Chapter 3

Inner Products and Best Approximations

3.1 Inner Product Spaces

In $\mathbb{R}^2$, one is able to calculate the angle between vectors using the dot product. Indeed, if $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$, the angle $\theta$ between $\mathbf{u}$ and $\mathbf{v}$ satisfies:

$$\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$$

where $\mathbf{u} \cdot \mathbf{v} = u_1v_1 + u_2v_2$ is the dot product of $\mathbf{u}$ and $\mathbf{v}$; and $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}} = \sqrt{u_1^2 + u_2^2}$ is the (usual) Euclidean length of a vector $\mathbf{u}$. In order to extend the geometric notions of “angle between vectors” and “length of a vector” to more general vector spaces it would seem enough to have a suitable generalization of the dot product to more general vector spaces. The inner product is such a generalization.

Let $V$ be a complex vector space. An inner product on $V$ is a function that maps pairs of vectors $\mathbf{u}, \mathbf{v} \in V$ to a complex scalar $\langle \mathbf{u}, \mathbf{v} \rangle$ in such a way that for all $\mathbf{u}, \mathbf{v} \in V$:

1. $\langle \mathbf{u}, \mathbf{u} \rangle$ is real and nonnegative, and $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ if and only if $\mathbf{u} = \mathbf{0}$.
2. $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$
3. $\langle \mathbf{u}, \alpha \mathbf{v} + \beta \mathbf{w} \rangle = \alpha \langle \mathbf{u}, \mathbf{v} \rangle + \beta \langle \mathbf{u}, \mathbf{w} \rangle$
If $V$ is a real vector space then for every pair of vectors $u, v \in V$, $\langle u, v \rangle$ is a real scalar so that conditions (1) and (3) continue to hold; condition (2) becomes:

$$\langle u, v \rangle = \langle v, u \rangle$$

A vector space $V$ on which an inner product is defined is called an inner product space. We define the norm of a vector in an inner product space as $\|u\| = \sqrt{\langle u, u \rangle}$.

Examples of inner product spaces:

- $\mathbb{R}^2$ – the set of real ordered pairs. For vectors $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ and $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ in $\mathbb{R}^2$,

  $$\langle u, v \rangle = u_1v_1 + u_2v_2 = u^t v$$

  defines an inner product on $\mathbb{R}^2$. This provides a model for generalizations to other vector spaces.

- $\mathbb{C}^n$ – the set of complex $n$-tuples. For vectors $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$ in $\mathbb{C}^n$,

  $$\langle u, v \rangle = \sum_{i=1}^{n} \overline{u_i}v_i = u^*v$$

  defines an inner product on $\mathbb{C}^n$. (The conjugate transpose of a matrix $A = [a_{ij}]$ is defined elementwise as $A^* = B = [b_{ij}]$ with $b_{ij} = \overline{a_{ji}}$).

- $\mathcal{P}_n$ – the set of (complex) polynomials of degree $n$ or less. Pick $n + 1$ distinct points $\{z_i\}_{i=0}^n$ in the complex plane. For polynomials $p, q \in \mathcal{P}_n$,

  $$\langle p, q \rangle = \sum_{i=0}^{n} p(z_i)q(z_i)$$

  defines an inner product on $\mathcal{P}_n$.

- $\mathbb{C}^{m \times n}$ – the set of $m \times n$ matrices with complex entries. For any matrix $T \in \mathbb{C}^{m \times n}$ define $\text{trace}(T) = \sum_{i=1}^{n} t_{ii}$. Then for any $A, B \in \mathbb{C}^{m \times n}$,

  $$\langle A, B \rangle = \text{trace}(A^*B)$$
defines an inner product on $\mathbb{C}^{m \times n}$. The associated norm: $\|A\|_F^2 = \text{trace}(A^* A) = \sum_{i,j} |a_{ij}|^2$ is called variously the Frobenius norm, the Hilbert-Schmidt norm, or the Schatten 2-norm. We adhere to the name “Frobenius” and will hang an “F” on such matrix norms to distinguish them from others yet to come.

- $C[a, b]$ – the set of real-valued continuous functions on $[a, b]$. For any functions $f, g \in C[a, b]$,$$
\langle f, g \rangle = \int_a^b f(x)g(x) \, dx
$$
defines an inner product on $C[a, b]$.

**Problem 3.1.1** Prove that in each case above, the function $\langle \cdot, \cdot \rangle$ satisfies all the conditions of an inner product.

It is a subtle but fundamental observation that the defining properties of the inner product are sufficient to guarantee that $|\langle u, v \rangle|/\|u\| \|v\| \leq 1$. This allows us to extend sensibly the notion of angle between vectors in general vector spaces so that the acute angle $\theta_{uv}$, between the vectors $u, v \in V$ (or more properly, the angle between the subspaces they generate) is defined so that
$$
\cos(\theta_{uv}) = |\langle u, v \rangle|/\|u\| \|v\|.
$$
The following theorem guarantees that the right-hand quantity is bounded by 1 in magnitude and is known as the Cauchy-Schwarz inequality.

**Theorem 3.1.2** Let $u, v \in V$, an inner product space. Then
$$
|\langle u, v \rangle|^2 \leq \|u\|^2 \|v\|^2
$$
Equality holds if and only if $u = kv$, for some scalar $k \in \mathbb{C}$.

**Proof:** Let $a = \|v\|^2$, $b = 2|\langle u, v \rangle|$ and $c = \|u\|^2$. First, if $a = 0$ then $v = 0$ (by property 1 of an inner product) which would then imply (by property 3 of an inner product) that $b = 0$. Clearly in this case the conclusion is true. Now, consider the case $a > 0$ and choose $\theta \in [0, 2\pi)$ so that $e^{i\theta} \langle u, v \rangle$ is a real nonnegative number. Then,
$$
e^{i\theta} \langle u, v \rangle = |e^{i\theta} \langle u, v \rangle| = |\langle u, v \rangle|
Pick a real number $t$ arbitrarily and define $z = te^{i\theta}$. One may calculate:

\[
0 \leq \|u + zv\|^2 = \langle u + zv, u + zv \rangle = \|u\|^2 + z\langle u, v \rangle + z\overline{\langle u, v \rangle} + |z|^2\|v\|^2 = c + bt + at^2.
\]

(3.1)

Since this is true for all $t \in \mathbb{R}$, the quadratic polynomial $at^2 + bt + c$ is always nonnegative; has only complex conjugate roots; and thus a nonpositive discriminant $b^2 - 4ac$. Thus

\[
b^2 \leq 4ac
\]

and the conclusion is obtained by taking square roots on each side.

Notice that equality in the Cauchy-Schwarz inequality is equivalent to a zero discriminant ($b^2 = 4ac$) which means that equality can be attained in (3.1) with $t = -b/2a$. But that means in turn that $u + zv = 0.$

Let $V$ be an inner product space. Two vectors $u$ and $v$ in $V$ are called orthogonal (or perpendicular) if $\langle u, v \rangle = 0$. A set of vectors $\{v_1, v_2, \ldots, v_r\} \subseteq V$ is called orthogonal if

\[
\langle v_i, v_j \rangle = 0, \quad i \neq j.
\]

An orthogonal set is called orthonormal if furthermore

\[
\langle v_i, v_i \rangle = \|v_i\|^2 = 1, \quad i = 1, 2, \ldots, r.
\]

For example the set \{\begin{pmatrix} 0 \\ i \\ (i + 1) \end{pmatrix}/\sqrt{2}, \begin{pmatrix} 0 \\ 1 + i \end{pmatrix}/\sqrt{2}, -1}\} is orthonormal.

Let $W$ be a subspace of an inner product space, $V$. The set of vectors

\[
W^\perp = \{u \in V \text{ such that } \langle u, w \rangle = 0 \text{ for all } w \in W\}
\]

is called the orthogonal complement of $W$.

**Problem 3.1.3** Show that

- $W^\perp$ is a subspace of $V$
- $W \subseteq (W^\perp)^\perp$
3.2. BEST APPROXIMATION AND PROJECTIONS

- $W \cap W^\perp = \{0\}$.

**Problem 3.1.4** Show that if $B$ is a set of orthogonal vectors, none of which is the zero vector, then $B$ is linearly independent.

**Problem 3.1.5** Show that if $B = \{w_1, w_2, \ldots, w_l\}$ is a set of linearly independent vectors, then $\ell \times \ell$ matrix, $G = [(w_i, w_j)]$ is nonsingular. $G$ is called the Gram matrix for $B$.

**Problem 3.1.6** Prove the Pythagorean Theorem in inner product spaces: If $u$ and $v$ are orthogonal then $\|u + v\|^2 = \|u\|^2 + \|v\|^2$.

3.2 Best approximation and projections

Very often in application settings a particular subspace, $W$ say, of an inner product space, $V$, may have some special properties that make it useful and interesting to approximate any given vector $v \in V$ as well as possible with a corresponding vector $w_\ast \in W$. That is, find $w_\ast \in W$ that solves

$$\min_{w \in W} \|w - v\| = \|w_\ast - v\|$$

(3.2)

We have the following characterization of the solution $w_\ast$.

**Theorem 3.2.1** Let $W$ be a subspace of an inner product space, $V$. The vector $w_\ast \in W$ is a solution to (3.2) if and only if $w_\ast - v \perp W$. Furthermore, for any given $v \in V$, there can be no more than one such solution $w_\ast$.

**Proof:** Suppose that $w_\ast \in W$ is a solution to (3.2) and pick an arbitrary vector $w \in W$. Choose a $\theta \in [0, 2\pi)$ so that $e^{i\theta}\langle w_\ast - v, w \rangle$ is real and nonnegative. Now, for any real $\varepsilon > 0$ define $z = -\varepsilon e^{i\theta}$ and notice that

$$\|w_\ast - v\|^2 \leq \|(w_\ast + z\rangle - v\|^2 = \|w_\ast - v\|^2 - 2\varepsilon\langle w_\ast - v, w \rangle + \varepsilon^2\|w\|^2$$

(3.3)

So for all $\varepsilon > 0$,

$$0 \leq -2\varepsilon\langle w_\ast - v, w \rangle + \varepsilon^2\|w\|^2$$

which means,

$$0 \leq 2\langle w_\ast - v, w \rangle \leq \varepsilon\|w\|^2$$
Inner Products and Best Approximation

Since we are free to make \( \varepsilon \) as small as we like, it must be that \( \langle w_s - v, w \rangle = 0 \) and this is true for each \( w \in W \).

To prove the converse, suppose \( w_s \in W \) satisfies \( w_s - v \perp W \). Then for any \( w \in W \), \( (w - w_s) \in W \), and

\[
\|w - v\|^2 = \|(w - w_s) + (w_s - v)\|^2 = \|(w - w_s)\|^2 + \|(w_s - v)\|^2 \geq \|(w_s - v)\|^2
\]

so \( w_s \) solves (3.2).

To prove uniqueness, suppose that there were two solutions to (3.2), say \( w_{1s} \) and \( w_{2s} \). Then \( (w_{1s} - v) \in W^\perp \) and \( (w_{2s} - v) \in W^\perp \). Since \( W^\perp \) is a subspace, we find

\[
(w_{1s} - v) - (w_{2s} - v) = (w_{1s} - w_{2s}) \in W^\perp.
\]

On the other hand, \( (w_{1s} - w_{2s}) \in W \). Since the only vector both in \( W \) and in \( W^\perp \) is \( 0 \) (see Problem 3.1.3), we find \( w_{1s} = w_{2s} \). \( \square \)

While this is a nice characterization of solutions to best approximation problems, this result leaves open the question of whether a solution \( w_s \) to (3.2) always exists and if so, how one might go about calculating it. The following theorem describes one way of obtaining a solution to (3.2).

**Theorem 3.2.2** Let \( V \) be an inner product space. If \( W \) denotes a subspace of \( V \), then every vector \( v \in V \) can be expressed uniquely as

\[
v = w_s + w^\perp,
\]

where \( w_s \in W \) and \( w^\perp \in W^\perp \). \( w_s \) is the unique solution to (3.2).

**Proof:** Let

\[
B = \{w_1, w_2, \ldots, w_r\}
\]

be a basis for \( W \). Define \( [\gamma_{ij}] \) to be the \( r \times r \) matrix inverse to \( [\langle w_i, w_j \rangle] \) (see Problem 3.1.5) and

\[
w_s = \sum_{i,j=1}^{r} w_i \gamma_{ij} \langle w_j, v \rangle
\]

and pick an arbitrary \( w \in W \). Then

\[
w = a_1 w_1 + \cdots + a_r w_r
\]
where \( a_i \in \mathbb{C}, i = 1, 2, \ldots, r \). We need only prove that \( v - w_* \) and \( w \) are orthogonal regardless of which \( w \in W \) was chosen and then we can define \( w^\perp = v - w_* \). Using properties of the inner product we see that

\[
\langle w, v - w_* \rangle = \sum_{k=1}^{r} a_k \langle w_k, v - w_* \rangle = \sum_{k=1}^{r} \overline{a_k} \langle w_k, v \rangle - \langle w_k, w_* \rangle
\]

Observe that

\[
\langle w_k, w_* \rangle = \sum_{i,j=1}^{r} \langle w_k, w_i \rangle \gamma_{ij} \langle w_j, v \rangle = \sum_{j=1}^{r} \left( \sum_{i=1}^{r} \langle w_k, w_i \rangle \gamma_{ij} \right) \langle w_j, v \rangle = \langle w_k, v \rangle
\]

Thus \( \langle w, v - w_* \rangle = 0 \). Since \( w_* - v \in W^\perp \), \( w_* \) is the unique solution to (3.2). \( \Box \)

**Corollary 3.2.3** \( W = (W^\perp)^\perp \)

**Proof:** We saw in Problem 3.1.3 that \( W \subset (W^\perp)^\perp \). Pick any vector \( v \in (W^\perp)^\perp \). Then \( v = w_* + w^\perp \) and since \( w_* \in W \subset (W^\perp)^\perp \), one finds that \( v - w_* \in (W^\perp)^\perp \). But \( v - w_* = w^\perp \in W^\perp \), so \( v - w_* = 0 \) and \( v = w_* \in W \). \( \Box \)

The vector \( w_* \) in the above decomposition of \( v \) is called the *orthogonal projection* of \( v \) onto \( W \). The vector \( w^\perp \) is called the component of \( v \) orthogonal to \( W \). From the proof of the theorem, one can see that best approximations to a subspace \( W \), can be found in a straightforward way if a basis for \( W \) is known. The mapping that carries the vector \( v \) to the vector \( w_* \) that solves (3.2) is called an *orthogonal projector*, which we will denote as \( w_* = P_W(v) \). \( P_W \) is a linear transformation, since evidently

\[
w_* = \sum_{i,j=1}^{r} w_i \gamma_{ij} \langle w_j, v \rangle
\]

inherits linearity with respect to \( v \) from the linearity of the inner product with respect to its second argument. If \( V = \mathbb{C}^n \), we have the following convenient matrix representation of \( P_W \) in terms of \( \{w_1, w_2, \ldots, w_r\} \):

\[
P_W = W G^{-1} W^*,
\]
where $W = [w_1, w_2, \ldots, w_r]$ and $G = W^*W$ is the Gram matrix for \{w_1, w_2, \ldots, w_r\}.

In general, what characterizes an orthogonal projector?

**Theorem 3.2.4** $P_W$ is an orthogonal projector onto a subspace $W$ of $\mathbb{C}^n$ if and only if:

1. $\text{Ran}(P_W) = W$
2. $P_W^2 = P_W$
3. $P_W^* = P_W$

**Proof:** Suppose that $P_W$ represents an orthogonal projection onto $W$, so that $w_s = P_W v$ solves (3.2) for each $v \in \mathbb{C}^n$. Then, in particular, for any vector $w \in W$, $v = w$ itself solves (3.2) so $w = P_W w$, and as a consequence $\text{Ran}(P_W) = W$. Furthermore, for any vector $v \in \mathbb{C}^n$, $w_s = P_W v \in W$, so

$$P_W^2 v = P_W (P_W v) = P_W w_s = w_s = P_W v,$$

implying that $P_W^2 = P_W$. Finally, for any vectors $u, v \in \mathbb{C}^n$, $u - P_W u \in W^\perp$ and $P_W v \in W$ so,

$$\langle u - P_W u, P_W v \rangle = 0$$
$$\langle u, P_W v \rangle - \langle P_W u, P_W v \rangle = 0$$
$$\langle u, P_W v \rangle - \langle u, P_W^* P_W v \rangle = 0$$
$$\langle u, (P_W - P_W^* P_W) v \rangle = 0$$

Thus, $P_W = P_W^* P_W$ and as a consequence,

$$(P_W)^* = (P_W^* P_W)^* = P_W^* P_W = P_W.$$

Conversely, suppose that $P_W$ is a matrix satisfying the three properties above. Then for any vector $v \in \mathbb{C}^n$ and any vector $w \in W$, we find

$$\langle v - P_W v, w \rangle = \langle v, w \rangle - \langle P_W v, w \rangle$$
$$= \langle v, w \rangle - \langle v, P_W^* P_W v \rangle$$
$$= \langle v, w \rangle - \langle v, P_W w \rangle$$
$$= \langle v, w \rangle - \langle v, w \rangle = 0$$

Thus, $P_W v$ solves (3.2) for each $v \in \mathbb{C}^n$ and so, represents an orthogonal projection. \(\square\)
Problem 3.2.5 If \( P_W \) represents an orthogonal projection onto a subspace \( W \) of \( \mathbb{C}^n \), show that \( I - P_W \) represents an orthogonal projection onto \( W^\perp \).

Problem 3.2.6 Given a matrix, \( C \in \mathbb{R}^{n \times r} \) such that \( \ker(C) = \{0\} \), show that

\[
P_{\mathrm{ran}(C)} = C(C^tC)^{-1}C^t
\]

represents an orthogonal projection onto \( \mathrm{ran}(C) \). What role does the assumption on \( \ker(C) \) play here?

Problem 3.2.7 Given a matrix, \( B \in \mathbb{R}^{r \times m} \) such that \( \mathrm{ran}(B) = \mathbb{R}^r \), show that

\[
P_{\ker(B)} = I - B^t(BB^t)^{-1}B
\]

represents an orthogonal projection onto \( \ker(B) \). What role does the assumption on \( \mathrm{ran}(B) \) play here?

Any linear transformation that \( Q_W \) that satisfies the two properties

1. \( \mathrm{ran}(Q_W) = W \)
2. \( Q_W^2 = Q_W \)

is called a skew projector (or just a projector) onto the subspace \( W \).

Problem 3.2.8 Show that if \( Q_W \) is a projector (either skew or orthogonal), then \( \mathrm{ran}(Q_W) = \ker(I - Q_W) \) and \( \ker(Q_W) = \mathrm{ran}(I - Q_W) \).

Problem 3.2.9 Suppose \( A \in \mathbb{R}^{m \times n} \) has a left inverse \( B_L \). Prove that \( Q = AB_L \) is a (possibly skew) projector onto \( \mathrm{ran}(A) \).

Problem 3.2.10 Suppose \( A \in \mathbb{R}^{m \times n} \) has a right inverse \( B_R \). Prove that \( Q = I - B_RA \) is a (possibly skew) projector onto \( \ker(A) \).

Let \( V = \mathbb{R}^3 \) and let \( W \) be spanned by \( \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \) and \( \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \), that is \( W \) is the x-y plane. For given \( u \in \mathbb{R}^3 \), the vectors \( u, w \) and \( w^\perp \) form a right triangle with \( u \) the hypotenuse, \( w \) in the x-y plane and \( w^\perp \) parallel to the z-axis.
Problem 3.2.11 Let $W$ be the subspace of $\mathbb{R}^3$ spanned by the vectors

$$v_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad v_2 = (-4/5, 0, -3/5).$$

Show that $v_1$ and $v_2$ form an orthonormal set and find $P_W u$, where $u = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$.

3.3 Pseudoinverses

Finding closest vectors out of subspaces can be used to extend the concepts of left and right inverses. Suppose that $A \in \mathbb{C}^{m \times n}$ has a right inverse and $m < n$. We know $nullity(A) > 0$ so $A$ cannot have a left inverse. Although $Ax = b$ is consistent for any $b$, each right hand side $b$ will be associated with an infinite family of solutions.

**Problem 1**: Find the smallest solution $\hat{x}$ to $Ax = b$.

Now, suppose instead that $A \in \mathbb{C}^{m \times n}$ has a left inverse and $m > n$. We know $\text{rank}(A) = n < m$ so $A$ cannot have a right inverse and $Ax = b$ will be inconsistent for some $b$.

**Problem 2**: Find a vector $\hat{x}$ that brings $Ax$ as close as possible to $b$.

In each of these cases, we seek vectors that are in some sense or other the most plausible substitute for a Given any (rectangular) matrix $A \in \mathbb{C}^{m \times n}$, the pseudoinverse of $A$ is a matrix $B \in \mathbb{C}^{n \times m}$, that provides for each $b \in \mathbb{C}^n$, a solution $Bb = x_\ast \in \mathbb{C}^n$ to the following best approximation problem:

$$\text{(3.4)} \quad \min_{x \in \mathbb{C}^m} \| Ax - b \|$$

such that $\|x\|$ is minimal.

A variety of notations are found for the pseudoinverse; the most common appears to be $B = \text{"}A^\dagger\text{"}$. The definition (3.4) does little to give insight into what actions are taken to transform $b$ into $x_\ast$. For that reason, the following prescription for constructing $x_\ast$ may be more useful as a definition of $A^\dagger$:

**The Action of the Pseudoinverse**

Define $P$ to be the orthogonal projection onto $\text{Ran}(A)$ and $Q$ to be the orthogonal projection onto $\text{Ker}(A)^\perp$.

1. Find the component of $b$ in $\text{Ran}(A)$: $Pb = y_\ast$. 

2. Find any one solution, \( \hat{x} \), to the linear system \( Ax = y \).

3. Find the component of \( \hat{x} \) in \( \text{Ker}(A)^\perp \): \( x_* = Q\hat{x} \).

Two issues immediately emerge:

- Is the construction above well-defined to the extent that the final result \( x_* \) is the same regardless of which intermediate result \( \hat{x} \) was picked?
- Does \( x_* \) solve \((3.4)\)?

**Theorem 3.3.1** The construction of \( x_* \) specified above is well-defined and produces the unique solution to \((3.4)\).

**Proof:** Step (1) defines \( y_* \) uniquely as the solution to \((3.2)\) with \( W = \text{Ran}(A) \) and \( v = b \):

\[
\|y_* - b\| = \min_{y \in \text{Ran}(A)} \|y - b\| = \min_{x \in \mathbb{C}^n} \|Ax - b\|
\]

Since \( y_* \in \text{Ran}(A) \), the linear system \( Ax = y_* \) of Step (2) must be consistent and has at least one solution, say \( \hat{x} \). Notice that \( \hat{x} \) is a solution to

\[
\min_{x \in \mathbb{C}^n} \|Ax - b\| = \|A\hat{x} - b\|
\]

and in fact, any solution to \( Ax = y_* \) will be a minimizer in the same sense.

To show that the outcome of Step (3) is independent of which solution \( \hat{x} \) was picked in Step (2), suppose that two solutions, \( \hat{x}_1 \) and \( \hat{x}_2 \), were known that solve \( Ax = y_* \) and these produce two outcomes in Step (3): \( x_{1*} = Q\hat{x}_1 \) and \( x_{2*} = Q\hat{x}_2 \). Since \( A\hat{x}_1 = y_* = A\hat{x}_2 \), rearrangement gives \( A(\hat{x}_1 - \hat{x}_2) = 0 \) so that \( (\hat{x}_1 - \hat{x}_2) \in \text{Ker}(A) \). We find

\[
x_{1*} - x_{2*} = Q(\hat{x}_1 - \hat{x}_2) = 0.
\]

Thus, \( x_{1*} = x_{2*} \) and the outcome of Step (3) is uniquely determined regardless of which solution in Step (2) was used. Furthermore, any such solution, \( \hat{x} \) can be decomposed as

\[
\hat{x} = x_* + \hat{n}
\]

where \( \hat{n} \in \text{Ker}(A) \). By the Pythagorean Theorem,

\[
\|\hat{x}\|^2 = \|x_*\|^2 + \|\hat{n}\|^2
\]
so that out of all possible solutions, \( \hat{x} \), the minimal norm solution must occur when \( \mathbf{n} = 0 \) — that is, when \( \hat{x} = x_s \).

In some cases, the prescription for \( x_s = A^\dagger b \) can be used to find what amounts to a formula for \( A^\dagger \).

**Theorem 3.3.2** If \( A \in \mathbb{C}^{m \times n} \) has the full rank factorization \( A = XY^* \), where \( X \in \mathbb{C}^{m \times p} \) and \( Y \in \mathbb{C}^{n \times p} \) are both of rank \( p \), then the pseudoinverse is given by

\[
A^\dagger = Y(Y^*Y)^{-1}(X^*X)^{-1}X^* 
\]

**Proof:** We first construct the projections \( P \) and \( Q \). Notice that \( \text{Ran}(A) = \text{Ran}(X) \) and that \( \text{Ker}(A) = \text{Ran}(A^*) = \text{Ran}(Y) \). Then, directly

\[
P = X(X^*X)^{-1}X^* , \quad Q = Y(Y^*Y)^{-1}Y^* 
\]

We now seek a solution to the system of equations

\[
A\hat{x} = XY^*\hat{x} = X(X^*X)^{-1}X^*b = Pb 
\]

Noticing that \( \text{rank}(X) = p \) implies that \( \text{Ker}(A) = \{0\} \), we can rearrange to find

\[
0 = XY^*\hat{x} - X(X^*X)^{-1}X^*b \\
0 = X(Y^*\hat{x} - (X^*X)^{-1}X^*b) \\
0 = Y^*\hat{x} - (X^*X)^{-1}X^*b 
\]

While this does not yield a formula for \( \hat{x} \) itself (recall that there won’t generally be just one solution), we may premultiply both sides of the final equation by \( Y(Y^*Y)^{-1} \) to get

\[
Y(Y^*Y)^{-1}Y^*\hat{x} = Y(Y^*Y)^{-1}(X^*X)^{-1}X^*b \\
x_s = Q\hat{x} = Y(Y^*Y)^{-1}(X^*X)^{-1}X^*b 
\]

**Problem 3.3.3** Using the permuted LU factorization, show that for any matrix \( A \in \mathbb{C}^{m \times n} \), if the rank of \( A \) is \( p \) then there are matrices \( X \in \mathbb{C}^{m \times p} \) and \( Y \in \mathbb{C}^{n \times p} \), both of rank \( p \), so that \( A = XY^* \).
3.4  Orthonormal Bases and the QR Decomposition

If \( S = \{w_1, \ldots, w_r\} \) is a basis for a subspace \( W \) of a vector space \( V \) and \( u \) is an arbitrary vector in \( W \), then \( u \) can be uniquely expressed as

\[
  u = k_1 w_1 + \cdots + k_r w_r
\]

In particular, if \( W \) is a subspace of \( \mathbb{C}^n \) then by defining a matrix \( W = [w_1, w_2, \ldots, w_r] \) and a vector of unknown coefficients, \( k = [k_1, \ldots, k_r]^t \), the coefficients can be found directly by solving the system of equations \( Wk = u \). If the system is inconsistent then \( u \) was not in the subspace \( W \) after all. In more general vector spaces, it may not be so clear how to proceed to find the coefficients \( \{k_1, \ldots, k_r\} \) or how to determine whether or not \( u \) is in \( W \). The next theorem tells us precisely how to find these coefficients in cases where the basis \( S \) is orthonormal. The general case will be considered later in the section.

**Theorem 3.4.1** If \( S = \{w_1, w_2, \ldots, w_r\} \) is an orthonormal basis for a subspace \( W \) of an inner product space \( V \), then for any \( u \in W \)

\[
  u = \langle w_1, u \rangle w_1 + \langle w_2, u \rangle w_2 + \cdots + \langle w_r, u \rangle w_r
\]

Furthermore, \( u \in W \) if and only if \( \|u\|^2 = \sum_{i=1}^{r} |\langle w_i, u \rangle|^2 \)

**Proof:** Since \( S \) is a basis for \( W \), there exist constants \( k_1, \ldots, k_r \) such that

\[
  u = k_1 w_1 + k_2 w_2 + \cdots + k_r w_r
\]

Thus by properties of the inner product and the fact that \( S \) is orthonormal, we obtain that

\[
  \langle w_1, u \rangle = \langle w_1, k_1 w_1 + k_2 w_2 + \cdots + k_r w_r \rangle = k_1 \langle w_1, w_1 \rangle + k_2 \langle w_1, w_2 \rangle + \cdots + k_r \langle w_1, w_r \rangle = k_1 \cdot 1 + k_2 \cdot 0 + \cdots + k_r \cdot 0
\]

Similarly,

\[
  \langle w_2, u \rangle = k_2, \ldots, \langle w_r, u \rangle = k_r.
\]
Also
\[ \| \mathbf{u} \|^2 = \left\| \sum_{i=1}^{r} k_i \mathbf{w}_i \right\|^2 = \langle \sum_{i=1}^{r} k_i \mathbf{w}_i, \sum_{j=1}^{r} k_j \mathbf{w}_j \rangle = \sum_{i=1}^{r} \sum_{j=1}^{r} k_i k_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle = \sum_{i=1}^{r} |k_i|^2 = \sum_{i=1}^{r} |\langle \mathbf{w}_i, \mathbf{u} \rangle|^2. \]

Any \( \mathbf{u} \in V \) can be written as \( \mathbf{u} = \mathbf{w}_s + \mathbf{w}^\perp \) with \( \mathbf{w}_s \in W \) and \( \mathbf{w}^\perp \perp W \).

Since
\[ \langle \mathbf{w}_i, \mathbf{u} \rangle = \langle \mathbf{w}_i, \mathbf{w}_s \rangle + \langle \mathbf{w}_i, \mathbf{w}^\perp \rangle = \langle \mathbf{w}_i, \mathbf{w}_s \rangle, \]
we have that \( \mathbf{w}_s = \langle \mathbf{w}_1, \mathbf{u} \rangle \mathbf{w}_1 + \langle \mathbf{w}_2, \mathbf{u} \rangle \mathbf{w}_2 + \cdots + \langle \mathbf{w}_r, \mathbf{u} \rangle \mathbf{w}_r \) and so
\[ \| \mathbf{u} \|^2 - \sum_{i=1}^{r} |\langle \mathbf{w}_i, \mathbf{u} \rangle|^2 = \| \mathbf{w}^\perp \|^2 \]
which means in turn that \( \| \mathbf{u} \|^2 = \sum_{i=1}^{r} |\langle \mathbf{w}_i, \mathbf{u} \rangle|^2 \) if and only if \( \mathbf{w}^\perp = 0 \), which occurs if and only if \( \mathbf{u} \in W \). \( \square \)

**Problem 3.4.2** Check that \( S = \{ \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -4/5, 0, 3/5 \end{pmatrix}, \begin{pmatrix} 3/5, 0, 4/5 \end{pmatrix} \} \) is an orthonormal basis for \( \mathbb{R}^3 \) and express the following vectors as linear combinations of the vectors in \( S \).

a. \( \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \)  
b. \( \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \)  
c. \( \begin{pmatrix} 3 \\ 2 \\ -1 \end{pmatrix} \).

The preceding theorem points out a benefit of having an orthonormal basis for a vector space. How do we go about finding such a basis? The next theorem demonstrates how to construct one starting from any given basis. The construction is called the **Gram-Schmidt Orthogonalization Process**.

**Theorem 3.4.3** Let \( V \) be an inner product space with a basis,
\[ S = \{ \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n \}. \]
Then $V$ has an orthonormal basis,

$$\{q_1, q_2, \ldots, q_n\},$$

so that for each $k = 1, \ldots, n$,

$$\text{span} \{u_1, u_2, \ldots, u_k\} = \text{span} \{q_1, q_2, \ldots, q_k\}.$$

Proof: First we note that since $S$ is a basis, $u_1 \neq 0$. Thus we may set

$$q_1 = \frac{u_1}{\|u_1\|}.$$

Then $\|q_1\| = 1$ and $\text{span} \{u_1\} = \text{span} \{q_1\}$. To construct $q_2$, compute the component of $u_2$ orthogonal to $\text{span} \{q_1\}$ as in Theorem 3.2.2 and divide by its length to produce a vector of length one. So we define

$$w_2 = u_2 - \langle q_1, u_2 \rangle q_1$$

To see that $w_2 \neq 0$, note that if it were zero, then $u_2$ would be a scalar multiple of $q_1$ which is in turn a scalar multiple of $u_1$. This would mean that $u_1, u_2$ is a linearly dependent set and could not be a part of a basis set for $V$, contradicting our initial hypothesis. So $w_2 \neq 0$ and we define

$$q_2 = \frac{w_2}{\|w_2\|}$$

Clearly $q_2$ has length 1 and since $q_2$ is orthogonal to $\text{span} \{q_1\}$, $\{q_1, q_2\}$ is an orthonormal set. Furthermore, $u_2 = r_1 q_1 + r_2 q_2$ where $r_1 = \langle q_1, u_2 \rangle$ and $r_2 = \|w_2\|$, so that $\text{span} \{u_1, u_2\} = \text{span} \{q_1, q_2\}$.

Now we continue the construction inductively. Suppose for some $k > 1$, we’ve produced a set of $k-1$ orthonormal vectors $\{q_1, q_2, \ldots, q_{k-1}\}$ so that for each $j = 1, \ldots, k-1$, $\text{span} \{u_1, u_2, \ldots, u_j\} = \text{span} \{q_1, q_2, \ldots, q_j\}$. (We’ve done this above for $k=3$). To construct $q_k$, we will compute the component of $u_k$ orthogonal to $\text{span} \{q_1, q_2, \ldots, q_{k-1}\}$ and then divide by its length, producing a vector of length one. Define

$$w_k = u_k - \sum_{j=1}^{k-1} \langle q_j, u_k \rangle q_j$$

Again we must check that $w_k \neq 0$. Were it so, then $u_k$ would be in the span of $\{q_1, q_2, \ldots, q_{k-1}\}$ and hence would be a linear combination
of \( \{u_1, u_2, \ldots, u_{k-1}\} \). This implies that \( \{u_1, u_2, \ldots, u_k\} \) is a linearly dependent set and could not be a part of a basis for \( V \), contradicting our starting hypothesis. Thus, \( w_k \neq 0 \) and we define

\[
q_k = \frac{w_k}{\|w_k\|}
\]

A quick calculation verifies for each \( j = 1, \ldots, k - 1 \)

\[
\langle q_j, q_k \rangle = \frac{1}{\|w_k\|} \langle q_j, w_k \rangle
\]

\[
= \frac{1}{\|w_k\|} \left( q_j, u_k - \sum_{\ell=1}^{k-1} \langle q_\ell, u_k \rangle q_\ell \right)
\]

\[
= \frac{1}{\|w_k\|} (\langle q_j, u_k \rangle - \sum_{\ell=1}^{k-1} \langle q_\ell, u_k \rangle \langle q_j, q_\ell \rangle)
\]

\[
= \frac{1}{\|w_k\|} (\langle q_j, u_k \rangle - \langle q_j, u_k \rangle) = 0,
\]

so \( \{q_1, q_2, \ldots, q_k\} \) is an orthonormal set. Furthermore, \( u_k = \sum_{\ell=1}^k r_{\ell k} q_\ell \), where \( r_{\ell k} = \langle q_\ell, u_k \rangle \) for \( \ell = 1, \ldots, k - 1 \) and \( r_{kk} = \|w_k\| \). Thus, \( u_k \in \text{span} \{q_1, q_2, \ldots, q_k\} \). Since we already know that

\[
\text{span} \{u_1, u_2, \ldots, u_{k-1}\} = \text{span} \{q_1, q_2, \ldots, q_{k-1}\} \subseteq \text{span} \{q_1, q_2, \ldots, q_k\},
\]

we find that \( \text{span} \{u_1, u_2, \ldots, u_j\} = \text{span} \{q_1, q_2, \ldots, q_j\} \) for each \( j = 1, \ldots, k \), which completes the induction step. \( \square \)

**Problem 3.4.4** Apply the Gram-Schmidt process to the vectors

\[
\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}
\]

**Problem 3.4.5** Let

\( W = \text{span}\{(-1,0,1,2),(0,1,0,1)\} \).

Find the solution to \( \min_{w \in W} \|w - v\| \) where \( v = (-1,2,6,0) \). (Hint: First use the Gram-Schmidt process.)

**Problem 3.4.6** Show that if \( \{q_1, \ldots, q_n\} \) is an orthonormal basis for an inner product space \( V \). Then for every vector \( v \in V \)

\[
\|v\|^2 = \|q_1, v\|^2 + \cdots + \|q_n, v\|^2
\]
3.5. UNITARY TRANSFORMATIONS AND THE SINGULAR VALUE DECOMPOSITION

Notice that if the original vectors \( \{u_1, u_2, \ldots, u_n\} \) are a basis for a subspace of \( \mathbb{C}^m \) then the conclusions of Theorem 3.4.3 can interpreted as giving a matrix decomposition, the QR decomposition, for a matrix having \( \{u_1, u_2, \ldots, u_n\} \) as columns.

**Theorem 3.4.7** Let \( A \in \mathbb{C}^{m \times n} \) have rank \( n \). Then there exist matrices \( Q \in \mathbb{C}^{m \times n} \) and \( R \in \mathbb{C}^{n \times n} \) so that \( Q^*Q = I \), \( R \) is upper triangular with strictly positive diagonal entries, and

\[
A = QR
\]

**Proof:** The columns of \( A, \{a_1, a_2, \ldots, a_n\} \) form a basis for \( \text{Ran}(A) \). Applying the Gram-Schmidt process to \( \{a_1, a_2, \ldots, a_n\} \) produces orthonormal vectors \( \{q_1, q_2, \ldots, q_n\} \) so that for each \( k = 1, 2, \ldots, n \)

\[
a_k = \sum_{j=1}^{k} r_{jk} q_j
\]

where, in particular, \( r_{kk} = \|w_k\| > 0 \) as defined in the proof of Theorem 3.4.3. This is just a column--by--column description of \( A = QR \) with \( Q = [q_1, \ldots, q_n] \). Orthonormality of \( \{q_1, q_2, \ldots, q_n\} \) is equivalent to \( Q^*Q = I \).

**Problem 3.4.8** Modify the Gram-Schmidt process so that it will produce an orthonormal basis for an inner product space \( V \), starting with any spanning set for \( V, \{u_1, u_2, \ldots, u_n\} \) (not necessarily a basis). How does this change Theorem 3.4.7?

### 3.5 Unitary Transformations and the Singular Value Decomposition

Consider a matrix \( U \in \mathbb{C}^{n \times n} \) satisfying any one of the properties defined below.

- **U preserves length** if for all \( x \in \mathbb{C}^n \),

  \[
  \|Ux\| = \|x\|.
  \]

- **U preserves inner products** if for all \( x \) and \( y \) in \( \mathbb{C}^n \),

  \[
  \langle Ux, Uy \rangle = \langle x, y \rangle.
  \]
• **U** is a unitary matrix if \( U^*L = I \) (that is, if the columns of \( U \) are orthonormal vectors in \( \mathbb{C}^n \)).

Our goal is to show that if \( U \) has any one of these properties, it has the remaining two as well. Since the inner product of two vectors is proportional to the cosine of the angle between them, equivalence of these three properties amounts to the observation that a length preserving transformation also preserves angles, and that such a transformation can be conveniently characterized by a unitary matrix. The action of this transformation is simply a rigid motion of the vectors of \( \mathbb{C}^n \). Such a motion involves only rotations and reflections through coordinate planes.

We introduce some standard notation. Let \( w = u + iv \) be a complex number with \( u \) and \( v \) real. We call \( u \) and \( v \) the real and imaginary parts of \( w \) and denote them by

\[
u = \text{Re}(w), \quad \text{and} \quad v = \text{Im}(w).
\]

Note that

\[
w + \overline{w} = 2\text{Re}(w) \quad \text{and} \quad w - \overline{w} = 2i\text{Im}(w),
\]

\[
\text{Re}(-iw) = \text{Im}(w)
\]

and

\[
ww^* = \text{Re}(w)^2 + \text{Im}(w)^2 := |w|^2.
\]

**Theorem 3.5.1** An \( n \times n \) matrix \( U \) preserves length, if and only if it preserves inner products, and if and only if it is unitary.

**Proof:** Suppose first that \( U \) preserves inner products. Then

\[
\|Ux\|^2 = \langle Ux, Ux \rangle = \langle x, x \rangle = \|x\|^2.
\]

Hence \( U \) preserves length.

Suppose now that \( U \) preserves lengths. We need the following equality whose proof we leave as a problem.

**Problem 3.5.2** Show for all complex numbers \( a \) and \( b \),

\[
|a + b|^2 = |a|^2 + i\overline{a}b + a\overline{b} + |b|^2.
\]
Singular Value Decomposition

By the problem 3.5.2 if \( x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \), then

\[
\|x + y\|^2 = \|x\|^2 + \sum_{i=1}^{n} (x_i y_i) + \|y\|^2
\]

Substituting \( iy \) for \( y \) in (3.7) gives

\[
\|x - iy\|^2 = \|x\|^2 + 2\text{Re}(x,-iy) + \|y\|^2
\]

Applying \( U \) to (3.7) and (3.8) and using the fact that \( U \) preserves lengths gives

\[
\|U(x + y)\|^2 = \|Ux\|^2 + 2\text{Re}(Ux,Uy) + \|Uy\|^2
\]

and

\[
\|U(x - iy)\|^2 = \|Ux\|^2 + 2\text{Im}(Ux,Uy) + \|Uy\|^2
\]

Thus we see that the real and imaginary parts of \( \langle x, y \rangle \) and \( \langle Ux, Uy \rangle \) agree. Hence

\[
\langle Ux, Uy \rangle = \langle x, y \rangle
\]

and \( U \) preserves inner products.

It is clear that if \( U \) is unitary, then it preserves length since

\[
\|Ux\|^2 = \langle Ux, Ux \rangle = \langle U^*Ux, x \rangle = \langle Ix, x \rangle = \langle x, x \rangle = \|x\|^2.
\]

Conversely, if \( U \) preserves length (and so, inner products) then for all \( x \) and \( y \) in \( \mathbb{C}^n \),

\[
\langle Ix, y \rangle = \langle x, y \rangle = \langle Ux, Uy \rangle = \langle U^*Ux, y \rangle
\]

so that \( \langle (U^*U - I)x, y \rangle = 0 \) for all \( x \) and \( y \) in \( \mathbb{C}^n \). But this is possible only if \( (U^*U - I) = 0 \), which is to say, only if \( U \) is unitary. \( \square \)
A matrix is unitary if and only if either the rows or columns of the matrix form an orthonormal basis for \( \mathbb{C}^n \).

Indeed, the columns of the unitary matrix, \( U \), are the vectors
\[
\{ U e_1, \ldots, U e_n \},
\]
where \( e_1, \ldots, e_n \) are the natural basis vectors for \( \mathbb{C}^n \). By the previous theorem, \( U \) preserves both lengths and inner products and so
\[
\langle U e_j, U e_k \rangle = \begin{cases} 
0 & \text{if } j \neq k \\
1 & \text{if } j = k 
\end{cases}
\]
This says that the columns of \( U \) form an orthonormal basis for \( \mathbb{C}^n \). We leave the converse to the reader.

Notice that \( U^* U = I \) implies \( U^{-1} = U^* \), so that \( U U^* = (U^*)^* U^* = I \) and \( U^* \) evidently is unitary as well. But by our preceding discussion this means that the columns of \( U^* \) (which are the conjugates of the rows of \( U \)) form an orthonormal basis for \( \mathbb{C}^n \). Hence the rows of \( U \) themselves form an orthonormal basis for \( \mathbb{C}^n \).

**Problem 3.5.5.3**

- Show that the product of unitary matrices is a unitary matrix.

- Show that \( U \) is a unitary matrix if and only if the larger partitioned matrix
\[
\begin{bmatrix}
I & 0 \\
0 & U
\end{bmatrix}
\]
is unitary.

Two matrices, \( A, B \in \mathbb{C}^{m \times n} \) are said to be \textit{unitarily equivalent} if there are unitary matrices \( U \in \mathbb{C}^{m \times m} \) and \( V \in \mathbb{C}^{n \times n} \), such that \( B = U^* A V \). Likewise, two square matrices, \( A, B \in \mathbb{C}^{n \times n} \) are said to be \textit{unitarily similar} if there is a unitary transformation \( U \in \mathbb{C}^{n \times n} \) such that \( B = U^* A U = U^{-1} A U \).

Perhaps one of the single most useful matrix representations in matrix theory is the \textit{Singular Value Decomposition} (SVD):

**Theorem 3.5.4** Every matrix is unitarily equivalent to a diagonal matrix (of the same size) having nonnegative entries on the diagonal. In particular, suppose \( A \in \mathbb{C}^{m \times n} \) and \( \text{rank}(A) = r \). There exist unitary matrices \( U \in \mathbb{C}^{m \times m} \) and \( V \in \mathbb{C}^{n \times n} \) so that
\[
(3.9) \quad A = U \Sigma V^*
\]
where \( \Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots) \in \mathbb{C}^{m \times n} \) with \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \) and \( \sigma_{r+1} = \cdots = \sigma_p = 0 \) for \( p = \min\{m, n\} \).
Singular Value Decomposition

The columns of $U = [u_1, u_2, \ldots, u_m]$ are called the left singular vectors; the columns of $V = [v_1, v_2, \ldots, v_n]$ are called the right singular vectors; and $\sigma_1, \sigma_2, \ldots$ are the singular values of $A$. We’ll prove this theorem while discussing some adjacent ideas.

The Frobenius norm of a matrix is invariant under unitary equivalence since for any unitary matrices $U \in \mathbb{C}^{m \times m}$ and $V \in \mathbb{C}^{n \times n},$

$$\|U^*AV\|_F^2 = \text{trace}(U^*AVU^*AV)$$
$$= \text{trace}(V^*A^*UU^*AV)$$
$$= \text{trace}(V^*A^*AV)$$
$$= \text{trace}(A^*AVV^*)$$
$$= \text{trace}(A^*A) = \|A\|_F^2$$

If $V$ is partitioned by columns as $V = [v_1, v_2, \ldots, v_n],$ notice that

$$\|A\|_F^2 = \|AV\|_F^2 = \sum_{i=1}^{n} \|Av_i\|^2$$

While different choices of unitary $V$ won’t change the overall sum, they can affect the distribution of magnitudes among the summands. For a given matrix $A$, we will seek to collect the “mass” of the sum as close to the beginning of the summation as possible. In particular, this means we’ll seek an orthonormal basis of $\mathbb{C}^n$ (the columns of $V$):

$$\{v_1, v_2, \ldots, v_n\},$$

that maximizes the sequence of quantities:

$$\|Av_1\|^2 \|Av_1\|^2 + \|Av_2\|^2 \|Av_1\|^2 + \|Av_2\|^2 + \|Av_3\|^2.$$

Although at first blush this may seem hopelessly complicated, notice that the first quantity maximized depends only on $v_1$, the second depends (in effect) only on $v_2$ since we’ve already gotten the best $v_1$, the third quantity depends only on $v_3$ in the same sense, and so on at each step we only are concerned with maximizing with respect to the next $v_k$ in line, having already chosen the best values of all previous $v$s.

The first step proceeds as follows: Define $\sigma_1 = \max_{\|v\|=1} \|Av\|$, let $v_1$ be the maximizing vector, and let $u_1 = \frac{1}{\sigma_1}Av_1$. Now starting with $v_1$, complete an orthonormal basis for $\mathbb{C}^n$ and fill out the associated unitary matrix $V_1$ having $v_1$ as its first column. Likewise, starting with $u_1$ complete
an orthonormal basis for $\mathbb{C}^m$ and fill out the associated unitary matrix $U_1$ having $u_1$ as its first column. Examining the partitioned matrix product of $U_1^*A V_1$ yields

$$U_1^*A V_1 = \begin{bmatrix} \sigma_1 & w^* \\ 0 & \hat{A}_2 \end{bmatrix}$$

The 0 in the $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ location comes from the orthogonality of $u_1$ (which is a multiple of $A v_1$) to all the remaining columns of $U_1$. We now will show that $w = 0$. Suppose we define the vectors $x = \left\{ \sigma_1 \right\}$ and $v = \frac{1}{\|x\|} x$. Then we find

$$A v = \begin{bmatrix} \sigma_1^2 + w^*w \\ \vdots \\ \text{other stuff} \\ \vdots \end{bmatrix}$$

In particular, we have that

$$\|A v\| \geq \sqrt{\sigma_1^2 + w^*w} \geq \sigma_1 = \max_{\|v\|=1} \|A v\|$$

But the last expression is the largest possible value that the previous expressions can attain, so in fact all inequalities are actually equalities, which in turn means that it must be that $w = 0$. At this point, we’ve shown that any matrix is unitarily equivalent to a matrix having first row and column zero except for a nonnegative diagonal entry.

Continuing to the next step, we go through the same construction on $\hat{A}_2$, and define $\sigma_2 = \max_{\|v\|=1} \|\hat{A}_2 v\|$, let $v_2$ be the maximizing vector, and let $\hat{u}_2 = \frac{1}{\sigma_2} \hat{A}_2 v_2$. Now starting with $v_2$, complete an orthonormal basis for $\mathbb{C}^{n-1}$ and fill out the associated unitary matrix $V_2$ having $v_2$ as its first column. Likewise, starting with $u_1$ complete an orthonormal basis for $\mathbb{C}^{m-1}$ and fill out the associated unitary matrix $U_2$ having $u_2$ as its first column. Similar reasoning to that found in the first step above reveals

$$U_2^* \hat{A}_2 V_2 = \begin{bmatrix} \sigma_2 & 0 \\ 0 & \hat{A}_3 \end{bmatrix}$$

**Problem 3.5.5** Explain why $\sigma_2$ as defined above satisfies

$$\sigma_2 \geq \|A v\|$$

for all vectors $v$ with $\|v\| = 1$ and $\langle v, v_1 \rangle = 0$
Singular Value Decomposition

We finish the second step by constructing,

\[ \mathbf{V}_2 = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \hat{\mathbf{V}}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{U}_2 = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \hat{\mathbf{U}}_2 \end{bmatrix} \]

Then one has

\[(\mathbf{U}_1 \mathbf{U}_2)^* \mathbf{A} \mathbf{V}_1 \mathbf{V}_2 = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \mathbf{A}_3 \end{bmatrix} .\]

The construction of the SVD continues in this way.

As an interesting by-product we’ve also solved our problem of “front-loading” of the Frobenius norm sum. In fact, we find

**Theorem 3.5.6** For any matrix \( \mathbf{A} \in \mathbb{C}^{m \times n} \) let \( \sigma_1 \geq \sigma_2 \geq \ldots \) denote its singular values and \( \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \) denote its columns. Then for each \( \ell = 1, 2, \ldots, n - 1 \)

\[ \sum_{i=1}^{\ell} \| \mathbf{a}_i \|^2 \leq \sum_{i=1}^{\ell} \sigma_i^2 \]

with equality for \( \ell = n \).