h by zero. Therefore it is a banded Hermitian Toeplitz matrix, with symbol

$$a(\omega) = \left| \sum_{\ell=0}^{L-1} h_\ell e^{-i\ell\omega} \right|^2 .$$

The dimensionally appropriate matrix is $G_n^* G_n$; the output covariance $G_n G_n^*$ has different dimensions and boundary entries. Let $C = (\sum |h_\ell|)^2$. The triangle inequality for shifted vectors proves $\|G_n x\| \le \sum |h_\ell| \|x\|$, so all λ_j and $a(\omega)$ lie in $[0, C]$. Here is a direct proof of the needed spectral limit. For any integer $k \ge 0$, indices farther than $k(L-1)$ from either boundary have the same closed-walk expansion in the diagonal of $(G_n^* G_n)^k$ as the doubly infinite convolution operator. That common value is the constant Fourier coefficient $(2\pi)^{-1} \int a(\omega)^k \, d\omega$. At most $2k(L-1)$ indices are affected, and each diagonal value and the common coefficient have magnitude at most C^k. Hence their normalized trace difference is at most $4k(L-1)C^k/n$. Linearity proves convergence for every polynomial. Continuous functions on $[0, C]$ are uniformly approximated by Bernstein polynomials: uniform continuity controls arguments within δ, and the binomial variance bound $1/(4m)$ bounds the remaining probability by $1/(4m\delta^2)$. This proves uniform convergence of that approximation and thus, by adding two uniform errors, establishes

$$\frac{1}{n} \sum_j f(\lambda_j) \longrightarrow \frac{1}{2\pi} \int_{-\pi}^{\pi} f(a(\omega)) \, d\omega$$

for every continuous f. This supplies the Toeplitz limit required here with its hypotheses and proof. For fixed ν, $g_\nu(t) = (\nu - N_0/t)_+$, defined as zero at $t = 0$, is continuous on $[0, C]$ and one-Lipschitz in ν. The limiting power function is continuous and strictly increasing after it becomes positive; it tends to infinity because a nonzero finite polynomial has positive gain on a set of positive measure. Spectral convergence applied to g_ν and a pair of levels bracketing the unique limiting solution proves $\nu_n \to \nu$, where

$$\frac{1}{2\pi} \int g_\nu(a(\omega))\, d\omega = P.$$

The function $f_\nu(t) = \log \max\{1, \nu t/N_0\}$ is also continuous, and is locally uniformly Lipschitz in positive ν. Apply the same convergence to obtain

$$\boxed{\lim_{n\to\infty} \frac{1}{n} C_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \max\{1, \nu a(\omega)/N_0\}\, d\omega.}$$

All rates are nats per complex use. Physical rate per second requires the corresponding sampling-rate factor and consistent spectral-density normalization. For $P = 0$ or the zero channel the capacity is zero. The argument holds for fixed finite memory L; memory growing with n requires a different boundary estimate.

Structural or analytic formulation

Fourier symbols turn a sequence of growing Toeplitz optimizations into scalar spectral allocation. The finite-block diagonalization remains exact; a circulant/FFT approximation is justified by the proved limiting argument. For a two-tap channel, a separate continuant identity gives an explicit finite-n determinant and a sharper $O(1/n)$ fixed-input error, included in the computation appendix.

Comparative efficiency and scope

Dense diagonalization costs $O(n^3)$ arithmetic and $O(n^2)$ storage. A banded eigensolver is a stronger finite-memory baseline. Sampling a fixed-length symbol on M frequencies costs $O(ML)$ directly, or an FFT after padding, followed by $O(M \log M)$ water-level sorting. This computes an approximation to the limiting problem, not an exactly equivalent finite-block solution; its quadrature and finite-block errors are explicit separate obligations. The polynomial trace estimate supplies a rigorous error strategy with a chosen continuous-function approximation.

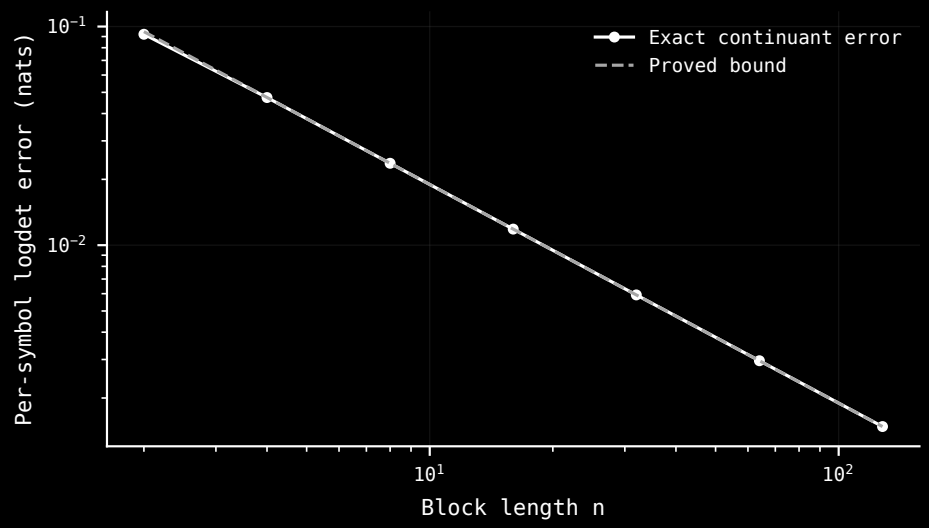


Figure 2.11: A two-tap numerical check of the general finite-memory theorem. The complete solution treats arbitrary fixed finite tap length and optimized spectral power.

2.30 SC-103 | Fixed-association power control and its minimal feasible vector

Reference: [2], exercise 4.5.

Source task and conventions

Form the interference matrix for a fixed base-station association, solve the fixed-target equations, and identify the feasibility condition.

Conventional solution and complete derivation

Write $N = N_0 W > 0$ and $\tau_i = \beta_i / G$ when β_i denotes the physical post-despreading SINR target. For the fixed receiving station c_i, the inequality is

$$p_i \geq \sum_{j \neq i} \tau_i \frac{g_{jc_i}}{g_{ic_i}} p_j + \tau_i \frac{N}{g_{ic_i}}.$$

Thus $F_{ii} = 0$, $F_{ij} = \tau_i g_{jc_i} / g_{ic_i}$ and $b_i = \tau_i N / g_{ic_i} > 0$. The source's displayed F, b are obtained when its target symbol has already absorbed $1/G$. Retaining both interpretations without this normalization would be inconsistent. If $\rho(F) < 1$, the matrix series $S = \sum_{k=0}^{\infty} F^k$ converges and $(I - F)S = S(I - F) = I$, by taking limits of the finite telescoping products. Hence

$$p_* = Sb = (I - F)^{-1} b > 0, \qquad p_* = F p_* + b.$$

Every component attains its target. If $q \geq Fq + b$, repeated substitution gives $q \geq \sum_{k=0}^{m-1} F^k b + F^m q \geq \sum_{k=0}^{m-1} F^k b$, and passage to the limit proves $q \geq p_*$. This proves componentwise minimality and therefore minimality for every nondecreasing power objective. Conversely, if a positive feasible q exists, $D^{-1} F D$, $D = \operatorname{diag} q$, has row sums at most $1 - \min_i b_i / q_i < 1$. Its infinity norm bounds its spectral radius, so $\rho(F) < 1$. Thus feasibility is exactly the stated spectral condition.

Structural or analytic formulation

The resolvent $(I-F)^{-1}$ is the sum of all interference walks. Its convergence is a finite-dimensional instance of a Neumann-series argument used in the functional-analytic synopses. A feasible vector itself supplies the weighted-norm contraction certificate, so a separate eigenvalue computation is optional.

Comparative efficiency and scope

A dense direct solution costs $O(K^3)$ arithmetic. A sparse iteration costs $O(\operatorname{nnz} F)$ per sweep. If a positive weight q gives row bound $c < 1$, the weighted error after t sweeps is at most c^t times its initial value. The certificate gives an explicit iteration count for a specified tolerance; an unquantified spectral-radius statement alone does not.

2.31 SC-104 | Perron–Frobenius terminology and a feasibility equivalence

Reference: [2], exercise 4.6.

Source task and conventions

Examine the positivity claim and prove equivalence of a positive feasible vector, spectral radius below one, and a positive inverse.

Conventional solution and complete derivation

For a matrix with strictly positive off-diagonal entries, the directed graph is strongly connected. For $K \geq 3$, every entry of F^2 is positive: a diagonal entry uses a two-edge return path and an off-diagonal entry uses a distinct intermediate index. For $K = 2$, however,

$$F = \begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix}, \quad a, b > 0$$

has eigenvalues $\pm\sqrt{ab}$ and no strictly positive power. The source's definition by a positive power describes a *primitive* matrix, not merely an irreducible one. Its strict spectral-dominance assertion therefore needs the primitive hypothesis. The ordinary irreducible Perron–Frobenius statement supplies a positive simple Perron eigenvalue and positive eigenvectors; other eigenvalues can have the same modulus in the periodic case. The $K = 1$ no-interference case is immediate. For $b > 0$ and a strongly connected nonnegative F, prove the requested equivalence directly. If $q \geq Fq + b$ with $q > 0$, weighted row sums as in the preceding question give $\rho(F) < 1$. If $\rho(F) < 1$, then $(I - F)^{-1} = \sum_{k \geq 0} F^k \geq 0$. Strong connectivity supplies a positive path from each index to each other; the corresponding power has a positive entry, and the identity supplies the diagonal. The sum is therefore entrywise strictly positive. Finally, a strictly positive inverse gives $q = (I - F)^{-1} b > 0$ and $q = Fq + b$. This proves all implications without relying on the source's erroneous primitive claim.

Structural or analytic formulation

A row-weight certificate is a constructive alternative to appealing to an eigenvector. The inverse entries are sums of nonnegative path weights; positivity follows from connectivity and does not require a single common positive power.

Comparative efficiency and scope

Testing a supplied positive weight uses $O(\operatorname{nnz} F)$ arithmetic, against a fresh dense eigendecomposition's $O(K^3)$. Finding a good weight can itself require a solve or iteration. These verification and construction costs are kept separate.

2.32 SC-105 | Joint association and componentwise power minima

Reference: [2], exercise 4.7.

Source task and conventions

Prove downward target feasibility, closure under componentwise power minima, and existence of a globally minimal assignment.

Conventional solution and complete derivation

Assume finitely many admissible receiving stations, positive direct gains and noise, and a feasible target vector. Lowering any target retains the same feasible power vector, proving the first assertion. Let $p^{(1)}, p^{(2)}$ be feasible with associations $c^{(1)}, c^{(2)}$. Put $q_i = \min\{p_i^{(1)}, p_i^{(2)}\}$ and choose c_i from an index attaining this minimum. At that station its desired received power is unchanged from the chosen feasible design, while every interfering power weakly decreases. Thus its SINR weakly increases and q is feasible. Ties may be broken arbitrarily. This proves the requested construction, including the simultaneous changes of association. For each feasible association take its unique minimal fixed-target power vector, proved in the preceding questions. There are finitely many associations. Iterating the componentwise-minimum construction over these vectors gives a feasible q that is no larger than any of them. Choose the resulting feasible association c; its minimal vector $p_*(c)$ is at most q. Since q was no larger than every feasible association's minimal vector, it is also at most $p_*(c)$, so equality holds. Thus one association attains the unique globally minimal power vector. Several associations may attain that same vector; uniqueness of the association does not follow.

Structural or analytic formulation

This is an order-theoretic closure argument on the positive cone. The interference map's lower envelope retains monotonicity, even though the identity of the minimizing station can change. No differentiability at a station switch is required.

Comparative efficiency and scope

Enumerating all L^K associations proves existence but is not an efficient algorithm. The minimum-interference iteration of the next question evaluates KL local alternatives per sweep and converges to the same minimal vector under feasibility. Its advantage requires a convergence bound or measured stopping criterion, not the existence proof alone.

2.33 SC-106 | Convergence of synchronous minimum-interference iteration

Reference: [2], exercise 4.8.

Source task and conventions

Prove positivity, monotonicity, scalability, uniqueness, and convergence from every nonnegative initial vector.

Conventional solution and complete derivation

For target $\tau_i > 0$, define

$$T_i(p) = \min_{\ell \in S_i} \frac{\tau_i}{g_{i\ell}} \left(N + \sum_{j \neq i} g_{j\ell} p_j \right), \quad N > 0.$$

It is strictly positive. Every term is nondecreasing in each coordinate, hence the minimum is monotone. For $\alpha > 1$, each affine term at αp equals α times its value at p minus $(\alpha - 1)\tau_i N / g_{i\ell}$. Choosing a minimizer at p shows $T_i(\alpha p) < \alpha T_i(p)$. This strict inequality, supplied by positive noise, is stronger than the source's weak scalability statement. The finite minimum is continuous. If p_*, q_* are distinct positive fixed points, interchange them if necessary so that $\alpha = \max_i p_{*,i}/q_{*,i} > 1$. At an index of equality, monotonicity and strict scalability give $p_{*,i} = T_i(p_*) \leq T_i(\alpha q_*) < \alpha T_i(q_*) = p_{*,i}$, a contradiction. Hence the fixed point is unique. Existence follows from any feasible vector $q \geq T(q)$: starting at zero, $\ell^{(t)} = T^t(0)$ increases and is bounded above by q. Its coordinatewise limit is a fixed point by continuity. For an arbitrary nonnegative start $p^{(0)}$, choose $\alpha \geq 1$ so that $p^{(0)} \leq \alpha p_*$. The sequence $u^{(t)} = T^t(\alpha p_*)$ decreases, stays above p_*, and converges to a fixed point, hence to p_*. Monotonicity gives $\ell^{(t)} \leq T^t(p^{(0)}) \leq u^{(t)}$; the squeeze proves convergence. Every feasible q bounds $\ell^{(t)}$, so $p_* \leq q$: the limit is globally power-minimal. Selecting a minimizing station in each coordinate realizes it as an association.

Structural or analytic formulation

The proof uses positive-cone order, continuity and strict subhomogeneity. It remains valid at nonsmooth association boundaries. Local station changes do not require a holomorphic or gradient-based optimization model.

Comparative efficiency and scope

A direct implementation forms station interference sums $I_\ell = N + \sum_j g_{j\ell} p_j$ in $O(KL)$ work and reuses $I_\ell - g_{i\ell} p_i$, so a sweep costs $O(KL)$. Recomputing each sum for each user costs $O(K^2 L)$. This is a proved per-sweep saving. Feasibility yields convergence but no uniform finite iteration bound without quantitative gain and margin assumptions.

2.34 SC-107 | Totally asynchronous power updates

Reference: [2], exercise 4.9.

Source task and conventions

Extend the power-control convergence proof to updates with unbounded but eventually fresh delays.

Conventional solution and complete derivation

Assume each coordinate is updated infinitely often and every read time used by later updates tends to infinity (the total-asynchrony hypotheses). Retain the monotone envelopes $\ell^{(k)} = T^k 0$ and $u^{(k)} = T^k(\alpha p_*)$ from the synchronous proof, choosing α large enough to bound the finite initial history. All iterates initially lie in the zeroth envelope; any update with inputs in that envelope lies in $[\ell^{(1)}, u^{(1)}] \subseteq [\ell^{(0)}, u^{(0)}]$, so it cannot leave it. Inductively suppose all sufficiently late stored values lie in $[\ell^{(k)}, u^{(k)}]$. Eventual freshness gives a time after which every read uses such a value. The infinite-update condition then gives a later finite time by which every coordinate has updated at least once using only those reads. Monotonicity places each resulting coordinate in $[T_i(\ell^{(k)}), T_i(u^{(k)})]$. All subsequent updates remain there, proving eventual inclusion in the $(k+1)$st envelope. Both envelopes converge to p_*, so for every $\epsilon > 0$ some k makes their coordinate widths below ϵ, and all sufficiently late asynchronous iterates are within ϵ of p_*. This proves convergence. Neither periodic schedules nor a common finite delay bound was used.

Structural or analytic formulation

The proof is a nested-order-interval argument. The epochs are existential synchronization points in the analysis; the implementation does not have to synchronize. Failure of eventual freshness can preserve obsolete interference indefinitely and invalidates the argument.

Comparative efficiency and scope

Unbounded gaps between updates exclude a universal elapsed-time or update-count rate. With bounded delay and bounded inter-update gaps, the same envelopes yield a rate in epochs once a contraction factor is available. Stating only a synchronous spectral rate for the totally asynchronous case would overclaim.

2.35 SC-108 | Two-user examples, trajectories, and association counterexamples

Reference: [2], exercise 4.10.

Source task and conventions

Construct feasible and infeasible two-user examples, trace updates, and test strongest-gain association.

Conventional solution and complete derivation

Use $N = 1$ and normalized equal target $\tau = 1/2$. With one base station and unit gains, feasibility is the intersection $p_1 \geq p_2/2 + 1/2$, $p_2 \geq p_1/2 + 1/2$; its componentwise minimum is $(1, 1)$. At target $\tau = 1$, the two inequalities imply $p_1 \geq p_2 + 1$ and $p_2 \geq p_1 + 1$, whose sum is impossible. This explicitly gives both cases and their geometric certificate. For the feasible unit-gain case, from $(0, 3)$ the synchronous iterates are $(2, 1/2)$, $(3/4, 3/2)$, $(5/4, 7/8)$, $(15/16, 9/8)$, and converge

to $(1,1)$. Indeed $p^{(t)} - p_* = F^t(-1,2)^\mathsf{T}$ and $F^2 = I/4$, proving the entire trajectory. For two stations take gain matrix (rows users, columns stations)

$$g = \begin{pmatrix} 1 & 2/5 \\ 1/5 & 1 \end{pmatrix}.$$

The association $(1,2)$ has $F_{12} = 1/10, F_{21} = 1/5$ and $b = (1/2, 1/2)$, giving $p_* = (55/98, 30/49)$. Associations $(1,1)$ and $(2,2)$ give $(1,5)$ and $(5/2,1)$; $(2,1)$ is infeasible because its interference-product exceeds one. Direct comparison proves $(1,2)$ is globally minimal in this example; the lower-envelope iteration converges to it. Strongest direct gain is not generally optimal. Take instead

$$g = \begin{pmatrix} 1 & 9/10 \\ 100 & 1/10 \end{pmatrix}.$$

Choosing the strongest station for both users gives $(p_1, p_2) = (1, 1/100)$. Association $(2,1)$ yields equations $p_1 = 5/9 + p_2/18$ and $p_2 = 1/200 + p_1/200$, hence $p_1 = 2001/3599$, $p_2 = 28/3599$. Both are strictly smaller than the strongest-gain design, despite user one's weaker desired gain. Reduced interference makes that association preferable. Every numerical pair is checked by substitution in its original SINR inequalities.

Structural or analytic formulation

The two-dimensional drawings are exact affine halfspace geometry, while the switched system is the lower envelope of those affine maps. The rational solutions and $F^2 = I/4$ identity provide independent certificates for the plots.

Comparative efficiency and scope

Four associations suffice for exhaustive verification in the two-user/two-station example. Generally enumeration has L^K cases, while a reused-interference sweep costs $O(KL)$. The latter converges by the preceding proof, with no undocumented global speed factor.

Iteration	p1	p2
0	0	3
1	2	0.5
2	0.75	1.5
3	1.25	0.875
4	0.9375	1.125
5	1.0625	0.96875
6	0.984375	1.03125
7	1.015625	0.9921875
8	0.99609375	1.0078125
9	1.00390625	0.998046875
10	0.999023438	1.00195312

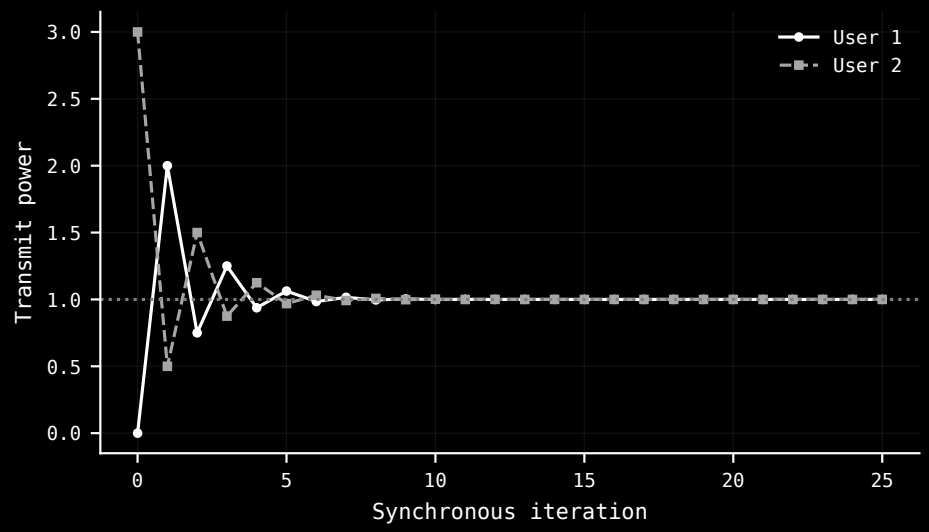


Figure 2.12: The source example starts at (0,3); alternating coordinate errors decay by 1/2, and two iterations scale the error by 1/4.

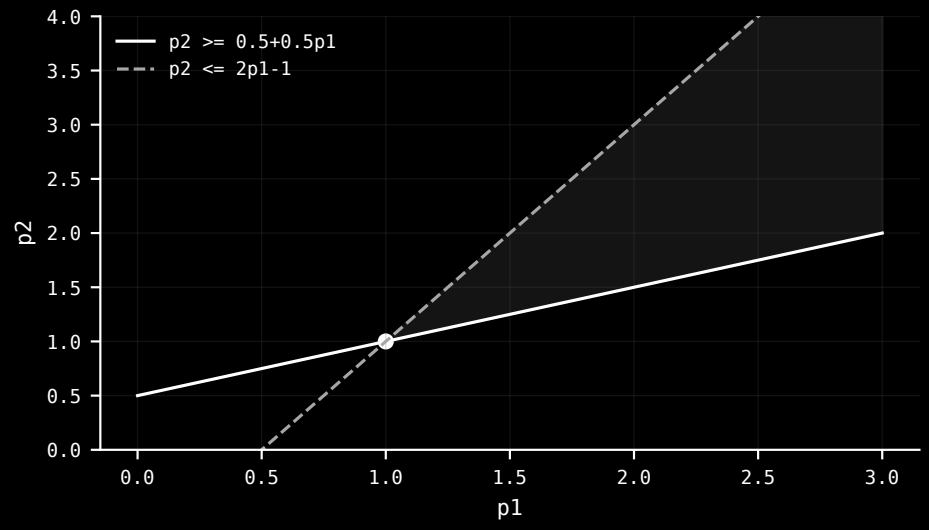


Figure 2.13: The feasible region for tau=1/2 has componentwise minimal point (1,1). At tau=1 the two affine inequalities contradict each other.

2.36 SC-109 | The chi-square density and its recurrence

Reference: [2], exercise A.1.

Source task and conventions

Derive the one-dimensional density, the n to $n+2$ recurrence, and the general density.

Conventional solution and complete derivation

For $A = W^2$ with $W \sim N(0,1)$, the two inverse branches $W = \pm\sqrt{a}$ each have derivative magnitude $1/(2\sqrt{a})$. Thus $f_1(a) = e^{-a/2}/\sqrt{2\pi a}$ for $a > 0$, and zero for $a < 0$. Two independent normal coordinates give the radial integral $P(W_1^2 + W_2^2 \le a) = 1 - e^{-a/2}$, so $f_2(a) = e^{-a/2}/2$. Let $\Gamma(t) = \int_0^\infty u^{t-1}e^{-u}du$; integration by parts gives $\Gamma(t+1) = t\Gamma(t)$ and the Gaussian integral gives $\Gamma(1/2) = \sqrt{\pi}$. Inductively suppose

$$f_n(a) = \frac{a^{n/2-1}e^{-a/2}}{2^{n/2}\Gamma(n/2)}.$$

Adding two independent squared normal coordinates convolves this with f_2. For $a > 0$ the convolution is

$$\frac{e^{-a/2}}{2^{n/2+1}\Gamma(n/2)} \int_0^a u^{n/2-1} du = \frac{a}{n} f_n(a) = f_{n+2}(a).$$

The bases one and two establish the density for every positive integer n, as well as the requested recurrence. Its integral is one by the substitution $u = a/2$. Density values at the single endpoint zero may be defined arbitrarily; the one-dimensional density has an integrable singularity there.

Structural or analytic formulation

The squared radius is a gamma variable with shape $n/2$ and scale two. Rotational symmetry reduces an n-dimensional density calculation to this one radial coordinate.

Comparative efficiency and scope

One gamma-density formula replaces integration over all coordinate tuples. Evaluating a radius from a supplied n-vector still requires n squares and $n-1$ additions; that dense-input cost cannot be bypassed.

2.37 SC-110 | Sphere hardening with a controlled shell approximation

Reference: [2], exercise B.10.

Source task and conventions

Answer the seven shell-volume and concentration questions, including the limits of the proposed approximation.

Conventional solution and complete derivation

Let $\widetilde{x} = x/\sqrt{N}$ with independent $x_i \sim N(0, P)$, $P > 0$. (1) Linearity gives $E\|\widetilde{x}\|^2 = P$. (2) Since $EX_i^4 = 3P^2$ (two integrations by parts in the normal density), $\mathrm{Var}(x_i^2) = 2P^2$; independence gives $\mathrm{Var}\,\|\widetilde{x}\|^2 = 2P^2/N$. Chebyshev therefore proves

$$P(|\|\widetilde{x}\|^2 - P| \geq \epsilon) \leq \frac{2P^2}{N\epsilon^2} \longrightarrow 0.$$

(3) Scaling a unit ball by radius r multiplies volume by r^N, so the shell volume is $B_N[\rho^N - (\rho-\delta)^N]$. Polar integration of the Gaussian integral $\int_{\mathbb{R}^N} e^{-\|u\|^2} du = \pi^{N/2}$ gives $B_N = \pi^{N/2}/\Gamma(N/2+1)$. The gamma recurrence produces exactly the even and odd formulas printed in the question. (4) By Taylor's theorem or integrating Nr^{N-1},

$$N\delta(\rho-\delta)^{N-1} \leq \rho^N - (\rho-\delta)^N \leq N\delta\rho^{N-1}.$$

Thus $NB_N\rho^{N-1}\delta$ is the thin-shell first-order approximation for fixed N and $\delta/\rho \to 0$. Uniform small relative error as N grows requires $N\delta/\rho \to 0$, not just a fixed small relative width. (5) The density is $(N/(2\pi P))^{N/2}\exp[-N\|a\|^2/(2P)]$. Replacing it throughout the shell by its outer-boundary value and using the first-order volume gives a constant times $\rho^{N-1}\delta\exp[-N\rho^2/(2P)]$. If $\delta = c\rho$, its dependence on ρ is $[\rho\exp(-\rho^2/(2P))]^N$. The source's equality sign suppresses a factor depending on N, P, c; this expression is a proportionality approximation, not a normalized probability formula. Density replacement has ratio bounded by $\exp[N(2\rho\delta - \delta^2)/(2P)]$, so it too is not uniform for fixed c as $N \to \infty$. (6) The derivative of $\log\rho - \rho^2/(2P)$ is $1/\rho - \rho/P$, positive below $\sqrt{P}$ and negative above. Hence that approximate expression has its unique positive maximum at $\rho^2 = P$. (7) The rigorous sphere-hardening conclusion already follows from (2), without an invalid interchange of the thin-shell and large-N limits. An exact shell probability is

$$F_{\chi^2_N}(N\rho^2/P) - F_{\chi^2_N}(N(\rho-\delta)^2/P),$$

using the gamma density derived in A.1. The exact radial-density mode is $\sqrt{P(N-1)/N}$ for $N > 1$, and it tends to $\sqrt{P}$.

Structural or analytic formulation

The normalized energy is a mean of independent squared normals. Its variance certificate is dimension-explicit and avoids the uncontrolled shell approximations in the heuristic argument. The gamma transform also gives exponential concentration bounds when stronger tails are required.

Comparative efficiency and scope

For relative squared-radius tolerance ε, Chebyshev guarantees failure probability at most $2/(N\varepsilon^2)$. This is a proved metric, with no Monte Carlo sampling required. Exact gamma probabilities cost special-function evaluation. The supplied radial plots illustrate these formulas and are not used as proofs.

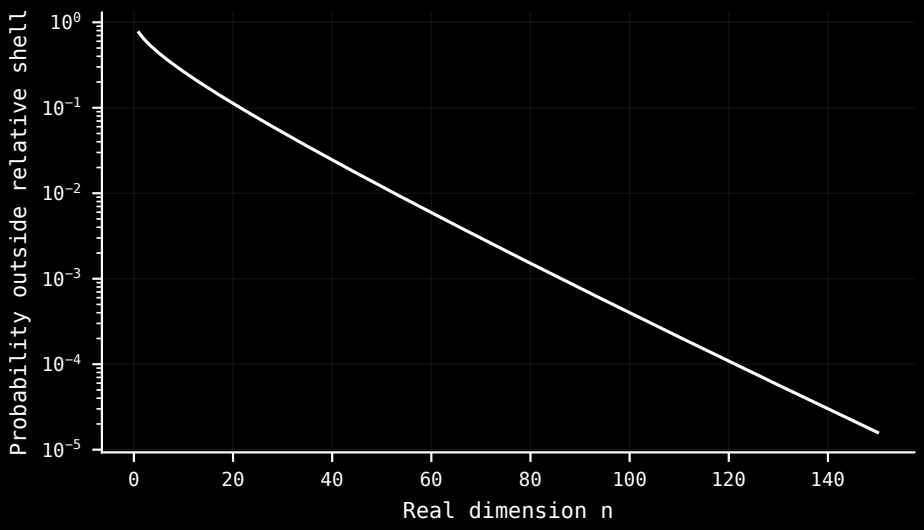


Figure 2.14: Exact chi-square radial probabilities for the shell [(1-epsilon)sqrt(n),(1+epsilon)sqrt(n)] with epsilon=0.25; the solution provides the asymptotic and finite moment reasoning.

Chapter 3

Analytic response, passivity and branch selection

3.1 All-pass filters: a phase budget from a proper map

3.1.1 A causal convention, not just a unit-circle picture

For real $|a| < 1$ take the discrete-time transfer function

$$H_a(z) = \frac{z^{-1} - a}{1 - az^{-1}}, \qquad y[k] - ay[k-1] = u[k-1] - au[k].$$

Its impulse response is causal and absolutely summable: $h[0] = -a$ and $h[k] = (1 - a^2)a^{k-1}$ for $k \geq 1$. The pole is $z = a$ inside the unit circle. The zeros and the pole belong to the transfer-variable convention z, whereas the proper disk factor $b_a(\zeta) = (\zeta - a)/(1 - a\zeta)$ uses $\zeta = z^{-1}$. Confusing these coordinates reverses a stability statement.

Theorem 3.1 (E01: All-pass magnitude and positive group delay)**.** *With $H_a(e^{i\omega}) = b_a(e^{-i\omega})$, its magnitude is one and*

$$\tau_a(\omega) = -\frac{d}{d\omega} \arg H_a(e^{i\omega}) = \frac{1 - a^2}{1 - 2a\cos\omega + a^2} > 0.$$

The delay integral over a complete frequency period is 2π. For a cascade of d stable elementary factors the delays add and the integral is $2\pi d$.

Proof. On $|\zeta| = 1$, the identity in M04 gives $|\zeta - a| = |1 - a\zeta|$. Logarithmic differentiation along $\zeta = e^{i\theta}$ gives $d \arg b_a(e^{i\theta})/d\theta = (1 - a^2)/|e^{i\theta} - a|^2$. Substituting $\theta = -\omega$ and the minus sign in the definition of delay proves the formula. Its integral is 2π either by the single boundary winding or by integrating the uniformly convergent Fourier expansion $1 + 2\sum_{n\geq 1} a^n \cos(n\omega)$. Logarithmic derivatives add under multiplication. □

At $a = 1/2$, $\tau(0) = 3$ samples and $\tau(\pi) = 1/3$ sample. Constant magnitude is therefore compatible with substantial phase dispersion. These are delays in sample units; a sampling period T_s converts the result to seconds by multiplication.

3.1.2 A functional-analytic interpretation with a truncation boundary

By Parseval, the absolutely summable convolution above defines an isometry on two-sided ℓ^2 signals, since multiplication by its boundary response preserves magnitude. With zero initial state, causal ℓ^2 inputs also have the same total output energy when the entire infinite tail is included. A finite recording window can lose energy into the omitted tail.

Theorem 3.2 (E02: Exact impulse-tail energy)**.** *For the single factor, the energy after sample $N \geq 0$ is*

$$\sum_{k>N} |h[k]|^2 = (1-a^2)a^{2N}.$$

The prefix energy plus this expression equals one.

Proof. Sum $(1-a^2)^2 \sum_{k=N+1}^{\infty} a^{2k-2}$. The geometric denominator is $1-a^2$, nonzero and positive. Including $h[0]^2 = a^2$ supplies the total energy identity. □

This gives an exact stopping rule, instead of claiming that a finite Toeplitz matrix is unitary because the infinite multiplier is inner. At $a = 1/2, N = 4$, the missing energy is $3/1024$. At $a \uparrow 1$, any fixed window behaves differently from a uniform infinite-tail estimate.

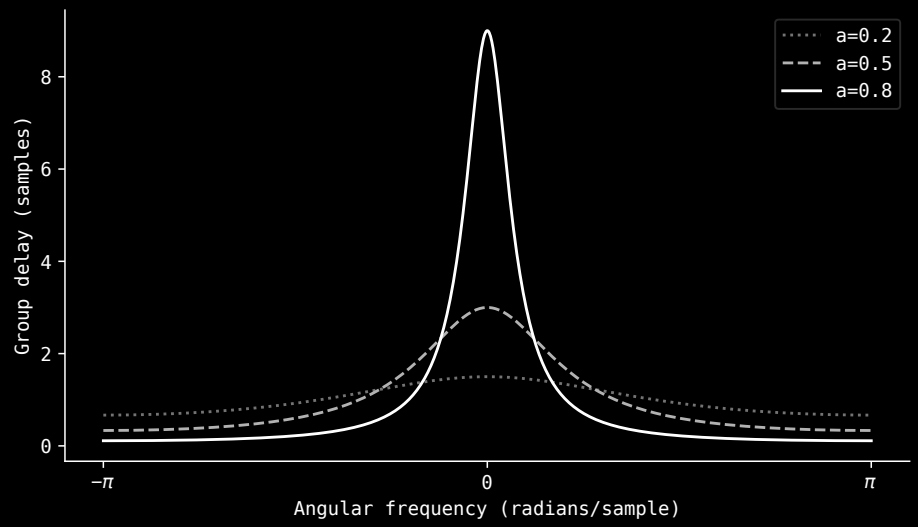


Figure 3.1: The exact delay curves for three interior values of the all-pass parameter. The height changes while the integral remains 2π. The endpoint $a = 1$ is excluded: a cancelled constant factor is a different-degree map, not a counterexample to the integral theorem.

Applicability. Unit magnitude at a finite frequency grid does not establish all-pass behavior on the whole circle. A product must be reduced and its pole convention fixed before a degree or stability claim. Inner multiplication is an isometry; composition with an arbitrary disk map is not automatically an isometry on the same Hardy norm.

3.2 From upper-half-plane maps to positive-real impedances

Let $F : \mathbb{H} \to \mathbb{H}$ be the rational proper map in M05. For $\operatorname{Re} s > 0$ define

$$Z(s) = -iF(is).$$

Then $is \in \mathbb{H}$ and $\operatorname{Re} Z(s) = \operatorname{Im} F(is) > 0$. This algebraic conversion establishes positivity in a half-plane; realizing an impedance with real physical components requires the additional reality symmetry $Z(\bar{s}) = \overline{Z(s)}$ and appropriate dimensional scaling.

Theorem 3.3 (E03: A lossless series LC model and a paired-pole model)**.** *For $F(z) = Lz - 1/(Cz)$, $L, C > 0$,*

$$Z(s) = Ls + \frac{1}{Cs}$$

is the impedance of an ideal series inductor and capacitor. More generally

$$F(z) = Lz - \frac{\lambda}{z-\Omega} - \frac{\lambda}{z+\Omega} \quad \Longrightarrow \quad Z(s) = Ls + \frac{2\lambda s}{s^2+\Omega^2}$$

for $L, \lambda, \Omega > 0$. The second term is the impedance of an ideal parallel LC branch with $C_p = 1/(2\lambda)$ and $L_p = 2\lambda/\Omega^2$ in normalized units.

Proof. The first identity is direct substitution. For the second, sum the two fractions to $2\lambda z/(z^2 - \Omega^2)$ and substitute $z = is$. A parallel LC impedance is $sL_p/(1 + s^2 L_p C_p)$; the stated choices give $2\lambda s/(s^2 + \Omega^2)$. Positivity in the right half-plane also follows from the original upper-half-plane formula; the imaginary-axis poles represent lossless resonances and do not contradict analyticity on the open half-plane. □

For actual units, choose a reference impedance Z_0 and frequency ω_0 and write $Z_{\text{physical}}(s) = Z_0 \widehat{Z}(s/\omega_0)$. A term $\widehat{L}\, s/\omega_0$ has physical inductance $Z_0\widehat{L}/\omega_0$; a term $1/(\widehat{C}s/\omega_0)$ has physical capacitance $\widehat{C}/(Z_0\omega_0)$. Omitting these conversions can turn a correct normalized rational identity into a dimensionally wrong circuit claim.

A single unpaired real pole of F generally breaks the required real-time symmetry after the Cayley rotation. The subclass with paired $\pm\Omega$ poles and no incompatible imaginary constant is physically special. General positive-real impedances include dissipative elements; a resistor is not a nonconstant proper holomorphic self-cover and should not be excluded from circuit theory merely because it is outside the classification in source Exercise 3.3.

3.3 A stability certificate with a simpler independent baseline

Consider the normalized real characteristic polynomial

$$p_K(s) = s^3 + 6s^2 + 11s + 6 + K.$$

It may arise from a specified feedback model, but no actual plant is asserted here. We assess its roots, not wiring or a hidden state-space realization.

Theorem 3.4 (E04: Exact stability interval and crossing)**.** *All roots lie in the open left half-plane exactly when $-6 < K < 60$. At $K = 60$, $p_{60} = (s+6)(s^2+11)$; the imaginary-axis pair is $\pm i\sqrt{11}$.*

Proof. The first column of the cubic Routh array is

$$1, \quad 6, \quad \frac{60-K}{6}, \quad 6+K.$$

Strict positivity gives the displayed interval by the cubic Routh criterion. The endpoint factorizations are direct algebra: at $K = -6$ there is a zero root; at $K = 60$ the quadratic factor gives the stated imaginary roots. Alternatively the cubic Hurwitz determinants are 6, $60 - K$ and $(6 + K)(60 - K)$, with the same positivity conditions. □

The Routh criterion is an imported finite-polynomial theorem, not derived from a root plot. The educational branches derive the cubic boundary by imaginary substitution and explain its relation to a continuous root count. Routh is the simpler baseline for this family; the contour method below adds a reusable whole-boundary certificate format.

Theorem 3.5 (E05: Uniform contour exclusion for a gain interval)**.** *The supplied exact certificate for the convex family p_K, $0 \le K \le 50$, proves that the number of zeros in the rectangle $0 < \operatorname{Re} s < 10$, $|\operatorname{Im} s| < 10$ is zero. All roots satisfy $|s| < 10$, so every member of that gain interval is strictly stable. The separate certificate for $K = 61$ counts two right-half-plane roots.*

Proof. CSE C02–C04 prove the checker theorem: exact Bernstein control points for every polynomial vertex lie in a common strict open half-plane on each contour segment. Convexity gives the same exclusion for every intermediate gain, and homotopy invariance fixes the winding. The verifier returns zero on $[0, 50]$ and two at 61. To see that this rectangle suffices, for $r \geq 10$,

$$r^3 > 6r^2 + 11r + 67.$$

At $r = 10$ the difference is $223 > 0$ and its derivative $3r^2 - 12r - 11$ is positive thereafter. The leading term therefore dominates the remaining coefficients for every gain considered. No root can have modulus at least ten. Boundary exclusion includes the imaginary-axis segment and hence gives strict, not marginal, stability for the zero-count case. □

The contour is positively oriented around the right-half-plane rectangle. Its imaginary-axis side is traversed downwards. A conventional Nyquist plot with the opposite orientation would change a winding sign; the relationship between zeros, poles and encirclements must be rederived with that orientation, not recalled without its contour convention.

At $K = 60$ the boundary hypothesis fails. The code rejects the attempted certificate instead of rounding a nearly imaginary root to one side. A solver's inability to certify another input does not by itself prove a boundary root; here the exact factorization supplies that additional fact.

Internal versus external stability. The state matrix $A = \mathrm{diag}(1, -1)$ with input $B = (0, 1)^T$ and output $C = (0, 1)$ has transfer function $1/(s + 1)$ but an internally unstable mode e^t if its first state is excited initially. A reduced transfer polynomial omits that state. A winding theorem about the transfer denominator cannot certify a different characteristic polynomial. This is the precise reason to name the object being counted.

3.4 Operating branches: a fold is not a numerical convention

In the normalized scalar model

$$V^2 - V + \lambda = 0, \qquad V(0) = 1,$$

the operating condition selects the high branch

$$V_+(\lambda) = \frac{1 + \sqrt{1 - 4\lambda}}{2} \quad (0 \leq \lambda < 1/4).$$

It is a deliberately restricted load-flow-like algebraic model, not the full AC network equations, which generally involve conjugates and multiple real variables.

Theorem 3.6 (E06: Residuals do not select the operating branch)**.** *Both roots have zero residual. Their separation is* $\sqrt{1 - 4\lambda}$ *and*

$$\frac{dV_+}{d\lambda} = -\frac{1}{\sqrt{1 - 4\lambda}}.$$

For $\lambda \in [0, \Lambda]$ *with* $\Lambda < 1/4$*, the branch is Lipschitz with constant* $(1 - 4\Lambda)^{-1/2}$*. No such uniform constant persists up to the fold.*

Proof. Solve the quadratic and differentiate the selected root. The derivative is negative and its magnitude increases with λ. The real mean-value theorem gives the stated uniform bound. Equivalently implicit differentiation gives $(2V - 1)V' = -1$; its denominator is precisely the root separation on this branch. □

For an approximate v, factor the residual as $(v - V_+)(v - V_-)$. A small residual controls distance to one of the roots, not automatically to V_+. A branch certificate adds an isolating neighborhood connected to the initial condition along an admissible parameter path. For example $\lambda = 3/16$ gives $V_+ = 3/4$ and $V_- = 1/4$; the lower answer is algebraically exact and operationally wrong under the specified initial branch.

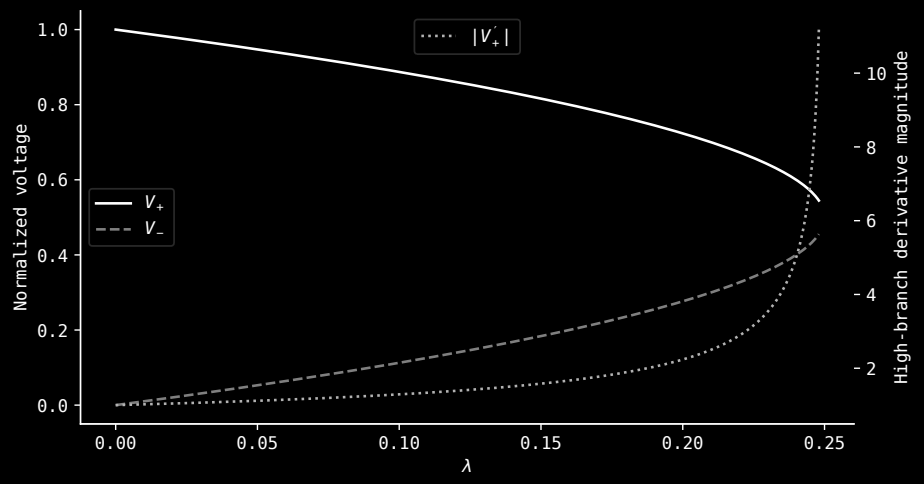


Figure 3.2: The two exact real solutions of the normalized fold model and the growth of high-branch sensitivity. The curve terminates before the branch point in numerical plots; the divergence follows from the displayed derivative, not from the picture.

3.5 Transmission lines and label-free propagation

For a uniform line with per-unit-length parameters $R, G \geq 0$ and $L, C > 0$, define

$$Z_\ell(s) = R + sL, \quad Y_\ell(s) = G + sC, \quad q(s) = Z_\ell(s)Y_\ell(s).$$

The propagation constant satisfies $\gamma^2 = q$. It has dimensions of inverse length; an expression $e^{-\gamma x}$ must have dimensionless exponent.

Theorem 3.7 (E07: Passive analytic branch in the right half-plane)**.** *On* $\operatorname{Re} s > 0$ *the factors have positive real part. There is an analytic square root* γ *of their product with* $\operatorname{Re}\gamma > 0$*, fixed by its positive value on positive real* s*. It is obtained by adding their analytic logarithms and halving:*

$$\gamma = \exp\big((\log Z_\ell + \log Y_\ell)/2\big).$$

Proof. Each factor is nonzero and its argument lies strictly between $-\pi/2$ and $\pi/2$. Their logarithms are analytic on this simply connected domain. The half sum of the arguments lies in the same interval, giving $\operatorname{Re}\gamma > 0$. Squaring yields q. Any other analytic square root differs by a constant sign on a connected domain; the positive-real normalization removes it. □

On the imaginary boundary, limiting cases may be purely imaginary or reach a zero. They require limits from the right half-plane, not arbitrary repeated principal-root calls on discontinuous data. A passive branch is a statement about an analytic model over a domain, not a conclusion from finitely many measurements with positive real parts.

Theorem 3.8 (E08: A propagation matrix can remain analytic when its eigenvalue labels branch)**.** *Let*

$$A(s) = \begin{pmatrix} 0 & Z_\ell(s) \\ Y_\ell(s) & 0 \end{pmatrix}, \qquad A(s)^2 = q(s)I.$$

For a fixed line length x,

$$e^{-xA} = c_x(q)I - s_x(q)A, \quad c_x(q) = \sum_{n\geq 0} \frac{x^{2n}q^n}{(2n)!}, \quad s_x(q) = \sum_{n\geq 0} \frac{x^{2n+1}q^n}{(2n+1)!}.$$

Both scalar coefficients are entire functions of q.

Proof. Separate the even and odd powers in the matrix exponential. Absolute convergence on bounded sets justifies this regrouping. Away from $q = 0$ the same expression reads $\cosh(x\sqrt{q})I - \sinh(x\sqrt{q})A/\sqrt{q}$. The series shows that the apparent quotient singularity at zero is removable, with $s_x(0) = x$. □

The sign of the off-diagonal entries depends on the chosen voltage/current direction and telegrapher convention; the formula here is for the equation $v_x = -Av$. A smooth propagator does not make individually labelled modes well conditioned at a collision. Conversely, divergent eigenvector conditioning need not make this basis-independent observable singular.

3.6 Energy, cascades and implemented state evolution

3.6.1 A continuous-time all-pass realization

Theorem 3.9 (E09: A state-storage certificate)**.** *For $a > 0$, take*

$$\dot{x} = -ax + \sqrt{2a}\,u, \qquad y = u - \sqrt{2a}\,x.$$

The transfer function under zero initial conditions is $(s-a)/(s+a)$ and

$$\frac{d}{dt}x^2 = u^2 - y^2.$$

For real signals on $[0,T]$, output energy plus final storage equals input energy plus initial storage.

Proof. Eliminating x in the Laplace domain gives $1 - 2a/(s+a)$. Differentiating x^2 gives $-2ax^2 + 2\sqrt{2a}xu$. Expanding $u^2 - (u - \sqrt{2a}x)^2$ gives the same polynomial. Integration proves the balance. For complex signals use absolute squares and real parts in the cross term. □

The pole is $-a$ and the zero is $+a$; the stable system is nonminimum phase. Setting $x(0) = 0$ does not make $x(T) = 0$ at an arbitrary truncation time. In a cascade the intermediate signal energies telescope and the total storage is the sum of the stage storages. This is a genuine functional-analytic passivity statement, distinct from source M24's theorem about composition of conformal domains, although both require a correctly typed intermediate object.

3.6.2 Stable stages do not imply stable switching

Theorem 3.10 (E10: A two-stage counterexample)**.** *The matrices*

$$A = \begin{pmatrix} 1/2 & 1 \\ 0 & 1/2 \end{pmatrix}, \quad B = \begin{pmatrix} 1/2 & 0 \\ 1 & 1/2 \end{pmatrix}$$

each have spectral radius $1/2$, but AB has spectral radius $(3+2\sqrt{2})/4 > 1$.

Proof. Their triangular forms give the individual spectra. Their product is $\left(\begin{smallmatrix} 5/4 & 1/2 \\ 1/2 & 1/4 \end{smallmatrix}\right)$, with trace $3/2$ and determinant $1/16$. The quadratic formula gives eigenvalues $(3 \pm 2\sqrt{2})/4$. Repetition of the same two-stage switching pattern therefore expands an eigenvector of the product. □

A norm contraction common to both stages would imply product contraction; separate spectral-radius statements do not supply that common norm. This is the implementation analogue of preserving all intermediate domain hypotheses before composing local theorems.

3.6.3 Quantization changes the equation

Theorem 3.11 (E11: A rounded-state invariant and a nonzero fixed point)**.** *Let* $k_{n+1} = \mathrm{round}(3k_n/4)$ *for integer states, with nearest rounding and ties away from zero. Then* $k = 1$ *is fixed, although the exact real recurrence tends to zero. Every step obeys*

$$|k_{n+1}| \le \tfrac{3}{4}|k_n| + \tfrac{1}{2},$$

and the integer set $\{-2, -1, 0, 1, 2\}$ *is invariant.*

Proof. The nearest-rounding error is at most $1/2$, giving the inequality. Direct rounding at $3/4$ gives one. For $|k| \le 2$, the inequality gives $|k'| \le 2$; integrality keeps the state in the displayed set. At $k = \pm 2$ the tie convention matters and has been specified. □

Theorem 3.12 (E12: An implemented error budget)**.** *If a model map contracts differences by* $q < 1$ *and each implemented step has discrepancy at most* η*, while all states remain in the declared domain, then*

$$e_n \le q^n e_0 + \eta \frac{1 - q^n}{1 - q}.$$

Proof. Use $e_{n+1} \le q e_n + \eta$ and induct. For the induction step, multiply the finite geometric sum by q and add the new η. The domain condition is needed at every application of the contraction hypothesis. □

For a quantization step Δ in physical units, $\eta = \Delta/2$ when the exact and implemented states are compared through this nearest-rounding model. The asymptotic bound is $\Delta/(2(1 - q))$. It is an error bound, not a guarantee that all trajectories converge to zero or that overflow and saturation never occur.

Chapter 4

Explicit circuits and finite implementations

4.1 EU1: an LC calculation with dimensions intact

For an ideal series circuit with $L = 2\,\mathrm{mH}$ and $C = 5\,\mu\mathrm{F}$,

$$Z(s) = Ls + \frac{1}{Cs}.$$

On $s = i\omega$, $Z = i(\omega L - 1/(\omega C))$. The resonance satisfies $\omega_0 = 1/\sqrt{LC} = 10^4\,\mathrm{s}^{-1}$. At $\omega = 5000\,\mathrm{s}^{-1}$ the impedance is $-30i\,\Omega$; at $20000\,\mathrm{s}^{-1}$ it is $30i\,\Omega$. The signs distinguish capacitive and inductive behavior under this convention.

For $s = \sigma + i\omega$ with $\sigma > 0$,

$$\operatorname{Re} Z(s) = L\sigma + \frac{\sigma}{C(\sigma^2 + \omega^2)} > 0.$$

This is a whole-domain positivity statement. The transformation $F(z) = iZ(-iz)$ maps the upper half-plane to itself and gives the rational structure in source Exercise 3.3(ii). The electrical calculation supplies the model and units; the complex geometry supplies an independent global constraint.

Exercise 4.1 (EU1a)**.** Why does a zero impedance at $s = i\omega_0$ not mean that the function has a pole there? Identify the poles of Z as a rational function on the sphere.

Solution. $Z(s) = (LCs^2 + 1)/(Cs)$ has zeros at $\pm i\omega_0$ and a finite pole at zero. It also has a pole at infinity because of the Ls term. A zero and a pole have opposite local meanings; neither should be called a "resonance singularity" without saying which response is being measured. The admittance $1/Z$ does have poles at the impedance zeros in this ideal model.

Exercise 4.2 (EU1b)**.** Calculate $Z'(s)$ and its positive-real zero. Does that point represent an unstable pole?

Solution. $Z'(s) = L - 1/(Cs^2)$ vanishes at $s = 1/\sqrt{LC}$ and its negative. At the positive point Z is finite and positive. This is a critical point of the impedance map, not a pole of a time-evolution transfer function. A branch or critical point in one representation must not be relabelled an unstable dynamical mode.

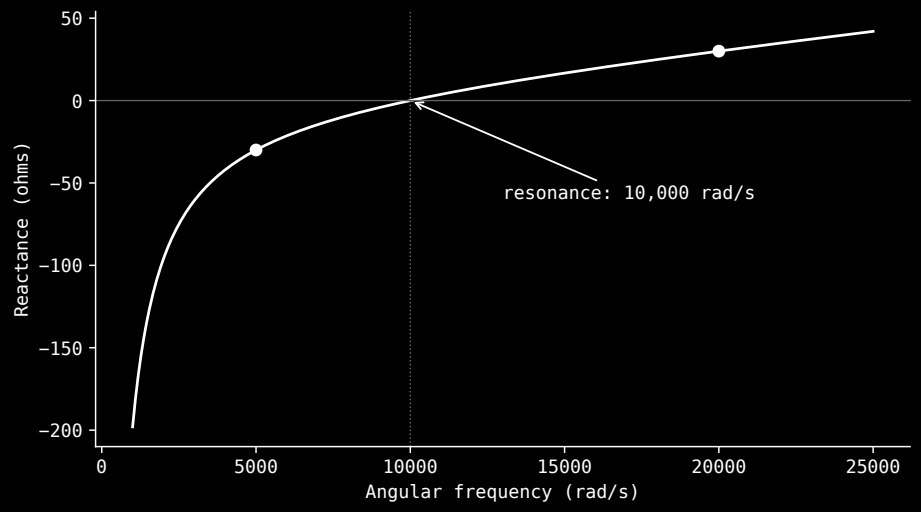


Figure 4.1: The ideal LC reactance with physical units fixed. Its zero at resonance is not a pole of the impedance. The corresponding admittance has different poles and therefore answers a different physical question.

4.2 EU2: a filter with constant magnitude and nonconstant delay

Take the discrete causal all-pass factor with $a = -1/3$:

$$H(z) = \frac{z^{-1} + 1/3}{1 + z^{-1}/3}.$$

Its pole is $-1/3$, inside the unit circle. The impulse response is $h[0] = 1/3$ and $h[k] = (8/9)(-1/3)^{k-1}$ for $k \geq 1$. The alternating sign does not prevent absolute summability.

The group delay is

$$\tau(\omega) = \frac{1 - a^2}{1 - 2a\cos\omega + a^2}.$$

Here $\tau(0) = 1/2$ sample and $\tau(\pi) = 2$ samples. The delay is concentrated toward the high-frequency end instead of the low-frequency end because the real pole is negative. The complete delay integral remains 2π for one factor. Thus equal magnitude responses do not imply equal time-domain pulse shapes.

Exercise 4.3 (EU2a)**.** Find the energy beyond sample three. Explain why a measurement ending at sample three can report output energy less than input energy without violating all-pass energy preservation.

Solution. The exact tail is $(1 - a^2)a^6 = (8/9)/729 = 8/6561$. An all-pass theorem concerns the complete response, including its infinite tail. The truncated record omits that positive energy. A finite-window comparison needs a tail bound or an explicit storage balance.

4.3 EU3: a cubic stability boundary derived from an imaginary root

Consider $p_K(s) = s^3 + 4s^2 + 5s + 2 + K$. At $K = 0$ it factors as $(s + 1)^2(s + 2)$, so all roots are stable. A root can change half-plane only by crossing the imaginary axis as K varies continuously. Substitute $s = i\omega$:

$$2 + K - 4\omega^2 = 0, \qquad \omega(5 - \omega^2) = 0.$$

The possible boundary gains are $K = -2$ at $\omega = 0$ and $K = 18$ at $\omega = \pm\sqrt{5}$. There is no imaginary-axis crossing between these values. Continuity of polynomial roots therefore preserves the stable count throughout $-2 < K < 18$.

For completeness, the Routh first column is $1, 4, (18 - K)/4, 2 + K$, giving exactly the same strict stability interval. The argument with imaginary substitution identifies the mechanism; the Routh criterion supplies a compact algebraic count for the whole real gain line.

Exercise 4.4 (EU3a)**.** At $K = 18$, factor the polynomial. At $K = 20$, determine the number of right-half-plane roots using the first-column sign changes.

Solution. At eighteen, $p = (s + 4)(s^2 + 5)$. At twenty, the first-column signs are $+, +, -, +$, so there are two sign changes and two right-half-plane roots. The crossing pair is not an arithmetic rounding error. A contour passing through those roots at the endpoint gain cannot be used for a strict argument-principle certificate.

Exercise 4.5 (EU3b)**.** A transfer function reduces to $1/(s + 2)$, but its state realization also contains an unobserved mode $\dot{x}_1 = x_1$. What stability statement does the reduced transfer function fail to establish?

Solution. It establishes no decay of arbitrary initial conditions in the hidden unstable mode. A pole count of the reduced input-output response omits that internal state. Internal stability must use the actual state matrix or an equivalent minimality and realization argument, not merely the simplified transfer denominator.

4.4 EU4: a square root can simplify when zeros have even order

In a normalized line model choose $Z_\ell = s + 2$ and $Y_\ell = 4(s + 2)$. Then

$$\gamma^2 = 4(s + 2)^2, \qquad \gamma = 2(s + 2)$$

on the passive right-half-plane branch, with characteristic impedance $Z_c = 1/2$. Although the generic equation involves a square root, this particular product has a double zero at $s = -2$. The selected square root is an entire linear function, not a genuinely two-valued branched function of s around that point.

Exercise 4.6 (EU4a)**.** Contrast this with $\gamma^2 = s + 2$. Follow a root once around a small circle centered at $s = -2$ in each model.

Solution. For $s + 2 = \rho e^{it}$, the first model admits $\gamma = 2\rho e^{it}$ and returns after one loop. The second gives $\gamma = \sqrt{\rho} e^{it/2}$ and changes sign. The order of the zero in the radicand determines whether the pullback of the square-root cover is branched. The mere presence of a square-root symbol in code does not settle the geometry.

4.5 EU5: sampling loses a logarithmic label

A continuous mode e^{st} sampled every T seconds becomes the discrete multiplier $z = e^{sT}$. For every integer k,

$$e^{(s+2\pi ik/T)T} = e^{sT}.$$

Thus the sampled multiplier does not determine a unique continuous frequency. With $T = 10^{-3}$ seconds, imaginary exponents $200\pi i$ and $2200\pi i$ produce identical sampled multipliers, corresponding to frequencies 100 and 1100 Hz in this constructed example.

The exponential map is an infinite covering of $\mathbb{C}^*$ and is not proper. The finite degree and finite ramification formulas of source section 3 must not be applied unchanged. The common lesson is the need for a label or domain restriction; the two mathematical map classes remain distinct.

Exercise 4.7 (EU5a)**.** What additional condition can select one logarithm of a sampled multiplier? Why is that condition model information instead of a consequence of the samples?

Solution. A prior frequency band of width less than $2\pi/T$, with an explicitly chosen boundary convention, can select one branch of the logarithm. The samples identify only the exponential value and cannot exclude the other frequencies on their own. Antialias filtering or a physical bandwidth assumption supplies the missing information.

4.6 EU6: a matrix response can avoid a mode label

For $A(\lambda) = \left(\begin{smallmatrix} 0 & 1 \\ \lambda & 0 \end{smallmatrix}\right)$, $A^2 = \lambda I$. The exponential is

$$e^{tA} = \sum_{n\geq 0} \frac{t^{2n}\lambda^n}{(2n)!} I + \sum_{n\geq 0} \frac{t^{2n+1}\lambda^n}{(2n+1)!} A.$$

This formula has no ambiguity at $\lambda = 0$, where it equals $I + tA(0)$. The eigenvalues $\pm\sqrt{\lambda}$ do have a branch-label issue. A model output computed from the whole propagator can therefore be smooth even when individually sorted modes jump. This is an exact algebraic calculation relevant to transmission and state models, not an assertion of physical adiabatic mode following.

Exercise 4.8 (EU6a)**.** Compute the exponential at $\lambda = 0$ directly from the nilpotent matrix and explain why diagonalization is unnecessary.

Solution. $A(0)^2 = 0$, so the exponential series stops at $I + tA(0) = \left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right)$. The matrix is not diagonalizable, but its exponential is fully determined. A failed eigenvector calculation need not imply a failed physical response model.

4.7 EU7: rounded implementation and a measurable error floor

Let a stored integer state evolve by $k' = \mathrm{round}(2k/3)$, using nearest rounding. The exact real recurrence tends to zero. The stored state $k = 1$ is fixed because $2/3$ rounds to one. Each step differs from the exact update by at most one half, giving

$$|k_n| \leq (2/3)^n |k_0| + \tfrac{3}{2}(1 - (2/3)^n).$$

This is a bound, not an assertion that every trajectory approaches the same nonzero value. Saturation, overflow and alternative tie conventions are additional implementation details that can change behavior.

Exercise 4.9 (EU7a)**.** For a physical quantization step $\Delta = 0.01$, derive the worst-case asymptotic deviation bound under $q = 2/3$ and nearest rounding. What further evidence is needed to claim a hardware implementation satisfies it?

Solution. The per-step discrepancy is at most $\Delta/2 = 0.005$, so the geometric error bound tends to $0.005/(1 - 2/3) = 0.015$. Hardware must implement the declared recurrence, rounding, range and timing without unmodelled saturation or overflow. A pole theorem for the exact recurrence does not establish these facts.

4.8 A six-part laboratory record

4.9 Undergraduate comparison laboratory: a model is not its implementation

4.9.1 EU8: constant magnitude does not mean constant delay

For a first-order all-pass factor with $a = 4/5$, the exact magnitude is one on the unit circle. At zero frequency,

$$\tau(0) = \frac{1+a}{1-a} = 9,$$

while at the Nyquist endpoint,

$$\tau(\pi) = \frac{1-a}{1+a} = \frac{1}{9}.$$

The delay is measured in samples with the stated discrete-time convention. An amplitude-only plot cannot distinguish this system from a pure sign change of gain one, even though their phases differ.

Conventional calculation. Sample the response, unwrap the phase and estimate its slope. The proposed calculation differentiates the known rational factor analytically. The metric of improvement is removal of finite-difference truncation error for that model. It is not a claim to remove uncertainty from measured data or to infer phase uniquely from magnitude.

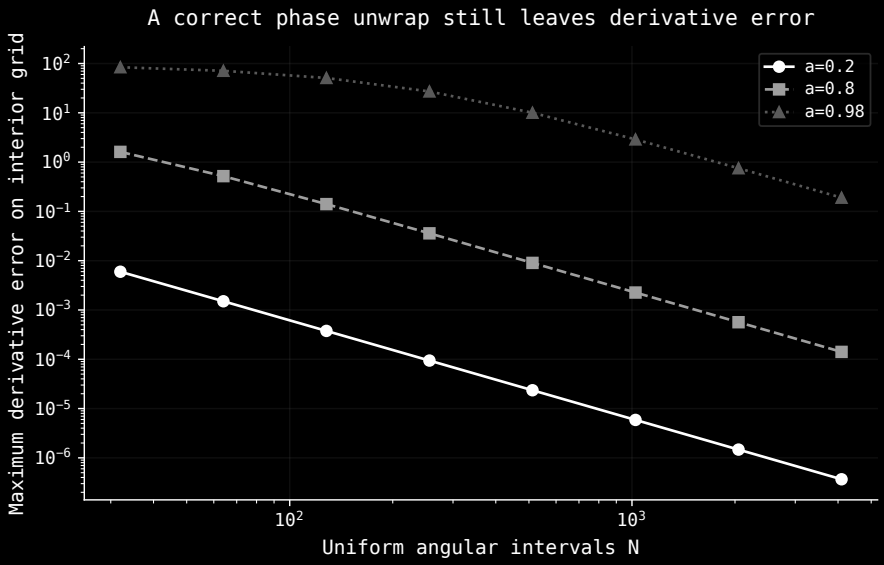


Figure 4.2: Finite-difference error after a correct unwrap. Near-boundary parameters have sharper phase variation and require greater resolution. The plotted errors are measured at grid points.

Exercise 4.10 (EU8a)**.** For $a = 4/5$, use E13 to give a sufficient uniform grid size for exact unwrap increments smaller than π. Does that grid guarantee one-percent delay accuracy?

Solution. The sufficient condition is $N > 2 \cdot 9 = 18$, so $N = 19$ intervals is sufficient for the exact phase-increment claim. It does not bound the error of a finite-difference derivative. One-percent delay accuracy needs a separate truncation and arithmetic analysis or comparison against a validated exact-model evaluation.

4.9.2 EU9: an impulse-tail bound is a separate finite-implementation choice

Suppose the ideal all-pass response is implemented by keeping only its first $N + 1$ impulse coefficients. The exact omitted impulse energy is $(1-a^2)a^{2N}$. At $a = 1/2$, this becomes $(3/4)4^{-N}$. To make it at most 10^{-4}, $N = 7$ is sufficient because $(3/4)/4^7 < 10^{-4}$, whereas $N = 6$ gives $(3/4)/4096 > 10^{-4}$.

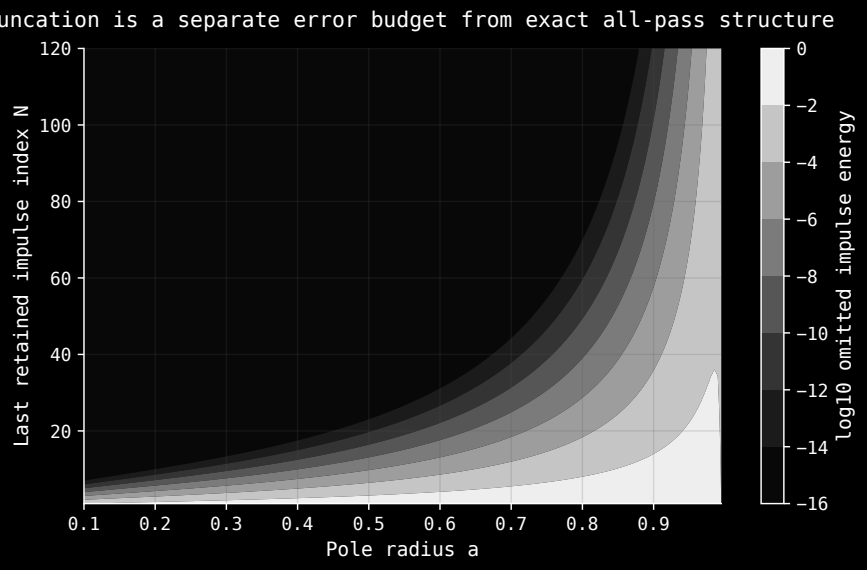


Figure 4.3: The tail energy as a function of radius and retained length. Its interpretation depends on the impulse norm; it is not automatically a worst-case gain error for every input.

Stronger output contract. For a general input, the ℓ^1 tail bound is $(1 + |a|)|a|^N$. At $a = 1/2$ and $N = 7$, this is $3/256$, about 0.0117. That is much larger than the square root of the impulse-tail energy in a different role. The system designer must decide whether the target is one impulse waveform or a gain guarantee over all allowed inputs.

Exercise 4.11 (EU9a)**.** Find the least N for which $(3/2)2^{-N} \leq 1/1000$. Compare with the impulse-energy target above.

Solution. The inequality is $2^N \geq 1500$, so the least integer is $N = 11$. The two problems have different targets and different norms; their retained lengths need not agree. A report that states only "error below 10^{-3}" is incomplete without defining the error quantity.

4.9.3 EU10: a sampled pole forgets frequency labels

Let $T > 0$ be a sampling interval and $s = \sigma + i\omega$ a continuous-time pole. Its sampled pole is $p = e^{sT}$. The magnitude $|p| = e^{\sigma T}$ retains the decay rate, but the phase retains ωT only modulo 2π.

An exact example. At $T = 1/1000$ second, frequencies 100 Hz and 1100 Hz give the same sampled phase because their angular-frequency difference times T is 2π. A numerical logarithm returning one principal value is not identifying the physical frequency without a prior band or additional measurement. The correct API reports a sheet convention or an admissible family $s_k = (\operatorname{Log} p + 2\pi ik)/T$.

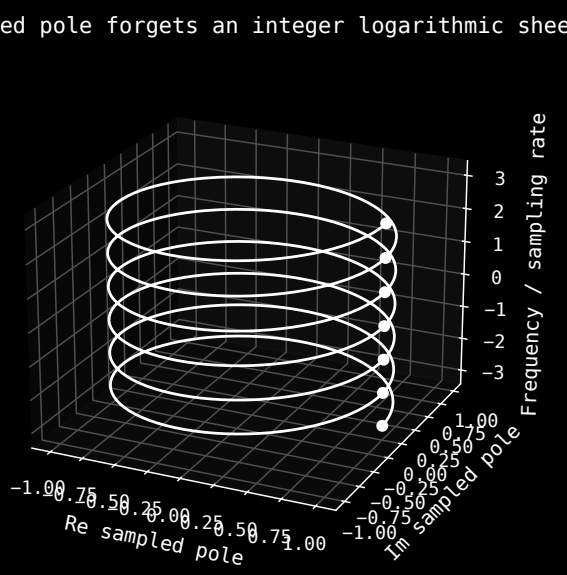


Figure 4.4: An unwrapped frequency coordinate over the sampled unit circle. Repeated vertical alignments are the logarithmic sheets lost by sampling.

Comparison metric. The proposed representation preserves the ambiguity explicitly, avoiding a falsely unique answer. It does not generate information lost by sampling. Anti-alias filtering, a sampling-rate change or a physical frequency prior belongs to the measurement model, not to a more elaborate complex logarithm routine.

Exercise 4.12 (EU10a)**.** What condition on an allowed frequency interval makes one logarithmic sheet sufficient? Why is an endpoint convention still needed?

Solution. An interval of angular-frequency width less than $2\pi/T$ can contain at most one representative of each sampled phase. At a width exactly $2\pi/T$, both endpoints correspond to the same phase, so one endpoint must be excluded or another convention supplied. This is a uniqueness condition, not a guarantee that noise will permit stable estimation near the interval boundary.

4.9.4 EU11: use the simplest complete stability check

For $s^3+6s^2+11s+6+K$, coefficient positivity and the cubic Routh condition give $-6 < K < 60$. In a gain design with $K \geq 0$, strict stability is $0 \leq K < 60$. At $K = 60$, the factorization $(s+6)(s^2+11)$ gives imaginary roots, so strict stability is lost.

Baseline retained. A root-locus plot is explanatory, and a contour certificate is a useful independent checker. Neither is simpler than these two scalar inequalities for this exact cubic. The lecture therefore includes a "retain the baseline" case instead of claiming advanced complex analysis should replace every elementary solution.

Where the branch geometry helps. If the characteristic relation is transcendental, has uncertain coefficients or requires tracking a physical branch, a global argument-principle or continuation certificate may answer a harder question. The prerequisite is the correct transfer or state model with no hidden unstable cancellation. The simple polynomial example does not certify those more complicated systems automatically.

Chapter 5

Energy, cascades and switched models

5.1 EG1: a lossless cascade has an additive storage function

For stage j, let

$$\dot{x}_j = -a_j x_j + \sqrt{2a_j}\, u_j, \qquad y_j = u_j - \sqrt{2a_j}\, x_j, \qquad a_j > 0.$$

Its storage $V_j = x_j^2$ satisfies $\dot{V}_j = u_j^2 - y_j^2$. Connect $u_{j+1} = y_j$. Summing over $j = 1, \dots, d$ cancels the internal signal energies and gives

$$\frac{d}{dt} \sum_j x_j^2 = u_1^2 - y_d^2.$$

The total transfer is $\prod_j (s - a_j)/(s + a_j)$. Thus a scalar all-pass factorization corresponds to a finite energy-conserving cascade with stable internal poles, under the declared state equations. The storage proof does not require a frequency sample grid.

Exercise 5.1 (EG1a)**.** For two stages $a_1 = 1, a_2 = 3$, derive the transfer numerator and denominator. Identify the zeros and poles, and explain why stable all-pass does not mean minimum phase.

Solution. The transfer is $(s^2 - 4s + 3)/(s^2 + 4s + 3)$. Its poles are $-1, -3$ and its zeros are $1, 3$. The poles are stable, but a causal stable inverse would have poles at the original right-half-plane zeros. Hence the system is not minimum phase. The whole-cascade energy balance coexists with this inverse obstruction.

5.2 EG2: a paired-pole rational map gives a Foster-type network

Consider the normalized impedance

$$Z(s) = \tfrac{1}{2} s + \frac{2s}{s^2 + 1} + \frac{6s}{s^2 + 9} + \frac{1}{2s}.$$

It is a series connection of an inductor, two parallel LC branches and a capacitor. The branch formula $sL/(1 + s^2 LC)$ identifies the first parallel branch with $L = 2, C = 1/2$, and the second with $L = 2/3, C = 1/6$. The final capacitor has $C = 2$ in these normalized units.

For $\operatorname{Re} s > 0$, each ideal LC impedance is positive real, and a series sum preserves nonnegative real part. More explicitly, for $s = \sigma + i\omega$,

$$\operatorname{Re} \frac{s}{s^2 + \Omega^2} = \frac{\sigma(\sigma^2 + \omega^2 + \Omega^2)}{|s^2 + \Omega^2|^2} > 0.$$

The positivity follows by multiplying numerator and denominator by the conjugate. Positive coefficients therefore give an analytic positive-real function on the open right half-plane. Poles on its boundary are permitted for this lossless idealization.

Exercise 5.2 (EG2a)**.** A fitted rational model changes the coefficient of $2s/(s^2+1)$ to -2. Why is it not justified to declare the modified model passive solely because its poles remain on the imaginary axis?

Solution. Pole locations do not enforce positive real part. The negative residue contributes negative real part near that resonance when approached from the right half-plane; for sufficiently small positive σ at $s = \sigma + i$, that divergent negative contribution dominates the bounded other terms. The original coefficient sign is a global passivity constraint, not a cosmetic fitting parameter.

5.3 EG3: robust cubic stability over a coefficient box

Let $p(s) = s^3 + as^2 + bs + c$ with independent real intervals

$$a \in [5, 6], \qquad b \in [8, 9], \qquad c \in [20, 30].$$

The cubic Hurwitz conditions are $a > 0$, $b > 0$, $c > 0$ and $ab > c$. On this box, $ab - c \geq 5 \cdot 8 - 30 = 10 > 0$, so every polynomial is strictly stable. This is a continuum statement obtained by interval monotonicity, not by testing only a few random draws.

A coefficient box with $c \in [20, 45]$ is not uniformly stable: the admissible corner $(a, b, c) = (5, 8, 45)$ has $ab - c = -5$. However, failure of the crude bound for a correlated parameter model would not necessarily imply a real unstable member. The uncertainty set must reflect the actual dependencies.

Exercise 5.3 (EG3a)**.** Suppose $a = 5 + t$, $b = 8 + t$ and $c = 40 + 10t$ for $0 \leq t \leq 1$. Prove or disprove strict stability throughout the interval, and explain why an independent box can be misleading.

Solution. $ab - c = (5 + t)(8 + t) - (40 + 10t) = 3t + t^2$. At $t = 0$ it vanishes, giving a boundary case, while it is positive for $t > 0$. All coefficients are positive, so the family is strictly stable only on $(0, 1]$. An independent box would permit combinations not realized by the common parameter and could obscure both the exact endpoint and the valid interior. The dependency is part of the model.

5.4 EG4: a transmission matrix without square-root cancellation

Let $A(s) = \left(\begin{smallmatrix} 0 & Z(s) \\ Y(s) & 0 \end{smallmatrix}\right)$ and $q = ZY$. For length ℓ,

$$T(\ell, s) = e^{-\ell A} = c_\ell(q) I - s_\ell(q) A,$$

where $c_\ell(q) = \sum_{n \geq 0} \ell^{2n} q^n / (2n)!$ and $s_\ell(q) = \sum_{n \geq 0} \ell^{2n+1} q^n / (2n+1)!$. Its determinant is one because $\operatorname{tr} A = 0$; direct multiplication also gives $c_\ell(q)^2 - q s_\ell(q)^2 = 1$.

The identity holds at $q = 0$ by the entire series. Computing $\sinh(\ell\sqrt{q})/\sqrt{q}$ directly can lose accuracy near zero even though the function is regular. A numerical routine should switch to its series or a stable entire-function evaluation, with an error bound in the declared parameter range.

Exercise 5.4 (EG4a)**.** For $|\ell^2 q| \leq 1/4$, bound the omitted terms of c_ℓ after its q^N term by a geometric majorant.

Solution. The first omitted magnitude is at most $r^{N+1}/(2N+2)!$, where $r = |\ell^2 q| \le 1/4$. Each later term divided by its predecessor is $r/((2n+2)(2n+1)) \le r/2 \le 1/8$. Therefore the tail is at most $\frac{8}{7} r^{N+1}/(2N+2)!$. This bound is conservative but uniform. It controls the scalar coefficient; the matrix error also includes the norm of A and the separate s_ℓ tail.

5.5 EG5: annular energy under a degree-d cover

Let $0 < r < 1$ and map the source annulus $A(r, 1)$ to $A(r^d, 1)$ by $w = z^d$. The map is unbranched on the annulus and has degree d. On the target, the harmonic potential with inner boundary one and outer boundary zero is

$$u(w) = \frac{\log |w|}{\log(r^d)}.$$

Its pullback is $v(z) = \log |z| / \log r$. Direct polar integration gives

$$\int_{A(r^d,1)} |\nabla u|^2 dA = \frac{2\pi}{d \log(1/r)}, \qquad \int_{A(r,1)} |\nabla v|^2 dA = \frac{2\pi}{\log(1/r)}.$$

The source energy is d times the target energy, in agreement with the general pullback-energy theorem. The computation simultaneously checks boundary values, topology and the Jacobian multiplicity.

In an ideal two-dimensional conducting sheet of constant conductivity and fixed thickness, the Dirichlet energy times the material factor determines dissipated power for the given voltage normalization. Mapping changes the geometry; the factor is not a universal claim that any physical capacitor or three-dimensional field scales by degree. Time-dependent heat equations also acquire metric factors instead of following the same steady harmonic rule.

Exercise 5.5 (EG5a)**.** Explain why a single-valued harmonic conjugate of v cannot exist on the source annulus.

Solution. Locally a conjugate is $\arg z / \log r$. Going once around the annulus changes it by $2\pi / \log r \neq 0$. A globally single-valued conjugate would have zero period on every loop. The logarithmic term is therefore a genuine topology-dependent component, not a poor choice of polynomial basis.

5.6 EG6: a periodic system is governed by its product

For a discrete two-stage period $x_{n+1} = A_j x_n$ with stages repeating, the period map is $M = A_2 A_1$. The eigenvalues of M, not the separate stage spectra, determine asymptotic behavior at period boundaries. The application volume gives an exact pair of stable triangular stages whose product has an eigenvalue greater than one.

A sufficient condition that does compose is a common positive-definite quadratic storage matrix P with $A_j^T P A_j \preceq q_j^2 P$ for each stage. Then

$$\|A_2 A_1 x\|_P \le q_2 q_1 \|x\|_P .$$

The proof applies the second inequality to $A_1 x$ and then the first. Using a different P_j at each stage requires explicit comparison constants; they cannot be omitted when multiplying gains.

Exercise 5.6 (EG6a)**.** If $P_1 \preceq \kappa P_2$ and $P_2 \preceq \kappa P_1$, explain how changing the norm between stages can introduce an extra gain factor.

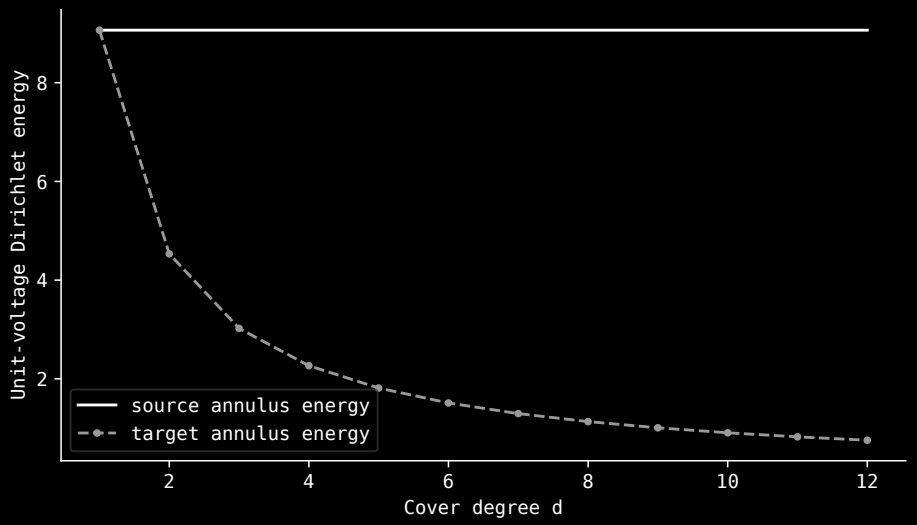


Figure 5.1: Exact Dirichlet energies for the annular covering family with source inner radius one half. The source and target domains differ with degree; their energy ratio is d. This does not assert a universal material-independent scaling law for devices.

Solution. The inequalities imply $\|x\|_{P_1} \le \sqrt{\kappa}\,\|x\|_{P_2}$ and the reverse. Every switch of the norm used in a bound may contribute $\sqrt{\kappa}$. A product of separately contracting estimates can therefore fail to contract once all norm changes are accounted for. The metric is part of the theorem, just as a domain is part of an inverse branch.

5.7 EG7: a two-layer certificate for quantized state evolution

Suppose an exact stable recurrence has a proven contraction in a norm, and the implemented update differs by a bounded rounding term. The first layer certifies the mathematical model; the second certifies an invariant state set and an arithmetic-error bound. The final recurrence estimate is valid only on that invariant set, where overflow and saturation assumptions have been checked.

Exercise 5.7 (EG7a)**.** For a scalar map with $q = 4/5$ and error at most $\Delta/2$ per update, show that the interval $[-R, R]$ is invariant whenever $R \ge 5\Delta/2$. Does this imply convergence to zero?

Solution. For $|x| \le R$, $|x'| \le (4/5)R + \Delta/2 \le R$ exactly under the displayed condition. This proves invariance. It does not prove zero convergence, because a bounded nonvanishing rounding disturbance can sustain a fixed point or cycle. To establish a finer property one must examine the finite quantized-state transition graph or prove an additional dissipativity condition.

5.8 Graduate submission standard

5.9 Graduate comparison: realization, whole-domain bounds and complexity

The graduate electrical branch compares methods only after fixing a physical or mathematical output: stored energy, phase increment, robust pole count, flux, or a finite implementation bound. Circuit identities and state-space semantics remain primary. The analytic method is inserted where it produces a specific additional guarantee.

5.9.1 EG8: a passive representation carries its own sign constraints

Consider

$$Z(s) = s + \frac{1}{s} + \frac{2s}{s^2+9}, \qquad \operatorname{Re} s > 0.$$

The paired term equals $1/(s-3i) + 1/(s+3i)$. Each reciprocal has positive real part in the right half-plane, and s does as well. Thus $\operatorname{Re} Z(s) > 0$ there. The poles lie on the boundary, and the terms have a lossless network interpretation under the normalization in the parent volume.

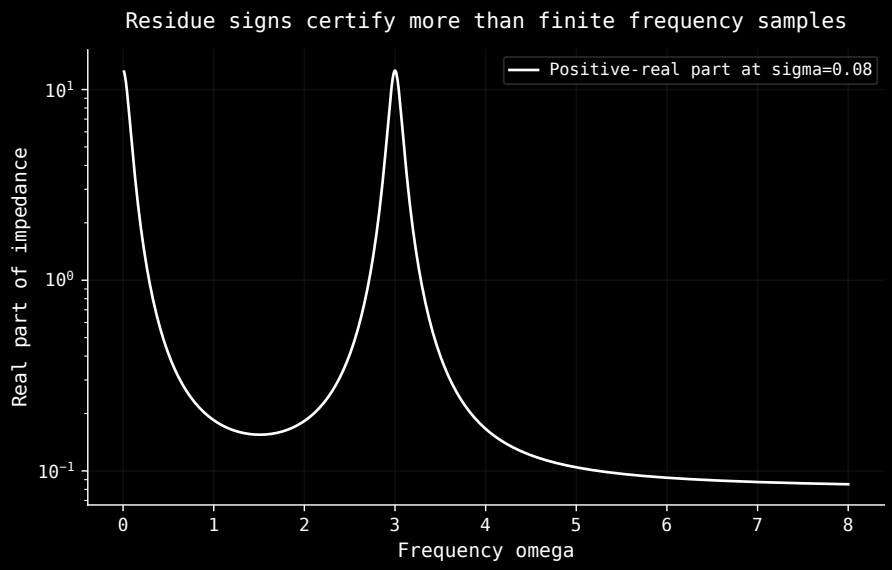


Figure 5.2: One vertical frequency slice of a positive-real function. The proof is the sign-constrained rational representation; finite sample positivity alone would not establish the full half-plane statement.

Conventional fit versus proposed fit. An unconstrained rational fit may match a finite set of impedances but introduce a right-half-plane pole or negative-real region between samples. A structured fit with admissible pole/residue data restricts the search to a class with a whole-domain property. That can improve feasibility reliability but may worsen the least-squares fit to noisy or model-inconsistent data. If the data cannot be represented by the chosen lossless family, forcing the model is not validation.

Exercise 5.8 (EG8a)**.** Change the coefficient of $1/s$ to -1. Show that the resulting function is not positive real by examining positive real s tending to zero.

Solution. On $s = x > 0$, the function is $x - 1/x + 2x/(x^2+9)$. The negative term diverges as $x \downarrow 0$ while the other two vanish, so the real part is negative for sufficiently small x. A residue-sign constraint has a genuine global role; it is not merely a stylistic parameterization choice.

5.9.2 EG9: a phase bound is not a group-delay accuracy bound

For a cascade of d real all-pass factors with $|a_j| \le \rho$, the delay satisfies

$$0 < \tau(\omega) \le d\frac{1+\rho}{1-\rho}.$$

Consequently $N > 2d(1+\rho)/(1-\rho)$ is a sufficient uniform phase-increment rule. A locally adaptive grid can use a tighter integrated or local derivative bound and spend fewer samples away from the narrow ridge. However, the derivative approximation still requires control of higher derivatives and roundoff.

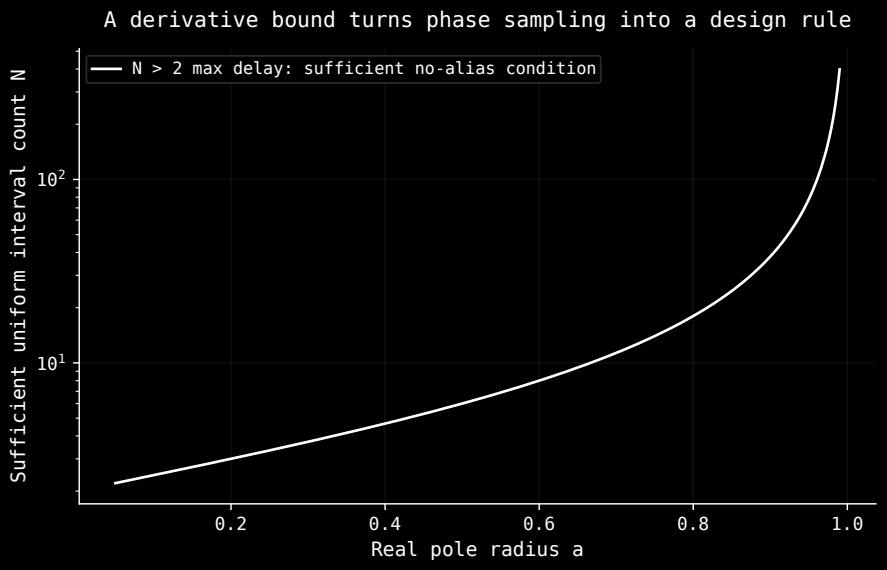


Figure 5.3: A sufficient no-alias phase-grid condition. The rapid growth is a consequence of allowing poles arbitrarily close to the unit circle, not evidence that all practical filters need such a grid.

Mathematically fair comparison. The baseline is a correctly unwrapped phase with a known resolution rule, not an unqualified phase plot. The proposed exact derivative removes stencil error only for an exactly specified factorization. With measured frequency samples, the factorization itself is an inference; any claimed gain must include identification uncertainty.

Exercise 5.9 (EG9a)**.** For $d = 3$ and $\rho = 0.9$, give the sufficient interval count. How does this compare to the one-factor bound?

Solution. The bound is $N > 2 \cdot 3 \cdot 19 = 114$, so $N = 115$ suffices for the exact-increment conclusion. It is three times the one-factor threshold before integer rounding. This worst-case bound allows all factor peaks to align; separated peak locations can permit a tighter model-specific bound.

5.9.3 EG10: a conformal energy identity is sensitive to normalization

For the annulus $a < r < 1$, the unit-voltage harmonic potential is $u = \log(r/a)/\log(1/a)$. Its radial gradient is $1/(r \log(1/a))$ and its Dirichlet energy is

$$\int_0^{2\pi} \int_a^1 |u_r|^2 r \, dr \, d\theta = \frac{2\pi}{\log(1/a)}.$$

This is an exact baseline calculation. A covering $z \mapsto z^d$ changes both the target annulus and the pulled-back function. The unnormalized potential plotted below has gradient multiplied by d; imposing equal boundary voltage requires dividing by d. Any comparison of physical energy must say which constraint is held fixed.

Advantage boundary. For this radial example, direct integration is simpler than a general conformal solver. The covering/energy theorem is useful when reusing an already known solution across a family of domains, or when checking an independently computed field. It is not a reason to replace the direct two-line radial calculation by a costly numerical map.

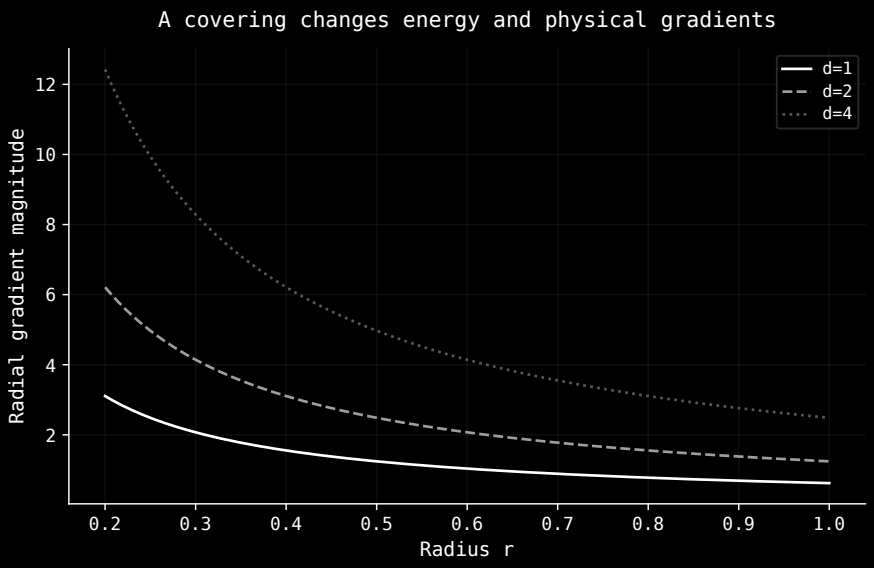


Figure 5.4: Radial gradients before voltage renormalization. A degree label alone cannot determine a same-voltage engineering comparison.

5.9.4 EG11: a common storage function certifies arbitrary switching

The original branch supplies a counterexample in which stable stages compose unstably. A positive result needs a stronger joint hypothesis. If each discrete-time stage A_i satisfies $A_i^* P A_i \preceq q^2 P$ for the same Hermitian $P \succ 0$ and some $q < 1$, then any switching product obeys

$$\|A_{i_n} \cdots A_{i_1} x\|_P \leq q^n \|x\|_P.$$

The proof repeatedly applies the same quadratic inequality. Separate matrices P_i do not permit the same induction without comparing the norms at switches.

Comparison. Checking each spectral radius is cheaper but does not answer arbitrary-switching stability. A common storage certificate is a stronger guarantee with more setup. This is an example where a method can be worth using without being faster. The state equation, switching admissibility and implementation error must still match the analyzed model.

Exercise 5.10 (EG11a)**.** If the norm changes at each switch with $\|x\|_{P_{i+1}} \leq c\|x\|_{P_i}$, what simple bound results from combining a per-stage contraction q with a switch every step?

Solution. Each step costs at most the factor cq, giving $(cq)^n$ in the resulting estimate. A sufficient contraction condition is $cq < 1$. This is a conservative estimate; it illustrates precisely why individually contracting systems in incompatible norms may not produce a useful switched-system bound.

5.9.5 EG12: implementation accuracy is part of the realization

A transfer-function certificate has no knowledge of overflow, saturating arithmetic, sensor delay or asynchronous updates unless those are modeled. The correct finite implementation analysis starts from a transition relation, not a picture of poles. The numerical recurrence can be bounded by an ideal contraction plus a quantified disturbance, while exact finite-state traversal identifies recurrent classes in a bounded invariant set. The two analyses complement instead of replace one another.

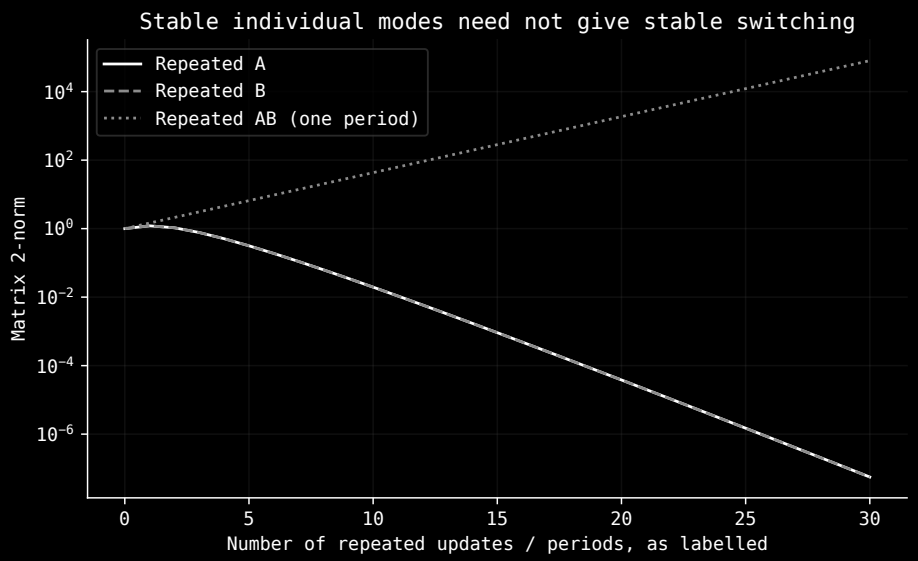


Figure 5.5: Norms of powers of the two matrices in E10 and their period map. Each factor has spectral radius one half, while the period map has spectral radius greater than one. Update count and period count are deliberately labelled separately.

Chapter 6

Robust response and implementation guarantees

6.1 EP1: uniform delay bounds over an all-pass parameter set

Let B be a degree-d finite Blaschke product with zeros satisfying $|a_j| \le \rho < 1$. Its boundary phase derivative is

$$\tau(\theta) = \sum_{j=1}^{d} \frac{1-|a_j|^2}{|e^{i\theta}-a_j|^2}.$$

For each summand, the triangle inequalities give

$$\frac{1-|a_j|}{1+|a_j|} \le \frac{1-|a_j|^2}{|e^{i\theta}-a_j|^2} \le \frac{1+|a_j|}{1-|a_j|}.$$

Consequently

$$d\frac{1-\rho}{1+\rho} \le \tau(\theta) \le d\frac{1+\rho}{1-\rho}, \qquad \int_0^{2\pi} \tau \, d\theta = 2\pi d.$$

This is a parameter-uniform phase budget. The upper bound diverges as zeros approach the unit circle; a bounded finite-degree family without a radius margin need not have a uniform peak delay.

Exercise 6.1 (EP1a)**.** For $d = 4$ and $\rho = 3/5$, give the lower and upper delay bounds. Why is the mean delay still four, instead of the midpoint of the two bounds?

Solution. The bounds are one and sixteen. The mean follows from the exact integral and is four. An upper/lower envelope does not specify the distribution of values or their mean; the topological winding gives additional global information.

6.1.1 The excluded degree-changing limit

As a zero reaches the boundary, a factor can cancel to a constant. Its phase derivative can concentrate in a narrow arc even while the rational expression tends locally to a constant away from that arc. The strong-operator calculation in the mathematical postgraduate branch makes the same limit precise: strong convergence need not preserve Fredholm index, while operator-norm convergence would impose a different constraint. This gives a functional-analytic explanation of why a parameter fit near the boundary can be sensitive despite an apparently simple limiting formula.

6.2 EP2: finite data define a feasible analytic class

Let $F : \mathbb{D} \to \overline{\mathbb{D}}$ be holomorphic. For interpolation sites z_j and values $w_j = F(z_j)$, the matrix

$$P_{jk} = \frac{1 - w_j\overline{w_k}}{1 - z_j\overline{z_k}}$$

is positive semidefinite. A functional-analytic proof uses multiplication by F as a contraction on H^2. On reproducing kernels, $M_F^* k_z = \overline{F(z)} k_z$, so the Gram matrix of the positive operator $I - M_F M_F^*$ on these kernels is exactly P. This proves necessity. General sufficiency is the Nevanlinna–Pick theorem, an additional theorem; it is not implied by a finite matrix calculation alone.

For two sites, a direct construction avoids invoking the full theorem. Normalize one site and its value to zero by disk automorphisms. Schwarz's lemma then gives

$$\left|\frac{w_2 - w_1}{1 - \overline{w_1} w_2}\right| \le \left|\frac{z_2 - z_1}{1 - \overline{z_1} z_2}\right|.$$

When the inequality holds, a linear contraction between the normalized points, conjugated back by the automorphisms, supplies an interpolant. This yields the exact two-point feasibility criterion and makes its geometric meaning explicit.

Exercise 6.2 (EP2a)**.** At sites $z_1 = 0, z_2 = 1/2$, are the real values $w_1 = 1/5, w_2 = 4/5$ feasible for a Schur function? Repeat with $w_2 = 3/5$ and construct one feasible function.

Solution. In the first case the output pseudohyperbolic distance is $(3/5)/(21/25) = 5/7 > 1/2$, so no such function exists despite both samples lying inside the unit disk. In the second case the distance is $(2/5)/(22/25) = 5/11 < 1/2$. With $c = 10/11$,

$$F(z) = \frac{1/5 + cz}{1 + (c/5)z}$$

is a disk-valued holomorphic interpolant and attains the two prescribed values. It is a contraction followed by a disk automorphism; it is not claimed to be an all-pass proper map.

6.3 EP3: the uncertainty disk at an unmeasured point

If only $F(0) = a$ is known, $|a| < 1$, Schwarz–Pick implies for $|z| = r < 1$

$$\left|\frac{F(z) - a}{1 - \bar{a} F(z)}\right| \le r.$$

The set of possible values is the Euclidean disk with center and radius

$$c = \frac{(1 - r^2)a}{1 - r^2|a|^2}, \qquad R = \frac{r(1 - |a|^2)}{1 - r^2|a|^2}.$$

Expand the squared inequality, complete the square in the unknown complex value, and use $1 - r^2|a|^2 > 0$ to obtain the formulas. Every boundary value of this disk is achieved by a disk automorphism after a rotation, so the one-observation uncertainty region is sharp.

Additional observations intersect necessary uncertainty regions and impose a larger Pick matrix. Pairwise inequalities alone are not asserted sufficient for three or more sites. Moreover a Schur-class uncertainty set is conditional on the physical transfer function actually belonging to that analytic class over the whole domain. Finite measurements inside the disk do not prove the class hypothesis.

Exercise 6.3 (EP3a)**.** For $a = 1/3$ and $r = 1/2$, calculate the center and radius. Explain why the uncertainty is not a disk centered at the observed value.

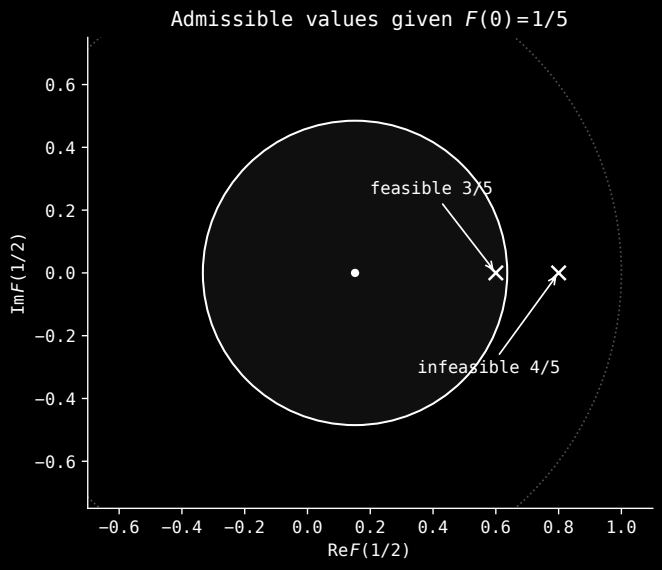


Figure 6.1: The exact feasible value disk at one half given the value one fifth at zero. Both marked observations are inside the unit disk, but only one satisfies the analytic interpolation constraint.

Solution. The denominator is $35/36$, giving $c = 9/35$ and $R = 16/35$. The disk is centered at $9/35$, not $1/3$, because the intrinsic pseudohyperbolic geometry is converted to Euclidean coordinates. Treating the measured value as the center of an arbitrary Euclidean tolerance loses the analytic constraint.

6.4 EP4: passive square-root continuation under uncertainty

For positive-real line factors $Z_\ell = R + sL$ and $Y_\ell = G + sC$, the right-half-plane logarithmic construction selects the passive γ. Suppose a parameter cell supplies a complex enclosure Q for $q = Z_\ell Y_\ell$. A proposed center c and radius r certify a local root through

$$\sup_{q \in Q} |c^2 - q| + r^2 < 2|c|r.$$

If the disk is also wholly in $\operatorname{Re}\gamma > 0$, the selected local branch agrees with the passive branch on overlaps. The sign condition is not optional: the algebraic equation has two roots and only one decays in the chosen propagation direction.

A measurement uncertainty set may cross a true zero of q or leave the class of positive-real factors. In that case the certificate must return a changed hypothesis or an unresolved branch, instead of silently applying the previous sign convention. A fit that is positive real at sample frequencies is not necessarily positive real on the open half-plane.

Exercise 6.4 (EP4a)**.** Certify a passive root for $|q - 4| \le 1$ using $c = 2$, $r = 1/2$. Give a uniform bound on the root error relative to the center.

Solution. The Rouché inequality is $1 + 1/4 < 2$, so there is exactly one root in $|\gamma - 2| < 1/2$. Its real part exceeds $3/2$, giving the passive sign. The root error is at most $1/2$ by the isolating disk; using $|\gamma - 2| = |q - 4|/|\gamma + 2|$ improves the bound to $1/(7/2) = 2/7$, since $|\gamma + 2| \ge \operatorname{Re}(\gamma + 2) > 7/2$.

6.5 EP5: robust stability is about the declared uncertainty set

For the cubic family $s^3 + as^2 + bs + c$, independent positive intervals give the sufficient and, for a rectangular box under these monotonicities, exact worst-corner condition $a_{\min} b_{\min} > c_{\max}$.

For a correlated nonlinear parameterization, this corner may not exist and a parameter-domain certificate is preferable. A tensor Bernstein enclosure of $ab - c$ can prove positivity over a box without treating all coefficients as independent.

For a rational closed-loop transfer, stability of the reduced denominator is not enough when pole-zero cancellations hide internal states. A research implementation should bind each certificate to the actual characteristic polynomial or a verified minimal realization. The recorded exact contour checker proves statements about the polynomial inputs it receives, not about an unprovided plant or controller.

Exercise 6.5 (EP5a)**.** Let $a = 2 + u$, $b = 3 + v$ and $c = 5 + u + v + uv$, with $u, v \in [0, 1]$. Prove strict stability uniformly without enumerating numerical roots.

Solution. All three coefficients are positive and $ab - c = (2 + u)(3 + v) - (5 + u + v + uv) = 1 + 2u + v \geq 1$. The Hurwitz determinant has a strict margin one. This is a continuum algebraic certificate. A numerical root grid is unnecessary for this particular proof, though it can serve as an implementation diagnostic.

6.6 EP6: finite-state refinement of an analytic certificate

A stable exact map plus a rounding bound yields an invariant interval or box. Once the stored states inside it are finite, build a directed graph with one outgoing edge for a deterministic update. Every trajectory eventually enters a directed cycle. This finite graph can distinguish zero convergence from nonzero fixed points and periodic cycles, properties not settled by the analytic contraction bound alone.

For nearest rounding of $3k/4$ on $k \in \{-2, -1, 0, 1, 2\}$ with ties away from zero, every state is fixed: ± 2 round from $\pm 3/2$ back to ± 2, ± 1 round from $\pm 3/4$ back to ± 1, and zero stays zero. The analytic invariant is correct, but a claim of zero convergence is maximally false on this small state set. Changing the tie convention changes part of the graph, so it must be stored as input data.

Exercise 6.6 (EP6a)**.** What does an exhaustive check on that five-state set prove, and what must be added before it proves behavior from an arbitrary machine state?

Solution. It proves every transition and eventual behavior for initial states in that set under the declared exact integer update. To cover arbitrary initial states, prove eventual entry into the invariant set without overflow or saturation, or exhaust a larger machine-state model. A finite graph restricted to a convenient subset is not an invariant of all possible initial conditions by itself.

6.7 EP7: translate mathematical error to the requested observable

Suppose a propagation calculation certifies $|\widehat{\gamma} - \gamma| \leq \varepsilon$ and $\operatorname{Re}\gamma \geq \alpha > 0$ over a parameter range. For length $\ell \geq 0$,

$$|e^{-\ell\widehat{\gamma}} - e^{-\ell\gamma}| \leq e^{-\ell\alpha}\big(e^{\ell\varepsilon} - 1\big).$$

Factor out $e^{-\ell\gamma}$ and use the exponential power series inequality $|e^z - 1| \leq e^{|z|} - 1$. This shows that a bound on the algebraic root can be translated to an amplitude response under a chosen physical length. It does not control a derivative, a group delay, or a reflected-wave realization without additional estimates.

Exercise 6.7 (EP7a)**.** Why is a fixed absolute error in γ not a fixed relative error in a very small transmitted amplitude? State what the displayed inequality actually controls.

Solution. The transmitted amplitude can be exponentially small in length, and a meaningful relative error depends on its exact magnitude and the error in the exponent. The inequality gives an absolute complex response error with a damping factor and an exponential sensitivity term. Dividing by $|e^{-\ell\gamma}|$ instead gives the separate relative bound $e^{\ell\varepsilon} - 1$. Both depend on length; neither is a length-independent guarantee.

6.8 Research acceptance criteria

A defensible research extension should specify the physical class, measured data and uncertainty, analytic continuation domain, realization constraints, arithmetic and requested observable. It should compare a native electrical baseline with the additional geometric certificate, including cases where the advanced method adds no advantage. New mathematical or application novelty would require a new guarantee or validated capability relative to prior methods, not the reuse of a classical Pick matrix or a visually intricate branch animation. No such priority claim or hardware benchmark is made in this test release.

6.9 Research comparison: uncertainty sets instead of fitted certainty

The research-level engineering question is often not whether a formula can be evaluated. It is what all admissible models permit, which measurements would reduce ambiguity, and which mathematical guarantee survives implementation. These questions extend the source's proper-map and lifting structures while preserving the native electrical model.

6.9.1 EP8: a deterministic feasible disk is not a confidence interval

Let $f : \mathbb{D} \to \mathbb{D}$ be holomorphic and suppose $f(0) = b$, $|b| < 1$. Schwarz's lemma applied to $\phi_b \circ f$ gives

$$\left| \frac{f(z) - b}{1 - \bar{b} f(z)} \right| \leq |z|.$$

For a target point of modulus r, the feasible output region is a pseudohyperbolic disk. Completing the square yields its Euclidean centre and radius:

$$c = \frac{(1 - r^2)b}{1 - r^2|b|^2}, \qquad R = \frac{r(1 - |b|^2)}{1 - r^2|b|^2}.$$

Every point in the closed feasible disk is realized by an appropriate Schur function for this one-point constraint: choose a constant α with $|\alpha| \leq 1$ in $\phi_b(f(z)) = \alpha z$ and solve for f.

Conventional fit versus proposed envelope. A least-squares fit returns one response curve. The analytic feasible set answers what is possible given the class and data. Its advantage is explicit non-identifiability and an impossibility certificate for excluded targets. It does not prove that the real system belongs to the assumed Schur class, and it is not a probabilistic confidence statement. With noisy data, replace exact interpolation by an explicitly bounded data set before deriving the region.

Exercise 6.8 (EP8a)**.** For $b = 1/3$ and $r = 1/2$, compute the exact centre and radius. Does the centre equal the original datum?

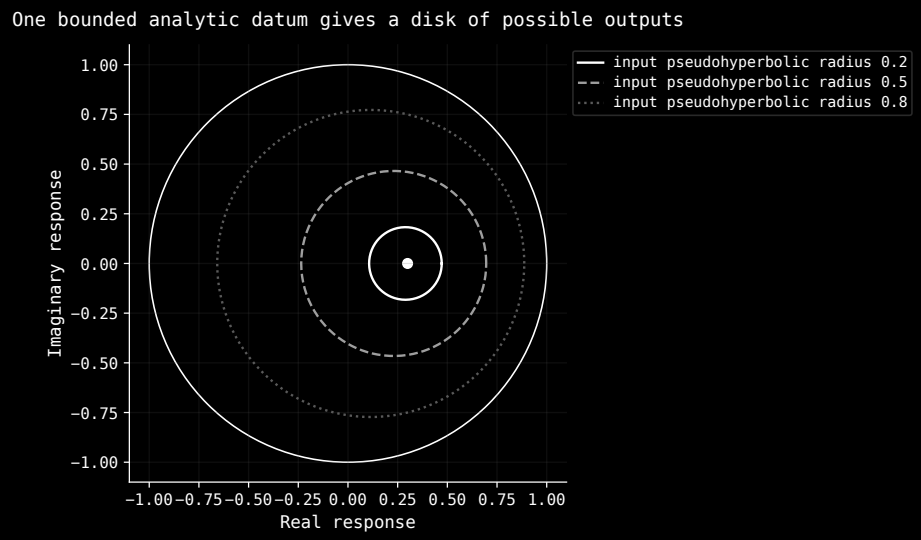


Figure 6.2: Nested feasible output disks for one exact datum. The region is determined by a function-class assumption and a measurement constraint, not by a statistical sampling distribution.

Solution. The denominator is $1 - 1/36 = 35/36$. Thus $c = (3/4)(1/3)/(35/36) = 9/35$ and $R = (1/2)(8/9)/(35/36) = 16/35$. The centre is not $1/3$. Pseudohyperbolic geometry is not the same as drawing a Euclidean disk with the data point at its centre.

6.9.2 EP9: model correlation is a resource

The family $s^3 + ts^2 + ts + t^2/2$, $t \in [1, 4]$, has a positive Routh margin. A coefficient box forgetting the common t contains unstable polynomials. The analytic margin and the box failure are both correct statements about different sets.

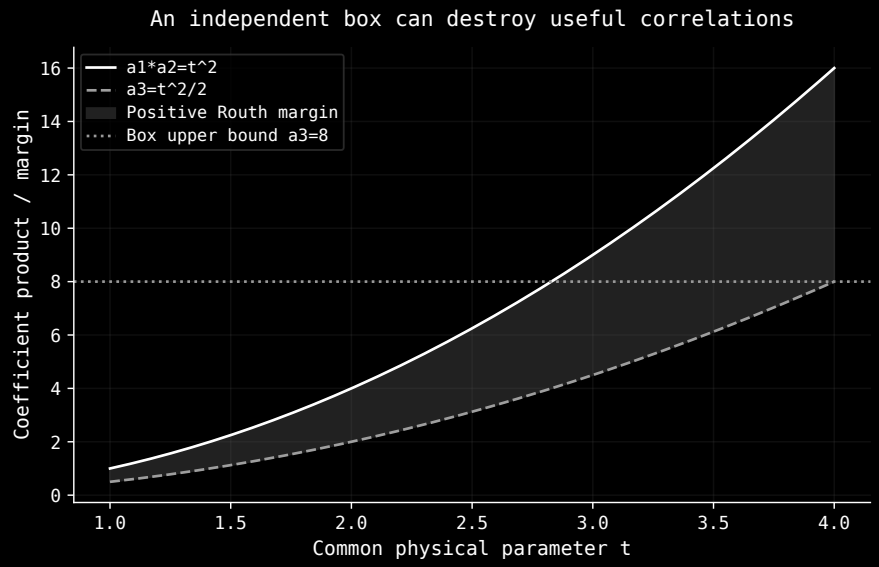


Figure 6.3: The structured family remains inside the strict Routh region while the independent box admits incompatible coefficient combinations. The improvement is faithful representation of uncertainty, not a replacement of the underlying stability theorem.

Proposed research problem. Given a physically derived nonlinear parameterization, construct an enclosure that preserves enough correlation to certify the desired output while remaining

cheap to verify. Candidate representations include affine forms, polynomial/Taylor models, tensor Bernstein coefficients or algebraic elimination with explicit remainder. The benchmark must compare the same physical parameter set, not a narrow proposed set against an unnecessarily broad conventional box without disclosing the change.

Exercise 6.9 (EP9a)**.** What evidence would justify replacing a correlated parameterization by independent intervals? Can a successful box certificate still be useful when the box is conservative?

Solution. The original physical uncertainty set must be contained in the box. Then a successful certificate on the larger box certifies the smaller set as well. Failure on the box is inconclusive for the smaller set. Independence as a statistical assertion is not needed for a deterministic enclosure, but the enclosure must not be treated as the exact physical feasible set unless that is established separately.

6.9.3 EP10: branch-certified propagation versus branch-dependent modes

A transmission model may contain a propagation constant $\gamma = \sqrt{ZY}$. Selecting a passive analytic branch in the right half-plane is a global condition. In contrast, the propagation matrix can often be expressed through entire functions of ZY, such as $\cosh(\ell\sqrt{ZY})$ and $\sinh(\ell\sqrt{ZY})/\sqrt{ZY}$ with the removable value at zero. A mode label can therefore be singular while the physical matrix response remains regular.

Comparison quantity. If the output is the matrix response, an invariant expression can avoid artificial branch jumps. If the output is the individually selected attenuation/phase constant, a branch label is essential. An invariant matrix calculation must not be advertised as solving a branch-identification problem it deliberately avoids. The CSE benchmark supplies the corresponding finite-matrix model and includes a robust general matrix exponential as a control.

6.9.4 EP11: two uncertainty budgets can be composed only after their domains agree

Let η_j^{model} bound the discrepancy between the physical model and the chosen analytic stage, and η_j^{num} bound implementation error. If both are measured in the same stage norm on a common valid tube, their sum can be inserted in the telescoping bound. If they are in different coordinates or over incompatible operating regions, adding their numerical values is meaningless.

A quantitative test. Suppose the downstream weights are $(1/64, 1/16, 1/4, 1)$ and a model discrepancy of 10^{-3} enters only at the last stage. It already contributes 10^{-3} to the final budget. Refining the earliest numerical stage below that scale may have negligible effect on the total. The integer precision optimum in MP8 is a theorem about numerical tolerances under its own cost model; a physical model-error floor changes that optimization problem.

Exercise 6.10 (EP11a)**.** If the target final error is 5×10^{-4} but an independently unavoidable last-stage model uncertainty is 10^{-3}, can increased numerical precision certify the target under a simple worst-case additive model?

Solution. No. The model uncertainty alone exceeds the target. One must reduce the model uncertainty, change the observable/target, or justify a less conservative dependence structure. Cancellation cannot be assumed in a worst-case additive model. A fast solver with arbitrarily small discretization error does not settle this issue.

6.9.5 EP12: proof and measurement remain separate acceptance gates

The present corpus proves properties of declared analytic, algebraic and finite-state models. It supplies reproducible numerical comparisons and rejects corrupted finite certificates. It does not report a hardware trial, a validated teaching intervention or a universal performance improvement. A research engineering submission would add calibrated measurements, operating constraints, uncertainty provenance and replication on an independent implementation.

The comparison slides make these gates visible: the conventional method, the proposed change, the common task, the quantitative result and the remaining limitation appear together. A method is retained when it is already the least expensive complete solution. This avoids turning the request for maximal mathematical scope into an unqualified recommendation to deploy every advanced tool in every design.

Chapter 7

Conformal fields, geometric uncertainty and response compression

7.1 E01. Conformal transformation of electrostatic potential

Let $\phi : U \to \Omega$ be a conformal isomorphism and let $u : U \to \mathbb{R}$ be harmonic. Define $v = u \circ \phi^{-1}$. In real coordinates the Jacobian has the form

$$J_\phi = \begin{pmatrix} \operatorname{Re}\phi' & -\operatorname{Im}\phi' \\ \operatorname{Im}\phi' & \operatorname{Re}\phi' \end{pmatrix}, \qquad J_\phi^T J_\phi = |\phi'|^2 I.$$

The chain rule and the Cauchy–Riemann equations give

$$\Delta(u \circ \phi^{-1}) = 0, \qquad |\nabla v(\phi(z))| = \frac{|\nabla u(z)|}{|\phi'(z)|}.$$

The potential is transported by composition. Electric-field magnitude includes the reciprocal derivative, and the area element includes $|\phi'|^2$. This separation is essential for field-strength interpretation.

For $\phi(z) = z + az^2$, $0 \le a < 1/2$, and $u(z) = V \operatorname{Re} z$, one obtains

$$|\nabla v(\phi(z))| = \frac{|V|}{|1 + 2az|}.$$

On $|z| \le r$, this field lies between $|V|/(1 + 2ar)$ and $|V|/(1 - 2ar)$. At $a = 1/5, r = 3/5$, the ratio of the upper and lower bounds is $31/19$. The parameter coordinates in this example are dimensionless. A length scale ℓ in $\phi_{\text{physical}} = \ell\phi$ inserts the factor $1/\ell$ in field magnitude.

Exercise 7.1. Use $V = 12$ volts, $\ell = 2$ millimetres, $a = 1/5$ and $r = 3/5$. Bound the field magnitude.

Solution. The scale is $|V|/\ell = 6000$ V/m. The exact field interval is

$$[6000/(31/25), 6000/(19/25)] = [150000/31, 150000/19] \quad \text{V/m}.$$

Its approximate endpoints are 4838.71 and 7894.74 V/m. These bounds apply to the specified normalized potential and mapped subdomain. □

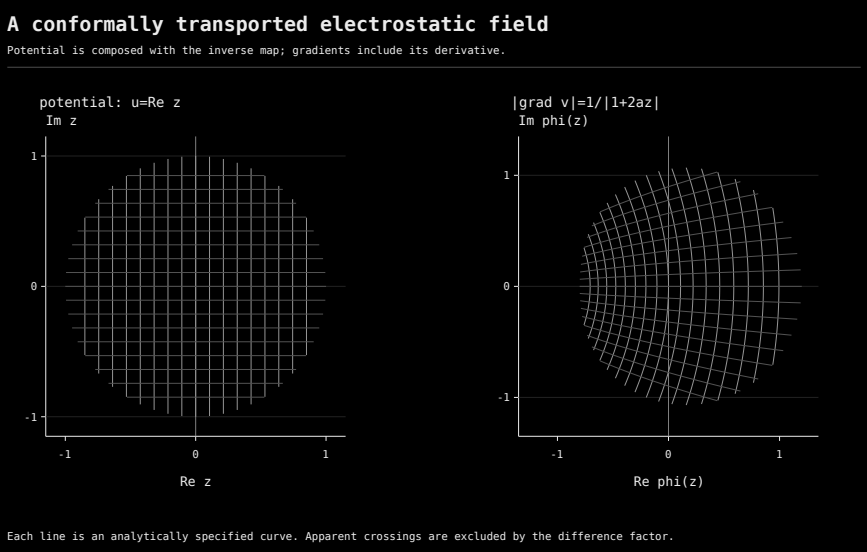


Figure 7.1: Equipotentials and their orthogonal trajectories are mapped together. The physical field calculation retains the derivative factor.

7.2 E02. Energy invariance and capacitance per unit length

For constant permittivity ϵ, the electrostatic energy per unit length is

$$\mathcal{E} = \frac{\epsilon}{2} \int_\Omega |\nabla v|^2 \, dA.$$

A conformal change of variables yields

$$\int_\Omega |\nabla v|^2 dA = \int_U \frac{|\nabla u|^2}{|\phi'|^2} |\phi'|^2 dA = \int_U |\nabla u|^2 dA.$$

This identity is specific to the two-dimensional Dirichlet integral. Physical time in a diffusion process is not preserved by this calculation.

On an annulus $r_0 < |z| < R_0$, prescribe inner potential zero and outer potential V. The solution is

$$u(z) = V \frac{\log(|z|/r_0)}{\log(R_0/r_0)}.$$

Direct polar integration gives

$$\int |\nabla u|^2 dA = \frac{2\pi V^2}{\log(R_0/r_0)}, \quad C' = \frac{2\pi\epsilon}{\log(R_0/r_0)} = \frac{\epsilon}{\operatorname{mod} A(r_0, R_0)}.$$

Here C' denotes capacitance per unit length, since $\mathcal{E} = C'V^2/2$. The source modulus convention includes $1/(2\pi)$; omitting that factor changes the physical result.

For spatially varying permittivity, the coefficient is transported as $\epsilon(\phi(z))$. An anisotropic constitutive tensor requires its own transformation law. The homogeneous isotropic formula does not establish an invariant scalar permittivity for arbitrary materials.

Exercise 7.2. If $\operatorname{mod} A = 1/5$, express C' and the energy for voltage V.

Solution. $C' = 5\epsilon$ and $\mathcal{E} = (5\epsilon/2)V^2$. The dimensions are F/m and J/m when ϵ is expressed in F/m. □

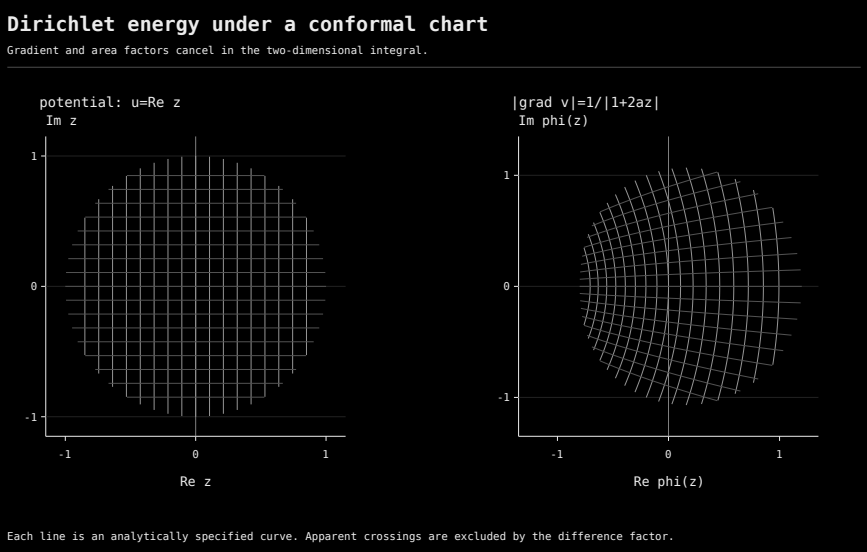


Figure 7.2: The cancellation occurs inside an integral. Pointwise field magnitudes and physical lengths generally change.

7.3 E03. Boundary certification and interior field errors

For two harmonic functions u, v continuous on a bounded domain's closure, a whole-boundary error $|u - v| \le \eta$ implies $|u - v| \le \eta$ throughout the domain by the maximum principle. A finite set of boundary values needs an additional regularity certificate. With N equally spaced angular samples, residual maximum s, and angular Lipschitz constant L, the sufficient whole-boundary bound is $s + \pi L/N$.

If the error is the real part of a holomorphic function h with $|h| \le \eta$ on $\mathbb{D}_R$, Cauchy's estimate gives

$$|h'(z)| \le \frac{\eta}{R - r}, \qquad |z| \le r < R.$$

This derivative estimate requires a bound on the complex potential error, including the conjugate normalization. A bound on its real part alone has a related harmonic gradient estimate with its own constant; the displayed complex estimate must not be applied without the stronger input.

Under a map ϕ with $\inf_{\mathbb{D}_r} |\phi'| \ge m > 0$, the corresponding physical gradient error is at most $\eta/[m(R - r)]$, after inserting the length scale. Near the boundary or near a small derivative, the same potential error permits a much larger field error.

Exercise 7.3. Suppose a complex-potential error is bounded by 10^{-3} volts on radius $R = 1$, the evaluation radius is $r = 3/4$, and the map derivative has lower bound $\ell/2$, with $\ell = 10^{-3}$ m. Bound the field error.

Solution. The bound is $10^{-3}/[(10^{-3}/2)(1/4)] = 8$ V/m. It is a deterministic bound conditional on the stated complex-potential certificate. □

Graphical reference. The shared construction is displayed in Volume II.

7.4 E04. Clearance, conformal radius, and geometric uncertainty

For a conformal isomorphism $\phi : (\mathbb{D}, 0) \to (\Omega, a)$, define $c = |\phi'(0)|$ and $d = \operatorname{dist}(a, \partial\Omega)$. The Koebe quarter theorem and the inverse Schwarz argument give

$$\frac{c}{4} \le d \le c.$$

Thus a verified derivative enclosure $c \in [c_-, c_+]$ implies $d \in [c_-/4, c_+]$. This is a clearance statement for the image domain. It does not certify the distance to an electrode absent from the domain model.

The geometric constants are sharp over the normalized univalent class. Particular models can improve them. For a round disk of radius R and a point at displacement $s < R$ from its centre,

$$c = R\left(1 - \frac{s^2}{R^2}\right), \qquad d = R - s, \qquad \frac{c}{d} = 1 + \frac{s}{R} < 2.$$

The round-disk ratio uses a smaller admissible class. It is not a new universal constant.

Collar separation provides another geometric control. For a full continuum K containing zero with $\operatorname{mod}(\mathbb{D} \setminus K) \geq \mu$, core M29 proves

$$\max_K |z| \leq 1 - e^{-\pi/(4\mu)}.$$

Combining this with derivative distortion yields explicit, possibly conservative, field ratios on compactly separated subdomains. The sharpness of each intermediate estimate and the physical usefulness of the resulting bound should be examined separately.

Exercise 7.4. If the conformal radius is enclosed by $[3.9, 4.1]$ mm, give the universal clearance enclosure.

Solution. $d \in [0.975, 4.1]$ mm. The lower endpoint is divided by four; the upper endpoint is unchanged. An additional geometric model can narrow the interval. □

Graphical reference. The shared construction is displayed in Volume I.

7.5 E05. Bounded analytic response functions and sensitivity

Let a dimensionless response S be holomorphic and disk-valued on $\mathbb{D}$. Schwarz–Pick gives

$$\frac{|S'(z)|}{1 - |S(z)|^2} \leq \frac{1}{1 - |z|^2}, \qquad \left|\frac{S(z) - S(w)}{1 - \overline{S(w)}S(z)}\right| \leq \left|\frac{z - w}{1 - \bar{w}z}\right|.$$

These are bounds in a function class with a global analytic constraint. Membership must be established from the model or a separate certificate. Samples inside the unit disk alone do not establish disk-valuedness on the full domain.

For a normalized positive-real impedance Z in the right half-plane, the Cayley transform $S = (Z - 1)/(Z + 1)$ maps the value constraint to the disk. Its inverse is $Z = (1 + S)/(1 - S)$, with

$$\frac{dZ}{dS} = \frac{2}{(1 - S)^2}.$$

Consequently a bounded response error can correspond to large impedance error near $S = 1$. The derivative is a coordinate-conditioning factor; it remains relevant to the measurement variable.

A state-space or transfer realization must separately establish stability, causality, and the applicable passivity property. For a finite-order rational response, root certificates and positive-real tests are natural companion obligations. Section 4 contributes normal-family and geometric compactness results; it does not determine hardware saturation or sampling effects.

Exercise 7.5. For $S(z) = (z + a)/(1 + az)$ with $a = 3/5$, compute $S'(-a)$ and the disk-metric derivative factor.

Solution. $S'(-a) = 1/(1 - a^2) = 25/16$. Since $S(-a) = 0$, the normalized factor is $|S'(-a)|(1 - a^2)/(1 - |S(-a)|^2) = 1$. The Euclidean derivative exceeds one while the intrinsic factor equals one. □

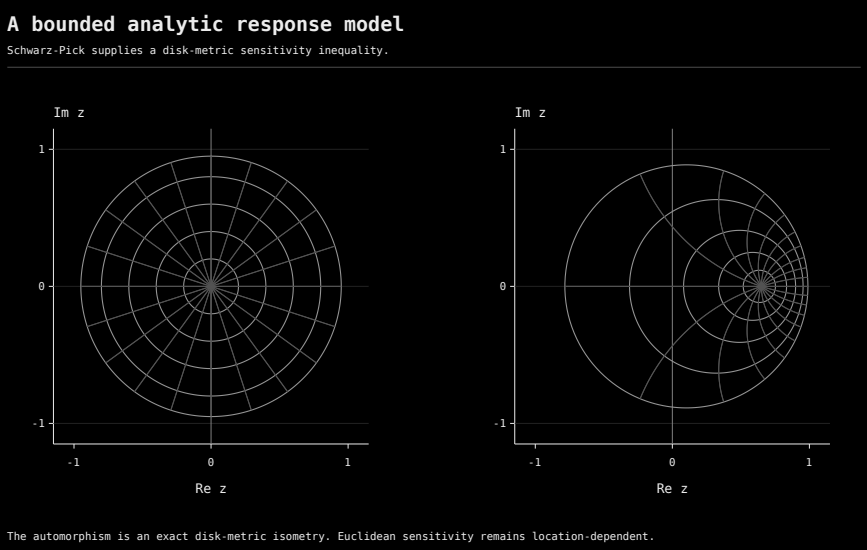


Figure 7.3: The disk automorphism transports an orthogonal family of coordinate curves and preserves the pseudohyperbolic distance exactly.

7.6 E06. Analytic response compression and noise

A square-summable causal discrete-time impulse response h_j defines the analytic generating function $H(z) = \sum_{j\geq 0} h_j z^j \in H^2(\mathbb{D})$. The convention uses nonnegative powers in the disk; the conventional z-transform in negative powers is obtained by reciprocal substitution. The convention must be stated before connecting pole locations and regions of convergence.

Restriction to rz multiplies the coefficients by r^j. It is equivalent to exponential weighting of the impulse response. The optimal linear rank-N error on the H^2 unit ball is r^N. This is a response-space approximation theorem. It does not automatically preserve passivity or minimal state dimension.

The familiar impulse truncation error $\sum_{j\geq N} |h_j|^2$ depends on the actual response. The universal restriction theorem supplies a worst-case class bound after exponential weighting. If measured coefficients have errors e_j, the noise energy after restriction is $\sum r^{2j}|e_j|^2$. These formulas permit a transparent separation of tail error and measurement error.

For a geometric sequence $h_j = \alpha^j$, $|\alpha| < 1$, the exact weighted tail is

$$\sum_{j\geq N} |r\alpha|^{2j} = \frac{|r\alpha|^{2N}}{1 - |r\alpha|^2}.$$

This model-specific value can be far below the class bound. Comparing the two requires the same input normalization.

Exercise 7.6. For $\alpha = r = 1/2$, compute the squared weighted tail after four retained coefficients.

Solution. $|r\alpha|^2 = 1/16$, so the tail is $(1/16)^4/(1 - 1/16) = 1/61440$. The corresponding norm is its square root. □

Graphical reference. The shared construction is displayed in Volume I.

7.7 E07. Model validation at singular and discretized limits

The Koebe map exhibits unbounded derivative near $z = 1$ even though every compact subdisk has a finite distortion bound. In a field computation, the evaluation region and mesh refinement

sequence determine which observable is being approximated. A growing maximum near a singular boundary need not indicate divergence of the potential or energy.

On a wedge of opening α, the harmonic mode $u(r,\theta) = r^{\pi/\alpha}\sin(\pi\theta/\alpha)$ has gradient magnitude proportional to $r^{\pi/\alpha-1}$. For a re-entrant wedge, $\alpha > \pi$, the gradient diverges as $r \downarrow 0$, while the local energy integral is proportional to $\int_0^{\varepsilon} r^{2\pi/\alpha-1}\,dr$, which is finite. This is a concrete separation of pointwise field and integral energy.

Exercise 7.7. For a wedge of angle $3\pi/2$, determine the field exponent and the order of energy in a radius-ε neighbourhood.

Solution. $\pi/\alpha = 2/3$. The field is of order $r^{-1/3}$, and the local energy is of order $\varepsilon^{4/3}$. A pointwise maximum can increase under refinement while the local integrated energy tends to zero. □

Graphical reference. The shared construction is displayed in Volume I.

Chapter 8

Field scales and explicit energy

8.1 EU01. Voltage coordinates and physical length

Mathematical reference. Volume III, conformal-field chapter E01

8.1.1 Prerequisites and governing statement

A potential gradient has units of volts per metre. A dimensionless conformal coordinate needs an explicit physical length scale.

$$\phi_{\mathrm{phys}}(z) = \ell\phi(z)$$

The physical derivative contains the factor ℓ.

$$|E(\ell\phi(z))| = \frac{|V|}{\ell|\phi'(z)|}.$$

8.1.2 Derivation and worked instance

For u(z)=V Re z the dimensionless gradient has magnitude |V|. The inverse coordinate Jacobian divides it by ℓ times |ϕ prime|. This produces the correct units and the physical field magnitude. A conformal transformation preserves angles but generally changes length and gradient size.

With V=12 volts, ℓ=2 millimetres and ϕ prime between 19/25 and 31/25, the field interval is [150000/31,150000/19] V/m.

8.1.3 Comparison and validity conditions

Baseline. Read a field magnitude from a dimensionless plot.

Proposed procedure. Insert voltage and length scales before interpreting the derivative.

Quantitative conclusion. The result is a unit-bearing field enclosure over a complete mapped subdomain.

Validity conditions. The example specifies a homogeneous two-dimensional electrostatic model and its boundary potential. A drawing alone supplies neither.

Exercise 8.1. How does the field change when all physical lengths are doubled at fixed voltage?

Solution. It is divided by two, since ℓ doubles while the dimensionless derivative is unchanged.□

Graphical reference. See Figure 7.1 for the shared construction.

8.2 EU02. Energy of a coaxial cross-section

Mathematical reference. Volume III, conformal-field chapter E02

8.2.1 Prerequisites and governing statement

The energy per unit length of a homogeneous electrostatic model is $\varepsilon/2$ times the integral of the squared field.

$$\mathcal{E} = \frac{\epsilon}{2}\int |\nabla u|^2 dA$$

For a capacitance per unit length C prime, the same energy is C prime V squared/2.

$$C' = \frac{2\pi\epsilon}{\log(R/r)} = \frac{\epsilon}{\operatorname{mod} A(r,R)}.$$

8.2.2 Derivation and worked instance

The potential V log(rho/r)/log(R/r) satisfies the two circular boundary conditions. Its radial derivative is V/[rho log(R/r)]. Squaring and integrating with the polar area factor 2pi rho gives 2pi V squared/log(R/r). Comparing the energy with C prime V squared/2 yields the formula.

If R/r=exp(2pi/5), the modulus is 1/5 and C prime=5epsilon. The exponential ratio is dimensionless.

8.2.3 Comparison and validity conditions

Baseline. Use a memorized coaxial-capacitance formula without a modulus convention.

Proposed procedure. Derive the logarithmic potential and integrate its energy explicitly.

Quantitative conclusion. The derivation fixes the otherwise ambiguous factor of 2pi.

Validity conditions. The formula assumes a homogeneous isotropic material and an effectively two-dimensional geometry per unit length.

Exercise 8.2. What happens to C prime if both radii are multiplied by the same positive number?

Solution. Their ratio is unchanged, so C prime is unchanged in the ideal two-dimensional homogeneous model. □

8.3 EU03. Impedance and reflection coordinates

8.3.1 Prerequisites and governing statement

A normalized impedance Z in the right half-plane is transformed to a disk-valued reflection coordinate.

$$S = \frac{Z-1}{Z+1}$$

For Z=1, the reflection coordinate is zero.

$$Z = \frac{1+S}{1-S}, \qquad \frac{dZ}{dS} = \frac{2}{(1-S)^2}.$$

8.3.2 Derivation and worked instance

Solving the first equation for Z gives the inverse. Differentiation shows that the inverse becomes sensitive as S approaches one. A small reflection error therefore has a location-dependent effect in impedance units. The disk geometry is useful for the analytic constraint, while the physical variable determines the relevant measurement error.

At S=0, the derivative magnitude is two. At the real value S=9/10, it is two hundred. These are exact local derivatives, not finite perturbation bounds of arbitrary size.

8.3.3 Comparison and validity conditions

Baseline. Interpret the same absolute tolerance in impedance and reflection coordinates.

Proposed procedure. Compute the coordinate transformation and its local sensitivity.

Quantitative conclusion. The coordinate-conditioning factor is explicit and depends on the operating point.

Validity conditions. For a finite perturbation, the entire segment must stay away from S=1 and the derivative bound must hold along it.

Exercise 8.3. Compute the normalized impedance corresponding to S=1/2.

Solution. Z=(1+1/2)/(1-1/2)=3. □

Graphical reference. See Figure 7.3 for the shared construction.

8.4 EU04. Certifying unsampled boundary angles

Mathematical reference. Volume III, conformal-field chapter E03; Volume II, analytic-approximation chapter C04.

8.4.1 Prerequisites and governing statement

A Lipschitz constant limits how fast a residual can vary with angle.

$$|g(\theta) - g(\vartheta)| \le L d_{\mathbb{T}}(\theta, \vartheta)$$

With N equally spaced angles, the nearest sample is at angular distance at most π/N.

$$\sup |g| \le s + \pi L/N.$$

8.4.2 Derivation and worked instance

The triangle inequality combines the residual at the nearest sample with the maximum allowed change over the intervening angular distance. The resulting bound covers every boundary angle. If the field error is harmonic and has this whole-boundary bound, the maximum principle transports it to the interior potential error.

For s=1/1000,L=1/5,N=70, replacing π by 22/7 gives a bound below 1/100. Seventy sample intervals have been certified through the derivative assumption.

8.4.3 Comparison and validity conditions

Baseline. Report the largest measured boundary residual as the exact maximum.

Proposed procedure. Add the full inter-sample variation permitted by the regularity bound.

Quantitative conclusion. A finite sample record yields a continuum certificate conditional on one regularity input.

Validity conditions. The derivative or Lipschitz bound must itself be justified for the actual residual. It cannot be inferred from differences of the same finite samples without further hypotheses.

Exercise 8.4. If N is doubled with s and L fixed, which part of the bound changes?

Solution. The inter-sample term π L/N is halved. The sample-error floor s remains unchanged.□

8.5 EU05. A clearance estimate from a conformal radius

Mathematical reference. Volume III, conformal-field chapter E04; source Corollary 4.18.

8.5.1 Prerequisites and governing statement

Conformal radius and Euclidean clearance are lengths associated with a marked point in a plane domain.

$$c = |\phi'(0)|, \qquad d = \operatorname{dist}(a, \partial\Omega)$$

The image domain must be Omega and $\phi(0)$=a.

$$c/4 \le d \le c.$$

8.5.2 Derivation and worked instance

The universal inequalities provide an interval for clearance. Uncertainty in c propagates monotonically: if c lies between c minus and c plus, then d lies between c minus/4 and c plus. A physical interpretation additionally needs the domain boundary to represent the actual geometric obstacle or conductor under consideration.

A conformal-radius enclosure [3.9,4.1] millimetres implies clearance in [0.975,4.1] millimetres. The interval can be conservative for nearly circular domains.

8.5.3 Comparison and validity conditions

Baseline. Estimate the nearest boundary distance from a rasterized geometry.

Proposed procedure. Use a verified conformal derivative together with the quarter theorem.

Quantitative conclusion. The geometric certificate includes all boundary points of the mathematical image.

Validity conditions. A clearance guarantee is not a dielectric-breakdown guarantee. Material, voltage and field enhancement need their own model.

Exercise 8.5. What is the universal clearance enclosure if c is exactly 8 millimetres?

Solution. It is [2,8] millimetres. □

Graphical reference. The shared construction is displayed in Volume I.

8.6 EU06. Exponential weighting of a causal response

Mathematical reference. Volume III, conformal-field chapter E06

8.6.1 Prerequisites and governing statement

A causal sequence h_j has a generating function with nonnegative powers in the disk.

$$H(z) = \sum_{j \geq 0} h_j z^j$$

The replacement z by rz multiplies the j-th coefficient by r^j.

$$\sum_{j \geq N} |r\alpha|^{2j} = \frac{|r\alpha|^{2N}}{1 - |r\alpha|^2}.$$

8.6.2 Derivation and worked instance

For the geometric impulse response h_j=α^j, exponential weighting and truncation leave another geometric series. The formula is valid when |r α|<1. It quantifies a squared sequence norm. Peak values, stability margins and passivity are different requested outputs and need suitable additional bounds.

With α=r=1/2 and N=4, the squared weighted tail is 1/61440. Retaining four coefficients means indices zero through three.

8.6.3 Comparison and validity conditions

Baseline. Truncate an impulse response without recording its remaining energy.

Proposed procedure. Compute an exact weighted tail for a specified response model.

Quantitative conclusion. The tail is exact for the geometric model.

Validity conditions. The generating-function convention uses positive powers. A conventional negative-power z-transform must be connected by reciprocal substitution.

Exercise 8.6. For α=1/2,r=1,N=3, compute the unweighted squared tail.

Solution. The ratio is 1/4, so the tail is $(1/4)^3$/(1-1/4)=1/48. □

8.7 EU07. A field singularity with finite energy

8.7.1 Prerequisites and governing statement

Polar area measure is r dr dtheta. A power singularity can have finite integral even when its pointwise maximum diverges.

$$\int_0^{\varepsilon} r^{\beta}\, dr < \infty \quad \text{if } \beta > -1$$

The integral equals $\varepsilon^{\beta+1}/(\beta+1)$.

$$u = r^{\pi/\alpha} \sin(\pi\theta/\alpha), \qquad |\nabla u| \asymp r^{\pi/\alpha - 1}.$$

8.7.2 Derivation and worked instance

The angular boundary conditions vanish on the wedge edges. Differentiating the radial and angular factors gives the stated gradient order. Squaring and multiplying by polar area gives an integrand of order $r^{2\pi/\alpha-1}$. Its exponent exceeds minus one for every positive wedge angle, so the local energy is finite.

At a re-entrant angle $3\pi/2$, the potential exponent is $2/3$, the field exponent is $-1/3$, and the energy in radius ε is of order $\varepsilon^{4/3}$.

8.7.3 Comparison and validity conditions

Baseline. Interpret every growing mesh maximum as divergence of the complete solution.

Proposed procedure. Track the potential, gradient and integrated energy with their own asymptotic orders.

Quantitative conclusion. The model predicts distinct convergence behavior for different observables.

Validity conditions. The wedge is an ideal local geometry. Finite corner rounding changes the peak field at sufficiently small scales.

Exercise 8.7. What is the field exponent at a right-angle wedge?

Solution. For $\alpha = \pi/2$, $\pi/\alpha = 2$, so the field is of order r and vanishes at the vertex. □

Graphical reference. The shared construction is displayed in Volume II.

8.8 EU08. Boundary conditions select the physical problem

Mathematical reference. Volume III, conformal-field chapter E01 and E03

8.8.1 Prerequisites and governing statement

A harmonic equation can have many solutions before boundary conditions are specified.

$$\Delta u = 0$$

Both zero and Re z satisfy this equation in the unit disk.

$$\Delta(u - v) = 0, \quad \|(u - v)|_{\partial D}\|_\infty \le \eta \implies \|u - v\|_{\overline{D}} \le \eta.$$

8.8.2 Derivation and worked instance

The maximum principle applies to the harmonic error and its negative. It yields an interior certificate only after the boundary error has been bounded. An interior residual test alone does not distinguish two exact harmonic solutions with different boundary values. The example also shows why a model fitted to the wrong electrodes can satisfy its own PDE perfectly.

On the unit disk, zero and Re z have identical zero Laplacian. Their boundary difference has supremum one, and their interior supremum difference is also one.

8.8.3 Comparison and validity conditions

Baseline. Accept a numerical potential after checking a small interior residual.

Proposed procedure. Check the differential equation, the complete boundary data, and the physical coefficients.

Quantitative conclusion. The certificate identifies the intended boundary-value problem.

Validity conditions. Finite collocation points do not prove that the Laplacian vanishes throughout the continuum.

Exercise 8.8. What proves uniqueness when the boundary error is zero?

Solution. The maximum principle gives $|u-v| \leq 0$ throughout the domain, so the two solutions coincide. □

Chapter 9

Wave coordinates and spectral constraints

9.1 EG01. Complex potential error and harmonic gradient error

Mathematical reference. Volume III, conformal-field chapter E03; Volume I, normal-family and mapping core M01 and M32.

9.1.1 Prerequisites and governing statement

The real part of a holomorphic function is harmonic. Bounds on the full complex function and its real part are different data.

$$u = \operatorname{Re} F$$

Adding an imaginary constant changes F but leaves u unchanged.

$$\|F - G\|_{\mathbb{D}_R} \le \eta \implies \|(F - G)'\|_{\mathbb{D}_r} \le \eta/(R - r).$$

9.1.2 Derivation and worked instance

The displayed estimate follows from Cauchy and requires a complex-valued supremum bound. For a real harmonic error, a gradient estimate can be derived from the Poisson kernel or harmonic interior estimates, with its own constant. An arbitrary imaginary offset obstructs a direct inference of a complex supremum from a real supremum until a conjugate normalization is selected.

If the full complex error is at most 10^{-3} on radius one, its derivative is at most $4*10^{-3}$ on radius 3/4. A physical map with derivative lower bound m multiplies this by at most 1/m.

9.1.3 Comparison and validity conditions

Baseline. Apply a complex Cauchy bound directly to a measured real-potential error.

Proposed procedure. Identify whether the certificate controls a complex potential or only its real part, then use the matching estimate.

Quantitative conclusion. The bound matches the actual information supplied by the potential certificate.

Validity conditions. The conjugate potential may be multivalued on a multiply connected domain. Topological periods must be included before constructing a global complex potential.

Exercise 9.1. Give a complex function whose real part is zero but whose modulus is arbitrarily large.

Solution. F(z)=iC with real C. Its real part vanishes and its modulus is |C|. Its derivative vanishes, illustrating the role of normalization without producing a gradient counterexample. □

Graphical reference. The shared construction is displayed in Volume II.

9.2 EG02. The transformed conductivity equation

Mathematical reference. Volume III, conformal-field chapter E01–E02

9.2.1 Prerequisites and governing statement

The divergence-form equation div(sigma grad u)=0 describes a steady conductivity model. A coordinate change transforms both gradient and area.

$$J_\phi = |\phi'|Q, \qquad Q^T Q = I$$

A conformal Jacobian is a scale multiplied by a rotation.

$$\widetilde{\sigma} = \det(J) J^{-1} (\sigma \circ \phi) J^{-T}.$$

9.2.2 Derivation and worked instance

For a scalar isotropic sigma and a conformal J in two dimensions, the scale factors cancel and the transformed tensor is the scalar sigma composed with ϕ times the identity. For a general coordinate map, the tensor becomes anisotropic. The formula follows by substituting the chain-rule gradient into the weak bilinear form and applying the Jacobian change of variables.

For the real scaling map J=diag(2,1), a unit scalar conductivity transforms to diag(1/2,2). The anisotropy is a direct consequence of unequal coordinate stretching.

9.2.3 Comparison and validity conditions

Baseline. Copy a constant-coefficient field equation through an arbitrary geometric warp.

Proposed procedure. Derive the transformed constitutive tensor from the weak energy form.

Quantitative conclusion. The physical equation is transported with its correct constitutive coefficients.

Validity conditions. Orientation and determinant conventions must be fixed. The displayed formula assumes an orientation-preserving map from computational to physical coordinates.

Exercise 9.2. What transformed tensor results from J=2I in two dimensions and sigma=1?

Solution. det J=4 and J inverse=(1/2)I, so the transformed tensor is 4*(1/2)^2 I=I. □

9.3 EG03. Disk geometry and response sensitivity

9.3.1 Prerequisites and governing statement

The pseudohyperbolic distance is invariant under disk automorphisms.

$$\delta(z,w)=\left|\frac{z-w}{1-\bar{w}z}\right|$$

At w=0 it equals |z|.

$$\frac{|S'(z)|(1-|z|^2)}{1-|S(z)|^2}\leq 1.$$

9.3.2 Derivation and worked instance

Normalize the selected input and output points to zero using disk automorphisms. Schwarz's lemma bounds the derivative of the normalized map by one. Undoing the coordinate changes gives the displayed inequality. Equality for an automorphism supplies a positive control: Euclidean expansion at some points is fully compatible with intrinsic isometry.

For S(z)=(z+3/5)/(1+3z/5), the derivative at z=-3/5 is 25/16 and the intrinsic factor equals one.

9.3.3 Comparison and validity conditions

Baseline. Compare sensitivity with a single Euclidean Lipschitz constant over the entire disk.

Proposed procedure. Use Schwarz-Pick and then convert the result to the required physical coordinate.

Quantitative conclusion. The sensitivity bound is invariant under the natural disk-coordinate automorphisms.

Validity conditions. The theorem requires a holomorphic disk-valued function on the complete domain. Measured values at finitely many frequencies do not establish that premise.

Exercise 9.3. Does intrinsic nonexpansiveness imply a unique attracting fixed point?

Solution. No. Rotations are isometries; the identity fixes every point. A strict contraction or an additional dynamical theorem is needed. □

Graphical reference. See Figure 7.3 for the shared construction.

9.4 EG04. Two measurements and an analytic feasibility test

9.4.1 Prerequisites and governing statement

A Schur function is holomorphic and bounded by one in modulus on the disk. Schwarz-Pick constrains any two prescribed values.

$$\delta(w_1,w_2)\leq\delta(z_1,z_2)$$

If both first coordinates are zero, the condition becomes |w2|≤|z2|.

$$S(0)=0,\quad S(r)=w\quad\Longleftrightarrow\quad|w|\leq r\quad(0<r<1).$$

9.4.2 Derivation and worked instance

Necessity is Schwarz's lemma. Sufficiency is supplied by S(z)=(w/r)z, which is disk-valued when |w|≤r. Disk automorphisms reduce a general two-point problem to this normalized case. The result is a complete feasibility criterion for two data in the scalar Schur class.

At r=1/2, the requested value w=3/5 is infeasible, while w=2/5 is realized by S(z)=4z/5. A least-squares polynomial can fit the infeasible data exactly without belonging to the required class.

9.4.3 Comparison and validity conditions

Baseline. Interpolate measured complex values with an unconstrained polynomial.

Proposed procedure. Test membership in the bounded analytic class before constructing an interpolant.

Quantitative conclusion. One scalar inequality provides either a feasible construction or an impossibility certificate.

Validity conditions. The analytic class must be justified by the physical model and transform convention. The test does not independently prove passivity of a device.

Exercise 9.4. Find a Schur interpolant for S(0)=0 and S(3/4)=1/2.

Solution. S(z)=(2/3)z satisfies both values and has modulus below one on the disk. □

9.5 EG05. Geometric approximation error in a field certificate

Mathematical reference. Volume I, normal-family and mapping core M32–M33; Volume III, conformal-field chapter E01 and E03.

9.5.1 Prerequisites and governing statement

A holomorphic map error controls its derivative on a smaller disk. A positive derivative margin controls inverse-coordinate amplification.

$$\delta = \varepsilon/(R-r) < m$$

The approximate derivative stays at least m-δ in modulus.

$$\left|\frac{1}{p'} - \frac{1}{\phi'}\right| \le \frac{\delta}{m(m-\delta)}.$$

9.5.2 Derivation and worked instance

The reciprocal difference equals the derivative difference divided by the product of the two derivatives. The exact derivative is bounded below by m and the approximate derivative by m-δ. This gives the displayed bound. A full physical field comparison also accounts for evaluation at displaced source coordinates and for any approximation error in the potential itself.

For m=19/25 and δ=1/20, substitute into the reciprocal-factor bound. The exact result is 125/1349, approximately 0.09266.

9.5.3 Comparison and validity conditions

Baseline. Use the same field bound after replacing an exact map by a numerical approximation.

Proposed procedure. Propagate derivative uncertainty into the reciprocal field factor.

Quantitative conclusion. The field certificate explicitly includes uncertainty in the geometric map.

Validity conditions. The bound is local to the certified subdomain and derivative margin. It diverges as the margin closes.

Exercise 9.5. Why does a small value error of the map require a larger-domain bound before controlling its derivative?

Solution. Cauchy differentiates through an enclosing circle. The available collar R-r supplies the denominator; a pointwise value error alone has no such derivative control. □

Graphical reference. The shared construction is displayed in Volume I.

9.6 EG06. Green functions as normalized geometric observables

Mathematical reference. Volume I, normal-family and mapping core M08; source section 4.2 following the Riemann mapping proof.

9.6.1 Prerequisites and governing statement

The Green function with a pole at a has logarithmic singularity and vanishes at the ideal boundary in the appropriate potential-theoretic sense.

$$G_a(\phi(z)) = -\log|z|$$

Equipotentials correspond to fixed |z|.

$$G_a(w) = -\log|w-a| + \log c + o(1), \quad c = |\phi'(0)|.$$

9.6.2 Derivation and worked instance

Near zero, ϕ(z)=a+ϕ prime(0)z+o(z). Inverting this relation and taking logarithms gives the expansion. The constant term log c therefore records the conformal radius. Radial lines and circles are orthogonal in the disk, and their images remain orthogonal where the conformal derivative is nonzero.

For a centred disk of radius R, G_a(w)=log(R/|w-a|), so the regular part is log R, agreeing with c=R.

9.6.3 Comparison and validity conditions

Baseline. Treat potential level sets as arbitrary plotted contours.

Proposed procedure. Identify their conformal parametrization and the normalization encoded in the singular part.

Quantitative conclusion. One normalized function determines both foliations and the conformal-radius constant.

Validity conditions. General domains need not have a classically smooth boundary. The source's Green-ray construction does not establish continuous landing at every boundary point.

Exercise 9.6. What is the radius of the equipotential G_a=t in a centred disk?

Solution. It is R exp(-t), obtained by solving log(R/rho)=t. □

9.7 EG07. Finite-frequency data and global response ambiguity

9.7.1 Prerequisites and governing statement

A finite set of interpolation conditions leaves infinitely many analytic degrees of freedom unless a model class is restricted.

$$p(z) = \prod_{k=1}^{N} (z - z_k)$$

The polynomial vanishes at every selected sample point.

$$\widetilde{H} = H + \delta p \quad \Longrightarrow \quad \widetilde{H}(z_k) = H(z_k).$$

9.7.2 Derivation and worked instance

Every δ produces identical sampled values. Stability or boundedness constraints can restrict the allowed δ but need not make the model unique. A finite-order rational assumption can reduce the degrees of freedom, yet its order and denominator conditions become part of the inference. The correct output may be a feasible family or an envelope over that family.

For equally spaced nodes on radius r, the perturbation is $\delta(z^N\text{-}r^N)$. Its maximum on that circle is $2|\delta|r^N$ despite zero perturbation at every sample.

9.7.3 Comparison and validity conditions

Baseline. Treat an exact fit at measured frequencies as a unique global model.

Proposed procedure. Construct analytic perturbations vanishing at all data and identify the constraints needed to exclude them.

Quantitative conclusion. The ambiguity has an explicit analytic witness.

Validity conditions. An estimated model class is an assumption about the physical system, not a consequence of interpolation accuracy.

Exercise 9.7. What additional condition makes a degree-(N-1) polynomial uniquely determined by N distinct values?

Solution. The degree bound. The difference of two such polynomials has N distinct zeros and degree below N, hence vanishes identically. □

Graphical reference. The shared construction is displayed in Volume II.

9.8 EG08. Conformal transformation of a wave equation

Mathematical reference. Volume III, conformal-field chapter physical-model specification.

9.8.1 Prerequisites and governing statement

A conformal map rescales the planar Laplacian by the squared derivative modulus.

$$\Delta_z(u \circ \phi) = |\phi'|^2(\Delta_w u) \circ \phi$$

The scalar factor is generally nonconstant.

$$(\Delta_w + k^2)u = 0 \quad \Longrightarrow \quad \Delta_z v + k^2|\phi'(z)|^2 v = 0.$$

9.8.2 Derivation and worked instance

Let v=u composed with ϕ. Substitute the Laplacian transformation and the original equation. The new zeroth-order coefficient is k squared times |ϕ prime| squared. Harmonicity is invariant because the zeroth-order term is absent when k=0. At nonzero wavenumber the transformed medium or equation must retain the spatial factor.

For ϕ(z)=z+$z^2/5$, the factor is $|1+2z/5|^2$. It varies across the disk and cannot be represented by one unchanged constant k.

9.8.3 Comparison and validity conditions

Baseline. Transfer a constant-coefficient Helmholtz equation through a conformal map unchanged.

Proposed procedure. Carry the induced spatial coefficient into the transformed equation.

Quantitative conclusion. The transformed model preserves the original equation's solutions with its correct spatial coefficient.

Validity conditions. This scalar two-dimensional calculation is not a full transformation rule for arbitrary vector Maxwell systems.

Exercise 9.8. What happens for the uniform scaling ϕ(z)=ℓ z?

Solution. The transformed wavenumber squared is k^2 ℓ^2, corresponding to the dimensionless wavenumber k ℓ. □

Chapter 10

Uncertainty, matrix response and model restrictions

10.1 EP01. Energy forms and admissible coordinate transformations

Mathematical reference. Volume III, conformal-field chapter E02; graduate dossier EG02.

10.1.1 Prerequisites and governing statement

A weak formulation is expressed through a bilinear energy form. Coordinate transformations act on its coefficient tensor.

$$a(u, v) = \int_{\Omega} (\nabla u)^{T} \sigma \nabla v \, dA$$

Both test and trial gradients transform by J inverse transpose.

$$\widetilde{\sigma} = \det J \, J^{-1} (\sigma \circ \phi) J^{-T}.$$

10.1.2 Derivation and worked instance

The transformation law follows by applying the chain rule to both gradients and the Jacobian determinant to the integration measure. A conformal two-dimensional map preserves scalar isotropy after composition of the material coefficient. A transient mass term instead acquires the Jacobian density |ϕ prime| squared. The distinction follows directly from which derivatives occur in each term.

For a heat equation with constant heat capacity c, the transformed weak mass coefficient is c times the squared modulus of the conformal derivative. Its spatial dependence enters the time-evolution problem.

10.1.3 Comparison and validity conditions

Baseline. Infer invariance of a physical model from local angle preservation alone.

Proposed procedure. Transform the full weak form and identify the coefficients and measures that survive.

Quantitative conclusion. The full transformed weak form identifies a precise class of invariant steady models.

Validity conditions. Material models, temporal variables, and boundary flux conventions must accompany any physical interpretation.

Exercise 10.1. Why does the mass term lack the gradient cancellation present in the energy term?

Solution. It contains no spatial derivatives. The area Jacobian remains without reciprocal derivative factors. □

Graphical reference. See Figure 7.2 for the shared construction.

10.2 EP02. Geometric uncertainty in an evaluation functional

Mathematical reference. Volume I, normal-family and mapping core M36; Volume III, conformal-field chapter E04.

10.2.1 Prerequisites and governing statement

The Bergman kernel diagonal equals the squared norm of point evaluation, and conformal radius determines it exactly.

$$K_D(a, a) = 1/(\pi c^2)$$

The function c mapped to $1/(\pi$ c squared) is decreasing for positive c.

$$c \in [c_-, c_+] \quad \Longrightarrow \quad K_D(a, a) \in [1/(\pi c_+^2), 1/(\pi c_-^2)].$$

10.2.2 Derivation and worked instance

The radius uncertainty and the functional norm enter different factors. Once a conformal-radius enclosure is proved, monotonicity gives the kernel enclosure without a new PDE solve. Evaluation of an A^2-normalized analytic observable then has magnitude at most sqrt(K_D(a,a)). The interpretation depends on the area normalization and on the chosen analytic-observable model.

For c in [2,3], the evaluation norm lies between $1/(3\text{sqrt } \pi)$ and $1/(2\text{sqrt } \pi)$. The interval is attained as a functional norm for the corresponding geometric endpoints when such domains are admitted.

10.2.3 Comparison and validity conditions

Baseline. Use a nominal conformal map to report one pointwise response bound.

Proposed procedure. Propagate a verified geometric-radius interval through the monotone kernel relation.

Quantitative conclusion. Geometric uncertainty becomes a quantitative functional-evaluation envelope.

Validity conditions. An electrical field is not automatically an arbitrary A^2 function with unit norm. A physical energy-to-function-space identification must be derived.

Exercise 10.2. What relative first-order sensitivity does K have with respect to c?

Solution. Differentiating logarithms gives d log K=-2 d log c. A small relative radius error produces approximately twice that relative kernel error with opposite sign. □

10.3 EP03. Approximation rank and realization constraints

Mathematical reference. Volume I, normal-family and mapping core M35; Volume III, conformal-field chapter E05–E06.

10.3.1 Prerequisites and governing statement

An analytic transfer representation may obey additional constraints such as positive realness or passivity.

$$\|T_r - T_r P_N\| = r^N$$

This equation is a Hilbert-space approximation statement, with no passivity premise.

$$\inf_{A \in \mathcal{A}_N \cap \mathcal{R}} \|T_r - A\| \geq \inf_{A \in \mathcal{A}_N} \|T_r - A\|.$$

10.3.2 Derivation and worked instance

Adding a constraint set R can only increase the infimum or leave it unchanged. The unconstrained rank optimum is therefore a lower bound for a constrained problem, provided the two are formulated in the same operator setting. Physical realization is often nonlinear in the coefficients and may not even define a linear operator class. In that case a new formulation is needed before transferring the rank theorem.

Truncating a power series preserves holomorphy inside its convergence disk. Positivity of the real part is a global inequality and is not a formal consequence of that truncation identity.

10.3.3 Comparison and validity conditions

Baseline. Assume that an optimal coefficient truncation preserves every physical property of a response.

Proposed procedure. Specify physical realization constraints as additional admissibility conditions and re-evaluate the optimization problem.

Quantitative conclusion. The unconstrained optimum is positioned correctly as a conditional lower bound.

Validity conditions. The dossier establishes an optimization-class relation, not a universal counterexample to every truncation of passive models.

Exercise 10.3. Can a constrained optimum be smaller than the unconstrained one over the same ambient class?

Solution. No. The constrained admissible set is a subset, so minimization over it cannot lower the infimum. □

Graphical reference. The shared construction is displayed in Volume I.

10.4 EP04. A local sampling certificate with uncertain node positions

Mathematical reference. Volume III, conformal-field chapter E03; Volume II, analytic-approximation chapter C04.

10.4.1 Prerequisites and governing statement

If sample positions are perturbed, the mesh distance to a true observation changes. A Lipschitz bound converts that displacement to value error.

$$|\widetilde{\theta}_k - \theta_k| \le \tau$$

Every target angle is within π/N+tau of some perturbed sample.

$$\sup |g| \le s + L(\pi/N + \tau).$$

10.4.2 Derivation and worked instance

Choose the closest nominal node to the target angle and then apply the triangle inequality for circular distance to its perturbed location. The Lipschitz estimate bounds the change in residual. The resulting expression separates sample-value uncertainty s, discretization π/N, and angular calibration tau. Increasing N eventually reaches the calibration and measurement floors.

For L=1/5 and tau=1/1000, the calibration contribution is 1/5000, independent of N. It remains after arbitrarily dense sampling.

10.4.3 Comparison and validity conditions

Baseline. Use the nominal equispaced-mesh bound despite uncertain measurement angles.

Proposed procedure. Include angular calibration error in the continuum residual certificate.

Quantitative conclusion. The certificate remains valid under a declared deterministic node-position uncertainty.

Validity conditions. Angular uncertainty must be expressed in radians in this formula. Conversion from degrees is part of the input specification.

Exercise 10.4. What sample count makes π/N no larger than tau=1/1000?

Solution. N≥1000pi, so N=3142 suffices. The conservative rational rule π<22/7 gives N=3143.□

10.5 EP05. Collar geometry and field-error allocation

Mathematical reference. Volume I, normal-family and mapping core M29–M30; source Theorem 4.16.

10.5.1 Prerequisites and governing statement

A conformal collar gives an interior radius after uniformization. Koebe bounds then control derivative variation on that radius.

$$k(\mu) = 1 - e^{-\pi/(4\mu)}$$

The radius is strictly below one for every positive μ.

$$\frac{|\phi'(z)|}{|\phi'(w)|} \le \left(\frac{1+k}{1-k}\right)^8.$$

10.5.2 Derivation and worked instance

The invariant-domain bound compares both the source uniformization and the composed target map. Each contributes a Koebe derivative quotient, producing the eighth power. The explicit collar formula makes the dependence on μ computable, though conservative. In field estimates, a small derivative margin can dominate the required approximation accuracy. A model-specific map can substantially narrow the bound.

For the quadratic chart the exact derivative quotient on a disk is (1+2ar)/(1-2ar). The universal invariant-domain estimate can be many orders larger because it admits a much broader geometric class.

10.5.3 Comparison and validity conditions

Baseline. Allocate the same error tolerance throughout a geometrically nonuniform region.

Proposed procedure. Propagate the collar-dependent derivative bounds into spatial error budgets.

Quantitative conclusion. The allocation formula exhibits every geometric amplification factor.

Validity conditions. A nonsharp universal enclosure may be too conservative for a practical design decision. Its validity and its numerical utility are separate questions.

Exercise 10.5. Why do two derivative quotients enter the invariant-domain result?

Solution. Writing ϕ=F composed with h inverse gives ϕ prime=F prime/h prime in the uniformizing coordinate. Both numerator and denominator vary across the comparison set. □

Graphical reference. The shared construction is displayed in Volume I.

10.6 EP06. Corner rounding and observable convergence

10.6.1 Prerequisites and governing statement

A local singular field depends on the distance to a geometric corner. A finite rounding radius supplies a smallest relevant scale.

$$|E(r)| \asymp r^{\beta-1}, \qquad 0 < \beta < 1$$

The field grows as r decreases, while its energy integrand is of order r^2beta-1.

$$E_{\max} \asymp \rho^{\beta-1}, \qquad \mathcal{E}(0,\rho) \asymp \rho^{2\beta}.$$

10.6.2 Derivation and worked instance

The local model predicts how a physical cutoff rho controls the maximum. The omitted energy inside the cutoff has a different positive exponent. A mesh that resolves a smaller mathematical radius can increase its peak without indicating that the energy approximation has failed. To validate a device-level field maximum, the geometric rounding must be part of the physical input.

For beta=2/3, reducing the cutoff by a factor eight multiplies the ideal peak by two and reduces the local energy contribution by a factor sixteen.

10.6.3 Comparison and validity conditions

Baseline. Compare maximum field values on meshes with different implicit corner cutoffs.

Proposed procedure. State the physical rounding radius and compare potential, peak field and energy separately.

Quantitative conclusion. The model yields distinct, testable scaling laws for peak and integral observables.

Validity conditions. The proportionality constants depend on the global boundary-value problem. The exponent calculation alone does not determine an absolute breakdown voltage.

Exercise 10.6. Check the two scaling factors for beta=2/3 and rho replaced by rho/8.

Solution. $(1/8)^{-1/3}=2$ for the peak and $(1/8)^{4/3}=1/16$ for local energy. □

10.7 EP07. Multiport analytic models and finite-dimensional compactness

Mathematical reference. Volume I, normal-family and mapping core M02; source Little Montel; EE response-model extension.

10.7.1 Prerequisites and governing statement

A matrix-valued analytic function is holomorphic entrywise. In fixed finite dimension, uniform operator-norm bounds control every entry.

$$|S_{ij}(z)| \le \|S(z)\|_2$$

There are only finitely many entries when the port count is fixed.

$$\sup_{S,z\in K} \|S(z)\|_2 < \infty \quad \Longrightarrow \quad \text{a locally uniform matrix subsequence.}$$

10.7.2 Derivation and worked instance

Apply scalar Little Montel successively to the finitely many entries and take a diagonal subsequence. Entrywise uniform convergence on compact sets implies operator-norm convergence in fixed finite dimension. The limit retains analyticity. For infinite-dimensional operator targets, boundedness need not provide compactness; the identity-basis examples in Hilbert space supply the obstruction.

For a two-port model, four scalar subsequence selections suffice. A norm estimate ||A||2≤||A||F≤2 max|aij| translates entrywise errors into an operator error.

10.7.3 Comparison and validity conditions

Baseline. Generalize a scalar normal-family result to arbitrary operator-valued models without a dimension assumption.

Proposed procedure. Prove compactness entrywise for a fixed port count and record the infinite-dimensional boundary.

Quantitative conclusion. The finite-dimensional extension has an explicit dimension and norm-conversion constant.

Validity conditions. The dimension-dependent factor cannot be treated as independent of the number of ports.

Exercise 10.7. For an m by m error matrix with every entry bounded by η, give an operator-norm upper bound.

Solution. The Frobenius norm is at most sqrt(m^2 η^2)=m η, and the operator norm is at most the Frobenius norm. □

Graphical reference. The shared construction is displayed in Volume I.

10.8 EP08. A physical certificate and its semantic premises

10.8.1 Prerequisites and governing statement

A theorem about a mathematical field or transfer function applies to a device only after the model variables and assumptions have been identified.

$$\text{geometry} + \text{constitutive law} + \text{boundary data} + \text{representation}$$

Each item carries its own uncertainty and evidence.

$$\text{model validity} \ \wedge \ \text{analytic hypotheses} \ \wedge \ \text{numerical enclosure} \ \Rightarrow \ \text{stated observable bound.}$$

10.8.2 Derivation and worked instance

The field example with V=12 volts and ℓ=2 millimetres gives an exact interval for the declared quadratic-domain model. It has no independent evidence about a manufactured electrode's dimensions or dielectric composition.

10.8.3 Comparison and validity conditions

Baseline. Promote a mathematically valid analytic bound into a complete device guarantee.

Proposed procedure. Preserve the model-to-measurement dependencies in a typed verification record.

Quantitative conclusion. The final bound has a traceable semantic interpretation and a declared evidence boundary.

Exercise 10.8. Which additional evidence is needed to apply the ideal field bound to a manufactured part?

Solution. At minimum, validated geometry, voltage and boundary configuration, material constitutive assumptions, dimensional calibration, and an assessment of neglected physical effects. □

References

[1] Stanley H. Mneney. *An Introduction to Digital Signal Processing: A Focus on Implementation.* River Publishers edition.

[2] David Tse, Pramod Viswanath. *Fundamentals of Wireless Communication.* AZW3 contents: draft dated 10 September 2004.