

Vibrations of mechanical structures: source localization and nonlinear eigenvalue problems for mode calculation

Jonathan P. Baker

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Mathematics

Mark Embree, Chair
Christopher Beattie
Serkan Gugercin
Rodrigo Sarlo

May 4, 2023
Blacksburg, Virginia

Keywords: mechanical vibrations, sensor networks, Cramér-Rao bound, controllability
matrices, nonlinear eigenvalue problems

Copyright 2023, Jonathan P. Baker

Vibrations of mechanical structures: source localization and nonlinear eigenvalue problems for mode calculation

Jonathan P. Baker

(ABSTRACT)

This work addresses two primary topics related to vibrations in structures. The first topic is the use of a spatially distributed sensor network for localization of vibration events. I use a received signal strength (RSS) framework that presumes exponential energy decay with distance to the source. I derive the Cramér–Rao bound (CRB) for this parameter estimation problem, with the unknown parameters being source location, source intensity, and the energy dissipation rate. In this framework, I show that the CRB matches the variance of maximum likelihood estimators (MLEs) in more computationally expensive Monte Carlo trials. I also compare the CRB to the results of physical experiments to test the power of the CRB to predict spatial areas where MLEs show practical evidence of being ill-conditioned. Supported by this evidence, I recommend the CRB as a simple measure of localization accuracy, which may be used to optimize sensor layouts before installation. I demonstrate how this numerical optimization may be performed for some regions of interest with simple geometries.

The second topic investigates modal vibrations of multi-body structures built from simple one-dimensional elements, with networks of elastic strings as the primary example. I introduce a method of using a nonlinear eigenvalue problem (NLEVP) to express boundary conditions of the vibrating elements so that the (infinitely many) eigenvalues of the full structure are the eigenvalues of the finite-dimensional NLEVP. The mode shapes of the structure can then be recovered in analytic form (not as a discretization) from the corresponding eigenvectors of the NLEVP. I show some advantages of this method over

dynamic stiffness matrices, which is another NLEVP framework for modal analysis. In numerical experiments, I test several contour integration solvers for NLEVPs on sample problems generated from string networks.

Vibrations of mechanical structures: source localization and nonlinear eigenvalue problems for mode calculation

Jonathan P. Baker

(GENERAL AUDIENCE ABSTRACT)

This work deals with two primary topics related to vibrations in structures. The first topic is the use of vibration sensors to detect movement or impact and to estimate the location of the detected event. Sensors that are close to the event will record a larger amount of energy than the sensors that are farther away, so comparing the signals of several sensors can approximately establish the event location. In this way, vibration sensors might be used to monitor activity in a building without the use of intrusive cameras. The accuracy of location estimates can be greatly affected by the relative positions of the sensors and the event. Generally, location estimates tend to be most accurate if the sensors closely surround the event, and less accurate if the event is outside of the sensor zone. These principles are useful, but not precise. Given a framework for how event energy and noise are picked up by the sensors, the Cramér–Rao bound (CRB) is a formula for the achievable accuracy of location estimates. I demonstrate that the CRB is usefully similar to the location estimate accuracy from experimental data collected from a volunteer walking through a sensor-rigged hallway. I then show how CRB computations may be used to find an optimal arrangement of sensors. The match between the CRB and the accuracy of the experiments suggests that the sensor layout that optimizes the CRB will also provide accurate location estimates in a real building.

The other main topic is how the vibrations of a structure can be understood through the structure’s natural vibration frequencies and corresponding vibration shapes, called the *modes* of the structure. I connect vibration modes to the abstract framework of *nonlinear*

eigenvalue problems (NLEVPs). An NLEVP is a square matrix-valued function for which one wants to find the inputs that make the matrix singular. But these singular matrices are usually isolated—distributed among the infinitely many matrices of the NLEVP in places that are difficult to predict. After discussing NLEVPs in general and some methods for solving them, I show how the vibration modes of certain structures can be represented by the solutions of NLEVPs. The structures I analyze are *multi-body* structures that are made of simple interconnected pieces, such as elastic strings strung together into a spider web. Once a multi-body structure has been cast into the NLEVP form, an NLEVP solver can be used to find the vibration modes. Finally, I demonstrate that this method can be computationally faster than many traditional modal analysis techniques.

Acknowledgments

I owe special thanks to Pablo Tarazaga for acting as my co-advisor on matters of vibration engineering. Thanks also to Sa'ed Ajlouni, Rodrigo Sarlo, Murat Ambarkutuk, and other members of the VT Smart Infrastructure Lab for generously sharing their time and expertise.

I am thankful to the members of my dissertation committee—Christopher Beattie, Serkan Gugercin, and Rodrigo Sarlo—for the great improvements to my dissertation made possible by their questions and suggestions. Most of all, I owe great thanks to my advisor Mark Embree for endless wisdom and support far beyond my deserving.

This project was supported in part by the National Science Foundation Research Traineeship Program grant DGE-1545362 for UrbComp, a program supporting graduate training in topics of urban computing. I would especially like to thank Naren Ramakrishnan and Wanawsha Shalaby for their work to make this program possible.

This project also received support from the National Science Foundation grant DMS-1720257 and from the 2018 Deloitte Foundation Data Analytics Fellowship.

Contents

List of Figures	x
List of Tables	xvii
1 Introduction	1
1.1 Project origins and summary of contributions	3
2 Vibration event detection	5
2.1 Goodwin Hall Experiment	5
2.2 Detection	9
2.3 Energy transfer model	12
3 Sensor placement	16
3.1 Cramér-Rao bound for energy-based localization	17
3.1.1 CRB-based heuristics	21
3.2 Optimal placement	24
3.2.1 Scalar representation of the CRB	25
3.3 Sensor optimization examples	29
3.3.1 Numerical optimization of many sensors on a line segment	39
3.4 CRB and Monte Carlo	45

3.5	CRB and Experimental Data	46
3.6	Conclusion	50
4	Nonlinear eigenvalue problems	51
4.1	Introduction	51
4.2	Contour methods	54
4.2.1	Single shift methods	56
4.2.2	Controllability index	59
4.2.3	Multiple shifts	67
5	NLEVPs for vibrating structures	70
5.1	Introduction	70
5.2	Elastic Web Model	72
5.2.1	Code repository	79
5.2.2	Damped systems	80
5.3	Three-string example	86
5.4	Deriving modes of three-string example	90
5.5	Numerical studies of three-strings	97
5.5.1	Parameterized three-string model	97
5.5.2	Effect of user choices	101
5.6	Computation time study	111
5.7	Dynamic stiffness matrix method	112

5.7.1	DSM of three-string	115
5.7.2	Poles in DSMs	116
6	Conclusion	123
	Bibliography	126

List of Figures

2.1	Locations of footstep assignments and active accelerometers during a practical experiment. Volunteers each walked from “Step 1” to “Step 42” <i>and back</i> ten times. The large view (top) shows the locations of all 60 sensors used in my experiment (some of which are cospatial but differently oriented). The individually assigned footstep locations are more clearly visible in the detail view (bottom).	7
2.2	Raw acceleration signals from two Goodwin Hall accelerometers during the walking experiment.	8
3.1	Example of collinear sensors. With sensors in this configuration, a strong impact at location B will be difficult to distinguish from a weak impact at location A. The localization is always expected to be poorly conditioned for events outside of the convex hull of the sensors, but the collinearity makes the problem especially apparent.	23
3.2	Example sensor layout. Three clusters of sensors (dots) are noticeable. The localization problem may be well-conditioned for sources near the sensor clusters but ill-conditioned for sources near the point marked $\times$. The fact that the marked point is inside the convex hull of the sensor locations does not guarantee accurate location estimation.	25

3.3	Reciprocal of $\mathcal{I}_x$ plotted against position in a one-dimensional corridor for three difference sensor layouts. In each plot, the five dashed lines represent the positions of the sensors. Peaks in $\mathcal{I}^{-1}$ form when there are too few sensors nearby. The center layout balances the outer and inner peaks, minimizing the maximum $\mathcal{I}^{-1}$ value.	30
3.4	Reciprocal of $\mathcal{I}_x$ plotted against position in a one-dimensional corridor with two sensors marked by dashed lines.	32
3.5	Reciprocal of $\mathcal{I}_x$ plotted against position in a one-dimensional corridor with three sensors marked by dashed lines.	33
3.6	Numerically determined optimal sensor placements for a range of region sizes and three values of β . Each optimal sensor configuration is represented by a set of horizontally aligned circles on the same axes. Curved vertical lines show how the optimal position of each sensor varies (relative to the region length) as the region half-length (x_M) changes. In all cases, the region is scaled to $[-1, 1]$. The lowest row of points in each plot displays the optimal sensor layout for the shortest region (smallest x_M), and the top row of points corresponds to the longest region.	38
3.7	Two sketches of a square region of interest using different coordinate systems. Sensors are located on the perpendicular bisectors of the sides. Assuming that α and β are known, the local maxima of the CRB occur at the corners and the midpoints of the sides of the region.	41
3.8	Example of a square region with labelled distances between one example sensor (marked with a circle) and the locations of the local CRB maxima (marked with stars).	43

3.9	MC-generated STDEV of MLEs of x (top left) and y (top right) and corresponding CRBs (bottom left and right). Distances and STDEVs are measured in meters.	46
3.10	Histogram of relative errors between top and bottom portions of Figure 3.9.	47
3.11	MLE STDEVs of MC study vs MLE STDEVs for first walking experiment. The results of the MC study are very similar to the CRB, so this plot also approximates the relationship between the CRB and the STDEVs of the first walking experiment.	48
3.12	CRB vs MLE STDEVs for the second walking experiment. Each point is labeled with the corresponding footstep target location number: 1 on the west (left) and 42 on the east (right) end of the walking path shown in Figure 3.13.	49
3.13	Reproduction of Figure 2.1 with several labeled steps locations. In Figure 3.12, steps east of location 11 are associated with noticeably higher location estimator variance. There is a particularly steep rise in the CRB for step locations east of location 23.	50
5.1	An example web structure described formally by (5.5).	75
5.2	The first six mode shapes of a simple regular spider web.	80
5.3	Modes 7–12 of a simple regular spider web.	81

5.4	Top: Eigenvalues of the three-string problem drawn in the complex plane, with color indicating the damping level, c . Bottom: Imaginary (left) and real (right) parts of three-string eigenvalues as functions of c , with a different color used for each eigenvalue. As damping increases, conjugate pairs of (underdamped) eigenvalues meet on the real axis, and two real (overdamped) eigenvalues are formed.	84
5.5	Eigenvalue loci of a three-string for a range of damping values. Left: damping applies equally to all directions of movement. Right: damping applies only to transverse movement. The gray lines show the integration contour used to find the eigenvalues, and the blue and red dots show the left and right interpolation points. On the right, the unmoving sixth eigenvalue is circled for visibility. The eigenvalue loci extend slightly beyond the contour because some eigenvalues were found outside of the contour, but this may not happen in general.	86
5.6	The three-string network studied in this section.	87
5.7	The first ten mode shapes of an example three-string.	94
5.8	Mode shapes 11–20 of an example three-string.	95
5.9	Mode shapes 21–30 of an example three-string.	96
5.10	Natural frequencies of the three-string shown in Figure 5.11 for a range of ϕ_2 values. The bottom plot is a detailed view of the small region marked with a red rectangle in the top plot where there is a “near miss” between the 15th and 16th eigenvalues. Overlaid on the bottom plot are the mode shapes of both eigenvalues for $\phi_2 = 33^\circ$ and 38° . The shapes of the modes appear to “jump” between eigenvalues: the shape of the higher mode at $\phi_2 = 33^\circ$ resembles the shape of the lower mode at $\phi_2 = 38^\circ$ and vice versa.	99

5.11	An asymmetric three-string with parameterized angles.	100
5.12	A sequence of circular contours and the eigenvalues they enclose. The right-most extent (maximum abscissa) of each contour is shown just outside the circle. For each contour, the number of moments K^* needed for accurate Hankel-based integrals is shown in Table 5.1.	103
5.13	Upper left and right: For elliptical contours of a range of sizes, eigenvalue convergence as the number of quadrature points increases, plotted as points per unit length of the contour. Each line is labeled with the vertical extent (maximum imaginary part) of the Ω used in that convergence study. The extents greater than 1 are shown separately on the right to more clearly show that the relationship between contour size and accuracy is not monotone. Bottom: Boundaries of elliptical Ω choices with vertical extents 2, 6, 8, . . . , 20 (grey); eigenvalues within Ω (black); and eigenvalues outside Ω (red).	104
5.14	Left: eigenvalue convergence as the number of quadrature points increases, plotted as points per unit length of the contour. For each contour, the same numbers of quadrature points are used: the first and last graphed points correspond to 10 and 1000 quadrature points. Right: eigenvalues, contours, and interpolation point.	105
5.15	Single-shift method with interpolation at $\sigma = \infty$. Upper left: approximation error of selected eigenvalues as a function of the number of moments used in the Hankel matrix ($K^* = 4$ marked with a dashed line). Upper right: error of all eigenvalues when using $K^* = 4$ moments. Bottom: integration contour.	106

5.16	Single-shift method with interpolation at $\sigma = 10 + 0.4i$. Upper left: approximation error of selected eigenvalues as a function of the number of moments used in the Hankel matrix ($K^* = 4$ marked with a dashed line). Upper right: error of all eigenvalues at $K^* = 4$ moments ($\Re(\sigma) = 10$ marked with a dashed line). Bottom: integration contour and shift value.	107
5.17	Multi-shift method. Eigenvalue convergence for shift counts $p = 31, 32, 64$ corresponding to the line labels.	108
5.18	Multi-shift method. One of the contour-shift configurations used to generate Figure 5.19. Left (red) and right (blue) shifts alternate at constant vertical distances on either side of the contour.	109
5.19	Multi-shift method. Left: Quadrature convergence for selected contour-to-shift distances ($N = 1000$ quadrature points marked with a dashed line). Each convergence line is labelled with the distance between the contour and the rows of shifts. Right: Clarifying view comparing the accuracy of computed eigenvalues with different shift positions for $N = 1000$ quadrature points.	110

5.20	Left: Elapsed wall time and CPU time required to calculate eigenvalues accurately to six decimal places. Right: Lowest discretization levels required by the two methods to achieve six places of accuracy. The discretization level of pseudospectral collocation is the number of subdivisions used on each string, while the “discretization” of the NLEVP method is the number of quadrature points. The fact that these very different quantities are close enough in value to fit reasonably on a plot together is a coincidence. In these plots, all of the lines rising noticeably for high frequencies correspond to the pseudospectral collocation method. The other lines correspond to solving the NLEVP formulation with the multi-shift contour method. . . .	111
5.21	Fifth dynamic mode of a elastic string with an artificial subdivision point two-fifths of the way along its length. This point is a node of the fifth mode.	119
5.22	The DSM $D(\omega)$ for the single-string system with several L_1 values. In every plot, dots show the true eigenvalues (5.41). Upper left: $L_1 = L/2$. Upper right: $L_1 = L/4$. Lower left: $L_1 = L/6$. Lower right: $L_1 = L\sqrt{1.1}/2$. . .	122
5.23	Detail of the upper left plot of Figure 5.22 ($L_1 = L/2$). The DSM $D(\omega)$ does not have an eigenvalue at $\omega = 0$ for this or any other subdivision point position.	122

List of Tables

2.1	Summary of variables introduced in this section.	15
4.1	Calculation of bounds (4.14) and (4.19) for example (4.20).	68
5.1	Comparison of bounds on K^* given in Section 4.2.2 for the contours shown in Figure 5.12. For each contour, $\alpha(\Omega)$ is the rightmost extent of the domain Ω (maximum abscissa) and K^* is the minimality index of the invariant pair associated with the eigenvalues inside the contour. In each case, the identity matrix as used for Q	102

Chapter 1

Introduction

Most machines and structures experience vibrations any time they are subjected to forces. Vehicles have engines with rotating parts that impose oscillatory forces that naturally cause vibration. Even if the force is not oscillatory, any material that tends to restore itself to its original position will usually vibrate after being pushed or pulled because—without sufficient damping—the restoring force will over-correct. The excess displacement causes the material to begin restoring itself in the direction of the original force, and so on. The result is vibration.

Vibrations can be destructive, most noticeably when resonance causes a structure to vibrate violently. But even small vibrations can cause material fatigue over time, so it is important to know how a structure will vibrate when subjected to the types of forces that are typical in its environment.

Vibrations carry information, both deliberately and incidentally. The sound waves used when speaking and listening carry intentional messages. The loudness of thunder gives listeners some information about how far away the associated lightning strike was. Careful vibration measurements can be used to precisely locate the origin of a high-energy event, such as when seismologists find the epicenter of an earthquake by comparing when the vibration waves reached different locations. Vibrations can also reveal information about the vibrating material. When struck, some melons have a deeper sound when they are more ripe. In more sophisticated settings, vibration testing may be used as a method of

structural health monitoring.

It is useful to classify vibration phenomena into two categories: standing waves and traveling waves. For a familiar example of standing waves, consider any musical instrument. When the instrument is played, a standing wave of the primary note frequency is produced, either in the substance of the instrument (as in strings and percussion) or in the air of a tuned chamber (as in woodwinds and brass). Ocean waves are typical traveling waves, as their peaks and troughs move continuously in a definite direction.

Traveling waves are “periodic” when they span an entire medium (as in a uniformly oscillating ocean), and are “localized” if they are limited in space (as in ripples spreading from a central source). Localized traveling waves are also called “wave packets.” Chapters 2 and 3 deal with using measurements of wave packets to locate their source(s).

Chapters 4 and 5 focus on vibrations that make standing (i.e., non-traveling) waves. Standing waves have characteristic “mode shapes” that are determined by physical traits of the medium. If the dynamic mode shapes of an object are measured, then information about the object’s physical traits can be inferred. Conversely, if a researcher has information about an object’s physical traits, then the mode shapes may be predicted. Objects can even be designed to have specific desired modal characteristics. For example, it is sometimes desirable to design objects that vibrate at certain frequencies (e.g., musical instruments and klaxons). Alternatively, in order to avoid energy feedback that could lead to damage, it is also common to design objects (e.g., buildings and vehicles) in order to *avoid* particular resonant frequencies that would be excited by their environment. Modal analysis is the topic of Chapters 4 and 5. Chapter 4 discusses solvers for nonlinear eigenvalue problems, and Chapter 5 shows how these solvers may be used to predict the modes of multi-body structures made of simple components.

1.1 Project origins and summary of contributions

This project grew out of my interest in using the versatile vibration data collected in Virginia Tech’s Goodwin Hall, which is a living laboratory building of the College of Engineering. Besides housing engineering labs, offices, and classrooms, Goodwin Hall is permanently outfitted with accelerometers that respond to ambient vibrations produced by activity in the building. I got in contact with Virginia Tech’s Smart Infrastructure Laboratory (VT-SIL) in order to work with others who used the accelerometer data to study Goodwin Hall’s traveling waves (especially wave packets produced by pedestrian footsteps) and standing waves (such as the characteristic modal movements induced by wind).

My own investigations of pedestrian-induced vibrations and the companion problems of tracking and sensor location optimization appear in Chapters 2 and 3. These chapters have two main contributions. First, I developed a new method of event identification across multiple sensors especially for nearly-simultaneous events that are difficult to separate. This method fits wave packet travel paths to the data to find combinations of events that explain the detections that took place at the sensors. Second, I collected and studied physical experimental data to demonstrate how the conditioning of the event localization problem varies over space, comparing practical, theoretical, and Monte Carlo results. A portion of these results were published in [2].

When I took a course on vibrations from VT-SIL’s founder Pablo Tarazaga, I noticed the similarity between determinant-based calculations of beam modes and the nonlinear eigenvalue problems (NLEVPs) that I was also studying. This led me to put more general modal problems in this framework in a way that is somewhat different from existing techniques. Chapter 4 discusses NLEVPs in general and presents some new theoretical results for moment-based contour methods. Chapter 5 explains my NLEVP scheme for vibrating

member networks and uses such models for numerical studies of modern NLEVP solution methods.

Chapter 2

Vibration event detection

2.1 Goodwin Hall Experiment

The practical experiments in this work were all conducted in Goodwin Hall, which was formerly known as the Signature Engineering Building. Goodwin Hall is a seismically instrumented building on the Virginia Tech campus. Over 240 accelerometers are permanently mounted to the underfloor girders throughout Goodwin Hall's 170,000 gross square footage and five stories, making it one of the most densely seismically instrumented buildings in the world. These accelerometers detect small local movements of the structure caused by indoor events like pedestrian footsteps, elevator operation, doors opening and closing, and the building ventilation system. The accelerometers also register outdoor events like wind, nearby motor vehicle traffic, and even distant earthquakes.

The accelerometers in Goodwin Hall are manufactured by PCB Piezotronics, Inc., model 352B. Technical specifications of these sensors appear on the company website [38]. When these accelerometers are subjected to an out-of-plane force, their voltage signals respond by 1000 mV/g, up to ± 5 g. Positive and negative voltage changes indicate acceleration in opposite directions along the sensitive axis. The accelerometers can respond properly to vibrations with frequency between 2 and 10^4 Hz. In the absence of vibration forces, the sensors emit a (mean-zero Gaussian) white noise signal with standard deviation equivalent to 8×10^{-4} m/s². When vibration signals are driving the sensors, such a noise signal appears

additively in the recorded result.

Each accelerometer transmits its signal through a cable to one of four data acquisition systems (DAQs). Each DAQ is connected to a nearby cluster of sensors (e.g., one DAQ takes data from all accelerometers on the second floor). The DAQs are controlled by a central computer, which can activate any or all of the DAQs and store the digitized accelerometer readings on a hard drive. However, the file writing speed is limited, and if data is collected from many sensors at once, the sample rate must be low to avoid corrupted data. When running experiments, it is common to record only the data from sensors that are near the events of interest (on the same building floor, for example) so that a relatively high sample rate can be used. While the DAQ runs and experimenters cause events of interest, the vibration signature of the events are recorded. The vibrations received by the various sensors can then be studied after the fact.

I conducted an experiment in which a volunteer walked back and forth in a hallway on the fourth floor of Goodwin Hall while the DAQ recorded the vibration data collected by the sensors at a rate of 4096 Hz. Before the experiment, assigned footstep locations were marked on the floor with small pieces of tape. The assigned step locations and the positions of the active sensors are shown in Figure 2.1. While walking, the volunteer attempted to plant their feet on consecutive pieces of tape on consecutive steps. Consequently, there is a record of where each footstep occurred. The volunteer walked the length of the hallway 20 times (10 times in each direction).

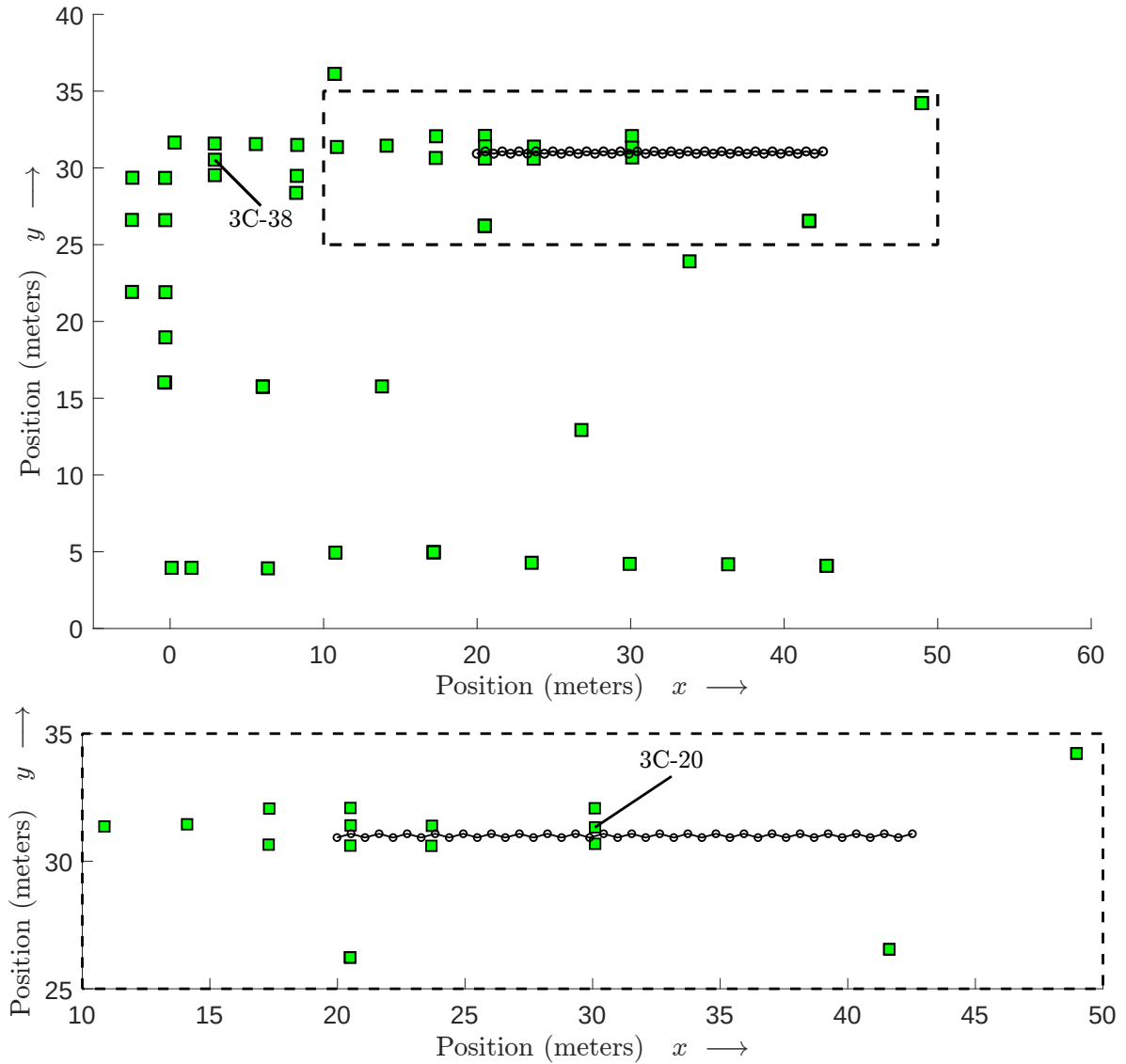


Figure 2.1: Locations of footstep assignments and active accelerometers during a practical experiment. Volunteers each walked from “Step 1” to “Step 42” *and back* ten times. The large view (top) shows the locations of all 60 sensors used in my experiment (some of which are cospatial but differently oriented). The individually assigned footstep locations are more clearly visible in the detail view (bottom).

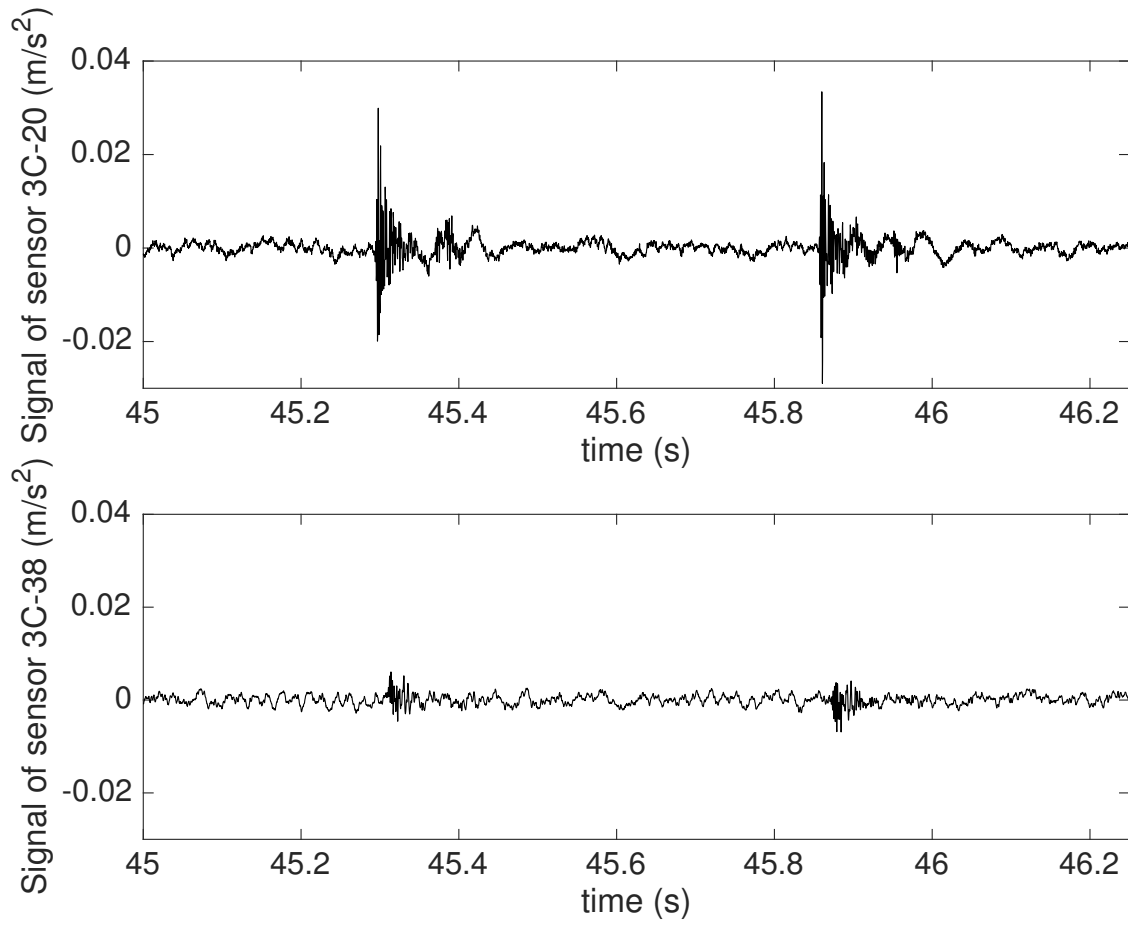


Figure 2.2: Raw acceleration signals from two Goodwin Hall accelerometers during the walking experiment.

Figure 2.2 shows raw acceleration signals from accelerometers “3C-20” and “3C-38” during the walking experiment. The positions of these two sensors are labeled in Figure 2.1. The vibrations caused by two consecutive footsteps are clearly visible in the prominent energy spikes in the signal of sensor 3C-20. Sensor 3C-38 also accelerated with the energy of these footsteps, but less intensely because sensor 3C-38 is farther from the impact than 3C-20 is. In both signals, relatively low-energy noise is visible in the time between footsteps. Received signal strength (RSS) localization methods use sudden changes in power to estimate sensor-to-source distances.

The impact waveforms also reach sensor 3C-20 slightly before they reach 3C-38. Time of arrival localization methods use such time differences to estimate sensor-to-source distances.

2.2 Detection

In order to locate the source of wave packets, the source event must register on the vibration sensors and cause their signals to be noticeably different from their baseline noise. Waves from impact events like footsteps typically appear in time series data as a rather sharp spike in energy that dissipates as the structure absorbs the energy during multiple oscillations. In this context, it is natural to choose a threshold and define an event as any signal sample that exceeds that threshold. However, it is not obvious how that threshold should be chosen or whether multiple nearby threshold-exceeding samples should be considered to be different events. Also, wave packets will arrive at the sensors at different times, so the energy peaks from different sensor signals must be matched in order to model them as originating with the same event. Similar issues are the focus of the field of event detection. Possibly the most frequent application of event detection is machine condition monitoring, where any change in the performance of moving parts may indicate

an imminent equipment failure [23, 31, 34].

For my event detection method, I use simple windowed energy threshold method to identify footstep events in the sensor time series. Use Δt to denote the inter-sample time (in seconds), so that the sample rate is $\frac{1}{\Delta t}$ Hz. In my the physical experiment, for example, $\Delta t = 1/4096$. For each sensor i , $i = 1, \dots, n$, use $x_i(k\Delta t)$ to denote the recorded acceleration at the k th sample time, $k\Delta t$. Use d_x to denote the width of time windows (in seconds) for combining sensor signals. I found that $d_x = 0.4$ was sufficient to enclose the noticeable energy contribution of experimental footsteps to each sensor signal. Use $M = d_x/\Delta t$ to denote the number of samples taken during a time window of length d_x . Finally, use σ_i for the standard deviation of the noise in the signal of sensor i . I chose to estimate each σ_i from the actual signal of sensor i rather than relying on the manufacturer-specified nominal noise level. This is because I consider ambient vibrations unrelated to the experiment to be noise, and an examination of the sensor signals revealed a noticeable difference in their realized noise levels.

Now define a new signal as the windowed sum of all sensor signals, each normalized by the sensor noise level:

$$s(k\Delta t) = \frac{1}{M} \sum_{i=1}^n \sum_{|k-j|\Delta t < d_x/2} \frac{|x_i(j\Delta t)|}{\sigma_i}.$$

Intuitively, $s(k\Delta t)$ is the sample-normalized total power appearing in any sensor signal near $k\Delta t$ in time. I consider s to be better than the simpler signal

$$\tilde{s}(t) = \frac{1}{M} \sum_{i=1}^n \frac{|x_i(t)|}{\sigma_i}$$

because the same event typically corresponds to peaks in different sensor signals that are slightly separated in time. When calculating $\tilde{s}$, such temporally offset peaks appear as a collection of individual peaks. When calculating s , temporally offset peaks are combined

into a single extra-prominent peak.

It remains to identify peaks in the combined signal s : one peak for each event. Use p_s to denote a threshold so that samples of s that exceed p_s correspond to events. It is not sensible to say that each sample of s that exceeds p_s corresponds to a separate event since the same event may produce many nearly-consecutive large terms in s . Instead, use a parameter d_s as a minimum time difference between detected events. I consider the times of detected events to be the times of the samples of the signal s that exceed the value p_s where the detection times that are separated by at least d_s seconds. Practically, this calculation can be performed by

1. reporting an event at the time of the highest sample in s ;
2. disqualifying all samples taken within d_s seconds of that sample;
3. repeating 1-2 until no eligible samples of s exceed the threshold p_s .

The range of sensible values for d_s depends on the frequency of footsteps. For example, typical walking pedestrian takes a step approximately every 0.5 seconds [18], but a sprinter may step every 0.2 seconds [35]. I used $d_s = 0.2$ (seconds), which was sufficient to avoid redundant event detections.

Choosing p_s is a more subtle matter since the energy recorded from an event that is far from most sensors may barely be noticeable above the noise. I tuned p_s manually to detect the maximum number of actual footsteps while not detecting any spurious events. This technique is used to detect the footstep events studied in Chapter 3.

2.3 Energy transfer model

This section describes the model of vibration detection introduced by Alajlouni in [1]. The model begins with a standard decomposition of the data into noise and a signal. Suppose power data is collected from sensor k for a duration of $M_k \Delta t$ seconds, beginning at time $J \Delta t$. The average power per sample received by the sensor is

$$Q_k = \frac{1}{M_k} \sum_{i=J}^{J+M_k} x_k(i \Delta t)^2. \quad (2.1)$$

The sensor signal $x_k(t)$ is presumed to be composed of a signal of interest $y_k(t)$ and an additive Gaussian white noise signal $\psi_k(t) \sim \mathcal{N}(0, \sigma_k^2)$:

$$x_k(t) = y_k(t) + \psi_k(t). \quad (2.2)$$

Substituting (2.2) into (2.1) gives

$$\begin{aligned} Q_k &= \frac{1}{M_k} \sum_{i=J}^{J+M_k} [y_k(i \Delta t) + \psi_k(i \Delta t)]^2 \\ &= \frac{1}{M_k} \left(\sum_{i=J}^{J+M_k} y_k(i \Delta t)^2 + \sum_{i=J}^{J+M_k} \psi_k(i \Delta t)^2 + 2 \sum_{i=J}^{J+M_k} y_k(i \Delta t) \psi_k(i \Delta t) \right). \end{aligned} \quad (2.3)$$

If M_k is large, the noise terms in (2.3) can be approximated with their expected values. For example, in my experiment $M_k = d_x / (\Delta t) = 0.4 / (2^{-12}) \approx 1638$ (the choice of how to round M_k to an integer number of samples appears inconsequential). Now we have

$$G_k := \frac{1}{M_k} \sum_{i=J}^{J+M_k} \psi_k(i \Delta t)^2 \approx \mathbb{E}(\psi_k^2) = \sigma_k^2, \quad (2.4a)$$

where ψ_k and y_k without time parameters denote generic samples. Also,

$$\begin{aligned} \frac{1}{M_k} \sum_{i=J}^{J+M_k} y_k(i\Delta t) \psi_k(i\Delta t) &\approx \mathbb{E}(y_k \psi_k) \\ &= \mathbb{E}(y_k) \mathbb{E}(\psi_k) = 0 \end{aligned} \quad (2.4b)$$

since y_k and ψ_k are assumed to be independent and $\mathbb{E}(\psi_k) = 0$.

It remains to relate the energy source to the average signal power term,

$$\alpha_k := \sum_{i=J}^{J+M_k} y_k(i\Delta t)^2. \quad (2.5)$$

Let (x_s, y_s) represent the location of the source (in two dimensions), let (x_k, y_k) represent the location of sensor k , and let $\Delta r_k := (x_k, y_k) - (x_s, y_s)$ denote the displacement vector from the location of the source to the location of sensor k . The (Euclidean) distance of from the source to sensor k is then $\|\Delta r_k\|$. An energy transfer model based on the inverse-square law (appropriate in a vibration medium spanning three dimensions) would resemble $\alpha_k = \frac{\alpha}{\|\Delta r_k\|^2}$, where α is an (initially unknown) parameter indicating the strength of the source. In an essentially two-dimensional medium like a floor, a linear decay expression might be expected instead: $\alpha_k = \frac{\alpha}{\|\Delta r_k\|}$. However, Alajlouni suggests using the exponential dissipation model

$$\alpha_k = \alpha \exp(\beta \|\Delta r_k\|), \quad (2.6)$$

where $\alpha > 0$ parameterizes the strength of the source and $\beta < 0$ parameterizes the rate at which power is lost over distance, due to both damping and the spatial spreading of the wavefront. Compared to the inverse-square law or linear decay models, a primary advantage of (2.6) is the lack of singularity as $\|\Delta r_k\|$ approaches zero. In the indoor context where sensors are placed in corridors, sources are expected to appear very near the sensors, and the resulting singularities of the rational models can be expected to harm

the conditioning of the localization problem.

Combining (2.3), (2.4b), and (2.6) gives $Q_k \approx \alpha_k + \sigma_k^2$, but Alajlouni goes on to model the distribution of Q_k using the distribution of the left-hand side of (2.4a) (called G_k). The stochastic effect of the term on the left-hand side of (2.4b) is ignored. Going forward, it will be assumed that

$$Q_k = \alpha_k + G_k. \quad (2.7)$$

Alajlouni derives $\mathbb{E}(G_k) = \sigma_k^2$ and $\text{var}(G_k) = 2\sigma_k^4/M_k$ [3, Claim 7.1.1]. Note that G_k is the empirical mean of a large number of independent identically distributed random variables. Assume that M_k is large enough that the distribution of G_k has converged satisfactorily toward a normal distribution by the central limit theorem, that is,

$$G_k \sim \mathcal{N}(\sigma_k^2, 2\sigma_k^4/M_k). \quad (2.8)$$

The definition of G_k in (2.4a) does not admit the possibility that $G_k < 0$, but under the modeling distribution (2.8), there is a positive probability that $G_k < 0$. However, the probability that $G_k < 0$ is vanishingly small under the assumption that M_k is large enough for G_k to have converged in distribution sufficiently.

Since α_k is considered to be deterministic, (2.7) and (2.8) give

$$Q_k \sim \mathcal{N}(\alpha_k + \sigma_k^2, 2\sigma_k^4/M_k). \quad (2.9)$$

As was just argued for G_k , the probability of the (impossible) event that $Q_k < 0$ is vanishingly small for large M_k .

The variables defined to this point that will continue to be used are summarized in Table 2.1.

Table 2.1: Summary of variables introduced in this section.

Q_k	Power per sample of sensor k
α_k	Signal (i.e., source-originated) power per sample of sensor k (unknown)
σ_k^2	Variance of noise for sensor k
(x_s, y_s)	Source location (unknown)
(x_k, y_k)	Location of sensor k
Δr_k	Spatial displacement from source to sensor k (unknown)
M_k	Number of samples in signal k used to calculate power (fixed)
α	Source intensity (unknown)
β	Power attenuation parameter (unknown)

In Chapter 3, it will be convenient to use a transformed version of the data that has a standard normal distribution. To that end, define

$$W_k = \frac{Q_k - \alpha_k - \sigma_k^2}{\sqrt{2/M_k \sigma_k^2}} = \frac{Q_k - \alpha \exp(\beta \|\Delta r_k\|) - \sigma_k^2}{\sqrt{2/M_k \sigma_k^2}}, \quad (2.10)$$

and note that $W_k \sim \mathcal{N}(0, 1)$. While the data Q_k is observed from actual sensor readings, W_k is not computed in practice since finding α_k requires knowledge of the source location (x_s, y_s) , intensity α , and dissipation rate β , all of which are typically unknown.

Now collect the power samples for all sensors Q_k into a single data vector $\mathbf{Q}$, and similarly collect the values W_k into a vector $\mathbf{W}$. Notice that, due to the independence of the noise across sensors, $\mathbf{W}$ has unit covariance:

$$\mathbf{W} \sim \mathcal{N}(0, I). \quad (2.11)$$

Chapter 3 will use the stochastic model of this section to derive lower bounds for variances of estimators of the unknown values x_s , y_s , α , and β .

Chapter 3

Sensor placement

This chapter is concerned with the conditioning of the source localization problem, i.e., the accuracy with which an energy source may be located. I will focus on received signal strength (RSS) as a means of estimating distances from the source to the sensors, and I will explore the Cramér-Rao bound (CRB) as a means of studying how the conditioning of the localization problem is affected by the placement of the sensors and the location of the source.

Section 3.1 introduces the CRB and applies it to the RSS framework of Alajlouni [1] (see Section 2.3). Based on the functional forms of the CRB for the RSS localization problem, Section 3.1.1 shows how the CRB suggests some heuristics for how the relative positions of the source and sensors affect localization accuracy. Section 3.2 further formalizes the CRB as an accuracy criterion. Section 3.3 gives examples of using the CRB to optimize sensor layouts in several simple scenarios. Sections 3.4 and 3.5 justify the use of the CRB as a measure of conditioning by comparing the CRB to the results of Monte Carlo samples and practical experiments.

Portions of Section 3.1, 3.4, and 3.5 were previously published in [2] in collaboration with Sa'ed Alajlouni and Pablo Tarazaga.

3.1 Cramér-Rao bound for energy-based localization

This chapter is concerned with how the placement of sensors affects the amount of information provided by their signals. I will use Fisher's model of information in the form of the Fisher information matrix (FIM) [20]. Cramér [16] and Rao [41] are known for interpreting the FIM as the covariance of *efficient* (minimum variance) parameter estimators. For syntactic convenience, this chapter will sometimes refer to $\mathcal{I}^{-1}$ (or its norm) as the CRB, even though the CRB is an inequality rather than a value.

Consider first the general parameter estimation problem. Say $\mathbf{X} \in \mathbb{R}^n$ is a random vector with a distribution that depends on a parameter $\theta \in \mathbb{R}^p$. Write the parameterized probability density function (PDF) of $\mathbf{X}$ as $f(x; \theta)$. If f is twice differentiable in θ , then the $p \times p$ FIM $\mathcal{I}$ of $\mathbf{X}$ can be defined entrywise in two equivalent ways:

$$\mathcal{I}_{i,j}(\theta) := \mathbb{E} \left[\frac{\partial}{\partial \theta_i} \log(f(x; \theta)) \cdot \frac{\partial}{\partial \theta_j} \log(f(x; \theta)) \right] \quad (3.1a)$$

$$= -\mathbb{E} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(f(x; \theta)) \right]. \quad (3.1b)$$

The definition of $\mathcal{I}$ guarantees that it is positive semi-definite, but it is possible for $\mathcal{I}$ to be singular.

If θ denotes the true parameter value, $\tilde{\theta}(\mathbf{X})$ is an unbiased estimator, and $\mathcal{I}$ is non-singular, then the well-known Cramér-Rao lower bound (CRB) for the variance of $\tilde{\theta}(\mathbf{X})$ is

$$\text{cov}_{\theta}(\tilde{\theta}(\mathbf{X})) \geq \mathcal{I}^{-1}, \quad (3.2)$$

meaning that the difference matrix $\text{cov}_{\theta}(\tilde{\theta}(\mathbf{X})) - \mathcal{I}^{-1}$ is positive semi-definite. In particular, the diagonal entries $(\mathcal{I}^{-1})_{i,i}$ ($i = 1, \dots, p$) are lower bounds for the variances of individual scalar estimators $\tilde{\theta}_i(\mathbf{X})$. It is usually desirable to use estimators with low vari-

ance, but low-variance estimators can only exist if $\mathcal{I}^{-1}$ (or its diagonal entries) are small. Importantly, this is a necessary but not sufficient condition. Good Fisher information does not guarantee that any particular estimator will have low variance, and there may not exist any unbiased estimator with covariance $\mathcal{I}^{-1}$.

If it is possible to control what data is collected, it makes sense to design the experiment to minimize $\mathcal{I}^{-1}$ in order to make it possible that a low-variance unbiased estimator exists. This chapter will recommend arranging vibration sensors with the objective of minimizing $\mathcal{I}^{-1}$, expecting that the estimates of vibration source locations will have lower variance, i.e., be more accurate.

In the context of source localization, the unknown parameters of Section 2.3 can be collected into the parameter vector.

$$\theta = \begin{bmatrix} x_s & y_s & \alpha & \beta \end{bmatrix}^T.$$

The Gaussian PDF of the random variable $\mathbf{W}$ is

$$f(\mathbf{w}) = \frac{1}{2\pi} e^{-\mathbf{w}^T \mathbf{w} / 2}, \quad (3.3)$$

where the dependence on θ is implied by the transformation (2.10). Lowercase letters $\mathbf{w}$, w_k , $\mathbf{q}$, and q_k will be used to distinguish possible realizations from the random variables $\mathbf{W}$, etc. Taking a log and partial derivative of (3.3),

$$\begin{aligned} \frac{\partial}{\partial \theta_i} \log f(\mathbf{w}; \theta) &= -\frac{1}{2} \frac{\partial}{\partial \theta_i} \mathbf{w}^T \mathbf{w} \\ &= -\sum_k w_k \frac{\partial}{\partial \theta_i} w_k. \end{aligned}$$

Next, compute $\frac{\partial}{\partial \theta_i} w_k$ for each i , that is, $\theta_1 = x_s$, $\theta_2 = y_s$, $\theta_3 = \alpha$, and $\theta_4 = \beta$:

$$\begin{aligned}
\frac{\partial}{\partial x_s} w_k &= \frac{\partial}{\partial x_s} \frac{q_k - \sigma_k^2 - \alpha e^{\beta \|\Delta r_k\|}}{\sqrt{2/M_k \sigma_k^2}} \\
&= -\alpha \frac{\sqrt{M_k/2}}{\sigma_k^2} e^{\beta \|\Delta r_k\|} \beta \frac{\partial}{\partial x_s} \|\Delta r_k\| \\
&= -\alpha \frac{\sqrt{M_k/2}}{\sigma_k^2} e^{\beta \|\Delta r_k\|} \beta \frac{\partial}{\partial x_s} \sqrt{(x_s - x_k)^2 + (y_s - y_k)^2} \\
&= -\alpha \frac{\sqrt{M_k/2}}{\sigma_k^2} e^{\beta \|\Delta r_k\|} \beta \frac{x_s - x_k}{\|\Delta r_k\|} \\
\frac{\partial}{\partial y_s} w_k &= -\alpha \frac{\sqrt{M_k/2}}{\sigma_k^2} e^{\beta \|\Delta r_k\|} \beta \frac{y_s - y_k}{\|\Delta r_k\|} \\
\frac{\partial}{\partial \alpha} w_k &= \frac{-\sqrt{M_k/2}}{\sigma_k^2} e^{\beta \|\Delta r_k\|} \\
\frac{\partial}{\partial \beta} w_k &= \frac{-\sqrt{M_k/2}}{\sigma_k^2} \alpha \|\Delta r_k\| e^{\beta \|\Delta r_k\|}.
\end{aligned}$$

Notice that the derivatives $\frac{\partial}{\partial \theta_i} w_k$ do not depend on the value of w_k . This is expected since $\mathbf{W}$ is an affine (shifted and scaled) transformation of the data $\mathbf{Q}$. Consequently, $\frac{\partial}{\partial \theta_i} W_k$ is a constant “random” variable, and it will be replaced with $\frac{\partial}{\partial \theta_i} w_k$ as needed in the next calculation.

Now, the entries of the FIM $\mathcal{I}$ can be expressed as

$$\begin{aligned}
\mathcal{I}_{i,j} &= \mathbb{E} \left[\left(\frac{\partial}{\partial \theta_i} \log(f(x; \theta)) \right) \left(\frac{\partial}{\partial \theta_j} \log(f(x; \theta)) \right) \right] \\
&= \mathbb{E} \left[\left(-\sum_k W_k \frac{\partial}{\partial \theta_i} W_k \right) \left(-\sum_k W_k \frac{\partial}{\partial \theta_j} W_k \right) \right] \\
&= \sum_k \left(\mathbb{E}[W_k^2] \frac{\partial}{\partial \theta_i} w_k \frac{\partial}{\partial \theta_j} w_k \right) + \sum_{k \neq \ell} \left(\mathbb{E}[W_k W_\ell] \left(\frac{\partial}{\partial \theta_i} w_k \frac{\partial}{\partial \theta_j} w_\ell + \frac{\partial}{\partial \theta_i} w_\ell \frac{\partial}{\partial \theta_j} w_k \right) \right) \\
&= \sum_k \left(\frac{\partial}{\partial \theta_i} w_k \frac{\partial}{\partial \theta_j} w_k \right).
\end{aligned}$$

The last expression above takes advantage of the fact that $\text{cov}(\mathbf{W}) = I$ by construction, and so $\mathbb{E}[W_k^2] = 1$ for $k = 1, \dots, n$ and $\mathbb{E}[W_k W_\ell] = 0$ for $k \neq \ell$. This FIM can be written in another convenient way,

$$\mathcal{I} = J^T J,$$

where $J = D_\theta \mathbf{W}$ is the Jacobian of $\mathbf{W}$ as a function of θ . The i th row of J is

$$J_i = \frac{-\sqrt{M_i/2}}{\sigma_i^2} \alpha e^{\beta \|\Delta r_i\|} \left[\beta \frac{x_s - x_i}{\|\Delta r_i\|}, \quad \beta \frac{y_s - y_i}{\|\Delta r_i\|}, \quad 1/\alpha, \quad \|\Delta r_i\| \right] \quad (3.4)$$

and

$$\mathcal{I} = \frac{\alpha^2}{2} \begin{bmatrix} \mathcal{I}_{x,y} & \mathcal{I}_* \\ \mathcal{I}_*^T & \mathcal{I}_{\alpha,\beta} \end{bmatrix}, \quad (3.5)$$

where

$$\mathcal{I}_{x,y} = \beta^2 \sum_{i=1}^n \frac{M_i}{\sigma_i^4} e^{2\beta \|\Delta r_i\|} \frac{1}{\|\Delta r_i\|^2} \begin{bmatrix} (x_s - x_i)^2 & (x_s - x_i)(y_s - y_i) \\ (x_s - x_i)(y_s - y_i) & (y_s - y_i)^2 \end{bmatrix}, \quad (3.6a)$$

$$\mathcal{I}_* = \beta \sum_{i=1}^n \frac{M_i}{\sigma_i^4} e^{2\beta \|\Delta r_i\|} \begin{bmatrix} \frac{x_s - x_i}{\alpha \|\Delta r_i\|} & x_s - x_i \\ \frac{y_s - y_i}{\alpha \|\Delta r_i\|} & y_s - y_i \end{bmatrix}, \quad (3.6b)$$

$$\mathcal{I}_{\alpha,\beta} = \sum_{i=1}^n \frac{M_i}{\sigma_i^4} e^{2\beta \|\Delta r_i\|} \begin{bmatrix} 1/\alpha^2 & \|\Delta r_i\|/\alpha \\ \|\Delta r_i\|/\alpha & \|\Delta r_i\|^2 \end{bmatrix}. \quad (3.6c)$$

If one is allowed to freely choose the locations of sensors, sensor arrangements could be designed to make the footstep localization accurate by maximizing the amount of information conveyed by the data, that is, by making the eigenvalues of $\mathcal{I}$ large for (x_s, y_s) in the area of interest. Equivalently, it is desirable for J to have large singular values (i.e., nearly orthogonal columns of large magnitude). If instead the eigenvalues of $\mathcal{I}$ are small, then the eigenvalues of $\mathcal{I}^{-1}$ are large, and (3.2) requires that any unbiased estimator $\tilde{\theta}(\mathbf{X})$ have large covariance. In practice, an estimator with large covariance should be considered

unreliable since there is a relatively high probability of the estimator differing greatly from the true parameter.

3.1.1 CRB-based heuristics

Heuristic arguments about how J and $\mathcal{I}$ relate to the sensor layouts yield some general principles that can guide sensor placement. Use $\lambda_1 \geq \dots \geq \lambda_4 \geq 0$ to denote the eigenvalues of $\mathcal{I}$.

- Localization accuracy may be poor if most or all of the sensors are far from the source. A far-distant source makes each $\alpha_i = \alpha e^{\beta \|\Delta r_i\|} / \sigma_i^2$ small (recall that $\beta < 0$). From (3.4), it is apparent that if every distance $\|\Delta r_i\|$ is sufficiently large, the first three columns of J will be small in magnitude. Asymptotically, the magnitude of the last column of J also shrinks when $\|\Delta r_i\| \rightarrow \infty$ for all i , and thus $J \rightarrow 0$. So, for far away sources, $\|J\|^2 = \lambda_1$ is small.
- Localization accuracy may be poor if all of the sensors and the source lie in a straight line. Mathematically, this situation is represented by the existence of a real γ such that for every i ,

$$y_s - y_i = \gamma(x_s - x_i). \quad (3.7)$$

In other words, the first two columns of J differ only by a factor of γ . Since two columns of J are linearly dependent, J is column-rank deficient and has at least one zero singular value, $\sqrt{\lambda_4} = 0$.

- If a straight line only comes close to containing the sensor positions and the source, then (3.7) is good approximation: the first two columns of J are *nearly* aligned, λ_4 is *small*, and $\mathcal{I}$ is *nearly* singular. Consequently, unbiased location estimators based on nearly collinear sensor layouts are bound to have high variance.

- More specifically, nearly collinear sensor placement produces high variance in the direction orthogonal to the line of sensors. Without loss of generality, suppose that all sensors *and* the source lie approximately in a horizontal line ($y_s \approx y_1 \approx \dots \approx y_n$). Then the second column of J is approximately zero and $\mathcal{I}_{2,2} \approx 0$. Since $\mathcal{I}$ is positive definite, Horn and Johnson [29, p. 472] find that for any conformal unit vector b ,

$$(b^* \mathcal{I} b)(b^* \mathcal{I}^{-1} b) \geq 1. \quad (3.8)$$

In particular if b is the second column of the identity matrix, (3.8) shows that $(\mathcal{I}^{-1})_{2,2}(\mathcal{I}_{2,2}) \geq 1$. Consequently $\sqrt{(\mathcal{I}^{-1})_{2,2}}$ is large, and any unbiased estimator for y_s will have high variance. The coordinate system is arbitrary and a similar argument applies to any orientation of nearly collinear sensors: when laid out in a line, a sensor network has poor resolution in the direction orthogonal to that line.

- The preceding argument about collinear sensors was essentially about the leading 2×2 submatrix of $\mathcal{I}$ corresponding to position (x and y), making no reference to α or β . Consequently, even if α and β were known (and $\mathcal{I}$ were 2×2), collinear sensors produce a localization problem that is poorly conditioned in the orthogonal direction. But when source intensity α is unknown, the estimated source location along the length of the line also has high variance if the source is “outside” of the sensor interval. Intuitively, if the source intensity is unknown and the source location is collinear with all sensors, it is impossible to tell how far the signal traveled before it reached the first sensor, so localization is not feasible. Continue considering the case where all sensors and the source lie along the same line: $y_s = y_1 = \dots = y_n$. Then every entry of the first column of J is the same, namely

$$\frac{-\sqrt{M_i/2}}{\sigma_i^2} \alpha e^{\beta \|\Delta r_i\|} \alpha \beta \frac{x_s - x_i}{\sqrt{(x_s - x_i)^2 + 0^2}} = \frac{-\sqrt{M_i/2}}{\sigma_i^2} \alpha e^{\beta \|\Delta r_i\|} \operatorname{sgn}(x_s - x_i) \alpha \beta.$$

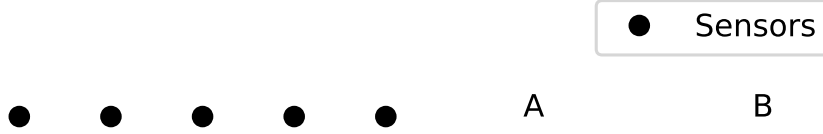


Figure 3.1: Example of collinear sensors. With sensors in this configuration, a strong impact at location B will be difficult to distinguish from a weak impact at location A. The localization is always expected to be poorly conditioned for events outside of the convex hull of the sensors, but the collinearity makes the problem especially apparent.

If the source lies to the left or right of all sensors, then $x_s - x_i$ will have the same sign for all i , and the first and third column of J differ only by a factor of $(\alpha\beta)$. Then J is column-rank deficient and $\mathcal{I}$ is singular. Figure 3.1 illustrates this problem with an example sensor layout and a pair of possible source locations that would be difficult to distinguish.

- Now consider a source outside the convex hull of the sensor positions. Without loss of generality, say $x_s - x_i > 0$ for all i . Then all entries of the first, third, and fourth columns of J have the same sign. While it is possible for a collection of vectors in the same orthant of $\mathbb{R}^n$ to be nearly orthogonal, they are typically more closely aligned with each other than with other randomly selected vectors the signs of whose entries vary freely. It is therefore likely that these columns are fairly closely aligned, which suggests that J is not far from column-rank deficient, and that J have a small singular value, $\sqrt{\lambda_4} \approx 0$.
- Adding a sensor to an existing sensor network increases Fisher information and decreases the CRB (possibly admitting unbiased estimators with lower variance). To see this, say that r^T is a row of the form (3.4) corresponding to the new sensor. Then the FIM of the augmented sensor network is a rank-one perturbation of $\mathcal{I}$,

specifically

$$\tilde{\mathcal{I}} = \begin{bmatrix} J^T & r \end{bmatrix} \begin{bmatrix} J \\ r^T \end{bmatrix} = J^T J + r r^T = \mathcal{I} + r r^T \geq \mathcal{I}.$$

The Sherman–Morrison formula for the inverse of such a matrix is

$$\tilde{\mathcal{I}}^{-1} = \mathcal{I}^{-1} - \frac{1}{1 + r^T \mathcal{I} r} \mathcal{I}^{-1} r r^T \mathcal{I}^{-1} \leq \mathcal{I}^{-1}.$$

In summary, the Fisher information (generally) supports common wisdom: for accurate localization, sensors should be near and surrounding the area of interest, and even a poorly placed sensor is better than no sensor.

3.2 Optimal placement

If a fixed region needs to be monitored and one has the opportunity to freely arrange a fixed number of sensors, where should the sensors be placed? One might think that the sensors should be installed around the perimeter of the region since other authors have observed that localization is typically only reliable within the convex hull of a sensor layout, e.g., [46, 51]. Yet it is a mistake to suppose that localization will be similarly accurate at *all locations* within the convex hull of a sensor layout. This is particularly apparent in some extreme cases such as the one illustrated in Figure 3.2. In this example, several groups of sensors are closely spaced, but there is such a distance between the groups that sources in the middle of the layout are indistinguishable from noise.

The previous section ended with heuristic arguments for some general principles that should guide sensor placement. This section explores the computation of optimal sensor arrangements for localization in fixed regions with simple geometry.

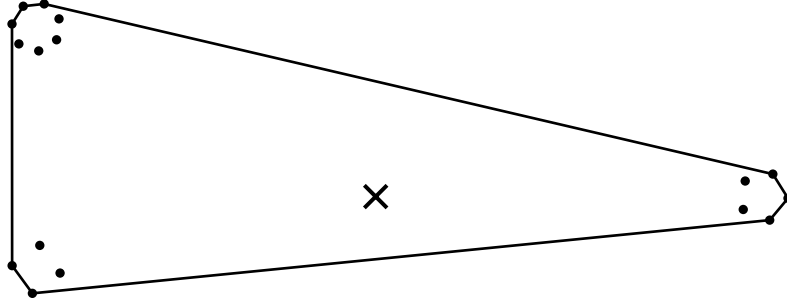


Figure 3.2: Example sensor layout. Three clusters of sensors (dots) are noticeable. The localization problem may be well-conditioned for sources near the sensor clusters but ill-conditioned for sources near the point marked $\times$. The fact that the marked point is inside the convex hull of the sensor locations does not guarantee accurate location estimation.

3.2.1 Scalar representation of the CRB

Positive definiteness only defines a partial ordering of matrices. It is only completely clear that the Fisher information at one location (or set of parameters) is better than another if the difference of the Fisher information matrices is a definite matrix. If the difference is indefinite, then the implications for the CRBs are not immediately apparent. For a concrete example, suppose that θ_A , θ_B , $\theta_C \in \mathbb{R}^3$ are possible parameter vectors in some parameter estimation problem, and that the corresponding Fisher information matrices are

$$\mathcal{I}_A = \mathcal{I}(\theta_A) = \begin{bmatrix} 10 & & \\ & 10 & \\ & & 10 \end{bmatrix},$$

$$\mathcal{I}_B = \mathcal{I}(\theta_B) = \begin{bmatrix} 5 & & \\ & 5 & \\ & & 5 \end{bmatrix}, \quad \mathcal{I}_C = \mathcal{I}(\theta_C) = \begin{bmatrix} 5 & & \\ & 5 & \\ & & 20 \end{bmatrix}.$$

Since $\mathcal{I}_A > \mathcal{I}_B$ (meaning $\mathcal{I}_A - \mathcal{I}_B$ is positive definite), the corresponding CRBs permit estimators of all three parameter entries to have lower variance when the true parameter is θ_A than when it is θ_B . Based on Fisher information, it appears to be unambiguously

easier to estimate the parameter when the true value is θ_A than when it is θ_B . Intuitively speaking, observations provide “more” information about θ_A than about θ_B .

On the other hand, the difference $\mathcal{I}_A - \mathcal{I}_C$ is neither positive nor negative definite. The CRB corresponding to $\mathcal{I}_A$ allows lower-variance estimators of the first two scalar parameters in θ than the CRB corresponding to $\mathcal{I}_C$ does, but the reverse is true of the third parameter. Based on Fisher information, there is an ambiguity in whether it is easier to estimate the parameter when the true value is θ_A than when it is θ_C . If it is necessary to make an evaluation of $\mathcal{I}_A$ versus $\mathcal{I}_C$, a problem-specific value judgment is needed.

Importantly, notice that the CRB is only a lower bound, and particular estimators may have higher variance. It may be that for some estimator $\tilde{\theta}$,

$$\text{cov}_{\theta_A}(\tilde{\theta}(\mathbf{X})) > \text{cov}_{\theta_B}(\tilde{\theta}(\mathbf{X}))$$

even though

$$\mathcal{I}(\theta_A)^{-1} < \mathcal{I}(\theta_B)^{-1}.$$

While the FIM may sometimes fail to predict the ordering of the variances of a particular estimator under different true parameter values, Monte Carlo experiments in Section 3.4 indicate that the CRB is an excellent predictor of the variance of the maximum likelihood location estimator under the exponential energy decay model presented in Section 2.3.

In order to resolve the problem of non-comparable Fisher information matrices, one might try to map the matrices to scalars. Obvious candidates for scalar measures of the strength of information include $\|\mathcal{I}^{-1}\|$ or the condition number $\|\mathcal{I}\|\|\mathcal{I}^{-1}\|$. However, the choice of units affects the eigenvalues of $\mathcal{I}$ unevenly, so the preferences established by scalar measures of information could be altered by the choice of units.

Covariance matrices are not dimensionless. The dimension (units) of an entry of a co-

variance matrix is the product of the dimensions of the corresponding variables. So, when comparing covariance, Fisher information, and Cramér-Rao bounds for different parameter values, it is important that the same units are used.

I suggest that a useful scalar measure of information (in the 2-dimensional localization context) is the matrix 2-norm of the upper 2×2 block of $\mathcal{I}^{-1}$. The 2-norm of this matrix provides a bound on the variance of the Euclidean error of location estimators. To be precise, use the block notation in (3.5) and write the Schur complement-based form of $\mathcal{I}^{-1}$ (see, for example, [29, p. 18]) as

$$\begin{aligned} \mathcal{I}^{-1} &= \begin{bmatrix} (\mathcal{I}_{x,y} - \mathcal{I}_* \mathcal{I}_{\alpha,\beta}^{-1} \mathcal{I}_*^T)^{-1} & -(\mathcal{I}_{x,y} - \mathcal{I}_* \mathcal{I}_{\alpha,\beta}^{-1} \mathcal{I}_*^T)^{-1} \mathcal{I}_* \mathcal{I}_{\alpha,\beta}^{-1} \\ -(\mathcal{I}_{\alpha,\beta} - \mathcal{I}_*^T \mathcal{I}_{x,y}^{-1} \mathcal{I}_*)^{-1} \mathcal{I}_*^T \mathcal{I}_{x,y}^{-1} & (\mathcal{I}_{\alpha,\beta} - \mathcal{I}_*^T \mathcal{I}_{x,y}^{-1} \mathcal{I}_*)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} (\mathcal{I}_{x,y} - \mathcal{I}_* \mathcal{I}_{\alpha,\beta}^{-1} \mathcal{I}_*^T)^{-1} & 0 \\ 0 & (\mathcal{I}_{\alpha,\beta} - \mathcal{I}_*^T \mathcal{I}_{x,y}^{-1} \mathcal{I}_*)^{-1} \end{bmatrix} \begin{bmatrix} I & -\mathcal{I}_* \mathcal{I}_{\alpha,\beta}^{-1} \\ -\mathcal{I}_*^T \mathcal{I}_{x,y}^{-1} & I \end{bmatrix}, \end{aligned} \quad (3.9)$$

assuming that the appropriate matrices are nonsingular. Call the upper left 2×2 submatrix

$$\mathcal{E} := (\mathcal{I}_{x,y} - \mathcal{I}_* \mathcal{I}_{\alpha,\beta}^{-1} \mathcal{I}_*^T)^{-1}. \quad (3.10)$$

The matrix $\mathcal{E}$ provides an analog to the CRB of θ that is specific to the location components. Specifically, if $\tilde{x}$ and $\tilde{y}$ denote the x and y components of the unbiased estimator $\tilde{\theta}$,

then for any $b \in \mathbb{R}^2$ with $\|b\| = 1$,

$$\begin{aligned}
 \text{var}_\theta \left(b^T \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \right) &= \mathbb{E}_\theta \left[b^T \left(\begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} - \begin{bmatrix} x_s \\ y_s \end{bmatrix} \right) \left(\begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} - \begin{bmatrix} x_s \\ y_s \end{bmatrix} \right)^T b \right] \\
 &= \mathbb{E}_\theta \left[\begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix}^T (\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T \begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix} \right] \\
 &= \begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix}^T \text{cov}_\theta(\tilde{\theta}) \begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix}.
 \end{aligned}$$

Applying the CRB now gives

$$\text{var}_\theta \left(b^T \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \right) \geq \begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix}^T \mathcal{I}^{-1} \begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix} = b^T \mathcal{E} b. \quad (3.11)$$

The lower bound $b^T \mathcal{E} b$ achieves its maximum when b is the dominant eigenvector of $\mathcal{E}$ so that

$$\text{var}_\theta \left(b^T \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \right) \geq b^T \mathcal{E} b = \|\mathcal{E}\|,$$

where $\|\mathcal{E}\|$ denotes the matrix 2-norm. In other words, there is a geometric direction b in which the variance of any unbiased location estimator must be at least $\|\mathcal{E}\|$. Consequently, it seems reasonable to use $\|\mathcal{E}\|$ as a scalar measure of localization accuracy.

It is not difficult to find a lower bound for $\|\mathcal{E}\|$ that involves only $\mathcal{I}_{x,y}$. Inverting both sides

of (3.10) and neglecting the negative definite term gives

$$\mathcal{E}^{-1} = \mathcal{I}_{x,y} - \mathcal{I}_* \mathcal{I}_{\alpha,\beta}^{-1} \mathcal{I}_*^T \leq \mathcal{I}_{x,y}.$$

Since $\|A\| \|A^{-1}\| \geq 1$ for any nonsingular matrix A , we also have

$$\|\mathcal{E}\| \geq \|\mathcal{I}_{x,y}^{-1}\| \geq \frac{1}{\|\mathcal{I}_{x,y}\|}.$$

One might think of $\|\mathcal{E}\| - \|\mathcal{I}_{x,y}^{-1}\|$ as the information cost of not knowing α and β . This is because, if α and β are known, then $\mathcal{I}$ is only the upper left 2×2 block $\mathcal{I} = \mathcal{I}_{x,y}$. Consequently, $\|\mathcal{E}\| = \|\mathcal{I}_{x,y}^{-1}\|$.

3.3 Sensor optimization examples

For general problems, it is difficult to produce formulas for optimal sensor layouts. But for a one-dimensional setting with just two or three sensors and a known energy decay rate β , one can prove optimality of a sensor layout in terms of the CRB.

Throughout this section, I take for granted that an optimal sensor layout exists. For the scenarios in one spatial dimension, assume that the optimal layout is symmetric about the center of the segment, as in the example in Figure 3.3. In each example, an optimal layout minimizes the maximum value of the geometric portion of $\mathcal{I}^{-1}$, that is $\mathcal{I}_x^{-1}$ for one-dimensional problems and $\|\mathcal{I}_{x,y}^{-1}\|$ in two dimensions. For simplicity in this section, assume that all sensors have the same noise STDEV σ and the same number of power calculation samples M .

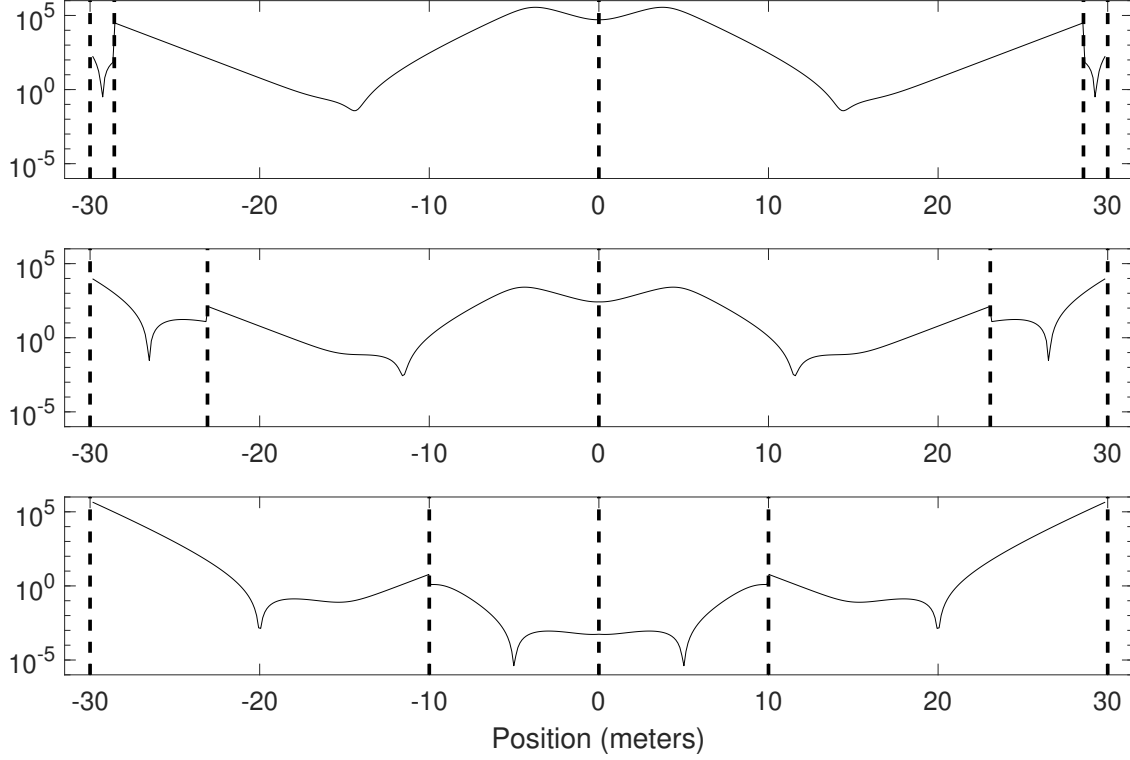


Figure 3.3: Reciprocal of $\mathcal{I}_x$ plotted against position in a one-dimensional corridor for three difference sensor layouts. In each plot, the five dashed lines represent the positions of the sensors. Peaks in $\mathcal{I}^{-1}$ form when there are too few sensors nearby. The center layout balances the outer and inner peaks, minimizing the maximum $\mathcal{I}^{-1}$ value.

Known intensity α and decay β , two sensors

The simplest sensor placement problem is to choose the locations of two sensors on a line segment when α and β are known beforehand and the source location x is constrained to the segment $[-x_M, x_M]$. We seek the positions for a pair of sensors that minimize the maximum $\mathcal{I}^{-1}$ value on the segment (the minimax or uniform minimization problem).

Moving the sensors farther apart increases the CRB at the local maximum at $x = 0$, and moving them closer together increases the CRB at the local maxima at the boundaries. Consequently, the sensor layout that minimizes the maximum CRB causes these local maxima to be equal. Since only the step location $x_s \in [-x_M, x_M]$ is unknown, $\mathcal{I}$ and $\mathcal{I}^{-1}$

are scalars, and this equal maxima condition can be written as

$$\mathcal{I}(0) = \mathcal{I}(x_M). \quad (3.12)$$

When only the source location x_s is unknown, J is simply the column vector

$$J = \frac{-\sqrt{M/2}}{\sigma^2} \alpha \beta \begin{bmatrix} \frac{x_s - x_1}{|x_s - x_1|} e^{\beta|x_s - x_1|} \\ \frac{x_s - x_2}{|x_s - x_2|} e^{\beta|x_s - x_2|} \end{bmatrix},$$

and so

$$\mathcal{I}(x_s) = J^T J = \frac{M}{2\sigma^4} \alpha^2 \beta^2 (e^{2\beta|x_s - x_1|} + e^{2\beta|x_s - x_2|}).$$

When this expression is used to expand condition (3.12), we get

$$\frac{M}{2\sigma^4} \alpha^2 \beta^2 (e^{2\beta|-x_1|} + e^{2\beta|-x_2|}) = \frac{M}{2\sigma^4} \alpha^2 \beta^2 (e^{2\beta|x_M - x_1|} + e^{2\beta|x_M - x_2|}).$$

The assumption of layout symmetry provides that $x_1 = -x_2$, so without loss of generality, $x_1 \geq 0$ and $x_2 = -x_1$. Combining like terms in the above equation and dividing by the like constant produces

$$2e^{2\beta x_1} = e^{2\beta(x_M - x_1)} + e^{2\beta(x_M + x_1)}.$$

This expression may be divided by $e^{2\beta(x_M + x_1)}$ to get

$$2e^{-2\beta x_M} = e^{-4\beta x_1} + 1$$

and solved for x_1 :

$$x_1 = \frac{-1}{4\beta} \ln(2e^{-2\beta x_M} - 1).$$

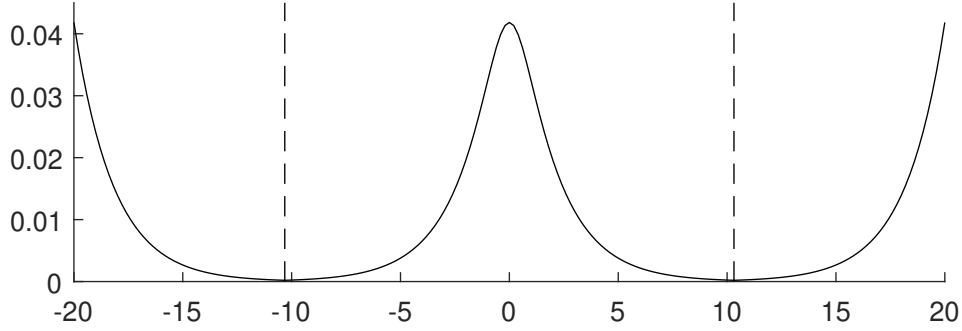


Figure 3.4: Reciprocal of $\mathcal{I}_x$ plotted against position in a one-dimensional corridor with two sensors marked by dashed lines.

For large x_M , x_1/x_M approaches $1/2$ (so that no point is farther from a sensor than $x_M/2$). Figure 3.4 shows an example of this scenario with $x_M = 20$. Placing sensors near the positions -10 and 10 balances the heights of the central $\mathcal{I}^{-1}$ peak and the peaks at the ends of the intervals, thereby minimizing the maximum peak height.

Known intensity α and decay β , three sensors

Suppose three sensors are positioned in the order $x_3 \leq x_2 \leq x_1$. The assumption of symmetry, two of the sensors will be symmetric about the origin ($x_3 = -x_1$) and the other will be at the origin ($x_2 = 0$). As Figure 3.5 shows, the inverse FIM will have a local maxima between each pair of adjacent sensors and at each end-point. Let x_p be the point between x_2 and x_1 where one of the interior maxima occurs ($0 = x_2 \leq x_p \leq x_1$). By symmetry, another local maximum appears at $-x_p$. Because α and β are known, $\mathcal{I}$ is a scalar. The first-order condition for a local maximum of $\mathcal{I}^{-1}$ at x_p is equivalent to the local minimum condition for $\mathcal{I}$, namely $\mathcal{I}'(x_p) = 0$.

Let $C = \frac{M\alpha^2\beta^2}{2\sigma^4}$. For the general case of n sensors in one dimension,

$$\mathcal{I}(x) = C \sum_{i=1}^n e^{2\beta|x-x_i|}, \quad (3.13)$$

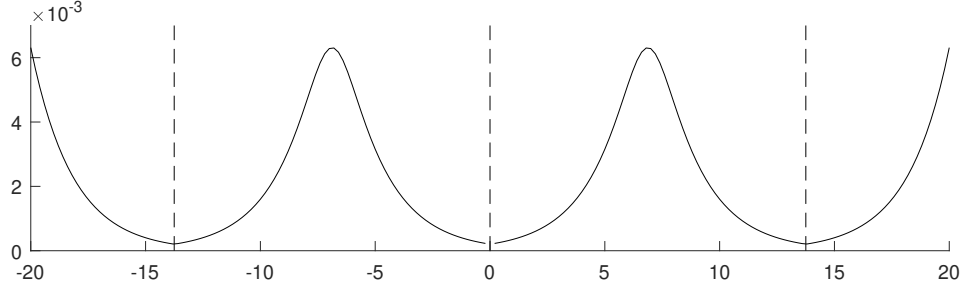


Figure 3.5: Reciprocal of $\mathcal{I}_x$ plotted against position in a one-dimensional corridor with three sensors marked by dashed lines.

and

$$\mathcal{I}'(x) = 2C\beta \sum_{i=1}^n \text{sgn}(x - x_i) e^{2\beta|x-x_i|},$$

so in the $n = 3$ case being considered,

$$0 = \mathcal{I}'(x_p) = 2C\beta \sum_{i=1}^3 \text{sgn}(x_p - x_i) e^{2\beta|x_p-x_i|}.$$

Dividing by $2C\beta$ and expanding the sum,

$$0 = e^{2\beta(x_p+x_1)} + e^{2\beta x_p} - e^{2\beta(x_1-x_p)}. \quad (3.14)$$

This may be divided by $e^{2\beta(x_p+x_1)}$ to get

$$0 = 1 + e^{-2\beta x_1} - e^{-4\beta x_p},$$

which may be solved for x_p :

$$x_p = \frac{-1}{4\beta} \ln(e^{-2\beta x_1} + 1).$$

Now expand (3.13) for three sensors to get

$$\mathcal{I}(x_p) = C(e^{2\beta(x_p+x_1)} + e^{2\beta x_p} + e^{2\beta(x_1-x_p)}),$$

and use (3.14) to make the substitution $e^{2\beta(x_p+x_1)} + e^{2\beta x_p} = e^{2\beta(x_1-x_p)}$ to find

$$\begin{aligned} \mathcal{I}(x_p) &= 2Ce^{2\beta(x_1-x_p)} \\ &= 2Ce^{2\beta\left[x_1 - \frac{-1}{4\beta} \ln(e^{-2\beta x_1} + 1)\right]} \\ &= 2Ce^{2\beta x_1} \left(e^{\ln(e^{-2\beta x_1} + 1)}\right)^{1/2} \\ &= 2Ce^{2\beta x_1} \sqrt{e^{-2\beta x_1} + 1}. \end{aligned} \tag{3.15}$$

The minimax condition requires

$$\mathcal{I}(x_p) = \mathcal{I}(x_M),$$

and substituting in (3.13) and (3.15) produces

$$2Ce^{2\beta x_1} \sqrt{e^{-2\beta x_1} + 1} = C(e^{2\beta(x_M+x_1)} + e^{2\beta x_M} + e^{2\beta(x_M-x_1)}).$$

For convenience, let $y := e^{2\beta x_1}$ so that

$$2y\sqrt{1/y + 1} = e^{2\beta x_M}(y + 1 + 1/y).$$

Remove the radical by squaring to get

$$4y + 4y^2 = e^{4\beta x_M}(y^2 + 1 + 1/y^2 + 2y + 2 + 2/y),$$

and then multiply by y^2 and combine like terms to obtain the polynomial

$$4y^3 + 4y^4 = e^{4\beta x_M} (y^4 + 2y^3 + 3y^2 + 2y + 1).$$

Now define $\gamma := e^{4\beta x_M}$ and notice that, since $\beta < 0$ and $x_M > 0$, $\gamma \in (0, 1)$. Rewriting the polynomial as

$$0 = (4 - \gamma)y^4 + (4 - 2\gamma)y^3 - 3\gamma y^2 - 2\gamma y - \gamma \quad (3.16)$$

makes it simple to use Descartes' rule of signs [4]:

The number of positive real roots of a polynomial is no more than the number sign changes in the sequence of nonzero coefficients when listed in order of degree. Furthermore, the difference between the number of positive real roots and the number of sign changes is even.

The possible range of γ values does not allow the signs of any coefficients to vary, and the sequence of signs is always $+, +, -, -, -$. Since there is only one sign change, (between the third and second order coefficients), Descartes' rule of signs reveals that (3.16) must have exactly $1 - 2k$ positive real roots for some $k \geq 0$. The number of roots may not be negative, so $k = 0$ and, for any $\gamma \in (0, 1)$, (3.16) has exactly one positive real root. Since $y := e^{2\beta x_1}$ must be positive, the relevant solution exists and is unique. The positive real solution y of (3.16) may be found numerically, and the optimal location for sensor 1 is $x_1 = \ln(y)/(2\beta)$. Placing the other sensors at $x_2 = 0$ and $x_3 = -x_1$ completes the optimal layout.

Known decay β and unknown intensity α

So far, the calculations in this section have assumed that both α and β are known. Since there has been only one spatial dimension in this section, $\mathcal{I}$ has been a scalar. This subsection treats the one-dimensional problem with known β but unknown α . This scenario corresponds to the case where the floor material is understood well enough that the power dissipation parameter β does not need to be estimated each time an event is located. In this case of two unknowns ($x_s \in [-x_M, x_M]$ and α), $\mathcal{I}$ is a 2×2 matrix. The i th row of J as a function of x_s is

$$J_i(x_s) = \frac{-\sqrt{M/2}}{\sigma^2} e^{\beta|x_s-x_i|} \begin{bmatrix} \alpha\beta \operatorname{sgn}(x_s - x_i) & 1 \end{bmatrix} \quad (3.17)$$

so that, with $C = M/(2\sigma^4)$,

$$\mathcal{I}(x_s) = J(x_s)^T J(x_s) = C \begin{bmatrix} \sum_i e^{2\beta|x_s-x_i|} & \frac{1}{\alpha\beta} \sum_i \operatorname{sgn}(x_s - x_i) e^{2\beta|x_s-x_i|} \\ \frac{1}{\alpha\beta} \sum_i \operatorname{sgn}(x_s - x_i) e^{2\beta|x_s-x_i|} & \frac{1}{\alpha^2\beta^2} \sum_i e^{2\beta|x_s-x_i|} \end{bmatrix}.$$

It will be convenient to define $\ell_i \equiv \ell_i(x_s) := e^{2\beta|x_s-x_i|}$. The CRB for estimators of x_s is the top left entry of $\mathcal{I}^{-1}$, which is

$$\begin{aligned} (\mathcal{I}^{-1})_{1,1} &= \frac{\mathcal{I}_{2,2}}{\det(\mathcal{I})} \\ &= \frac{\frac{1}{\alpha^2\beta^2} \sum_i e^{2\beta|x_s-x_i|}}{\frac{1}{\alpha^2\beta^2} (\sum_i e^{2\beta|x_s-x_i|})^2 - \frac{1}{\alpha^2\beta^2} (\sum_i \operatorname{sgn}(x_s - x_i) e^{2\beta|x_s-x_i|})^2} \\ &= \frac{\sum_i \ell_i}{\sum_i \ell_i^2 + 2 \sum_{x_i < x_j} \ell_i \ell_j - \left(\sum_i \ell_i^2 + 2 \sum_{x_i < x_j} \operatorname{sgn}(x_s - x_i) \operatorname{sgn}(x_s - x_j) \ell_i \ell_j \right)} \\ &= \frac{\sum_i \ell_i}{4 \sum_{x_i < x_s < x_j} \ell_i \ell_j}. \end{aligned} \quad (3.19)$$

From (3.19), the CRB is not defined for locations that are outside of the sensor interval because, for such locations, the denominator of (3.19) is zero (an empty sum) and

therefore $\mathcal{I}$ is singular.

When optimizing (3.19), it is useful to notice that the denominator is piecewise constant for x_s values between adjacent sensor locations, i.e., in (x_i, x_{i+1}) . To see this, first notice that for any sensor locations $x_i < x_j$ and any $x_s \in (x_i, x_j)$,

$$\begin{aligned}\ell_i(x_s)\ell_j(x_s) &= e^{2\beta(|x_s-x_i|+|x_s-x_j|)} \\ &= e^{2\beta(x_s-x_i-(x_s-x_j))} \\ &= e^{2\beta(x_j-x_i)},\end{aligned}$$

which is independent of x_s . Also, for any locations $\hat{x}_1, \hat{x}_2$ that are between adjacent sensors, the expressions $x_i < \hat{x}_1 < x_j$ and $x_i < \hat{x}_2 < x_j$ apply to the same sensor pairs (i, j) , therefore

$$\sum_{x_i < \hat{x}_1 < x_j} \ell_i(\hat{x}_1)\ell_j(\hat{x}_1) = \sum_{x_i < \hat{x}_2 < x_j} \ell_i(\hat{x}_2)\ell_j(\hat{x}_2).$$

Each ℓ_i is piecewise convex since for $x_s \neq x_i$,

$$\ell'_i(x_s) = 2\beta \operatorname{sgn}(x_s - x_i)e^{2\beta|x_s-x_i|}, \quad \text{and} \quad \ell''_i(x_s) = 4\beta^2 e^{2\beta|x_s-x_i|} > 0.$$

Consequently the sum $\sum_i \ell_i(x_s)$ is convex on intervals that do not include sensor locations. Therefore, $\mathcal{I}_{1,1}^{-1}$ is piecewise convex between sensors, and the maximum value of $\mathcal{I}_{1,1}^{-1}$ must occur in the limit as the source approaches a sensor, or at the region boundary. Anyone seeking to numerically optimize many sensors while knowing β beforehand might use this fact to reduce the cost of searching for the maximum $\mathcal{I}_{1,1}^{-1}$ value for a each candidate sensor layout.

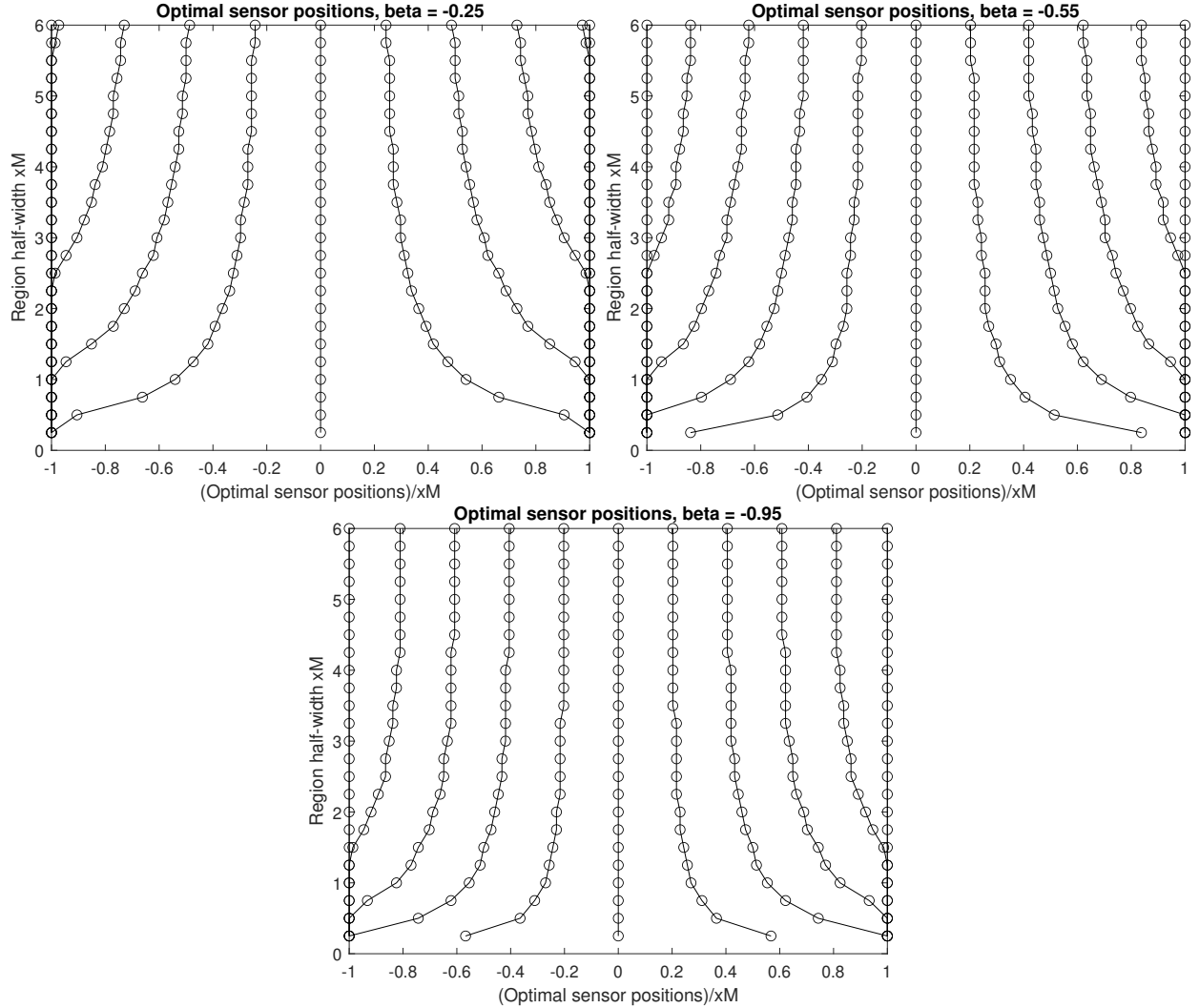


Figure 3.6: Numerically determined optimal sensor placements for a range of region sizes and three values of β . Each optimal sensor configuration is represented by a set of horizontally aligned circles on the same axes. Curved vertical lines show how the optimal position of each sensor varies (relative to the region length) as the region half-length (x_M) changes. In all cases, the region is scaled to $[-1, 1]$. The lowest row of points in each plot displays the optimal sensor layout for the shortest region (smallest x_M), and the top row of points corresponds to the longest region.

3.3.1 Numerical optimization of many sensors on a line segment

I optimized sensor locations numerically for a range of interval sizes and several values of β . I consider a sensor layout optimal if it minimizes the maximum CRB value on the interval.

Figure 3.6 shows optimal sensor locations for a range of region lengths and three β values. For large regions, sensors should be approximately evenly spaced. As the region gets smaller, optimal sensor locations move toward the boundaries. In fact, when optimizing over the discrete set of candidate sensor positions in this study, the boundary points themselves were reported as optimal for multiple sensors in many cases. For small enough intervals, it appears that half of the sensors should be placed very close to each boundary point, with a sensor in the center if n is odd.

Figure 3.6 also shows that, for a fixed region length, decreasing β makes optimal sensor layouts more evenly spaced. So decreasing β and lengthening the region of interest have similar effects on the optimal sensor layout. This is not surprising since decreasing β and stretching the region both result in greater signal power attenuation over a fixed proportion of the interval.

I would like to offer an intuitive explanation for why corridor length and power dissipation rate affect the optimal sensor layout as they appear to in Figure 3.6. In a long (or heavily damped) corridor, it is important to maximize the chance that at least one or two sensors “hear” the event clearly, like sentinels guarding a border. Evenly spaced sensors minimize the distance that an event could possibly be from the nearest sensor. On the other hand, if the region is short (or very lightly damped), then energy travels relatively freely from one end to the other. Every sensor will “hear” the event clearly, and comparing how much power reaches each end of the interval is the most reliable localization method.

A grid search was used to generate Figure 3.6, and the discretization is visible in the

“kinks” in the near-vertical lines of the figure. A simple local optimization can be used to improve the solution slightly, which makes the lines more smooth.

For the parameter range of used in the figure, a near-singularity can form at the interval end-points. When the sensors near an end-point of the interval are too close together, the CRB increases sharply at the outermost sensor location. I consider that this small area with a high CRB value is not typically relevant to experiments, so this value was ignored when determining the maximum CRB value on the interval when generating Figure 3.6. If this extreme end-point value is not ignored, other sensors should be moved even closer to the edges to reduce the CRB value at the end-points.

Figure 3.6 shows that it can be (at least approximately) optimal for multiple sensors to be placed at each end-point. It may seem as though duplicate sensors at the same location would provide no additional information. However, the noise recorded by each sensor is independent, so the estimate of the signal power at the shared sensor location is improved by additional sensors providing alternative (noisy) power readings. The effect of adding sensors at the same location is comparable to decreasing the noise level of a single sensor.

Four sensors on a square, known intensity α and decay β

Now consider optimal placement of four sensors for locating sources in a square region. Continue to assume that α and β are known. For an optimality criterion, minimize $\|\mathcal{I}^{-1}\|$. Suppose that the centers of the sides of the region are at a distance x_M from the region center at the origin. For example, the corners of the region might be (x_M, x_M) , $(-x_M, x_M)$, $(-x_M, -x_M)$, and $(x_M, -x_M)$. However, it will be convenient to consider another orientation as well.

I conducted several *ad hoc* numerical experiments, which suggested that optimal layouts have sensors on the orthogonal bisectors of the sides. For the example orientation just

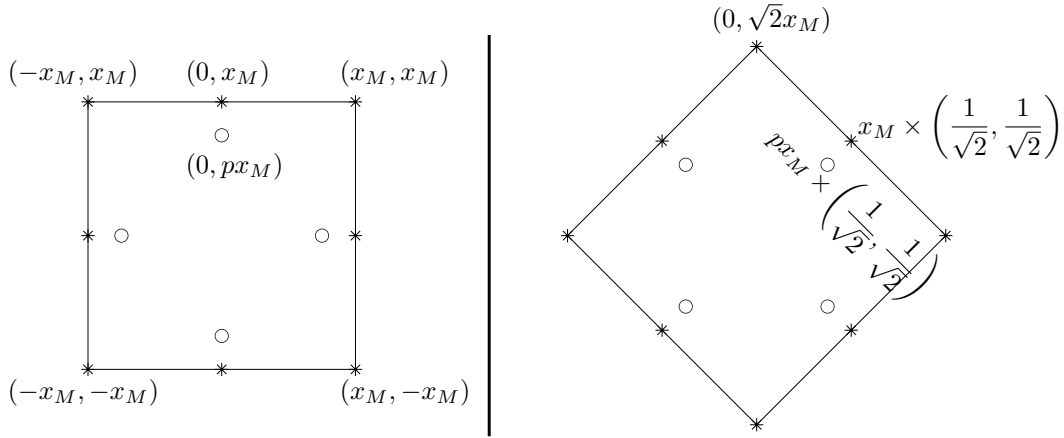


Figure 3.7: Two sketches of a square region of interest using different coordinate systems. Sensors are located on the perpendicular bisectors of the sides. Assuming that α and β are known, the local maxima of the CRB occur at the corners and the midpoints of the sides of the region.

given, this means the sensors are placed at $(px_M, 0)$, $(0, px_M)$, $(-px_M, 0)$, and $(0, px_M)$ for some scaling factor $p > 0$. The goal of this section is to explore a means of determining the value of p that optimizes such a layout. Perhaps counter-intuitively, the CRB at a point may decrease by moving all sensors farther away.

For such layouts, the CRB has local maxima at the corners and side midpoints of the region boundary. These locations of the local maxima are shown in Figure 3.7.

The CRB at the corner points shrinks as p grows, but the CRB at the side midpoints shrinks as p shrinks. Consequently, the optimal p value is the one that causes the CRB to have equal values at the corners and side midpoints.

Using the region layout described above, number the sensors in counterclockwise order, starting with the one on the right. The CRB at all four side-center points will be the same. Consider specifically the point in the middle of the top side, $(0, x_M)$. The distances from all side midpoints to a single sensor are illustrated in Figure 3.8.

Because of rotational symmetry, the distances from the edge midpoints to a single sensor are the same as the distances from a single edge midpoint to all sensors. Specifically, the

distances from $(0, x_M)$ to the sensors are, in counterclockwise order, d_1, d_3, d_5, d_3 , where

$$\begin{aligned} d_1 &= x_M(1 - p) \\ d_3 &= x_M\sqrt{p^2 + 1} \\ d_5 &= x_M(1 + p). \end{aligned}$$

While assuming that α and β are known (and that only x_s, y_s are unknown), only the first two columns of (3.4) are used to find J . Again assume that all sensors have the same noise STDEV σ and the same number of power calculation samples M . It is convenient to define $C = \frac{-\sqrt{M/2}}{\sigma^2}\alpha\beta$ so that J can be written as

$$J(0, x_M) = Cx_M \begin{bmatrix} \frac{0-p}{d_3}e^{\beta d_3} & \frac{1-0}{d_3}e^{\beta d_3} \\ \frac{0-0}{d_1}e^{\beta d_1} & \frac{1-p}{d_1}e^{\beta d_1} \\ \frac{0+p}{d_3}e^{\beta d_3} & \frac{1-0}{d_3}e^{\beta d_3} \\ \frac{0-0}{d_5}e^{\beta d_5} & \frac{1+p}{d_5}e^{\beta d_5} \end{bmatrix} = Cx_M \begin{bmatrix} \frac{-p}{d_3}e^{\beta d_3} & \frac{1}{d_3}e^{\beta d_3} \\ 0 & \frac{1}{x_M}e^{\beta d_1} \\ \frac{p}{d_3}e^{\beta d_3} & \frac{1}{d_3}e^{\beta d_3} \\ 0 & \frac{1}{x_M}e^{\beta d_5} \end{bmatrix}.$$

The FIM is then

$$\mathcal{I}(0, x_M) = J(0, x_M)^T J(0, x_M) = C^2 x_M^2 \begin{bmatrix} 2\frac{p^2}{d_3^2}e^{2\beta d_3} & 0 \\ 0 & 2\frac{1}{d_3^2}e^{2\beta d_3} + \frac{1}{x_M^2}(e^{2\beta d_1} + e^{2\beta d_5}) \end{bmatrix}.$$

Therefore, $\mathcal{I}(0, x_M)$ is diagonal, and it is easy to observe that the smaller eigenvalue is the upper left entry. Since α and β are assumed to be known, the suggested scalar proxy of the FIM is

$$\|\mathcal{E}\| = \|\mathcal{I}(0, x_M)^{-1}\| = \frac{d_3^2}{2C^2 x_M^2 p^2 e^{2\beta d_3}} = \frac{p^2 + 1}{2C^2 p^2 e^{2\beta d_3}}.$$

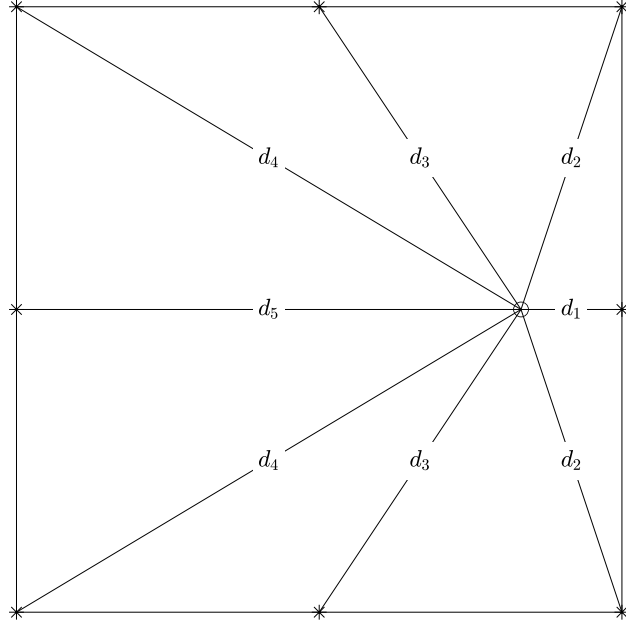


Figure 3.8: Example of a square region with labelled distances between one example sensor (marked with a circle) and the locations of the local CRB maxima (marked with stars).

Consider now the local CRB maxima at the corners of the region. The rotated coordinate system on the right-hand side of Figure 3.7 will cause the FIM at each corner to be a convenient diagonal matrix. In this coordinate system, the corners of the square are $(0, x_M\sqrt{2})$, $(-x_M\sqrt{2}, 0)$, $(0, -x_M\sqrt{2})$, and $(x_M\sqrt{2}, 0)$, and the sensors are located at $(px_M/\sqrt{2}, -px_M/\sqrt{2})$, $(px_M/\sqrt{2}, px_M/\sqrt{2})$, $(-px_M/\sqrt{2}, px_M/\sqrt{2})$, and $(-px_M/\sqrt{2}, -px_M/\sqrt{2})$. Take the corner at $(x_M\sqrt{2}, 0)$ as the example. As illustrated in Figure 3.8, the distances from this corner to the sensors are d_2 , d_2 , d_4 , and d_4 , where

$$d_2 = x_M \sqrt{(1-p)^2 + 1}$$

$$d_4 = x_M \sqrt{(1+p)^2 + 1}.$$

In the next calculation, a tilde indicates that the rotated reference frame is being used.

Then we have

$$\tilde{J}(0, x_M/\sqrt{2}) = C \frac{x_M}{\sqrt{2}} \begin{bmatrix} \frac{-p}{d_2} e^{\beta d_2} & \frac{2-p}{d_2} e^{\beta d_2} \\ \frac{p}{d_2} e^{\beta d_2} & \frac{2-p}{d_2} e^{\beta d_2} \\ \frac{-p}{d_4} e^{\beta d_4} & \frac{2+p}{d_4} e^{\beta d_4} \\ \frac{p}{d_4} e^{\beta d_4} & \frac{2+p}{d_4} e^{\beta d_4} \end{bmatrix},$$

and thus

$$\begin{aligned} \tilde{\mathcal{I}}(0, x_M/\sqrt{2}) &= \tilde{J}(0, x_M/\sqrt{2})^T \tilde{J}(0, x_M/\sqrt{2}) \\ &= C^2 x_M^2 \begin{bmatrix} \frac{p^2}{d_2^2} e^{2\beta d_2} + \frac{p^2}{d_4^2} e^{2\beta d_4} & 0 \\ 0 & \frac{(2-p)^2}{d_2^2} e^{2\beta d_2} + \frac{(2+p)^2}{d_4^2} e^{2\beta d_4} \end{bmatrix}. \end{aligned}$$

For $p \in (0, 1]$, the top-left entry of this matrix is less than the bottom-right entry. Therefore

$$\left\| \tilde{\mathcal{I}}(0, x_M/\sqrt{2})^{-1} \right\| = \frac{1}{C^2 x_M^2 \left(\frac{p^2}{d_2^2} e^{2\beta d_2} + \frac{p^2}{d_4^2} e^{2\beta d_4} \right)}.$$

The minimax problem is now to find p such that the CRB at the two types of local maximum locations is the same, that is

$$\left\| \mathcal{I}(0, x_M)^{-1} \right\| = \left\| \tilde{\mathcal{I}}(0, x_M/\sqrt{2})^{-1} \right\|.$$

This condition is equivalent to

$$2 \frac{e^{2\beta d_3}}{d_3^2} = \frac{e^{2\beta d_1}}{d_1^2} + \frac{e^{2\beta d_2}}{d_2^2},$$

which is simple to solve for p numerically when the distances d_1 , d_2 , and d_3 are treated as functions of p .

This section demonstrates that it is possible to algebraically derive expressions for optimal sensor positions in some simple cases. However, as in this case, numerical approximation may still be required to handle the resulting transcendental expression. I suggest that such algebraic manipulations are not generally worthwhile inasmuch as the CRB is simple to calculate and is amenable to more direct numerical optimization.

3.4 CRB and Monte Carlo

For the energy-based framework of Alajlouni [1] described in Section 2.3, I performed Monte Carlo experiments to test whether the CRB is a good predictor of the variance of maximum likelihood estimators (MLEs) of the location parameters x and y .

The model of Section 2.3, developed in [3], was used to simulate the transfer of energy from a theoretical impact at each point on a fine spatial grid (7500 points) to the real-world locations of sensors in Goodwin Hall. The parameters $\alpha = 4 \times 10^{-4}$ and $\beta = -0.55$ were chosen to approximately match a typical footstep, as found in [2]. The energy received by each sensor was then perturbed with additive noise, as indicated in Section 2.3, using $\sigma = 0.001$ for each sensor. The energy of each impact was affected by 6000 noise simulations to form a Monte Carlo (MC) experiment. For each noise simulation, the MLEs for x , y , α , and β were calculated. The standard deviation (STDEV) of the 6000 x and y estimators were calculated for comparison with the CRB. Figure 3.9 shows the layout of some sensors in Goodwin Hall overlaid on the results of the MC study, compared to the CRB for x and y in the same region. The top row of Figure 3.9 is visually difficult to distinguish from the bottom row, which indicates that the STDEV of the MLEs obtained by MC simulation (for both x and y) are very close to the CRB—the theoretical optimum.

Figure 3.10 shows a histogram of the relative error between the CRB and the MLE

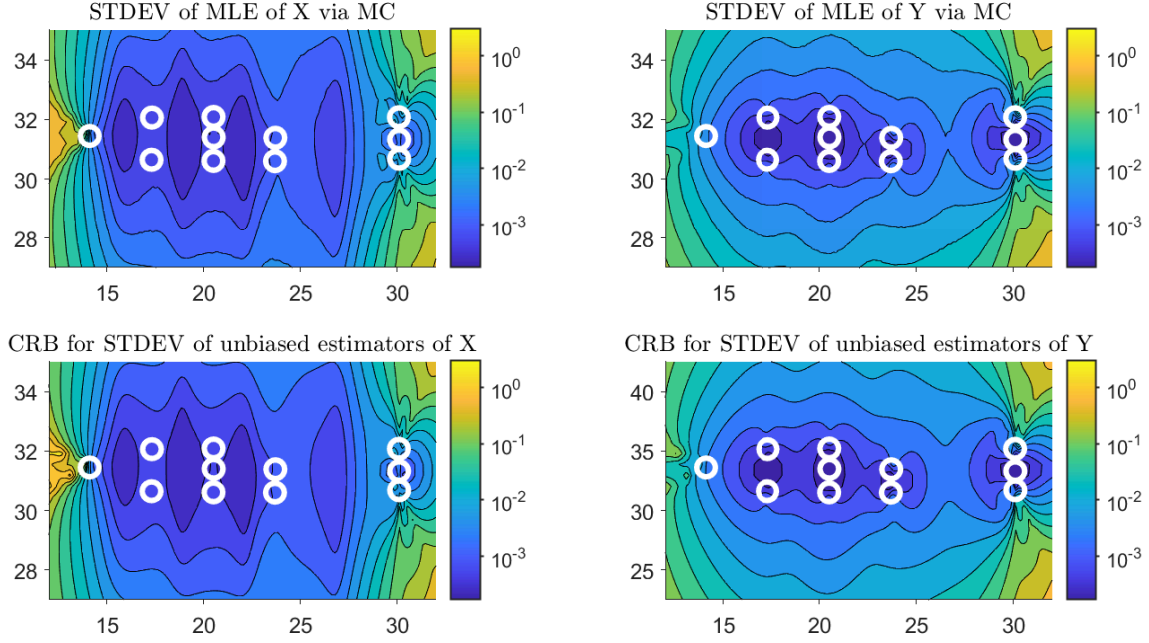


Figure 3.9: MC-generated STDEV of MLEs of x (top left) and y (top right) and corresponding CRBs (bottom left and right). Distances and STDEVs are measured in meters.

STDEVs. At approximately half of the 7500 grid points, the relative error was less than 10^{-2} , meaning the CRB and MLE STDEV differed by less than 1%. At only 159 points was the relative error greater than 10%.

Importantly, the CRB is much less computationally expensive than the millions of simulations and optimizations for the MC study. Since the CRB and MLE STDEV match each other so well for this typical sensor layout, I recommend the CRB as an inexpensive alternative to MC-based estimates of localization variance.

3.5 CRB and Experimental Data

This section will show that the CRB is useful for identifying regions in which high variance can be expected for the MLEs from practical experiments. The first experiment is the one

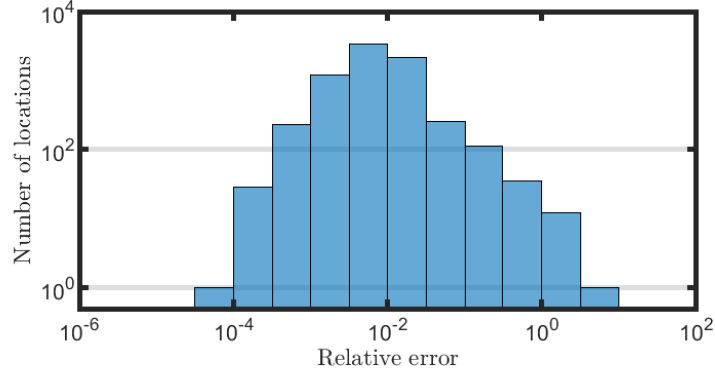


Figure 3.10: Histogram of relative errors between top and bottom portions of Figure 3.9.

described in [2], which used only the sensors shown in Figure 3.9 and involved two test subjects walking east and west between $x = 14$ and $x = 30$. Evenly-spaced bits of adhesive tape were placed on the floor before the experiment, and the subjects were instructed to attempt to step on tape so that their nominal step locations would be known. By stepping on the tape, each subject took 27 steps to cross the length of the experimental area. Each subject crossed this distance in both directions six times, so 12 footsteps theoretically took place at each of 27 x -coordinates.

Using the sensor readings and the model of Section 2.3, MLEs x^* , y^* , α^* , and β^* were calculated for each footstep. For each assigned footstep location, I calculated the STDEVs of the 12 corresponding x^* and y^* values. This STDEV is approximately 100 times larger than predicted by the MC experiments and CRB. Figure 3.11 compares the STDEV of the MLEs from the practical experiment to those of MC experiments. Figure 3.11 also shows the best fit linear regression line by ordinary least squares (OLS), which has a coefficient of determination of $R^2 = 0.230$. This slight correlation is only modest evidence of the efficacy of MC experiments for predicting localization accuracy. Since the MC STDEVs are seen to be nearly the same as the CRB, we may also say that the CRB has some correlation with localization accuracy.

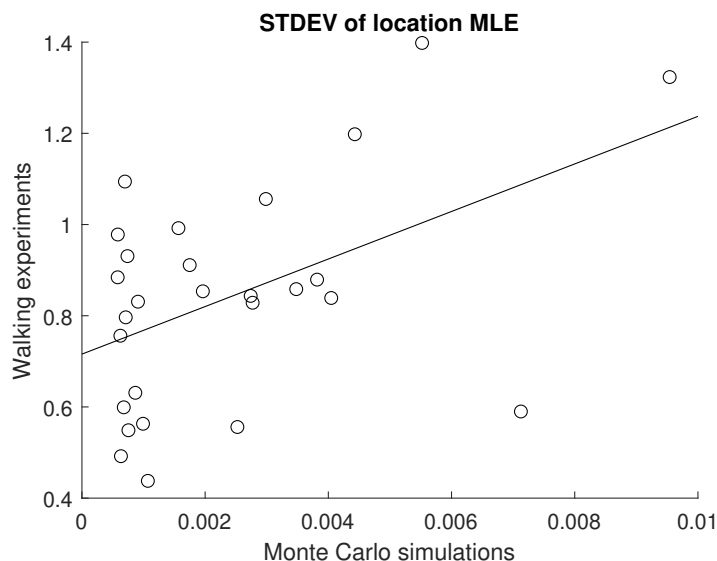


Figure 3.11: MLE STDEVs of MC study vs MLE STDEVs for first walking experiment. The results of the MC study are very similar to the CRB, so this plot also approximates the relationship between the CRB and the STDEVs of the first walking experiment.

However, this first experiment took place near a relatively closely spaced collection of sensors. It seemed plausible that the CRB would have power to predict areas of high localization accuracy (near many sensors) and low accuracy (far from the sensors). I conducted a second experiment with more steps at each assigned location and with some step locations farther from most sensors. In this second experiment, one test subject stepped at the assigned locations shown in Figure 2.1. The subject walked in each direction ten times, stepping on each target 20 times, which provides more samples for each STDEV calculation than the first experiment did. Figure 3.12 is analogous to Figure 3.11, but with the results of the second experiment and the CRB used in place of the MC results. Ill-conditioning affects the problem of localizing impacts that are far from most sensors, which explains why some CRB values for the second experiment are so much larger. In particular, the CRB is quite large for the higher-indexed footstep locations on the east (right) end of the walking path. Figure 3.12 also shows the OLS best fit line for the log-transformed data.

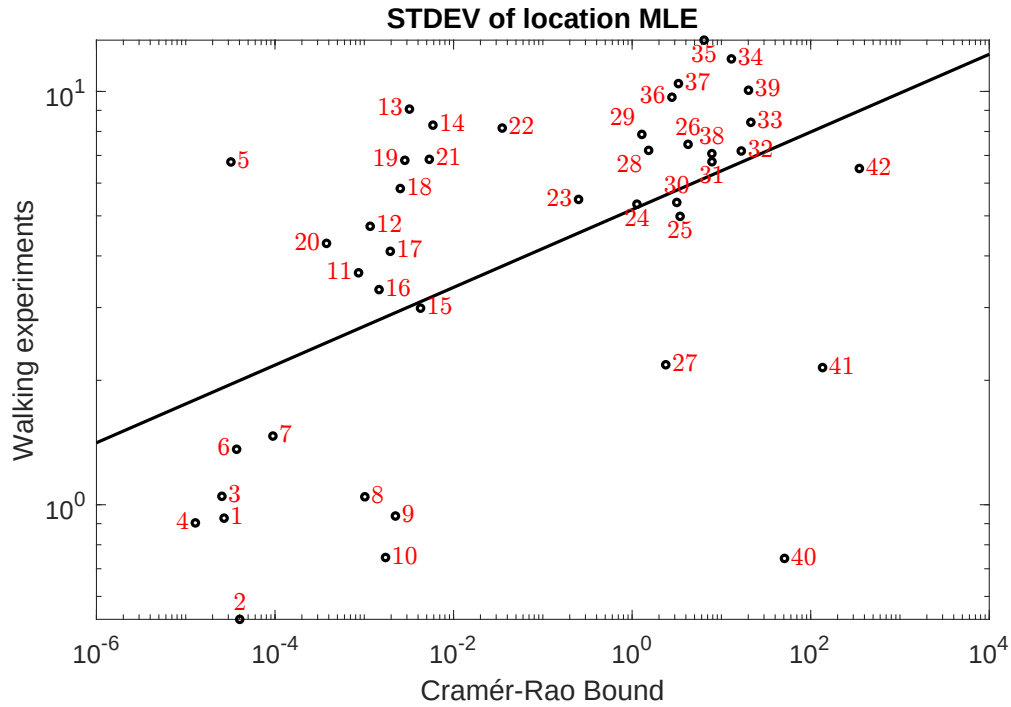


Figure 3.12: CRB vs MLE STDEVs for the second walking experiment. Each point is labeled with the corresponding footstep target location number: 1 on the west (left) and 42 on the east (right) end of the walking path shown in Figure 3.13.

In Figure 3.12 there are some clusters worth noting. The walking STDEVs seem to increase abruptly from step 10 to 11. Figure 3.13 shows that steps 1 to 7 are in the convex hull of a collection of nearby sensors. The walker was not in that sensitive region on steps 8 to 42, which which may explain the sudden accuracy drop slightly to the east of that position. Similarly, the walker passes another tight group of sensors between steps 19 and 20. The experimental STDEV does not appear to be much affected by this transition, but there is a noticeable difference in the CRB values of steps before 22 compared to those after step 23: a few steps east of several nearby sensors.

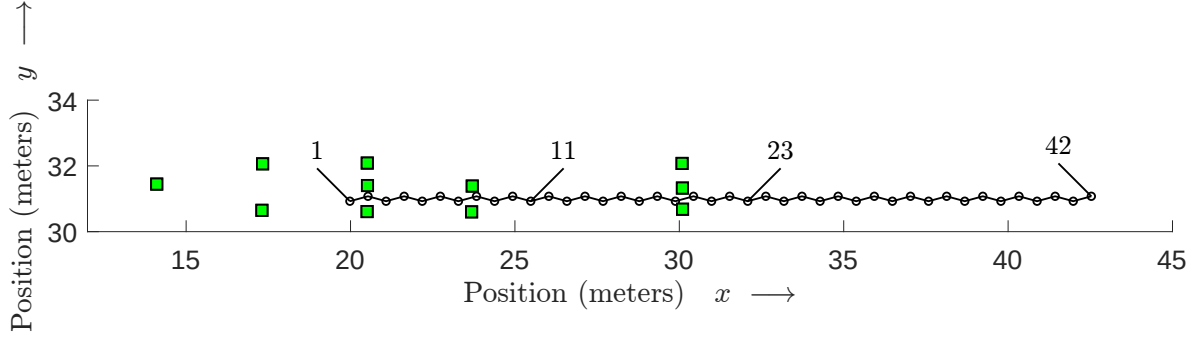


Figure 3.13: Reproduction of Figure 2.1 with several labeled steps locations. In Figure 3.12, steps east of location 11 are associated with noticeably higher location estimator variance. There is a particularly steep rise in the CRB for step locations east of location 23.

3.6 Conclusion

This chapter has explored how the Fisher information matrix may be used to evaluate the fitness of sensor layouts for the source localization problem. The FIM may be used to reveal regions where the localization problem is ill-conditioned and inaccurate estimates should be expected. For one-dimensional regions, it is quite feasible to iteratively adjust the sensor layout to numerically optimize the FIM. Figure 3.6 illustrates the particularly useful principle that long, heavily damped corridors are best monitored by uniformly distributed sensors, whereas sensors should be placed primarily at the ends of short, lightly damped corridors. Figures 3.11 and 3.12 summarize the results of the live experiments that give some evidence for the ability of the FIM to predict localization accuracy. This evidence provides partial justification for the use of the FIM as a readily-computable proxy for physical experiments.

Chapter 4

Nonlinear eigenvalue problems

4.1 Introduction

Given a square matrix $A \in \mathbb{C}^{n \times n}$, the standard linear eigenvalue problem (LEP) seeks scalar values $\lambda \in \mathbb{C}$ such that the matrix $A - \lambda I$ is singular. If $(A - \lambda I)v = 0$ for a vector $v \neq 0$, then v is an eigenvector corresponding to the eigenvalue λ . A nonlinear eigenvalue problem (NLEVP) is a generalization based on a nonlinear matrix-valued function of a scalar parameter,

$$T : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}.$$

A scalar $\lambda \in \mathbb{C}$ such that $T(\lambda)$ is singular is an eigenvalue of T , and if $T(\lambda)v = 0$ for a vector $v \neq 0$, then v is an eigenvector corresponding to the eigenvalue λ . The spectrum (the set of eigenvalues) of T will be denoted by $\sigma(T)$.

Unless T is singular for all inputs, T is said to be *regular*. Throughout these chapters, T will be presumed to be regular and analytic on the domain of interest. For the more restrictive case of polynomial eigenvalue problems (PEVPs), it is not difficult to invent a solution algorithm. Suppose T is a polynomial matrix function of degree d , meaning

$$T(z) = \sum_{i=0}^d z^i A_i,$$

where each $A_i \in \mathbb{C}^{n \times n}$ is a constant matrix. Then the NLEVP $T(\lambda)\mathbf{v} = 0$ is equivalent to

the $n(d+1) \times n(d+1)$ generalized linear eigenvalue problem $L(\lambda)\mathbf{w} = 0$, where

$$L(\lambda) := \begin{bmatrix} A_{d-1} & A_{d-2} & \cdots & A_1 & A_0 \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix} - \begin{bmatrix} -A_d & & & & \\ & I & & & \\ & & I & & \\ & & & \ddots & \\ & & & & I \end{bmatrix} \lambda.$$

The problems are equivalent since

$$T(\lambda)\mathbf{v} = 0 \quad \Longleftrightarrow \quad L(\lambda) \begin{bmatrix} \lambda^{d-1}\mathbf{v}^T & \lambda^{d-2}\mathbf{v}^T & \cdots & \mathbf{v}^T \end{bmatrix}^T = 0.$$

Any desired eigenvalue solver may then be applied to the linear problem. This approach is generalized in [47] to use any set of analytic scalar basis functions $f_0, \dots, f_d$ (in place of the monomial basis $z^0, \dots, z^d$).

The development of additional NLEVP solvers is currently an active topic in the numerical analysis community. For a recent survey of NLEVP methods, see [26].

There are important similarities between linear and nonlinear eigenvalue problems. The notion of algebraic multiplicity extends to nonlinear problems provided it is defined as the order of the root of $\det T(z)$ at $z = \lambda$, i.e., the least integer $k \geq 1$ such that

$$\left. \frac{d^k}{dz^k} \det T(z) \right|_{z=\lambda} \neq 0.$$

Likewise, the geometric multiplicity of the eigenvalue λ is defined for LEPs and NLEVPs as $\dim \ker T(\lambda)$, which makes it clear that the geometric multiplicity of an eigenvalue cannot exceed the problem dimension n . It is also not hard to show that the geometric multiplicity of an eigenvalue cannot exceed its algebraic multiplicity for LEPs or NLEVPs.

On the other hand, some familiar facts about LEPs have no analogue for NLEVPs. For an LEP, the sum of the algebraic multiplicities of the eigenvalues is the problem dimension n , but the size of an NLEVP gives no information about the algebraic multiplicities of its eigenvalues. For example, when $n = 1$ and k is any positive integer, the 1×1 NLEVP $T(z) = z^k$ has an eigenvalue at $\lambda = 0$ with algebraic multiplicity k . The number of eigenvalues of an NLEVP may not be bounded by n or even finite, as in the case of the 1×1 problem $T(z) = \sin(z)$. However, if T is analytic and regular on a domain Ω , then its eigenvalues in Ω have no accumulation points [26]. Consequently, if $\Omega \subset \mathbb{C}$ is bounded, then Ω contains only finitely many eigenvalues of T .

Keldysh's theorem establishes an important decomposition of NLEVPs into invariant subspaces and an analytic residual [26], and serves as the foundation of the contour methods discussed in Sections 4.2.1–4.2.3. In its general form, Keldysh's theorem states that the inverse of any regular analytic $T : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$ can be written as

$$T(z)^{-1} = V(zI - J)^{-1}W^* + E(z), \quad (4.1)$$

where $J \in \mathbb{C}^{m \times m}$ is a block-Jordan matrix whose eigenvalues are the eigenvalues of T that lie in Ω , the columns of $V, W \in \mathbb{C}^{n \times m}$ are corresponding right and left generalized eigenvectors of T , and $E : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$ is a residual function that is analytic on the domain Ω containing the captured eigenvalues. In essence, all non-analytic behavior of $T(z)^{-1}$ in Ω is captured by the resolvent $(zI - J)^{-1}$. Consequently, the decomposition (4.1) is “complete,” in the sense that the algebraic multiplicities of the eigenvalues of J are no smaller than the corresponding algebraic multiplicities of the eigenvalues of T .

The decomposition (4.1) is not unique because it may contain redundant spectral infor-

mation. For a simple example, consider that

$$V(zI - J)^{-1}W^* = \frac{1}{2} \begin{bmatrix} V & V \end{bmatrix} \begin{bmatrix} zI - J & 0 \\ 0 & zI - J \end{bmatrix}^{-1} \begin{bmatrix} W & W \end{bmatrix}^*.$$

An additional restriction is needed to make (4.1) essentially unique, up to eigenvector permutation and scaling. This restriction is that (J, V) must be a *minimal pair* for T , which means that the pair (J, V) must be observable in the usual sense of control theory, i.e., the observability matrix

$$\mathcal{O}(J, V) = \begin{bmatrix} V \\ VJ \\ \vdots \\ VJ^{m-1} \end{bmatrix}$$

is full rank. If (J, V) is not minimal, then a minimal pair representing the same invariant subspaces may be formed by the procedure in [8, Thm. 3]. Consequently, there exists a decomposition of the form (4.1) where (J, V) is a minimal pair.

If (J, V) is minimal, the decomposition (4.1) makes it clear that each eigenvalue has the same algebraic multiplicity for both J and T . In this case, Kressner [32] and Beyn [10] refer to (J, V) as a *simple invariant pair* of T .

4.2 Contour methods

Contour integral methods attempt to find the eigenvalues of $T : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$ that lie in a user-specified domain $\Omega \subset \mathbb{C}$. Let $m < \infty$ be the total number of eigenvalues of T that lie in Ω , that is, m is the sum of the algebraic multiplicities of the distinct eigenvalues in Ω .

Suppose that one can compute $T(z)$, or at least the actions $T(z)^{-1}$ and $\mathbf{w}^*T(z)^{-1}$, for any

desired input point $z \in \mathbb{C}$ and directions $\mathbf{v}, \mathbf{w} \in \mathbb{C}^n$. For any test function $f : \mathbb{C} \rightarrow \mathbb{C}$ analytic on Ω , (4.1) shows that

$$\frac{1}{2\pi i} \int_{\partial\Omega} f(z) T(z)^{-1} dz = \left(\frac{1}{2\pi i} \int_{\partial\Omega} f(z) V((zI - J)^{-1} W^* + E(z)) dz \right).$$

Since E and f are analytic on Ω , Cauchy's integral theorem shows that $\int_{\partial\Omega} f(z) E(z) dz = 0$.

Now, by Cauchy's integral formula for matrices [28, Thm. 6.2.28],

$$\frac{1}{2\pi i} \int_{\partial\Omega} f(z) T(z)^{-1} dz = \frac{1}{2\pi i} \int_{\partial\Omega} f(z) V(zI - J)^{-1} W^* dz = V f(J) W^*. \quad (4.2)$$

It is not hard to imagine that V , J , and W could be revealed by calculating $V f_k(J) W^*$ for several wisely chosen test functions f_k . Several specific methods will be outlined later in this section.

The matrix inverses above may be replaced with linear solves by applying p left and right “probing” directions assembled into $L, R \in \mathbb{C}^{n \times p}$ to give

$$\frac{1}{2\pi i} \int_{\partial\Omega} f(z) L^* T(z)^{-1} R dz = L^* V f(J) W^* R. \quad (4.3)$$

If the integral (4.3) could be computed exactly, the methods presented later in Sections 4.2.1 and 4.2.3 would exactly deliver some eigenvalues of T as the eigenvalues of derived linear eigenvalue problem. In actuality, the integral in (4.3) is approximated by quadrature. For properly chosen quadrature points $\{z_i\}_{i=1}^N$ on $\partial\Omega$ and corresponding weights $\{w_i\}_{i=1}^N$,

$$L^* V f(J) W^* R \approx \sum_{i=1}^N f(z_i) w_i L^* T(z_i)^{-1} R dz.$$

It seems worth noting that the integral in (4.3) resembles the computation of spectral projectors for the FEAST algorithm for linear eigenvalue problems [30, 40] and the extension of FEAST to NLEVPs in [22].

When interpreting $T(z)^{-1}$ as the transfer function of a dynamical system, the rational portion $H(z) := V(zI - J)^{-1}W^*$ can be seen as the (rational) transfer function of a linear time invariant (LTI) system that matches the poles of $T(z)^{-1}$ in Ω . This perspective offers connections to rational system identification theory. Sections 4.2.1 and 4.2.3 briefly review some results of [12], which applies the system theoretic results of [36] and [5] to NLEVPs. Section 4.2.2 makes additional connections to control theory and presents a new control-inspired result relevant to contour integration NLEVP solvers.

4.2.1 Single shift methods

This section briefly explains two closely related NLEVP solution methods based on rational interpolation of T at a single point. The first technique is often referred to as Beyn's method after its presentation in [9], although Asakura et al. derived a similar method for simple eigenvalues earlier in [6].

Rewriting $H(z) := V(zI - J)^{-1}W^*$ using a Neumann series expansion of $(zI - J)^{-1}$ gives

$$\begin{aligned}
 L^*H(z)R &= L^*V(zI - J)^{-1}W^*R \\
 &= \frac{1}{z}L^*V\left(I - \frac{1}{z}J\right)^{-1}W^*R \\
 &= \frac{1}{z}L^*V\left(\sum_{i=0}^{\infty}\left(\frac{1}{z}J\right)^i\right)W^*R \\
 &= \sum_{k=0}^{\infty}L^*VJ^kW^*R\frac{1}{z^{k+1}} =: \sum_{k=0}^{\infty}A_k\frac{1}{z^{k+1}}.
 \end{aligned} \tag{4.4}$$

Here, each $A_k = L^*VJ^kW^*R$ can be evaluated from (4.3) using the test function $f_k(z) = z^k$. Specifically,

$$A_k = \frac{1}{2\pi i} \int_{\partial\Omega} z^k L^*T(z)^{-1}R \, dz, \tag{4.5}$$

so each A_k may be approximated by quadrature with samples of $L^*T(z)^{-1}R$. The form (4.4) corresponds to Laurent series interpolation of the rational function $H(z)$ at $z = \infty$. The resulting algorithm will thus be referred to as being based on “interpolation at infinity.”

Now with any desired even number $2K$ of samples, form the block Hankel matrices

$$\mathbb{H} := \mathcal{H}(A_0, \dots, A_{2K-2}) := \begin{bmatrix} A_0 & A_1 & \cdots & A_{K-1} \\ A_1 & A_2 & \cdots & A_K \\ \vdots & \vdots & \ddots & \vdots \\ A_{K-1} & A_K & \cdots & A_{2K-2} \end{bmatrix} \in \mathbb{C}^{pK \times pK}, \quad (4.6a)$$

$$\mathbb{H}_s := \mathcal{H}(A_1, \dots, A_{2K-1}) := \begin{bmatrix} A_1 & A_2 & \cdots & A_K \\ A_2 & A_3 & \cdots & A_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ A_K & A_{K+1} & \cdots & A_{2K-1} \end{bmatrix} \in \mathbb{C}^{pK \times pK}, \quad (4.6b)$$

and take the (reduced) singular value decomposition $\mathbb{H} = U\Sigma V^*$. In exact arithmetic, it can be shown [9] that if $\text{rank } \mathbb{H} = m$ (so that $\Sigma \in \mathbb{R}^{m \times m}$), then the m (ordinary linear) eigenvalues of

$$U^* \mathbb{H}_s V \Sigma^{-1} \in \mathbb{C}^{m \times m}$$

are the m eigenvalues of T that lie in Ω .

Rather than interpolating H at infinity, a similar procedure produces eigenvalues through interpolation at a (finite) complex point $\sigma \in \mathbb{C} \setminus \overline{\Omega}$. The Taylor expansion of H at σ is

$$\begin{aligned} L^*H(z)R &= \sum_{k=0}^{\infty} \frac{(z - \sigma)^k}{k!} L^*H^{(k)}(\sigma)R \\ &= \sum_{k=0}^{\infty} (-L^*V(J - \sigma I)^{-k-1}W^*R)(z - \sigma)^k =: \sum_{k=0}^{\infty} M_k(z - \sigma)^k. \end{aligned} \quad (4.7)$$

Each $M_k = -L^*V(J - \sigma I)^{-k-1}W^*R$ can be evaluated from (4.3) using the test function $f_k(z) = (z - \sigma)^{-k-1}$. Specifically,

$$M_k = \frac{-1}{2\pi i} \int_{\partial\Omega} (z - \sigma)^{-k-1} L^*T(z)^{-1} R \, dz, \quad (4.8)$$

so each M_k may be approximated by quadrature with samples of $L^*T(z)^{-1}R$. Similar to the previous process, form the block Hankel matrices

$$\mathbb{L}_0 := \mathcal{H}(M_0, \dots, M_{2K-2})$$

$$\mathbb{L} := \mathcal{H}(M_1, \dots, M_{2K-1})$$

$$\mathbb{L}_s := \sigma \mathbb{L} + \mathbb{L}_0,$$

where $\mathbb{L}$ and $\mathbb{L}_s$ are also known as the *Loewner* and *shifted Loewner* matrices of order K for H at $\sigma \in \mathbb{C}$. Now take the (reduced) singular value decomposition $\mathbb{L} = U_L \Sigma_L V_L^*$. In exact arithmetic, it can be shown [12] that, if $\text{rank } \mathbb{L} = m$ (so that $\Sigma_L \in \mathbb{R}^{m \times m}$), then the m eigenvalues of the matrix

$$U_L^* \mathbb{L}_s V_L \Sigma_L^{-1} \in \mathbb{C}^{m \times m}$$

are the m eigenvalues of T that lie in Ω (with matching algebraic multiplicities).

Once an eigenvalue λ of T is known, corresponding approximate eigenvectors may be obtained in a separate process by examining the left and right nullspaces of $T(\lambda)$. Alternatively, the procedures described in this chapter may be altered slightly to calculate eigenvectors more directly, as in [12].

4.2.2 Controllability index

This section is concerned with the requirements in Section 4.2.1 that $\mathbb{H}$ and $\mathbb{L}$ have rank m to match the number of sought after eigenvalues in Ω . This section will focus on $\mathbb{H}$, but similar arguments apply to $\mathbb{L}$. To simplify notation in this section, take $p = n$ and use identity probing $R = L = I$.

From the definitions of A_k and $\mathbb{H}$ in (4.4) and (4.6a), we can write

$$\mathbb{H} = \mathcal{O}^K \mathcal{C}^K, \quad \mathbb{H}_s = \mathcal{O}^K J \mathcal{C}^K, \quad (4.9)$$

where $\mathcal{O}^K$ and $\mathcal{C}^K$ are the “partial” observability and controllability matrices

$$\mathcal{O}^K := \mathcal{O}^K(J, V) := \begin{bmatrix} V \\ VJ \\ \vdots \\ VJ^{K-1} \end{bmatrix} \in \mathbb{C}^{pK \times m}, \quad (4.10a)$$

$$\mathcal{C}^K := \mathcal{C}^K(J, W^*) := \begin{bmatrix} W^* & JW^* & \dots & J^{K-1}W^* \end{bmatrix} \in \mathbb{C}^{m \times pK}. \quad (4.10b)$$

The traditional (full) observability and controllability matrices are $\mathcal{O} := \mathcal{O}^m$ and $\mathcal{C} := \mathcal{C}^m$.

When the single-shift methods of Section 4.2.1 are used to calculate eigenvalues, the number of eigenvalues m is not generally known. Practitioners must choose a K value, use samples to construct $\mathbb{H}$, and attempt to determine whether their choice of K was appropriate. Since $\mathcal{O}^K$ and $(\mathcal{C}^K)^*$ are $pK \times m$, the rank of $\mathbb{H} = \mathcal{O}^K \mathcal{C}^K$ could be as large as $\min(pK, m)$. If the numerical rank of $\mathbb{H}$ is pK then it is possible that $m > pK$, and it is advisable to take more samples (increasing K) and form a new $\mathbb{H}$ with which to calculate eigenvalues. If one sees that the numerical rank of $\mathbb{H}$ is less than pK , it is *possible* that $\text{rank } \mathbb{H} = m$ and that m —the number of eigenvalues in Ω —has been discovered. However, it is also possible

that linear dependence among the eigenvectors has prevented $\text{rank } \mathcal{C}^K$ and $\text{rank } \mathcal{O}^K$ (and therefore $\text{rank } \mathbb{H}$) from reaching pK . One may wish to compare the numerical ranks of several $\mathbb{H}$ matrices using different numbers of samples to ensure that increasing K no longer affects $\text{rank } \mathbb{H}$.

It appears useful to know something of the circumstances under which many samples must be taken to achieve $\text{rank } \mathbb{H} = m$. It will be simpler to consider the least K for which just one of the factors achieves full rank, say $\text{rank } \mathcal{C}^K = m$. To that end, define the *controllability index*

$$K^* := K^*(J, W^*) := \min\{K : \text{rank } \mathcal{C}^K(J, W^*) = m\}. \quad (4.11)$$

If $\text{rank } \mathcal{C}^K(J, W^*) < m$ for all K , then say that $K^* = \infty$. When $K^*(J, W^*) \neq \infty$, the pair (J, W^*) may be called *controllable*. By taking the transposes of J and V , the same definition (4.11) can be used to write $K^*(J^T, V^T)$ for the least K such that $\text{rank } (\mathcal{O}^K) = m$, which might be called the *observability index*. Other works on NLEVPs focus on the right eigenvector matrix V and derive corresponding results about the observability matrix $\mathcal{O}^K$. For example, [9], [10], and [26] call the pair (J, V) *minimal* when the observability index $K^*(J^T, V^T)$ is finite. I will continue to focus on W^* and the controllability matrix $\mathcal{C}^K$ because $\mathcal{C}^K$ is also the K th block Krylov matrix generated by J acting on W^* . In the Krylov context, K^* is sometimes called the *grade* of W^* with respect to J [25]. However, unlike the definition just given in (4.11), the important grade for Krylov methods is that which “completes” the Krylov space. The maximal Krylov space may not be of dimension m , and the grade is usually defined as

$$\tilde{K}^* := \tilde{K}^*(J, W^*) := \min\{K : \text{rank } \mathcal{C}^K = \text{rank } \mathcal{C}^{K+1}\}.$$

Since $\text{rank } \mathcal{C}^K$ is non-decreasing in K , it is apparent that $K^* = \tilde{K}^*$ so long as K^* is finite. I will continue to use K^* because this section is motivated by achieving $\text{rank } \mathcal{C}^K = m$,

but the results apply to $\tilde{K}^*$ with little modification. Also, the results of this section can be easily modified to give bounds on the observability index since it is simply the controllability index of (J^T, V^T) . In the context of NLEVPs, the bounds in this section are primarily a tool to predict the number of moments K that must be used to achieve $\text{rank } \mathcal{O}^K \mathcal{C}^K = m$, and necessarily $\text{rank } \mathcal{O}^K = \text{rank } \mathcal{C}^K = m$. Thus, the minimum number of moments needed is no less than the larger of the controllability and observability indices: $\max(K^*(J, W^*), K^*(J^T, V^T))$. If W and V are arbitrary, then the controllability and observability indices might have nothing to do with each other. But in the present context, V and W are connected since they hold right and left eigenvectors of the same NLEVP. In fact, in the vibration-related NLEVP problems I have examined, the controllability and observability are equal.

Theorem 4.1 below is a standard result in control theory textbooks, e.g., [44, Sec. 3.3]. Condition 1 is commonly regarded as the definition of the (state space) controllability of (J, W^*) . The equivalence of Conditions 3 and 4 is apparent, and some sources present only one.

Theorem 4.1. *The following are equivalent.*

1. Definition of controllability

For any initial condition $\mathbf{x}(0) = \mathbf{x}_0$, there is a piecewise continuous control function $\mathbf{u}$ such that the (linear time-invariant) control system

$$\dot{\mathbf{x}}(t) = J\mathbf{x}(t) + W^*\mathbf{u}(t)$$

is driven to $\mathbf{x}(T) = 0$ in some finite time T .

2. Kalman rank criterion

$$\text{rank } \mathcal{C} = m.$$

3. Popov–Belevitch–Hautus (PBH) eigenvector criterion

For all left eigenvectors $\mathbf{x}$ of J : $\mathbf{x}^*W^* \neq 0$.

4. Hautus lemma

For all $\lambda \in \mathbb{C}$: $\text{rank} \begin{bmatrix} J - \lambda I & W^* \end{bmatrix} = m$.

To proceed further, some notation is needed to describe the sizes of the eigenspaces of J . Say that the distinct eigenvalues of J are $\lambda_1, \dots, \lambda_q$ and that the algebraic and geometric multiplicities of λ_i are m_i and d_i . For each i , fully decompose the algebraic multiplicity of λ_i into the partial multiplicities $m_{i,1} \geq \dots \geq m_{i,d_i}$ corresponding to distinct Jordan chains. Notice that $m_i = \sum_{j=1}^{d_i} m_{i,j}$ and $m = \sum_{i=1}^q m_i$.

The Kalman rank criterion connects K^* to the controllability of (J, W^*) . If $\mathcal{C}^K$ has full row rank for some K , then $K = m$ must be sufficient. This holds because the characteristic polynomial of J has degree m , and so from the Cayley–Hamilton theorem, the columns of $J^m W^*$ are linear combinations of the columns of $\mathcal{C} = \mathcal{C}^m$. Consequently,

$$K^* \leq m. \tag{4.12a}$$

Kressner found this bound in the NLEVP context in [32, Lem. 5].

Next, let $d := \sum_{i=1}^q m_{i,1}$ be the degree of the minimal polynomial of J . Then the columns of $J^d W^*$ are linear combinations of the columns of $\mathcal{C}^d$, and if (J, W^*) is controllable, then

$$K^* \leq d. \tag{4.12b}$$

Since $d \leq m$, (4.12b) is a stronger bound than (4.12a). Beyn included (4.12b) while

deriving the contour integral method with interpolation at infinity in [9, Lem. 5.1].

The improvement of (4.12b) over (4.12a) is most apparent when a single eigenvalue of J is associated with multiple Jordan blocks. If $J = I_m$ is the identity matrix, then $\lambda_1 = 1$ is associated with m 1×1 Jordan blocks, that is, $d_1 = m$ and $m_{1,j} = 1$ for $j = 1, \dots, m$. The right hand side of (4.12b) is $d = m_{1,1} = 1$ while the right hand side of (4.12a) is m .

The bounds (4.12a) and (4.12b) are independent of W^* and cannot capture the possibly intricate relationships W^* may have with J . In particular, these bounds do not make use of the fact that linearly independent columns of W^* may cause the rank of $\mathcal{C}^K(J, W^*)$ to grow quickly with K . For an extreme example, take $m = n$, $J = \text{diag}(1, \dots, m)$, and $W = I_m$. Then $\lambda_i = i$ and $m_i = m_{i,1} = 1$ for $i = 1, \dots, m$. The right hand sides of (4.12a) and (4.12b) are both m , but $\mathcal{C}^1 = W^*$ is full rank, and so $K^* = 1$.

One more upper bound is given in [25, Lem. 8]. If row i of W is denoted by W_i , then the Krylov grade can be bounded as

$$\tilde{K}^*(J, W^*) \leq \max_i \tilde{K}^*(J, W_i^*).$$

This bound is most informative if $\tilde{K}^*(J, W_i^*)$ is small for every i , as when each W_i^* lies in a low-dimensional (right) invariant subspace of J . The corresponding bound for the controllability index K^* ,

$$K^*(J, W^*) \leq \max_i K^*(J, W_i^*), \quad (4.13)$$

also holds, but is less interesting. For each i , the controllability matrix $\mathcal{C}^K(J, W_i^*)$ is only $m \times K$. Consequently, $K^*(J, W_i^*)$ is either ∞ or m . If $\text{rank } \mathcal{C}^m(J, W_i^*) = m$ for all i , then each $K^*(J, W_i^*) = m$, and (4.13) only replicates (4.12a). Otherwise, $K^*(J, W_i^*) = \infty$ for some i , then (4.13) provides no information.

The following lower bound for K^* appears to be new.

Theorem 4.2. *Suppose $J \in \mathbb{C}^{m \times m}$, $W \in \mathbb{C}^{n \times m}$, and (J, W^*) is controllable. Take $Q \in \mathbb{C}^{m \times \ell}$ such that its columns form a basis for some ℓ -dimensional left-invariant subspace of J , in other words*

$$Q^* J = U Q^*$$

for some invertible $U \in \mathbb{C}^{\ell \times \ell}$ (remember $\ell = \text{rank } Q$).

Then $\text{rank } WQ > 0$ and

$$K^* \geq \frac{\text{rank } Q}{\text{rank } WQ}. \quad (4.14)$$

Proof. Notice first that the range of Q includes a left eigenvector of J . Specifically, for any left eigenpair $(\lambda, \mathbf{x})$ of U ,

$$\mathbf{x}^* Q^* J = \mathbf{x}^* U Q^* = \lambda \mathbf{x}^* Q^*,$$

so $Q\mathbf{x}$ is a left eigenvector of J . By the PBH eigenvector criterion of Theorem 4.1, $\mathbf{x}^* Q^* W^* \neq 0$ and therefore $\text{rank } WQ > 0$.

Let $r := \text{rank } WQ$. Choose a matrix $Y \in \mathbb{C}^{r \times n}$ whose rows form a basis for the row space of $Q^* W^*$, so that $Q^* W^* = FY$ for some $F \in \mathbb{C}^{\ell \times r}$. Then for any $K \geq 1$ we have

$$\begin{aligned} Q^* \mathcal{C}^K &= \begin{bmatrix} Q^* W^* & Q^* J W^* & \dots & Q^* J^{K-1} W^* \end{bmatrix} \\ &= \begin{bmatrix} Q^* W^* & U Q^* W^* & \dots & U^{K-1} Q^* W^* \end{bmatrix} \\ &= \begin{bmatrix} F & UF & \dots & U^{K-1} F \end{bmatrix} \begin{bmatrix} Y \\ Y \\ \ddots \\ Y \end{bmatrix}. \end{aligned} \quad (4.15)$$

The inner dimension of the product in (4.15) is Kr , so $\text{rank } Q^* \mathcal{C}^K \leq Kr = K \text{rank}(WQ)$.

In other words, for every $K \geq 1$

$$K \geq \frac{\text{rank } Q^* \mathcal{C}^K}{\text{rank } WQ}. \quad (4.16)$$

Now define

$$k_Q^* := \min\{K : \text{rank } Q^* \mathcal{C}^K = \text{rank } Q\}. \quad (4.17)$$

Notice that the set on the right-hand side (RHS) of (4.17) is non-empty and in particular includes K^* (and so $K^* \leq k_Q^*$) since Q^* and $\mathcal{C}^{(K^*)}$ have linearly independent rows.

From (4.16) and (4.17),

$$k_Q^* \geq \min\left\{\frac{\text{rank } Q^* \mathcal{C}^K}{\text{rank } WQ} : \text{rank } Q^* \mathcal{C}^K = \text{rank } Q\right\} = \frac{\text{rank } Q}{\text{rank } WQ}. \quad (4.18)$$

Since $K^* \geq k_Q^*$, (4.18) shows that

$$K^* \geq k_Q^* \geq \frac{\text{rank } Q}{\text{rank } WQ}.$$

□

Since (4.14) holds for all full-rank Q that are left-invariant to J , the maximum of (4.14) over all such Q is also a lower bound for K^* . Of course, the RHS of (4.14) may not be an integer, so it could be replaced with its ceiling without affecting the conclusion:

$$K^* \geq \left\lceil \frac{\text{rank } Q}{\text{rank } WQ} \right\rceil. \quad (4.19)$$

Intuitively, we wish to increase K until the columns of $\mathcal{C}^K$ span every left-invariant subspace of J . The controllability index K^* is determined by the left-invariant subspace of J that

W^* takes the longest to “fill.” Each time K increases, some new columns of $\mathcal{C}^K$ may lie in that space ($\text{range } Q$), and some of these new columns may be linearly independent of previous such columns. But the number of those new columns is limited to $\text{rank } WQ$. Since at most $\text{rank } WQ$ vectors are added to the basis at each step, then at least $\frac{\text{rank } Q}{\text{rank } WQ}$ steps must take place before the basis covers $\text{range } Q$. Sometimes, fewer than $\text{rank } WQ$ new linearly independent vectors are contributed at some steps, in which case (4.16) does not hold with equality for some K . Finally, because of the limited growth rate, there is no way for $\text{rank } Q^* \mathcal{C}^K$ to “catch up” if it falls below the optimal Kr . In other words, if the sides of (4.16) are not equal for some $K < K^*$, then (4.16) is a strict inequality for all larger K (including K^*), and (4.14) is a strict inequality.

Return now to the example $J = \text{diag}(1, \dots, m)$ and $W = I$, for which the upper bounds (4.12a) and (4.12b) were pessimistic. For any Q with columns spanning a left-invariant subspace of J , $WQ = Q$, so the right hand side of (4.14) is 1. Since $K^* = 1$ as well, (4.16) is tight.

A pair of examples illustrate how (4.14) estimates K^* . Let J be a 6×6 Jordan block and consider two possible W matrices:

$$J = \begin{bmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ & & 0 & 1 & & \\ & & & 0 & 1 & \\ & & & & 0 & 1 \\ & & & & & 0 \end{bmatrix}, \quad \begin{aligned} W_1 &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \\ W_2 &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \end{aligned} \quad (4.20)$$

The smallest full-rank partial controllability matrices are

$$\mathcal{C}^3(J, W_1^*) = \left[\begin{array}{c|c|c|c} & & & 1 \\ & 1 & & \\ & & 1 & \\ 1 & & & 1 \\ & 1 & & \\ & & 1 & \end{array} \right], \quad \text{and} \quad \mathcal{C}^5(J, W_2^*) = \left[\begin{array}{c|c|c|c|c|c} & & & & & 1 \\ & & & & 1 & \\ & & 1 & & & \\ & 1 & & 1 & & \\ 1 & & & & 1 & \\ & 1 & & & & 1 \end{array} \right],$$

so the controllability indices are $K^*(J, W_1^*) = 3$ and $K^*(J, W_2^*) = 5$. Each increment of K increased the rank of $\mathcal{C}^K(J, W_1^*)$ by 2, but the rank of $\mathcal{C}^K(J, W_2^*)$ by only 1. Increasing the rank of $\mathcal{C}^K$ by 2 at each step is “optimal” for these W since their own ranks are 2, but the columns of W_1^* “cooperate” better than those of W_2^* to reach a full-rank controllability matrix more quickly. The result (4.14) bounds K^* with this “optimal” rate of rank increase.

For a desired dimension $1 \leq \ell \leq 6$, consider Q matrices of the form

$$Q = \begin{bmatrix} 0_{(6-\ell) \times \ell} \\ I_{\ell \times \ell} \end{bmatrix} \in \mathbb{C}^{6 \times \ell}.$$

Every left-invariant subspace of J is represented by such a Q . Table 4.1 shows the RHS of (4.16) for W_1 and W_2 and $\ell = 1, \dots, 6$. When $\ell = 3$, (4.14) reveals that $K^*(J, W_1^*) \geq 3$ since it will take at least three blocks formed by the first column of W_1^* to span a three-dimensional space.

When $\ell = 4$, then (4.14) can only show $K^*(J, W_1^*) \geq 2$ since $\text{rank } W_1 Q = 2$, and (potentially) only two blocks are needed when two vectors are used to build a basis for a four-dimensional space.

Now consider W_2 . As Table 4.1 shows, the RHS of (4.14) can never reach $K^*(J, W_2^*) = 5$ since the bound is based on the maximum rate of basis building, which the columns of W_2^* do not achieve.

4.2.3 Multiple shifts

The order- K interpolation methods described in Section 4.2.1 requires adding $(K - 1)th$ powers of complex numbers as part of the quadrature in (4.5) and (4.8). If K is large, then the summed values may differ by orders of magnitude, which may lead to a loss of precision.

Table 4.1: Calculation of bounds (4.14) and (4.19) for example (4.20).

$\ell = \text{rank } Q$	rank $W_i Q$		RHS of (4.14)		RHS of (4.19)	
	W_1	W_2	W_1	W_2	W_1	W_2
1	1	1	1	1	1	1
2	1	2	2	1	2	1
3	1	2	3	3/2	3	2
4	2	2	2	2	2	2
5	2	2	5/2	5/2	3	3
6	2	2	3	3	3	3

In the interest of greater numerical stability, a multi-shift method is suggested by [12] based on the Loewner framework for system identification by Mayo and Antoulas [36]. The user specifies a collection of p left and right interpolation points

$$\{\theta_i\}_{i=1}^p \subset \mathbb{C} \setminus \overline{\Omega} \quad \text{and} \quad \{\sigma_i\}_{i=1}^p \subset \mathbb{C} \setminus \overline{\Omega},$$

which are assumed to be jointly distinct. The user also specifies corresponding left and right interpolation matrices

$$L = \begin{bmatrix} \ell_1 & \cdots & \ell_p \end{bmatrix} \in \mathbb{C}^{n \times p} \quad \text{and} \quad R = \begin{bmatrix} r_1 & \cdots & r_p \end{bmatrix} \in \mathbb{C}^{n \times p}.$$

Then define the (i, j) entry of the Loewner and shifted Loewner matrices by

$$\begin{aligned} \mathbb{L}_{i,j} &:= \ell_i^* \left(\frac{H(\theta_i) - H(\sigma_j)}{\theta_i - \sigma_j} \right) r_j \\ (\mathbb{L}_s)_{i,j} &:= \ell_i^* \left(\frac{\theta_i H(\theta_i) - \sigma_j H(\sigma_j)}{\theta_i - \sigma_j} \right) r_j \end{aligned}$$

for $i, j = 1, \dots, p$. Each sample required to construct the Loewner matrices can be obtained by an integral: taking z_0 to be any θ_i or σ_j ,

$$\ell_i^* H(z_0) r_j = \frac{1}{2\pi i} \int_{\partial\Omega} \frac{\ell_i^* T(z)^{-1} r_j}{z_0 - z} dz.$$

The rest of the multipoint contour method proceeds like the single shift methods in Section 4.2.1. The sample integrals may be approximated by quadrature, requiring little more work than a linear solve with p right-hand sides for each quadrature point. If the Loewner matrix has rank m , then the eigenvalues of T that lie in Ω are the eigenvalues of the matrix

$$U_L^* \mathbb{L}_s V_L \Sigma_L^{-1} \in \mathbb{C}^{m \times m},$$

where $\mathbb{L} = U_L \Sigma_L V_L^*$ is the (reduced) singular value decomposition ($U_L, V_L \in \mathbb{C}^{p \times m}$ and $\Sigma_L \in \mathbb{C}^{m \times m}$).

Section 5.5 illustrates several advantages of this multi-shift method over the single shift methods.

Chapter 5

NLEVPs for vibrating structures

5.1 Introduction

The focus of this section is the determination of the characteristic vibrating shapes (or “mode shapes”) of a structure and the corresponding “natural” or “resonant frequencies” that may induce them. All vibration behavior of a structure can be described as a combination of its modal vibrations.

Vibrations are often studied through laboratory experiments that subject the structure of interest (or important substructures) to prescribed forces, such as tuned shaking or precise impacts. Meanwhile, carefully-placed vibration sensors monitor the response of the structure.

Other studies are conducted while the entire structure is simply subjected to normal use. For example, some researchers use the Goodwin Hall accelerometers discussed in [Section 2.1](#) to study the mode shapes of the building itself while it is vibrating (slightly) due to the forces imparted by the pedestrians inside and nearby vehicle traffic. Studying modes using the vibrations caused by normal use is called operational modal analysis (OMA). A thorough introduction to OMA can be found in [\[13\]](#)

This chapter deals with predicting mode shapes and natural frequencies mathematically based on material properties of the structure. When a structure is modeled as a collection of finitely many moving masses, the result is a multiple degree of freedom (MDOF) system

whose dynamics can be predicted by solving a (linear) eigenvalue problem. The situation is more complex when using a continuous model to fully represent the shapes of flexible components with theoretically infinite degrees of freedom. The dynamics of such a continuous model is called a distributed parameter system (DPS), and representations of DPSs typically have nonlinear expressions that cannot be solved as linear eigenvalue problems.

Finite element methods (FEMs) are a very common approach to solving DPSs. FEMs discretize DPSs by modelling continuous structural components as chains of many small moving masses, thus converting the DPS to a (large) MDOF system. Modes of the approximating MDOF system may then be found by a linear eigenvalue solver.

Alternatively, spectral methods represent the vibrations of each flexible component as sums of many continuous basis functions. Like FEMs, spectral methods produce large linear eigenvalue problems that may be solved with any desired method.

Typically, the eigenvalues of smallest magnitude (corresponding to the lowest natural frequencies) are of the most interest. But if information about high frequency modes is needed, the FEM or spectral discretizations may need to be very fine (and the linearized system very large) to calculate these modes accurately. Discretization approaches like FEM and spectral methods also have the drawback of producing spurious eigenvalues. Finesse is needed to distinguish the eigenvalues that are good approximations from those that are artifacts of the discretization.

The method used in this section avoids discretization and large linearized systems by forming an NLEVP that describes the boundary conditions of coupled vibrating systems, each portion of which can be solved analytically.

Contour methods for solving NLEVPs can find eigenvalues in any desired region of the complex plane. When performed properly, contour methods produce no spurious eigenvalues and their cost is essentially independent of the range of frequencies that are being

explored.

The examples in this section will be networks of elastic strings, similar to spiderwebs. Similar problem formulations are discussed in [7].

This chapter deals with NLEVP methods for finding dynamic modes of vibrating structures. Three approaches will be presented: an apparently original method, the dynamic stiffness matrix (DSM) method, and the slope-deflection (SD) method. The original method presented first requires no algebraic reprocessing of the problem but at the expense of solving a larger NLEVP than is required by the DSM and SD methods. The DSM and SD methods typically begin with symbolic algebra work that the new method avoids.

The example structures of this chapter will be connected networks of elastic strings comparable to spiderwebs.

5.2 Elastic Web Model

The string network (or web) model described in this section is functionally the same as used in [17], [33, Ch. II], and [42] to study the controllability of networks of vibrating elastic strings. This model was also used in [24] as the basis for a spectral method study of vibrational modes of spiderwebs.

Consider a web that is anchored to an unmoving frame. Some strands of the web connect to the frame on one end and connect to one or more other strands at the other end. The remaining strands are not directly connected to the frame but are connected to other strands at each end.

Model the web as a weakly connected directed graph with n_e edges and n_n nodes. Label the tail and head nodes with vectors $\mathbf{t}$ and $\mathbf{h}$ so that the tail and head of edge i are t_i and h_i for $i = 1, \dots, n_e$. For any node w , let $N(w) = \{i : h_i = w \text{ or } t_i = w\}$ be the set of edges

connected to w . The unmoving frame may be treated as a single node which will be called F .

Say strand i has length $L_i > 0$ when relaxed and linear density ρ_i . When the structure is at rest, say that string i is oriented in the direction of the unit vector $v_i \in \mathbb{R}^3$. For the structure to maintain its shape, each string must be under tension. Therefore, let string i be stretched by a factor of $s_i > 1$, so that it has a physical length of $s_i L_i$ when the structure is at rest. Hooke's constant for string i will be written as k_i , and k will be used when only one string is being studied or if all strings are of the same stiffness. However, different authors refer to several different quantities as "Hooke's constant." I will briefly describe these to simplify comparison with other work.

A quantity often called Hooke's constant is $\hat{k}$ in the expression

$$T = \hat{k} \Delta x, \quad (5.1)$$

where T is the magnitude of the restoring tension in the (uniformly stretched) string, $\Delta x = L_0(s - 1)$ is the change in length of the string from its relaxed state, and L_0 is the relaxed length of the string. This version of Hooke's constant is convenient for many problems with a single degree of freedom. The quantity $\hat{k}$ has the advantage of being constant even as the string is stretched, so it is used in the classic expression of harmonic motion for a point mass m attached to a massless elastic string, $m\ddot{x}(t) = -\hat{k}x(t)$. However, $\hat{k}$ may be different for strings made of the same material. If a string with constant $\hat{k}$ is cut in half, Hooke's constant for the half-string is $2\hat{k}$.

Some authors normalize Hooke's constant with the length of the string. Lagnese, Leugering, and Schmidt [33, Ch. II] use

$$\tilde{k} = \hat{k} L, \quad (5.2)$$

where $L = sL_0$ is the length of the taut string. For dynamic problems with a larger range of motion, the quantity $\tilde{k}$ might be more difficult to use than $\hat{k}$ or k because $\tilde{k}$ changes as the string stretches. However, there is no problem with using the quantity $\tilde{k}$ for modal calculations where motion (and stretching beyond the taut state) is assumed to be infinitesimal.

This chapter and the next will refer to the quantity

$$k = \hat{k}L_0 \quad (5.3)$$

as Hooke's constant, effectively normalizing by the *relaxed* length of the string. It is useful to consider k as a 1-dimensional analog of Young's modulus. For a 3-dimensional rod with cross-sectional area A , Young's modulus is defined by

$$E := \frac{\sigma}{\varepsilon}, \quad \sigma := \frac{T}{A}, \quad \varepsilon := \frac{\Delta x}{L_0}, \quad (5.4)$$

where σ is stress and ε is (engineering) strain. If the cross-sectional area is normalized to $A = 1$, combining (5.1), (5.3), and (5.4) gives $k = E$. Numerical results in Section 5.6 will be compared to those of Green [24] who also uses (5.3) as Hooke's constant, citing Cox [15]. Like Young's modulus under the assumption of constant A , the quantity k is not affected by stretching or by the length to which the string was cut during manufacture. These invariant properties are convenient when modeling many strings of the same material that are differently stretched.

The span of each string i will be parameterized by $\eta \in [0, L_i]$ with $\eta = 0$ referring to the position of node t_i and $\eta = L_i$ referring to the position of node h_i . It will be useful to be able to refer to the η value that corresponds to a node position in the context of a

particular mode function, so define

$$\ell_i(w) := \begin{cases} 0 & \text{if } w = t_i, \\ L_i & \text{if } w = h_i. \end{cases}$$

If edge i is not connected to node w , then $\ell_i(w)$ is not defined.

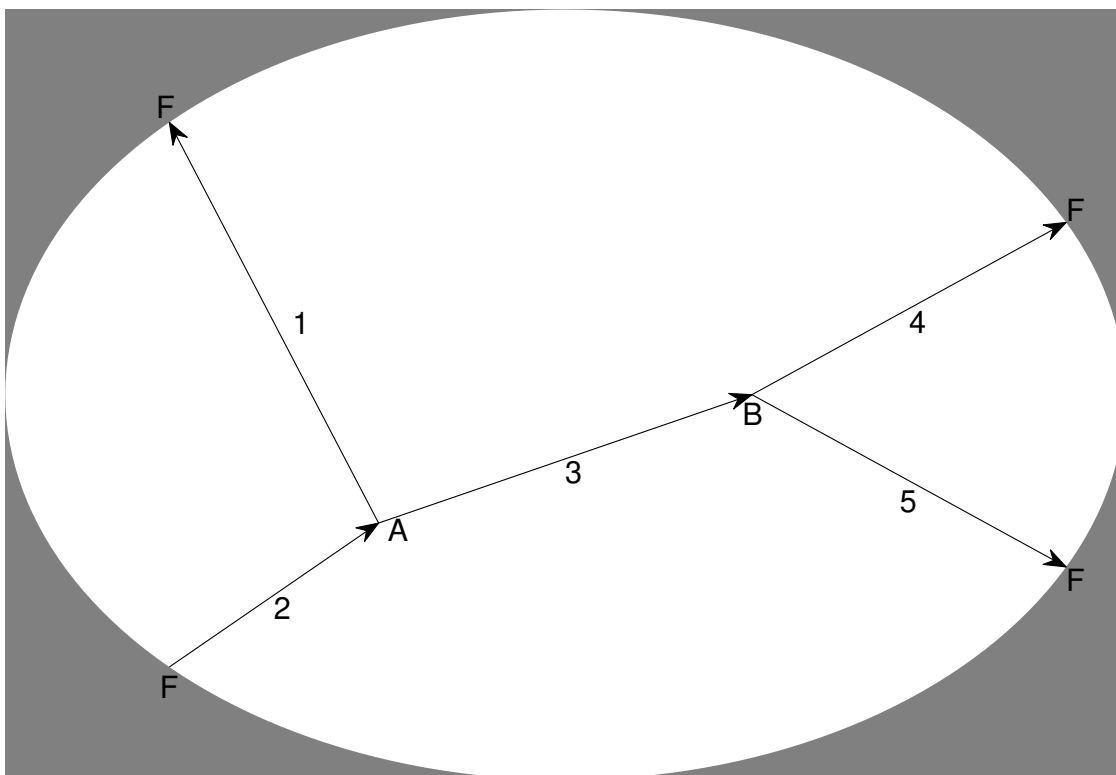


Figure 5.1: An example web structure described formally by (5.5).

For an example graph, consider Figure 5.1. This graph has three nodes called A , B , and F connected by five edges. (Remember that frame connection points will be treated as a single unmovable node, F .) The tail and head vectors, node neighborhoods, and ℓ values

are

$$\mathbf{t} = \begin{bmatrix} A \\ F \\ A \\ B \\ B \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} F \\ A \\ B \\ F \\ F \end{bmatrix}, \quad \begin{aligned} N(A) &= \{1, 2, 3\}, & \ell_1(A) &= 0, & \ell_1(F) &= L_2, \\ N(B) &= \{3, 4, 5\}, & \ell_2(A) &= L_2, & \ell_2(F) &= 0, \\ N(F) &= \{1, 2, 4, 5\}. & \ell_3(A) &= 0, & \ell_3(B) &= L_3, \\ & & \ell_4(B) &= 0, & \ell_4(F) &= L_4, \\ & & \ell_5(B) &= 0, & \ell_5(F) &= L_5. \end{aligned} \quad (5.5)$$

Assume that small vibrating deformations take place (in three dimensions) so that the movement of each strand follows the wave equation without damping. Let the displacement of string i (compared to its resting position) be $X_i(\eta) \in \mathbb{R}^3$, where $\eta \in [0, L_i]$ parameterizes the position along the resting string. Say that $X_i(0)$ is the displacement at the tail node t_i and $X_i(L_i)$ is the displacement at the head node h_i . When the structure vibrates according to one of its modes, the collection of X_i constitute the eigenfunction of the structure. In order to find the form of X_i corresponding to modal vibration, the shape of string i will be described as a product $r^{(i)}(\eta, t) = u(t)X_i(\eta) \in \mathbb{R}^3$, where u , an oscillating scalar function of time, is the same for all strings. A different u is applicable for each vibration mode.

Using subscripts to denote partial derivatives, the linear dynamics of string i , as derived in [33, Ch. II] and [42], are described by

$$\rho_i r_{tt}^{(i)}(\eta, t) = H_i r_{\eta\eta}^{(i)}(\eta, t), \quad (5.6)$$

where

$$H_i := k_i s_i v_i v_i^* + k_i (s_i - 1) I$$

and I is the 3×3 identity matrix. (Remember that v_i is the unit vector describing the orientation of string i .) Since $s_i > 1$, H_i is positive definite with eigenvalues $\gamma_{i,1} = k_i s_i$ and

$\gamma_{i,2} = \gamma_{i,3} = k_i(s_i - 1)$, so (5.6) is a hyperbolic partial differential equation.

Continuing from (5.6), separating variables gives

$$r^{(i)}(\eta, t) = e^{i\omega t} X_i(\eta), \quad (5.7)$$

and from (5.6),

$$-\omega^2 \rho_i X_i(\eta) = H_i X_i''(\eta). \quad (5.8)$$

Now diagonalize $H_i = V_i \Gamma_i V_i^T$ with $\Gamma_i = \text{diag}(\gamma_{i,1}, \gamma_{i,2}, \gamma_{i,3})$ and write $Y_i(\eta) := V_i^T X_i(\eta)$.

Then $X_i(\eta) = V_i Y_i(\eta)$, and premultiplying both sides of (5.8) by V_i^T produces

$$-\omega^2 \rho_i Y_i(\eta) = \Gamma_i Y_i''(\eta). \quad (5.9)$$

The general real solution of each row (5.9) provides an expression for the j th entry of the vector Y_i :

$$Y_{i,j}(\eta) = A_{i,j} \sin\left(\omega \eta \sqrt{\frac{\rho_i}{\gamma_{i,j}}}\right) + B_{i,j} \cos\left(\omega \eta \sqrt{\frac{\rho_i}{\gamma_{i,j}}}\right), \quad (5.10)$$

where $A_{i,j}$ and $B_{i,j}$ are constants whose values will be determined by the constraints (boundary conditions) imposed by the physical system. When the constraints are established, a collection of $A_{i,j}$ and $B_{i,j}$ values that satisfies all constraints simultaneously exists if and only if ω is a natural frequency of the network.

Next, constraints the of a string network will be enumerated. The number of constraints matches the number of degrees of freedom, and writing the constraints in (square) matrix form establishes an NLEVP whose eigenvalues are the natural frequencies of the system.

The node connectivity condition is that, for all free nodes $G \neq F$, there is a common displacement $d(G)$ so that for all strings $i \in N(G)$ connected to G , $X_i(\ell_i(G)) = d(G)$. These conditions are equivalent to $|N(G)| - 1$ vector equalities for each free node $G \neq F$.

The number of such equalities across all free nodes is then

$$\sum_{G \neq F} (|N(G)| - 1) = \sum_{G \neq F} |N(G)| - (n_n - 1).$$

(Recall that n_n denotes the number of nodes, including the frame, and n_e denotes the number of edges.) The sum of all node degrees (including the frame) is $\sum_G |N(G)| = 2n_e$, so the number of node connectivity conditions is

$$2n_e - |N(F)| - n_n + 1.$$

The fixed-frame condition is that $d(F) = 0$, or that for all $i \in N(F)$, $X_i(\ell_i(F)) = 0$. This gives one constraint for each edge connected to the frame, or $|N(F)|$ conditions.

The force-balance condition at each free node is that the sum of the tension forces at the node is zero. For each free node G ,

$$\sum_{i : t_i = G} H_i X'_i(0) - \sum_{i : h_i = G} H_i X'_i(L_i) = 0. \quad (5.11)$$

This gives one condition for each free node, or $n_n - 1$ conditions.

The total number of vector constraints is then

$$[2n_e - |N(F)| - (n_n - 1)] + |N(F)| + n_n - 1 = 2n_e. \quad (5.12)$$

For problems in D spatial dimensions, these vector conditions can be thought of as $2Dn_e$ scalar conditions.

Each string satisfies a second-order differential equation in D dimensions and so has $2D$ degrees of freedom. The degrees of freedom of string i are made explicit in the coeffi-

cients $A_{i,1}, B_{i,1}, A_{i,2}, \dots, B_{i,D}$. The overall system then has $2Dn_e$ degrees of freedom, which matches the total number of constraints.

Each boundary condition is linear in the undetermined coefficients (though nonlinear in ω). Consequently, the boundary conditions above can be written as a square NLEVP for the function $T(\omega) \in \mathbb{R}^{2Dn_e \times 2Dn_e}$ with the undetermined coefficients forming the eigenvector:

$$T(\omega) \begin{bmatrix} A_{1,1} \\ B_{1,1} \\ \vdots \\ B_{n_e,D} \end{bmatrix} = 0. \quad (5.13)$$

5.2.1 Code repository

To assist anyone wishing to use this method to study elastic webs, I have created a repository of code to support the framework just presented. It is available at

https://github.com/jpeterbaker/nlevp_webs.

The repository includes general code for constructing the NLEVP associated with any elastic web, as well as several example problems. The examples are the three-string studied in detail in the rest of this chapter, “regular” spider webs with equally spaced spokes and rings, and two “realistic” spider webs based on tracings of actual webs presented in [24].

The repository includes code to assist in visualizing mode shapes of string networks. As an example of the capability of the tools, Figures 5.2 and 5.3 show modes of a “regular” web that were calculated with the code of the repository.

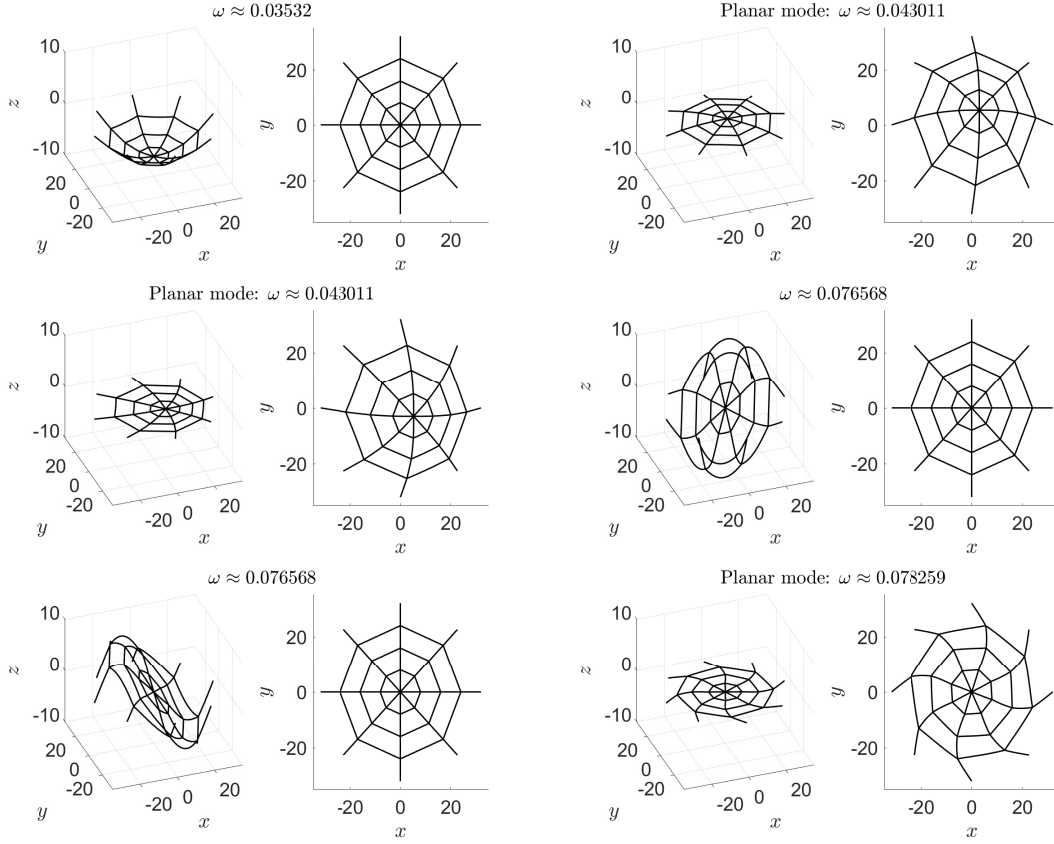


Figure 5.2: The first six mode shapes of a simple regular spider web.

5.2.2 Damped systems

It is not difficult to add a viscous damping force to (5.6). For a thorough discussion of viscously damped wave equations, see [14]. If constant viscous damping is applied naively in all directions, the damped PDE for string i is

$$\rho_i r_{tt}^{(i)}(\eta, t) + c_i r_t^{(i)}(\eta, t) = H_i r_{\eta\eta}^{(i)}(\eta, t), \quad (5.14)$$

where $c_i \geq 0$ is the damping coefficient of string i .

For the undamped wave equation of Section 5.3, it was convenient to refer to the real value ω as the system eigenvalue since ω represents a natural frequency. For damped systems,

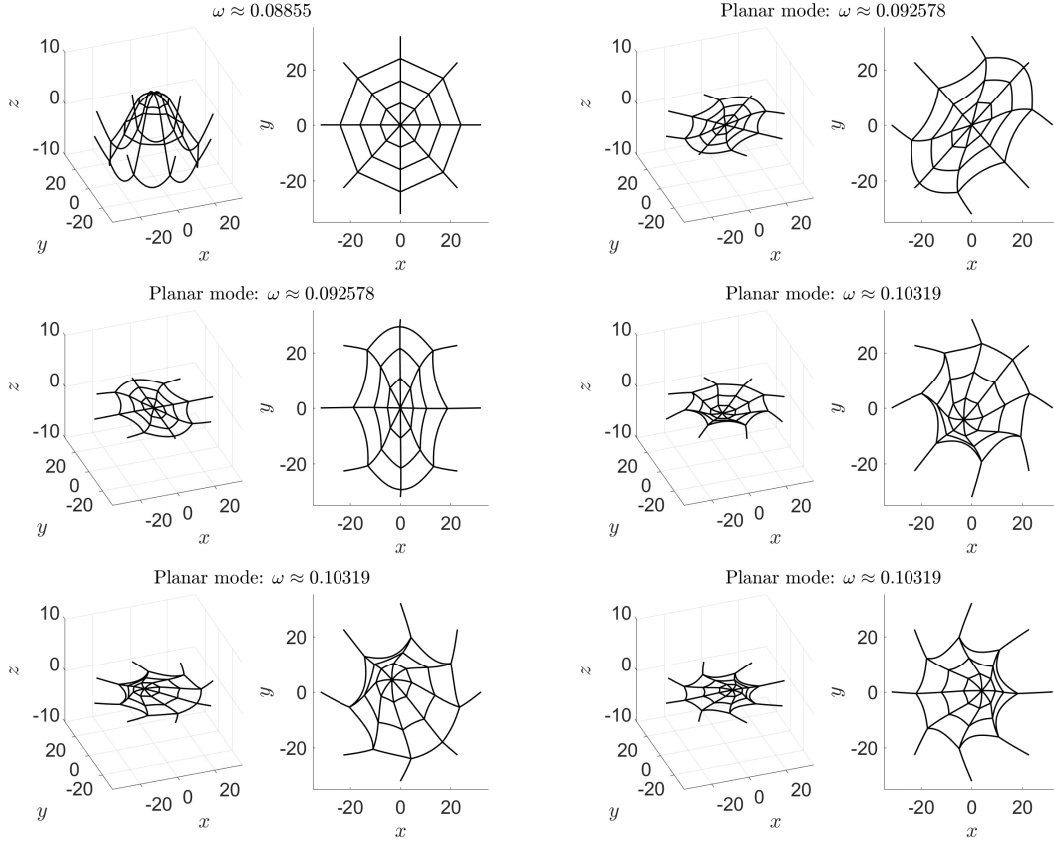


Figure 5.3: Modes 7–12 of a simple regular spider web.

I will use $\lambda = i\omega$ to represent eigenvalues. As long as each $c_i \geq 0$, λ will lie in the left half-plane. If each $c_i = 0$, the system is undamped and every eigenvalue λ will lie on the imaginary axis.

In λ notation, displacement $r^{(i)}$ is decomposed as

$$r^{(i)}(\eta, t) = e^{\lambda t} X_i(\eta),$$

and the damped version of (5.8) is

$$(\lambda^2 \rho_i + \lambda c_i) X_i(\eta) = H_i X_i''(\eta). \quad (5.15)$$

Defining $Y_i := V_i^T X_i$ as before, the j th entry of Y_i is now

$$Y_{i,j}(\eta) = A_{i,j} \sin\left(\eta \sqrt{\frac{\lambda^2 \rho_i + \lambda c_i}{-\gamma_{i,j}}}\right) + B_{i,j} \cos\left(\eta \sqrt{\frac{\lambda^2 \rho_i + \lambda c_i}{-\gamma_{i,j}}}\right). \quad (5.16)$$

The constraints can now be assembled into an NLEVP just as in Section 5.2.

Proposition 5.1. *Let there be a connected network of elastic strings modeled as above. Let the strings have the same linear density ρ and the same viscous damping coefficient c . Let T be the NLEVP matrix function associated with the system, and let T_0 be the NLEVP matrix function associated with the undamped version of the same system (i.e., $c = 0$). Let λ_0 be an eigenvalue of T_0 . Let λ be either of the values*

$$\lambda = \frac{-c}{2\rho} \pm \sqrt{\lambda_0^2 + \frac{c^2}{4\rho^2}}. \quad (5.17)$$

Then $T(\lambda) = T_0(\lambda_0)$ up to the arbitrary choices involved in the construction of these NLEVPs, i.e., the bases V_i ($i = 1, \dots, n_e$), the orders of the rows and columns, and the signs of the square-root terms. In particular, λ is an eigenvalue of T .

Proof. Every appearance of λ and c in (5.16) is part of the expression

$$g(\lambda, c) := \lambda^2 \rho + \lambda c. \quad (5.18)$$

Likewise, λ and c affect $Y'_{i,j}(\eta)$ only through (5.18). Since evaluations of $Y_{i,j}$ and $Y'_{i,j}$ make up the non-constant entries of $T(\lambda)$, T depends on λ and c only through (5.18).

To finish the proof, it is only a matter of algebra to find λ satisfying $g(\lambda, c) = g(\lambda_0, 0)$, so that $T(\lambda) = T_0(\lambda_0)$. If we set

$$\lambda^2 \rho + \lambda c = g(\lambda, c) = g(\lambda_0, 0) = \lambda_0^2 \rho,$$

then the quadratic formula provides the two solutions in (5.17). $\square$

The critical damping level for each mode (or eigenvalue) is the value of c that nullifies the portion of (5.17) under the radical, specifically $c = 2\rho\Im(\lambda_0)$. If the damping coefficient is $c < 2\rho\Im(\lambda_0)$, then the corresponding (under)damped eigenvalue λ has the imaginary part $\Im(\lambda) = \pm\sqrt{\Im(\lambda_0)^2 - \frac{c^2}{4\rho}}$. (Remember that λ_0 is on the imaginary axis since it is an eigenvalue of an undamped system.) Furthermore,

$$|\lambda|^2 = \left(\frac{-c}{2\rho}\right)^2 + \Im(\lambda_0)^2 - \frac{c^2}{4\rho} = |\lambda_0|^2,$$

and thus subcritical damping does not affect the modulus of an eigenvalue. Figure 5.4 shows the damped eigenvalues of the three-string problem introduced in Section 5.3. The circular arcs in the top plot of the figure illustrate that, as damping increases, the modulus of each eigenvalue is constant until damping surpasses the critical value for that eigenvalue. When damping reaches the critical value for an eigenvalue, that eigenvalue meets its conjugate pair on the real axis. If damping increases further, then that mode is overdamped, and the pair of eigenvalues remains on the real axis, moving apart from each other. The bottom plots of Figure 5.4 illustrate this behavior by showing the real and imaginary portions of the eigenvalues separately.

Eigenvalues of critically damped and overdamped systems are on the real axis, and contours that contain these eigenvalues must cross the real axis. Contour integral methods require that the NLEVP matrix function be analytic, but the square-root expressions in (5.16) are not analytic for some real λ , particularly at $\lambda = -c/\rho_i$. Consequently, it may be difficult to find critical and overdamped eigenvalues via the NLEVP formulation (5.13) and contour integral methods. Instead, these methods may be used to find undamped eigenvalues (λ_0), and (5.17) may be used to calculate corresponding damped eigenvalues.

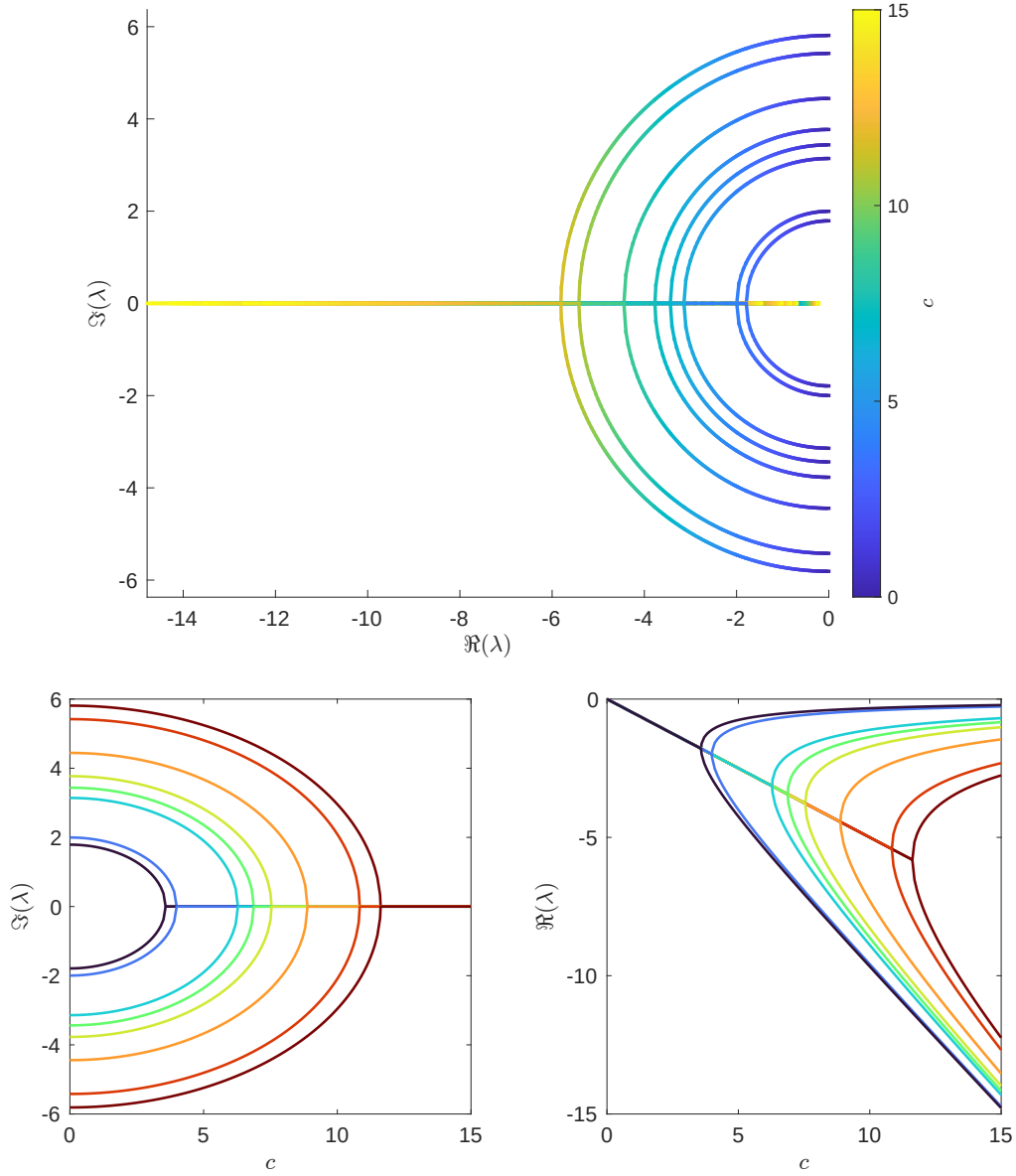


Figure 5.4: **Top:** Eigenvalues of the three-string problem drawn in the complex plane, with color indicating the damping level, c . **Bottom:** Imaginary (left) and real (right) parts of three-string eigenvalues as functions of c , with a different color used for each eigenvalue. As damping increases, conjugate pairs of (underdamped) eigenvalues meet on the real axis, and two real (overdamped) eigenvalues are formed.

It might be expected that damping will not affect all directions of motion equally. In particular, if damping is caused by viscosity of a surrounding fluid, then a string with a small axial profile should experience little resistance to its motion along its axis, and it might be more accurate to consider damping only in transverse directions. In other words, the motion of string i in the direction v_i is undamped. Such direction-dependent damping can be incorporated by introducing more damping parameters $c_{i,j}$ for $i = 1, \dots, n_e$ and $j = 1, \dots, D$. As long as the damping directions are aligned with the orientation of the strings, the new dynamics are described by

$$\rho_i r_{tt}^{(i)}(\eta, t) + C_i r_t^{(i)}(\eta, t) = H_i r_{\eta\eta}^{(i)}(\eta, t), \quad (5.19)$$

where

$$C_i := V_i \text{diag}(c_{i,1}, \dots, c_{i,D}) V_i^T$$

is the damping matrix for string i . Since C_i and H_i are simultaneously diagonalizable, the homogeneous solution of the new system (5.19) is very similar to (5.16) except that $c_{i,j}$ appears on row j instead of c_i . If the orientation v_i of string i is the first column of V_i , then transverse-only damping is represented by setting $c_{i,1} = 0$.

Figure 5.5 compares the two damping models just discussed for the three-string problem introduced in Section 5.3. In the left plot, damping is applied equally in all directions (so the eigenvalues shown match those in Figure 5.4). In the right plot, damping is only applied in the transverse direction. In both cases, the eigenvalues were calculated by the multi-shift contour integral method discussed in Section 4.2.3. In this case, Gauss–Legendre quadrature was used on each segment of a rectangular contour. The contour is entirely in the top half-plane to avoid crossing the real axis where the NLEVP is not analytic. As damping increases, the eigenvalues of the omnidirectionally damped problem change as predicted by (5.17), but the transversely damped eigenvalues do not follow an obvious

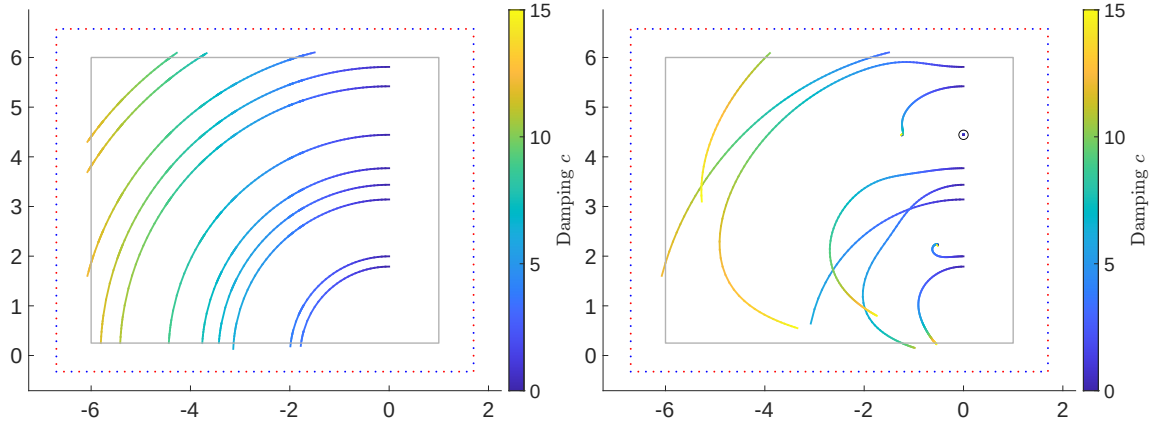


Figure 5.5: Eigenvalue loci of a three-string for a range of damping values. **Left:** damping applies equally to all directions of movement. **Right:** damping applies only to transverse movement. The gray lines show the integration contour used to find the eigenvalues, and the blue and red dots show the left and right interpolation points. On the right, the unmoving sixth eigenvalue is circled for visibility. The eigenvalue loci extend slightly beyond the contour because some eigenvalues were found outside of the contour, but this may not happen in general.

pattern. Of special note is the sixth eigenvalue, which corresponds to a compression mode with no transverse motion. Transverse damping does not affect this mode, so its eigenvalue does not move in the plot as damping changes.

5.3 Three-string example

A “tritare” is an experimental stringed musical instrument introduced in [21]. Each string of a tritare consists of three taut strings that are connected at a point (a “three-string”). The other end of each individual string is mounted to the instrument frame. In [21], Gaudet et al. analyzed the natural frequencies of vibration of such three-strings, which, unlike the overtones of ordinary single vibrating strings, do not form a harmonic sequence of rational multiples of a fundamental mode. Three-strings are a useful introductory example for understanding vibrating webs because, despite their simplicity, they exhibit theoretical subtleties that do not arise in the study of individual vibrating strings.

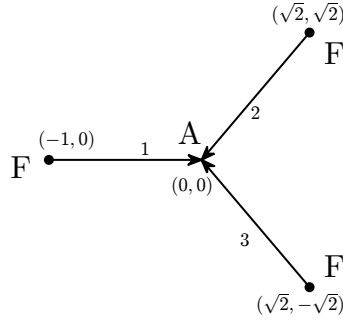


Figure 5.6: The three-string network studied in this section.

Consider the three-string illustrated in Figure 5.6.

For simplicity, only two dimensions are used in this example. The string orientation vectors are

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad v_3 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

All three strings meet at a node called A , and the other end of each string is connected to the frame F . In the notation of the previous section, $\ell_i(F) = 0$ and $\ell_i(A) = 1$ for $i = 1, 2, 3$.

For this example, the physical properties of the strings are identical, specifically $L = 1$, $\rho = 1$, $k = 1$, and $s = 2$, where the subscripts have been omitted for simplicity. This setting is mathematically convenient, but note that it is not physically feasible. Physically identical strings that are stretched by the same factor exert tension forces of the same magnitude $\mathcal{T}$. When arranged as in Figure 5.6, the horizontal component of the net tension force acting on the center point is nonzero. Specifically, the net force at the center point is

$$\mathcal{T} \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \mathcal{T} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} + \mathcal{T} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \mathcal{T} \begin{bmatrix} \sqrt{2} - 1 \\ 0 \end{bmatrix}.$$

To physically realize the vibrations of such a system, a constant canceling external force would have to be applied to the center point without adding mass or otherwise interfering with vibrations.

Since the strings have the same physical properties, and each H_i matrix will have the same eigenvalues, $\gamma_1 = sk = 2$ and $\gamma_2 = k(s - 1) = 1$, where the second subscript has been omitted. Similarly, the subscript may be left off of the diagonal matrix

$$\Gamma = \text{diag}(\gamma_1, \gamma_2) = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix},$$

which is the same for all strings. As in Section 5.2, an orthogonal matrix V_i must be formed for each string by supplementing each orientation vector v_i with an orthogonal vector: $V_i = \begin{bmatrix} v_i & v_i^\perp \end{bmatrix}$. Conveniently, the signs of the columns of each V_i can be chosen to make V_i symmetric, and then the matrices $H_i = V_i \Gamma V_i^T$ can be formed. Explicitly, the H_i and V_i matrices are

$$\begin{aligned} H_1 &= \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, & H_2 &= \begin{bmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{3}{2} \end{bmatrix}, & H_3 &= \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix}, \\ V_1 &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & V_2 &= \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix}, & V_3 &= \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}. \end{aligned}$$

The Dirichlet boundary conditions at the frame node are $X_i(0) = 0$ for $i = 1, 2, 3$, and so $Y_i(0) = V_i^T X_i(0) = 0$ as well. Using this fact in (5.10) reveals that $B_{i,j} = 0$ for $i = 1, 2, 3$ and $j = 1, 2$.

The continuity conditions at node A can be expressed as

$$\begin{aligned} 0 &= X_1(1) - X_2(1) \\ &= \begin{bmatrix} A_{1,1} \sin\left(\frac{\omega}{\sqrt{2}}\right) \\ A_{1,2} \sin(\omega) \end{bmatrix} - \frac{1}{\sqrt{2}} \begin{bmatrix} -A_{2,1} \sin\left(\frac{\omega}{\sqrt{2}}\right) - A_{2,2} \sin(\omega) \\ -A_{2,1} \sin\left(\frac{\omega}{\sqrt{2}}\right) + A_{2,2} \sin(\omega) \end{bmatrix}, \end{aligned} \quad (5.20a)$$

$$\begin{aligned}
0 &= X_1(1) - X_3(1) \\
&= \begin{bmatrix} A_{1,1} \sin\left(\frac{\omega}{\sqrt{2}}\right) \\ A_{1,2} \sin(\omega) \end{bmatrix} - \frac{1}{\sqrt{2}} \begin{bmatrix} -A_{2,1} \sin\left(\frac{\omega}{\sqrt{2}}\right) + A_{2,2} \sin(\omega) \\ A_{2,1} \sin\left(\frac{\omega}{\sqrt{2}}\right) + A_{2,2} \sin(\omega) \end{bmatrix}.
\end{aligned} \tag{5.20b}$$

Next, the force-balancing condition at node A is

$$H_1 X'_1(1) + H_2 X'_2(1) + H_3 X'_3(1) = 0,$$

or equivalently

$$V_1 \Gamma Y'_1(1) + V_2 \Gamma Y'_2(1) + V_3 \Gamma Y'_3(1) = 0. \tag{5.21}$$

The derivative of the j th row of Y_i is

$$Y'_{i,j}(\eta) = \omega A_{i,j} \sqrt{\frac{1}{\gamma_j}} \cos\left(\omega \eta \sqrt{\frac{1}{\gamma_j}}\right).$$

Substituting this expression into (5.21),

$$\begin{aligned}
&\omega \left(\begin{bmatrix} A_{1,1} \sqrt{2} \cos\left(\frac{\omega}{\sqrt{2}}\right) \\ A_{1,2} \cos(\omega) \end{bmatrix} \right. \\
&\quad + \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} A_{2,1} \sqrt{2} \cos\left(\frac{\omega}{\sqrt{2}}\right) \\ A_{2,2} \cos(\omega) \end{bmatrix} \\
&\quad \left. + \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} A_{3,1} \sqrt{2} \cos\left(\frac{\omega}{\sqrt{2}}\right) \\ A_{3,2} \cos(\omega) \end{bmatrix} \right) = 0.
\end{aligned} \tag{5.22}$$

The conditions in (5.20) and (5.22) can now be combined into the order-6 NLEVP

$$\begin{bmatrix} \sin \omega & 0 & \frac{1}{\sqrt{2}} \sin \omega & \frac{1}{\sqrt{2}} \sin \frac{\omega}{\sqrt{2}} & 0 & 0 \\ 0 & \sin \frac{\omega}{\sqrt{2}} & \frac{1}{\sqrt{2}} \sin \omega & -\frac{1}{\sqrt{2}} \sin \frac{\omega}{\sqrt{2}} & 0 & 0 \\ \sin \omega & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} \sin \omega & \frac{1}{\sqrt{2}} \sin \frac{\omega}{\sqrt{2}} \\ 0 & \sin \frac{\omega}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \sin \omega & \frac{1}{\sqrt{2}} \sin \frac{\omega}{\sqrt{2}} \\ \omega \sqrt{2} \cos \frac{\omega}{\sqrt{2}} & 0 & -\omega \cos \frac{\omega}{\sqrt{2}} & -\frac{\omega}{\sqrt{2}} \cos \omega & -\omega \cos \frac{\omega}{\sqrt{2}} & \frac{\omega}{\sqrt{2}} \cos \omega \\ 0 & \omega \cos \omega & -\omega \cos \frac{\omega}{\sqrt{2}} & \frac{\omega}{\sqrt{2}} \cos \omega & \omega \cos \frac{\omega}{\sqrt{2}} & \frac{\omega}{\sqrt{2}} \cos \omega \end{bmatrix} \begin{bmatrix} A_{1,1} \\ A_{1,2} \\ A_{2,1} \\ A_{2,2} \\ A_{3,1} \\ A_{3,2} \end{bmatrix} = 0. \quad (5.23)$$

Remember that (5.12) predicted that a three-string problem in $D = 2$ dimensions should have 12 degrees of freedom (and an order-12 NLEVP). The six of degrees of freedom corresponding to the B -coefficients were omitted for brevity since the B -coefficients are all zero for this system. This convenience is a result of every string being having its “tail” connected to the frame where a homogeneous Dirichlet boundary condition is imposed. The degrees of freedom are not so easily eliminated for strings that connect two free nodes, such as string 3 in Figure 5.1.

5.4 Deriving modes of three-string example

This section algebraically finds the natural frequencies of the undamped three-string system described in Section 5.3. It will be simpler to derive the modes directly from the constraints, rather than starting from the NLEVP (5.23).

As Section 5.3 established for this special example, the Dirichlet boundary conditions on the frame imply $B_{i,j} = 0$ for each i, j . These conditions and the B -coefficients will not need to be referenced again.

The central continuity condition $X_1(1) = X_2(1) = X_3(1)$ can be written as

$$V_1 \begin{bmatrix} A_{1,1} \sin \omega \\ A_{1,2} \sin \frac{\omega}{\sqrt{2}} \end{bmatrix} = V_2 \begin{bmatrix} A_{2,1} \sin \omega \\ A_{2,2} \sin \frac{\omega}{\sqrt{2}} \end{bmatrix} = V_3 \begin{bmatrix} A_{3,1} \sin \omega \\ A_{3,2} \sin \frac{\omega}{\sqrt{2}} \end{bmatrix}. \quad (5.24)$$

Consider first that the four scalar conditions expressed in (5.24) are all satisfied if $\sin \omega = 0$ and $A_{i,2} = 0$ for all i . There remain only two scalar conditions of the force-balance equation (5.22). These conditions are underdetermined in the *three* remaining variables ($A_{i,1}$ for all i), and so there is a nontrivial solution to these conditions corresponding to an eigenvector of (5.23). Consequently, all ω satisfying

$$\sin \omega = 0 \quad (5.25)$$

are natural frequencies. A parallel argument holds for ω satisfying

$$\sin \frac{\omega}{\sqrt{2}} = 0. \quad (5.26)$$

In order to find other natural frequencies, assume for the rest of the calculation that (5.25) and (5.26) are not satisfied.

Expanding and expressing all coefficients in terms of $A_{2,1}$, $A_{2,2}$ gives

$$A_{1,1} = \frac{1}{\sqrt{2}} \left(-A_{2,1} - A_{2,2} \frac{\sin \omega}{\sin \frac{\omega}{\sqrt{2}}} \right) \quad (5.27a)$$

$$A_{1,2} = \frac{1}{\sqrt{2}} \left(-A_{2,1} \frac{\sin \frac{\omega}{\sqrt{2}}}{\sin \omega} + A_{2,2} \right) \quad (5.27b)$$

$$A_{3,1} = A_{2,2} \frac{\sin \omega}{\sin \frac{\omega}{\sqrt{2}}} \quad (5.27c)$$

$$A_{3,2} = -A_{2,1} \frac{\sin \frac{\omega}{\sqrt{2}}}{\sin \omega}. \quad (5.27d)$$

The force-balance condition (5.11) for this system is

$$0 = \omega \left(V_1 \begin{bmatrix} A_{1,1}\sqrt{2}\cos\frac{\omega}{\sqrt{2}} \\ A_{1,2}\cos\omega \end{bmatrix} + V_2 \begin{bmatrix} A_{2,1}\sqrt{2}\cos\frac{\omega}{\sqrt{2}} \\ A_{2,2}\cos\omega \end{bmatrix} + V_3 \begin{bmatrix} A_{3,1}\sqrt{2}\cos\frac{\omega}{\sqrt{2}} \\ A_{3,2}\cos\omega \end{bmatrix} \right). \quad (5.28)$$

The case $\omega = 0$ is included in both (5.25) and (5.26), so the factor of ω may be ignored.

Expand the terms multiplied the V_i matrices and substitute (5.27) into the top row of (5.28) to find

$$\begin{aligned} 0 = & \left[\frac{1}{\sqrt{2}} \left(-A_{2,1} - A_{2,2} \frac{\sin\omega}{\sin\frac{\omega}{\sqrt{2}}} \right) \right] \sqrt{2}\cos\frac{\omega}{\sqrt{2}} \\ & + \frac{1}{\sqrt{2}} \left(-A_{2,1}\sqrt{2}\cos\frac{\omega}{\sqrt{2}} - A_{2,2}\cos\omega \right) \\ & + \frac{1}{\sqrt{2}} \left(- \left[A_{2,2} \frac{\sin\omega}{\sin\frac{\omega}{\sqrt{2}}} \right] \sqrt{2}\cos\frac{\omega}{\sqrt{2}} + \left[-A_{2,1} \frac{\sin\frac{\omega}{\sqrt{2}}}{\sin\omega} \right] \cos\omega \right). \end{aligned} \quad (5.29)$$

Simplifying and factoring (5.29) yields

$$0 = - \left(A_{2,1} \sin\frac{\omega}{\sqrt{2}} + A_{2,2} \sin\omega \right) \left(2 \cot\frac{\omega}{\sqrt{2}} + \frac{1}{\sqrt{2}} \cot\omega \right). \quad (5.30)$$

Substituting (5.27) into the bottom row of (5.28) produces

$$\begin{aligned} 0 = & \left[\frac{1}{\sqrt{2}} \left(-A_{2,1} \frac{\sin\frac{\omega}{\sqrt{2}}}{\sin\omega} + A_{2,2} \right) \right] \cos\omega \\ & + \frac{1}{\sqrt{2}} \left(-A_{2,1}\sqrt{2}\cos\frac{\omega}{\sqrt{2}} + A_{2,2}\cos\omega \right) \\ & + \frac{1}{\sqrt{2}} \left(\left[A_{2,2} \frac{\sin\omega}{\sin\frac{\omega}{\sqrt{2}}} \right] \sqrt{2}\cos\frac{\omega}{\sqrt{2}} + \left[-A_{2,1} \frac{\sin\frac{\omega}{\sqrt{2}}}{\sin\omega} \right] \cos\omega \right). \end{aligned} \quad (5.31)$$

Similar to the previous calculation, (5.31) may be simplified and factored as

$$0 = -\sqrt{2} \left(A_{2,1} \sin\frac{\omega}{\sqrt{2}} - A_{2,2} \sin\omega \right) \left(\sqrt{2} \cot\omega + \cot\frac{\omega}{\sqrt{2}} \right). \quad (5.32)$$

Thus (5.27) is equivalent to (5.30) and (5.32) together. Besides the trivial solution $A_{2,1} = A_{2,2} = 0$, there are two possibilities. Either

$$2 \cot \frac{\omega}{\sqrt{2}} + \frac{1}{\sqrt{2}} \cot \omega = 0 \quad (5.33)$$

and $\left(A_{2,1} \sin \frac{\omega}{\sqrt{2}} - A_{2,2} \sin \omega\right) = 0$, or else

$$\sqrt{2} \cot \omega + \cot \frac{\omega}{\sqrt{2}} = 0 \quad (5.34)$$

and $\left(A_{2,1} \sin \frac{\omega}{\sqrt{2}} + A_{2,2} \sin \omega\right) = 0$. All of the natural frequencies satisfy one of (5.25), (5.26), (5.33), or (5.34), which are the transcendental equations used to find true eigenvalues of this three-string system in Section 5.3.

Figures 5.7, 5.8, and 5.9 show two perspectives of each of the first 30 mode shapes of the three-string example. The mode shapes that remain in the x - y plane are labeled as “planar” for clarity. The dots plotted on the modes are *not* a representation of any kind of discretization to aid numerical calculations—a major advantage of the NLEVP formulation presented here is that it avoids discretization. Instead, dots are plotted on the modes to make compressed regions easier to distinguish. In particular, the modes of $\omega \approx 4.4429$ and $\omega \approx 8.8858$ have only longitudinal displacement for each string, which is visible in the unequal spacing of the dots in those diagrams. Some natural frequencies appear in the figures multiple times because multiple linearly independent mode shapes correspond to those frequencies. Any mode shape in the span of those pictured is also a mode shape. In other words, if ω is an eigenvalue of T , then *any* member of the space $\ker(T(\omega))$ corresponds to a mode shape. The mode shapes in the figures correspond to an arbitrary basis for $\ker(T(\omega))$. The first 24 modes of the three-string example presented here are also shown in [24, Figs 2.5–2.7], as calculated with Chebyshev pseudospectral collocation.

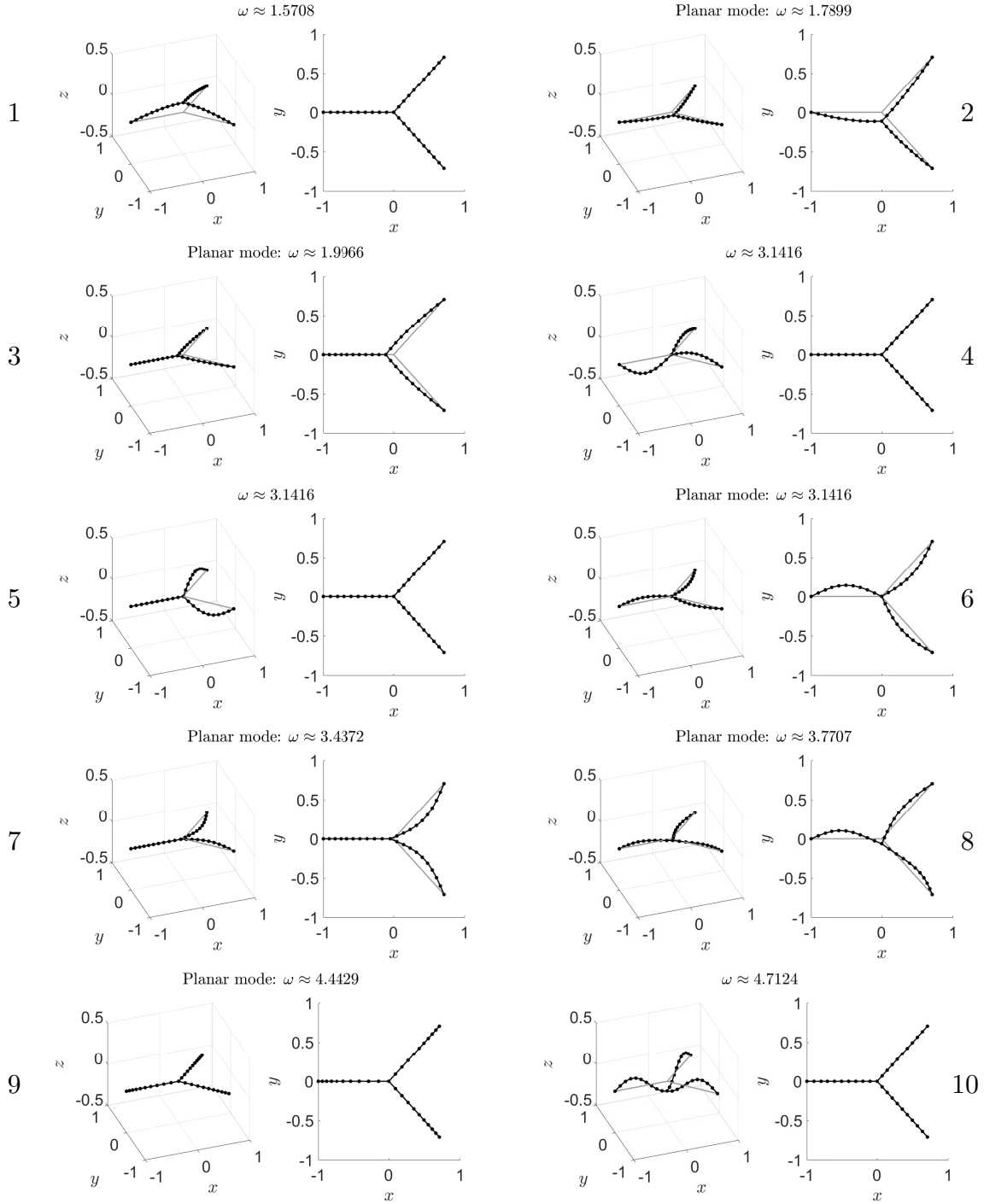


Figure 5.7: The first ten mode shapes of an example three-string.

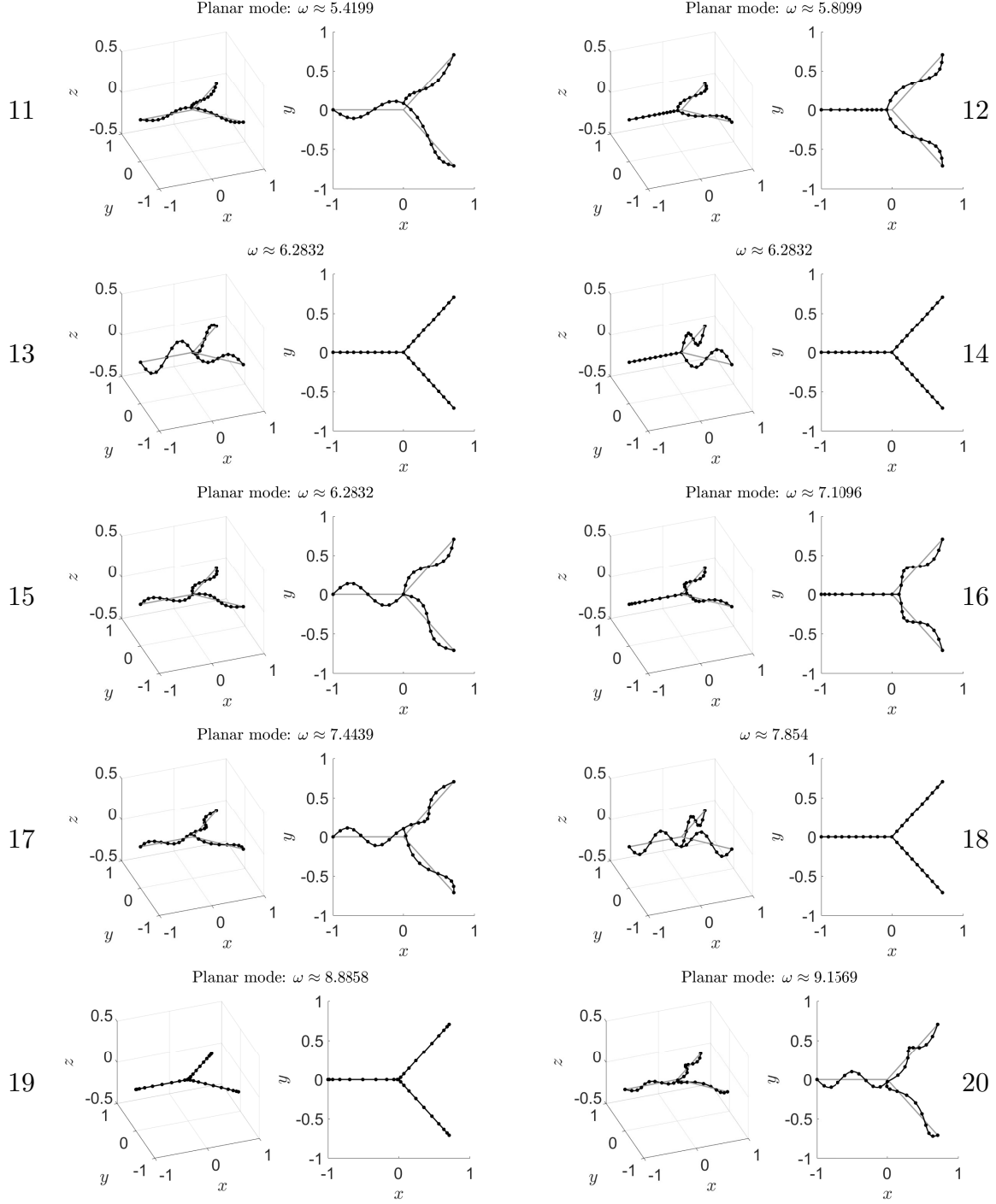


Figure 5.8: Mode shapes 11–20 of an example three-string.

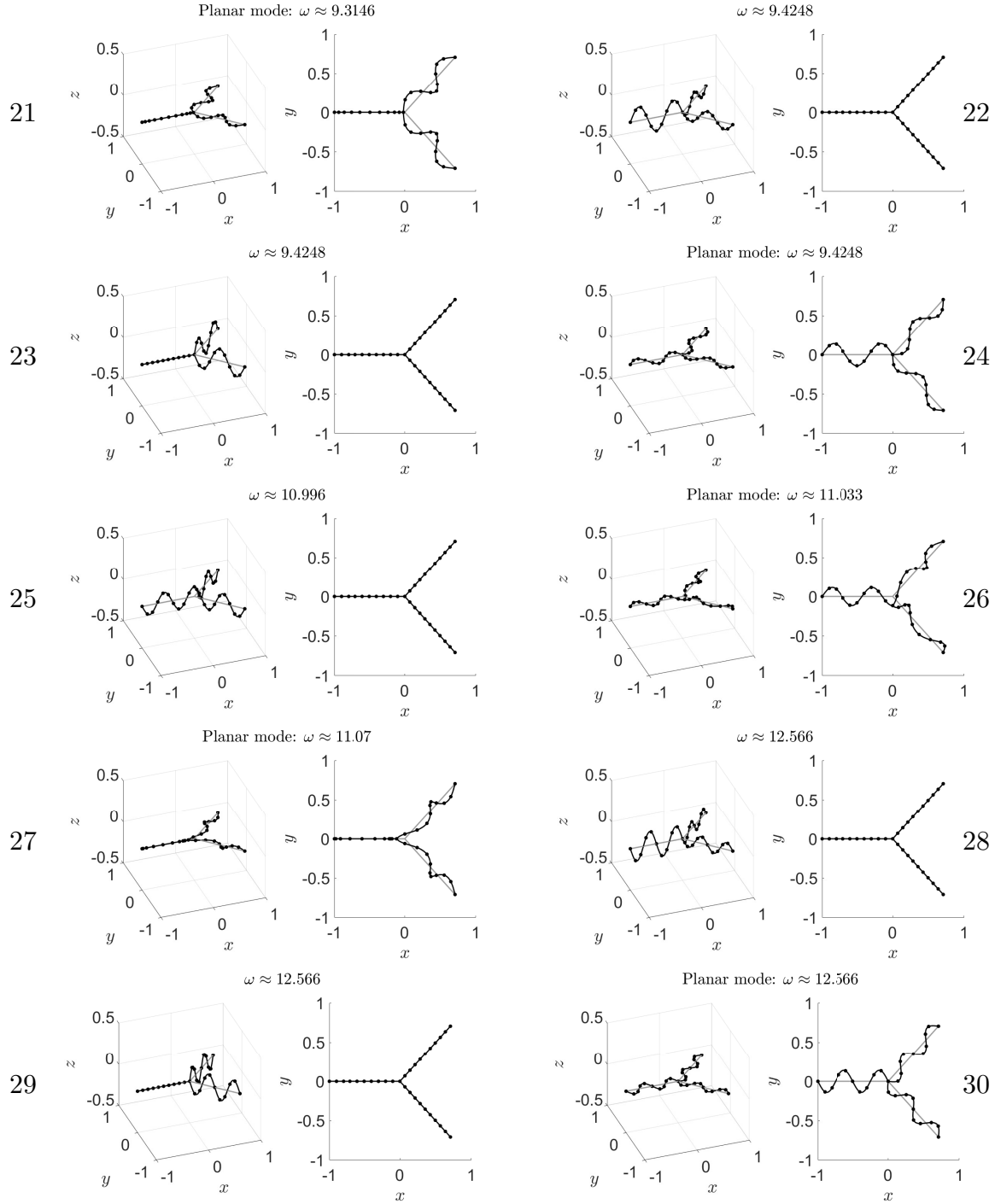


Figure 5.9: Mode shapes 21–30 of an example three-string.

5.5 Numerical studies of three-strings

5.5.1 Parameterized three-string model

Eigenvalue “avoidance” or “veering” is classic phenomenon of parameterized linear eigenvalue problems. As a Hermitian matrix is transformed continuously into another, the loci of the (real) eigenvalues do not typically intersect. Rather, the eigenvalues seem to avoid each other. The eigenvalues of real (non-Hermitian) matrices are real or appear in conjugate pairs, and the conjugate pairs often meet on the real axis if a matrix is modified. This behavior is notably explored in [39] and [48].

It is not surprising to find similar eigenvalue avoidance in parameterized NLEVPs. Teytel observed that eigenvalue crossings are rare for self-adjoint operators on function spaces [45]. Since the NLEVPs studied in this chapter are designed to have the same eigenvalues as such operators, we can expect multiple eigenvalues to be atypical. In [10], Beyn et al. observed eigenvalue avoidance except at the meeting of conjugate pairs for “real” NLEVPs, defined as those satisfying $T(\bar{z}) = \overline{T(z)}$ for all $z \in \mathbb{C}$. Examining (5.16) shows that even the damped versions of the NLEVPs of this chapter are what Beyn calls “real” (conjugating λ in (5.16) produces the conjugate of $Y_{i,j}$ from which T is built). Thus, eigenvalue veering is expected if these NLEVPs are parameterized. Relatedly, Bindel and Hood extend several classic eigenvalue localization theorems to nonlinear problems, bounding the movement of eigenvalues as the problem changes [11].

Furthermore, in linear parameterized eigenvalue problems, mode shapes tend to “jump” between eigenvalues. For example, consider a matrix A and pencil $A + \varepsilon B$ for small $\varepsilon > 0$. To see this clearly, consider the linear eigenvector perturbation calculations of Wilkinson [49, p. 70]. Say the eigenvalues of A are all simple $\lambda_1, \dots, \lambda_n$ and have corresponding right eigenvectors $x_1, \dots, x_n$ and left eigenvectors $y_1, \dots, y_n$. Wilkinson finds that, up to a

scaling factor, the eigenvector of $A + \varepsilon B$ that corresponds to the eigenvector x_1 of A is

$$x_1(\varepsilon) = x_1 + \varepsilon \left[\sum_{i=2}^n \frac{\beta_{i,1} x_i}{(\lambda_1 - \lambda_i) s_i} \right] + O(\varepsilon^2).$$

The presence of $(\lambda_1 - \lambda_i)$ in the denominator suggests that, if two eigenvalues have come near one another ($\lambda_1 \approx \lambda_i$), then their eigenvectors will dominate one another's perturbed eigenvectors. Figure 5.10 illustrates a similar phenomenon for the NLEVPs of a parameterized three-string introduced later.

One way to parameterize the vibration-derived NLEVPs of this chapter is to vary the damping parameter c . The effects of two forms of parameterized damping on the three-string example are illustrated Figure 5.4. When damping is applied in all directions, the conjugate pairs of eigenvalues meet on the real axis when critical damping is reached for that eigenvalue. This is similar to the meeting of conjugate pairs for families of real non-Hermitian matrices.

The rest of this section illustrates eigenvalue avoidance with an undamped parameterized model. In this section, physical plausibility will be enforced by requiring that the tension forces on the junction point balance even in the resting state. This requirement does not allow the example three-string described in Section 5.3, which was chosen to be algebraically convenient rather than realistic.

The three-string model can be altered to produce parameterized NLEVPs. If the position of one arm of the three-string is fixed and the position of another arm is changed, then the third string should be moved to balance the tension forces. To be specific, suppose that the strings are arranged as in Figure 5.11 so that

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} -\cos \phi_2 \\ -\sin \phi_2 \end{bmatrix}, \quad v_3 = \begin{bmatrix} -\cos \phi_3 \\ \sin \phi_3 \end{bmatrix}. \quad (5.35)$$

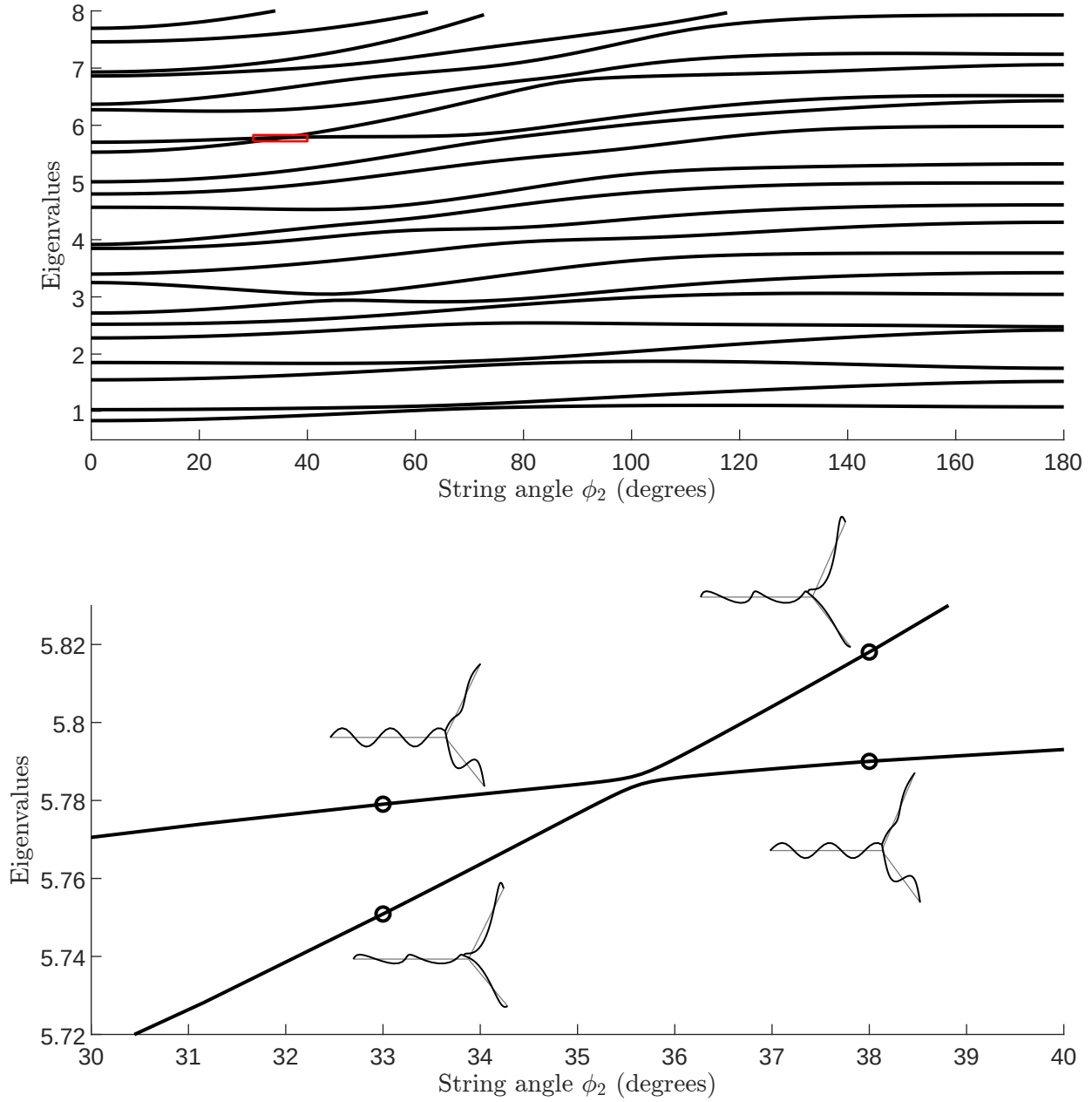


Figure 5.10: Natural frequencies of the three-string shown in Figure 5.11 for a range of ϕ_2 values. The bottom plot is a detailed view of the small region marked with a red rectangle in the top plot where there is a “near miss” between the 15th and 16th eigenvalues. Overlaid on the bottom plot are the mode shapes of both eigenvalues for $\phi_2 = 33^\circ$ and 38° . The shapes of the modes appear to “jump” between eigenvalues: the shape of the higher mode at $\phi_2 = 33^\circ$ resembles the shape of the lower mode at $\phi_2 = 38^\circ$ and vice versa.

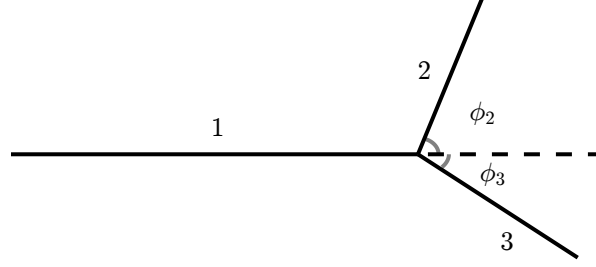


Figure 5.11: An asymmetric three-string with parameterized angles.

If the magnitude of the tension force in string i is T_i , then the static force-balance condition is

$$\sum_{i=1}^3 T_i v_i = 0. \quad (5.36)$$

Combining (5.35) and (5.36) gives

$$T_2 \sin \phi_2 = T_3 \sin \phi_3, \quad T_1 = T_2 \cos \phi_2 + T_3 \cos \phi_3. \quad (5.37)$$

Take T_2 and T_3 to be fixed and parameterize the system by ϕ_2 while allowing T_1 and ϕ_3 to vary in order to balance the tension forces. Rearranging (5.37) produces expressions for the determined variables:

$$\phi_3 = \sin^{-1} \left(\frac{T_2}{T_3} \sin \phi_2 \right)$$

and

$$\begin{aligned} T_1 &= T_2 \cos \phi_2 + T_3 \cos \left(\sin^{-1} \left(\frac{T_2}{T_3} \sin \phi_2 \right) \right) \\ &= T_2 \cos \phi_2 + T_3 \sqrt{1 - \left(\frac{T_2}{T_3} \sin \phi_2 \right)^2}. \end{aligned}$$

The expression under the radical must be positive, which produces a constraint on the fixed variables, namely $T_3 \geq T_2 \sin \phi_2$. Intuitively, if the tension force T_3 is too small compared to T_2 , no positioning of string 3 can nullify the net vertical tension force.

Rather than tension T_i , the formulation of Section 5.2 uses the stiffness k_i , relaxed length L_i , and stretch s_i of the strings. So, use Hooke's law to write $T_i = k_i L_i (s_i - 1)$ and eliminate the tension terms. Then fix each k_i and L_i , fix s_2 and s_3 , and allow s_1 (rather than T_1) and ϕ_3 to vary as functions of ϕ_2 .

Now the three-string can be studied as a parameterized NLEVP. For each desired value of ϕ_2 , the corresponding s_1 and ϕ_3 can be calculated and the NLEVP (5.13) can be constructed (this example uses samples from $\phi_2 \in [0, \pi]$). In order to use the contour methods of Section 4.2, a complex contour containing a region of interest must be specified. This section will use a contour surrounding the rectangle $[0.1, 8] + [-0.1, 0.1]i$, although no eigenvalues appeared between 0.1 and 2. The examples in this section will use the multi-point Loewner method (with $p = 30$) explained in Section 4.2.3 because it appears to be more robust than the single-point methods. Alternating equally-spaced left- and right-expansion points were placed above the contour on the segment $0.6i + [0.1, 8]$

Figure 5.10 shows an example where each string had different physical parameters and no symmetries occurred. Duplicate eigenvalues are avoided because the strings have different stiffnesses: $k_1 = 1$, $k_2 = 2$, $k_3 = 3$. The other physical properties are $L = \rho = 1$ for all strings, $s_2 = s_3 = 2$, and s_1 varied freely as needed to nullify the net force on the center point.

5.5.2 Effect of user choices

The contour-based NLEVP solvers discussed in Chapter 4 have a variety of user-selected parameters. This section briefly explores how user choices affect the accuracy of the calculated eigenvalues. For all examples in this section, I use the undamped three-string NLEVP introduced in Section 5.3. The NLEVP is treated as a function of real frequencies ω , and the eigenvalues are on the real axis. The true eigenvalues of this problem can

Table 5.1: Comparison of bounds on K^* given in Section 4.2.2 for the contours shown in Figure 5.12. For each contour, $\alpha(\Omega)$ is the rightmost extent of the domain Ω (maximum abscissa) and K^* is the minimality index of the invariant pair associated with the eigenvalues inside the contour. In each case, the identity matrix is used for Q .

$\alpha(\Omega)$	(4.14)	(4.19)	K^*	(4.12b)
2.5	1	1	1	2
5	1	1	1	6
6.5	1.5	2	2	9
8	1.833	2	2	11
10	2.5	3	3	15
12	2.833	3	3	17
14	3.5	4	4	21
16	4	4	5	24
18	4.5	5	5	27
19.5	5	5	6	30

be found to high precision by solving the transcendental equations (5.25), (5.26), (5.33), and (5.34). In all cases, accuracy is measured by *maximum forward error*, which is the maximum distance between a calculated eigenvalue and the true eigenvalue closest to it.

Changing the region of interest Ω may change how many test functions (or moments) are needed to solve the same NLEVP. Table 5.1 shows the minimality index for the three-string problem when considered in each of the ten nested domains shown in Figure 5.12. The lower and upper bounds (4.14) and (4.12b) are also displayed. For this problem, all eigenvalues are simple, so (4.12b) is just the number of eigenvalues inside Ω . For simplicity, the left-invariant subspace used in Theorem 4.1 to calculate (4.14) is $\mathbb{C}^m$ in every case, i.e., $Q = I_m$. For the example values in Table 5.1, the bound (4.19) matches K^* exactly in eight out of ten cases, whereas (4.12b) grows noticeably more quickly than K^* as Ω expands. When used as a guide for the number of test functions to use for single-shift contour solvers, (4.19) appears more accurate than (4.12b) for this problem.

The choice of Ω also has a numerical effect on NLEVP solvers, even if the same eigenvalues are enclosed. Figure 5.13 shows the convergence of quadrature for the three-string NLEVP

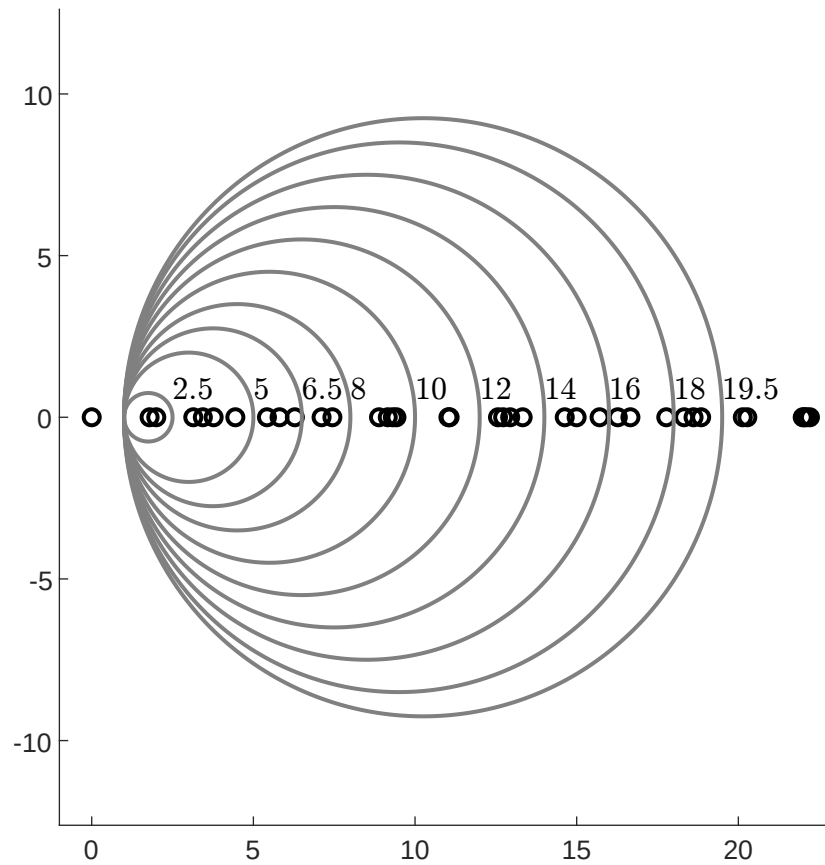


Figure 5.12: A sequence of circular contours and the eigenvalues they enclose. The rightmost extent (maximum abscissa) of each contour is shown just outside the circle. For each contour, the number of moments K^* needed for accurate Hankel-based integrals is shown in Table 5.1.

with elliptical Ω of varying vertical axis lengths. Convergence tends to be more rapid for contours that enclose the eigenvalues somewhat tightly, but contours that pass extremely close to eigenvalues would likely suffer numerical error since the eigenvalues are poles of $T(\lambda)^{-1}$. The optimal vertical axis for this example appears to be approximately 1.

The accuracy of the computed eigenvalues is also affected by the shape of Ω and the quadrature method used on its boundary. Figure 5.14 illustrates quadrature convergence for the single-shift method with several different Ω domains enclosing the same eigenvalues. The interpolation point used in each case is $\sigma = 4 + 4i$. The rectangular and elliptical contours appear similar, but the elliptical quadrature method converges noticeably faster.

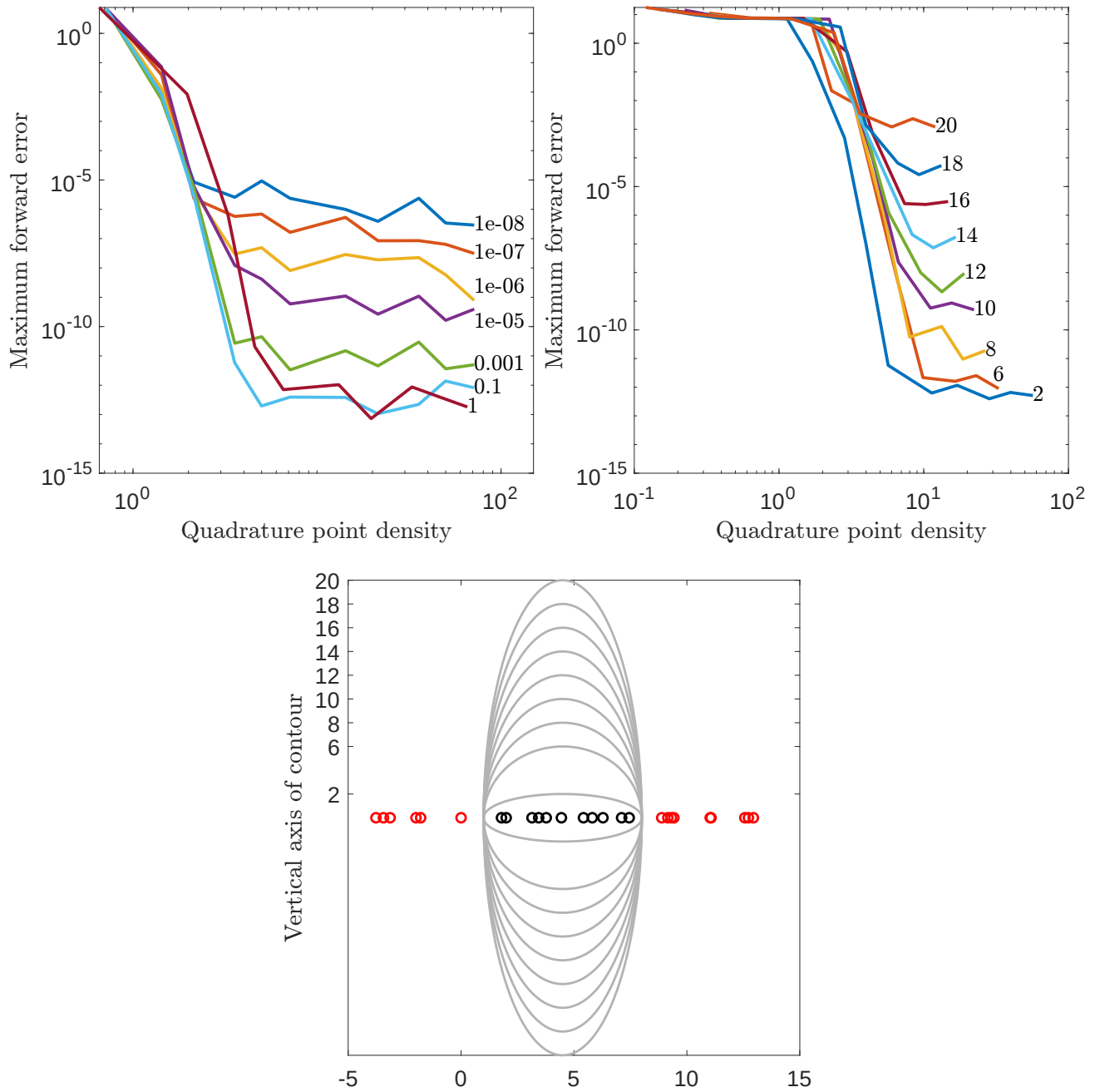


Figure 5.13: **Upper left and right:** For elliptical contours of a range of sizes, eigenvalue convergence as the number of quadrature points increases, plotted as points per unit length of the contour. Each line is labeled with the vertical extent (maximum imaginary part) of the Ω used in that convergence study. The extents greater than 1 are shown separately on the right to more clearly show that the relationship between contour size and accuracy is not monotone. **Bottom:** Boundaries of elliptical Ω choices with vertical extents 2, 6, 8, $\dots$, 20 (grey); eigenvalues within Ω (black); and eigenvalues outside Ω (red).

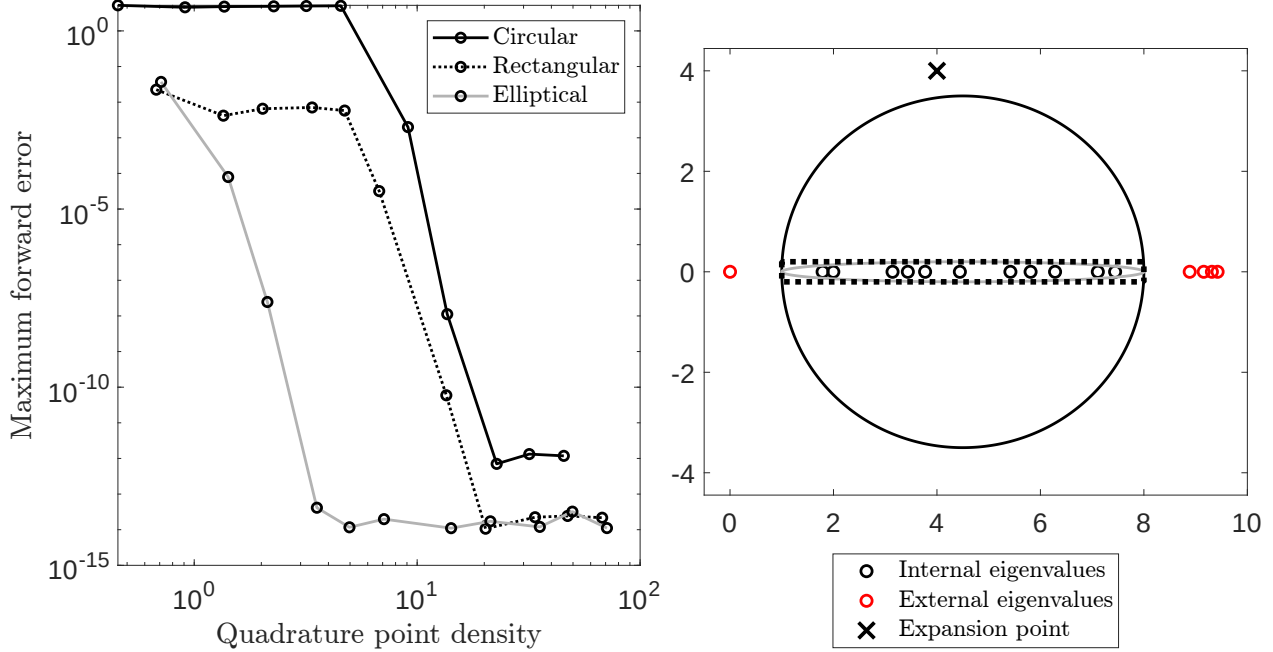


Figure 5.14: **Left:** eigenvalue convergence as the number of quadrature points increases, plotted as points per unit length of the contour. For each contour, the same numbers of quadrature points are used: the first and last graphed points correspond to 10 and 1000 quadrature points. **Right:** eigenvalues, contours, and interpolation point.

In the example shown in Figure 5.15, the infinite-shift method of Section 4.2.1 was to find eigenvalues of the three-string in a fixed domain Ω but with a varying number of moments K (For all single-shift experiments in this section, $p = 8$ random probing directions are used.) Four moments were needed to resolve the eigenvalues ($K^* = 4$), but including more moments harms accuracy because larger powers create a greater disparity in the summed values used in the quadrature. There is some tendency for eigenvalues of small magnitude to be calculated with the least accuracy, which is not surprising as the small eigenvalues are “farthest” from the interpolation point $\sigma = \infty$. In this example, the accuracy difference is most visible when “too many” moments are being used (upper left): With a large number of moments, the largest eigenvalues (e.g., ω_{27}) are still calculated accurately, while the quadrature contributions of the small eigenvalues are numerically lost due to the large powers involved.

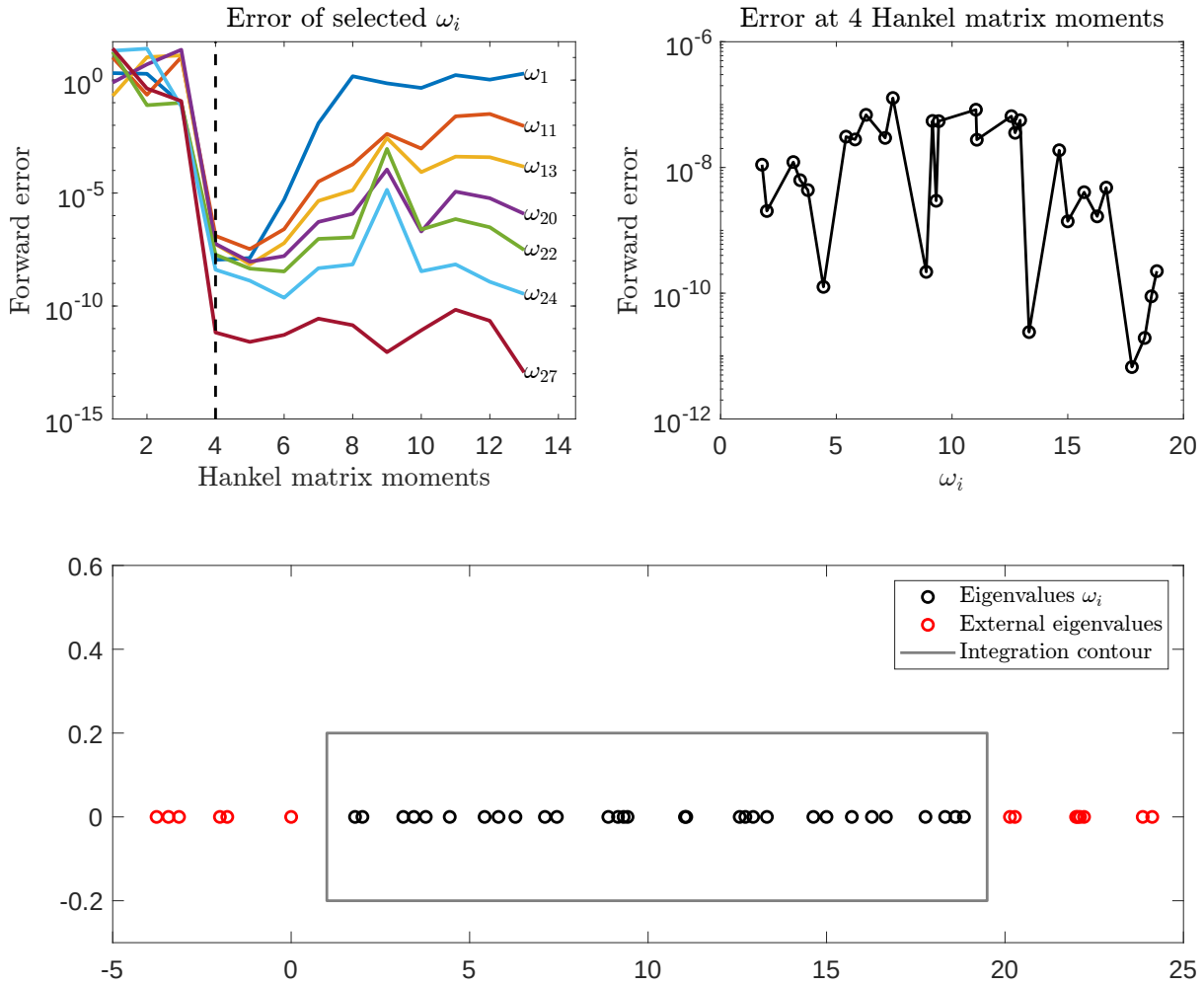


Figure 5.15: Single-shift method with interpolation at $\sigma = \infty$. **Upper left:** approximation error of selected eigenvalues as a function of the number of moments used in the Hankel matrix ($K^* = 4$ marked with a dashed line). **Upper right:** error of all eigenvalues when using $K^* = 4$ moments. **Bottom:** integration contour.

Figure 5.16 is analogous to Figure 5.15, but now using the finite interpolation point $\sigma = 10 + 0.4i$. Again, $K^* = 4$ moments are needed to resolve the eigenvalues, and including more moments harms accuracy. The eigenvalues closest to σ tend to be calculated most accurately.

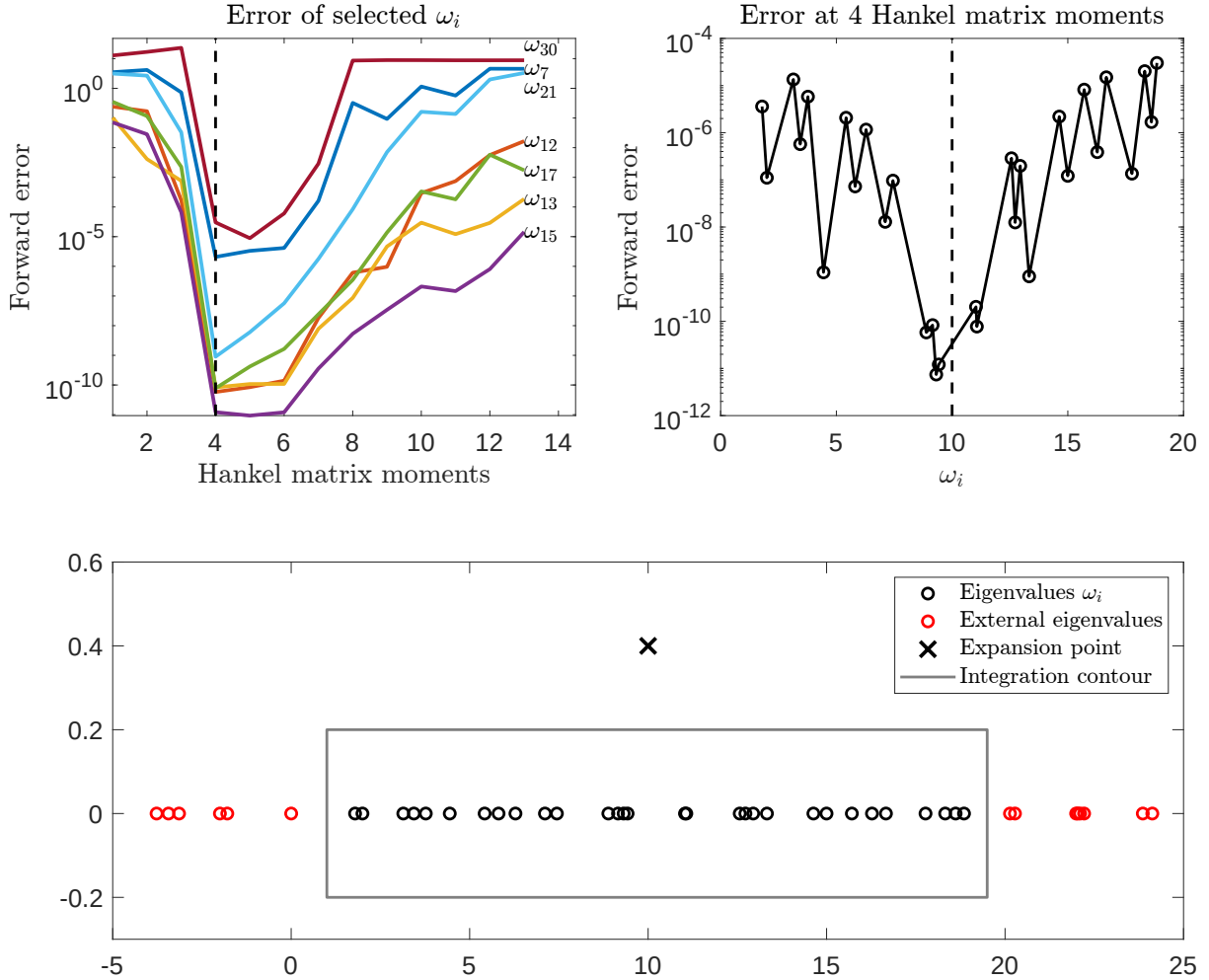


Figure 5.16: Single-shift method with interpolation at $\sigma = 10 + 0.4i$. **Upper left:** approximation error of selected eigenvalues as a function of the number of moments used in the Hankel matrix ($K^* = 4$ marked with a dashed line). **Upper right:** error of all eigenvalues at $K^* = 4$ moments ($\Re(\sigma) = 10$ marked with a dashed line). **Bottom:** integration contour and shift value.

Figure 5.17 illustrates a study of the effect of the number of shifts used in the multi-shift method of Section 4.2.3. For this experiment, the contour and contour-to-shift distance are the same as in Figure 5.18. Since 32 eigenvalues lie within this contour, the method may not produce any eigenvalues with accuracy if $p = 31$ (or fewer) shifts are used. Indeed, Figure 5.17 shows that, with only $p = 31$ shifts, some calculated eigenvalues were unsatisfactorily far from any true eigenvalues. Using the minimum number of shifts ($p = 32$) or many more shifts than necessary ($p = 64$) achieved similar results. This may be contrasted with the single-shift methods just discussed for which using more than the necessary number of moments ($K > K^*$) appears to cause the computed eigenvalues to be less accurate.

Finally, note that in Figure 5.18, the shifts are distributed just outside and along the length of a long domain. Consequently, each eigenvalue is approximately equally close to the nearest interpolation points. This appears to cause the accuracy of the computed

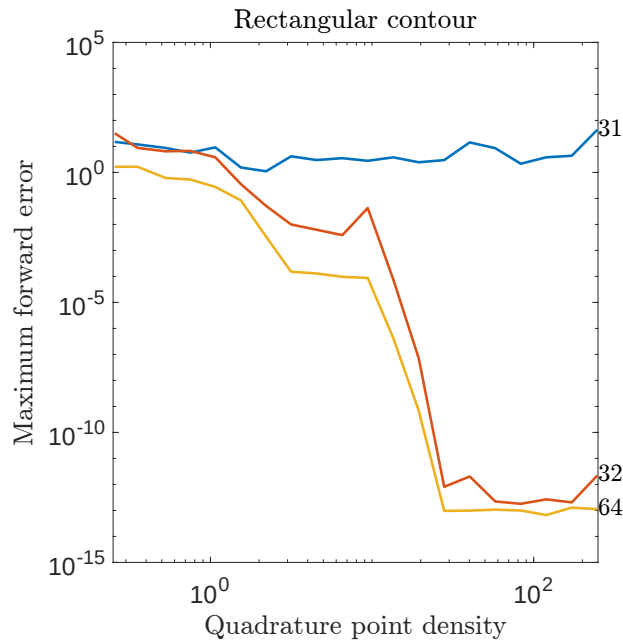


Figure 5.17: Multi-shift method. Eigenvalue convergence for shift counts $p = 31, 32, 64$ corresponding to the line labels.

eigenvalues to be approximately constant throughout the domain. This result is in contrast to the single-shift methods, which show a tendency to exhibit greater error for eigenvalues far from the shift.

Figure 5.19 shows the results of an experiment with the multi-shift method of Section 4.2.3. The shifts $\{\sigma_i\}_{i=1}^{32}$ and $\{\theta_i\}_{i=1}^{32}$ are placed in two rows above and below the contour, each row alternating between left and right shifts as illustrated in Figure 5.18. The rows of shifts were placed at a variety of distances from the integration contour, and for each shift placement, quadratures of varying fineness were used to solve the NLEVP. The integration contour was a rectangle with horizontal sides having imaginary parts ± 0.1 .

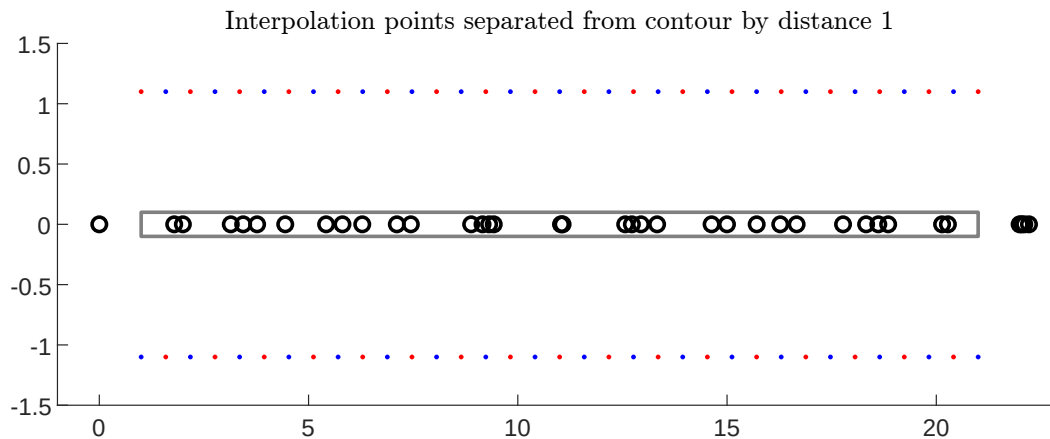


Figure 5.18: Multi-shift method. One of the contour-shift configurations used to generate Figure 5.19. Left (red) and right (blue) shifts alternate at constant vertical distances on either side of the contour.

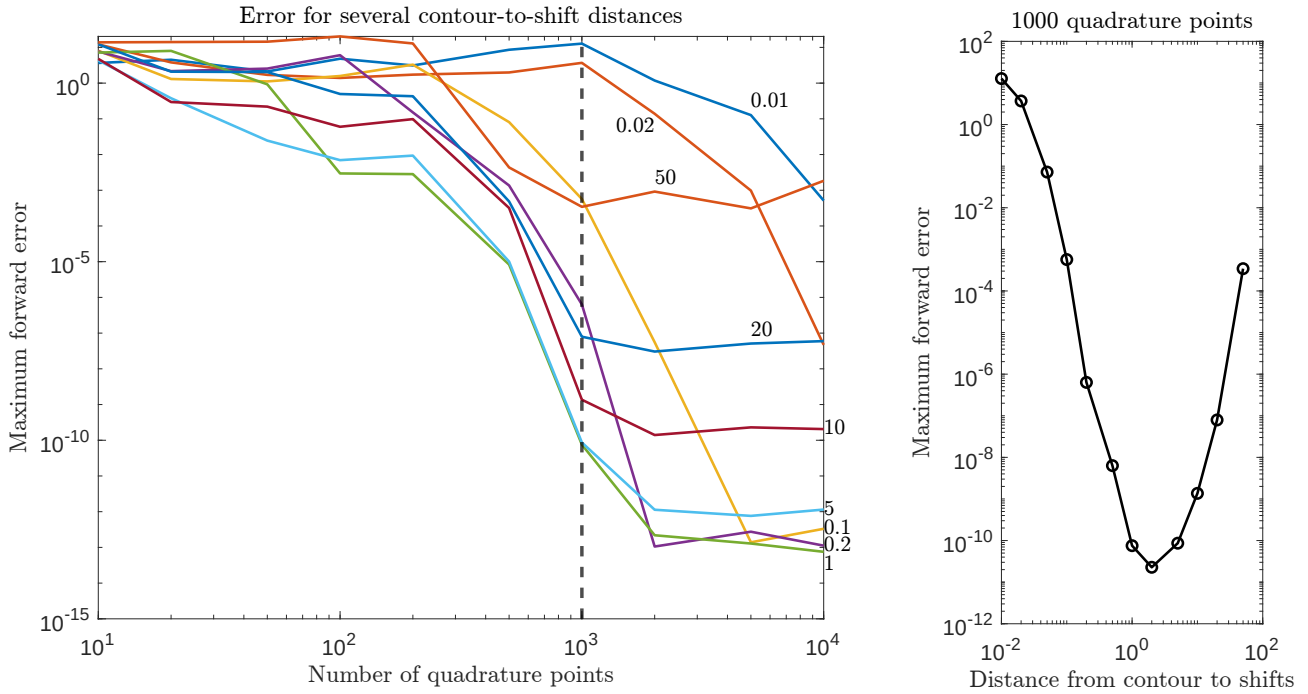


Figure 5.19: Multi-shift method. **Left:** Quadrature convergence for selected contour-to-shift distances ($N = 1000$ quadrature points marked with a dashed line). Each convergence line is labelled with the distance between the contour and the rows of shifts. **Right:** Clarifying view comparing the accuracy of computed eigenvalues with different shift positions for $N = 1000$ quadrature points.

In some experiments, the shifts were placed very close to the contour, such as with imaginary part 0.11 at a distance of just 0.01 from the top of the contour. Shifts very close to the contour produced inaccurate results, even for very fine quadrature (e.g., $N = 10^4$). Shifts that were placed too far from the contour also performed poorly. Moderate contour-to-shift distances (around 0.1 to 5) performed best. This non-monotone relationship between accuracy and shift placement is illustrated in the right plot of Figure 5.19. It is apparent that choosing the positions of shifts is a matter of finesse, requiring further study. For additional results on shift selection, see [19].

5.6 Computation time study

This section compares the computational costs of two methods when calculating high-frequency eigenvalues of the undamped three-string example of Section 5.3. The first method is the NLEVP-based method proposed by this chapter, with the multi-shift contour integration solver described in Section 4.2.3 for calculating eigenvalues. The second method is the pseudospectral collocation method of [24]. It is expected that other methods that rely on discretizing the members of the structure will perform somewhat similarly.

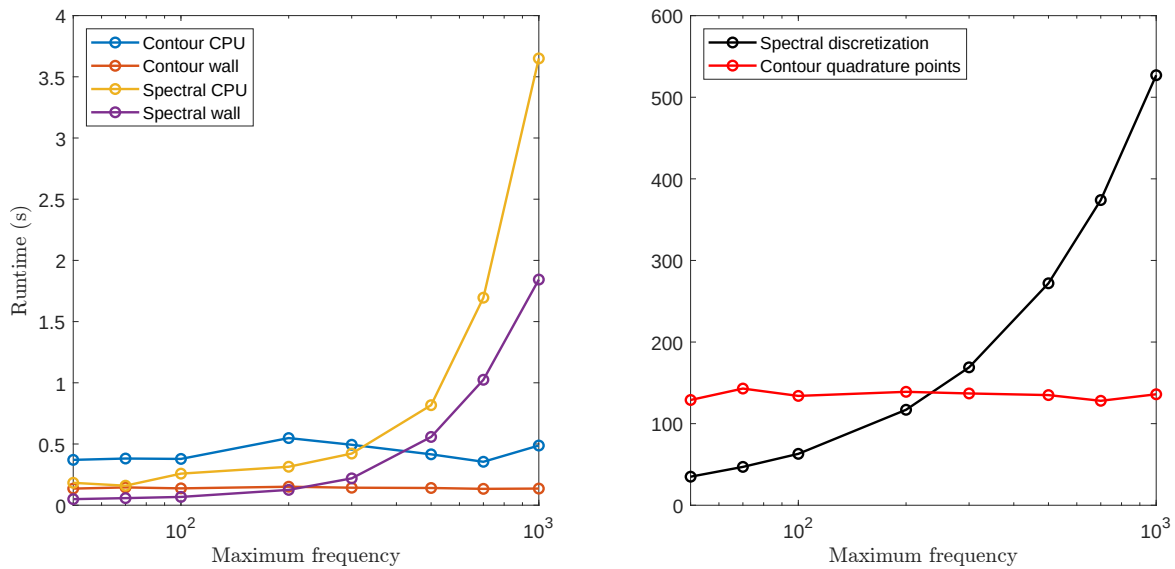


Figure 5.20: **Left:** Elapsed wall time and CPU time required to calculate eigenvalues accurately to six decimal places. **Right:** Lowest discretization levels required by the two methods to achieve six places of accuracy. The discretization level of pseudospectral collocation is the number of subdivisions used on each string, while the “discretization” of the NLEVP method is the number of quadrature points. The fact that these very different quantities are close enough in value to fit reasonably on a plot together is a coincidence. In these plots, all of the lines rising noticeably for high frequencies correspond to the pseudospectral collocation method. The other lines correspond to solving the NLEVP formulation with the multi-shift contour method.

Figure 5.20 shows the two methods' computation time (measured by MATLAB's built-in timing methods) and complexity (measured by discretization level and number of quadrature points). The frequency of the eigenvalues has little effect on the complexity of the multi-shift contour method; changing the frequency of the calculated eigenvalues is as simple as translating the integration contour. In this study, the contours used are thin ellipses enclosing intervals of length 20 just below the maximum desired frequency. The $p = 30$ left and right shifts were arranged in an alternating pattern on both sides of the contour as in Figure 5.18. In order to compute high-frequency eigenvalues, the pseudospectral collocation method requires a fine discretization of the elements, which produces a large linear eigenvalue problem that must be solved. Consequently, calculating even a single high-frequency eigenvalue by pseudospectral collocation (or other discretization methods) can be quite costly.

For low frequencies, discretization methods like pseudospectral collocation are likely to be more efficient than the contour integration method. However, the mode shapes produced by discretization methods are approximations that are only as accurate as the discretization. The mode shapes produced by the contour method can be calculated to any desired accuracy from (5.10) and the relationship $X_i(\eta) = V_i Y_i(\eta)$. The only limitation is the accuracy to which eigenvalues ω have been calculated.

5.7 Dynamic stiffness matrix method

Another equivalent NLEVP that produces system modes is the “dynamic stiffness matrix” (DSM), which is common in the vibration engineering literature. A DSM defines a frequency-dependent linear map from responses to driving forces. Suppose that oscillating forces and moments are applied to a structure at discrete points (nodes) with frequency ω . Also assume that the structure experiences no damping. Let d be the total number of

degrees of freedom of the force-points of the structure. Let $\mathbf{f} \in \mathbb{R}^d$ be a vector with each entry corresponding to a degree of freedom of one node, so that the forces applied to the nodes of the structure at time t are described by $\Re(\mathbf{f}e^{i\omega t})$. Let the corresponding response vector be $\mathbf{x} \in \mathbb{C}^d$, so that the vector of displacements and rotations of the nodes is $\Re(\mathbf{x}e^{i\omega t})$. Then the dynamic stiffness matrix for the structure at the frequency ω is $D(\omega) \in \mathbb{C}^{d \times d}$ defined by the relationship

$$D(\omega)\mathbf{x} = \mathbf{f}. \quad (5.38)$$

Natural frequencies can be identified as the ω values such that $D(\omega)$ is singular. Intuitively, if $D(\omega)$ is singular, then there is a nonzero oscillating response vector $\mathbf{x}e^{i\omega t}$ that can be maintained with zero input because $D(\omega)\mathbf{x} = \mathbf{0}$.

The Wittrick–Williams (WW) algorithm introduced in [50] is the basis of most methods used by engineers for solving the NLEVPs associated with dynamic stiffness matrices. Like the framework of Section 5.2 paired with contour-based NLEVP solvers, WW has the advantage of being immediately applicable to any frequency range of interest without the need to produce a fine discretization when high-frequency modes are desired. Although an extension of WW to damped systems has been proposed [27], the widely known original WW algorithm is applicable only to problems with real eigenvalues (which necessarily correspond to undamped systems). Compared to WW, the framework of Section 5.2 has the advantage of handling damped systems very naturally.

The engineering literature does not yet appear to have many examples of using modern NLEVP solvers in the dynamic stiffness matrix context, nor does it appear that many numerical analysts have noticed that dynamic stiffness matrices provide compelling examples of NLEVPs. A contribution of this dissertation is to highlight that each community could benefit from greater familiarity with the other's work.

Deriving the dynamic stiffness matrix

Many problems, including elastic webs, have boundary conditions that each involve either displacement or force, but not both. In these cases, the DSM may be derived from the NLEVP T introduced in Section 5.2. Suppose the number of free nodes in the structure is n_f , the number of spatial dimensions is d , the number of strings is n_e , and $n = 2dn_e$ is the dimension of $T(\omega)$. For the string problems being considered, the total number of degrees of freedom of the free nodes—which is also the size of the dynamic stiffness matrix—is dn_f .

Suppose that the rows of $T(\omega)$ corresponding to displacement boundary conditions are assembled in $T_1(\omega) \in \mathbb{R}^{(n-dn_f) \times n}$ and the rows corresponding to force-based conditions are assembled in $T_2(\omega) \in \mathbb{R}^{dn_f \times n}$. For any vector $\mathbf{c} = \begin{bmatrix} A_{1,1} & \cdots & B_{n_e,d} \end{bmatrix}^T \in \mathbb{R}^n$ of integration coefficients, the condition $T_1(\omega)\mathbf{c} = 0$ is required for the displacement boundary conditions to be satisfied (i.e., continuity). For such $\mathbf{c}$, the vector of net forces on the junction points is $\mathbf{f} = T_2(\omega)\mathbf{c}$. An eigenpair $(\omega, \mathbf{c})$ of T satisfies $T(\omega)\mathbf{c} = 0$, which means $T_1(\omega)\mathbf{c} = 0$ and $\mathbf{f} = T_2(\omega)\mathbf{c} = 0$.

For any ω , let $\mathbf{N}$ be a matrix whose columns are a basis for the null space of $T_1(\omega)$, in other words, a basis for the space of integration coefficients that satisfy the displacement boundary conditions. Let N be the nullity of $T_1(\omega)$ so that $\mathbf{N} \in \mathbb{R}^{n \times N}$. If $T_1(\omega)$ has full row rank, then $N = dn_f$.

Choose $\mathbf{Y} \in \mathbb{R}^{dn_f \times n}$ so that, for any vector $\mathbf{c}$ of integration coefficients, $\mathbf{x} = \mathbf{Y}\mathbf{c}$ describes the corresponding displacements of the free nodes. The only acceptable displacements $\mathbf{x}$ are those that satisfy the boundary conditions, which means they can be represented as $\mathbf{x} = \mathbf{Y}\mathbf{N}\mathbf{q}$, for some $\mathbf{q} \in \mathbb{R}^N$. The DSM is given by $D := D(\omega)$ in (5.38), which is equivalent to

$$D\mathbf{Y}\mathbf{N}\mathbf{q} = T_2\mathbf{N}\mathbf{q}$$

for all $\mathbf{q}$. If $\mathbf{YN}$ is square ($N = dn_f$) and nonsingular, then

$$D = T_2 \mathbf{N} [\mathbf{YN}]^{-1}.$$

For any ω such that the matrix $\mathbf{YN}$ is not invertible, and in particular if $N > dn_f$, the dynamic stiffness matrix D will not normally be defined at ω . Instead, D can be expected to have a pole. To see this, suppose $\mathbf{q} \in \mathbb{R}^N$ is a nonzero null vector of $\mathbf{YN}$. Then

$$0 = D \mathbf{YN} \mathbf{q} = T_2 \mathbf{N} \mathbf{q}. \quad (5.39)$$

For some problems, it may be possible that $T_2 \mathbf{N} \mathbf{q} = 0$ for all $\mathbf{q} \in \mathcal{N}(\mathbf{YN})$, but this is hardly universal, as upcoming examples will show. Typically, $T_2 \mathbf{N} \mathbf{q} \neq 0$, contradicting (5.39), and no $D(\omega)$ can satisfy the definition of the dynamic stiffness matrix at ω . Instead, the form of (5.39) suggests that D will have a pole at ω .

The failure of $\mathbf{YN}$ to be invertible corresponds to an oscillating regime in which the free nodes of the structure are all completely stationary. This null vector $\mathbf{q}$ encodes the (nonzero) vibrations of the strings via the integration coefficients $\mathbf{N} \mathbf{q}$. The fact that $\mathbf{x} = \mathbf{YN} \mathbf{q} = 0$ illustrates that the free nodes are stationary.

5.7.1 DSM of three-string

The calculations in Section 5.4 eliminated the degrees of freedom of the strings except for string 2, so it is convenient to represent the displacement of the junction point of the three-string with A_2, C_2 . Modal displacement is simply

$$\mathbf{x} = X_2(1) = V_2 \begin{bmatrix} A_2 \sin(\omega/\sqrt{2}) \\ C_2 \sin \omega \end{bmatrix}.$$

When the junction point is being driven by a sinusoidally oscillating force, instead of the tension forces summing to zero at the junction as in (5.28), the driving force must balance the tension forces, i.e., the net force on the massless center point must always be zero. Because the right-hand side of (5.28) is the force at the “right” endpoint of each string ($\eta = L$), it actually expresses the negative of the restoring force of the strings at the junction. Thus the force balancing condition in the driven scenario is that the right-hand side of (5.28) is $-\mathbf{f}$, and the first and second entries of $-\mathbf{f}$ are displayed in the right-hand sides of (5.30) and (5.32). These may be combined into

$$-\mathbf{f} = \begin{bmatrix} 2 \cot \frac{\omega}{\sqrt{2}} + \frac{1}{\sqrt{2}} \cot \omega & 0 \\ 0 & 2 \cot \omega + \sqrt{2} \cot \frac{\omega}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} A_2 \sin \frac{\omega}{\sqrt{2}} \\ C_2 \sin \omega \end{bmatrix}.$$

Comparing this last expression with the definition of $D(\omega)$ reveals

$$\begin{aligned} D(\omega) &= \begin{bmatrix} 2 \cot \frac{\omega}{\sqrt{2}} + \frac{1}{\sqrt{2}} \cot \omega & 0 \\ 0 & 2 \cot \omega + \sqrt{2} \cot \frac{\omega}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} V_2^{-1} \\ &= \begin{bmatrix} -2\sqrt{2} \cot \frac{\omega}{\sqrt{2}} - \cot \omega & 0 \\ 0 & -2\sqrt{2} \cot \omega - 2 \cot \frac{\omega}{\sqrt{2}} \end{bmatrix}. \end{aligned} \tag{5.40}$$

All dynamic stiffness matrices are symmetric [50], so the symmetry of (5.40) is expected. However, (5.40) has poles rather than eigenvalues at some of the natural frequencies of the structure. The next section examines this phenomenon further.

5.7.2 Poles in DSMs

Dynamic stiffness matrices have poles that can occasionally mask the natural frequencies of the system they represent. Consider modifying a multi-body system by imposing a clamping condition at each of its junction points. Let ω be a natural frequency of this

modified system. Since the clamping conditions isolate the structure members, ω is also a natural frequency of some individual member under double clamping conditions. When the modified structure is modally oscillating at frequency ω , let $\mathbf{f}$ be the normalized vector of forces imparted by the clamps. When the only moving member is one with natural frequency ω , a nonzero force must be provided by the clamps to maintain force-balance at its free end(s). Consequently, if the element with natural frequency ω has at least one free end, then $\mathbf{f}$ may be taken to not be the zero vector.

The original “free” structure may of course oscillate in the same way as the clamped structure when the oscillating external forces $\mathbf{f}$ are applied to mimic the effects of the clamps. Just as the junction points of the clamped structure are stationary, the displacement vector $\mathbf{x}$ of the free structure is zero. The defining expression (5.38) for the DSM at ω is then $\mathbf{f} = D(\omega)\mathbf{x} = D(\omega)\mathbf{0}$, which cannot hold for finite $D(\omega)$ since $\mathbf{f}$ is not the zero vector. Conclude that D has a pole at ω . If ω also happens to be a natural frequency of the free multi-body system, then the presence of a pole at ω causes D to fail to have an eigenvalue at ω . Therefore, finding the eigenvalues of the DSM is not always sufficient to reveal the natural frequencies of the system. In other words, although each eigenvalue of the dynamic stiffness matrix is a natural frequency, the converse is not always true.

When the DSM is used to find natural frequencies in the engineering literature, it is typical to use the Wittrick-Williams algorithm (WW) [50]. WW does not calculate eigenvalues of D directly, and can therefore avoid the problem of natural frequencies hidden by poles. For a given $\omega \in \mathbb{R}$, let $J(\omega)$ be the number of natural frequencies of the multi-body system that are less than ω . Let $J_0(\omega)$ be the number of poles of D that are less than ω , that is, the number of natural frequencies of the individual members when clamped. If each member is simple enough that its clamped natural frequencies are known, then $J_0(\omega)$ may be calculated easily. Let $m_-(D(\omega))$ be the number of (strictly) negative eigenvalues of $D(\omega)$. If $D(\omega)$ can be formed, then counting its negative eigenvalues is not difficult. WW

uses a spectrum slicing method based on an argument through Sylvester’s law of inertia [28, Thm. 4.5.8] that $J(\omega) = J_0(\omega) + m_-(D(\omega))$. By sampling $J(\omega)$ values systematically, WW calculates the natural frequencies of the system without specific concern for whether those natural frequencies are also eigenvalues of D .

For one example of DSM poles coinciding with natural frequencies, consider the eigenvalues of the three-string that satisfy (5.25) or (5.26), e.g., integer multiples of π or $\pi\sqrt{2}$. Figures 5.7–5.9 show that such natural frequencies correspond to mode shapes in which the junction point is stationary. The planar modes corresponding to eigenvalues that are multiples of π are labelled 6, 15, 24, and 30, while the modes for multiples of $\pi\sqrt{2}$ are 9 and 19. Modes 9 and 19 are *compression* or *longitudinal* modes, meaning that each string moves only in its own axial direction. The dots in the figure make the stretching and compression of these modes visible. Although the figures depict the 3-dimensional modes of the three-string, the planar modes are the same as for the 2-dimensional system in the derivations of (5.25) and (5.26). In the DSM (5.40) of the three-string, both diagonal entries have poles at each of these natural frequencies, and therefore these are not eigenvalues of the DSM.

To further understand the poles of a DSM, consider lateral vibrations of a single elastic string stretched between two walls. In the notation of Section 5.2,

$$\mathbf{t} = \mathbf{h} = \begin{bmatrix} F \end{bmatrix},$$

where F is the “frame” node. Say the string has linear density $\rho = 1$ and length-normalized Hooke’s constant $k = 1$. For this one-dimensional problem, $X(\eta) = Y(\eta)$ and, as in (5.10),

$$X(\eta) = A \sin(\omega\eta) + B \cos(\omega\eta).$$

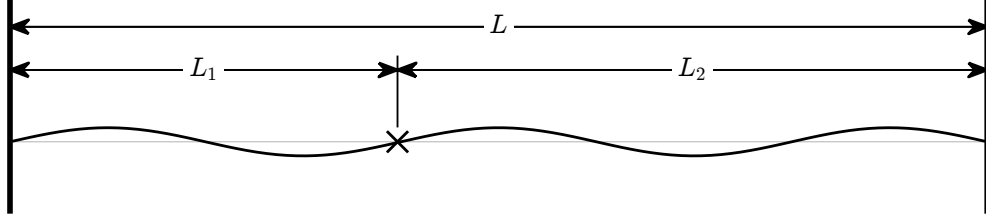


Figure 5.21: Fifth dynamic mode of an elastic string with an artificial subdivision point two-fifths of the way along its length. This point is a node of the fifth mode.

The boundary conditions $X(0) = X(L) = 0$ reveal that $B = 0$ and

$$0 = A \sin(\omega L),$$

and therefore the natural frequencies are

$$\omega = \frac{n\pi}{L}, \quad n \in \mathbb{N}. \quad (5.41)$$

Consider now a reformulation of this problem with a node dividing the string in two. Seeking natural frequencies through the DSM will produce a discrepancy. Say the strings have lengths L_1 and L_2 such that $L_1 + L_2 = L$. Because k is the length-normalized Hooke's constant, the divided strings have the same stiffness: $k_1 = k_2 = 1$. Let $\eta = 0$ correspond to the fixed end of each string so that, in the notation of Section 5.2,

$$\mathbf{t} = \mathbf{h} = \begin{bmatrix} F \\ J \end{bmatrix},$$

where F is the frame node and J is the free node where the strings meet. This setup is illustrated in Figure 5.21.

Let string 1 have positive orientation, meaning $V_1 = 1$ and $V_2 = -1$. As in Section 5.2, the eigenvalue associated with longitudinal vibrations is $k = 1$, so $H_i = V_i k V_i^T = 1$ for

$i = 1, 2$. As usual,

$$Y_i(\eta) = A_i \sin(\omega\eta) + B_i \cos(\omega\eta),$$

and $X_i(\eta) = V_i Y_i(\eta)$. The frame boundary conditions $X_i(0) = 0$ show that $B_1 = B_2 = 0$. The continuity condition at the junction point is

$$X_1(L_1) = X_2(L_2) = -Y_2(L_2),$$

so

$$X_1(L_1) = A_1 \sin(\omega L_1), \quad X_2(L_2) = -A_2 \sin(\omega L_2)$$

and

$$A_1 = -A_2 \frac{\sin(\omega L_2)}{\sin(\omega L_1)}.$$

To construct the DSM, suppose an external force $\mathbf{f}$ cancels the restoring forces of the strings given in (5.11), in other words

$$\begin{aligned} \mathbf{f} &= kX'_1(L_1) + kX'_2(L_2) \\ &= \omega(A_1 \cos(\omega L_1) - A_2 \cos(\omega L_2)) \\ &= \omega \left(-A_2 \frac{\sin(\omega L_2)}{\sin(\omega L_1)} \cos(\omega L_1) - A_2 \cos(\omega L_2) \right). \end{aligned}$$

The defining expression of the DSM is $D(\omega)\mathbf{x} = \mathbf{f}$, where $\mathbf{x} = X_1(L_1) = X_2(L_2)$. In the current scalar case,

$$\begin{aligned} D(\omega) = \mathbf{f}/\mathbf{x} &= \frac{-\omega\sqrt{k} \left(A_2 \frac{\sin(\omega L_2)}{\sin(\omega L_1)} \cos(\omega L_1) + A_2 \cos(\omega L_2) \right)}{-A_2 \sin(\omega L_2)} \\ &= \omega\sqrt{k} [\cot(\omega L_1) + \cot(\omega L_2)]. \end{aligned}$$

The simplest case $L_1 = L_2 = L/2$ best illustrates the problem of using the DSM to find

natural frequencies. In this case, $D(\omega) = 2\omega\sqrt{k} \cot(\omega L/2)$, which has zeros (eigenvalues) when

$$\omega = \frac{\pi(2n+1)}{L}, \quad n \in \mathbb{Z},$$

which only includes every second natural frequency in (5.41). The zeros of D do not include the “even” dynamic modes, in which the center point J is stationary (displacement $\mathbf{x}$ is 0). In other words, the original system is continuous with infinitely many degrees of freedom, but the reduction to this particular single degree of freedom is singular.

More generally, subdividing the string at $L_1 = Lp/q$ places a pole instead of the desired zero at every q th natural frequency (if p/q is a reduced fraction). In other words, as long as L_1/L is rational, there will be modes where node J is stationary and the DSM does not have a zero to reflect that natural frequency. Figure 5.22 illustrates this principle with plots of $D(\omega)$ for several L_1 values. When L_1/L is irrational, the DSM reveals all of the natural frequencies except for $\omega = 0$. When $L_1 = L/2$, every pole of D is a natural frequency. When $L_1 = L/4$ or $L/6$, only some of the poles are natural frequencies. When $L_1/L = \sqrt{1.1}/2$, none of the poles are natural frequencies.

Furthermore, the DSM does not have a zero at the eigenvalue $\omega = 0$ for any L_1 value. Instead, $D(\omega)$ has a removable singularity at $\omega = 0$ because, although cotangent has a pole at 0,

$$\lim_{\omega \rightarrow 0} D(\omega) = \lim_{\omega \rightarrow 0} \omega [\cot(\omega L_1) + \cot(\omega L_2)] = \frac{1}{L_1} + \frac{1}{L_2}.$$

Figure 5.23 shows that $D(0) \neq 0$ more clearly than Figure 5.22 does.

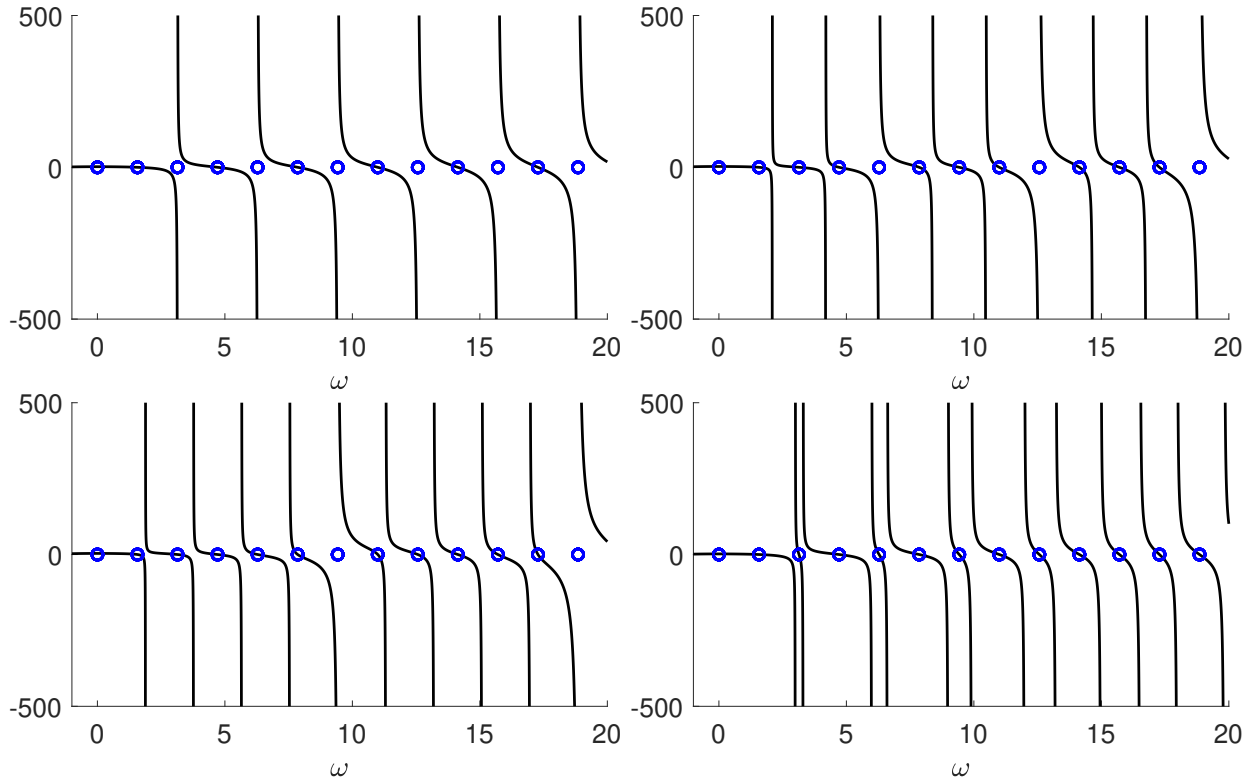


Figure 5.22: The DSM $D(\omega)$ for the single-string system with several L_1 values. In every plot, dots show the true eigenvalues (5.41). **Upper left:** $L_1 = L/2$. **Upper right:** $L_1 = L/4$. **Lower left:** $L_1 = L/6$. **Lower right:** $L_1 = L\sqrt{1.1}/2$.

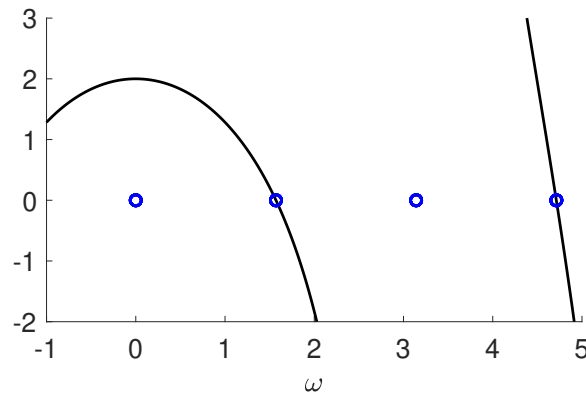


Figure 5.23: Detail of the upper left plot of Figure 5.22 ($L_1 = L/2$). The DSM $D(\omega)$ does not have an eigenvalue at $\omega = 0$ for this or any other subdivision point position.

Chapter 6

Conclusion

In this work, I explore several topics in mechanical vibrations. I use the results of physical experiments to support the use of the exponential energy decay model of the received signal strength (RSS) localization method proposed by Alajlouni in [1]. The experiments also support the use of the Cramér–Rao bound (CRB) as a tool for identifying spatial areas where the source localization problem will be ill-conditioned. I also show how the CRB can be used as a criterion for optimizing sensor layouts. The conditioning of location estimators could also be estimated with physical experiments or Monte Carlo simulations, but the CRB is computationally cheaper than either of these approaches.

My numerical CRB minimization for sensors on a line indicate some principles for arranging a fixed number of sensors in a corridor. The sensors should be approximately evenly spread across the length of the corridor if energy from an event does not travel readily across the entire area. Three factors contribute to energy remaining relatively local: the corridor is long (in the terms of Section 3.3, x_M is large), energy dissipates rapidly over distance ($-\beta$ is large), or the event has low energy content (α is small). If energy *does* travel easily through the area, then the minimax CRB is optimized by clustering the sensors near the ends of the corridor.

Before discussing a connection between vibrations and nonlinear eigenvalue problems (NLEVPs), I review several NLEVP solution methods. One does not typically know *a priori* the number K^* of test functions that will be needed by single-shift contour integra-

tion algorithms for NLEVPs. Theorem 4.2 is a new theoretical lower bound for K^* , which may assist practitioners who have some knowledge of the NLEVP they are studying.

NLEVPs are the basis of my method of modal analysis for multi-body structures. My method has several advantages over dynamic stiffness matrices (DSMs), which are an established tool for modal analysis. DSMs traditionally only support structures whose element dynamics are described by the positions, forces, and moments of the endpoints. My method readily supports other linearized dynamics, such as viscous damping. I also show that the DSM approach can produce ambiguous results due to the DSM having singularities at any mode where the junction points of the structure are stationary—a problem that my NLEVP approach avoids naturally.

I use networks of elastic strings as examples of multi-body structures amenable to my method, and I use the associated NLEVPs for numerical tests of contour-based solvers. I demonstrate several advantages of the recently-proposed multi-shift method, and show that contour methods can be expected to calculate high-frequency eigenvalues more efficiently than discretization methods.

There are several natural ways in which to continue the work presented here. The CRB could be applied to other localization frameworks, such as RSS methods with different (e.g., quadratic) energy dissipation models and time of arrival methods. When optimizing a sensor layout, the CRB could be amalgamated over the region in ways besides minimax, such as the mean CRB value.

The NLEVP solvers described in Chapter 4 have many interesting features that are not completely understood. In Sections 4.2.1 and 4.2.3, the contour-based NLEVP solution methods require the specification of one or more interpolation points. For the small problems studied in this work, it was not difficult to perform *ad hoc* experiments to select satisfactory shifts. A more thorough study might reveal a reasonable systematic approach

to shift selection. Similarly, Section 5.5 illustrates only briefly how the choice of domain Ω affects the accuracy of computed eigenvalues. There is certainly more to learn about how the proximity of eigenvalues (internal or external to Ω) affect accuracy.

Sections 5.2.2 and 5.5.1 illustrate that the NLEVPs related to vibrating string networks may be parameterized by damping or by movement of the at-rest members. The eigenvalue continuation method of [10] could be applied to such scenarios, improving eigenvalue calculations for a collections of related problems.

The multi-body structures featured in this work are made of elastic strings, but the framework of Section 5.2 may be applied to other simple one-dimensional members such as beams. Beam models require additional boundary conditions related to torque and bending moments, which result in a larger but conceptually similar NLEVPs. Structural networks formed by a combination of beams and tense cables are sometimes called *tensegrity* structures, seemingly as a shortening of “tensional integrity.” Tensegrity structures are being actively studied for their uses in architecture and robotics as well as their natural occurrence in biological structures [37, 43]. With adaptations for the introduction of beams, the NLEVP method Section 5.2 is naturally suited to studying the modes of tensegrity structures.

Bibliography

- [1] Sa’ed Alajlouni, Murat Ambarkutuk, and Pablo Tarazaga. Towards the detection and localization of multiple occupant footsteps from vibroacoustic measurements. In *Dynamics of Civil Structures, Volume 2: Proceedings of the 38th IMAC, A Conference and Exposition on Structural Dynamics 2020*, pages 193–200. Springer, 2021.
- [2] Sa’ed Alajlouni, Jonathan Baker, and Pablo Tarazaga. Maximum likelihood estimation for passive energy-based footstep localization. *Mechanical Systems and Signal Processing*, 163:108158, 2022.
- [3] Sa’ed Ahmad Alajlouni. *Energy-based Footstep Localization using Floor Vibration Measurements from Accelerometers*. PhD thesis, Virginia Tech, 2017.
- [4] Bruce Anderson, Jeffrey Jackson, and Meera Sitharam. Descartes’ rule of signs revisited. *The American Mathematical Monthly*, 105(5):447–451, 1998.
- [5] Athanasios C Antoulas. *Approximation of Large-Scale Dynamical Systems*. SIAM, 2005.
- [6] Junko Asakura, Tetsuya Sakurai, Hiroto Tadano, Tsutomu Ikegami, and Kinji Kimura. A numerical method for nonlinear eigenvalue problems using contour integrals. *JSIAM Letters*, 1:52–55, 2009.
- [7] JR Banerjee. Dynamic stiffness formulation for structural elements: a general approach. *Computers & Structures*, 63(1):101–103, 1997.
- [8] Timo Betcke and Daniel Kressner. Perturbation, extraction and refinement of invariant pairs for matrix polynomials. *Linear Algebra and its Applications*, 435(3):514–536, 2011.

- [9] Wolf-Jürgen Beyn. An integral method for solving nonlinear eigenvalue problems. *Linear Algebra and its Applications*, 436(10):3839–3863, 2012.
- [10] Wolf-Jürgen Beyn, Cedric Effenberger, and Daniel Kressner. Continuation of eigenvalues and invariant pairs for parameterized nonlinear eigenvalue problems. *Numerische Mathematik*, 119:489–516, 2011.
- [11] David Bindel and Amanda Hood. Localization theorems for nonlinear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1728–1749, 2013.
- [12] Michael C Brennan, Mark Embree, and Serkan Gugercin. Contour integral methods for nonlinear eigenvalue problems: A systems theoretic approach. *SIAM Review*, 65(2):439–470, 2023.
- [13] Rune Brincker and Carlos Ventura. *Introduction to Operational Modal Analysis*. John Wiley & Sons, 2015.
- [14] Steven Cox and Enrique Zuazua. The rate at which energy decays in a damped string. *Communications in partial differential equations*, 19(1-2):213–243, 1994.
- [15] Steven J. Cox. Setting and solving the string net equations. Unpublished notes, 2006.
- [16] Harald Cramér. *Mathematical Methods of Statistics*, volume 26. Princeton university press, 1999.
- [17] René Dáger and Enrique Zuazua. *Wave Propagation, Observation and Control in 1-D Flexible Multi-Structures*. Springer, 2006.
- [18] Frederic Danion, E Varraine, Mireille Bonnard, and Jean Pailhous. Stride variability in human gait: the effect of stride frequency and stride length. *Gait & Posture*, 18(1):69–77, 2003.

- [19] Mark Embree and A. Cosmin Ioniță. Pseudospectra of Loewner pencils. In C. Beattie, P. Benner, M. Embree, S. Gugergin, and S. Lefteriu, editors, *Realization and Model Reduction of Dynamical Systems: A Festschrift in Honor of the 70th Birthday of Thanos Antoulas*, pages 59–78. Springer, Cham, Switzerland, 2022.
- [20] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.
- [21] Samuel Gaudet, Claude Gauthier, Sophie Léger, and Cory Walker. The vibrations of a real 3-string: the timbre of the tritare. *Journal of Sound and Vibration*, 281(1-2): 219–234, 2005.
- [22] Brendan Gavin, Agnieszka Międlar, and Eric Polizzi. Feast eigensolver for nonlinear eigenvalue problems. *Journal of Computational Science*, 27:107–117, 2018.
- [23] Deepam Goyal and BS Pabla. Condition based maintenance of machine tools—a review. *CIRP Journal of Manufacturing Science and Technology*, 10:24–35, 2015.
- [24] Jennifer Neal Green. Modeling spider webs as multilinked structures using a chebyshev pseudospectral collocation method. Master’s thesis, Virginia Tech, 2018.
- [25] Martin H Gutknecht and Thomas Schmelzer. The block grade of a block Krylov space. *Linear Algebra and its Applications*, 430(1):174–185, 2009.
- [26] Stefan Güttel and Françoise Tisseur. The nonlinear eigenvalue problem. *Acta Numerica*, 26:1–94, 2017.
- [27] Fei Han, Danhui Dan, and Wei Cheng. Extension of dynamic stiffness method to complicated damped structures. *Computers & Structures*, 208:143–150, 2018.

- [28] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994.
- [29] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2012.
- [30] James Kestyn, Eric Polizzi, and Ping Tak Peter Tang. Feast eigensolver for non-hermitian problems. *SIAM Journal on Scientific Computing*, 38(5):S772–S799, 2016.
- [31] Yuma Koizumi, Yohei Kawaguchi, Keisuke Imoto, Toshiki Nakamura, Yuki Nikaido, Ryo Tanabe, Harsh Purohit, Kaori Suefusa, Takashi Endo, Masahiro Yasuda, et al. Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring. *arXiv preprint arXiv:2006.05822*, 2020.
- [32] Daniel Kressner. A block Newton method for nonlinear eigenvalue problems. *Numerische Mathematik*, 114:355–372, 2009.
- [33] John E Lagnese, Günter Leugering, and EJP Georg Schmidt. *Modeling, Analysis and Control of Dynamic Elastic Multi-Link Structures*. Springer, 2012.
- [34] Benzhe Li, Yindi Jing, and Wilsun Xu. A generic waveform abnormality detection method for utility equipment condition monitoring. *IEEE Transactions on Power Delivery*, 32(1):162–171, 2016.
- [35] Krzysztof Mackala. Optimisation of performance through kinematic analysis of the different phases of the 100 metres. *New Studies in Athletics*, 22(2):7–16, 2007.
- [36] AJ Mayo and AC Antoulas. A framework for the solution of the generalized realization problem. *Linear Algebra and its Applications*, 425(2-3):634–662, 2007.
- [37] Reno Motro. *Tensegrity: Structural Systems for the Future*. Elsevier, 2003.

- [38] PCB Piezotronics. Model 352b. Online: <https://www.pcb.com/products?model=352B>. Accessed: 30 Apr. 2021.
- [39] NC Perkins and CD Mote, Jr. Comments on curve veering in eigenvalue problems. *Journal of Sound and Vibration*, 106(3):451–463, 1986.
- [40] Eric Polizzi. Density-matrix-based algorithm for solving eigenvalue problems. *Phys. Rev. B*, 79:115112, 2009.
- [41] C Radhakrishna Rao et al. Information and the accuracy attainable in the estimation of statistical parameters. *Breakthroughs in Statistics*, pages 235–247, 1992.
- [42] EJP Georg Schmidt. On the modelling and exact controllability of networks of vibrating strings. *SIAM Journal on Control and Optimization*, 30(1):229–245, 1992.
- [43] Robert E Skelton and Mauricio C De Oliveira. *Tensegrity Systems*, volume 1. Springer, 2009.
- [44] Eduardo D Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, volume 6. Springer, 2013.
- [45] Mikhail Teytel. How rare are multiple eigenvalues? *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 52(8):917–934, 1999.
- [46] Slavisa Tomic, Marko Beko, and Rui Dinis. Rss-based localization in wireless sensor networks using convex relaxation: Noncooperative and cooperative schemes. *IEEE Transactions on Vehicular Technology*, 64(5):2037–2050, 2014.
- [47] Roel Van Beeumen, Karl Meerbergen, and Wim Michiels. Compact rational krylov methods for nonlinear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 36(2):820–838, 2015.

- [48] Johann von Neumann and Eugene P Wigner. Über das Verhalten von Eigenwerten bei adiabatischen Prozessen. In *The Collected Works of Eugene Paul Wigner*, pages 294–297. Springer, 1993.
- [49] James Hardy Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, 1965.
- [50] W. H. Wittrick and F. W. Williams. A general algorithm for computing natural frequencies of elastic structures. *The Quarterly Journal of Mechanics and Applied Mathematics*, 24(3):263–284, 1971.
- [51] Yongsheng Yan, Xiaohong Shen, Fei Hua, and Xionghu Zhong. On the semidefinite programming algorithm for energy-based acoustic source localization in sensor networks. *IEEE Sensors Journal*, 18(21):8835–8846, 2018.