# Note 4: Special Classes of Matrices

Math 198: Math for Machine Learning

## 1 Normal Matrices

### 1.1 Definition

We start our discussion of special classes of matrices with the introduction of the *normal matrix*. A normal matrix is any matrix $\mathbf{A}$ which commutes with its transpose:

$$\mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}^\top$$

This deceptively simple property actually has very important ramifications. Notably, the Spectral Theorem from Note 3 applies to normal matrices, as we will see after considering a few other classes of matrices. For now, just remember that a normal matrix commutes with its transpose.

## 2 Orthogonal (Unitary) Matrices

### 2.1 Definition

*Orthogonal* matrices are a subclass of sorts of normal matrices; every orthogonal matrix is normal, but the converse is not always true. Formally, a orthogonal matrix is any matrix $\mathbf{Q}$ whose transpose is its inverse:

$$\mathbf{Q}^\top = \mathbf{Q}^{\mathbf{-1}}$$

Note that this implies that $\mathbf{Q}$ is normal:

$$\mathbf{Q}^\top \mathbf{Q} = \mathbf{Q}^{\mathbf{-1}}\mathbf{Q} = \mathbf{I} = \mathbf{Q}\mathbf{Q}^{\mathbf{-1}} = \mathbf{Q}\mathbf{Q}^\top$$

In complex vector spaces, a matrix whose inverse is equal to its conjugate transpose is known as *unitary*. Because the conjugate of a real number is itself, in real vector spaces, unitary matrices are exactly the same as orthogonal matrices, and we will use the two terms interchangeably.

### 2.2 Properties

Since $\mathbf{Q}^\top = \mathbf{Q}^{\mathbf{-1}}$, we have that $\det(\mathbf{Q}) = 1$. We can also show the stronger result that all eigenvalues of $\mathbf{Q}$ are $\pm 1$. Let $\lambda$ be an eigenvalue of $\mathbf{Q}$ with corresponding eigenvector $\mathbf{v}$. Then $||\mathbf{Qv}||^2 = |\lambda|^2||\mathbf{v}||^2$. Rewriting the left side in terms of the inner product, we get $||\mathbf{Qv}||^2 = \langle \mathbf{Qv}, \mathbf{Qv}\rangle = \mathbf{v}^\top \mathbf{Q}^\top \mathbf{Qv} = \mathbf{v}^\top \mathbf{v} = ||\mathbf{v}||^2$. So, $||\mathbf{v}||^2 = |\lambda|^2||\mathbf{v}||^2$; this implies $|\lambda| = 1$, and so $\lambda = \pm 1$. Because all of the eigenvalues of $\mathbf{Q}$ are 1, orthogonal matrices can be though of as matrices which rotate or reflect space rather than scaling it. Also note that in the process of the previous proof, we stumbled upon the result $||\mathbf{Qv}||^2 = ||\mathbf{v}||^2$; this is true in general for orthogonal matrices.

Orthogonal matrices have one hugely important property lurking behind the previous results. Notably, we wrote that $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$ when proving the normality of orthogonal matrices. But this implies that the dot product of the $i$-th and $j$-th columns of $\mathbf{Q}$ is $\mathbf{I}_{ij}$. Formally,

$$\langle \mathbf{Q_i}, \mathbf{Q_j}\rangle = (\mathbf{Q}^\top \mathbf{Q})_{ij} = \mathbf{I}_{ij} = \delta_{ij}$$

In other words, the columns of $\mathbf{Q}$ are pairwise orthonormal. Therefore, they form an orthonormal basis for $\mathbb{R}^n$.

# 3 Diagonal Matrices

## 3.1 Definition and Properties

A *diagonal* matrix $\mathbf{D}$ is a matrix whose only non-zero elements are on its diagonal. Note that this immediately implies that $\mathbf{D}^\top = \mathbf{D}$, and so diagonal matrices are normal. Diagonal matrices actually commute with any other diagonal matrix, and not just themselves. Additionally, it is not much work to see that the eigenvectors of a diagonal matrix are the basis vectors it is defined with respect to, and the eigenvalues are the elements along the diagonal. The determinant is therefore the product of the elements along the diagonal. Additionally, there are extremely simple formulas for the inverse of a diagonal matrix, as well as the sum and product of two diagonal matrices:

$$\begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix}^{-1} = \begin{bmatrix} a_1^{-1} & 0 & 0 \\ 0 & a_2^{-1} & 0 \\ 0 & 0 & a_3^{-1} \end{bmatrix}$$

$$\begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix} + \begin{bmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix} = \begin{bmatrix} a_1 + b_1 & 0 & 0 \\ 0 & a_2 + b_2 & 0 \\ 0 & 0 & a_3 + b_3 \end{bmatrix}$$

$$\begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix} \begin{bmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix} = \begin{bmatrix} a_1 b_1 & 0 & 0 \\ 0 & a_2 b_2 & 0 \\ 0 & 0 & a_3 b_3 \end{bmatrix}$$

Therefore, it can be desirable to express matrices in a similar diagonal form when trying to perform computations such as finding the determinant or eigenvalues, taking large powers or inverses, or multiplying matrices together. The *Spectral Theorem* will determine exactly when doing so is possible.

# 4 Spectral Theorem

## 4.1 Similarity and Diagonalization

Recall that two matrices $\mathbf{X}, \mathbf{Y}$ are similar if there exists an invertible matrix $\mathbf{P}$ such that

$$\mathbf{X} = \mathbf{P}\mathbf{Y}\mathbf{P}^{\mathbf{-1}}$$

Similarity is a very important notion because similar matrices share the same rank, determinant, trace, and eigenvalues. This is because similar matrices represent the same linear operation, just with respect to different bases. Therefore, if we can establish that a matrix is similar to another, simpler matrix, we can derive information about the more complex matrix by studying the simpler one.

A matrix $\mathbf{Z}$ is *diagonalizable* if it is similar to a diagonal matrix. That is, there exists an invertible matrix $\mathbf{P}$ and a diagonal matrix $\mathbf{D}$ such that

$$\mathbf{Z} = \mathbf{P}\mathbf{D}\mathbf{P}^{\mathbf{-1}}$$

Because $\mathbf{Z}$ and $\mathbf{D}$ are similar, they share the same eigenvalues; so, the eigenvalues of $\mathbf{Z}$ can be read off the diagonal of $\mathbf{D}$. This in turn implies that the columns of $\mathbf{P}$ are eigenvectors of $\mathbf{Z}$; the action of $\mathbf{P}\mathbf{D}\mathbf{P}^{\mathbf{-1}}$ on a vector $\mathbf{v}$ is to first change the basis which $\mathbf{v}$ is written with respect to to the eigenbasis which spans the eigenvectors of $\mathbf{Z}$ via the change-of-basis matrix $\mathbf{P}^{\mathbf{-1}}$; then, each component of $\mathbf{P}^{-1}\mathbf{v}$ is scaled by the appropriate eigenvalue in $\mathbf{D}$; and finally, $\mathbf{D}\mathbf{P}^{\mathbf{-1}}\mathbf{v}$ is converted back to the original basis by the change-of-basis matrix $\mathbf{P}$. This has the same effect as just applying the original matrix $\mathbf{Z}$ to $\mathbf{v}$. Note that this does not imply that $\mathbf{P}$ is orthogonal, nor does it imply that the eigenvectors of $\mathbf{Z}$ span $\mathbb{R}^n$. The situations in which an arbitrary matrix is diagonalizable are out of scope of this class, and are covered in Math 110; we will only ever encounter diagonalization in the case of one of these special classes of matrices, via the Spectral Theorem.

### 4.2 Normal Matrices

In the case of normal matrices, the statement of the Spectral Theorem is as follows:

A matrix $\mathbf{A}$ is normal if and only if it is unitarily diagonalizable.

A matrix $\mathbf{A}$ is *unitarily diagonalizable* if there exists a unitary (orthogonal) matrix $\mathbf{Q}$ and a diagonal matrix $\mathbf{\Lambda}$ such that

$$\mathbf{A} = \mathbf{Q\Lambda Q^{-1}} = \mathbf{Q\Lambda Q}^\top$$

$\mathbf{\Lambda}$ contains the eigenvalues of $\mathbf{A}$ on its diagonal, and $\mathbf{Q}$ contains the eigenvectors of $\mathbf{A}$ as its columns. The eigenvectors of $\mathbf{A}$ therefore form an orthonormal basis for $\mathbf{R}^n$, as they are the columns of an orthogonal matrix. The form $\mathbf{Q\Lambda Q^{-1}}$ is known as the *eigendecomposition* of $\mathbf{A}$.

The proof of the Spectral Theorem relies on a more basic result, the Schur decomposition, a proof of which is not included here but can be found on Wikipedia.[1] By the Schur decomposition, for any square matrix $\mathbf{A}$ we can write $\mathbf{A} = \mathbf{UTU}^\top$, where $\mathbf{U}$ is unitary and $\mathbf{T}$ is *upper-triangular*, which should be familiar to anyone who has solved systems of equations with matrices before. (A matrix is upper-triangular iff it is in row-echelon form, i.e. all entries below the diagonal are 0.) If $\mathbf{A}$ is normal, we can write

$$\mathbf{TT}^\top = \mathbf{U^{-1}AUU^{-1}A}^\top\mathbf{U} = \mathbf{U^{-1}AA}^\top\mathbf{U} = \mathbf{U^{-1}A}^\top\mathbf{AU} = \mathbf{U^{-1}A}^\top\mathbf{UU^{-1}AU} = \mathbf{T}^\top\mathbf{T}$$

So if $\mathbf{A}$ is normal, $\mathbf{T}$ is normal. It suffices now to show that any normal upper-triangular matrix is diagonal. Let $\mathbf{e_i}$ denote the $i$-th standard basis vector. Suppose $\mathbf{T}$ is upper-triangular and normal. Then since $\mathbf{TT}^\top = \mathbf{T}^\top\mathbf{T}$, we have that

$$\langle \mathbf{e_i}, \mathbf{TT}^\top\mathbf{e_i}\rangle = \langle \mathbf{e_i}, \mathbf{T}^\top\mathbf{Te_i}\rangle$$
$$\mathbf{e_i}^\top\mathbf{TT}^\top\mathbf{e_i} = \mathbf{e_i}^\top\mathbf{T}^\top\mathbf{Te_i}$$
$$||\mathbf{T}^\top\mathbf{e_i}||^2 = ||\mathbf{Te_i}||^2$$

This implies that the $i$-th row of $T$ has the same magnitude as the $i$-th column. Consider the first row and column. These share one element $\mathbf{T}_{11}$, and every other element in the first column is zero (since $\mathbf{T}$ is upper-triangular). Therefore, every other element in the first row must be zero as well. This argument can then be extended inductively to every row and column of $\mathbf{T}$. So if $\mathbf{T}$ is normal and upper-triangular, it is diagonal. Therefore, any normal matrix has an eigendecomposition. □

The converse of this statement is easy enough to prove – if $\mathbf{A} = \mathbf{Q\Lambda Q}^\top$, then

$$\mathbf{AA}^\top = \mathbf{Q\Lambda Q}^\top(\mathbf{Q\Lambda Q}^\top)^\top = \mathbf{Q\Lambda Q}^\top\mathbf{Q\Lambda}^\top\mathbf{Q}^\top = \mathbf{Q\Lambda\Lambda}^\top\mathbf{Q}^\top$$

$$= \mathbf{Q\Lambda}^\top\mathbf{\Lambda Q}^\top = \mathbf{Q\Lambda}^\top\mathbf{Q}^\top\mathbf{Q\Lambda Q}^\top = (\mathbf{Q\Lambda Q}^\top)^\top\mathbf{Q\Lambda Q}^\top = \mathbf{A}^\top\mathbf{A}$$

so $\mathbf{A}$ is normal. □

## 5 Symmetric Matrices

### 5.1 Definition

A matrix $\mathbf{S}$ is *symmetric* if it is equal to its own transpose; that is, $\mathbf{S} = \mathbf{S}^\top$. This implies that $\mathbf{S}$ is normal, and so we can take an eigendecomposition using the Spectral Theorem.

[1] https://en.wikipedia.org/wiki/Schur_decomposition

## 5.2 Properties

The most important property of symmetric matrices is that all of their eigenvalues are real. Matrices with only real entries can have complex eigenvalues if their characteristic polynomial has non-real zeros. Consider the rotation matrix

$$\begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix}$$

This matrix has no real eigenvalues, as it is a rotation of all of the 2D plane – there is no real vector in the 2D plane left fixed or stretched by such a rotation. Therefore, if we take its eigendecomposition, we will end up with a diagonal matrix containing complex numbers. However, if we restrict our eigendecompositions to symmetric matrices, we are guaranteed that all the eigenvalues of our matrix are real.

To prove that a symmetric matrix has only real eigenvalues, we need to step into the framework of complex vector spaces for a moment. In a real inner product space, the inner product is endowed with three properties – linearity in the first coordinate, symmetry, and that it is positive semi-definite. Complex inner product spaces are mostly the same, except symmetry is replaced with conjugate symmetry, i.e.

$$\langle \mathbf{v}, \mathbf{w} \rangle = \overline{\langle \mathbf{w}, \mathbf{v} \rangle}$$

In a complex vector space, a matrix which is equal to its conjugate transpose is called *Hermitian.* Symbolically, this is represented as $\mathbf{H}^* = \mathbf{H}$ for a Hermitian matrix $\mathbf{H}$. Symmetric matrices are therefore Hermitian matrices with only real entries, and any results on Hermitian matrices will apply to symmetric matrices as well. Let $\mathbf{H}$ be a Hermitian matrix on $\mathbb{C}^n$ and $\mathbf{v_i}$ be an eigenvector of $\mathbf{H}$. Then

$$\lambda_i \langle \mathbf{v_i}, \mathbf{v_i} \rangle = \langle \mathbf{H}\mathbf{v_i}, \mathbf{v_i} \rangle = \mathbf{v_i^*} \mathbf{H}^* \mathbf{v_i} = \mathbf{v_i^*} \mathbf{H} \mathbf{v_i} = \langle v_i, \mathbf{H}\mathbf{v_i} \rangle = \overline{\lambda_i} \langle \mathbf{v_i}, \mathbf{v_i} \rangle$$

This implies $\lambda_i = \overline{\lambda_i}$, which implies $\lambda_i$ is real. So, all Hermitian matrices have real eigenvalues, and thus all symmetric matrices have real eigenvalues.

Since symmetric matrices are normal, we can take the same eigendecomposition $\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ for symmetric matrices as we did with normal matrices. However, in the case where $\mathbf{S}$ is symmetric, we are guaranteed that the entries of $\mathbf{\Lambda}$ are all real.

## 5.3 Rayleigh Quotients

For a symmetric matrix $\mathbf{S}$, the expression $\mathbf{x}^\top \mathbf{S} \mathbf{x}$ is known as a *quadratic form.* The quadratic form of a symmetric matrix can give us insight into its eigenvalues. Define the *Rayleigh quotient* as

$$R_{\mathbf{S}}(\mathbf{x}) = \frac{\mathbf{x}^\top \mathbf{S} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

Note that the Rayleigh quotient is scale-invariant (i.e. $R_{\mathbf{S}}(\alpha \mathbf{x}) = R_{\mathbf{S}}(\mathbf{x})$) and that, for any eigenvector $\mathbf{v_i}$ of $\mathbf{S}$, $R_{\mathbf{S}}(\mathbf{v_i}) = \lambda_i$. It turns out that the Rayleigh quotient is bounded by the largest and smallest eigenvalues of $\mathbf{S}$. Let $\mathbf{x}$ be a vector with norm 1; by scale invariance of the Rayleigh quotient, the following argument will apply to any vector $\mathbf{x}$ of arbitrary length. We begin by decomposing $\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$. We then use the change of variable $\mathbf{y} = \mathbf{Q}\mathbf{x}$; because $\mathbf{Q}$ is an orthogonal matrix, $||\mathbf{y}|| = ||\mathbf{x}|| = 1$, and this mapping is invertible. Therefore,

$$\max_{||x||=1} R_{\mathbf{S}}(\mathbf{x}) = \max_{||y||=1} \mathbf{y}^\top \mathbf{\Lambda} \mathbf{y} = \max_{||y||=1} \sum_{i=1}^{n} \lambda_i y^2$$

Since $||y|| = 1$, the vector which maximizes this summation is the one with a 1 in the position corresponding to the largest eigenvalue and 0 elsewhere. Therefore, $\max\limits_{||x||=1} R_{\mathbf{S}}(\mathbf{x}) = \lambda_{\max}$. An analogous argument can be used to show that $\min\limits_{||x||=1} R_{\mathbf{S}}(\mathbf{x}) = \lambda_{\min}$. So, in general, we have

$$\lambda_{\min} \leq R_{\mathbf{S}}(\mathbf{x}) \leq \lambda_{\max}$$

# 6 Positive Semi-Definite (PSD) Matrices

## 6.1 Definition

A symmetric matrix is *positive semi-definite* if its eigenvalues are all nonnegative. If its eigenvalues are all positive, it is *positive definite.* Equivalently, a matrix $\mathbf{A}$ is positive semi-definite if $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$. $\mathbf{A}$ is positive definite if it is positive semi-definite and the only $\mathbf{x}$ for which $\mathbf{x}^\top \mathbf{A}\mathbf{x} = 0$ is $\mathbf{0}$. Note that any PSD matrix can be made positive-definite by perturbing its diagonal; that is, if $\mathbf{A}$ is PSD and $\epsilon > 0$, then $\mathbf{A} + \epsilon\mathbf{I}$ is positive-definite. This is particularly useful in light of the fact that positive definite matrices are invertible, since all their eigenvalues are nonzero.

## 6.2 Properties

PSD matrices have all the same properties as symmetric matrices and the properties in their definition. Additionally, PSD matrices always have a unique matrix square root. We can demonstrate this by using the eigendecomposition. For a diagonal matrix $\mathbf{D}$ with non-negative entries, let $\mathbf{D}^{\frac{1}{2}}$ be the diagonal matrix whose diagonal contains the square roots of the entries on the original's diagonal. Then $\mathbf{A}^{\frac{1}{2}} = \mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^\top$:

$$(\mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^\top)^2 = \mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^\top\mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^\top = \mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^\top = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top = \mathbf{A}$$

Since the eigenvalues of a PSD matrix are non-negative, $\Lambda^{\frac{1}{2}}$ exists.

Finally, note that for any matrix $\mathbf{X}$, $\mathbf{X}^\top\mathbf{X}$ is positive semi-definite – letting $\mathbf{v}$ denote any vector in $\mathbb{R}^n$,

$$\mathbf{v}^\top\mathbf{X}^\top\mathbf{X}\mathbf{v} = ||\mathbf{X}\mathbf{v}||^2 \geq 0$$

This in turn implies that for *any* matrix $\mathbf{X}$ and $\epsilon > 0$, $\mathbf{X}^\top\mathbf{X} + \epsilon\mathbf{I}$ is positive definite, and therefore invertible. Additionally, if $\mathbf{X}$ is full rank, then it has a trivial nullspace, and so the only vector $\mathbf{v}$ for which $\mathbf{X}\mathbf{v} = 0$ is $\mathbf{0}$. This implies that $\mathbf{v}^\top\mathbf{X}^\top\mathbf{X}\mathbf{v}$ is only 0 when $\mathbf{v} = \mathbf{0}$, so if $\mathbf{X}$ is full rank, $\mathbf{X}^\top\mathbf{X}$ is positive definite.

# 7 Application: Finishing the OLS Proof

We previously started the OLS proof, and reduced the problem of finding the closest vector $\mathbf{Xw} \in \text{range}(\mathbf{X})$ to a vector $\mathbf{y} \in \mathbb{R}^n$ to the problem of finding an orthonormal basis for range($\mathbf{X}$). We could now develop an algorithmic method to complete the OLS dervation – we could generate a spanning set for range($\mathbf{X}$) by taking the columns of $\mathbf{X}$, remove extraneous elements until we have a basis for range($\mathbf{X}$), and then orthonormalize this basis. However, we can improve on this method by instead deriving an expression for $\mathbf{Xw}$ directly. To do so, we will derive an important result regarding matrices and their transposes. We will then confirm that our derivation is correct using another special class of matrices – *projection matrices.*

## 7.1 OLS Proof, Completed

To complete the proof, we first require the following result:

$$\text{null}(\mathbf{X}^\top) = \text{range}(\mathbf{X})^\top$$

*Proof.* Let $\mathbf{X}$ be a $m \times n$ matrix and let $\mathbf{x_1}, \ldots, \mathbf{x_n}$ denote its columns (and thus the rows of $\mathbf{X}^\top$). For all $\mathbf{v} \in \text{null}(\mathbf{X}^\top)$, $\mathbf{X}^\top\mathbf{v} = 0$. This is true iff each component of $\mathbf{X}^\top\mathbf{v}$ is 0, which is the same as saying $\mathbf{x_i}^\top\mathbf{v} = 0$ for all $i$. This is equivalent to stating that any linear combination of the $\mathbf{x_i}$ is orthogonal to $\mathbf{v}$, i.e.,

$$\langle\sum_{i=1}^{n}\alpha_i\mathbf{x_i}, \mathbf{v}\rangle = 0$$

However,

$$\{\sum_{i=1}^{n}\alpha_i\mathbf{x_i} \mid \alpha_i \in \mathbb{R}\} = \text{range}(\mathbf{X})$$

So, any $\mathbf{v} \in \text{null}(\mathbf{X}^\top)$ is orthogonal to the entire range of $\mathbf{X}$; additionally, since all our statements were equivalent, the converse is also true, and any vector orthogonal to the range of $\mathbf{X}$ is in the nullspace of $\mathbf{X}^\top$; therefore, $\text{null}(\mathbf{X}^\top) = \text{range}(\mathbf{X})^\top$. □

Of what use is this result to us? We know that $\mathbf{y} - \mathbf{Xw} \in \text{range}(\mathbf{X})^\top$. Using this result, we can now say that this implies that $\mathbf{y} - \mathbf{Xw} \in \text{null}(\mathbf{X}^\top)$. So, $\mathbf{X}^\top(\mathbf{y} - \mathbf{Xw}) = 0$. This implies that $\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{Xw}$. By assumption, $\mathbf{X}$ is full rank, so $\mathbf{X}^\top\mathbf{X}$ is positive definite and therefore invertible, and so we have that

$$\mathbf{w} = (\mathbf{X}^\top\mathbf{X})^{\mathbf{-1}}\mathbf{X}^\top\mathbf{y}$$

which completes the derivation of OLS. The closest vector in range($\mathbf{X}$) to $\mathbf{y}$ is given by

$$\mathbf{Xw} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{\mathbf{-1}}\mathbf{X}^\top\mathbf{y}$$

## 7.2 Projection Matrices

At the end of the previous application, we had determined that if we had an orthonormal basis for $W$, $\beta_W = \{\mathbf{v_1}, \ldots, \mathbf{v_k}\}$, which we could extend to an orthonormal basis for $V$, $\beta = \{\mathbf{v_1}, \ldots, \mathbf{v_n}\}$, we could then write

$$\mathbf{v} = \sum_{i=1}^{n}\alpha_i\mathbf{v_i},\ \mathbf{v_W} = \sum_{i=1}^{k}\alpha_i\mathbf{v_i}$$

In homework, we noted that we could write $\mathbf{v_W}$ without explicitly keeping track of the $\alpha_i$:

$$\mathbf{v_W} = \sum_{i=1}^{k}\langle\mathbf{v}, \mathbf{v_i}\rangle\mathbf{v_i}$$

We then expressed this sum as a matrix denoted $\mathbf{P_W}$, and derived that $\mathbf{P_W}^2 = \mathbf{P_W}$ and that $\mathbf{P_W}^\top = \mathbf{P_W}$. These properties can be generalized to special classes of matrices. A projection matrix (or *projector*) $\mathbf{P}$ is any matrix equal to its square, i.e., $\mathbf{P}^2 = \mathbf{P}$. An *orthogonal projection matrix (orthogonal projector)* is a symmetric projection matrix. (Note that orthogonal projection matrices are **not** orthogonal matrices. They are, in general, not invertible.)

Orthogonal projectors have various important properties. Primarily, because $\mathbf{P}^2 = \mathbf{P}$, all eigenvalues of $\mathbf{P}$ must be either 0 or 1. This implies (via the Spectral Theorem) that $\mathbf{P}$ is similar to a diagonal matrix $\mathbf{D}$ containing only 0s and 1s on the diagonal. This in turn implies that $\mathbf{D}^2 = \mathbf{D}$; so, any orthogonal projector is similar to a diagonal orthogonal projector. Additionally, $\mathbf{P}$ is an orthogonal projector if and only if there exists a matrix $\mathbf{U}$ such that $\mathbf{U}\mathbf{U}^\top = \mathbf{P}$ and $\mathbf{U}^\top\mathbf{U} = \mathbf{I}'$, where $\mathbf{I}'$ is the identity matrix restricted to range($\mathbf{P}$). The proof of this statement is as follows.

*Proof.* We start with the statement "If $\mathbf{P}$ is an orthogonal projector, then there exists a matrix..." Because $\mathbf{P}$ is an orthogonal projector, it is symmetric and can be decomposed by the Spectral Theorem. Let $\mathbf{P} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top$ be the eigendecomposition of $\mathbf{P}$. As noted in the previous paragraph, $\Lambda$ is a diagonal orthogonal projector. So

$$\mathbf{P} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top = \mathbf{Q}\boldsymbol{\Lambda}^{\mathbf{2}}\mathbf{Q}^\top = \mathbf{Q}\boldsymbol{\Lambda}\boldsymbol{\Lambda}\mathbf{Q}^\top = \mathbf{Q}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\mathbf{Q}^\top = (\mathbf{Q}\boldsymbol{\Lambda})(\mathbf{Q}\boldsymbol{\Lambda})^\top$$

Note that $(\mathbf{Q}\boldsymbol{\Lambda})^\top(\mathbf{Q}\boldsymbol{\Lambda}) = \boldsymbol{\Lambda}\mathbf{Q}^\top\mathbf{Q}\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{\mathbf{2}} = \boldsymbol{\Lambda}$ Since $\boldsymbol{\Lambda}$ is similar to $\mathbf{P}$, it is a projection matrix with range $\mathbf{P}$; since it is a diagonal matrix containing only 1s and 0s, it is the identity matrix restricted to range($\mathbf{P}$), a.k.a. $\mathbf{I}'$. So, the desired $\mathbf{U}$ is $\mathbf{Q}\boldsymbol{\Lambda}$. □

We now prove the converse. Let $\mathbf{P} = \mathbf{U}\mathbf{U}^\top$. Then

$$\mathbf{P}^\top = (\mathbf{U}\mathbf{U}^\top)^\top = \mathbf{U}\mathbf{U}^\top = \mathbf{P}$$

and

$$\mathbf{P}^2 = (\mathbf{U}\mathbf{U}^\top)^2 = \mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top = \mathbf{U}\mathbf{I}'\mathbf{U}^\top = \mathbf{P}$$

as desired. □

This proof proves a slightly stronger statement, as well – that an orthogonal projector can be decomposed into the form $\mathbf{P} = \mathbf{U}\mathbf{U}^\top$ where the non-zero columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{P}$ (and thus an orthonormal basis for range($\mathbf{P}$)).

## 7.3 Checking OLS

In our derivation, we determined that the closest vector in range($\mathbf{X}$) to $\mathbf{y}$ is given by $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{y}$. We can confirm our result by checking that $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top$ is, indeed, an orthogonal projector – that is, $\mathbf{P}_{\text{range}(\mathbf{X})}\mathbf{y} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{y}$. In the next note, after we have developed the machinery of the Singular Value Decomposition, we can even determine the matrix $\mathbf{U}$ such that $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top = \mathbf{U}\mathbf{U}^\top$. But first, we check the two requirements for an orthogonal projector:

$$(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top)^2 = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{I}\mathbf{X}^\top$$

$$(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top)^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top$$

We can additionally confirm that range($\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top$) = range($\mathbf{X}$), since range($\mathbf{X}^\top$) $\perp$ null($\mathbf{X}$).

# 8 Appendix: The OLS Derivation, Condensed

For convenience, we now present the entire OLS derivation, including proofs, from start to finish. We have an $n \times d$ full-rank data matrix $\mathbf{X}$ and an $n$-dimensional observation vector $\mathbf{y}$, and we seek a $d$-dimensional weight vector $\mathbf{w}^*$ such that $\mathbf{w}^* = \underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min} ||\mathbf{y} - \mathbf{X}\mathbf{w}||$.[2] We start by proving that $\mathbf{w}^*$ is optimal if and only if $\mathbf{y} - \mathbf{X}\mathbf{w}^* \perp \text{range}(\mathbf{X})$.

*Proof.* Fix some arbitrary $\mathbf{v} \in \text{range}(\mathbf{X})$, and define the function $f_v(t) = ||\mathbf{y} - (\mathbf{X}\mathbf{w}^* + t\mathbf{v})||^2$. Then $f$ is the square of the distance between $\mathbf{X}\mathbf{w}^* + t\mathbf{v}$, a vector in range($\mathbf{X}$), and $\mathbf{y}$. It should be clear that $f$ is minimized when $t = 0$, as $\mathbf{w}^*$ is assumed to be the optimal solution. So, the derivative of $f_v$ at $t = 0$ is 0. To determine the derivative of $f_v$, we first expand it by rewriting it as an inner product:

$$f_v(t) = \langle(\mathbf{y} - \mathbf{X}\mathbf{w}^*) - t\mathbf{v}, (\mathbf{y} - \mathbf{X}\mathbf{w}^*) - t\mathbf{v}\rangle$$

$$= \langle \mathbf{y} - \mathbf{X}\mathbf{w}^*, \mathbf{y} - \mathbf{X}\mathbf{w}^*\rangle - 2\langle \mathbf{y} - \mathbf{X}\mathbf{w}^*, t\mathbf{v}\rangle + \langle t\mathbf{v}, t\mathbf{v}\rangle$$

$$= ||\mathbf{y} - \mathbf{X}\mathbf{w}^*||^2 - 2t\langle \mathbf{y} - \mathbf{X}\mathbf{w}^*, \mathbf{v}\rangle + t^2||\mathbf{v}||^2$$

We then take the derivative with respect to $t$:

$$f_v'(t) = -2\langle \mathbf{y} - \mathbf{X}\mathbf{w}^*, \mathbf{v}\rangle + 2t||\mathbf{v}||^2$$

and so

$$0 = f_v'(0) = -2\langle \mathbf{y} - \mathbf{X}\mathbf{w}^*, \mathbf{v}\rangle$$

and so $\mathbf{y} - \mathbf{X}\mathbf{w}^*$ is orthogonal to $\mathbf{v}$. Since our choice of $\mathbf{v}$ was arbitrary, we conclude that $\mathbf{y} - \mathbf{X}\mathbf{w}^*$ is orthogonal to every vector in range($\mathbf{X}$). To prove the converse, note that $f_v$ is quadratic in its input and non-negative; so if $\mathbf{y} - \mathbf{X}\mathbf{w}^*$ is orthogonal to every vector in range($\mathbf{X}$), then $f_v'(0) = 0$, and so $t = 0$ must be the global minimum of $f_v$ for all $\mathbf{v}$; so, $||\mathbf{y} - (\mathbf{X}\mathbf{w}^* + t\mathbf{v})||^2$ is minimized for $t = 0$, and thus $\mathbf{X}\mathbf{w}^*$ is the closest vector in range($\mathbf{X}$) to $\mathbf{y}$. □

We now prove that $\text{null}(\mathbf{X}^\top) = \text{range}(\mathbf{X})^\top$, and thus that $\mathbf{y} - \mathbf{X}\mathbf{w}^* \in \text{null}(\mathbf{X}^\top)$.

*Proof.* Let $\mathbf{X}$ be a $m \times n$ matrix and let $\mathbf{x_1}, \ldots, \mathbf{x_n}$ denote its columns (and thus the rows of $\mathbf{X}^\top$). For all $\mathbf{v} \in \text{null}(\mathbf{X}^\top)$, $\mathbf{X}^\top\mathbf{v} = 0$. This is true iff each component of $\mathbf{X}^\top\mathbf{v}$ is 0, which is the same as saying $\mathbf{x_i}^\top\mathbf{v} = 0$ for all $i$. This is equivalent to stating that any linear combination of the $\mathbf{x_i}$ is orthogonal to $\mathbf{v}$, i.e.,

$$\langle\sum_{i=1}^{n} \alpha_i \mathbf{x_i}, \mathbf{v}\rangle = 0$$

However,

$$\{\sum_{i=1}^{n} \alpha_i \mathbf{x_i} \mid \alpha_i \in \mathbb{R}\} = \text{range}(\mathbf{X})$$

So, any $\mathbf{v} \in \text{null}(\mathbf{X}^\top)$ is orthogonal to the entire range of $\mathbf{X}$; additionally, since all our statements were equivalent, the converse is also true, and any vector orthogonal to the range of $\mathbf{X}$ is in the nullspace of $\mathbf{X}^\top$; therefore, $\text{null}(\mathbf{X}^\top) = \text{range}(\mathbf{X})^\top$. □

Since $\mathbf{y} - \mathbf{X}\mathbf{w}^* \in \text{null}(\mathbf{X}^\top)$, we have that $\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w}^*) = 0$. Therefore, $\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X}\mathbf{w}^*$. Because $\mathbf{X}$ is full rank, $\mathbf{X}^\top\mathbf{X}$ is positive definite, and therefore invertible. So, the optimal weight vector $\mathbf{w}^*$ is given by

$$\mathbf{w}^* = (\mathbf{X}^\top\mathbf{X})^{\mathbf{-1}}\mathbf{X}^\top\mathbf{y}$$

which completes the derivation.

[2] Note that since "distance from $\mathbf{y}$" induces a total order on vectors in range($\mathbf{X}$), we are guaranteed that a minimum exists, and therefore that a solution exists. Additionally, since $\mathbf{X}\mathbf{w}^*$ is unique and $\mathbf{X}$ is full rank, $\mathbf{w}^*$ is unique.