

Appendix A. Linear Algebra

In this appendix, we gather together some useful properties and identities involving matrices and determinants. This is not intended to be an introductory tutorial, and it is assumed that the reader is already familiar with basic linear algebra. For some results, we indicate how to prove them, whereas in more complex cases we leave the interested reader to refer to standard textbooks on the subject. In all cases, we assume that inverses exist and that matrix dimensions are such that the formulae are correctly defined. A comprehensive discussion of linear algebra can be found in Golub and Van Loan (1996), and an extensive collection of matrix properties is given by Lütkepohl (1996). Matrix derivatives are discussed in Magnus and Neudecker (1999).

A.1. Matrix Identities

A matrix \mathbf{A} has elements A_{ij} where i indexes the rows and j indexes the columns. We use \mathbf{I}_N to denote the $N \times N$ identity matrix (also called the unit matrix), and if there is no ambiguity over dimensionality, we simply use \mathbf{I} . The transpose matrix \mathbf{A}^T has elements $(\mathbf{A}^T)_{ij} = A_{ji}$. From the definition of a transpose, we have

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T, \quad (\text{A.1})$$

which can be verified by writing out the indices. The inverse of \mathbf{A} , denoted \mathbf{A}^{-1} , satisfies

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \quad (\text{A.2})$$

Because $\mathbf{ABB}^{-1}\mathbf{A}^{-1} = \mathbf{I}$, we have

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}. \quad (\text{A.3})$$

Also we have

$$(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T, \quad (\text{A.4})$$

which is easily proven by taking the transpose of (A.2) and applying (A.1).

A useful identity involving matrix inverses is the following:

$$(\mathbf{P}^{-1} + \mathbf{B}^T \mathbf{R}^{-1} \mathbf{B})^{-1} \mathbf{B}^T \mathbf{R}^{-1} = \mathbf{P} \mathbf{B}^T (\mathbf{B} \mathbf{P} \mathbf{B}^T + \mathbf{R})^{-1}, \quad (\text{A.5})$$

which is easily verified by right-multiplying both sides by $(\mathbf{B} \mathbf{P} \mathbf{B}^T + \mathbf{R})$. Suppose that \mathbf{P} has dimensionality $N \times N$ and that \mathbf{R} has dimensionality $M \times M$, so that \mathbf{B} is $M \times N$. Then if $M \ll N$, it will be much cheaper to evaluate the right-hand side of (A.5) than the left-hand side. A special case that sometimes arises is

$$(\mathbf{I} + \mathbf{A} \mathbf{B})^{-1} \mathbf{A} = \mathbf{A} (\mathbf{I} + \mathbf{B} \mathbf{A})^{-1}. \quad (\text{A.6})$$

Another useful identity involving inverses is the following:

$$(\mathbf{A} + \mathbf{B} \mathbf{D}^{-1} \mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B} (\mathbf{D} + \mathbf{C} \mathbf{A}^{-1} \mathbf{B})^{-1} \mathbf{C} \mathbf{A}^{-1}, \quad (\text{A.7})$$

which is known as the *Woodbury identity*. It can be verified by multiplying both sides by $(\mathbf{A} + \mathbf{B} \mathbf{D}^{-1} \mathbf{C})$. This is useful, for instance, when \mathbf{A} is large and diagonal and hence easy to invert, and when \mathbf{B} has many rows but few columns (and conversely for \mathbf{C}), so that the right-hand side is much cheaper to evaluate than the left-hand side.

A set of vectors $\{\mathbf{a}_1, \dots, \mathbf{a}_N\}$ is said to be *linearly independent* if the relation $\sum_n \alpha_n \mathbf{a}_n = 0$ holds only if all $\alpha_n = 0$. This implies that none of the vectors can be expressed as a linear combination of the remainder. The rank of a matrix is the maximum number of linearly independent rows (or equivalently the maximum number of linearly independent columns).

A.2. Traces and Determinants

Square matrices have traces and determinants. The trace $\text{Tr}(\mathbf{A})$ of a matrix \mathbf{A} is defined as the sum of the elements on the leading diagonal. By writing out the indices, we see that

$$\text{Tr}(\mathbf{A} \mathbf{B}) = \text{Tr}(\mathbf{B} \mathbf{A}). \quad (\text{A.8})$$

By applying this formula multiple times to the product of three matrices, we see that

$$\text{Tr}(\mathbf{A} \mathbf{B} \mathbf{C}) = \text{Tr}(\mathbf{C} \mathbf{A} \mathbf{B}) = \text{Tr}(\mathbf{B} \mathbf{C} \mathbf{A}), \quad (\text{A.9})$$

which is known as the *cyclic* property of the trace operator. It clearly extends to the product of any number of matrices. The determinant $|\mathbf{A}|$ of an $N \times N$ matrix \mathbf{A} is defined by

$$|\mathbf{A}| = \sum (\pm 1) A_{1i_1} A_{2i_2} \cdots A_{Ni_N} \quad (\text{A.10})$$

in which the sum is taken over all products consisting of precisely one element from each row and one element from each column, with a coefficient $+1$ or -1 according to whether the permutation $i_1 i_2 \dots i_N$ is even or odd, respectively. Note that $|\mathbf{I}| = 1$,

and that the determinant of a diagonal matrix is given by the product of the elements on the leading diagonal. Thus, for a 2×2 matrix, the determinant takes the form

$$|\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \quad (\text{A.11})$$

The determinant of a product of two matrices is given by

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}| \quad (\text{A.12})$$

as can be shown from (A.10). Also, the determinant of an inverse matrix is given by

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}, \quad (\text{A.13})$$

which can be shown by taking the determinant of (A.2) and applying (A.12).

If \mathbf{A} and \mathbf{B} are matrices of size $N \times M$, then

$$|\mathbf{I}_N + \mathbf{AB}^T| = |\mathbf{I}_M + \mathbf{A}^T\mathbf{B}|. \quad (\text{A.14})$$

A useful special case is

$$|\mathbf{I}_N + \mathbf{ab}^T| = 1 + \mathbf{a}^T\mathbf{b} \quad (\text{A.15})$$

where \mathbf{a} and \mathbf{b} are N -dimensional column vectors.

A.3. Matrix Derivatives

Sometimes we need to consider derivatives of vectors and matrices with respect to scalars. The derivative of a vector \mathbf{a} with respect to a scalar x is a vector whose components are given by

$$\left(\frac{\partial \mathbf{a}}{\partial x} \right)_i = \frac{\partial a_i}{\partial x} \quad (\text{A.16})$$

with an analogous definition for the derivative of a matrix. Derivatives with respect to vectors and matrices can also be defined, for instance

$$\left(\frac{\partial x}{\partial \mathbf{a}} \right)_i = \frac{\partial x}{\partial a_i} \quad (\text{A.17})$$

and similarly

$$\left(\frac{\partial \mathbf{a}}{\partial \mathbf{b}} \right)_{ij} = \frac{\partial a_i}{\partial b_j}. \quad (\text{A.18})$$

The following is easily proven by writing out the components:

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^T \mathbf{a}) = \frac{\partial}{\partial \mathbf{x}} (\mathbf{a}^T \mathbf{x}) = \mathbf{a}. \quad (\text{A.19})$$

Similarly

$$\frac{\partial}{\partial x} (\mathbf{A}\mathbf{B}) = \frac{\partial \mathbf{A}}{\partial x} \mathbf{B} + \mathbf{A} \frac{\partial \mathbf{B}}{\partial x}. \quad (\text{A.20})$$

The derivative of the inverse of a matrix can be expressed as

$$\frac{\partial}{\partial x} (\mathbf{A}^{-1}) = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{-1} \quad (\text{A.21})$$

as can be shown by differentiating the equation $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ using (A.20) and then right-multiplying by \mathbf{A}^{-1} . Also

$$\frac{\partial}{\partial x} \ln |\mathbf{A}| = \text{Tr} \left(\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \right), \quad (\text{A.22})$$

which we shall prove later. If we choose x to be one of the elements of \mathbf{A} , we have

$$\frac{\partial}{\partial A_{ij}} \text{Tr}(\mathbf{A}\mathbf{B}) = B_{ji} \quad (\text{A.23})$$

as can be seen by writing out the matrices using index notation. We can write this result more compactly in the form

$$\frac{\partial}{\partial \mathbf{A}} \text{Tr}(\mathbf{A}\mathbf{B}) = \mathbf{B}^T. \quad (\text{A.24})$$

With this notation, we have the following properties:

$$\frac{\partial}{\partial \mathbf{A}} \text{Tr}(\mathbf{A}^T \mathbf{B}) = \mathbf{B}, \quad (\text{A.25})$$

$$\frac{\partial}{\partial \mathbf{A}} \text{Tr}(\mathbf{A}) = \mathbf{I}, \quad (\text{A.26})$$

$$\frac{\partial}{\partial \mathbf{A}} \text{Tr}(\mathbf{A}\mathbf{B}\mathbf{A}^T) = \mathbf{A}(\mathbf{B} + \mathbf{B}^T), \quad (\text{A.27})$$

which can again be proven by writing out the matrix indices. We also have

$$\frac{\partial}{\partial \mathbf{A}} \ln |\mathbf{A}| = (\mathbf{A}^{-1})^T, \quad (\text{A.28})$$

which follows from (A.22) and (A.24).

A.4. Eigenvectors

For a square matrix \mathbf{A} of size $M \times M$, the eigenvector equation is defined by

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i \quad (\text{A.29})$$

for $i = 1, \dots, M$, where \mathbf{u}_i is an *eigenvector* and λ_i is the corresponding *eigenvalue*. This can be viewed as a set of M simultaneous homogeneous linear equations, and the condition for a solution is that

$$|\mathbf{A} - \lambda_i \mathbf{I}| = 0, \quad (\text{A.30})$$

which is known as the *characteristic equation*. Because this is a polynomial of order M in λ_i , it must have M solutions (though these need not all be distinct). The rank of \mathbf{A} is equal to the number of non-zero eigenvalues.

Of particular interest are symmetric matrices, which arise as covariance matrices, kernel matrices, and Hessians. Symmetric matrices have the property that $A_{ij} = A_{ji}$, or equivalently $\mathbf{A}^T = \mathbf{A}$. The inverse of a symmetric matrix is also symmetric, as can be seen by taking the transpose of $\mathbf{A}^{-1} \mathbf{A} = \mathbf{I}$ and using $\mathbf{A} \mathbf{A}^{-1} = \mathbf{I}$ together with the symmetry of \mathbf{I} .

In general, the eigenvalues of a matrix are complex numbers, but for symmetric matrices, the eigenvalues λ_i are real. This can be seen by first left-multiplying (A.29) by $(\mathbf{u}_i^*)^T$, where \star denotes the complex conjugate, to give

$$(\mathbf{u}_i^*)^T \mathbf{A} \mathbf{u}_i = \lambda_i (\mathbf{u}_i^*)^T \mathbf{u}_i. \quad (\text{A.31})$$

Next we take the complex conjugate of (A.29) and left-multiply by \mathbf{u}_i^T to give

$$\mathbf{u}_i^T \mathbf{A} \mathbf{u}_i^* = \lambda_i^* \mathbf{u}_i^T \mathbf{u}_i^* \quad (\text{A.32})$$

where we have used $\mathbf{A}^* = \mathbf{A}$ because we are considering only real matrices \mathbf{A} . Taking the transpose of the second of these equations and using $\mathbf{A}^T = \mathbf{A}$, we see that the left-hand sides of the two equations are equal and hence that $\lambda_i^* = \lambda_i$, and so λ_i must be real.

The eigenvectors \mathbf{u}_i of a real symmetric matrix can be chosen to be orthonormal (i.e., orthogonal and of unit length) so that

$$\mathbf{u}_i^T \mathbf{u}_j = I_{ij} \quad (\text{A.33})$$

where I_{ij} are the elements of the identity matrix \mathbf{I} . To show this, we first left-multiply (A.29) by \mathbf{u}_j^T to give

$$\mathbf{u}_j^T \mathbf{A} \mathbf{u}_i = \lambda_i \mathbf{u}_j^T \mathbf{u}_i \quad (\text{A.34})$$

and hence, by exchanging the indices, we have

$$\mathbf{u}_i^T \mathbf{A} \mathbf{u}_j = \lambda_j \mathbf{u}_i^T \mathbf{u}_j. \quad (\text{A.35})$$

We now take the transpose of the second equation and make use of the symmetry property $\mathbf{A}^T = \mathbf{A}$, and then subtract the two equations to give

$$(\lambda_i - \lambda_j) \mathbf{u}_i^T \mathbf{u}_j = 0. \quad (\text{A.36})$$

Hence, for $\lambda_i \neq \lambda_j$, we have $\mathbf{u}_i^T \mathbf{u}_j = 0$ so that \mathbf{u}_i and \mathbf{u}_j are orthogonal. If the two eigenvalues are equal, then any linear combination $\alpha \mathbf{u}_i + \beta \mathbf{u}_j$ is also an eigenvector

with the same eigenvalue, so we can select one linear combination arbitrarily, and then choose the second to be orthogonal to the first (it can be shown that the degenerate eigenvectors are never linearly dependent). Hence, the eigenvectors can be chosen to be orthogonal, and by normalizing can be set to unit length. Because there are M eigenvalues, the corresponding M orthogonal eigenvectors form a complete set and so any M -dimensional vector can be expressed as a linear combination of the eigenvectors.

We can take the eigenvectors \mathbf{u}_i to be the columns of an $M \times M$ matrix \mathbf{U} , which from orthonormality satisfies

$$\mathbf{U}^T \mathbf{U} = \mathbf{I}. \quad (\text{A.37})$$

Such a matrix is said to be *orthogonal*. Interestingly, the rows of this matrix are also orthogonal, so that $\mathbf{U}\mathbf{U}^T = \mathbf{I}$. To show this, note that (A.37) implies $\mathbf{U}^T \mathbf{U} \mathbf{U}^{-1} = \mathbf{U}^{-1} = \mathbf{U}^T$ and so $\mathbf{U}\mathbf{U}^{-1} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$. Using (A.12), it also follows that $|\mathbf{U}| = 1$.

The eigenvector equation (A.29) can be expressed in terms of \mathbf{U} in the form

$$\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{\Lambda} \quad (\text{A.38})$$

where $\mathbf{\Lambda}$ is an $M \times M$ diagonal matrix whose diagonal elements are given by the eigenvalues λ_i .

If we consider a column vector \mathbf{x} that is transformed by an orthogonal matrix \mathbf{U} to give a new vector

$$\tilde{\mathbf{x}} = \mathbf{U}\mathbf{x} \quad (\text{A.39})$$

then the length of the vector is preserved because

$$\tilde{\mathbf{x}}^T \tilde{\mathbf{x}} = \mathbf{x}^T \mathbf{U}^T \mathbf{U} \mathbf{x} = \mathbf{x}^T \mathbf{x} \quad (\text{A.40})$$

and similarly the angle between any two such vectors is preserved because

$$\tilde{\mathbf{x}}^T \tilde{\mathbf{y}} = \mathbf{x}^T \mathbf{U}^T \mathbf{U} \mathbf{y} = \mathbf{x}^T \mathbf{y}. \quad (\text{A.41})$$

Thus, multiplication by \mathbf{U} can be interpreted as a rigid rotation of the coordinate system.

From (A.38), it follows that

$$\mathbf{U}^T \mathbf{A} \mathbf{U} = \mathbf{\Lambda} \quad (\text{A.42})$$

and because $\mathbf{\Lambda}$ is a diagonal matrix, we say that the matrix \mathbf{A} is *diagonalized* by the matrix \mathbf{U} . If we left-multiply by \mathbf{U} and right-multiply by \mathbf{U}^T , we obtain

$$\mathbf{A} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T. \quad (\text{A.43})$$

Taking the inverse of this equation and using (A.3) together with $\mathbf{U}^{-1} = \mathbf{U}^T$, we have

$$\mathbf{A}^{-1} = \mathbf{U} \mathbf{\Lambda}^{-1} \mathbf{U}^T. \quad (\text{A.44})$$

These last two equations can also be written in the form

$$\mathbf{A} = \sum_{i=1}^M \lambda_i \mathbf{u}_i \mathbf{u}_i^T \quad (\text{A.45})$$

$$\mathbf{A}^{-1} = \sum_{i=1}^M \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T. \quad (\text{A.46})$$

If we take the determinant of (A.43) and use (A.12), we obtain

$$|\mathbf{A}| = \prod_{i=1}^M \lambda_i. \quad (\text{A.47})$$

Similarly, taking the trace of (A.43), and using the cyclic property (A.8) of the trace operator together with $\mathbf{U}^T \mathbf{U} = \mathbf{I}$, we have

$$\text{Tr}(\mathbf{A}) = \sum_{i=1}^M \lambda_i. \quad (\text{A.48})$$

We leave it as an exercise for the reader to verify (A.22) by making use of the results (A.33), (A.45), (A.46), and (A.47).

A matrix \mathbf{A} is said to be *positive definite*, denoted by $\mathbf{A} \succ 0$, if $\mathbf{w}^T \mathbf{A} \mathbf{w} > 0$ for all non-zero values of the vector \mathbf{w} . Equivalently, a positive definite matrix has $\lambda_i > 0$ for all of its eigenvalues (as can be seen by setting \mathbf{w} to each of the eigenvectors in turn and noting that an arbitrary vector can be expanded as a linear combination of the eigenvectors). Note that having all positive elements does not necessarily mean that a matrix is that positive definite. For example, the matrix

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad (\text{A.49})$$

has eigenvalues $\lambda_1 \simeq 5.37$ and $\lambda_2 \simeq -0.37$. A matrix is said to be *positive semidefinite* if $\mathbf{w}^T \mathbf{A} \mathbf{w} \geq 0$ holds for all values of \mathbf{w} , which is denoted $\mathbf{A} \succeq 0$ and is equivalent to $\lambda_i \geq 0$.

The condition number of a matrix is given by

$$\text{CN} = \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{1/2} \quad (\text{A.50})$$

where λ_{\max} is the largest eigenvalue and λ_{\min} is the smallest eigenvalue.