Shift-Invariant Linear Systems

Many of the ideas in this book rely on properties of shift-invariant linear systems. In the text, I introduced these properties without any indication of how to prove that they are true. In this appendix, I will sketch proofs of several important properties of shift-invariant linear systems. Then I will describe convolution and the discrete Fourier series, two tools that help us take advantage of these properties.¹

Shift-Invariant Linear Systems: Definitions

Shift-invariance is a system property that can be verified by experimental measurement. For example, in Chapter 2, I described how to check whether the optics of the eye are shift-invariant by the following measurements. Image a line through the optics onto the retina, and measure the linespread function. Then, shift the line to a new position, forming a new retinal image. Compare the new image with the original. If the images have the same shape, differing only by a shift in position, then the optics are shift-invariant over the measured range.

¹The results in this appendix are expressed using real harmonic functions. This is in contrast to the common practice of using complex notation, specifically Euler's complex exponential, as a shorthand to represent sums of harmonic functions. The exposition using complex exponentials is brief and elegant, but I believe it obscures the connection with experimental measurements. For this reason, I have avoided the development based on complex notation.
We can express the empirical property of shift-invariance using the following mathematical notation. Choose a stimulus, \( p_j \), and measure the system response, \( r_i \). Now, shift the stimulus by an amount \( j \) and measure again. If the response is shifted by \( j \) as well, then the system may be shift-invariant. Try this experiment for many values of the shift parameter, \( j \). If the experiment succeeds for all shifts, then the system is shift-invariant.

If you think about this definition as an experimentalist, you can see that there are some technical problems in making the measurements needed to verify shift-invariance. Suppose that the original stimulus and response are represented at \( N \) distinct points, \( i = 1, \ldots, N \). If we shift the stimulus three locations so that now the fourth location contains the first entry, the fifth the second, and so forth, how do we fill in the first three locations in the new stimulus? And, what do we do with the last three values, \( N - 2, N - 1, N \), which have nowhere to go?

Theorists avoid this problem by treating the real observations as if they are part of an infinite periodic set of observations. They assume that the stimuli and data are part of an infinite periodic series with a period of \( N \), equal to the number of original observations. If the data are infinite, the first three entries of the shifted vector are the three values at locations \(-3, -2, -1\). If the data are periodic with period \( N \), these values are the same as the values at \( N - 3, N - 2, N - 1 \).

The assumption that the measurements are part of an infinite and periodic sequence permits the theorist to avoid the experimentalist's practical problem. The assumption is also essential for obtaining several of the simple closed-form results concerning the properties of shift-invariant systems. The assumption is not consistent with real measurements, since real measurements cannot be made using infinite stimuli: there is always a beginning and an end to any real experiment. As an experimentalist you must always be aware that many theoretical calculations using shift-invariant methods are not valid near the boundaries of data sets, such as near the edge of an image.

Suppose we refer to the finite input as \( I \), and suppose that the measured output, \( r \), is finite. In the theoretical analysis we extend both of these functions to be infinite and periodic. We will use a hat symbol (\( \hat{\ } \)) to denote the extended functions,

\[
\hat{I}_i = I_i \quad \text{and} \quad \hat{p}_i = p_i
\]  
(A.1)

The extended functions \( \hat{I} \) and \( \hat{p} \) agree with our measurements over the measurement range from 1 to \( N \). By the periodicity assumption, the values outside of the measurement range are filled in by looking at the values within the measurement range. For example,

\[
(\ldots, \hat{I}_{-1} = \hat{I}_{N-1}, \hat{I}_0 = \hat{I}_N, \hat{I}_1, \ldots, \hat{I}_N, \hat{I}_{N+1} = \hat{I}_1, \ldots)
\]

Next, I will derive some of the properties of linear shift-invariant systems. I begin by describing these properties in terms of the system matrix (see Chapter 2). Then I will show how the simple structure of the shift-invariant system matrix permits us to relate the input and output by a summation formula called cyclic convolution. The convolution formula is so important that shift-invariant systems are sometimes called convolution systems.

In Chapter 2, I reviewed how to measure the system matrix of an optical system for one-dimensional input stimuli. We measured the image resulting from a single line at a series of uniformly spaced input locations. If the system is shift-invariant, then the columns of the system matrix are shifted copies of one another (except for edge artifacts). To create the system matrix, we extend the inputs and outputs to be periodic functions (Equation A.1). Then, we select a central block of size \( N \times N \) to be the system matrix, and we use the corresponding entries of the extended stimulus. For example, if the input stimulus consists of six values, \( p = (0, 0, 0, 1, 0, 0) \), and the response to this stimulus is the vector, \( I = (0.0, 0.3, 0.6, 0.2, 0.1, 0.0) \), then the \( 6 \times 6 \) system matrix is

\[
\hat{C} = \begin{bmatrix}
0.2 & 0.6 & 0.3 & 0.0 & 0.0 & 0.1 \\
0.1 & 0.2 & 0.6 & 0.3 & 0.0 & 0.0 \\
0.0 & 0.1 & 0.2 & 0.6 & 0.3 & 0.0 \\
0.0 & 0.0 & 0.1 & 0.2 & 0.6 & 0.3 \\
0.3 & 0.0 & 0.0 & 0.1 & 0.2 & 0.6 \\
0.6 & 0.3 & 0.0 & 0.0 & 0.1 & 0.2 \\
\end{bmatrix}
\]  
(A.2)

For a general linear system, we calculate the output using the summation formula for matrix multiplication in Equation 2.4,

\[
\hat{r}_i = \sum_{j=1}^{N} C_{ij} \hat{p}_j
\]  
(A.3)

When the linear system is shift-invariant, this summation formula simplifies for two reasons. First, because of the assumed periodicity, the summation is precisely the same when we sum over any \( N \) consecutive integers. It is useful to incorporate this generalization into the summation formula as

\[
\hat{r}_i = \sum_{j=(N)} C_{ij} \hat{p}_j
\]  
(A.4)

Since we use only cyclic convolution here, I will drop the word "cyclic" and refer to the formula simply as convolution. This is slightly abusive, but it conforms to common practice in many fields.
where the notation \( f = i(N) \) means that summation can take place over any \( N \) consecutive integers. Second, notice that for whichever \( N \times N \) block of values we choose, the typical entry of the system matrix will be

\[
C_{ij} = \hat{I}_{i-j}
\]  
(A.5)

We can use this relationship to simplify the summation further:

\[
\hat{r}_i = \sum_{j=(N)} \hat{I}_{i-j} \hat{p}_j
\]  
(A.6)

In this form, we see that the response depends only on the input and the linespread. The summation formula in Equation A.6 is called cyclic convolution. Hence, we have shown that, to compute the response of a shift-invariant linear system to any stimulus, we need measure only the linespread function.

### Convolution and Harmonic Functions

Next, we study some of the properties of the convolution formula. Most important, we will see why harmonic functions have a special role in the analysis of convolution systems.

Beginning with our analysis of optics in Chapter 2, we have relied on the fact that the response of a shift-invariant system to a harmonic function at frequency \( f \) is also a harmonic function at \( f \). In that chapter, the result was stated in two equivalent ways:

1. If the input is a harmonic at frequency \( f \), the output is a shifted and scaled copy of the harmonic.
2. The response to a harmonic at frequency \( f \) will be the weighted sum of a sinusoid and a cosinusoid at the same frequency (Equation 2.8).

We can derive this result from the convolution formula. Define a new variable, \( k = i - j \), and substitute \( k \) into Equation A.6. Remember that the summation can take place over any adjacent \( N \) values. Hence, the substitution yields a modified convolution formula,

\[
\hat{r}_i = \sum_{k=(N)} \hat{I}_k \hat{p}_{i-k}
\]  
(A.7)

Next, we use the convolution formula in Equation A.7 to compute the response to a sinusoidal input \( \sin(2\pi f_j/N) \). From trigonometry, we have that

\[
\sin\left(\frac{2\pi f_j}{N}\right) = \sin\left(\frac{2\pi f_j}{N}\right)\cos\left(\frac{2\pi f_i}{N}\right) + \sin\left(\frac{2\pi f_i}{N}\right)\cos\left(\frac{2\pi f_j}{N}\right)
\]  
(A.8)

Substitute Equation A.8 into Equation A.7, remembering that \( \sin(-k) = -\sin(k) \) and \( \cos(-k) = \cos(k) \):

\[
\hat{r}_i = \sum_{k=(N)} \hat{I}_k \sin\left(\frac{2\pi f_j}{N}\right)\cos\left(\frac{2\pi f_k}{N}\right) + \sum_{k=(N)} \hat{I}_k \sin\left(-\frac{2\pi f_k}{N}\right)\cos\left(\frac{2\pi f_j}{N}\right) \\
= \sin\left(\frac{2\pi f_j}{N}\right) \sum_{k=(N)} \hat{I}_k \cos\left(\frac{2\pi f_k}{N}\right) - \cos\left(\frac{2\pi f_i}{N}\right) \sum_{k=(N)} \hat{I}_k \sin\left(\frac{2\pi f_k}{N}\right)
\]  
(A.9)

We can simplify this expression to the form

\[
\hat{r}_i = a \sin\left(\frac{2\pi f_j}{N}\right) - b \cos\left(\frac{2\pi f_j}{N}\right)
\]  
(A.10)

where

\[
a = \sum_{k=(N)} \hat{I}_k \cos\left(\frac{2\pi f_k}{N}\right) \\
b = \sum_{k=(N)} \hat{I}_k \sin\left(\frac{2\pi f_k}{N}\right)
\]

We have shown that when the input to the system is a sinusoidal function at frequency \( f \), the output of the system is the weighted sum of a sinusoid and a cosinusoid, both at frequency \( f \). This is equivalent to showing that when the input is a sinusoid at frequency \( f \), the output will be a scaled and shifted copy of the input, \( s_f \sin\left(\frac{2\pi f}{N} + \phi_f\right) \) (see Equation 2.8). As we shall see below, it is easy to generalize this result to all harmonic functions.

### The Discrete Fourier Series: Definitions

In general, when we measure the response of a shift-invariant linear system we measure \( N \) output values. When the input is a sinusoid, or more generally a harmonic, we can specify the response using only the two numbers, \( a \) and \( b \), in Equation A.10. We would like to take advantage of this special property of shift-invariant systems. To do so, we need a method of representing input stimuli as the weighted sum of harmonic functions.

The method used to transform a stimulus into the weighted sum of harmonic functions is called the discrete Fourier transform (DFT). The representation of the stimulus as the weighted sum of harmonic functions is called the discrete Fourier series (DFS). We use the DFS to represent an extended stimulus, \( \hat{p} \):

\[
\hat{p}_i = \sum_{f=0}^{N-1} a_f \cos\left(\frac{2\pi f}{N}\right) + b_f \sin\left(\frac{2\pi f}{N}\right)
\]  
(A.11)
We are interested in that part of the extended stimulus that coincides with our measurements. We can express the relationship between the harmonic functions and the original stimulus, \( p \), using a matrix equation,

\[
p = Ca + Sb
\]

which has the matrix tableau form

\[
\begin{pmatrix}
p
\end{pmatrix} = \begin{pmatrix}
C \\
S
\end{pmatrix} \begin{pmatrix}
a \\
b
\end{pmatrix}
\]

The vectors \( a \) and \( b \) contain the coefficients \( a_f \) and \( b_f \), respectively. The columns of the matrices \( C \) and \( S \) contain the relevant portions of the cosinusoidal and sinusoidal terms, \( \cos(2\pi f t/N) \) and \( \sin(2\pi f t/N) \), that are used in the DFS representation.

The DFS represents the original stimulus as the weighted sum of a set of harmonic functions (i.e., sampled sine and cosine functions). We call these sampled harmonic functions the **basis functions** of the DFS representation. The vectors \( a \) and \( b \) contain the basis functions' weights or coefficients, which specify how much of each basis function must be added in to recreate the original stimulus, \( p \).

### The Discrete Fourier Series: Properties

Figure A.1 shows the sampled sine and cosine functions for a period of \( N = 8 \). The functions are arrayed in a circle to show how they relate to one another. There are a total of 16 basis functions. But, as you can see from Figure A.1, they are redundant. The sampled sinusoids in the columns of \( C \) repeat themselves (in reverse order); for example, when \( N = 8 \) the cosinusoids for \( f = 1, 2, 3 \) are the same as the cosinusoids for \( f = 7, 6, 5 \). The sampled sinusoids in the columns of \( S \) also repeat themselves except for a sign reversal (multiplication by \(-1\)). There are only four independent sampled cosinusoids and four independent sampled sinusoids. As a result of this redundancy, neither the \( S \) nor the matrix \( C \) is invertible.

Nevertheless, the properties of these harmonic basis functions make it simple to calculate the vectors containing the weights of the harmonic functions from the original stimulus, \( p \). To compute \( a \) and \( b \), we multiply the input by the basis functions, as in

\[
a = \frac{1}{N} C^T p \quad \text{and} \quad b = \frac{1}{N} S^T p
\]

We can derive the relationship in Equation A.13 from two observations. First, the matrix sum, \( H = S + C \) has a simple inverse. The columns of \( H \) are orthogonal to one another, so that the inverse of \( H \) is simply:

\[
H^{-1} = \frac{1}{N} H^T = \frac{1}{N} (S + C)^T
\]

Second, the columns of the matrices \( C \) and \( S \) are perpendicular to one another: \( 0 = CS \). This observation should not be surprising, since continuous sinusoids and cosinusoids are also orthogonal to one another.

We can use these two observations to derive Equation A.13 as follows. Express the fact that \( H \) and \( N/H \) are inverses as follows:

\[
I_{N \times N} = H(N/H)
\]

\[
= \frac{1}{N} (C + S)(C + S)^T
\]

\[
= \frac{1}{N} (CC^T + SS^T)
\]

where \( I_{N \times N} \) is the identity matrix. Then, multiply both sides of Equation A.14 by \( p \):

\[
p = \frac{1}{N} (CC^T p + SS^T p)
\]

Compare Equation A.17 and Equation A.12. Notice that the equations become identical if we make the assignments in Equation A.13. This completes the sketch of the proof.

---

3 This observation is the basis of another useful linear transform method called the **Hartley transform** (Bracewell, 1986).
Measurements with Harmonic Functions

Finally, I will show how to predict the response of a shift-invariant linear system to any stimulus using only the responses to unit amplitude sinusoidal inputs. This result is the logical basis for describing system performance from measurements of the response to harmonic functions. When the system is linear and shift-invariant, its responses to harmonic functions are a complete description of the system; but this is not true for arbitrary linear systems.

Because cosinusoids and sinusoids are shifted copies of one another, the response of a shift-invariant linear system to these functions is the same except for a shift. From a calculation like the one in Equation A.8, except using a cosinusoidal input, we can calculate the following result: if the response to a cosinusoid at \( f \) is the sum of a cosinusoid and sinusoid with weights \((a_f, b_f)\), then the response to a sinusoid at frequency \( f \) will have weights \((b_f, -a_f)\). Hence, if we know the response to a sinusoid at \( f \), we also know the response to a sinusoid at \( f \).

Next, we can use our knowledge of the response to sinusoids and cosinusoids at \( f \) to predict the response to any harmonic function at \( f \). Suppose that the input is a harmonic function \( a \cos(2\pi f_1 i/N) + b \sin(2\pi f_1 i/N) \), and the output is \( a' \cos(2\pi f_1 i/N) + b' \sin(2\pi f_1 i/N) \). If the response to a unit amplitude cosinusoid is \( u_f \cos(2\pi f_1 i/N) + v_f \sin(2\pi f_1 i/N) \), then the response to a unit amplitude sinuosoid is \( u_f \cos(2\pi f_1 i/N) - v_f \sin(2\pi f_1 i/N) \). Using these two facts and linearity, we calculate the coefficients of the response:

\[
a' = au_f + bv_f \\
b' = av_f - bu_f
\]  

(A.18)

We have shown that if we measure the system response to unit amplitude sinusoidal inputs, we can compute the system response to an arbitrary input stimulus as follows:

2. Calculate the output DFS coefficients using Equation A.18.
3. Reconstruct the output using Equation A.11.

You will find this series of calculations used implicitly at several points in the text. For example, I followed this organization when I described measurements of the optical quality of the lens (Chapter 2) and when I described measurements of behavioral sensitivity to spatiotemporal patterns (Chapter 7).

Display Calibration

Visual displays based on a cathode ray tube (CRT) are widely used in business, education, and entertainment. The CRT reproduces color images using principles embodied in the color-matching experiment (Chapter 4).

The design of the color CRT is one of the most important applications of vision science; thus, it is worth understanding the design as an engineering achievement. Also, because the CRT is used widely in experimental vision science, understanding how to control the CRT display is an essential skill for all vision scientists. This appendix reviews several of the principles of monitor calibration.

An Overview of a CRT Display

Figure B.1A shows the main components of a color CRT display. The display contains a cathode, or electron gun, that provides a source of electrons. The electrons are focused into a beam whose direction is deflected back and forth in a raster pattern so that it scans the faceplate of the display.

Light is emitted by a process of absorption and emission that occurs at the faceplate of the display. The faceplate consists of a phosphor painted onto a glass substrate. The phosphor absorbs electrons from the scanning beam and emits light. A signal, generally controlled from a computer, modulates the intensity of the electron beam as it scans across the faceplate. The intensity of the light emitted by the phosphor