Linear Algebra for Theoretical Neuroscience (Part 1)

Ken Miller

Current versions of all parts of this work can be found at http://www.neurotheory.columbia.edu/~ken/math-notes. Please feel free to link to this site.

I would appreciate any and all feedback that would help improve these notes as a teaching tool – what was particularly helpful, where you got stuck and what might have helped get you unstuck. I already know that more figures, problems, and neurobiological examples are needed in a future incarnation – for the most part I didn't have time to make figures – but that shouldn't discourage contributions of or suggestions as to useful figures, problems, examples. There are also many missing mathematical pieces I would like to fill in, as described on the home page for these notes. If anyone wants to turn this into a collaboration and help, I'd be open to discussing that too. Feedback can be sent to me by email, ken@neurotheory.columbia.edu

Reading These Notes (Instructions as written for classes I've taught that used these notes)

I have tried to begin at the beginning and make things clear enough that everyone can follow assuming basic college math as background. Some of it will be trivial for you; I hope none of it will be over your head, but some might. My suggested rules for reading this are:

- Read and *work through* everything. Read with pen and paper beside you. Never let yourself read through anything you don't completely understand; work through it until it is crystal clear to you. Go at your own pace; breeze through whatever is trivial for you.

- Do all of the "problems". Talk among yourselves as much as desired in coming to an understanding of them, but then actually write up the answers by yourself. Most or all of the problems are very simple; many only require one line as an answer.

  If you find a problem to be so obvious for you that it is a waste of your time or annoying to write it down, go ahead and skip it. But do be conservative in your judgements – it can be surprising how much you can learn by working out in detail what you think you understand in a general way.

  You can't understand the material without doing. In most cases, I have led you step by step through what is required. The purpose of the problems is not to test your math ability, but simply to make sure you "do" enough to achieve understanding.

- The "exercises" do not require a written answer. But — except where one is prefaced by something like "for those interested" — you should read them, make sure you understand them, and if possible solve them in your head or on paper.

- As you read these notes, mark them with feedback: things you don't understand, things you get confused by, things that seem trivial or unnecessary, suggestions, whatever. Then turn in to me a copy of your annotated notes.

References

If you want to consult other references on this material: an excellent text, although fairly mathematical, is **Differential equations, dynamical systems and linear algebra**, by Morris W. Hirsch and Steven Smale (Academic Press, NY, 1974). Gilbert Strang has written several very nice texts that are strong on intuition, including a couple of different linear algebra texts – I'm not sure of their relative strengths and weaknesses – and an **Introduction to Applied Mathematics**. A good practical reference — sort of a cheat sheet of basic results, plus computer algorithms and practical advice on doing computations — is **Numerical Recipes in C**, *2nd Edition*, by W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery (Cambridge University Press, 1992). Part 3 of these notes, which deals with non-normal matrices – matrices that do not have a complete orthonormal basis of eigenvectors – needs to be completely rewritten: since it was written, I've learned that non-normal matrices have many features not predicted by the eigenvalues that are of great relevance in neurobiology and in biology more generally, and the notes don't deal with this. In the meantime, for mathematical aspects of non-normal matrix behavior, see the book by L.N. Trefethen and M. Embree, Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators. Princeton University Press, 2005.

1 Introduction to Vectors and Matrices

We will start out by reviewing basic notation describing, and basic operations of, vectors and matrices. Why do we care about such things? In neurobiological modeling we are often dealing with arrays of variables: the activities of all of the neurons in a network at a given time; the firing rate of a neuron in each of many small epochs of time; the weights of all of the synapses impinging on a postsynaptic cell. The natural language for thinking about and analyzing the behavior of such arrays of variables is the language of vectors and matrices.

1.1 Notation

A **scalar** is simply a number – we use the term scalar to distinguish numbers from vectors, which are arrays of numbers. Scalars will be written without boldface: x, y, etc.

We will write a **vector** as a bold-faced small letter, *e.g.* $\mathbf{v}$; this denotes a *column vector*. Its elements v_i are written without bold-face:

$$\mathbf{v} = \begin{pmatrix} v_0 \\ v_1 \\ \dots \\ v_{N-1} \end{pmatrix} \tag{1.1}$$

Here N, the number of elements, is the **dimension** of $\mathbf{v}$. The **transpose** of $\mathbf{v}$, $\mathbf{v}^{\mathrm{T}}$, is a row vector:

$$\mathbf{v}^{\mathrm{T}} = (v_0, v_1, \dots, v_{N-1}). \tag{1.2}$$

The transpose of a row vector, in turn, is a column vector; in particular, $(\mathbf{v}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{v}$. Thus, to keep things easier to write, we can also write $\mathbf{v}$ as $\mathbf{v} = (v_0, v_1, \dots, v_{N-1})^{\mathrm{T}}$.[1]

We will write a **matrix** as a bold-faced capital letter, *e.g.* $\mathbf{M}$; its elements M_{ij}, where i indicates the row and j indicates the column, are written without boldface:

$$\mathbf{M} = \begin{pmatrix} M_{00} & M_{01} & \dots & M_{0(N-1)} \\ M_{10} & M_{11} & \dots & M_{1(N-1)} \\ \dots & \dots & \dots & \dots \\ M_{(N-1)0} & M_{(N-1)1} & \dots & M_{(N-1)(N-1)} \end{pmatrix} \tag{1.3}$$

This is a square, $N \times N$ matrix. A matrix can also be rectangular, *e.g.* a $P \times N$ matrix would have P rows and N columns. In particular, an N-dimensional vector can be regarded as an $N \times 1$ matrix, while its transpose can be regarded as a $1 \times N$ matrix. For the most part, we will only be concerned with square matrices and with vectors, although we will eventually return to non-square matrices.

The **transpose** of $\mathbf{M}$, $\mathbf{M}^{\mathrm{T}}$, is the matrix with elements $M^{\mathrm{T}}_{ij} = M_{ji}$:

$$\mathbf{M}^{\mathrm{T}} = \begin{pmatrix} M_{00} & M_{10} & \dots & M_{(N-1)0} \\ M_{01} & M_{11} & \dots & M_{(N-1)1} \\ \dots & \dots & \dots & \dots \\ M_{0(N-1)} & M_{1(N-1)} & \dots & M_{(N-1)(N-1)} \end{pmatrix} \tag{1.4}$$

[1] Those of you who have taken upper-level physics courses may have seen the "bra" and "ket" notation, $|\mathbf{v}\rangle$ ("ket") and $\langle\mathbf{v}|$ ("bra"). For vectors, these are just another notation for a vector and its transpose: $\mathbf{v} = |\mathbf{v}\rangle$, $\mathbf{v}^{\mathrm{T}} = \langle\mathbf{v}|$. The bra and ket notation is useful because one can effortlessly move between vectors and functions using the same notation, making transparent the fact – which we will eventually discuss in these notes – that vector spaces and function spaces can all be dealt with using the same formalism of linear algebra. But we will be focusing on vectors and will stick to the simple notation $\mathbf{v}$ and $\mathbf{v}^{\mathrm{T}}$.

Note, under this definition, the transpose of a $P \times N$ matrix is an $N \times P$ matrix.

Definition 1.1 *A square matrix* $\mathbf{M}$ *is called* **symmetric** *if* $\mathbf{M} = \mathbf{M}^{\mathrm{T}}$*; that is, if* $M_{ij} = M_{ji}$ *for all* i *and* j*.*

Example: The matrix $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ is not symmetric. Its transpose is $\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^{\mathrm{T}}$. The matrix $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ is symmetric; it is equal to its own transpose.

A final point about notation: we will generally use 0 to mean any object all of whose entries are 0. It should be clear from context whether the thing that is set equal to zero is just a number, or a vector all of whose elements are 0, or a matrix all of whose elements are 0. So we abuse notation by using the same symbol 0 for all of these cases.

1.2 Matrix and vector addition

The definitions of matrix and vector addition are simple: you can only add objects of the same type and size, and things add element-wise:

- **Addition of two vectors:** $\mathbf{v} + \mathbf{x}$ is the vector with elements $(\mathbf{v} + \mathbf{x})_i = v_i + x_i$.
- **Addition of two matrices:** $\mathbf{M} + \mathbf{P}$ is the matrix with elements $(\mathbf{M} + \mathbf{P})_{ij} = M_{ij} + P_{ij}$.

Subtraction works the same way: $(\mathbf{v} - \mathbf{x})_i = v_i - x_i$, $(\mathbf{M} - \mathbf{P})_{ij} = M_{ij} - P_{ij}$.

Addition or subtraction of two vectors has a simple geometrical interpretation ... (illustrate).

1.3 Multiplication by a scalar

Vectors or matrices can be multiplied by a scalar, which is just defined to mean multiplying every element by the scalar:

- **Multiplication of a vector or matrix by a scalar**: Let k be a scalar (an ordinary number). The vector $k\mathbf{v} = \mathbf{v}k = (kv_0, kv_1, \ldots, kv_{N-1})^{\mathrm{T}}$. The matrix $k\mathbf{M} = \mathbf{M}k$ is the matrix with entries $(k\mathbf{M})_{ij} = kM_{ij}$.

1.4 Linear Mappings of Vectors

Consider a function $\mathbf{M}(\mathbf{v})$ that maps an N-dimensional vector $\mathbf{v}$ to a P-dimensional vector $\mathbf{M}(\mathbf{v}) = (M_0(\mathbf{v}), M_1(\mathbf{v}), \ldots, M_{P-1}(\mathbf{v}))^{\mathrm{T}}$. We say that this mapping is **linear** if (1) for all scalars a, $\mathbf{M}(a\mathbf{v}) = a\mathbf{M}(\mathbf{v})$ and (2) for all pairs of N-dimensional vectors $\mathbf{v}$ and $\mathbf{w}$, $\mathbf{M}(\mathbf{v} + \mathbf{w}) = \mathbf{M}(\mathbf{v}) + \mathbf{M}(\mathbf{w})$. It turns out that the most general linear mapping can be written in the following form: each element of $\mathbf{M}(\mathbf{v})$ is determined by a linear combination of the elements of $\mathbf{v}$, so that for each i, $M_i(\mathbf{v}) = M_{i0}v_0 + M_{i1}v_1 + \ldots + M_{i(P-1)}v_{P-1} = \sum_j M_{ij}v_j$ for some constants M_{ij}.

This motivates the definition of matrices and matrix multiplication. We define the $P \times N$ matrix $\mathbf{M}$ to have the elements M_{ij}, and the product of $\mathbf{M}$ with $\mathbf{v}$, $\mathbf{M}\mathbf{v}$, is defined by $(\mathbf{M}\mathbf{v})_i = \sum_j M_{ij}v_j$. Thus, the set of all possible linear functions corresponds precisely to the set of all possible matrices, and matrix multiplication of a vector corresponds to a linear transformation of the vector. This motivates the definition of matrix multiplication, to which we now turn.

1.5 Matrix and vector multiplication

The definitions of matrix and vector multiplication sound complicated, but it gets easy when you actually do it (see examples below, and Problem 1.1). The basic idea is this:

- The multiplication of two objects **A** and **B** to form **AB** is only defined if the number of columns of **A** (the object on the left) equals the number of rows of **B** (the object on the right). Note that this means that order matters! (In general, even if both **AB** and **BA** are defined, they need not be the same thing: $\mathbf{AB} \neq \mathbf{BA}$).

- To form **AB**, take row (i) of **A**; rotate it clockwise to form a column, and multiply each element with the corresponding element of column (j) of **B**. Sum the results of these multiplications, and that gives a single number, entry (ij) of the resulting output structure **AB**.

Let's see what this means by defining the various possible allowed cases (if this is confusing, just keep plowing on through; working through Problem 1.1 should clear things up):

- **Multiplication of two matrices:** $\mathbf{MP}$ is the matrix with elements $(\mathbf{MP})_{ik} = \sum_j M_{ij}P_{jk}$.

  Example:
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}$$

- **Multiplication of a column vector by a matrix:** $\mathbf{Mv} = ((\mathbf{Mv})_0, (\mathbf{Mv})_1, \ldots, (\mathbf{Mv})_{N-1})^{\mathrm{T}}$ where $(\mathbf{Mv})_i = \sum_j M_{ij}v_j$. $\mathbf{Mv}$ is a column vector.

  Example:
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax+by \\ cx+dy \end{pmatrix}$$

- **Multiplication of a matrix by a row vector.** $\mathbf{v}^{\mathrm{T}}\mathbf{M} = ((\mathbf{v}^{\mathrm{T}}\mathbf{M})_0, (\mathbf{v}^{\mathrm{T}}\mathbf{M})_1, \ldots, (\mathbf{v}^{\mathrm{T}}\mathbf{M})_{N-1})$ where $(\mathbf{v}^{\mathrm{T}}\mathbf{M})_j = \sum_i v_iM_{ij}$. $\mathbf{v}^{\mathrm{T}}\mathbf{M}$ is a row vector.

  Example:
$$\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} xa+yc & xb+yd \end{pmatrix}$$

- **Dot or inner product of two vectors:** multiplication by a row vector on the left of a column vector on the right. $\mathbf{v}\cdot\mathbf{x}$ is a notation for the dot product, which is defined by $\mathbf{v}\cdot\mathbf{x} = \mathbf{v}^{\mathrm{T}}\mathbf{x} = \sum_i v_ix_i$. $\mathbf{v}^{\mathrm{T}}\mathbf{x}$ is a *scalar*, that is, a single number. Note from this definition that $\mathbf{v}^{\mathrm{T}}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{v}$.

  Example:
$$\begin{pmatrix} x \\ y \end{pmatrix}\cdot\begin{pmatrix} z \\ w \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} z \\ w \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} z \\ w \end{pmatrix} = xz+yw$$

- **Outer product of two vectors:** multiplication by a column vector on the left of a row vector on the right. $\mathbf{vx}^{\mathrm{T}}$ is a *matrix*, with elements $(\mathbf{vx}^{\mathrm{T}})_{ij} = v_ix_j$.
$$\begin{pmatrix} x \\ y \end{pmatrix}\begin{pmatrix} z \\ w \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} x \\ y \end{pmatrix}\begin{pmatrix} z & w \end{pmatrix} = \begin{pmatrix} xz & xw \\ yz & yw \end{pmatrix}$$

These rules will all become obvious with a tiny bit of practice, as follows:

Problem 1.1 *Let* $\mathbf{v} = (1, 2, 3)^{\mathrm{T}}$, $\mathbf{x} = (4, 5, 6)^{\mathrm{T}}$.

- *Compute the inner product* $\mathbf{v}^{\mathrm{T}}\mathbf{x}$ *and the outer products* $\mathbf{v}\mathbf{x}^{\mathrm{T}}$ *and* $\mathbf{x}\mathbf{v}^{\mathrm{T}}$*. To compute* $\mathbf{v}^{\mathrm{T}}\mathbf{x}$*, begin by writing the row vector* $\mathbf{v}^{\mathrm{T}}$ *to the left of the column vector* $\mathbf{x}$*, so you can* **see** *the multiplication that the inner product consists of, and why it results in a single number, a scalar. Similarly, to compute the outer products, say* $\mathbf{v}\mathbf{x}^{\mathrm{T}}$*, begin by writing the column vector* $\mathbf{v}$ *to the left of the row vector* $\mathbf{x}^{\mathrm{T}}$*, so you can* **see** *the multiplication, and why it results in a matrix of numbers. Finally, let* $A = \mathbf{v}\mathbf{x}^{\mathrm{T}}$*, and note that* $A^{\mathrm{T}} = \mathbf{x}\mathbf{v}^{\mathrm{T}}$*; that is,* $(\mathbf{v}\mathbf{x}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{x}\mathbf{v}^{\mathrm{T}}$.

- *Compute the matrix* $AA^{\mathrm{T}} = \mathbf{v}\mathbf{x}^{\mathrm{T}}\mathbf{x}\mathbf{v}^{\mathrm{T}}$ *in two ways: as a product of two matrices,* $(\mathbf{v}\mathbf{x}^{\mathrm{T}})(\mathbf{x}\mathbf{v}^{\mathrm{T}})$*, and as a scalar times the outer product of two vectors:* $\mathbf{v}(\mathbf{x}^{\mathrm{T}}\mathbf{x})\mathbf{v}^{\mathrm{T}} = (\mathbf{x}^{\mathrm{T}}\mathbf{x})(\mathbf{v}\mathbf{v}^{\mathrm{T}})$ *(note, in the last step we have made use of the fact that a scalar,* $(\mathbf{x}^{\mathrm{T}}\mathbf{x})$*, commutes with anything and so can be pulled out front). Show that the outcomes are identical.*

- *Show that* $AA^{\mathrm{T}} \neq A^{\mathrm{T}}A$*; that is, matrix multiplication need not commute. Note that* $A^{\mathrm{T}}A$ *can also be written* $\mathbf{x}(\mathbf{v}^{\mathrm{T}}\mathbf{v})\mathbf{x}^{\mathrm{T}} = (\mathbf{v}^{\mathrm{T}}\mathbf{v})(\mathbf{x}\mathbf{x}^{\mathrm{T}})$.

- *Compute the row vector* $\mathbf{x}^{\mathrm{T}}\mathbf{v}\mathbf{x}^{\mathrm{T}}$ *in two ways, as a row vector times a matrix:* $\mathbf{x}^{\mathrm{T}}(\mathbf{v}\mathbf{x}^{\mathrm{T}})$*; and as a scalar times a row vector:* $(\mathbf{x}^{\mathrm{T}}\mathbf{v})\mathbf{x}^{\mathrm{T}}$*. Show that the outcomes are identical, and proportional to the vector* $\mathbf{x}^{\mathrm{T}}$.

- *Compute the column vector* $\mathbf{v}\mathbf{x}^{\mathrm{T}}\mathbf{v}$ *in two ways: as a matrix times a column vector:* $(\mathbf{v}\mathbf{x}^{\mathrm{T}})\mathbf{v}$*; and as a column vector times a scalar* $\mathbf{v}(\mathbf{x}^{\mathrm{T}}\mathbf{v})$*. Show that the outcomes are identical, and proportional to* $\mathbf{v}$.

Exercise 1.1 *Make up more examples as needed to make sure the definitions above of matrix and vector multiplication are intuitively clear to you.*

Problem 1.2

1. *Prove that for any vectors* $\mathbf{v}$ *and* $\mathbf{x}$ *and matrices* $\mathbf{M}$ *and* $\mathbf{P}$*:* $(\mathbf{v}\mathbf{x}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{x}\mathbf{v}^{\mathrm{T}}$*,* $(\mathbf{M}\mathbf{v})^{\mathrm{T}} = \mathbf{v}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}$*, and* $(\mathbf{M}\mathbf{P})^{\mathrm{T}} = \mathbf{P}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}$*. Hint: in general, the way to get started in a proof is to write down precisely what you need to prove. In this case, it helps to write this down in terms of indices. For example, here's how to solve the first one: we need to show that* $((\mathbf{v}\mathbf{x}^{\mathrm{T}})^{\mathrm{T}})_{ij} = (\mathbf{x}\mathbf{v}^{\mathrm{T}})_{ij}$ *for any* i *and* j*. So write down what each side means:* $((\mathbf{v}\mathbf{x}^{\mathrm{T}})^{\mathrm{T}})_{ij} = (\mathbf{v}\mathbf{x}^{\mathrm{T}})_{ji} = v_j x_i$*, while* $(\mathbf{x}\mathbf{v}^{\mathrm{T}})_{ij} = x_i v_j$*. We're done! –* $v_j x_i = x_i v_j$*, so just writing down what the proof requires, in terms of indices, is enough to solve the problem.*

2. *Show that* $(\mathbf{M}\mathbf{P}\mathbf{Q})^{\mathrm{T}} = \mathbf{Q}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}$ *for any matrices* $\mathbf{M}$*,* $\mathbf{P}$ *and* $\mathbf{Q}$*. (Hint: apply the two-matrix result first to the product of the two matrices* $\mathbf{M}$ *and* $(\mathbf{P}\mathbf{Q})$*; then apply it again to the product of the two matrices* $\mathbf{P}$ *and* $\mathbf{Q}$*.)*

As you might guess, or easily prove, this result extends to a product of any number of matrices: you form the transpose of the product by reversing their order and taking the transpose of each element.

As the above problems and exercises suggest, matrix and vector multiplication are *associative*: $\mathbf{ABC} = (\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$, etc.; but they are not in general *commutative*: $\mathbf{AB} \neq \mathbf{BA}$. However, a scalar — a number — always commutes with anything.

From the dot product, we can also define two other important concepts:

Definition 1.2 *The* **length** *or absolute value* $|\mathbf{v}|$ *of a vector* $\mathbf{v}$ *is given by* $|\mathbf{v}| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{\sum_i v_i^2}$.

This is just the standard Euclidean length of the vector: the distance from the origin (the vector 0) to the end of the vector. This might also be a good place to remind you of your high school geometry: the dot product of any two vectors $\mathbf{v}$ and $\mathbf{w}$ can be expressed $\mathbf{v} \cdot \mathbf{w} = |v||w| \cos\theta$ where θ is the angle between the two vectors.

Definition 1.3 *Two vectors* $\mathbf{v}$ *and* $\mathbf{w}$ *are said to be* **orthogonal** *if* $\mathbf{v} \cdot \mathbf{w} = 0$*.*

Geometrically, two vectors are orthogonal when the angle between them is 90^o, so that the cosine of the angle between them is 0.

Problem 1.3 Better understanding matrix multiplication: *Let the* $N \times N$ *matrix* $\mathbf{M}$ *have columns* $\mathbf{c}_i$*:* $\mathbf{M} = (\ \mathbf{c}_0\ \mathbf{c}_1\ \dots\ \mathbf{c}_{N-1}\ )$ *where each* $\mathbf{c}_i$ *is an N-dimensional column vector. Let it have rows* $\mathbf{r}_i^{\mathrm{T}}$*:* $\mathbf{M} = (\ \mathbf{r}_0\ r_1\ \dots\ \mathbf{r}_{N-1}\ )^{\mathrm{T}}$*.*

1. *Show that for any vector* $\mathbf{v}$*,* $\mathbf{Mv} = (\mathbf{r}_0 \cdot \mathbf{v}\ \mathbf{r}_1 \cdot \mathbf{v}\ \dots\ \mathbf{r}_{N-1} \cdot \mathbf{v})^{\mathrm{T}}$*. (Hint: note that* $M_{ij} = (\mathbf{r}_i)_j$*, and show that* $(\mathbf{Mv})_k = \mathbf{r}_k \cdot \mathbf{v}$*; that is,* $(\mathbf{Mv})_k = \sum_i M_{ki} v_i$*, while* $\mathbf{r}_k \cdot \mathbf{v} = \sum_i (\mathbf{r}_k)_i v_i$*, so show that these are equal.) Thus, any vector* $\mathbf{v}$ *that is orthogonal to all the rows of* $\mathbf{M}$*, that is, for which* $\mathbf{r}_i \cdot \mathbf{v} = 0\ \forall i$*, is mapped to the zero vector.*

2. *Show that for any vector* $\mathbf{v}$*,* $\mathbf{Mv} = \sum_i v_i \mathbf{c}_i$*. (Hint: note that* $M_{ij} = (\mathbf{c}_j)_i$*, where* $(\mathbf{c}_j)_i$ *is the* i^{th} *component of* $\mathbf{c}_j$*; and show that* $(\mathbf{Mv})_k = (\sum_i v_i \mathbf{c}_i)_k = \sum_i v_i (\mathbf{c}_i)_k$*.) Thus, the* **range** *of* $\mathbf{M}$ *– the set of vectors* $\{\mathbf{w} : \mathbf{w} = \mathbf{Mv}$ *for some vector* $\mathbf{v}\}$ *– is composed of all linear combinations of the columns of* $\mathbf{M}$ *(a linear combination of the* $\mathbf{c}_i$ *is a combination* $\sum_i a_i \mathbf{c}_i$ *for some constants* a_i*). You can gain some intuition for this result by noting that, in the matrix multiplication* $\mathbf{Mv}$*,* v_0 *only multiplies elements of* $\mathbf{c}_0$*,* v_1 *only multiplies elements of* $\mathbf{c}_1$*, etc.*

3. *Let's make this concrete: consider the matrix* $\mathbf{M} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ *and the vector* $\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$*. Compute* $\mathbf{Mv}$ *the ordinary way, which corresponds to the format of item 1 above. Now instead write* $\sum_i v_i \mathbf{c}_i$ *where* $\mathbf{c}_i$ *are the columns of* $\mathbf{M}$*, and show that this gives the same answer.*

4. *Consider another* $N \times N$ *matrix* $\mathbf{P}$*, with columns* $\mathbf{d}_i$ *and rows* $\mathbf{s}_i^{\mathrm{T}}$*.*

   - *Show that* $(\mathbf{MP})_{ij} = \mathbf{r}_i \cdot \mathbf{d}_j$*. (Hint:* $(\mathbf{MP})_{ij} = \sum_k M_{ik} P_{kj}$*, while* $\mathbf{r}_i \cdot \mathbf{d}_j = \sum_k (\mathbf{r}_i)_k (\mathbf{d}_j)_k$*; show that these are equal.)*
   - *Show that* $\mathbf{MP} = \sum_i \mathbf{c}_i \mathbf{s}_i^{\mathrm{T}}$*, by showing that* $(\mathbf{MP})_{kj} = (\sum_i \mathbf{c}_i \mathbf{s}_i^{\mathrm{T}})_{kj} = \sum_i (\mathbf{c}_i)_k (\mathbf{s}_i)_j$*. Note that each term* $\mathbf{c}_i \mathbf{s}_i^{\mathrm{T}}$ *is a matrix. Again, you can gain some intuition for this result by noticing that elements of* $\mathbf{s}_i$ *only multiply elements of* $\mathbf{c}_i$ *in the matrix multiplication.*

5. *Let's make this concrete: consider* $\mathbf{M} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ *and* $\mathbf{P} = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$*. Compute* $\mathbf{MP}$ *the ordinary way, which amounts to* $(\mathbf{MP})_{ij} = \mathbf{r}_i \cdot \mathbf{d}_j$*. Now instead write it as* $\mathbf{MP} = \sum_i \mathbf{c}_i \mathbf{s}_i^{\mathrm{T}}$*, and show that this sums to the same thing.*

1.6 The Identity Matrix

The *identity matrix* will be written as $\mathbf{1}$. This is the matrix that is 1 on the diagonal and zero otherwise:

$$\mathbf{1} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{1.5}$$

Note that $\mathbf{1v} = \mathbf{v}$ and $\mathbf{v}^{\mathrm{T}}\mathbf{1} = \mathbf{v}^{\mathrm{T}}$ for any vector $\mathbf{v}$, and $\mathbf{1M} = \mathbf{M1} = \mathbf{M}$ for any matrix $\mathbf{M}$. (The dimension of the matrix $\mathbf{1}$ is generally to be inferred from context; at any point, we are referring to that identity matrix with the same dimension as the other vectors and matrices being considered).

Exercise 1.2 *Verify that* $\mathbf{1v} = \mathbf{v}$ *and* $\mathbf{v}^{\mathrm{T}}\mathbf{1} = \mathbf{v}^{\mathrm{T}}$ *for any vector* $\mathbf{v}$*, and* $\mathbf{1M} = \mathbf{M1} = \mathbf{M}$ *for any matrix* $\mathbf{M}$.

1.7 The Inverse of a Matrix

Definition 1.4 *The* **inverse** *of a square matrix* $\mathbf{M}$ *is a matrix* $\mathbf{M}^{-1}$ *satisfying* $\mathbf{M}^{-1}\mathbf{M} = \mathbf{M}\mathbf{M}^{-1} = \mathbf{1}$.

Fact 1.1 *For square matrices* $\mathbf{A}$ *and* $\mathbf{B}$*, if* $\mathbf{AB} = \mathbf{1}$*, then* $\mathbf{BA} = 1$*; so knowing either* $\mathbf{AB} = \mathbf{1}$ *or* $\mathbf{BA} = \mathbf{1}$ *is enough to establish that* $\mathbf{A} = \mathbf{B}^{-1}$ *and* $\mathbf{B} = \mathbf{A}^{-1}$.

Not all matrices $\mathbf{M}$ have an inverse; but if a matrix has an inverse, that inverse is unique (there is at most one matrix that is the inverse of $\mathbf{M}$, proof for square matrices: suppose $\mathbf{C}$ and $\mathbf{B}$ are both inverses of $\mathbf{A}$. Then $\mathbf{CAB} = \mathbf{C}(\mathbf{AB}) = \mathbf{C1} = \mathbf{C}$; but also $\mathbf{CAB} = (\mathbf{CA})\mathbf{B} = \mathbf{1B} = \mathbf{B}$; hence $\mathbf{C} = \mathbf{B}$). Intuitively, the inverse of $\mathbf{M}$ "undoes" whatever $\mathbf{M}$ does: if you apply $\mathbf{M}$ to a vector or matrix, and then apply $\mathbf{M}^{-1}$ to the result, you end up having applied the identity matrix, that is, not having changed anything. If a matrix has an inverse, we say that it is **invertible**.

A matrix fails to have an inverse when it maps some nonzero vector(s) to the zero vector, 0. Suppose $\mathbf{Mv} = 0$ for $\mathbf{v} \neq 0$. Then, since matrix multiplication is a linear operation, for any other vector $\mathbf{w}$, $\mathbf{M}(a\mathbf{v} + \mathbf{w}) = a\mathbf{Mv} + \mathbf{Mw} = \mathbf{Mw}$, so all input vectors of the form $a\mathbf{v} + \mathbf{w}$ are mapped to the same output vector $\mathbf{Mw}$. Hence in this case the action of M cannot be undone – given the output vector $\mathbf{Mw}$, we cannot say which input vector produced it.

You may notice that above, we defined addition, subtraction, and multiplication for matrices, but not division. Ordinary division is really multiplying by the inverse of a number: $x/y = y^{-1}x$ where $y^{-1} = 1/y$. As you might imagine, the generalization for matrices would be multiplying by the inverse of a matrix. Since not all matrices have inverses, it turns out to be more sensible to leave it at that, and not define division as a separate operation for matrices.

Exercise 1.3 *Suppose* $\mathbf{A}$ *and* $\mathbf{B}$ *are both invertible* $N \times N$ *matrices. Show that* $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. *(Hint: just multiply* $\mathbf{AB}$ *times* $\mathbf{B}^{-1}\mathbf{A}^{-1}$ *and see what you get.) Similarly if* $\mathbf{C}$ *is another invertible* $N \times N$ *matrix,* $(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$*; etc. This should remind you of the result of problem 1.2 for transposes.*

Exercise 1.4 *Show that* $(\mathbf{A}^{\mathrm{T}})^{-1} = (\mathbf{A}^{-1})^{\mathrm{T}}$. *Hint: take the equation* $(\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}^{\mathrm{T}} = \mathbf{1}$*, and take the transpose.*

1.8 Why Vectors and Matrices? – Two Toy Problems

As mentioned at the outset, in problems of theoretical neuroscience, we are often dealing with large sets of variables — the activities of a large set of neurons in a network; the development of a large set of synaptic strengths impinging on a neuron. The equations to describe models of these systems are usually best expressed and analyzed in terms of vectors and matrices. Here are two simple examples of the formulation of problems in these terms; as we go along we will develop the tools to analyze them.

- **Development in a set of synapses.** Consider a set of N presynaptic neurons with activities a_i making synapses w_i onto a single postsynaptic cell. Take the activity of the postsynaptic cell to be $b = \sum_j w_j a_j$. Suppose there is a simple linear Hebb-like plasticity rule of the form $\tau dw_i/dt = ba_i$ for some time constant τ that determines how quickly weights change. Substituting in the expression for b, this becomes

  $$\tau \frac{dw_i}{dt} = \sum_j (a_i a_j) w_j \tag{1.6}$$

  or

  $$\tau \frac{d\mathbf{w}}{dt} = \mathbf{a}\mathbf{a}^{\mathrm{T}}\mathbf{w}. \tag{1.7}$$

  Now, suppose that input activity patterns occur with some overall statistical structure, *e.g.* some overall patterns as to which neurons tend to be coactive (or not) with one another. For example, suppose the input neurons represent the lateral geniculate nucleus (LGN), which receives visual input from the eyes and projects to primary visual cortex. We may consider spontaneous activity in the LGN before vision; or we might consider visually-induced LGN activity patterns as an animal explores its natural environment. In either case, averaged over some short time (perhaps ranging from a few minutes to a few hours), the tendency of different neurons to be coactive or not may be quite reproducible. If τ is much larger than this time, so that weights change little over this time, then we can average Eq. 1.7 and replace $\mathbf{a}\mathbf{a}^{\mathrm{T}}$ by $\langle \mathbf{a}\mathbf{a}^{\mathrm{T}} \rangle$ where $\langle \mathbf{x} \rangle$ represents the average over input activity patterns of $\mathbf{x}$. Defining $\mathbf{C} = \langle \mathbf{a}\mathbf{a}^{\mathrm{T}} \rangle$ to be the matrix of correlations between activities of the different inputs, we arrive at the equation[2]

  $$\tau \frac{d\mathbf{w}}{dt} = \mathbf{C}\mathbf{w}. \tag{1.8}$$

  Of course, this is only a toy model: weights are unbounded and can change their signs, and more generally we don't expect postsynaptic activity or plasticity to be determined by such simple linear equations. But it's useful to play with toy cars before driving real ones; as with cars, we'll find out that they do have something in common with the real thing. We will return to this model as we develop the tools to understand its behavior.

- **Activity in a network of neurons.** Consider two layers of N neurons each, an input layer and an output layer. Label the activities of the input layer neurons by a_i, $i = 0, \ldots, N-1$, and similarly label the activities of the output layer neurons by b_i. Let W_{ij} by the strength of the synaptic connection from input neuron j to output neuron i. Also let there be synaptic connections between the output neurons: let B_{ij} be the strength of the connection from output neuron j to output neuron i (we can define $B_{ii} = 0$ for all i, if we want to exclude self-synapses). Let τ be a time constant of integration in the postsynaptic neuron. Then a

[2] Equation 1.8 can also be derived starting from slightly more complicated models. For example, we might assume that the learning depends on the covariance rather than product of the postsynaptic and presynaptic activities: $\tau dw_i/dt = (b - \langle b \rangle)(a_i - \langle a_i \rangle)$. This means that, if the post- and pre-synaptic activities fluctuate up from their mean activities at the same time, the weight gets stronger (this also happens if the activites fluctuate down together, which is certainly not realistic); while if one activity goes up from its mean while the other goes down, the weight gets weaker. After averaging, this gives Eq. 1.8, but with $\mathbf{C}$ now defined by $\mathbf{C} = (\mathbf{a} - \langle \mathbf{a} \rangle)(\mathbf{a}^{\mathrm{T}} - \langle \mathbf{a}^{\mathrm{T}} \rangle)$ (check that this is so). More generally, any rules in which the postsynaptic activity depends linearly on presynaptic activity, and the weight change depends linearly on postsynaptic activity (though perhaps nonlinearly on presynaptic activity), will yield an equation of the form $\tau \frac{d\mathbf{w}}{dt} = \mathbf{C}\mathbf{w} + \mathbf{h}$ for some matrix $\mathbf{C}$ defined by the input activities and some constant vector $\mathbf{h}$. Equations of this form can also sometimes be derived to describe aspects of development starting from more nonlinear rules.

very simple, linear model of activity in the output layer, given the activity in the input layer, would be:

$$\tau \frac{db_i}{dt} = -b_i + \sum_j W_{ij} a_j + \sum_j B_{ij} b_j. \tag{1.9}$$

The $-b_i$ term on the right just says that, in the absence of input from other cells, the neuron's activity b_i decays to zero (with time constant τ). Again, this is only a toy model, *e.g.* rates can go positive or negative and are unbounded in magnitude.

Eq. 1.9 can be written as a vector equation:

$$\begin{aligned} \tau \frac{d\mathbf{b}}{dt} &= -\mathbf{b} + \mathbf{W}\mathbf{a} + \mathbf{B}\mathbf{b} \qquad (1.10) \\ &= -(\mathbf{1} - \mathbf{B})\mathbf{b} + \mathbf{W}\mathbf{a} \end{aligned}$$

$\mathbf{Wa}$ is a vector that is independent of $\mathbf{b}$: $(\mathbf{Wa})_i = \sum_j W_{ij} a_j$ is the external input to output neuron i. So, let's give it a name: we'll call the vector of external inputs $\mathbf{h} = \mathbf{Wa}$. Thus, our equation finally is

$$\tau \frac{d\mathbf{b}}{dt} = -(\mathbf{1} - \mathbf{B})\mathbf{b} + \mathbf{h} \tag{1.11}$$

This is very similar in form to Eq. 1.8 for the previous model: the right side has a term in which the variable whose time derivative we are studying ($\mathbf{b}$ or $\mathbf{w}$) is multiplied by a matrix (here, $-(\mathbf{1} - \mathbf{B})$; previously, $\mathbf{C}$). In addition, this equation now has a term $\mathbf{h}$ independent of that variable. (In general, an equation of the form $\frac{d}{dt}\mathbf{x} = \mathbf{C}\mathbf{x}$ is called homogeneous, while one with an added constant term, $\frac{d}{dt}\mathbf{x} = \mathbf{C}\mathbf{x} + \mathbf{h}$, is called inhomogeneous.)

We can also write down an equation for the steady-state or fixed-point output activity pattern $\mathbf{b}^{\text{FP}}$ for a given input activity pattern $\mathbf{h}$: by definition, a steady state or fixed point is a point where $\frac{d\mathbf{b}}{dt} = 0$. Thus, the fixed point is determined by

$$(\mathbf{1} - \mathbf{B})\mathbf{b}^{\text{FP}} = \mathbf{h} \tag{1.12}$$

If the matrix $(\mathbf{1} - \mathbf{B})$ has an inverse, $(\mathbf{1} - \mathbf{B})^{-1}$, then we can multiply both sides of Eq. 1.12 by this inverse to obtain

$$\mathbf{b}^{\text{FP}} = (\mathbf{1} - \mathbf{B})^{-1}\mathbf{h} \tag{1.13}$$

We'll return to this later to better understand what this equation means.

2 Coordinate Systems, Orthogonal Basis Vectors, and Orthogonal Change of Basis

To solve the equations that arise in the toy models just introduced, and in many other models, it will be critical to be able to view the problem in alternative coordinate systems. Choice of the right coordinate system will greatly simplify the equations and allow us to solve them. So, in this section we address the topic of coordinate systems: what they are, what it means to change coordinates, and how we change them. We begin by addressing the problem in two dimensions, where one can draw pictures and things are more intuitively clear. We'll then generalize our results to higher dimensions, as needed to address problems involving many variables such as our toy models. For now we are only going to consider coordinate systems in which each coordinate axis is orthogonal to all the other coordinate axes; much later we will consider more general coordinate systems.

2.1 Coordinate Systems and Orthogonal Basis Vectors in Two Dimensions

When we write $\mathbf{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix}$, we are working in some coordinate system. For example, in Fig. 2.1, v_x and v_y are the coordinates of $\mathbf{v}$ along the x and y axes, respectively, so these are the coordinates of $\mathbf{v}$ in the x, y coordinate system. What do these coordinates mean? v_x is the extent of $\mathbf{v}$ in the x direction, while v_y is its extent in the y direction. How do we compute v_x and v_y? If ϕ is the angle between the x axis and $\mathbf{v}$, then from trigonometry, $v_x = |v| \cos\phi$, while $v_y = |v| \sin\phi$.

We can express this in more general form by defining *basis vectors*: vectors of unit length along each of our orthogonal coordinate axes. The basis vectors along the x and y directions, when expressed in the x and y coordinate system, are $\mathbf{e}_x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathbf{e}_y = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, respectively; that is, $\mathbf{e}_x$ is the vector with extent 1 in the x direction and 0 in the y direction, and similarly for $\mathbf{e}_y$. Note that these basis vectors are orthogonal: $\mathbf{e}_x \cdot \mathbf{e}_y = 0$. Then the same geometry gives $\mathbf{e}_x \cdot \mathbf{v} = |e_x||v| \cos\phi = |v| \cos\phi$. That is, $\mathbf{e}_x \cdot \mathbf{v}$ gives the component of $\mathbf{v}$ along the x axis, v_x. We can also see this directly from the definition of the dot product: $\mathbf{e}_x^{\mathrm{T}}\mathbf{v} = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix} = v_x$. Similarly, $\mathbf{e}_y \cdot \mathbf{v} = |v| \sin\phi = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix} = v_y$.

So, we can understand the statement that $\mathbf{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix}$ in the x, y coordinate system to mean that $\mathbf{v}$ has v_x units of the $\mathbf{e}_x$ basis vector, and v_y units of the $\mathbf{e}_y$ basis vector, where $v_x = \mathbf{e}_x^{\mathrm{T}}\mathbf{v}$ and $v_y = \mathbf{e}_y^{\mathrm{T}}\mathbf{v}$:

$$\mathbf{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix} = v_x \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_y \begin{pmatrix} 0 \\ 1 \end{pmatrix} = v_x\mathbf{e}_x + v_y\mathbf{e}_y = (\mathbf{e}_x^{\mathrm{T}}\mathbf{v})\mathbf{e}_x + (\mathbf{e}_y^{\mathrm{T}}\mathbf{v})\mathbf{e}_y \tag{2.1}$$

We call $\mathbf{e}_x$ and $\mathbf{e}_y$ basis vectors, because together they form a basis for our space: any vector in our two-dimensional space can be expressed as a linear combination of $\mathbf{e}_x$ and $\mathbf{e}_y$ – a weighted sum of these basis vectors. For orthogonal basis vectors, the weighting of each basis vector in the sum is just that basis vector's dot product with the vector being expressed (note that $\mathbf{v}$ was an arbitrary vector, so Eq. 2.1 is true for any arbitrary vector in our space). Note that we can use the orthogonality of the basis vectors to show that this is the correct weighting: $\mathbf{e}_x \cdot \mathbf{v} = v_x\mathbf{e}_x \cdot \mathbf{e}_x + v_y\mathbf{e}_x \cdot \mathbf{e}_y = v_x$, and similarly $\mathbf{e}_y \cdot \mathbf{v} = v_y$.

Notice that the statement $\mathbf{v} = v_x\mathbf{e}_x + v_y\mathbf{e}_y$ is a geometric statement about the relationship between vectors – between the vector $\mathbf{v}$, and the vectors $\mathbf{e}_x$ and $\mathbf{e}_y$. This states that you can build $\mathbf{v}$ by multiplying $\mathbf{e}_x$ by v_x, multiplying $\mathbf{e}_y$ by v_y, and adding the two resulting vectors (make sure this is clear to you both geometrically – look at Fig. 2.1 – and algebraically, Eq. 2.1). This statement about vectors will be true no matter what coordinate system we express these vectors in. When we express this as $\mathbf{v} = (\mathbf{e}_x^{\mathrm{T}}\mathbf{v})\mathbf{e}_x + (\mathbf{e}_y^{\mathrm{T}}\mathbf{v})\mathbf{e}_y$, there are no numbers in the equation – this is an equation entirely about the relationship between vectors. Again, this statement will be true in any particular coordinate system in which we choose to express these vectors. But since the dot product, $\mathbf{e}_x^{\mathrm{T}}\mathbf{v}$, is a scalar – its value is independent of the coordinates in which we express the vectors – then in any coordinate system, the equation $\mathbf{v} = (\mathbf{e}_x^{\mathrm{T}}\mathbf{v})\mathbf{e}_x + (\mathbf{e}_y^{\mathrm{T}}\mathbf{v})\mathbf{e}_y$ will yield the equation $\mathbf{v} = v_x\mathbf{e}_x + v_y\mathbf{e}_y$.

2.2 Rigid Change of Basis in Two Dimensions

Equations are generally written in some coordinate system — for example, the x, y coordinate system in Fig. 2.1. But we could certainly describe the same biology equally well in other coordinate

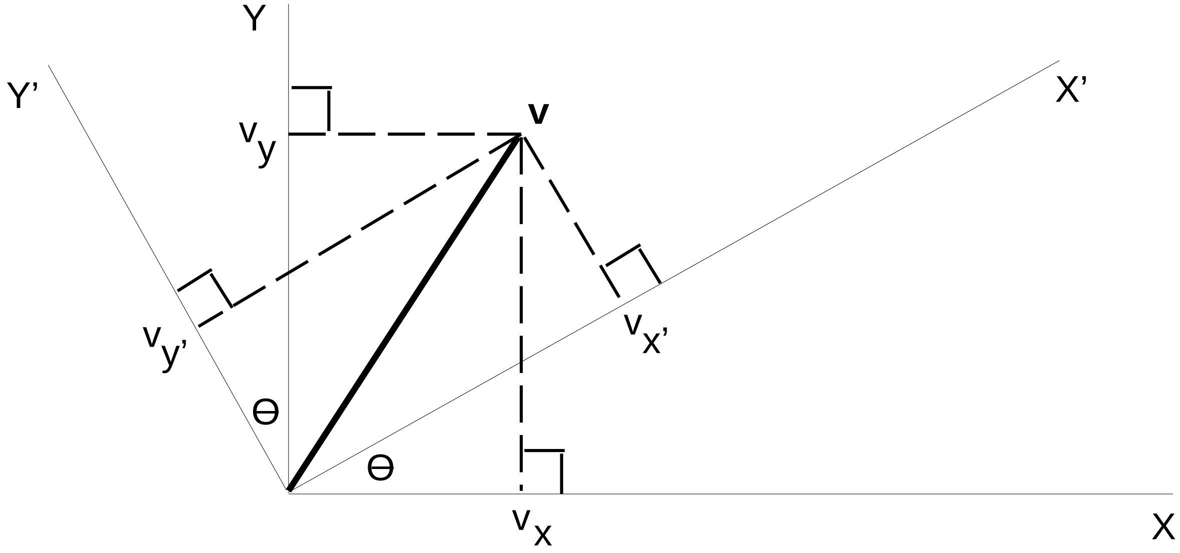

Figure 2.1: **Representation of a vector in two coordinate systems**
The vector **v** is shown represented in two coordinate systems. The (x', y') coordinate system is rotated by an angle θ from the (x, y) coordinate system. The coordinates of **v** in a given coordinate system are given by the perpendicular projections of **v** onto the coordinate axes, as illustrated by the dashed lines. Thus, in the (x, y) basis, **v** has coordinates (v_x, v_y), while in the (x', y') basis, it has coordinates $(v_{x'}, v_{y'})$.

systems. Suppose we want to describe things in the new coordinate axes, x', y', determined by a rigid rotation by an angle θ from the x, y coordinate axes, Fig. 2.1. How do we define coordinates in this new coordinate system?

Let's first define basis vectors $\mathbf{e}_{x'}$, $\mathbf{e}_{y'}$ to be the vectors of unit length along the x' and y' axes, respectively. Like any other vectors, we can write these vectors as linear combinations of $\mathbf{e}_x$ and $\mathbf{e}_y$:

$$\mathbf{e}_{x'} = (\mathbf{e}_x^{\mathrm{T}}\mathbf{e}_{x'})\mathbf{e}_x + (\mathbf{e}_y^{\mathrm{T}}\mathbf{e}_{x'})\mathbf{e}_y \tag{2.2}$$
$$\mathbf{e}_{y'} = (\mathbf{e}_x^{\mathrm{T}}\mathbf{e}_{y'})\mathbf{e}_x + (\mathbf{e}_y^{\mathrm{T}}\mathbf{e}_{y'})\mathbf{e}_y \tag{2.3}$$

From the geometry, and the fact that the basis vectors have unit length, we find the following dot products:

$$\mathbf{e}_x^{\mathrm{T}}\mathbf{e}_{x'} = \cos\theta \tag{2.4}$$
$$\mathbf{e}_y^{\mathrm{T}}\mathbf{e}_{x'} = \sin\theta \tag{2.5}$$
$$\mathbf{e}_x^{\mathrm{T}}\mathbf{e}_{y'} = -\sin\theta \tag{2.6}$$
$$\mathbf{e}_y^{\mathrm{T}}\mathbf{e}_{y'} = \cos\theta \tag{2.7}$$

Thus, we can write our new basis vectors as

$$\mathbf{e}_{x'} = \cos\theta\mathbf{e}_x + \sin\theta\mathbf{e}_y \tag{2.8}$$
$$\mathbf{e}_{y'} = -\sin\theta\mathbf{e}_x + \cos\theta\mathbf{e}_y \tag{2.9}$$

(Check, from the geometry of Fig. 2.1, that this makes sense.)

Exercise 2.1 *Using the expressions for $\mathbf{e}_{x'}$ and $\mathbf{e}_{y'}$ in Eqs. 2.8-2.9, check that $\mathbf{e}_{x'}$ and $\mathbf{e}_{y'}$ are orthogonal to one another – that is, that $\mathbf{e}_{x'}^{\mathrm{T}}\mathbf{e}_{y'} = 0$ – and that they each have unit length – that is, that $\mathbf{e}_{x'}^{\mathrm{T}}\mathbf{e}_{x'} = \mathbf{e}_{y'}^{\mathrm{T}}\mathbf{e}_{y'} = 1$.*

Problem 2.1 *We've seen that, in a given coordinate system with orthonormal (mutually orthogonal and unit length) basis vectors $\mathbf{e}_0$, $\mathbf{e}_1$, any vector $\mathbf{v}$ has the representation $\mathbf{v} = \begin{pmatrix} \mathbf{e}_0^{\mathrm{T}}\mathbf{v} \\ \mathbf{e}_1^{\mathrm{T}}\mathbf{v} \end{pmatrix}$, which is just shorthand for $\mathbf{v} = (\mathbf{e}_0^{\mathrm{T}}\mathbf{v})\mathbf{e}_0 + (\mathbf{e}_1^{\mathrm{T}}\mathbf{v})\mathbf{e}_1$. Based on this and Eqs. 2.8-2.9, we know that,* in the x, y coordinate system, $\mathbf{e}_{x'} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$, $\mathbf{e}_{y'} = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}$, $\mathbf{e}_x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\mathbf{e}_y = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

Now, show that, in the x', y' coordinate system, $\mathbf{e}_{x'} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\mathbf{e}_{y'} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $\mathbf{e}_x = \begin{pmatrix} \cos\theta \\ -\sin\theta \end{pmatrix}$, $\mathbf{e}_y = \begin{pmatrix} \sin\theta \\ \cos\theta \end{pmatrix}$. *(Note, for each of these four vectors $\mathbf{v}$, you just have to form $\begin{pmatrix} \mathbf{e}_{x'}^{\mathrm{T}}\mathbf{v} \\ \mathbf{e}_{y'}^{\mathrm{T}}\mathbf{v} \end{pmatrix}$.) You can compute the necessary dot products using the representations in the x, y coordinate system, since dot products are coordinate-independent (although you can also just look them up from Eqs. 2.4-2.7). Note also that these equations should make intuitive sense: the x, y coordinate system is rotated by $-\theta$ from the x', y' system, so expressing $\mathbf{e}_x, \mathbf{e}_y$ in terms of $\mathbf{e}_{x'}, \mathbf{e}_{y'}$ should look exactly like expressing $\mathbf{e}_{x'}, \mathbf{e}_{y'}$ in terms of $\mathbf{e}_x, \mathbf{e}_y$, except that we must substitute $-\theta$ for θ; and note that* $\cos(-\theta) = \cos(\theta)$, $\sin(-\theta) = -\sin(\theta)$.*)*

We can reexpress the above equations for each set of basis vectors in the other's coordinate system in the coordinate-independent form:

$$\begin{aligned}
\mathbf{e}_{x'} &= \cos\theta\mathbf{e}_x + \sin\theta\mathbf{e}_y && (2.10)\\
\mathbf{e}_{y'} &= -\sin\theta\mathbf{e}_x + \cos\theta\mathbf{e}_y && (2.11)\\
\mathbf{e}_{x} &= \cos\theta\mathbf{e}_{x'} - \sin\theta\mathbf{e}_{y'} && (2.12)\\
\mathbf{e}_{y} &= \sin\theta\mathbf{e}_{x'} + \cos\theta\mathbf{e}_{y'} && (2.13)
\end{aligned}$$

Now, verify these equations in each *coordinate system. That is, first, using the x, y representation, substitute the coordinates of each vector and show that each equation is true. Then do the same thing again using the x', y' representation. The numbers change, but the equations, which are statements about geometry that are true in any coordinate system, remain true.*

OK, back to our original problem: we want to find the representation $\begin{pmatrix} v_{x'} \\ v_{y'} \end{pmatrix}$ of $\mathbf{v}$ in the new coordinate system. As we've seen, this is really just a short way of saying that $\mathbf{v} = v_{x'}\mathbf{e}_{x'} + v_{y'}\mathbf{e}_{y'}$ where $v_{x'} = \mathbf{e}_{x'}^{\mathrm{T}}\mathbf{v}$ and $v_{y'} = \mathbf{e}_{y'}^{\mathrm{T}}\mathbf{v}$. But we also know that $\mathbf{v} = v_x\mathbf{e}_x + v_y\mathbf{e}_y$. So, using Eqs. 2.4-2.7, we're ready to compute:

$$\begin{aligned}
v_{x'} &= \mathbf{e}_{x'}^{\mathrm{T}}\mathbf{v} = \mathbf{e}_{x'}^{\mathrm{T}}(v_x\mathbf{e}_x + v_y\mathbf{e}_y) = v_x(\mathbf{e}_{x'}^{\mathrm{T}}\mathbf{e}_x) + v_y(\mathbf{e}_{x'}^{\mathrm{T}}\mathbf{e}_y) = v_x\cos\theta + v_y\sin\theta && (2.14)\\
v_{y'} &= \mathbf{e}_{y'}^{\mathrm{T}}\mathbf{v} = \mathbf{e}_{y'}^{\mathrm{T}}(v_x\mathbf{e}_x + v_y\mathbf{e}_y) = v_x(\mathbf{e}_{y'}^{\mathrm{T}}\mathbf{e}_x) + v_y(\mathbf{e}_{y'}^{\mathrm{T}}\mathbf{e}_y) = -v_x\sin\theta + v_y\cos\theta && (2.15)
\end{aligned}$$

or in matrix form

$$\begin{pmatrix} v_{x'} \\ v_{y'} \end{pmatrix} = \begin{pmatrix} \mathbf{e}_{x'}^{\mathrm{T}}\mathbf{e}_x & \mathbf{e}_{x'}^{\mathrm{T}}\mathbf{e}_y \\ \mathbf{e}_{y'}^{\mathrm{T}}\mathbf{e}_x & \mathbf{e}_{y'}^{\mathrm{T}}\mathbf{e}_y \end{pmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix} \qquad (2.16)$$

Note that the first row of the matrix is just $\mathbf{e}_{x'}^{\mathrm{T}}$ as expressed in the $\mathbf{e}_x, \mathbf{e}_y$ coordinate system, and similarly the second row is just $\mathbf{e}_{y'}^{\mathrm{T}}$ as expressed in the $\mathbf{e}_x, \mathbf{e}_y$ coordinate system. This should make intuitive sense: to find $v_{x'}$, we want to find $\mathbf{e}_{x'}^{\mathrm{T}}\mathbf{v}$, which is obtained by applying the first row of the matrix to $\mathbf{v}$ as written in the $\mathbf{e}_x, \mathbf{e}_y$ coordinate system; and similarly $v_{y'}$ is found as $\mathbf{e}_{y'}^{\mathrm{T}}\mathbf{v}$, which is just the second row of the matrix applied to $\mathbf{v}$, all carried out in the $\mathbf{e}_x, \mathbf{e}_y$ coordinate system.

We can give a name to the above matrix: $\mathbf{R}_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$. This is a commonly encountered matrix known as a "rotation matrix". $\mathbf{R}_\theta$ represents rotation of coordinates by an angle θ: it is the matrix that transforms coordinates to a new set of coordinate axes rotated by θ from the previous coordinate axes.

Problem 2.2 *Verify the equation $\mathbf{v} = v_x\mathbf{e}_x + v_y\mathbf{e}_y$ in the x', y' coordinate system. That is, substitute the x', y' coordinate representation of $\mathbf{v}$ (from Eq. 2.16), $\mathbf{e}_x$, and $\mathbf{e}_y$, and verify that this equation is true. It's not quite as obvious as it was when it was expressed in the x, y coordinate system (Eq. 2.1), but it's still just as true.*

Problem 2.3 *Show that $\mathbf{R}_\theta^{\mathrm{T}}\mathbf{R}_\theta = \mathbf{R}_\theta\mathbf{R}_\theta^{\mathrm{T}} = \mathbf{1}$, that is, that $\mathbf{R}_\theta^{\mathrm{T}} = \mathbf{R}_\theta^{-1}$. (Note that this makes intuitive sense, because $\mathbf{R}_\theta^{\mathrm{T}} = \mathbf{R}_{-\theta}$; this follows from $\cos(-\theta) = \cos\theta$, $\sin(-\theta) = -\sin\theta$).*

To summarize, we've learned how a vector $\mathbf{v}$ transforms under a rigid change of basis, in which our coordinate axes are rotated counterclockwise by an angle θ. If $\mathbf{v}'$ is the representation of $\mathbf{v}$ in the new coordinate system, then $\mathbf{v}' = \mathbf{R}_\theta\mathbf{v}$. Furthermore, using the fact that $\mathbf{R}_\theta^{\mathrm{T}}\mathbf{R}_\theta = \mathbf{1}$, we can also find the inverse transform: $\mathbf{R}_\theta^{\mathrm{T}}\mathbf{v}' = \mathbf{R}_\theta^{\mathrm{T}}\mathbf{R}_\theta\mathbf{v} = \mathbf{v}$, *i.e.* $\mathbf{v} = \mathbf{R}_\theta^{\mathrm{T}}\mathbf{v}'$.

Now, we face a final question: how should matrices be transformed under this change of basis? For any matrix $\mathbf{M}$, let $\mathbf{M}'$ be its representation in the rotated coordinate system. To see how this should be transformed, note that $\mathbf{Mv}$ is a vector for any vector $\mathbf{v}$; so we know that $(\mathbf{Mv})' = \mathbf{R}_\theta\mathbf{Mv}$. But the transformation of the vector $\mathbf{Mv}$ should be the same as the vector we get from operating on the transformed vector $\mathbf{v}'$ with the transformed matrix $\mathbf{M}'$; that is, $(\mathbf{Mv})' = \mathbf{M}'\mathbf{v}'$. And we know $\mathbf{v}' = \mathbf{R}_\theta\mathbf{v}$. So, we find that $\mathbf{M}'\mathbf{R}_\theta\mathbf{v} = \mathbf{R}_\theta\mathbf{Mv}$, for every vector $\mathbf{v}$. But this can only true if $\mathbf{M}'\mathbf{R}_\theta$ and $\mathbf{R}_\theta\mathbf{M}$ are the same matrix[3]: $\mathbf{M}'\mathbf{R}_\theta = \mathbf{R}_\theta\mathbf{M}$. Finally, multiplying on the right by $\mathbf{R}_\theta^{\mathrm{T}}$, and using $\mathbf{R}_\theta\mathbf{R}_\theta^{\mathrm{T}} = \mathbf{1}$, we find

$$\mathbf{M}' = \mathbf{R}_\theta\mathbf{M}\mathbf{R}_\theta^{\mathrm{T}} \tag{2.17}$$

Intuitively, you can think of this as follows: to compute $\mathbf{M}'\mathbf{v}'$, which is just $\mathbf{Mv}$ in the new coordinate system, you first multiply $\mathbf{v}'$ by $\mathbf{R}_\theta^{\mathrm{T}}$, the inverse of $\mathbf{R}_\theta$. This takes $\mathbf{v}'$ back to $\mathbf{v}$, *i.e.* moves us back from the new coordinate system to the old coordinate system. You then apply $\mathbf{M}$ to $\mathbf{v}$ in the old coordinate system. Finally, you apply $\mathbf{R}_\theta$ to the result, to transform the result back into the new coordinate system.

2.3 Rigid Change of Basis in Arbitrary Dimensions

As our toy models should make clear, in neural modeling we are generally dealing with vectors of large dimensions. The above results in two dimensions generalize nicely to N dimensions. Suppose we want to consider only changes of basis consisting of rigid rotations. How shall we define these? We define these as the class of transformations $\mathbf{O}$ that *preserve all inner products*: that is, the transformations $\mathbf{O}$ such that, for any vectors $\mathbf{v}$ and $\mathbf{x}$, $\mathbf{v} \cdot \mathbf{x} = (\mathbf{Ov}) \cdot (\mathbf{Ox})$. Transformations satisfying this are called *orthogonal transformations*.

Why are these rigid? The dot product of two vectors of unit length gives the cosine of the angle between them, in any dimensions; and the dot product of a vector with itself tells you its length (squared). So, a dot-product-preserving transformation preserves the angles between all pairs of vectors and the lengths of all vectors. This coincides with what we mean by a rigid rotation — no stretching, no shrinking, no distortions.

We can rewrite the dot product, $(\mathbf{Ov}) \cdot (\mathbf{Ox}) = (\mathbf{Ov})^{\mathrm{T}}(\mathbf{Ox}) = \mathbf{v}^{\mathrm{T}}\mathbf{O}^{\mathrm{T}}\mathbf{Ox}$. The requirement that this be equal to $\mathbf{v}^{\mathrm{T}}\mathbf{x}$ for *any* vectors $\mathbf{v}$ and $\mathbf{x}$ can only be satisfied if $\mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{1}$.

Thus, we define:

Definition 2.1 *An* **orthogonal matrix** *is a matrix* $\mathbf{O}$ *satisfying* $\mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{O}\mathbf{O}^{\mathrm{T}} = \mathbf{1}$.

Note that the rotation matrix $\mathbf{R}_\theta$ in two dimensions is an example of an orthogonal matrix. Under an orthogonal transformation $\mathbf{O}$, a column vector is transformed $\mathbf{v} \mapsto \mathbf{Ov}$; a row vector is transformed $\mathbf{v}^{\mathrm{T}} \mapsto \mathbf{v}^{\mathrm{T}}\mathbf{O}^{\mathrm{T}}$ (as can be seen by considering $(\mathbf{v})^{\mathrm{T}} \mapsto (\mathbf{Ov})^{\mathrm{T}} = \mathbf{v}^{\mathrm{T}}\mathbf{O}^{\mathrm{T}}$); and a matrix is transformed $\mathbf{M} \mapsto \mathbf{OMO}^{\mathrm{T}}$.

The argument as to why $\mathbf{M}$ is mapped to $\mathbf{OMO}^{\mathrm{T}}$ is just as we worked out for two dimensions; the argument goes through unchanged for arbitrary dimensions. Here are two other ways to see it:

- The outer product $\mathbf{vx}^{\mathrm{T}}$ is a matrix. Under an orthogonal change of basis, $\mathbf{v} \mapsto \mathbf{Ov}$, $\mathbf{x} \mapsto \mathbf{Ox}$, so the outer product is mapped $\mathbf{vx}^{\mathrm{T}} \mapsto (\mathbf{Ov})(\mathbf{Ox})^{\mathrm{T}} = \mathbf{Ovx}^{\mathrm{T}}\mathbf{O}^{\mathrm{T}} = \mathbf{O}(\mathbf{vx}^{\mathrm{T}})\mathbf{O}^{\mathrm{T}}$. Thus, the matrix $\mathbf{vx}^{\mathrm{T}}$ transforms as indicated.

[3] Given that $\mathbf{Av} = \mathbf{Bv}$ for all vectors $\mathbf{v}$, suppose the i^{th} column of $\mathbf{A}$ is not identical to the i^{th} column of $\mathbf{B}$. Then choose $\mathbf{v}$ to be the vector that is all 0's except a 1 in the i^{th} position. Then $\mathbf{Av}$ is just the i^{th} column of $\mathbf{A}$, and similarly for $\mathbf{Bv}$, so $\mathbf{Av} \neq \mathbf{Bv}$ for this vector. Contradiction. Therefore every column of $\mathbf{A}$ and $\mathbf{B}$ must be identical, *i.e.* $\mathbf{A}$ and $\mathbf{B}$ must be identical.

- An expression of the form $\mathbf{v}^{\mathrm{T}}\mathbf{M}\mathbf{x}$ is a scalar, so it is unchanged by a coordinate transformation. In the new coordinates, this is $(\mathbf{O}\mathbf{v})^{\mathrm{T}}\tilde{\mathbf{M}}\mathbf{O}\mathbf{x}$, where $\tilde{\mathbf{M}}$ is the represention of $\mathbf{M}$ in the new coordinate system. Thus, $(\mathbf{O}\mathbf{v})^{\mathrm{T}}\tilde{\mathbf{M}}\mathbf{O}\mathbf{x} = \mathbf{v}^{\mathrm{T}}\mathbf{M}\mathbf{x}$, for any $\mathbf{v}$, $\mathbf{x}$, and $\mathbf{M}$ and orthogonal transform $\mathbf{O}$. We can rewrite $\mathbf{v}^{\mathrm{T}}\mathbf{M}x$ by inserting the identity, $\mathbf{1} = \mathbf{O}^{\mathrm{T}}\mathbf{O}$, as follows: $\mathbf{v}^{\mathrm{T}}\mathbf{M}x = \mathbf{v}^{\mathrm{T}}\mathbf{1}\mathbf{M}\mathbf{1}\mathbf{x} = \mathbf{v}^{\mathrm{T}}(\mathbf{O}^{\mathrm{T}}\mathbf{O})\mathbf{M}(\mathbf{O}^{\mathrm{T}}\mathbf{O})\mathbf{x} = (\mathbf{O}\mathbf{v})^{\mathrm{T}}(\mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}})\mathbf{O}\mathbf{x}$. The only way this can be equal to $(\mathbf{O}\mathbf{v})^{\mathrm{T}}\tilde{\mathbf{M}}\mathbf{O}\mathbf{x}$ for any $\mathbf{v}$ and $\mathbf{x}$ is if $\tilde{\mathbf{M}} = (\mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}})$.

Exercise 2.2 *Show that the property "$\mathbf{M}$ is the identity matrix" is basis-independent, that is, $\mathbf{O}\mathbf{1}\mathbf{O}^{\mathrm{T}} = \mathbf{1}$. Thus, the identity matrix looks the same in any basis.*

Exercise 2.3 *Note that the property "$\mathbf{x}$ is the zero vector" ($\mathbf{x} = 0$; $\mathbf{x}$ is the vector all of whose elements are zero) is basis-independent; that is, if $\mathbf{x} = 0$, then $\mathbf{O}\mathbf{x} = 0$ for any $\mathbf{O}$. Similarly, "$\mathbf{M}$ is the zero matrix" ($\mathbf{M} = 0$; $\mathbf{M}$ is the matrix all of whose elements are zero) is basis-independent: if $\mathbf{M} = 0$, then $\mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}} = 0$ for any $\mathbf{O}$.*

Problem 2.4 *1. Show that the property "$\mathbf{P}$ is the inverse of $\mathbf{M}$" is basis-independent. That is, if $\mathbf{P} = \mathbf{M}^{-1}$, then $\mathbf{O}\mathbf{P}\mathbf{O}^{\mathrm{T}} = (\mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}})^{-1}$, where $\mathbf{O}$ is orthogonal. (Hint: to show that $\mathbf{A} = \mathbf{B}^{-1}$, just show that $\mathbf{A}\mathbf{B} = \mathbf{1}$.)*

2. Note, from problem 1.2, that $(\mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{O}\mathbf{M}^{\mathrm{T}}\mathbf{O}^{\mathrm{T}}$. Use this result to prove two immediate corollaries:

- *The property "$\mathbf{P}$ is the transpose of $\mathbf{M}$" is invariant under orthogonal changes of basis: that is, $\mathbf{O}\mathbf{P}\mathbf{O}^{\mathrm{T}} = (\mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}})^{\mathrm{T}}$ for $\mathbf{P} = \mathbf{M}^{\mathrm{T}}$.*
- *The property "$\mathbf{M}$ is symmetric" is invariant under orthogonal changes of basis: that is, if $\mathbf{M} = \mathbf{M}^{\mathrm{T}}$, $(\mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}}$.*

Problem 2.5 *Write down arguments to show that (1) a dot-product preserving transformation is one for which $\mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{1}$; and (2) under this transformation, $\mathbf{M} \mapsto \mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}}$* — without looking at these notes. *You can look at these notes as much as you want in preliminary tries, but the last try you have to go from beginning to end without looking at the notes.*

2.4 Complete Orthonormal Bases

Consider the standard basis vectors in N dimensions: $\mathbf{e}_0 = (1, 0, \ldots, 0)^{\mathrm{T}}$, $\mathbf{e}_1 = (0, 1, \ldots, 0)^{\mathrm{T}}$, ..., $\mathbf{e}_{N-1} = (0, 0, \ldots, 1)^{\mathrm{T}}$. These form an *orthonormal basis*. This means: (1) The $\mathbf{e}_i$ are mutually *orthogonal*: $\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = 0$ for $i \neq j$; and (2) the $\mathbf{e}_i$ are each *normalized* to length 1: $\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_i = 1$, $i = 0, \ldots, N-1$. We can summarize and generalize this by use of the Kronecker delta:

Definition 2.2 *The* **Kronecker delta** δ_{ij} *is defined by* $\delta_{ij} = 1$, $i = j$; $\delta_{ij} = 0$, $i \neq j$.

Note that δ_{ij} describes the elements of the identity matrix: $(\mathbf{1})_{ij} = \delta_{ij}$.

Problem 2.6 *Show that, for any vector $\mathbf{x}$, $\sum_j \delta_{ij}x_j = x_i$. This ability of the Kronecker delta to "collapse" a sum to a single term is something that will be used over and over again. (Note that this equation is just the equation $\mathbf{1}\mathbf{x} = \mathbf{x}$, in component form.)*

Definition 2.3 *A set of N vectors $\mathbf{e}_i$, $i = 0, \ldots, N-1$, form an* **orthonormal basis** *for an N-dimensional vector space if $\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = \delta_{ij}$.*

Exercise 2.4 *Show that in two dimensions, the vectors* $\mathbf{e}_0 = \mathbf{R}_{(\theta)}(1,0)^{\mathrm{T}} = (\cos\theta, -\sin\theta)^{\mathrm{T}}$*, and* $\mathbf{e}_1 = \mathbf{R}_{(\theta)}(0,1)^{\mathrm{T}} = (\sin\theta, \cos\theta)^{\mathrm{T}}$*, form an orthonormal basis, for any angle* θ*.*

Exercise 2.5 *Prove that an orthonormal basis remains an orthonormal basis after transformation by an orthogonal matrix. Your proof is likely to consist of writing down one sentence about what orthogonal transforms preserve.*

Let's restate more generally what we learned in two dimensions: when we state that $\mathbf{v} = (v_0, v_1, \ldots, v_{N-1})^{\mathrm{T}}$ in some orthonormal basis $\mathbf{e}_i$, we mean that $\mathbf{v}$ has extent v_0 in the $\mathbf{e}_0$ direction, etc. We can state this more formally by writing

$$\mathbf{v} = v_0\mathbf{e}_0 + \ldots + v_{N-1}\mathbf{e}_{N-1} = \sum_i v_i\mathbf{e}_i \tag{2.18}$$

This is an expansion of the vector $\mathbf{v}$ in the $\mathbf{e}_i$ basis: an expression of $\mathbf{v}$ as a weighted sum of the $\mathbf{e}_i$. This is, in essence, what it means for the $\mathbf{e}_i$ to be a basis: any vector $\mathbf{v}$ can be written as a weighted sum of the $\mathbf{e}_i$. The coefficients of the expansion, v_i, are the components of $\mathbf{v}$ in the basis of the $\mathbf{e}_i$; we summarize all of this when we state that $\mathbf{v} = (v_0, v_1, \ldots, v_{N-1})^{\mathrm{T}}$ in the $\mathbf{e}_i$ basis. The coefficients v_i are given by the dot product of $\mathbf{v}$ and $\mathbf{e}_i$: $v_i = \mathbf{e}_i^{\mathrm{T}}\mathbf{v}$:

Problem 2.7 *Show that* $v_j = \mathbf{e}_j^{\mathrm{T}}\mathbf{v}$*. (Hint: multiply Eq. 2.18 from the left by* $\mathbf{e}_j^{\mathrm{T}}$*, and use the result of Problem 2.6.)*

In particular, we can expand the basis vectors in themselves:

$$\mathbf{e}_i = (\mathbf{e}_0^{\mathrm{T}}\mathbf{e}_i)\mathbf{e}_0 + \ldots + (\mathbf{e}_{N-1}^{\mathrm{T}}\mathbf{e}_i)\mathbf{e}_{N-1} = \sum_j (\mathbf{e}_j^{\mathrm{T}}\mathbf{e}_i)\mathbf{e}_j = \sum_j \delta_{ij}\mathbf{e}_j = \mathbf{e}_i. \tag{2.19}$$

That is, the basis vectors, when expressed in their own basis, are always just written $\mathbf{e}_0 = (1, 0, \ldots, 0)^{\mathrm{T}}$, $\mathbf{e}_1 = (0, 1, \ldots, 0)^{\mathrm{T}}$, …, $\mathbf{e}_{N-1} = (0, 0, \ldots, 1)^{\mathrm{T}}$. Thus, the equation $\mathbf{v} = \sum_i v_i\mathbf{e}_i$ (Eq. 2.18), when written in the $\mathbf{e}_i$ basis, just represents the intuitive statement

$$\mathbf{v} = \begin{pmatrix} v_0 \\ v_1 \\ \ldots \\ v_{N-1} \end{pmatrix} = v_0\begin{pmatrix} 1 \\ 0 \\ \ldots \\ 0 \end{pmatrix} + v_1\begin{pmatrix} 0 \\ 1 \\ \ldots \\ 0 \end{pmatrix} + \ldots + v_{N-1}\begin{pmatrix} 0 \\ 0 \\ \ldots \\ 1 \end{pmatrix} = \sum_i v_i\mathbf{e}_i \tag{2.20}$$

In summary, for any vector $\mathbf{v}$ and orthonormal basis $\mathbf{e}_i$, we can write

$$\mathbf{v} = \sum_i \mathbf{e}_i(\mathbf{e}_i^{\mathrm{T}}\mathbf{v}) = \sum_i v_i\mathbf{e}_i \tag{2.21}$$

In particular, any orthonormal basis vectors $\mathbf{e}_i$, when expressed in their own basis, have the simple representation $\mathbf{e}_0 = (1, 0, \ldots, 0)^{\mathrm{T}}$, $\mathbf{e}_1 = (0, 1, \ldots, 0)^{\mathrm{T}}$, …, $\mathbf{e}_{N-1} = (0, 0, \ldots, 1)^{\mathrm{T}}$.

We can rewrite $\mathbf{v} = \sum_i \mathbf{e}_i(\mathbf{e}_i^{\mathrm{T}}\mathbf{v})$ as $\mathbf{v} = \sum_i(\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}})\mathbf{v} = (\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}})\mathbf{v}$. Since this is true for any vector $\mathbf{v}$, this means that $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$, the identity matrix. This is true for *any* orthonormal basis.

Problem 2.8 *For any orthonormal basis* $\mathbf{e}_i$*,* $i = 0, \ldots, N-1$*: Show that* $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$*, by working in the* $\mathbf{e}_i$ *basis, as follows. In that basis, show that* $\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}$ *is the matrix composed of all 0's, except for a 1 on the diagonal in the* i^{th} *row/column. Do the summation to show that* $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$*.*

Exercise 2.6 *Make sure you understand the following. Although you have derived* $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$ *in Problem 2.8 by working in a particular basis, the result is general: it is true no matter in which orthonormal basis you express the* $\mathbf{e}_i$*. This follows immediately from exercise 2.2. Or, you can see this explicitly, for example, by transforming the equation to another orthonormal basis by applying an orthogonal matrix* $\mathbf{O}$ *on the left and* $\mathbf{O}^{\mathrm{T}}$ *on the right. This gives* $\sum_i \mathbf{O}\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}\mathbf{O}^{\mathrm{T}} = \mathbf{O}\mathbf{1}\mathbf{O}^{\mathrm{T}}$*, which becomes* $\sum_i(\mathbf{O}\mathbf{e}_i)(\mathbf{O}\mathbf{e}_i)^{\mathrm{T}} = \mathbf{1}$*. Thus, the equation holds for the* $\mathbf{e}_i$ *as expressed in the new coordinate system.*

We can restate the fact that $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$ in words to, hopefully, make things more intuitive, as follows. The matrix $\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}$, when applied to the vector $\mathbf{v}$, finds the component of $\mathbf{v}$ along the $\mathbf{e}_i$ direction, and multiplies this by the vector $\mathbf{e}_i$: $(\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}})\mathbf{v} = \mathbf{e}_i(\mathbf{e}_i^{\mathrm{T}}\mathbf{v}) = v_i\mathbf{e}_i$. That is, $\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}$ finds the *projection* of $\mathbf{v}$ along the $\mathbf{e}_i$ axis. When the $\mathbf{e}_i$ form an orthonormal basis, these separate projections are independent: any $\mathbf{v}$ is just the sum of its projections onto each of the $\mathbf{e}_i$: $\mathbf{v} = \sum_i \mathbf{e}_i(\mathbf{e}_i^{\mathrm{T}}\mathbf{v})$. Taking the projections of $\mathbf{v}$ onto each axis of a complete orthonormal basis, and adding up the results, just reconstitutes the vector $\mathbf{v}$. (For example, Fig. 2.1 illustrates that in two dimensions, adding the vectors $v_x\mathbf{e}_x$ and $v_y\mathbf{e}_y$, the projections of $\mathbf{v}$ on the x and y axes, reconstitutes $\mathbf{v}$.) That is, the operation of taking the projections of $\mathbf{v}$ on each axis, and then summing the projections, is just the identity operation; so $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$.

The property $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$ represents a pithy summation of the fact that an orthonormal basis is *complete*:

Definition 2.4 *A* **complete basis** *for a vector space is a set of vectors* $\mathbf{e}_i$ *such that any vector* $\mathbf{v}$ *can be uniquely expanded as a weighted sum of the* $\mathbf{e}_i$*:* $\mathbf{v} = \sum_i v_i\mathbf{e}_i$*, where there is only one set of* v_i *for a given* $\mathbf{v}$ *that will satisfy this equation.*

Fact 2.1 *An orthonormal set of vectors* $\mathbf{e}_i$ *forms a complete basis if and only if* $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$*.*

Intuitively: if we have an incomplete basis – we are missing some directions – then $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}$ will give 0 when applied to vectors representing the missing directions, so it can't be the identity; saying $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$ means that it reconstitutes *any* vector, so there are no missing directions.

More formally, we can prove this as follows: if $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$, then for any vector $\mathbf{v}$, $\mathbf{v} = \mathbf{1}\mathbf{v} = \sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}\mathbf{v} = \sum_i v_i\mathbf{e}_i$ where $v_i = \mathbf{e}_i^{\mathrm{T}}\mathbf{v}$. So any vector $\mathbf{v}$ can be represented as a linear combination of the $\mathbf{e}_i$, so they form a complete basis. Conversely, if the $\mathbf{e}_i$ form a complete basis, then for any vector $\mathbf{v}$, $\mathbf{v} = \sum_i v_i\mathbf{e}_i$ for some v_i. By the orthonormality of the $\mathbf{e}_i$, taking the dot product with $\mathbf{e}_j$ gives $\mathbf{e}_j \cdot \mathbf{v} = \sum_i v_i\mathbf{e}_j \cdot \mathbf{e}_i = \sum_i v_i\delta_{ji} = v_j$. So for any $\mathbf{v}$, $\mathbf{v} = \sum_i \mathbf{e}_i v_i = \sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}\mathbf{v} = \left(\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}\right)\mathbf{v}$. This can only be true for every vector $\mathbf{v}$ if $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$.

Fact 2.2 *In an N-dimensional vector space, a set of orthonormal vectors forms a complete basis if and only if the set contains N vectors.*

That is, any set of N orthonormal vectors constitutes a complete basis; you can't have more than N mutually orthonormal vectors in an N-dimensional space; and if you only have N-1 (or fewer) orthonormal vectors, you're missing a direction and so can't represent vectors pointing in that direction or that have a component in that direction.

Finally, we've interpreted the components of a vector, $\mathbf{v} = (v_0, v_1, \ldots, v_{N-1})^{\mathrm{T}}$, as describing $\mathbf{v}$ only in some particular basis; the more general statements, given some underlying basis vectors $\mathbf{e}_i$, are $\mathbf{v} = \sum_i v_i\mathbf{e}_i$, where $v_i = \mathbf{e}_i^{\mathrm{T}}\mathbf{v}$. We now do the same for a matrix. We write $\mathbf{M} = \mathbf{1}\mathbf{M}\mathbf{1} =$

$(\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}})\mathbf{M}(\sum_j \mathbf{e}_j\mathbf{e}_j^{\mathrm{T}}) = \sum_{i,j} \mathbf{e}_i(\mathbf{e}_i^{\mathrm{T}}\mathbf{M}\mathbf{e}_j)\mathbf{e}_j^{\mathrm{T}}$. But $(\mathbf{e}_i^{\mathrm{T}}\mathbf{M}\mathbf{e}_j)$ is a scalar; call it M_{ij}. Since a scalar commutes with anything, we can pull this out front; thus, we have obtained

$$\mathbf{M} = \sum_{ij} M_{ij}\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}} \qquad \text{where } M_{ij} = \mathbf{e}_i^{\mathrm{T}}\mathbf{M}\mathbf{e}_j \tag{2.22}$$

When working in the basis of the $\mathbf{e}_i$ vectors, $\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}}$ is the matrix that is all 0's except for a 1 in the i^{th} row, j^{th} column (verify this!). Thus, in the basis of the $\mathbf{e}_i$ vectors,

$$\mathbf{M} = \begin{pmatrix} M_{00} & M_{01} & \dots & M_{0(N-1)} \\ M_{10} & M_{11} & \dots & M_{1(N-1)} \\ \dots & \dots & \dots & \dots \\ M_{(N-1)0} & M_{(N-1)1} & \dots & M_{(N-1)(N-1)} \end{pmatrix} \tag{2.23}$$

Thus, the $M_{ij} = \mathbf{e}_i^{\mathrm{T}}\mathbf{M}\mathbf{e}_j$ are the elements of $\mathbf{M}$ in the $\mathbf{e}_i$ basis, just as $v_i = \mathbf{e}_i^{\mathrm{T}}\mathbf{v}$ are the elements of $\mathbf{v}$ in the $\mathbf{e}_i$ basis. The more general description of $\mathbf{M}$ is given by Eq. 2.22.

2.5 Which Basis Does an Orthogonal Matrix Map To?

Suppose we change basis by some orthogonal matrix $\mathbf{O}$: $\mathbf{v} \mapsto \mathbf{O}\mathbf{v}$, $\mathbf{M} \mapsto \mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}}$. What basis are we mapping to? The answer is: in our current basis, $\mathbf{O}$ is the matrix each of whose rows is one of the new basis vectors, as expressed in our current basis. This should be intuitive: applying the first row of $\mathbf{O}$ to a vector $\mathbf{v}$, we should get the coordinate of $\mathbf{v}$ along the first new basis vector $\mathbf{e}_0$; but this coordinate is $\mathbf{e}_0^{\mathrm{T}}\mathbf{v}$, hence the first row should be $\mathbf{e}_0^{\mathrm{T}}$. We can write this as $\mathbf{O} = (\ \mathbf{e}_0\ \mathbf{e}_1\ \dots\ \mathbf{e}_{N-1}\ )^{\mathrm{T}}$, where $\mathbf{e}_0$ means a column of our matrix corresponding to the new basis vector $\mathbf{e}_0$ as expressed in our current basis. To be precise, we mean the following: letting $(\mathbf{O})_{ij}$ be the $(ij)^{\text{th}}$ component of the matrix $\mathbf{O}$, and letting $(\mathbf{e}_i)_j$ be the j^{th} component of new basis vector $\mathbf{e}_i$ (all of these components expressed in our current basis), then $(\mathbf{O})_{ij} = (\mathbf{e}_i)_j$. It of course follows that each column of $\mathbf{O}^{\mathrm{T}}$ is one of the new basis vectors, that is, $\mathbf{O}^{\mathrm{T}} = (\ \mathbf{e}_0\ \mathbf{e}_1\ \dots\ \mathbf{e}_{N-1}\ )$.

Problem 2.9 *Use the results of problem 1.3, or rederive from scratch, to show the following:*

1. *Show that the statement* $\mathbf{O}\mathbf{O}^{\mathrm{T}} = \mathbf{1}$ *simply states the orthonormality of the new basis vectors:* $\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = \delta_{ij}$.

2. *Similarly, show that the statement* $\mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{1}$ *simply expresses the completeness of the new basis vectors:* $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$.

2.6 Recapitulation: The Transformation From One Orthogonal Basis To Another

We have seen that, for any orthonormal basis $\{\mathbf{e}_i\}$, any vector $\mathbf{v}$ can be expressed $\mathbf{v} = \sum_i v_i\mathbf{e}_i$ where $\mathbf{v}_i = \mathbf{e}_i^{\mathrm{T}}\mathbf{v}$, and any matrix $\mathbf{M}$ can be expressed $\mathbf{M} = \sum_{ij} M_{ij}\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}}$ where $M_{ij} = \mathbf{e}_i^{\mathrm{T}}\mathbf{M}\mathbf{e}_j$. Consider another orthonormal basis $\{\mathbf{f}_i\}$. Using $\mathbf{1} = \sum_k \mathbf{f}_k\mathbf{f}_k^{\mathrm{T}}$, we can derive the rules for transforming coordinates from the $\{\mathbf{e}_i\}$ basis to the $\{\mathbf{f}_i\}$ basis, and in so doing recapitulate the results of this chapter, as follows:

- Transformation of a vector: write $\mathbf{v} = \sum_i v_i\mathbf{e}_i = \sum_i v_i\mathbf{1}\mathbf{e}_i = \sum_{ik} v_i\mathbf{f}_k\mathbf{f}_k^{\mathrm{T}}\mathbf{e}_i = \sum_k v'_k\mathbf{f}_k$, where $v'_k = \sum_i \mathbf{f}_k^{\mathrm{T}}\mathbf{e}_i v_i = \sum_i O_{ki}v_i$, and the matrix $\mathbf{O}$ is defined by $O_{ki} = \mathbf{f}_k^{\mathrm{T}}\mathbf{e}_i$. That is, the coordinates v'_i of $\mathbf{v}$ in the $\{\mathbf{f}_k\}$ coordinate system are given, in terms of the coordinates v_i in the $\{\mathbf{e}_i\}$ coordinate system, by $\mathbf{v}' = \mathbf{O}\mathbf{v}$.

Note that $\mathbf{O}$ is indeed orthogonal: $(\mathbf{O}\mathbf{O}^{\mathrm{T}})_{ij} = \sum_k O_{ik}O_{jk} = \sum_k \mathbf{f}_i^{\mathrm{T}}\mathbf{e}_k\mathbf{f}_j^{\mathrm{T}}\mathbf{e}_k = \sum_k \mathbf{f}_i^{\mathrm{T}}\mathbf{e}_k\mathbf{e}_k^{\mathrm{T}}\mathbf{f}_j = \mathbf{f}_i^{\mathrm{T}}(\sum_k \mathbf{e}_k\mathbf{e}_k^{\mathrm{T}})\mathbf{f}_j = \mathbf{f}_i^{\mathrm{T}}\mathbf{f}_j = \delta_{ij}$; while $(\mathbf{O}^{\mathrm{T}}\mathbf{O})_{ij} = \sum_k O_{ki}O_{kj} = \sum_k \mathbf{f}_k^{\mathrm{T}}\mathbf{e}_i\mathbf{f}_k^{\mathrm{T}}\mathbf{e}_j = \sum_k \mathbf{e}_i^{\mathrm{T}}\mathbf{f}_k\mathbf{f}_k^{\mathrm{T}}\mathbf{e}_j = \mathbf{e}_i^{\mathrm{T}}(\sum_k \mathbf{f}_k\mathbf{f}_k^{\mathrm{T}})\mathbf{e}_j = \mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = \delta_{ij}$.

Note also that the i^{th} row of $\mathbf{O}$ has elements $O_{ij} = \mathbf{f}_i^{\mathrm{T}}\mathbf{e}_j$, with j varying across the row; these are just the coordinates of $\mathbf{f}_i$ in the $\{\mathbf{e}_j\}$ basis – that is, the i^{th} row of $\mathbf{O}$ is precisely the vector $\mathbf{f}_i^{\mathrm{T}}$ as expressed in the $\{\mathbf{e}_j\}$ basis. The i^{th} column of $\mathbf{O}$ has elements $O_{ji} = \mathbf{f}_j^{\mathrm{T}}\mathbf{e}_i$, with j varying across the column – these are the coordinates of $\mathbf{e}_i$ in the $\mathbf{f}_j$ basis. So the i^{th} column is just the i^{th} old basis vector, written in the coordinates of the new basis, while the i^{th} row is the i^{th} new basis vector, written in the coordinates of the old basis. Thus, when we take the transpose of $\mathbf{O}$, the roles of the two basis sets are reversed; so $\mathbf{O}^{\mathrm{T}}$ is the mapping from the $\{f_i\}$ basis to the $\{\mathbf{e}_i\}$ basis, and thus is the inverse of $\mathbf{O}$.

- Transformation of a matrix: write $\mathbf{M} = \sum_{ij} M_{ij}\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}} = \sum_{ij} M_{ij}\mathbf{1}\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}}\mathbf{1} = \sum_{ijkl} M_{ij}\mathbf{f}_k\mathbf{f}_k^{\mathrm{T}}\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}}\mathbf{f}_l\mathbf{f}_l^{\mathrm{T}} = \sum_{kl} M'_{kl}\mathbf{f}_k\mathbf{f}_l^{\mathrm{T}}$ where $M'_{kl} = \sum_{ij} f_k^{\mathrm{T}}\mathbf{e}_i M_{ij}\mathbf{e}_j^{\mathrm{T}}\mathbf{f}_l = \sum_{ij} O_{ki}M_{ij}O_{lj} = \sum_{ij} O_{ki}M_{ij}O^{\mathrm{T}}_{jl}$, where again the matrix $\mathbf{O}$ is defined by $O_{ij} = \mathbf{f}_i^{\mathrm{T}}\mathbf{e}_j$. That is, the coordinates M'_{ij} of $\mathbf{M}$ in the $\{\mathbf{f}_i\}$ coordinate system are given, in terms of the coordinates M_{ij} in the $\{\mathbf{e}_i\}$ coordinate system, by $\mathbf{M}' = \mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}}$.

2.7 Summary

Vectors and matrices are geometrical objects. The vector $\mathbf{v}$ has some length and points in some direction in the world, independent of any basis. Similarly, a given matrix represents the same transformation – for example, the one that takes $\mathbf{e}_{x'}$ to $\mathbf{e}_x$ and $\mathbf{e}_{y'}$ to $\mathbf{e}_y$ in Fig. 2.1 – in any basis.

To talk about vectors and matrices, we generally define some complete orthonormal basis. This is a set of N vectors $\mathbf{e}_i$, where N is the dimension of the space, that satisfy $\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = \delta_{ij}$. The fact that the basis is complete means that any vector can be written as a weighted sum of these basis vectors: $\mathbf{v} = \sum_i v_i\mathbf{e}_i$ where $\mathbf{v}_i = \mathbf{e}_i^{\mathrm{T}}\mathbf{v}$. This completeness is summarized by the fact that $\sum_i \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{1}$, where $\mathbf{1}$ is the identity matrix.

The choice of basis is, in principal, arbitrary. Transformations between orthonormal bases are given by *orthogonal* transformations, determined by matrices $\mathbf{O}$ satisfying $\mathbf{O}\mathbf{O}^{\mathrm{T}} = \mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{1}$. Vectors transform as $\mathbf{v} \mapsto \mathbf{O}\mathbf{v}$, while matrices transform as $\mathbf{M} \mapsto \mathbf{O}\mathbf{M}\mathbf{O}^{\mathrm{T}}$. The rows of $\mathbf{O}$ are the new basis vectors, as written in the coordinate system of the current basis vectors.

The interesting properties of vectors and matrices are those that are geometric, that is, independent of basis. Any scalars formed from vector and matrix operations are invariant under orthogonal changes of basis, for example the dot product $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ of any two vectors $\mathbf{x}$ and $\mathbf{y}$, or the quantity $\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{y}$ for any two vectors $\mathbf{x}$ and $\mathbf{y}$ and matrix $\mathbf{M}$ (note that the latter is just a dot product of two vectors, $\mathbf{x}^{\mathrm{T}}(\mathbf{M}\mathbf{y})$. From this follows the orthogonal-basis-independence of such geometric quantities as the length $|v|$ of a vector $\mathbf{v}$ ($|v| = \sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}$) or the angle θ between two vectors $\mathbf{x}$ and $\mathbf{y}$ (which is the inverse cosine of $\mathbf{x}^{\mathrm{T}}\mathbf{y}/|x||y| = \cos\theta$). Similarly, equalities between vectors or matrices are basis-independent: *e.g.*, if $\mathbf{M}\mathbf{v} = \mathbf{y}$ in one basis, the same is true in any basis. Thus, a matrix $\mathbf{M}$ represents the same transformation in any basis – it takes the same vectors $\mathbf{v}$ to the same vectors $\mathbf{y}$. A number of other properties are also preserved under orthogonal transformations, such as whether or not a set of vectors is orthonormal (this follows from the preservations of length and angles), whether or not a matrix is symmetric, and whether or not a matrix is orthogonal.

In the next section, we will see that both the familiarity we have gained with vectors and matrices, and in particular the freedom we have developed to switch between bases, will help us to greatly simplify and solve linear differential equations, such as those that arise in studying simple models of neural activity and synaptic development.

3 Linear Differential Equations, Eigenvectors and Eigenvalues

The formulation of our toy models led us to linear differential equations of the form $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v} + \mathbf{h}$. Here $\mathbf{v}$ is the vector whose time evolution we are studying, like the vector of weights in our model of synaptic development, or the vector of neural activities in our model of activity in a circuit; $\mathbf{M}$ is a matrix; and $\mathbf{h}$ is a constant vector. An equation of this form is called an inhomogeneous equation; an equation of the form $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$ is a homogeneous equation. We will focus on the homogeneous equation, because once we understand how to solve this, solving the inhomogeneous case is easy. At the end of this section, we'll return to the inhomogeneous case and show how it's solved. Solving $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$ is easy if we can find the *eigenvectors* and *eigenvalues* of the matrix $\mathbf{M}$, so much of this section will be devoted to understanding what these are. But we'll begin, once again, by thinking about the problem in one or two dimensions.

3.1 Linear Differential Equations in Vector Form

The solution to the simple linear differential equation,

$$\frac{d}{dt}v = kv \tag{3.1}$$

is

$$v(t) = v(0)e^{kt} \tag{3.2}$$

where $v(0)$ is the value of v at $t = 0$.

Exercise 3.1 *If this is not obvious to you, show that Eq. 3.2 is indeed a solution to Eq. 3.1.*

Now, consider two independent equations:

$$\frac{d}{dt}v_0 = k_0 v_0 \tag{3.3}$$

$$\frac{d}{dt}v_1 = k_1 v_1 \tag{3.4}$$

We can rewrite these as the matrix equation

$$\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v} \tag{3.5}$$

where $\mathbf{v} = \begin{pmatrix} v_0 \\ v_1 \end{pmatrix}$ and $\mathbf{M} = \begin{pmatrix} k_0 & 0 \\ 0 & k_1 \end{pmatrix}$. That is:

$$\frac{d}{dt}\begin{pmatrix} v_0 \\ v_1 \end{pmatrix} = \begin{pmatrix} k_0 & 0 \\ 0 & k_1 \end{pmatrix}\begin{pmatrix} v_0 \\ v_1 \end{pmatrix}. \tag{3.6}$$

Exercise 3.2 *Satisfy yourself that Eq. 3.6 is, component-wise, identical to Eqs. 3.3-3.4.*

Of course, Eq. 3.6 has the solution

$$v_0(t) = v_0(0)e^{k_0 t} \tag{3.7}$$

$$v_1(t) = v_1(0)e^{k_1 t} \tag{3.8}$$

Congratulations!! You've just solved your first matrix differential equation. Pretty easy, eh?

Moral 3.1 *When a matrix* $\mathbf{M}$ *is* diagonal *(that is, has nonzero entries only along the diagonal), the equation* $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$ *is trivial — it is just a set of independent, one-dimensional equations.*

Exercise 3.3 *Let's clarify the meaning of Eqs. 3.5-3.6. What does it mean to take the time derivative of a vector,* $\frac{d}{dt}\mathbf{v}$*? First, it means that* $\mathbf{v}$ *is a vector function of time,* $\mathbf{v}(t)$ *(but we generally won't explicitly write the '(t)'); that is,* $\mathbf{v}$ *represents a different vector at each time t. The equation* $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$ *tells the vector change in* $\mathbf{v}(t)$ *in a short time* Δt*:* $\mathbf{v}(t+\Delta t) = \mathbf{v}(t) + \mathbf{M}\mathbf{v}(t)\Delta t$*. Now, expand* $\mathbf{v} = \sum_i v_i \mathbf{e}_i$ *in some basis* $\mathbf{e}_i$*. Note that the* $\mathbf{e}_i$ *are fixed, time-invariant vectors:* $\frac{d}{dt}\mathbf{e}_i = 0$*. The time-dependence of* $\mathbf{v}$ *is reflected in the time-dependence of the* v_i*. Thus, we can write* $\frac{d}{dt}\mathbf{v} = \frac{d}{dt}\left(\sum_i v_i \mathbf{e}_i\right) = \sum_i \mathbf{e}_i \frac{d}{dt} v_i$*. In the* $\mathbf{e}_i$ *basis,* $\sum_i \mathbf{e}_i \frac{d}{dt} v_i = \begin{pmatrix} \frac{d}{dt}v_0 \\ \frac{d}{dt}v_1 \end{pmatrix} = \frac{d}{dt}\begin{pmatrix} v_0 \\ v_1 \end{pmatrix}$*.*

Now, suppose you've been given the set of two independent equations in Eqs. 3.3–3.6; but you've been given them in *the wrong coordinate system.* This could happen if somebody didn't know anything about v_0 and v_1, and instead measured things in some different coordinates that seemed natural from the viewpoint of an experiment. We're going to find that that's exactly the case in our toy models: in the coordinates in which we're given the problem – the weights, or the activities – the relevant matrix is not diagonal; but there *is* a coordinate system in which the matrix is diagonal. So, let's say the coordinates that were measured turn out to be $w_0 = (v_0 + v_1)/\sqrt{2}$, $w_1 = (-v_0 + v_1)/\sqrt{2}$. (The factors of $\sqrt{2}$ are included to make this an orthogonal – length-preserving – change of coordinates.) We can express this as a matrix equation:

$$\begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} v_0 \\ v_1 \end{pmatrix}. \tag{3.9}$$

We could also find this transformation matrix by thinking geometrically about the change of basis involved in going from $\mathbf{v}$ to $\mathbf{w}$. It's not hard to see (draw a picture of $\mathbf{v}$ and $\mathbf{w}$! – for example, set $\mathbf{v}$ along the x axis, and work in x, y coordinates) that this represents a rotation of coordinates by $45° = \pi/4$. That is, our transformation matrix is

$$\mathbf{R}_{\pi/4} = \begin{pmatrix} \cos\pi/4 & \sin\pi/4 \\ -\sin\pi/4 & \cos\pi/4 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \tag{3.10}$$

Thus, the equation $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$ will be transformed into $\frac{d}{dt}(\mathbf{R}_{\pi/4}\mathbf{v}) = (\mathbf{R}_{\pi/4}\mathbf{M}\mathbf{R}^{\mathrm{T}}_{\pi/4})(\mathbf{R}_{\pi/4}\mathbf{v})$, or

$$\frac{d}{dt}\mathbf{w} = \tilde{\mathbf{M}}\mathbf{w} \tag{3.11}$$

where

$$\mathbf{w} = \mathbf{R}_{\pi/4}\mathbf{v}, \qquad \tilde{\mathbf{M}} = \mathbf{R}_{\pi/4}\mathbf{M}\mathbf{R}^{\mathrm{T}}_{\pi/4} = \frac{1}{2}\begin{pmatrix} k_1 + k_0 & k_1 - k_0 \\ k_1 - k_0 & k_1 + k_0 \end{pmatrix} \tag{3.12}$$

In components, this is

$$\frac{d}{dt}\begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} k_1 + k_0 & k_1 - k_0 \\ k_1 - k_0 & k_1 + k_0 \end{pmatrix}\begin{pmatrix} w_0 \\ w_1 \end{pmatrix} \tag{3.13}$$

Problem 3.1
- *Show that the elements of* $\mathbf{R}_{\pi/4}\mathbf{M}\mathbf{R}^{\mathrm{T}}_{\pi/4}$ *are as given in Eq. 3.12.*
- *Show that the equation* $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$*, after multiplying both sides from the left by* $\mathbf{R}_{\pi/4}$*, transforms into the equation* $\frac{d}{dt}\mathbf{w} = \tilde{\mathbf{M}}\mathbf{w}$*. Note, to achieve this, you can insert* $\mathbf{1} = \mathbf{R}^{\mathrm{T}}_{\pi/4}\mathbf{R}_{\pi/4}$ *between* $\mathbf{M}$ *and* $\mathbf{v}$*.*

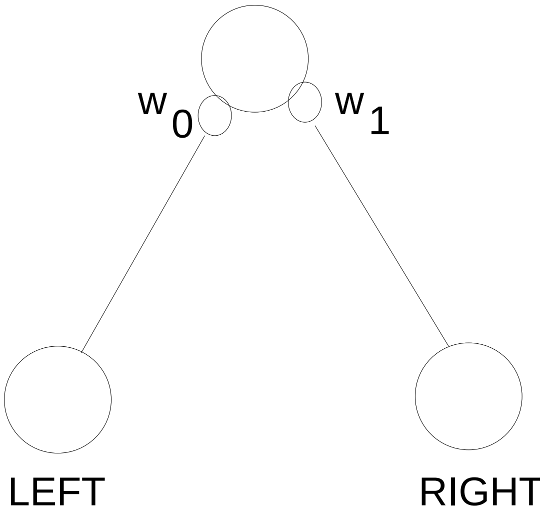

Figure 3.1: **A Very Simple Model of Ocular Dominance**
Two input cells, one from each eye, project to one output cell. The synapse from the left-eye cell is w_0; that from the right-eye cell is w_1.

So, we have a messy-looking matrix equation for $\mathbf{w}$. The developments of w_0 and w_1 are *coupled*: the development of w_0 depends on both w_0 and w_1, and similarly for the development of w_1. But we know that really, there are two independent, uncoupled one-dimensional equations hidden here: the development of v_0 depends only on v_0, that of v_1 only on v_1. Things are really simple, if we can only find our way back. How do we find our way back, assuming we don't know the transformation that got us here in the first place? That is, given Eq. 3.13, how could we ever realize that, by a change of basis, we could change it into Eq. 3.6, where the matrix is diagonal and the equations trivial?

The answer is, we have to find the *eigenvectors* of the matrix $\tilde{M}$, as explained in the following sections. This is a general method for finding the coordinates (if any exist — more on that in a bit) in which the matrix determining time development becomes diagonal. In this coordinate system, the equations become trivial — just a set of independent, uncoupled, one-dimensional equations.

Before considering how to do that in general, however, let's consider our example problems again.

3.2 Two Examples

3.2.1 A Simple Correlation-Based Model of Ocular Dominance

We consider perhaps the simplest possible model of ocular dominance: one postsynaptic cell, two presynaptic cells, one from each eye. There are two synapses, one from each presynaptic cell onto the postsynaptic cell. Let the synaptic strength from the left eye be w_0, that from the right eye, w_1 (Fig. 3.1).

Assume we have a correlation-based rule for synaptic development of the form $\tau \frac{d}{dt} w_i = \sum_j C_{ij} w_j$, where $\mathbf{C}$ is the matrix of correlations between the inputs, and τ is a constant determining the speed of development (Eq. 1.8). Let the self-correlation be 1, and let the between-eye correlation be ϵ.

Then the synaptic development equations are

$$\tau \frac{d}{dt} w_0 \quad = \quad (w_0 + \epsilon w_1) \tag{3.14}$$

$$\tau \frac{d}{dt} w_1 \quad = \quad (\epsilon w_0 + w_1) \tag{3.15}$$

or, in matrix notation,

$$\tau \frac{d}{dt} \mathbf{w} = \mathbf{C}\mathbf{w} \tag{3.16}$$

where the correlation matrix is

$$\mathbf{C} = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix} \tag{3.17}$$

Exercise 3.4 *Make sure you understand exactly where every term in Eqs. 3.14-3.17 comes from.*

Eq. 3.16 has exactly the same form as Eq. 3.11. The equations are identical if we set $\frac{1}{\tau}\mathbf{C} = \tilde{\mathbf{M}}$; this requires $(k_1 + k_0)/2 = 1/\tau$, $(k_1 - k_0)/2 = \epsilon/\tau$, which we can solve to find $k_0 = (1 - \epsilon)/\tau$, $k_1 = (1 + \epsilon)/\tau$. Thus, with this identification, Eq. 3.16 *is* Eq. 3.11.

Exercise 3.5 *Don't just take my word for it: show that the equations are equivalent when k_0 and k_1 are as stated.*

In this case, the natural experimental variables were the synaptic weights — w_0 and w_1. But, by the derivation of Eq. 3.11 from Eq. 3.6, we know that the variables in which the equations simplify — in which they become independent, one-dimensional equations — are $v_0 = (1/\sqrt{2})(w_0 - w_1)$, and $v_1 = (1/\sqrt{2})(w_0 + w_1)$. These correspond, respectively, to the ocular dominance or difference between the strength of the two eyes, v_0, and the sum of the two eyes' strength, v_1. We know the solutions to this model: from Eqs. 3.7-3.8, they are

$$v_0(t) \quad = \quad v_0(0) e^{\frac{(1-\epsilon)}{\tau} t} \tag{3.18}$$

$$v_1(t) \quad = \quad v_1(0) e^{\frac{(1+\epsilon)}{\tau} t} \tag{3.19}$$

So, when the two eyes are anticorrelated — when $\epsilon < 0$ — then the ocular dominance v_0 outgrows the sum of the two eyes' strengths v_1. But when the two eyes are correlated — when $\epsilon > 0$ — then the sum outgrows the ocular dominance. In either case, the sum and the ocular dominance grow *independently* – each grows along its merry way, oblivious to the presence of the other.

Problem 3.2 *Show that, when the ocular dominance v_0 outgrows the sum v_1, the eye whose synaptic strength is initially strongest takes over — its synapse grows, and the other eye's synapse shrinks (or grows more negative). Show that when the sum v_1 outgrows the ocular dominance v_0, both eyes' synapses grow (although the difference between their strengths — the ocular dominance — also grows, for $\epsilon < 1$).*

To show these results, note that (1) the left eye's synaptic strength w_0, is proportional to $v_1 + v_0$, while the right eye's strength w_1 is proportional to $v_1 - v_0$; (2) if the left eye's synapse is initially stronger, $v_0(0) > 0$ and v_0 grows increasingly more positive with time; while if the right eye's synapse is initially stronger, $v_0(0) < 0$ and v_0 grows increasingly more negative with time.

Note that we have not incorporated anything in this model to make it competitive (for example, conserving synaptic weight, so that when one eye gains strength, the other eye must lose strength) — both eyes' synapses can gain strength, even though one may be growing faster than the other. Nor have we incorporated any limits on synaptic weights, for example, restricting them to remain positive or to remain less than some maximum strength. So this is a very simplified model, even beyond the fact that there are only two presynaptic and one postsynaptic cells. Nonetheless, it already captures a bit of the flavor of a model for development of ocular dominance.

3.2.2 Two symmetrically coupled linear neurons

We return to Eq. 1.9 for activity in a linear network of neurons, and consider a case in which there are just two neurons, which make identical weights onto each other: $B_{01} = B_{10} = B$ (this is what I mean by "symmetrically coupled"). We exclude self-synapses: $B_{00} = B_{11} = 0$. So Eq. 1.9 becomes

$$\tau \frac{d}{dt} b_0 = -b_0 + B b_1 + h_0 \tag{3.20}$$

$$\tau \frac{d}{dt} b_1 = -b_1 + B b_0 + h_1 \tag{3.21}$$

or, in matrix notation,

$$\tau \frac{d}{dt} \mathbf{b} = -(\mathbf{1} - \mathbf{B})\mathbf{b} + \mathbf{h} \tag{3.22}$$

where the matrix (**1-B**) is

$$\mathbf{1} - \mathbf{B} = \begin{pmatrix} 1 & -B \\ -B & 1 \end{pmatrix} \tag{3.23}$$

Consider the case of no external input: $\mathbf{h} = 0$. Then Eqs. 3.22-3.23 are identical to Eqs. 3.16-3.17 for the ocular dominance model, except for two changes: (1) There is a minus sign in front of the right hand side and (2) The parameter ϵ, the between-eye correlation, is replaced by $-B$, the negative of the between-neuron synaptic weight. One way to see what the minus sign does is that it is equivalent to replacing τ with $-\tau$. So from the solutions, Eqs. 3.18-3.19, of the ocular dominance model, we can immediately write down the solutions for the two-neuron activity model by substituting $\epsilon \to -B$ and $\tau \to -\tau$. Thus, letting $v_0 = (1/\sqrt{2})(b_0 - b_1)$ and $v_1 = (1/\sqrt{2})(b_0 + b_1)$ be the difference and sum, respectively, of the two neurons' activities, the solutions are

$$v_0(t) = v_0(0) e^{-\frac{(1+B)}{\tau} t} \tag{3.24}$$

$$v_1(t) = v_1(0) e^{-\frac{(1-B)}{\tau} t} \tag{3.25}$$

What does this solution mean? Consider two cases:

- Case 1: $|B| < 1$. In this case, both the sum and the difference of the activities decay to zero. If B is excitatory ($B > 0$), the sum v_1 decays more slowly than the difference v_0, meaning that the two activities quickly approach one another and more slowly move together to zero. If B is inhibitory ($B < 0$), the sum decays more quickly than the difference: the two activities quickly approach being equal in magnitude and opposite in sign (so that their sum is near 0), while their magnitudes move more slowly toward zero.

- Case 2: $|B| > 1$. In this case,the system is unstable: one of the two terms, v_0 or v_1, will grow exponentially, while the other will decay to zero. In this case, if B is excitatory, the sum grows while the difference shrinks, so that the two activities approach one another while both grow without bound; while if B is inhibitory, the difference grows while the sum shrinks, so that the two activities approach having equal magnitude but opposite sign, while the magnitude of each grows without bound.

This should all make intuitive sense: cells that equally excite one another ought to approach similar activity values, while cells that equally inhibit one another ought to approach opposite activity values; and feedback with a gain of less than one gives a stable outcome, while a gain of greater than one gives an unstable outcome.

We'll deal with the case of a nonzero external input $\mathbf{h}$ a little later.

3.2.3 Generalizing from these examples

To solve these problems, we had to know how to get from the $\mathbf{w}$ or $\mathbf{b}$ representation back to the $\mathbf{v}$ representation — the representation in which the matrix $\mathbf{C}$ or $(\mathbf{1}-\mathbf{B})$ became diagonal. We happened to know the way in this case, because we had already come the other way: starting from $\mathbf{v}$, we had found $\mathbf{w}$ or $\mathbf{b}$. Now we have to figure out how to solve this problem more generally.

3.3 Eigenvalues and Eigenvectors: The Coordinate System in Which a Matrix is Diagonal

Suppose we are faced with an equation $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$. Suppose there is an orthonormal basis $\mathbf{e}_i$, $i = 0, \ldots, N-1$, in which $\mathbf{M}$ is diagonal:

$$\mathbf{M}_{\{\mathbf{e}_\mathrm{i}\ \mathrm{basis}\}} = \begin{pmatrix} \lambda_0 & 0 & \ldots & 0 \\ 0 & \lambda_1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_{N-1} \end{pmatrix} \tag{3.26}$$

In the $\mathbf{e}_i$ basis, the $\mathbf{e}_i$ are just $\mathbf{e}_0 = (1, 0, \ldots, 0)^\mathrm{T}$, $\mathbf{e}_1 = (0, 1, \ldots, 0)^\mathrm{T}$, …, $\mathbf{e}_{N-1} = (0, 0, \ldots, 1)^\mathrm{T}$. Thus, by working in the $\mathbf{e}_i$ basis, we can see that, for each $i = 0, \ldots, N-1$,

$$\mathbf{M}\mathbf{e}_i = \lambda_i \mathbf{e}_i \tag{3.27}$$

Problem 3.3 *Show that Eq. 3.27 holds in any coordinate system: apply $\mathbf{O}$ from the left to both sides of the equation, and insert $\mathbf{O}^\mathrm{T}\mathbf{O}$ between $\mathbf{M}$ and $\mathbf{e}_i$; and note that, in the new coordinate system, $\mathbf{M}$ is transformed to $\mathbf{O}\mathbf{M}\mathbf{O}^\mathrm{T}$, while $\mathbf{e}_i$ is transformed to $\mathbf{O}\mathbf{e}_i$.*

This brings us to define the *eigenvectors* and *eigenvalues* of a matrix:

Definition 3.1 *The* **eigenvectors** *of a matrix $\mathbf{M}$ are vectors $\mathbf{e}_i$ satisfying $\mathbf{M}\mathbf{e}_i = \lambda_i \mathbf{e}_i$ for some scalar λ_i. The λ_i are known as the* **eigenvalues** *of the matrix $\mathbf{M}$.*

Thus, if $\mathbf{M}$ is diagonal in some orthonormal basis $\mathbf{e}_i$, then the $\mathbf{e}_i$ are eigenvectors of $\mathbf{M}$. Therefore, $\mathbf{M}$ has a complete, orthonormal basis of eigenvectors. But this means we can immediately solve our original problem, $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$, as follows.

We expand $\mathbf{v}$ as $\mathbf{v} = \sum_i v_i \mathbf{e}_i$. As discussed previously in exercise 3.3, $\frac{d}{dt}\mathbf{v} = \sum_i \mathbf{e}_i \frac{d}{dt} v_i$. $\mathbf{M}\mathbf{v} = \mathbf{M}\sum_i v_i \mathbf{e}_i = \sum_i v_i \mathbf{M}\mathbf{e}_i = \sum_i v_i \lambda_i \mathbf{e}_i$. Thus $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$ becomes

$$\sum_i \mathbf{e}_i \frac{d}{dt} v_i = \sum_i \mathbf{e}_i v_i \lambda_i \tag{3.28}$$

Each side of this equation is a vector. We pick out one component of this vector (in the eigenvector basis), let's call it the j^{th} one, by applying $\mathbf{e}_j^{\mathrm{T}}$ to both sides of the equation. Thus, we obtain

$$\frac{d}{dt}v_j = v_j\lambda_j \tag{3.29}$$

Exercise 3.6 *Derive Eq. 3.29 from Eq. 3.28.*

Eq. 3.29 has the solution

$$v_j(t) = v_j(0)e^{\lambda_j t}. \tag{3.30}$$

Here $v_j(0)$ is the projection of $\mathbf{v}$ on the eigenvector $\mathbf{e}_j$ at time 0: $v_j(0) = \mathbf{e}_j^{\mathrm{T}}\mathbf{v}(0)$, where $\mathbf{v}(0)$ is the vector $\mathbf{v}$ at time 0. Thus, the equations decompose into N independent one-dimensional equations, one describing each v_j. The v_j's grow exponentially, independently of one another. Thus, the component of $\mathbf{v}$ in the $\mathbf{e}_j$ direction grows independently and exponentially at the rate λ_j.

Putting it all together, we obtain

$$\mathbf{v}(t) = \sum_i v_i(t)\mathbf{e}_i = \sum_i v_i(0)e^{\lambda_i t}\mathbf{e}_i = \sum_i [\mathbf{e}_i^{\mathrm{T}}\mathbf{v}(0)]e^{\lambda_i t}\mathbf{e}_i \tag{3.31}$$

It must be emphasized that this solution is expressed in terms of a specific set of vectors, the eigenvectors $\mathbf{e}_i$ of $\mathbf{M}$; the $\mathbf{e}_i$ do not represent any orthonormal basis, but only the eigenvector basis.

Problem 3.4 *Assume that* $\mathbf{M}$ *has a complete orthonormal basis of eigenvectors,* $\mathbf{e}_i$*, with eigenvalues* λ_i*.* Without looking at these notes, *write down the procedure for solving the equation* $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$*. The steps are*

1. *Expand* $\mathbf{v}$ *in terms of the* $\mathbf{e}_i$*;*
2. *Apply* $\frac{d}{dt}$ *and* $\mathbf{M}$ *to this expanded form of* $\mathbf{v}$*;*
3. *Apply* $\mathbf{e}_j^{\mathrm{T}}$ *to pull out the equation for* v_j*;*
4. *Write down the solution for component* $v_j(t)$*;*
5. *Use this to write down the solution for* $\mathbf{v}(t)$*.*

If necessary, you may look at this list, but don't otherwise look at these notes, when you solve this problem for the last time.

It turns out that eigenvalues can sometimes be complex numbers, and the corresponding eigenvectors then are vectors of complex numbers. We'll deal with that in a while. For the moment, let's assume that we're dealing with a matrix whose eigenvalues are real numbers. Then our solution, Eq. 3.31, shows several things: (1) If all the eigenvalues are negative, then $\mathbf{v}(t)$ decays to zero. (2) More generally, the components of $\mathbf{v}$ in the direction of eigenvectors with positive eigenvalue grow in time, while those in the direction of eigenvectors with negative eigenvalue decay. (3) Assuming there is at least one positive eigenvalue: after long times, the solution $\mathbf{v}(t)$ points more or less in the direction of the fastest-growing eigenvector, the one with the largest eigenvalue. For example, the ratio of any two components $v_i(t)$ and $v_j(t)$ in the eigenvector basis is $v_i(t)/v_j(t) = [v_i(0)/v_j(0)]e^{(\lambda_i-\lambda_j)t}$. If $\lambda_i > \lambda_j$, then this ratio grows exponentially with time, and will eventually grow as large as you like. Thus, the component corresponding to the eigenvector with largest eigenvalue becomes exponentially larger than any other components, and dominates the development over long times (of course, if all of the eigenvalues are negative, then all of the components are going to zero, so this would only mean that this component is going to zero more slowly than the others).

Because this eigenvector with largest eigenvalue plays a special role, we give it a name, the **principal eigenvector** of $\mathbf{M}$.

Problem 3.5 *We return to the example of section 3.2.1:*

$$\tau \frac{d}{dt}\mathbf{w} = \mathbf{C}\mathbf{w} = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix} \mathbf{w} \tag{3.32}$$

- *We can rewrite this as $\frac{d}{dt}\mathbf{w} = \frac{1}{\tau}\mathbf{C}\mathbf{w}$. Show that the eigenvectors of $\frac{1}{\tau}\mathbf{C}$ are $\mathbf{e}_S = \frac{1}{\sqrt{2}}(1,1)^{\mathrm{T}}$, with eigenvalue $\lambda_S = (1+\epsilon)/\tau$, and $\mathbf{e}_D = \frac{1}{\sqrt{2}}(1,-1)^{\mathrm{T}}$, with eigenvalue $\lambda_D = (1-\epsilon)/\tau$. The factors of $\frac{1}{\sqrt{2}}$ are just there to normalize the eigenvectors: they make $\mathbf{e}_S^{\mathrm{T}}\mathbf{e}_S = 1$ and $\mathbf{e}_D^{\mathrm{T}}\mathbf{e}_D = 1$. (Hint: all that's required here is to show that $\frac{1}{\tau}\mathbf{C}\mathbf{e}_0 = \lambda_0\mathbf{e}_0$ and $\frac{1}{\tau}\mathbf{C}\mathbf{e}_1 = \lambda_1\mathbf{e}_1$.)*

  Note: these expressions for the eigenvectors are written in the basis of the left-eye and right-eye weights, w_0 and w_1 (which is the same basis in which $\mathbf{C}$ is written in Eq. 3.32). I label the eigenvectors with S for sum and D for difference rather than calling them $\mathbf{e}_0$ and $\mathbf{e}_1$, so that I can reserve 0 and 1 for the basis of left-eye and right-eye weights w_0 and w_1; so w_0 represents the left-eye strength, whereas w_S represents the component of $\mathbf{w}$ in the $\mathbf{e}_S$ direction.) .

- *Thus, the solution of Eq. 3.32 is*

$$\mathbf{w}(t) = \mathbf{e}_S^{\mathrm{T}}\mathbf{w}(0)e^{\lambda_S t}\mathbf{e}_S + \mathbf{e}_D^{\mathrm{T}}\mathbf{w}(0)e^{\lambda_D t}\mathbf{e}_D \tag{3.33}$$

  You don't have to write anything down for this section, but some points to notice: note the correspondence between this result and Eqs. 3.18-3.19, as follows. The $\mathbf{e}_S$ component of $\mathbf{w}$, $\mathbf{e}_S^{\mathrm{T}}\mathbf{w}$, corresponds to the sum of the left-eye plus right-eye weights, which we called v_1; while $\mathbf{e}_D^{\mathrm{T}}\mathbf{w}$ corresponds to their difference, the ocular dominance, which we called v_0. The time course of these components is $\mathbf{e}_S^{\mathrm{T}}\mathbf{w}(t) = \mathbf{e}_S^{\mathrm{T}}\mathbf{w}(0)e^{\lambda_S t}$, and $\mathbf{e}_D^{\mathrm{T}}\mathbf{w}(t) = \mathbf{e}_D^{\mathrm{T}}\mathbf{w}(0)e^{\lambda_D t}$; note the correspondence of these to Eqs. 3.18-3.19.

  Also, understand the following: (1) If $\epsilon > 0$, then $\mathbf{e}_S$ is the principal eigenvector, so over long time the weights approach the $\mathbf{e}_S$ direction: that is, the two weights become equal; (2) If $\epsilon < 0$, then $\mathbf{e}_D$ is the principal eigenvector, so over long time the weights approach the $\mathbf{e}_D$ direction: that is, the two weights become equal in magnitude but opposite in sign; (3) The sign of the component in the $\mathbf{e}_D$ direction doesn't change with time (i.e. $\mathbf{e}_D^{\mathrm{T}}\mathbf{w}(t) = \mathbf{e}_D^{\mathrm{T}}\mathbf{w}(0)e^{\lambda_D t}$, so the sign of $\mathbf{e}_D^{\mathrm{T}}\mathbf{w}(t)$ is the same as the sign of $\mathbf{e}_D^{\mathrm{T}}\mathbf{w}(0)$); therefore, whichever synapse is initially largest stays largest. In particular, for $\epsilon < 0$, this means that the initially larger synapse grows strong and positive, while the other synapse becomes strong and negative.

- *Write down equation 3.33 in the w_0, w_1 basis, to derive the solution for the left-eye and right-eye weights, $w_0(t)$ and $w_1(t)$:*

$$w_0(t) = \frac{1}{2}\left\{[w_0(0) + w_1(0)]\, e^{\frac{(1+\epsilon)t}{\tau}} + [w_0(0) - w_1(0)]\, e^{\frac{(1-\epsilon)t}{\tau}}\right\} \tag{3.34}$$

$$w_1(t) = \frac{1}{2}\left\{[w_0(0) + w_1(0)]\, e^{\frac{(1+\epsilon)t}{\tau}} - [w_0(0) - w_1(0)]\, e^{\frac{(1-\epsilon)t}{\tau}}\right\} \tag{3.35}$$

  Confirm that substituting $t = 0$ on the right side gives back $w_0(0)$ and $w_1(0)$ for Eqs. 3.34 and 3.35, respectively, as it should.

Again, just understand the following, no need to write anything down: we can draw the same conclusions from these equations as we drew from Eq. 3.33: (1) For $\epsilon > 0$, $w_0(t)$ and $w_1(t)$ become roughly equal as $t \to \infty$, that is, the growth of the sum dominates the growth of the difference, because the $e^{\frac{(1+\epsilon)t}{\tau}}$ term dominates the $e^{\frac{(1-\epsilon)t}{\tau}}$ term (2) For $\epsilon < 0$, $w_0(t)$ and $w_1(t)$ become roughly equal in magnitude but opposite in sign, that is, the growth of the difference dominates the growth of the sum, because the $e^{\frac{(1-\epsilon)t}{\tau}}$ term dominates the $e^{\frac{(1+\epsilon)t}{\tau}}$ term; (3) Whichever synapse is initially stronger will always remain stronger, because the first term on the right is identical for the two weights, while the second term is always positive for the initially stronger synapse (whichever is larger of $w_0(0)$ and $w_1(0)$) and always negative for the initially weaker synapse. In particular, for $\epsilon < 0$, the initially stronger synapse will ultimately grow strong and positive, while the initially weaker will ultimately become strong and negative.

In summary, if we can find a basis in which a matrix $\mathbf{M}$ is diagonal — that is, if we can find a complete orthonormal basis of eigenvectors of $\mathbf{M}$ — we have solved our problem: we have found our way back to the basis in which the matrix equation is really just a set of independent one-dimensional equations. That basis *is* just the basis of eigenvectors. We no longer need to have started from that basis, as in sections 3.1-3.2, in order to find our way back. Starting from the matrix $\mathbf{M}$, our task is to find its eigenvectors.

3.4 A Matrix Is Characterized By Its Eigenvectors and Eigenvalues

The word "eigen" in German translates as "own"; that is, the eigenvectors are the matrix's own vectors, the vectors that belong to it ("eigenvector" is also sometimes translated as "characteristic vector"). The following problems should make help make clear why this is so:

Problem 3.6 • *Show that a set of orthonormal eigenvectors and their eigenvalues uniquely characterize a matrix, as follows. If $\mathbf{e}_i$ is a complete orthonormal basis of eigenvectors of $\mathbf{M}$, with eigenvalues λ_i, then*

$$\mathbf{M} = \sum_i \lambda_i \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}} \tag{3.36}$$

This just says that, for any vector $\mathbf{v}$, if $\mathbf{v} = \sum_i v_i \mathbf{e}_i$, then $\mathbf{Mv} = \sum_i \lambda_i v_i \mathbf{e}_i$. That is, $\mathbf{M}$ is precisely the matrix that breaks any vector $\mathbf{v}$ down into its components along each eigenvector, multiplies the i^{th} component by λ_i, and then puts the components back together to give $\mathbf{Mv}$.

To show that $\mathbf{M} = \sum_i \lambda_i \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}$, go back to the definition in Eq. 2.22 of $\mathbf{M}$ with respect to any basis set $\mathbf{e}_k$: $\mathbf{M} = \sum_{i,j} M_{ij} \mathbf{e}_i \mathbf{e}_j^{\mathrm{T}}$ where $M_{ij} = \mathbf{e}_i^{\mathrm{T}} M \mathbf{e}_j$. In the eigenvector basis, show that $M_{ij} = \lambda_j \delta_{ij}$. Plug this in to the sum to get $\mathbf{M} = \sum_j \lambda_j \mathbf{e}_j \mathbf{e}_j^{\mathrm{T}}$. (An alternate proof: use the expansion $\mathbf{v} = \sum_i \mathbf{e}_i(\mathbf{e}_i^{\mathrm{T}} \mathbf{v})$ to show that for any *vector $\mathbf{v}$, $\mathbf{Mv} = \sum_i \lambda_i \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}} \mathbf{v}$; and see footnote 3 to see why, for two matrices $\mathbf{M}$ and $\mathbf{P}$, if $\mathbf{Mv} = \mathbf{Pv}$ for* any *vector $\mathbf{v}$, then the two matrices are equal).*

- *Now let's see a concrete example of this. Take the matrix $\mathbf{C} = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}$, which has eigenvectors $\mathbf{e}_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ with eigenvalue $\lambda_0 = 1 + \epsilon$ and $\mathbf{e}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ with eigenvalue $\lambda_0 = 1 - \epsilon$. Write down $\sum_i \lambda_i \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}$ for this matrix (in the basis in which the eigenvectors are as I've just written them) – this is a sum of two matrices – and show that you get back the original matrix $\mathbf{C}$.*

Now work in the eigenvector basis, so that $\mathbf{e}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ *and* $\mathbf{e}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$*. Write down* $\sum_i \lambda_i \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}$ *in this basis and show that you get* $\mathbf{C}$ *as written in the eigenvector basis (recall that in the eigenvector basis,* $\mathbf{C}$ *is a diagonal matrix whose diagonal entries are the eigenvalues). Hopefully, this should help to convince you that the equation* $\mathbf{C} = \sum_i \lambda_i \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}$ *is a general, coordinate-invariant statement about the relationship between* $\mathbf{C}$ *and its eigenvectors/eigenvalues, and is true in particular in any coordinate system in which we wish to work.*

Problem 3.7 *Show that if* $\mathbf{M}$ *is a matrix with a complete orthonormal basis of eigenvectors* $\mathbf{e}_i$*, with corresponding eigenvalues* λ_i*, and if* $\lambda_i \neq 0$ *for all* i*, then the inverse of* $\mathbf{M}$ *exists and is given by*

$$\mathbf{M}^{-1} = \sum_i \frac{1}{\lambda_i} \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}} \tag{3.37}$$

To do this, simply show that $\mathbf{M}\mathbf{M}^{-1} = \mathbf{M}^{-1}\mathbf{M} = \mathbf{1}$.

Now show that this works for a specific case, the matrix $\mathbf{C}$ *of the second part of problem 3.6. Write down* $\sum_i \frac{1}{\lambda_i} \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}$ *for* $\mathbf{C}$*, in the basis in which the eigenvectors are* $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ \pm 1 \end{pmatrix}$*, and show that the resulting matrix is the inverse of* $\mathbf{C}$ *(that is, show that multiplying it by* $\mathbf{C}$ *from either side gives the identity).*

Recall that we said that a matrix fails to have an inverse when it takes some nonzero matrix to 0: if $\mathbf{M}\mathbf{v} = 0$ for some $\mathbf{v} \neq 0$, then $\mathbf{M}$ is not invertible. But $\mathbf{M}\mathbf{v} = 0$ for $\mathbf{v} \neq 0$ precisely means that $\mathbf{M}$ has a zero eigenvalue. A matrix is invertible if and only if it has no zero eigenvalues. This is a general truth about matrices; Eq. 3.37 shows this for the specific case of matrices with complete orthonormal bases of eigenvectors, by explicitly writing down the inverse when no eigenvalues are zero.

Intuitively, Eq. 3.37 should make sense: $\mathbf{M}$ acts on any vector by taking the vector's component along each eigenvector and multiplying it by the corresponding eigenvalue. So $\mathbf{M}^{-1}$ is the vector that "undoes" this: it acts on any vector by taking the vector's component along each eigenvector and multiplying it by the *inverse* of the corresponding eigenvalue. Thus, following $\mathbf{M}$ by $\mathbf{M}^{-1}$ leaves everything unchanged, as does following $\mathbf{M}^{-1}$ by $\mathbf{M}$; that is, $\mathbf{M}\mathbf{M}^{-1} = \mathbf{M}^{-1}\mathbf{M} = \mathbf{1}$.

3.5 When does a Matrix Have a Complete Orthonormal Basis of Eigenvectors?

Any **symmetric** matrix always has a complete, orthonormal basis of eigenvectors. This is convenient for simple correlation-based models: for example, the correlation of input i to input j is the same as the correlation of input j to input i, so the matrix describing correlations between inputs is a symmetric matrix.

Exercise 3.7 *For those who are interested: here's how to show that eigenvectors of a symmetric matrix* $\mathbf{M}$ *are mutually orthogonal. Let* $\mathbf{e}_i$*,* $\mathbf{e}_j$ *be two eigenvectors, with eigenvalues* λ_i*,* λ_j*. Then* $\mathbf{e}_i^{\mathrm{T}}\mathbf{M}\mathbf{e}_j = \mathbf{e}_i^{\mathrm{T}}(\mathbf{M}\mathbf{e}_j) = \mathbf{e}_i^{\mathrm{T}}\lambda_j\mathbf{e}_j = \lambda_j\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j$*. But also,* $\mathbf{e}_i^{\mathrm{T}}\mathbf{M}\mathbf{e}_j = (\mathbf{e}_i^{\mathrm{T}}\mathbf{M})\mathbf{e}_j = (\mathbf{M}\mathbf{e}_i)^{\mathrm{T}}\mathbf{e}_j = \lambda_i\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j$ *(note, we used the fact that* $\mathbf{M}$ *is symmetric to set* $(\mathbf{e}_i^{\mathrm{T}}\mathbf{M}) = (\mathbf{M}\mathbf{e}_i)^{\mathrm{T}}$*). Thus,* $\lambda_j\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = \lambda_i\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j$*, or* $(\lambda_j - \lambda_i)\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = 0$*. If* $\lambda_j \neq \lambda_i$*, then* $\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = 0$*.*

If $\lambda_i = \lambda_j = \lambda$*, then any linear combination of* $\mathbf{e}_i$ *and* $\mathbf{e}_j$ *is also an eigenvector with the same eigenvalue:* $\mathbf{M}(a\mathbf{e}_i + b\mathbf{e}_j) = \lambda(a\mathbf{e}_i + b\mathbf{e}_j)$*. By a process called Gram-Schmidt orthogonalization, we can replace* $\mathbf{e}_j$ *by a linear combination that is orthogonal to* $\mathbf{e}_i$*. This can be extended to arbitrary* *numbers of eigenvectors that share an eigenvalue. Thus, we choose eigenvectors belonging to a single eigenvalue to be orthogonal, while eigenvectors belonging to different eigenvalues are automatically orthogonal. In this way, all of the eigenvectors can be chosen to be mutually orthogonal.*

Symmetric matrices have another nice property: all of the eigenvectors and eigenvalues of a real symmetric matrix are real (in general, the eigenvalues and eigenvectors of a real matrix may be complex).

When matrices are not symmetric, things can get somewhat more complicated. To describe this in more detail, we will have to think about complex rather than real vector spaces. We will get to this soon enough. The basic answer is that "most" matrices do have a complete basis of eigenvectors, though not necessarily an orthonormal basis and not necessarily a real one. For quite a while, we're only going to worry about matrices that have orthonormal bases, but we will soon have to deal with complex eigenvectors — for example, as soon as we think about Fourier transforms. For the moment, though, we'll just restrict ourselves to thinking about symmetric matrices.

3.6 The Matrix That Transforms to the Eigenvector Basis

Suppose $\mathbf{M}$ has a complete orthonormal basis of eigenvectors $\mathbf{e}_i$, with eigenvalues λ_i. We saw in section 2.5 that the orthogonal transformation $\mathbf{O}$ that takes us from our current basis to this eigenvector basis is $\mathbf{O} = (\ \mathbf{e}_0\ \mathbf{e}_1\ \dots\ \mathbf{e}_{N-1}\ )^{\mathrm{T}}$. In this basis, $\mathbf{M}$ is diagonal, that is, $\mathbf{OMO}^{\mathrm{T}}$ is diagonal.

Exercise 3.8 *Show that this is true, by computing* $\mathbf{OMO}^{\mathrm{T}}$*. Your final result should be*

$$\mathbf{OMO}^{\mathrm{T}} = \begin{pmatrix} \lambda_0 & 0 & \dots & 0 \\ 0 & \lambda_1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_{N-1} \end{pmatrix} \tag{3.38}$$

Here's how to do it. First, compute $\mathbf{MO}^{\mathrm{T}}$ *by showing that* $\mathbf{M}(\ \mathbf{e}_0\ \mathbf{e}_1\ \dots\ \mathbf{e}_{N-1}\ ) = (\ \mathbf{Me}_0\ \mathbf{Me}_1\ \dots\ \mathbf{Me}_{N-1}\ )$*. You should be able to think through why this is so, by thinking about the operation of successive rows of* $\mathbf{M}$ *on successive columns of* $\mathbf{O}^{\mathrm{T}}$*: the first row of* $\mathbf{M}$ *acts successively on each column of* $\mathbf{O}^{\mathrm{T}}$ *to produce the entry in the first row of that column in the product; the second row of* $\mathbf{M}$ *acts successively on each column of* $\mathbf{O}^{\mathrm{T}}$ *to produce the entry in the second row of that column in the product; etc. You can also prove it in components: the left-hand side is a matrix with* $(ij)^{\mathrm{th}}$ *component* $\sum_k M_{ik}(\mathbf{O}^{\mathrm{T}})_{kj} = \sum_k M_{ik}(\mathbf{e}_j)_k$*; while the right-hand side has* $(ij)^{\mathrm{th}}$ *component* $(\mathbf{Me}_j)_i = \sum_k M_{ik}(\mathbf{e}_j)_k$*. Now, use the fact that the* $\mathbf{e}_i$ *are the eigenvectors of* $\mathbf{M}$*, to convert this to* $(\ \lambda_0\mathbf{e}_0\ \lambda_1\mathbf{e}_1\ \dots\ \lambda_{N-1}\mathbf{e}_{N-1}\ )$*. Now multiply this from the left by* $\mathbf{O}$*, the matrix whose* rows *are the eigenvectors as row vectors, and use the orthonormality of the eigenvectors.*

3.7 The Determinant and Trace of a Matrix

To find the eigenvalues and eigenvectors of a matrix, we are going to need to deal with the determinant of a matrix; we write the determinant of the matrix $\mathbf{M}$ as $\det \mathbf{M}$. The determinant is a coordinate-invariant scalar function of a matrix (where by a scalar function we mean that it is a single number, rather than a vector or a matrix), composed of a sum of terms, each of which is a product of N elements of the matrix, where N is the dimension of the matrix. The determinant turns out to be equal to the product of the matrix's eigenvalues; so in particular, $\det \mathbf{M}$ is zero if and only if $\mathbf{M}$ has at least one zero eigenvalue.

The determinant can be defined as the unique scalar function that satisfies 3 properties:

1. $\det(\mathbf{MN}) = (\det \mathbf{M})(\det \mathbf{N})$ for all matrices $\mathbf{M}$ and $\mathbf{N}$;

2. $\det \mathbf{1} = 1$;

3. $\det \mathbf{M} \neq 0$ if and only if $\mathbf{M}$ has an inverse $\mathbf{M}^{-1}$

It can be shown that there is only one scalar function with these three properties, and that is the determinant. The 3rd condition becomes more intuitive if you know that a matrix is invertible if and only if it has no zero eigenvalues. The first two conditions guarantee that the determinant is coordinate-invariant: $\det(\mathbf{OMO}^{\mathrm{T}}) == (\det \mathbf{O})(\det \mathbf{O}^{\mathrm{T}})(\det \mathbf{M}) = (\det(\mathbf{OO}^{\mathrm{T}}))(\det \mathbf{M}) = \det \mathbf{M}$.

The formula for computing the determinant is best stated recursively. If $\mathbf{M}$ is $N \times N$, let $\mathbf{M}^{ij}$ be the $(N-1) \times (N-1)$ matrix obtained by deleting the i^{th} row and j^{th} column from $\mathbf{M}$. Then, for any row i,

$$\det \mathbf{M} = \sum_j (-1)^{i+j} M_{ij} \det \mathbf{M}^{ij} \tag{3.39}$$

(In particular, it's usually convenient to use the top row, $i = 0$). Alternatively one can pick any column j:

$$\det \mathbf{M} = \sum_i (-1)^{i+j} M_{ij} \det \mathbf{M}^{ij} \tag{3.40}$$

Both formulas yield the same answer, and they yield the same answer no matter which row or which column is chosen. These formulas reduce the problem of computing the determinant of an $N \times N$ matrix to one of computing the determinant of an $(N-1) \times (N-1)$ matrix. Finally we stop the recursion by stating that the determinant of the 1×1 matrix with the single element a is equal to a.

Problem 3.8 *1. Show that* $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = (ad - bc)$.

2. Show that $\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = a(ei - fh) - b(di - fg) + c(dh - eg)$

3. Show that for a diagonal matrix $\mathbf{D}$*,* $\det \mathbf{D}$ *is just the product of the diagonal entries, e.g.* $\det \begin{pmatrix} \lambda_0 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \end{pmatrix} = \lambda_0 \lambda_1 \lambda_2$*; this along with the coordinate-invariance of the determinant explains why the determinant of a matrix is equal to the product of the matrix's eigenvalues.*

4. Consider again the matrix $\mathbf{C} = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}$ *(e.g., problem 3.6, 2nd part). Compute its determinant, and show that it is equal to the product of the eigenvalues of* $\mathbf{C}$*.*

Exercise 3.9 *Some useful facts about the determinant that you might want to prove:*

- *If all the elements of one row or one column of a matrix are multiplied by k, the determinant is also multiplied by k.*
- $\det \mathbf{M}^T = \det \mathbf{M}$.
- $\det \mathbf{M}^{-1} = 1/(\det \mathbf{M})$ *(hint: use* $\det(\mathbf{MN}) = (\det \mathbf{M})(\det \mathbf{N})$*).*

Use the last two facts to prove that, for any orthogonal matrix $\mathbf{O}$*,* $\det \mathbf{O} = \pm 1$.

Although we will not be making use of it, this is also a good place to introduce another commonly-encountered, coordinate-invariant scalar function of a matrix, the **trace**. The trace of a matrix is the sum of its diagonal components: letting $\mathrm{Tr}\,\mathbf{M}$ signify the trace of $\mathbf{M}$, $\mathrm{Tr}\,\mathbf{M} = \sum_i M_{ii}$. It is easy to show that $\mathrm{Tr}\,(\mathbf{MN}) = \mathrm{Tr}\,(\mathbf{NM})$: $\mathrm{Tr}\,(\mathbf{MN}) = \sum_i (\mathbf{MN})_{ii} = \sum_{ij} M_{ij}N_{ji} = \sum_{ji} N_{ji}M_{ij} = \sum_j (\mathbf{NM})_{jj} = \mathrm{Tr}\,(\mathbf{NM})$. From this it follows that $\mathrm{Tr}\,(\mathbf{MNP}) = \mathrm{Tr}\,(\mathbf{PMN})$ (by considering $\mathbf{MN}$ as one matrix) and therefore that the trace is coordinate-invariant: $\mathrm{Tr}\,\mathbf{OMO}^{\mathrm{T}} = \mathrm{Tr}\,\mathbf{O}^{\mathrm{T}}\mathbf{OM} = \mathrm{Tr}\,\mathbf{M}$. The trace of any matrix is equal to the sum of its eigenvalues, as should be clear for symmetric matrices from taking the trace in the coordinate system in which the matrix is diagonal.

3.8 How To Find the Eigenvalues and Eigenvectors of a Matrix

How to do it in principle: the equation $\mathbf{Mv} = \lambda\mathbf{v}$ means $(\mathbf{M} - \lambda\mathbf{1})\mathbf{v} = 0$. This can only be true if $\det(\mathbf{M} - \lambda\mathbf{1}) = 0$. This is because, if $\mathbf{M}$ has an eigenvalue λ, then $\mathbf{M} - \lambda\mathbf{1}$ has a corresponding eigenvalue 0, so $\det(\mathbf{M} - \lambda\mathbf{1}) = 0$ for that value of λ. Thus we can find the eigenvalues of $\mathbf{M}$ by finding the values of λ that make $\det(\mathbf{M} - \lambda\mathbf{1}) = 0$. The equation $\det(\mathbf{M} - \lambda\mathbf{1}) = 0$ gives an N-th order polynomial equation for λ, known as the **characteristic equation** for $\mathbf{M}$ (and the polynomial $\det(\mathbf{M} - \lambda\mathbf{1})$ is known as the **characteristic polynomial** for $\mathbf{M}$); this has N solutions, corresponding to the N eigenvalues of $\mathbf{M}$. For each such solution – that is, for each eigenvalue λ – you can solve $\mathbf{Mv} = \lambda\mathbf{v}$ for the corresponding eigenvector.[4]

How to do it in practice: In some cases, by understanding/analyzing the mathematical structure and symmetries of the matrix, you can find an analytical solution, or make an inspired guess, that reveals the eigenvectors and eigenvalues. Otherwise, get a computer to do it for you. You can use a stand-alone program like Maple or Mathematica; or, you can write a program calling standard routines. See the book *Numerical Recipes in C, 2nd Edition*, by W. Press et al., Cambridge University Press, 1992. The "in principle" method outlined above is very inefficient and is only practical for very simple cases, like 2-dimensional matrices.

Exercise 3.10 *To make things clear, let's think through how to do it for our pet 2-dimensional case. Consider the two-dimensional symmetric matrix:*

$$\mathbf{M} = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix} \tag{3.41}$$

To find its eigenvalues, we need to find solutions to

$$\det(\mathbf{M} - \lambda\mathbf{1}) = \det\begin{pmatrix} 1-\lambda & \epsilon \\ \epsilon & 1-\lambda \end{pmatrix} = 0 \tag{3.42}$$

This yields $(1-\lambda)^2 - \epsilon^2 = 0$. We solve this quadratic equation for λ, giving $\lambda = 1 \pm \epsilon$. There are two solutions, the two eigenvalues.

Then, for each value of λ, we solve $\mathbf{Me} = \lambda\mathbf{e}$ for the corresponding eigenvector $\mathbf{e}$. Since the length of $\mathbf{e}$ is irrelevant, we can write $\mathbf{e}$ in terms of a single parameter, and solve for this parameter: for example, we could write $\mathbf{e} = (k, \sqrt{1-k^2})^{\mathrm{T}}$. However, a simpler form is $\mathbf{e} = (k, 1)^{\mathrm{T}}$ (this can't be used to find an eigenvector proportional to $(1,0)^{\mathrm{T}}$, but if $\epsilon \neq 0$, $(1,0)^{\mathrm{T}}$ can't be an eigenvector (check that this is true!), while if $\epsilon = 0$, the matrix is diagonal so the eigenvectors and eigenvalues are found trivially). Thus we need to solve $\begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}\begin{pmatrix} k \\ 1 \end{pmatrix} = (1 \pm \epsilon)\begin{pmatrix} k \\ 1 \end{pmatrix}$ or $\begin{pmatrix} k+\epsilon \\ 1+k\epsilon \end{pmatrix} = (1 \pm \epsilon)\begin{pmatrix} k \\ 1 \end{pmatrix}$. Check that the solutions are given by $k = 1$ for eigenvalue $1+\epsilon$ and $k = -1$ for eigenvalue $1-\epsilon$.

[4] Note that the eigenvector $\mathbf{v}$ is arbitrary up to an overall scalar multiple (that is, if $\mathbf{v}$ is an eigenvector with eigenvalue λ, so is $k\mathbf{v}$ for any scalar k), so you must fix the length of $\mathbf{v}$, say $\mathbf{v} \cdot \mathbf{v} = 1$.

Problem 3.9 *Find the eigenvectors and eigenvalues for the matrix* $\mathbf{M} = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$

Problem 3.10 *Show that if an* $N \times N$ *matrix* $\mathbf{M}$ *has eigenvalues* λ_i*,* $i = 0, \ldots, N-1$*, then the matrix* $\mathbf{M} + k\mathbf{1}$ *for a scalar* k *has eigenvalues* $\lambda_i + k$*,* $i = 0, \ldots, N-1$*. (Hint: how does adding* $k\mathbf{1}$ *modify the characteristic polynomial and its solutions? – show that if the characteristic equation of* $\mathbf{M}$ *has a solution* λ*, then the characteristic equation of* $\mathbf{M} + k\mathbf{1}$ *has a solution* $\lambda + k$*.) Show also that the eigenvectors are preserved: if* $\mathbf{e}_i$ *is an eigenvector of* $\mathbf{M}$ *with eigenvalue* λ_i*, then it is also an eigenvector of* $\mathbf{M} + k\mathbf{1}$ *with eigenvalue* $\lambda_i + k$*. Thus, adding a multiple of the identity matrix to a matrix just moves all the eigenvalues by a constant amount, leaving the action of the matrix otherwise unchanged.*

3.9 Ocular Dominance Again: Two Eyes That Each Fire Synchonously

Let's again consider the ocular dominance model, but now let there be N input cells from each eye, a total of 2N inputs. Again, we'll restrict ourselves to one postsynaptic cell. Suppose each eye fires as a unit — as when Mike Stryker and Sheri Harris put TTX in the eyes to silence spontaneous activity, and fired the optic nerves as units. Let's let the left-eye synapses onto the postsynaptic cell be $w_0, w_1, \ldots, w_{N-1}$, and the right-eye synapses be $w_N, w_{N+1}, \ldots, w_{2N-1}$. Let the value of the correlation in firing between any two left-eye inputs be 1, and similarly that for any two right-eye units is 1; and let the interocular correlation be ϵ. The matrix of input correlations is $\mathbf{C}$, whose components C_{ij} represent the correlation between input i and input j. This matrix is

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & \ldots & 1 & 1 & \epsilon & \epsilon & \ldots & \epsilon & \epsilon \\ 1 & 1 & \ldots & 1 & 1 & \epsilon & \epsilon & \ldots & \epsilon & \epsilon \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & \ldots & 1 & 1 & \epsilon & \epsilon & \ldots & \epsilon & \epsilon \\ 1 & 1 & \ldots & 1 & 1 & \epsilon & \epsilon & \ldots & \epsilon & \epsilon \\ \epsilon & \epsilon & \ldots & \epsilon & \epsilon & 1 & 1 & \ldots & 1 & 1 \\ \epsilon & \epsilon & \ldots & \epsilon & \epsilon & 1 & 1 & \ldots & 1 & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \epsilon & \epsilon & \ldots & \epsilon & \epsilon & 1 & 1 & \ldots & 1 & 1 \\ \epsilon & \epsilon & \ldots & \epsilon & \epsilon & 1 & 1 & \ldots & 1 & 1 \end{pmatrix} \tag{3.43}$$

Let $\mathbf{J}$ be the NxN matrix whose entries are all 1's. Then, we can rewrite $\mathbf{C}$ as

$$\mathbf{C} = \begin{pmatrix} \mathbf{J} & \epsilon\mathbf{J} \\ \epsilon\mathbf{J} & \mathbf{J} \end{pmatrix} \tag{3.44}$$

We want to solve the equation $\frac{d}{dt}\mathbf{w} = \mathbf{C}\mathbf{w}$. We'll do this in two steps. First, suppose that $\mathbf{J}$ has an orthonormal basis of (N-dimensional) eigenvectors, $\mathbf{j}_i$, $i = 0, \ldots, N-1$ with eigenvalues λ_i.

Problem 3.11 *Find the eigenvectors of* $\mathbf{C}$*, as follows:*

- *Show that* $\frac{1}{\sqrt{2}}\begin{pmatrix} \mathbf{j}_i \\ \mathbf{j}_i \end{pmatrix}$ *is an eigenvector of* $\mathbf{C}$*, with eigenvalue* $\lambda_i(1+\epsilon)$*; and that* $\frac{1}{\sqrt{2}}\begin{pmatrix} \mathbf{j}_i \\ -\mathbf{j}_i \end{pmatrix}$ *is an eigenvector of* $\mathbf{C}$*, with eigenvalue* $\lambda_i(1-\epsilon)$*. This should remind you of the eigenvectors and eigenvalues of the two-dimensional case we considered previously.*

Note: to do this, it will help to realize the following: if $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ *are each* $N \times N$ *matrices, and* $\mathbf{v}$ *and* $\mathbf{w}$ *are* N*-dimensional vectors, then* $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{Av} + \mathbf{Bw} \\ \mathbf{Cv} + \mathbf{Dw} \end{pmatrix}$. *Convince yourself intuitively that this is true, but you don't need to prove it, you may assume it.*

- *Show that these eigenvectors are orthonormal, and that there are 2N of them. Thus, we have found a complete orthonormal basis for* $\mathbf{C}$.

  Again, it will help to realize that, if $\mathbf{v}$, $\mathbf{w}$, $\mathbf{x}$, $\mathbf{y}$ *are all N-dimensional vectors, then the dot product* $\begin{pmatrix} \mathbf{v} \\ \mathbf{w} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \mathbf{v} \cdot \mathbf{x} + \mathbf{w} \cdot \mathbf{y}$. *You can try to prove this by writing these equations in terms of indices, or just assume it (but convince yourself intuitively that it is true).*

Second, we'll find the eigenvectors of $\mathbf{J}$.

Problem 3.12 *Find the eigenvectors of* $\mathbf{J}$, *as follows:*

- *Show that the N-dimensional vector* $\mathbf{j}_0 = (1, 1, \ldots, 1)^{\mathrm{T}}/\sqrt{N}$ *is an eigenvector of* $\mathbf{J}$, *with eigenvalue* $\lambda_0 = N$.
- *Show that any N-dimensional vector whose elements add up to zero is an eigenvector of* $\mathbf{J}$, *with eigenvalue 0. Show that any such vector is orthogonal to* $\mathbf{j}_0$.
- *Show that any N-dimensional vector orthogonal to* $\mathbf{j}_0$ *is a vector whose elements sum to zero. To show this, show that for any vector* $\mathbf{v}$, $\mathbf{j}_0 \cdot \mathbf{v} = (\sum_i v_i)/\sqrt{N}$.
- *Take my word for it that one can select exactly* $(N-1)$ *orthonormal N-dimensional vectors that are orthogonal to* $\mathbf{j}_0$ *(Reason: the subspace of N-dimensional vectors orthogonal to* j_0 *is an* $(N-1)$*-dimensional subspace. One can choose an orthonormal basis for this subspace; these are* $N-1$ *orthonormal vectors.); each of these is an eigenvector of* $\mathbf{J}$ *with eigenvalue 0.*

Thus, the eigenvectors of $\mathbf{J}$ are j_0, with eigenvalue N; and $(N-1)$ other vectors, each with eigenvalue 0.

Now, the solutions of $\frac{d}{dt}\mathbf{v} = \mathbf{Cv}$ are given by

$$\mathbf{v}(t) = \sum_{i=0}^{N-1} \left(\frac{v_i^+(0)}{\sqrt{2}} \begin{pmatrix} \mathbf{j}_i \\ \mathbf{j}_i \end{pmatrix} e^{\lambda_i(1+\epsilon)t} + \frac{v_i^-(0)}{\sqrt{2}} \begin{pmatrix} \mathbf{j}_i \\ -\mathbf{j}_i \end{pmatrix} e^{\lambda_i(1-\epsilon)t} \right) \tag{3.45}$$

where $v_i^+(0)$ is the value of $\mathbf{v} \cdot \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{j}_i \\ \mathbf{j}_i \end{pmatrix}$ at $t = 0$, and $v_i^-(0)$ is the value of $\mathbf{v} \cdot \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{j}_i \\ -\mathbf{j}_i \end{pmatrix}$ at $t = 0$. Plugging in the solutions we have found for the eigenvectors and eigenvalues of $\mathbf{J}$, this becomes

$$\mathbf{v}(t) = \frac{v_0^+(0)}{\sqrt{2}} \begin{pmatrix} \mathbf{j}_0 \\ \mathbf{j}_0 \end{pmatrix} e^{N(1+\epsilon)t} + \frac{v_0^-(0)}{\sqrt{2}} \begin{pmatrix} \mathbf{j}_0 \\ -\mathbf{j}_0 \end{pmatrix} e^{N(1-\epsilon)t} + \mathbf{c} \tag{3.46}$$

where $\mathbf{c}$ is a constant vector (since the eigenvectors for $i \neq 0$ have eigenvalue 0, the coefficients of these vectors do not change in time, hence the sum of all terms for $i \neq 0$ is a constant vector). After sufficient time, the exponential growth will cause the first two terms to swamp the constant term, so we can neglect $\mathbf{c}$.

The vector $\begin{pmatrix} \mathbf{j}_0 \\ \mathbf{j}_0 \end{pmatrix}$ represents equal strengths of all synapses, while the vector $\begin{pmatrix} \mathbf{j}_0 \\ -\mathbf{j}_0 \end{pmatrix}$ represents equal strengths for all left-eye synapses, and equal and opposite strengths for all right-eye synapses. Thus, if the two eyes are correlated ($\epsilon > 0$), the sum of the two eyes' strengths grows faster than the difference of their strengths; while if the two eyes are anticorrelated, the difference of the two eyes' strengths grows faster than their sum, meaning that one eye's strengths will grow and the other eye's strengths will shrink. Each eye grows as a unit, all synapses within an eye growing identically. Any variations in synaptic strengths within an eye are incorporated in the constant vector $\mathbf{c}$; these stem from the initial condition and do not change in time. Except for this constant vector, which is negligible, the model behaves just like the two-input model we studied previously, with each eye behaving like one input.

It is not hard to guess that, as we make the correlations more localized within each eye, eigenvectors incorporating variation in synaptic strength within each eye will acquire finite growth rates, and differences between the two-input case and the many-input case will become noticeable.

3.10 Higher-Order Differential Equations

We have spent a lot of time on the first-order differential equation $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$ ("first-order" means that it contains only first derivatives). But what about equations with higher-order derivatives, like the equation for the harmonic oscillator, $\frac{d^2}{dt^2}x = -kx$? These can always be turned into a first-order equation just by increasing the number of variables. For example, for the harmonic oscillator, define $x_0 = x$ and $x_1 = \frac{d}{dt}x$. Then the harmonic oscillator equation can be expressed as $\frac{d}{dt}x_0 = x_1$, $\frac{d}{dt}x_1 = -kx_0$, or

$$\frac{d}{dt}\begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -k & 0 \end{pmatrix}\begin{pmatrix} x_0 \\ x_1 \end{pmatrix} \tag{3.47}$$

More generally, if we had a k^{th}-order equation, $\frac{d^n}{dt^n}x + a_1\frac{d^{n-1}}{dt^{n-1}}x + \ldots + a_{n-1}\frac{d}{dt}x + a_n x = 0$, we could define $x_0 = x$, $x_i = \frac{d^i}{dt^i}x$ for $i = 1, \ldots, n-1$, and obtain the equation

$$\frac{d}{dt}\begin{pmatrix} x_0 \\ x_1 \\ \ldots \\ x_{n-2} \\ x_{n-1} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \ldots & -a_1 \end{pmatrix}\begin{pmatrix} x_0 \\ x_1 \\ \ldots \\ x_{n-2} \\ x_{n-1} \end{pmatrix} \tag{3.48}$$

Thus, if an equation involves n^{th}-order derivatives, we just multiply the number of variables by n – defining one variable for each derivative up to $n-1$ – and the vector equation in terms of these variables is a first-order equation. So if we can understand first-order vector equations, we have a completely general understanding of linear differential equations. (Of course, at the moment we only understand first-order equations $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$ for *symmetric* $\mathbf{M}$, which will not let us solve Eqs. 3.47-3.48. But be patient, we will get to general matrices eventually.)

3.11 Inhomogeous Equations

So far we've only dealt with equations of the form $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$, which are called homogeneous first-order linear differential equations. But a first-order linear differential equation may also have a driving term:

$$\frac{d}{dt}\mathbf{v}(t) = \mathbf{M}\mathbf{v}(t) + \mathbf{h}(t) \tag{3.49}$$

This is called an inhomogeneous first-order linear differential equation. It's easy to extend our framework to this case.

First, recall from Eq. 0.50 that the solution to the equation

$$\frac{d}{dt}v(t) = mv(t) + h(t) \tag{3.50}$$

is given by

$$v(t) = e^{mt}\left[\int_0^t ds\, e^{-ms}h(s) + v(0)\right] \tag{3.51}$$

or

$$v(t) = v(0)e^{mt} + \int_0^t ds\, e^{m(t-s)}h(s) \tag{3.52}$$

Recall also that if $h(t) \equiv h$ is a time-independent constant, then the solution of Eq. 3.51 becomes

$$v(t) = v(0)e^{mt} - (h/m)(1 - e^{mt}), \qquad m \neq 0 \tag{3.53}$$

$$v(t) = v(0) + ht, \qquad m = 0 \tag{3.54}$$

where $v^{\text{FP}} = -h/m$ is the fixed point of Eq. 3.50, defined as the point where $\frac{d}{dt}v(t) = 0$.

Now assume that $\mathbf{M}$ has a complete basis of N eigenvectors $\mathbf{e}_i$ with eigenvalues λ_i. Express $\mathbf{v}$ and $\mathbf{h}$ in this basis: $\mathbf{v}(t) = \sum_i v_i(t)\mathbf{e}_i$, $\mathbf{h}(t) = \sum_i h_i(t)\mathbf{e}_i$, where $v_i(t) = \mathbf{e}_i^{\text{T}}\mathbf{v}(t)$ and $h_i(t) = \mathbf{e}_i^{\text{T}}\mathbf{h}(t)$. Then Eq. 3.49 becomes a set of N independent 1-dimensional equations,

$$\frac{d}{dt}v_i(t) = \lambda_i v_i + h_i(t) \quad \text{ for } i = 0, \ldots, N-1 \tag{3.55}$$

Each 1-d equation has the solution Eq. 3.51, so we can write the general solution for $\mathbf{v}(t)$:

$$\mathbf{v}(t) = \sum_i v_i(t)\mathbf{e}_i = \sum_i \mathbf{e}_i e^{\lambda_i t}\left[\int_0^t ds\, e^{-\lambda_i s}h_i(s) + v_i(0)\right] \tag{3.56}$$

If $\mathbf{h}$ is a constant and none of the λ_i are zero, the solution becomes

$$\mathbf{v}(t) = \sum_i \mathbf{e}_i\left[v_i(0)e^{\lambda_i t} + v_i^{\text{FP}}(1 - e^{\lambda_i t})\right] \tag{3.57}$$

where the $v_i^{\text{FP}} = -h_i/\lambda_i$ are the components in the eigenvector basis of $\mathbf{v}^{\text{FP}} = -\mathbf{M}^{-1}\mathbf{h}$.

Problem 3.13 *Let's return finally to the case of activity in a linear network of neurons (section 3.2.2). Recall that our equation (Eq. 3.22) is*

$$\tau\frac{d}{dt}\mathbf{b} = -(\mathbf{1} - \mathbf{B})\mathbf{b} + \mathbf{h} \tag{3.58}$$

Let $\mathbf{e}_i$ be the eigenvectors of $\mathbf{B}$, with eigenvalues λ_i: $\mathbf{B}\mathbf{e}_i = \lambda_i$. Each eigenvector represents some pattern of output-cell activity that reproduces itself, multiplied by a constant, under the connectivity $\mathbf{B}$. Show that the $\mathbf{e}_i$ are also eigenvectors of $-(\mathbf{1}-\mathbf{B})$, and determine the corresponding eigenvalues of $-(\mathbf{1} - \mathbf{B})$. Assume $\mathbf{B}$ has a complete orthonormal basis of eigenvectors, and that none of its eigenvalues is equal to 1; show that this means that $\mathbf{1}-\mathbf{B}$ is invertible. Thus, show that the solution to Eq. 3.58 is (see Eq. 3.57):

$$\mathbf{b}(t) = \sum_i \mathbf{e}_i\left[b_i(0)e^{-(1-\lambda_i)t/\tau} + b_i^{\text{FP}}\left(1 - e^{-(1-\lambda_i)t/\tau}\right)\right] \tag{3.59}$$

Here, $\mathbf{b}^{\mathrm{FP}} = (\mathbf{1} - \mathbf{B})^{-1}\mathbf{h}$, in agreement with Eq. 1.13.

Expand $\mathbf{h}$ in the eigenvector basis, $\mathbf{h} = \sum_j h_j \mathbf{e}_j$ for $h_j = \mathbf{e}_j^{\mathrm{T}}\mathbf{h}$; and use Eq. 3.37 to write $(\mathbf{1} - \mathbf{B})^{-1}$ in terms of the $\mathbf{e}_i$ and λ_i. Use these to find $\mathbf{b}^{\mathrm{FP}}$ in terms of the h_i, λ_i, and $\mathbf{e}_i$. Thus arrive finally at the equation

$$\mathbf{b}(t) = \sum_i \mathbf{e}_i \left[b_i(0) e^{-(1-\lambda_i)t/\tau} + \frac{h_i}{1-\lambda_i}\left(1 - e^{-(1-\lambda_i)t/\tau}\right)\right] \qquad (3.60)$$

This ends the problem, but here are some comments, which you should verify for yourself. What does this equation tell us?

- *The equation is* stable*, and flows to the fixed point as $t \to \infty$, if $\lambda_i < 1$ for all i, that is, if the connectivity matrix $\mathbf{B}$ has no eigenvalues greater than or equal to 1.*

- *The fixed point is found by taking the component of the input, $\mathbf{h}$, along the i^{th} eigenvector, and dividing it by $1 - \lambda_i$ (which is the corresponding eigenvalue of $\mathbf{1} - \mathbf{B}$). Thus, if the equation is stable ($\lambda_i < 1$ for all i), then the effect of the connectivity is as follows: given a fixed input, sustained for a long time, then the connectivity takes the component of the input along each eigenvector, $\mathbf{e}_i$, and multiplies that component by $\frac{1}{1-\lambda_i}$. In particular, in this case, eigenvectors $\mathbf{e}_i$ with eigenvalues $\lambda_i > 0$ (but with $\lambda_i < 1$) are amplified relative to the corresponding input h_i, while those with eigenvalues $\lambda_i < 0$ are diminished in size relative to the input h_i; and the eigenvector with eigenvalue closest to 1 is most amplified (or least diminished).*

- *The component in the direction of any eigenvector with eigenvalue $\lambda_i > 1$ is* unstable*: as $t \to \infty$, the component along such a direction becomes exponentially large. This corresponds to the intuition that linear feedback with gain greater than one is unstable. In this case, the gain is the amplification under $\mathbf{B}$ of a* pattern *of activity, rather than the size of the feedback onto any particular cell.*

Exercise 3.11 *We can further understand the fixed point as follow. You might recall that, for a number x with $|x| < 1$, one can write*

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \ldots = \sum_{i=0}^{\infty} x^i \qquad (3.61)$$

One way to see that this is true is to multiply the right side by $1 - x$; you should convince yourself that x times the right side is just the right side minus 1, from which it follows that $1 - x$ times the right side equals 1. More formally, you can obtain this as the Taylor series of $\frac{1}{1-x}$, expanded about $x = 0$. The condition $|x| < 1$ is required, because otherwise the series on the right side does not converge, that is, it does not go to a finite sum as the number of terms goes to infinity.

One can formally write the same expression for $\mathbf{B}$: if, for all i, $|\lambda_i| < 1$, then

$$(\mathbf{1} - \mathbf{B})^{-1} = \mathbf{1} + \mathbf{B} + \mathbf{B}^2 + \mathbf{B}^3 + \ldots = \sum_{i=0}^{\infty} \mathbf{B}^i \qquad (3.62)$$

One way to see that this is true is to go to the basis in which $\mathbf{B}$ is diagonal, that is, in which $\mathbf{B} = \begin{pmatrix} \lambda_0 & 0 & \ldots & 0 \\ 0 & \lambda_1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_{N-1} \end{pmatrix}$. In this basis, Eq. 3.62 becomes N independent equations just like

Eq. 3.61, one equation for each eigenvalue (make sure you understand this); so if each eigenvalue satisfies $|\lambda_i| < 1$, each of these N equations is valid, so Eq. 3.62 is true in that basis. But this is a matrix equation, so it is basis-independent.

Using Eq. 3.62, we can rewrite the fixed point as

$$\mathbf{b}^{\mathrm{FP}} = \mathbf{h} + \mathbf{B}\mathbf{h} + \mathbf{B}^2\mathbf{h} + \mathbf{B}^3\mathbf{h} + \ldots = \sum_{i=0}^{\infty} \mathbf{B}^i\mathbf{h} \tag{3.63}$$

That is, the fixed point is the activity pattern corresponding to the input, plus the input transformed once by the connectivity, plus the input transformed twice by the connectivity, and so on. This makes intuitive sense: to be a fixed point, the output activity must not change, while we leave the input clamped on. But the input propagates through the connectivity, and the output of this is added to the continuing input. Then this sum is propagated through the connectivity, and added to the continuing input ... and so on. The fixed point is the point at which this process converges, so that the activity can remain unchanging as we keep propagating the cortical activity through $\mathbf{B}$ *and re-adding the input. That is, it is the point at which* $\mathbf{h} + \mathbf{B}\mathbf{b}^{\mathrm{FP}} = \mathbf{b}^{\mathrm{FP}}$ *(which is another way of writing* $\mathbf{b}^{\mathrm{FP}} = (\mathbf{1} - \mathbf{B})^{-1}\mathbf{h}$*). This process in turn can only converge if all activity patterns are multiplied at each iteration by something with absolute value less than 1. (If* $\lambda_i \leq -1$*, this series description of the fixed point is not valid, but the statement* $\mathbf{b}^{\mathrm{FP}} = \sum_i \frac{h_i}{1-\lambda_i}\mathbf{e}_i$ *is still correct).*

Yet another way to understand Eq. 3.63 is to write it in components in the original (cellular) basis, in which b_i^{FP} *is the activity of the* i^{th} *cell at the fixed point. Equation 3.63 becomes*

$$\mathbf{b}_i^{\mathrm{FP}} = h_i + \sum_j B_{ij}h_j + \sum_{k,j} B_{ik}B_{kj}h_j + \ldots \tag{3.64}$$

In this form, the equation can be interpreted as follows: clamping on the input h_i *to each cell* i*, and letting the network respond until it reaches the fixed point (steady-state activity), one finds that the steady state activity of the* i^{th} *cell is the direct input* h_i *to cell* i*, plus the input* h_j *to each other cell* j *propagated through one synapse* B_{ij} *to* i*, plus the input to each other cell* j *propagated through two synapses to* i*, and so on – the sum of all possible polysynaptic intracortical contributions of every length from 0 to* ∞*.*

3.12 Summary

A linear differential equation, $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$, becomes very simple when the matrix $\mathbf{M}$ is diagonal — the equation then becomes a set of independent, one-dimensional differential equations. If $\mathbf{M}$ is not diagonal, this may just be because we're working in the wrong coordinates — there may be a basis in which $\mathbf{M}$ is diagonal. To solve $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v}$, we wish to find such a basis.

An eigenvector of $\mathbf{M}$ is a vector $\mathbf{e}_i$ such that $\mathbf{M}\mathbf{e}_i = \lambda_i\mathbf{e}_i$ for some scalar λ_i. The basis in which $\mathbf{M}$ becomes diagonal is the basis of eigenvectors of $\mathbf{M}$. A symmetric matrix always has a complete orthonormal basis of eigenvectors, with real eigenvalues. Given a complete orthonormal basis of eigenvectors, it is easy to solve explicitly for $\mathbf{v}(t)$. The same methods extend easily to solving inhomogeneous equations of the form $\frac{d}{dt}\mathbf{v} = \mathbf{M}\mathbf{v} + \mathbf{h}(t)$.