EECS 16B Designing Information Systems and Devices II UC Berkeley Fall 2022

Note 13: Orthonormalization

1 Overview and Motivation

Thus far, in controls, we have introduced linear models (Note 9), discussed how to learn them from data (Note 10), calculated the asymptotical behavior of the state (Note 11), and developed a way to reach a given target state from a given initial state (Note 12). The next step in our journey is to find a way to reach a given target state from a given initial state *in the best way* (for some reasonable definition of "best"). We will do this in Note 16, but first we will need to introduce some mathematical tools and ideas which are very powerful in their own right.

First among these tools is the idea of *orthonormalization*.

Key Idea 1 (Orthonormalization)
Orthonormalization is the technique of making a set of vectors easier to work with. Given a set of vectors, it allows us to generate a set of orthogonal and normalized vectors which has the same span.

In this note, we will cover what orthonormal vectors are, the idea of projection and its connections to orthonormality, the process of orthonormalization via the Gram-Schmidt algorithm, some consequences of orthonormalization, and some applications to speeding up system identification via the QR decomposition.

2 Orthogonality and Orthonormality

Recall that in EECS 16A we defined two vectors as orthogonal if they are at 90° to each other, or equivalently if their dot product is 0. This can be extended to inner products $\langle\cdot,\ \cdot\rangle$, which are themselves generalizations of dot products, defined in EECS 16A Note 21.

Definition 2 (Orthogonal Vectors)

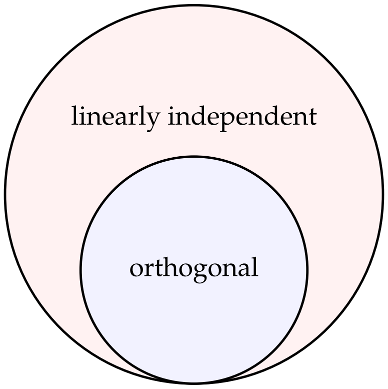
- Let $\vec{x}, \vec{y} \in \mathbb{R}^n$. Then $\vec{x}$ and $\vec{y}$ are *orthogonal* if and only if
$$\langle\vec{x},\ \vec{y}\rangle = 0. \tag{1}$$
- Let S be a set of vectors. Then S is an *orthogonal set* if and only if every pair of distinct vectors in S is orthogonal, i..e,
$$\langle\vec{x},\ \vec{y}\rangle = 0 \qquad \text{for all } \vec{x}, \vec{y} \in S \text{ with } \vec{x} \neq \vec{y}. \tag{2}$$

Orthogonality is a very restrictive condition, in the sense that orthogonal sets are necessarily very structured. For example, we have the following result:

Theorem 3
Suppose $S \subseteq \mathbb{R}^n$ is an orthogonal set of nonzero vectors. Then S is linearly independent.

Concept Check: Prove Theorem 3.

Theorem 3 says that orthogonality is a stricter condition than linear independence; every orthogonal set is linearly independent, but also has some additional structure that makes the inner products 0.



Corollary 4

Suppose $S \subseteq \mathbb{R}^n$ is an orthogonal set of nonzero vectors. Then S contains at most n vectors.

This is true because linearly independent sets in $\mathbb{R}^n$ have at most n vectors.

In addition, we may also talk about two sets of vectors being orthogonal to each other.

Definition 5 (Orthogonality of Sets)

Let $S_1, S_2 \subseteq \mathbb{R}^n$ be two sets of vectors. Then S_1 and S_2 are *orthogonal to each other* if and only if, for all $\vec{x} \in S_1$ and $\vec{y} \in S_2$, we have that $\vec{x}$ and $\vec{y}$ are orthogonal.

NOTE: We can say *one* set is orthogonal – this means that the vectors in the set are pairwise orthogonal, in the sense of Definition 2. Or we can say *two* sets are orthogonal to one another – this means that the vectors in the first set are orthogonal to the vectors in the second set, in the sense of Definition 5. We will keep the meaning clear in these notes, but usually the two different notions are distinguished from context.

For example, if the two sets S_1 and S_2 are subspaces, then by saying S_1 and S_2 are orthogonal subspaces we mean that every vector in S_1 is orthogonal to every vector in S_2 (i.e., Definition 5). This is because non-trivial subspaces have infinitely many vectors, and thus by Corollary 4, they cannot be orthogonal sets in the sense of Definition 2.

Definition 5 is usually useful when S_1 and S_2 are subspaces. In this context, we have the following result, which will be useful in later notes.

Proposition 6

Let B_1 be a basis for the subspace S_1, and B_2 be a basis for the subspace S_2. Then S_1 and S_2 are orthogonal to each other if and only if B_1 and B_2 are orthogonal to each other.

Concept Check: Prove Proposition 6.

In other words, Proposition 6 says that to check orthogonality of subspaces, it is only required to check that their respective bases are orthogonal.

We may also talk about a vector $\vec{x}$ being orthogonal to a set S, which is just shorthand for saying that the sets $\{\vec{x}\}$ and S are orthogonal.

There is also a second part to the definition of orthonormality – that is, normality.

Definition 7 (Normalized Vectors)
Let $\vec{x} \in \mathbb{R}^n$. Then $\vec{x}$ is *normalized* (i.e., has *unit norm*) if and only if $\|\vec{x}\| = 1$.

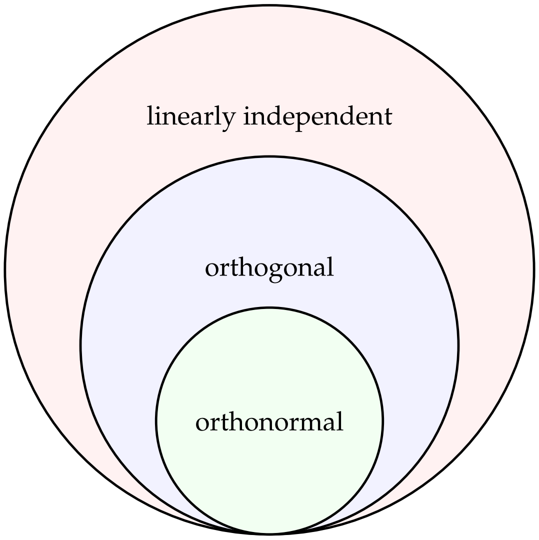
Now we can define orthonormality.

Definition 8 (Orthonormal Vectors)

- Let $\vec{x}, \vec{y} \in \mathbb{R}^n$. Then $\vec{x}$ and $\vec{y}$ are orthonormal if and only if:
  - $\vec{x}$ and $\vec{y}$ are orthogonal; and
  - both $\vec{x}$ and $\vec{y}$ are unit norm.
- Let $S \subseteq \mathbb{R}^n$. Then S is an *orthonormal set* if and only if S is an orthogonal set and all vectors in S have unit norm.

The second definition is equivalent to saying that all vectors in S are unit norm, and also are pairwise orthogonal.

Sets of orthonormal vectors have large amounts of structure, and they are particularly nice to work with. Along with inheriting all properties of orthogonal sets like Theorem 3, i.e.,



We also have properties that are specific to orthonormality, such as the following result, which also provides a way to check for orthonormality.

Theorem 9
Let S be a set of vectors. Then S is an orthonormal set if and only if

$$\langle \vec{x}, \vec{y} \rangle = \begin{cases} 1 & \text{if } \vec{x} = \vec{y} \\ 0 & \text{if } \vec{x} \neq \vec{y} \end{cases} \qquad \text{for all } \vec{x}, \vec{y} \in S. \tag{3}$$

Proof. Indeed, $\vec{x}$ and $\vec{y}$ are orthogonal if and only if $\langle \vec{x}, \vec{y} \rangle = 0$. And $\langle \vec{x}, \vec{x} \rangle = \|\vec{x}\|^2 = 1$ if and only if $\vec{x}$ is normalized. Applying this to all $\vec{x}, \vec{y} \in S$ obtains the claim. □

NOTE: This gives a way to test whether a given set of vectors is orthonormal. If we have an expression for each vector, we can take pairwise inner products and see that they equal to 1 or 0 depending on the vectors we take the inner product of.

Some orthonormality properties manifest nicely as matrix equations with special matrices. For this reason, we define the classes of matrices whose rows or columns are orthonormal sets.

Definition 10 (Orthonormal Matrices)

- A square matrix $Q \in \mathbb{R}^{n \times n}$ whose columns or rows form an orthonormal set is called an *orthonormal matrix*.[a]
- A tall matrix $Q \in \mathbb{R}^{m \times n}$ with $m \geq n$ whose columns form an orthonormal set is said to have *orthonormal columns*.
- A wide matrix $Q \in \mathbb{R}^{m \times n}$ with $m \leq n$ whose rows form an orthonormal set is said to have *orthonormal rows*.

[a] Mathematics literature has some unfortunate terminology for this; namely, in mathematics, such matrices are called *orthogonal*, even if the rows or columns are normalized.

NOTE: By Corollary 4, we cannot have a tall matrix with orthonormal rows, or a wide matrix with orthonormal columns.

We now present the most foundational property of matrices with orthonormal rows or columns.

Theorem 11

(i) If $Q \in \mathbb{R}^{m \times n}$ with $m \geq n$ is a tall matrix, then Q has orthonormal columns if and only if $Q^\top Q = I_n$.

(ii) If $Q \in \mathbb{R}^{m \times n}$ with $m \leq n$ is a wide matrix, then Q has orthonormal rows if and only if $QQ^\top = I_m$.

(iii) If $Q \in \mathbb{R}^{n \times n}$ is a square matrix, then Q is orthonormal if and only if $Q^\top = Q^{-1}$.

The proof of Theorem 11 is on the longer side and may distract from the overall flow of this note, so it is left to Appendix A.1. We fully expect you to read the proof and understand it. It is completely in-scope for the course.

The property $Q^\top = Q^{-1}$ is possibly *the most important property* of orthonormal matrices.

Corollary 12

If $Q \in \mathbb{R}^{n \times n}$ is a square matrix, then it is orthonormal if and only if it has both orthonormal rows and orthonormal columns.

A geometric interpretation of orthonormal matrices, which will be important in the future, is that multiplying a vector by an orthonormal matrix just rotates (and possibly reflects) this vector around the origin, i.e., the length of the vector does not change. In particular, we have the following theorem, which is a little stronger because the desired matrix doesn't need to be square, just to have orthonormal columns.

Theorem 13 (Matrices with Orthonormal Columns are Reflections and Rotations)

(i) Let $Q \in \mathbb{R}^{m\times n}$ with $m \geq n$ have orthonormal columns, and $x \in \mathbb{R}^n$. Then $\|Q\vec{x}\| = \|x\|$.

(ii) Let $Q \in \mathbb{R}^{m\times n}$ with $m \geq n$ have orthonormal columns, and $X \in \mathbb{R}^{n\times p}$. Then $\|QX\|_F = \|X\|_F$. (Recall the *Frobenius norm* from Note 10).

(iii) Let $Q \in \mathbb{R}^{m\times n}$ with $m \leq n$ have orthonormal rows, and $X \in \mathbb{R}^{p\times m}$. Then $\|XQ\|_F = \|X\|_F$.

The proof of Theorem 13 is on the longer side and may distract from the overall flow of this note, so it is left to Appendix A.2. We fully expect you to read the proof and understand it. It is completely in-scope for the course.

3 Projections

Now we will review the concept of projections, which is very much intertwined with the idea of orthonormality. Recall the following definition from EECS 16A Note 21.

Definition 14 (Projection)
Let $\vec{x} \in \mathbb{R}^n$ and define $S \subseteq \mathbb{R}^n$ to be a subspace. Then we define *the projection of $\vec{x}$ onto S* to be the unique point

$$\operatorname{proj}_S(\vec{x}) := \operatorname*{argmin}_{\vec{y}\in S} \|\vec{x} - \vec{y}\|. \tag{4}$$

In other words, $\operatorname{proj}_S(\vec{x})$ is the closest point in S to $\vec{x}$.

Recall from EECS 16A Note 23 the following property of least squares.

Theorem 15 (Least Squares is Projection)
Let $A \in \mathbb{R}^{m\times n}$ with $m \geq n$ have full column rank, and $\vec{y} \in \mathbb{R}^m$. Then

$$\operatorname{proj}_{\operatorname{Col}(A)}(\vec{y}) = A(A^\top A)^{-1}A^\top \vec{y}. \tag{5}$$

We can now augment this theorem with our knowledge of orthonormal matrices.

Theorem 16 (Least Squares with Orthonormal Columns)
Let $Q \in \mathbb{R}^{m\times n}$ with $m \geq n$ have orthonormal columns $\vec{q}_1, \ldots, \vec{q}_n \in \mathbb{R}^m$, and $\vec{y} \in \mathbb{R}^m$. Then

$$\operatorname{proj}_{\operatorname{Col}(Q)}(\vec{y}) = QQ^\top \vec{y} = \sum_{i=1}^{n} \langle \vec{y}, \vec{q}_i \rangle \vec{q}_i. \tag{6}$$

The proof of Theorem 16 is on the longer side and may distract from the overall flow of this note, so it is left to Appendix A.3. We fully expect you to read the proof and understand it. It is completely in-scope for the course.

The following fact will be useful when doing our own orthonormalization computations.

Corollary 17 (Least Squares with Orthogonal Columns)
Let $Z \in \mathbb{R}^{m \times n}$ with $m \geq n$ have orthogonal and nonzero columns $\vec{z}_1, \ldots, \vec{z}_n \in \mathbb{R}^m$. Then

$$\text{proj}_{\text{Col}(Z)}(\vec{y}) = \sum_{i=1}^{n} \frac{\langle \vec{y}, \vec{z}_i \rangle}{\|\vec{z}_i\|^2} \vec{z}_i = \sum_{i=1}^{n} \frac{\langle \vec{y}, \vec{z}_i \rangle}{\langle \vec{z}_i, \vec{z}_i \rangle} \vec{z}_i. \tag{7}$$

Proof. Since $\{\vec{z}_1, \ldots, \vec{z}_n\}$ is an orthogonal set of nonzero vectors, $\{\frac{\vec{z}_1}{\|\vec{z}_1\|}, \ldots, \frac{\vec{z}_n}{\|\vec{z}_n\|}\}$ is an orthonormal set with the same span, which is still $\text{Col}(Z)$. Starting with Theorem 16, we have

$$\text{proj}_{\text{Col}(Z)}(\vec{y}) = \sum_{i=1}^{n} \left\langle \vec{y}, \frac{\vec{z}_i}{\|\vec{z}_i\|} \right\rangle \frac{\vec{z}_i}{\|\vec{z}_i\|} = \sum_{i=1}^{n} \frac{\langle \vec{y}, \vec{z}_i \rangle}{\|\vec{z}_i\|^2} \vec{z}_i = \sum_{i=1}^{n} \frac{\langle \vec{y}, \vec{z}_i \rangle}{\langle \vec{z}_i, \vec{z}_i \rangle} \vec{z}_i. \tag{8}$$

□

We can, in fact, abstract away the idea of projecting onto column spaces and instead discuss projections onto arbitrary subspaces.

Theorem 18 (Projection with Orthogonal or Orthonormal Columns)
Let $S \subseteq \mathbb{R}^m$ be a subspace and $\{\vec{s}_1, \ldots, \vec{s}_n\}$ be a basis of orthogonal vectors for this subspace. Then

$$\text{proj}_S(\vec{x}) = \sum_{i=1}^{n} \frac{\langle \vec{x}, \vec{s}_i \rangle}{\|\vec{s}_i\|^2} \vec{s}_i = \sum_{i=1}^{n} \frac{\langle \vec{x}, \vec{s}_i \rangle}{\langle \vec{s}_i, \vec{s}_i \rangle} \vec{s}_i. \tag{9}$$

If $\{\vec{s}_1, \ldots, \vec{s}_n\}$ is an orthonormal set, then

$$\text{proj}_S(\vec{x}) = \sum_{i=1}^{n} \langle \vec{x}, \vec{s}_i \rangle \vec{s}_i. \tag{10}$$

Concept Check: Use Theorem 16 and Corollary 17 to prove Theorem 18.

This formula is useful in practice – after all, it allows us to mechanically compute projections just by having the basis vectors. It is also useful in theoretical aspects - it lets us prove things, such as the following result, which is inspired by the exposition in EECS 16A Note 23. Namely, the projection formula above allows us to give a much shorter proof of the following theorem, which allows us to determine whether a given vector is a projection onto a given subspace.

Theorem 19 (Orthogonality Principle)
Let $S \subseteq \mathbb{R}^m$ be a linear subspace, and $\vec{x} \in \mathbb{R}^m$ be any vector. Then the following are equivalent:

(i) $\vec{y} = \text{proj}_S(\vec{x})$;

(ii) $\vec{y} - \vec{x}$ is orthogonal to S.

The proof of Theorem 19 is on the longer side and may distract from the overall flow of this note, so it is left to Appendix A.4. We fully expect you to read the proof and understand it. It is completely in-scope for the course.

Since the proof uses theorems that we haven't stated or proved yet, but will do so by the end of the note, we recommend you read the proof after reading the rest of the note. No other theorems in this note depend on Theorem 19.

4 Gram-Schmidt Orthonormalization

Now that we have defined orthonormal vectors and sets, and shown off some of their properties, we can now learn how to convert sets of vectors into orthonormal sets of vectors which have the same span. The easiest way to do this is, for each vector in turn, we subtract off the component of this vector which is in the direction of all the previous vectors we processed. Mathematically, this component is the projection of the vector onto the span of the previous vectors.

So really, we can execute the following pseudocode, given some vectors $\vec{a}_1, \dots, \vec{a}_\ell$ that we want to orthonormalize:

```
• Set $\vec{q}_1 := \frac{\vec{a}_1}{\|\vec{a}_1\|}$.
• For $i = 2, 3, \dots, \ell$:
    – Set $\vec{z}_i := \vec{a}_i - \text{proj}_{\text{Span}(\vec{q}_1,\dots,\vec{q}_{i-1})}(\vec{a}_i)$.
    – Set $\vec{q}_i := \frac{z_i}{\|\vec{z}_i\|}$.
```

Intuitively, this gives us a nice picture (in the case $\ell = n = 2$):

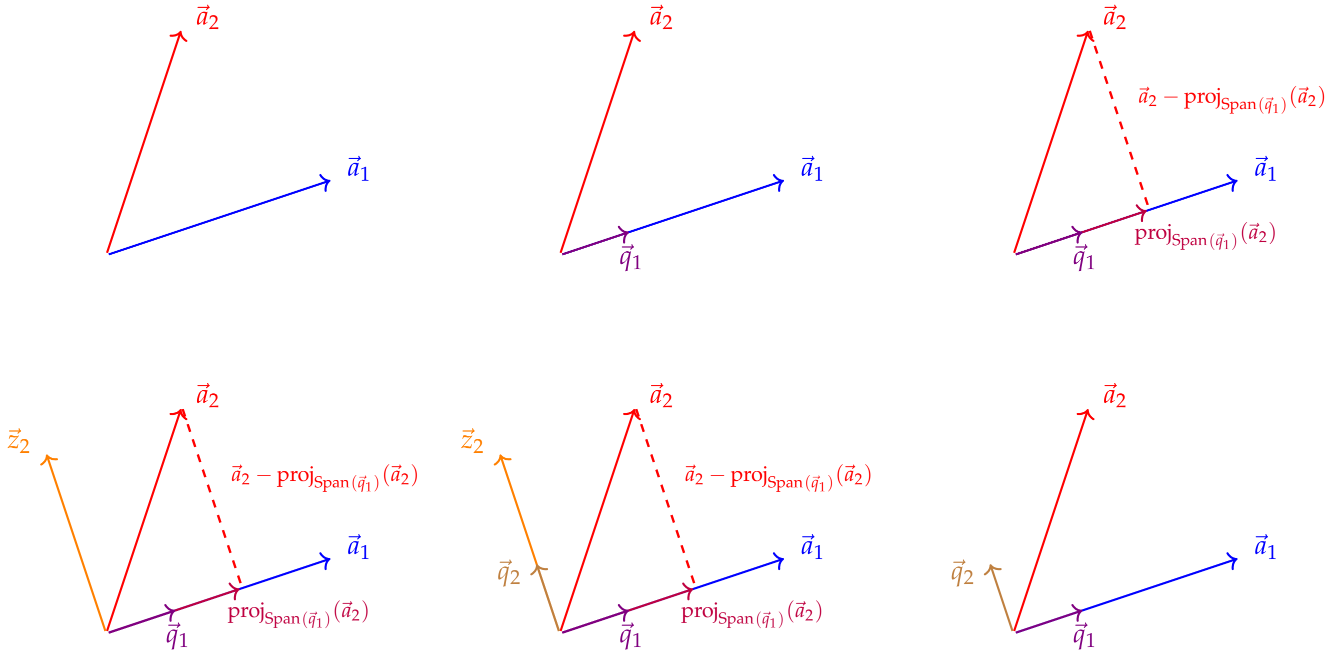

$\vec{a}_2$
$\vec{a}_1$
$\vec{q}_1$
$\vec{a}_2 - \text{proj}_{\text{Span}(\vec{q}_1)}(\vec{a}_2)$
$\text{proj}_{\text{Span}(\vec{q}_1)}(\vec{a}_2)$
$\vec{z}_2$
$\vec{q}_2$

Figure 1: Orthonormalization in action, going left to right then top to bottom.

Wikipedia also has a very nice animation of this process.

There are two questions remaining:

- How we would calculate $\text{proj}_{\text{Span}(\vec{q}_1,\dots,\vec{q}_{i-1})}(\vec{a}_k)$. Indeed, Theorem 16 tells us that

$$\text{proj}_{\text{Span}(\vec{q}_1,\dots,\vec{q}_{i-1})}(\vec{a}_i) = \sum_{k=1}^{i-1} \langle \vec{a}_i,\ \vec{q}_k \rangle\, \vec{q}_k. \tag{11}$$

- What to do if any $\vec{z}_i = \vec{0}$ (and so we cannot divide by its norm to get $\vec{q}_i$). In this case it depends what exactly we want to do with the algorithm, but usually we can just throw it out, as it does not contribute to the span.

All of this put together exactly forms the *Gram-Schmidt orthogonalization algorithm.*

Algorithm 20 Gram-Schmidt Orthogonalization Algorithm

Input: A set of vectors $\{\vec{a}_1, \ldots, \vec{a}_\ell\} \subseteq \mathbb{R}^n$.

Output: An orthonormal set of vectors $\{\vec{q}_1, \ldots, \vec{q}_l\}$ which spans the same set.

```
function GramSchmidt($\{\vec{a}_1, \ldots, \vec{a}_\ell\}$)
    for $i = 1, \ldots, \ell$ do
        $\vec{z}_i := \vec{a}_i - \sum_{k=1}^{i-1} \langle \vec{a}_i, \vec{p}_k \rangle \vec{p}_k$        ▷ If $i = 1$ then there are no $k$ to sum over, so $\vec{z}_1 = \vec{a}_1$.
        if $\vec{z}_i = \vec{0}_n$ then
            $\vec{p}_i = \vec{0}_n$
        else
            $\vec{p}_i := \frac{\vec{z}_i}{\|\vec{z}_i\|}$
        end if
    end for
    return $\{\vec{q}_1, \ldots, \vec{q}_l\} := \{\vec{p}_1, \ldots, \vec{p}_\ell\}$ with the $\ell - l$ zero vectors discarded.
end function
```

Here we are being very pedantic with the checks that $\vec{z}_i = \vec{0}$, etc, because this function will be extremely critical in a variety of procedures we will come up with. As such, we would like to have well-defined behavior in every possible case.

The main result related to Algorithm 20 is the following theorem, which states that the Gram-Schmidt process accomplishes what we want to do. This type of theorem is called a *proof of correctness* in the literature.

Theorem 21 (Correctness of Gram-Schmidt Orthogonalization Algorithm)
Let $\{\vec{a}_1, \ldots, \vec{a}_\ell\} \subseteq \mathbb{R}^n$ be a set of vectors and $\{\vec{q}_1, \ldots, \vec{q}_l\} := \text{GramSchmidt}(\{\vec{a}_1, \ldots, \vec{a}_\ell\})$. Then

(i) $\{\vec{q}_1, \ldots, \vec{q}_l\}$ is an orthonormal set of vectors.

(ii) $\text{Span}(\vec{q}_1, \ldots, \vec{q}_l) = \text{Span}(\vec{a}_1, \ldots, \vec{a}_\ell)$.

Further, if $\{\vec{a}_1, \ldots, \vec{a}_\ell\}$ are linearly independent then

(iii) $\ell = l$. (That is, no vectors are discarded.)

(iv) $\text{Span}(\vec{q}_1, \ldots, \vec{q}_i) = \text{Span}(\vec{a}_1, \ldots, \vec{a}_i)$ for all $1 \leq i \leq \ell$.

The proof of Theorem 21 is on the longer side and may distract from the overall flow of this note, so it is left to Appendix A.5. We fully expect you to read the proof and understand it. It is completely in-scope for the course.

An example of this algorithm being run is in Section 8.1.

5 Consequences of Orthonormalization

Once we know that we are able to orthonormalize finite sets of vectors, we have the following abstract result.

Theorem 22 (Existence of Orthonormal Basis)
Let $S \subseteq \mathbb{R}^n$ be a subspace. There exists an orthonormal basis for S.

Proof. Take any basis for S and run Gram-Schmidt on it. □

There is also a constructive way to *extend* orthonormal bases for subspaces to other subspaces.

Theorem 23 (Extending an Orthonormal Basis)
Let $S \subseteq T \subseteq \mathbb{R}^n$ be subspaces. Then for every orthonormal basis B_S of S, there exists an orthonormal basis B_T for T that contains B_S.

NOTE: We mean that B_T contains B_S in the following sense: each of the $\dim(S)$ vectors in B_S is also one of the $\dim(T)$ vectors in B_T.

Proof. Take any basis B of T and run Gram-Schmidt on $B_S \cup B$ to get B_T. Because the orthonormal vectors in B_S are processed first by Gram-Schmidt, B_S is unchanged by Gram-Schmidt, so B_S is contained in B_T. And since B spans T, we know that $B_S \cup B$ spans T, so B_T is a basis for T. □

NOTE: Here we assume that we are using *ordered sets* (which can essentially be thought of as lists or arrays) to hold our bases. In this context, $B_S \cup B$ can be thought of as the concatenation of the lists representing B_S and B, with the elements of B_S going first.

Algorithmically, we can write this process in the following way.

Algorithm 24 Basis Extension

Input: Subspaces $S \subseteq T \subseteq \mathbb{R}^n$, an orthonormal basis B_S for S.

Output: An orthonormal basis B_T for T that contains B_S.

```
function ExtendBasis($B_S$, $T$)
    $B$ := any basis for $T$.
    return $B_T$ := GramSchmidt($B_S \cup B$).
end function
```

NOTE: The basis extension algorithm is most commonly used with $T = \mathbb{R}^n$, i.e., in the context of extending orthonormal bases for a subspace S to an orthonormal basis for $\mathbb{R}^n$. One common choice for an arbitrary basis B of T is to pick B to be the columns of the identity matrix, i.e., $B = \{\vec{e}_1, \ldots, \vec{e}_n\}$.

An example of this algorithm being run is in Section 8.2.

6 (OPTIONAL) The QR Decomposition

Theorem 21 tells us that, if $\{\vec{a}_1, \ldots, \vec{a}_\ell\}$ is linearly independent, then each $\vec{a}_i$ can be written in terms of $\vec{q}_1, \ldots, \vec{q}_i$ only. This gives the linear system writing $\vec{a}_1, \ldots, \vec{a}_\ell$ in terms of the $\vec{q}_k$ a *triangular* structure. More formally, we have the following decomposition:

Theorem 25 (QR Decomposition)
Let $A \in \mathbb{R}^{n \times \ell}$ have full column rank. Then there exists a matrix $Q \in \mathbb{R}^{n \times \ell}$ with orthonormal columns, and an *upper triangular* matrix $R \in \mathbb{R}^{\ell \times \ell}$ (i.e., $r_{ij} = 0$ for $i > j$), such that

$$A = QR. \tag{12}$$

The proof of Theorem 25 is on the longer side and may distract from the overall flow of this note, so it is left to Appendix A.6.

The proof also reveals a constructive way to reach a QR decomposition.

Algorithm 26 QR Decomposition

Input: A matrix $A \in \mathbb{R}^{n \times \ell}$ of full column rank.

Output: A matrix Q with orthonormal columns, and an upper-triangular matrix R, such that $A = QR$.

```
function QR($A$)
    $\{\vec{a}_1, \ldots, \vec{a}_\ell\}$ := columns of $A$
    $\{\vec{q}_1, \ldots, \vec{q}_\ell\}$ := GramSchmidt($\{\vec{a}_1, \ldots, \vec{a}_\ell\}$)
    $Q := \begin{bmatrix} \vec{q}_1 & \cdots & \vec{q}_\ell \end{bmatrix} \in \mathbb{R}^{n \times \ell}$
    $R := \begin{bmatrix} \langle \vec{a}_1, \vec{q}_1 \rangle & \langle \vec{a}_2, \vec{q}_1 \rangle & \cdots & \langle \vec{a}_\ell, \vec{q}_1 \rangle \\ 0 & \langle \vec{a}_2, \vec{q}_2 \rangle & \cdots & \langle \vec{a}_\ell, \vec{q}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle \vec{a}_\ell, \vec{q}_\ell \rangle \end{bmatrix} \in \mathbb{R}^{\ell \times \ell}$
    return $Q, R$
end function
```

7 (OPTIONAL) Application: Speeding Up Model Selection

Previously, in system identification (see Note 10), we knew exactly how many parameters our model had, and were able to design an algorithm to learn the parameters from data. In practice, we sometimes do not know how many parameters our model has; this is related to the problem of *model selection* in machine learning literature.

More formally, suppose we have several datasets $D_1 \in \mathbb{R}^{n \times 1}, D_2 \in \mathbb{R}^{n \times 2}, \ldots, D_\ell \in \mathbb{R}^{n \times \ell}$, where the columns of D_i are a strict subset of the columns of D_{i+1}. In the context of system ID, D_i is the dataset if we include the i "most relevant" states and try to build a model off of that.[1] We also have an output vector $\vec{s} \in \mathbb{R}^n$. We want to learn parameter vectors $\vec{p}_1, \ldots, \vec{p}_\ell$ where $\vec{p}_i \in \mathbb{R}^i$ such that $D_i \vec{p}_i \approx \vec{s}$. With these estimates, we can evaluate which $\vec{p}_i$ performs best in practice (or via *validation*), and thereby learn a compact (minimal states used) yet useful model for the system.

[1] There are two questions here. First, why would we not use all the states? The answer is that in most modern applications, there are usually many possible states to include, and most of them don't have any effect on the results, so including them in the model is both computationally inefficient and can cause an "overfitting" phenomenon where the parameters use the irrelevant states to erroneously capture the effects of noise. Second, how would we pick the "most relevant" states? The answer is not not simple; there are many ways to do this, and it is a very large problem in statistics and control theory, but it suffices to say that there are many approaches to this "feature selection" that work in practice.

One way to do this is our system identification technique of least squares. In particular, we solve a sequence of least squares problems

$$\vec{\widehat{p}}_i \in \operatorname*{argmin}_{\vec{p}_i \in \mathbb{R}^i} \|D_i \vec{p}_i - \vec{s}\| \tag{13}$$

which has solution

$$\vec{\widehat{p}}_i := (D_i^\top D_i)^{-1} D_i^\top \vec{s}. \tag{14}$$

The issue is that for large n, computing $(D_i^\top D_i)^{-1}$ for every i is computationally inefficient, as measured by the number of floating point operations required for an algorithm to do the computation. Matrix multiplication to compute $D_i^\top D_i$ takes roughly in^2 floating point operations, and then taking the inverse takes roughly i^3 floating point operations. Summing this up over all $i \in \{1, \dots, \ell\}$ means that this process takes roughly $\ell^4 + \ell^2 n^2$ floating point operations, which is extremely large (and thus extremely slow to compute) when n and ℓ are large.

Instead, what we can do is use the fact that least squares is a projection. The following algorithm speeds up this repeated least squares a lot.

- Take the QR decomposition D_ℓ to get matrices $Q_\ell := \begin{bmatrix} \vec{q}_1 & \dots & \vec{q}_\ell \end{bmatrix}$ and R_ℓ as defined in the QR decomposition, which takes around $\ell^2 n$ floating point operations. In addition, for notation's sake we can define $Q_i := \begin{bmatrix} \vec{q}_1 & \dots & \vec{q}_i \end{bmatrix}$.

- Let $\vec{\widehat{w}}_1 := \vec{q}_1^\top \vec{s}$; this amounts to efficiently calculating $Q_1^\top \vec{s}$.

- For each $i \in \{2, \dots, \ell\}$, form $\vec{\widehat{w}}_i$ by appending the entry $\vec{q}_i^\top \vec{s}$ to $\vec{\widehat{w}}_{i-1}$; this amounts to efficiently calculating $Q_i^\top \vec{s}$.

- Note that $\vec{\widehat{w}}_i \neq \vec{\widehat{p}}_i$, because we changed the data from D_i to Q_i. To relate the two, we note that

$$D_i \vec{\widehat{p}}_i = D_i (D_i^\top D_i)^{-1} D_i^\top \vec{s} = \text{proj}_{\text{Col}(D_i)}(\vec{s}) = \text{proj}_{\text{Col}(Q_i)}(\vec{s}) = Q_i Q_i^\top \vec{s} = Q_i \vec{\widehat{w}}_i. \tag{15}$$

  In particular, decomposing $D_i = Q_i R_i$ where R_i is the upper left $i \times i$ sub-block of R_ℓ (and is in particular upper triangular), we have

$$Q_i R_i \vec{\widehat{p}}_i = Q_i \vec{\widehat{w}}_i. \tag{16}$$

  Left-multiplying by Q_i on both sides, we get

$$R_i \vec{\widehat{p}}_i = \vec{\widehat{w}}_i. \tag{17}$$

  Since R_i is upper triangular, we can solve for $\vec{\widehat{p}}_i$ efficiently using backward substitution.

Overall this takes much fewer floating point operations – certainly fewer than $\ell^4 + \ell^2 n^2$.

Concept Check: How many floating point operations does our new algorithm require, in terms of ℓ and n?

8 Examples

We give an extended example for the two main algorithms we present in this note.

8.1 Gram-Schmidt

Say we have two vectors in $\mathbb{R}^3$:

$$\vec{a}_1 := \begin{bmatrix} 3 \\ 4 \\ 0 \end{bmatrix} \qquad \vec{a}_2 := \begin{bmatrix} 4 \\ 3 \\ 0 \end{bmatrix} \tag{18}$$

that we want to run Gram-Schmidt on. Mechanically, we have

$$\vec{p}_1 := \frac{\vec{a}_1}{\|\vec{a}_1\|} = \frac{1}{5}\begin{bmatrix} 3 \\ 4 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix}. \tag{19}$$

Now we compute the projection residual:

$$\begin{aligned}
\vec{z}_2 &= \vec{a}_2 - \langle \vec{a}_2, \vec{p}_1 \rangle \vec{p}_1 && (20)\\
&= \begin{bmatrix} 4 \\ 3 \\ 0 \end{bmatrix} - \left\langle \begin{bmatrix} 4 \\ 3 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} \right\rangle \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} && (21)\\
&= \begin{bmatrix} 4 \\ 3 \\ 0 \end{bmatrix} - \left(4 \cdot \frac{3}{5} + 3 \cdot \frac{4}{5} + 0 \cdot 0\right) \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} && (22)\\
&= \begin{bmatrix} 4 \\ 3 \\ 0 \end{bmatrix} - \frac{24}{5} \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} && (23)\\
&= \begin{bmatrix} 4 \\ 3 \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{72}{25} \\ \frac{96}{25} \\ 0 \end{bmatrix} && (24)\\
&= \begin{bmatrix} \frac{28}{25} \\ -\frac{21}{25} \\ 0 \end{bmatrix}. && (25)
\end{aligned}$$

Then we can normalize it to get $\vec{p}_2$ as

$$\begin{aligned}
\vec{p}_2 &= \frac{\vec{z}_2}{\|\vec{z}_2\|} && (26)\\
&= \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix}. && (27)
\end{aligned}$$

Since $\vec{p}_1 \neq \vec{0}_3$ and $\vec{p}_2 \neq \vec{0}_3$, we output

$$\vec{q}_1 = \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} \qquad \text{and} \qquad \vec{q}_2 = \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix}. \tag{28}$$

8.2 Extending Basis

Let's say we wanted to extend our basis $\{\vec{q}_1, \vec{q}_2\}$ to a basis for $\mathbb{R}^3$. To do this, we would pick a basis for $\mathbb{R}^3$, usually the identity basis $\{\vec{e}_1, \vec{e}_2, \vec{e}_3\}$. Then we would run Gram-Schmidt on $\{\vec{q}_1, \vec{q}_2, \vec{e}_1, \vec{e}_2, \vec{e}_3\}$. Since $\mathbb{R}^3$ is three-dimensional, any basis for $\mathbb{R}^3$ contains 3 vectors, so the output of Gram-Schmidt will also contain 3 vectors, which is an orthonormal basis for $\mathbb{R}^3$.

Anyways, we re-run Gram-Schmidt on this basis. The computation for $\vec{p}_1$ and $\vec{p}_2$ goes the same way as the previous section since we are running it on the same vectors in the same order, so we get

$$\vec{p}_1 := \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} \qquad \text{and} \qquad \vec{p}_2 := \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix}. \tag{29}$$

Now $\vec{p}_3, \vec{p}_4, \vec{p}_5$ are new. Let's start with computing the projection residual:

$$\begin{aligned}
\vec{z}_3 &= \vec{e}_1 - \langle \vec{e}_1,\ \vec{p}_1 \rangle \vec{p}_1 - \langle \vec{e}_1,\ \vec{p}_2 \rangle \vec{p}_2 && (30)\\
&= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \left\langle \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} \right\rangle \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} - \left\langle \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix} \right\rangle \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix} && (31)\\
&= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \frac{3}{5} \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} - \frac{4}{5} \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix} && (32)\\
&= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{9}{25} \\ \frac{12}{25} \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{16}{25} \\ -\frac{12}{25} \\ 0 \end{bmatrix} && (33)\\
&= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. && (34)
\end{aligned}$$

This implies that $\vec{p}_3 = \vec{0}_3$ and therefore there is no corresponding $\vec{q}$ vector. Similarly,

$$\begin{aligned}
\vec{z}_4 &= \vec{e}_2 - \langle \vec{e}_2,\ \vec{p}_1 \rangle \vec{p}_1 - \langle \vec{e}_2,\ \vec{p}_2 \rangle \vec{p}_2 - \langle \vec{e}_2,\ \vec{p}_3 \rangle \underbrace{\vec{p}_3}_{=\vec{0}_3} && (35)\\
&= \vec{e}_2 - \langle \vec{e}_2,\ \vec{p}_1 \rangle \vec{p}_1 - \langle \vec{e}_2,\ \vec{p}_2 \rangle \vec{p}_2 && (36)\\
&= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} - \left\langle \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} \right\rangle \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} - \left\langle \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix} \right\rangle \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix} && (37)\\
&= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} - \frac{4}{5} \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} - \left(-\frac{3}{5}\right) \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix} && (38)\\
&= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{12}{25} \\ \frac{16}{25} \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{12}{5} \\ -\frac{9}{25} \\ 0 \end{bmatrix} && (39)
\end{aligned}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{40}$$

Again, this implies that $\vec{p}_4 = \vec{0}_3$ and therefore there is no corresponding $\vec{q}$ vector.

Finally, computing the residual for $\vec{e}_3$ gets us

$$\vec{z}_5 = \vec{e}_3 - \langle \vec{e}_3, \vec{p}_1 \rangle \vec{p}_1 - \langle \vec{e}_3, \vec{p}_2 \rangle \vec{p}_2 - \langle \vec{e}_3, \vec{p}_3 \rangle \underbrace{\vec{p}_3}_{=\vec{0}_3} - \langle \vec{e}_3, \vec{p}_4 \rangle \underbrace{\vec{p}_4}_{=\vec{0}_3} \tag{41}$$

$$= \vec{e}_3 - \langle \vec{e}_3, \vec{p}_1 \rangle \vec{p}_1 - \langle \vec{e}_3, \vec{p}_2 \rangle \vec{p}_2 \tag{42}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - \left\langle \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} \right\rangle \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} - \left\langle \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix} \right\rangle \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix} \tag{43}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - 0 \cdot \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix} - 0 \cdot \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix} \tag{44}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{45}$$

Then we have

$$\vec{p}_5 := \frac{\vec{z}_5}{\|\vec{z}_5\|} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{46}$$

Thus $\vec{q}_3 = \vec{p}_5 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. Gram-Schmidt outputs

$$\vec{q}_1 = \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \\ 0 \end{bmatrix}, \qquad \vec{q}_2 = \begin{bmatrix} \frac{4}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix}, \qquad \vec{q}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{47}$$

which is indeed an orthonormal basis for $\mathbb{R}^3$.

A Longer Proofs

A.1 Proof of Theorem 11

Proof of Theorem 11.

(i) Suppose the columns of Q are $\vec{q}_1, \ldots, \vec{q}_n \in \mathbb{R}^m$. Then we have

$$\begin{aligned} Q^\top Q &= \begin{bmatrix} \vec{q}_1 & \cdots & \vec{q}_n \end{bmatrix}^\top \begin{bmatrix} \vec{q}_1 & \cdots & \vec{q}_n \end{bmatrix} && (48) \\ &= \begin{bmatrix} \vec{q}_1^\top \\ \vdots \\ \vec{q}_n^\top \end{bmatrix} \begin{bmatrix} \vec{q}_1 & \cdots & \vec{q}_n \end{bmatrix} && (49) \\ &= \begin{bmatrix} \vec{q}_1^\top \vec{q}_1 & \cdots & \vec{q}_1^\top \vec{q}_n \\ \vdots & \ddots & \vdots \\ \vec{q}_n^\top \vec{q}_1 & \cdots & \vec{q}_n^\top \vec{q}_n \end{bmatrix} && (50) \\ &= \begin{bmatrix} \langle \vec{q}_1, \vec{q}_1 \rangle & \cdots & \langle \vec{q}_n, \vec{q}_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \vec{q}_1, \vec{q}_n \rangle & \cdots & \langle \vec{q}_n, \vec{q}_n \rangle \end{bmatrix} && (51) \end{aligned}$$

By Theorem 9, we see that $Q^\top Q = I_n$ if and only if $\{\vec{q}_1, \ldots, \vec{q}_n\}$ are orthonormal.

(ii) Apply (i) to $Q^\top \in \mathbb{R}^{n \times m}$, which is a tall matrix. Note that

$$QQ^\top = (Q^\top)^\top (Q^\top). \tag{52}$$

This tells us that

$$QQ^\top = I_m \tag{53}$$

if and only if $Q^\top$ has orthonormal columns, i.e., Q has orthonormal rows.

(iii) Suppose Q is orthogonal. Then it has orthonormal columns or orthonormal rows. If it has orthonormal columns then, by (i), $Q^\top Q = I_n$; if it has orthonormal rows then, by (ii), $QQ^\top = I_n$. Either way, $Q^\top = Q^{-1}$ as desired. Going the other way, if $Q^\top = Q^{-1}$ then $Q^\top Q = I_n$, which confirms by (i) that Q has orthonormal columns and is thus orthonormal.

□

A.2 Proof of Theorem 13

Proof of Theorem 13.

(i)

$$\|Q\vec{x}\| = \sqrt{\|Q\vec{x}\|^2} = \sqrt{\langle Q\vec{x}, Q\vec{x} \rangle} = \sqrt{\langle \vec{x}, Q^\top Q\vec{x} \rangle} = \sqrt{\langle \vec{x}, \vec{x} \rangle} = \sqrt{\|\vec{x}\|^2} = \|\vec{x}\|. \tag{54}$$

(ii)

$$\|QX\|_F = \sqrt{\|QX\|_F^2} = \sqrt{\sum_{i=1}^{p} \|(QX)_i\|^2} = \sqrt{\sum_{i=1}^{p} \|Q\vec{x}_i\|^2} = \sqrt{\sum_{i=1}^{p} \|\vec{x}_i\|^2} = \sqrt{\|X\|_F^2} = \|X\|_F. \tag{55}$$

(iii) $Q^\top$ has orthonormal columns. Then

$$\|XQ\|_F = \left\|(XQ)^\top\right\|_F = \left\|Q^\top X^\top\right\|_F = \left\|X^\top\right\|_F = \|X\|_F. \tag{56}$$

□

A.3 Proof of Theorem 16

Proof of Theorem 16. Since Q has orthonormal columns, by Theorem 3, we know that the columns of Q are linearly independent, so Q has full column rank. By Theorem 15, we have

$$\text{proj}_{\text{Col}(Q)}(\vec{y}) = Q(Q^\top Q)^{-1}Q^\top \vec{y}. \tag{57}$$

By Theorem 11, since Q has orthonormal columns, we know that $Q^\top Q = I_n$. Thus

$$\begin{aligned}
\text{proj}_{\text{Col}(Q)}(\vec{y}) &= Q(Q^\top Q)^{-1}Q^\top \vec{y} && (58)\\
&= Q(I_n)^{-1}Q^\top \vec{y} && (59)\\
&= QQ^\top \vec{y} && (60)\\
&= \begin{bmatrix} \vec{q}_1 & \cdots & \vec{q}_n \end{bmatrix} \begin{bmatrix} \vec{q}_1^\top \\ \vdots \\ \vec{q}_n^\top \end{bmatrix} \vec{y} && (61)\\
&= \begin{bmatrix} \vec{q}_1 & \cdots & \vec{q}_n \end{bmatrix} \begin{bmatrix} \vec{q}_1^\top \vec{y} \\ \vdots \\ \vec{q}_n^\top \vec{y} \end{bmatrix} && (62)\\
&= \begin{bmatrix} \vec{q}_1 & \cdots & \vec{q}_n \end{bmatrix} \begin{bmatrix} \langle \vec{y}, \vec{q}_1 \rangle \\ \vdots \\ \langle \vec{y}, \vec{q}_n \rangle \end{bmatrix} && (63)\\
&= \sum_{i=1}^{n} \langle \vec{y}, \vec{q}_i \rangle \vec{q}_i. && (64)
\end{aligned}$$

□

A.4 Proof of Theorem 19

Proof of Theorem 19. For now, we will take for granted the fact that S has an orthonormal basis $\{\vec{s}_1, \ldots, \vec{s}_n\} \subset \mathbb{R}^m$; the existence of such a basis is proved in Theorem 22, which does not rely on this theorem. We will also take for granted the fact that this basis can be extended to an orthonormal basis $\{\vec{s}_1, \ldots, \vec{s}_m\}$ for $\mathbb{R}^m$; this extension exists by Theorem 23, which also does not rely on this theorem.

The proof technique used here is very common – *when we have an orthonormal basis, write everything in terms of this basis, and lots of stuff will cancel.*

(i) $\Longrightarrow$ (ii) By Theorem 18, we know that

$$\vec{y} = \text{proj}_S(\vec{x}) = \sum_{i=1}^{n} \langle \vec{x}, \vec{s}_i \rangle \vec{s}_i. \tag{65}$$

Now $\vec{x}$ can be written as its projection onto $\mathbb{R}^m$, that is,

$$\vec{x} = \text{proj}_{\mathbb{R}^m}(\vec{x}) = \sum_{i=1}^{m} \langle \vec{x},\, \vec{s}_i \rangle \vec{s}_i. \tag{66}$$

Now

$$\vec{x} - \vec{y} = \sum_{i=1}^{m} \langle \vec{x},\, \vec{s}_i \rangle \vec{s}_i - \sum_{i=1}^{n} \langle \vec{x},\, \vec{s}_i \rangle \vec{s}_i = \sum_{i=n+1}^{m} \langle \vec{x},\, \vec{s}_i \rangle \vec{s}_i. \tag{67}$$

Now $\vec{x} - \vec{y}$ is a linear combination of $\vec{s}_{n+1}, \dots, \vec{s}_m$, and so $\vec{x} - \vec{y} \in \text{Span}(\vec{s}_{n+1}, \dots, \vec{s}_m)$. Since $\{\vec{s}_1, \dots, \vec{s}_m\}$ is an orthonormal set, we have that $\{\vec{s}_1, \dots, \vec{s}_n\}$ and $\{\vec{s}_{n+1}, \dots, \vec{s}_m\}$ are orthogonal as sets, so by Proposition 6, $S = \text{Span}(\vec{s}_1, \dots, \vec{s}_n)$ and $\text{Span}(\vec{s}_{n+1}, \dots, \vec{s}_m)$ are orthogonal subspaces. Thus S is orthogonal to $\vec{x} - \vec{y}$ and (ii) holds.

(ii) $\implies$ (i) Let $\vec{z}$ be any vector in S other than $\vec{y}$. Then

$$\begin{aligned}
\|\vec{z} - \vec{x}\|^2 &= \|(\vec{z} - \vec{y}) + (\vec{y} - \vec{x})\|^2 && (68)\\
&= \langle (\vec{z} - \vec{y}) + (\vec{y} - \vec{x}),\, (\vec{z} - \vec{y}) + (\vec{y} - \vec{x}) \rangle && (69)\\
&= \langle \vec{z} - \vec{y},\, \vec{z} - \vec{y} \rangle + \langle \vec{z} - \vec{y},\, \vec{y} - \vec{x} \rangle + \langle \vec{y} - \vec{x},\, \vec{z} - \vec{y} \rangle + \langle \vec{y} - \vec{x},\, \vec{y} - \vec{x} \rangle && (70)\\
&= \|\vec{z} - \vec{y}\|^2 + 2\langle \vec{z} - \vec{y},\, \vec{y} - \vec{x} \rangle + \|\vec{y} - \vec{x}\|^2 && (71)\\
&= \|\vec{z} - \vec{y}\|^2 + \|\vec{y} - \vec{x}\|^2 && (72)\\
&> \|\vec{y} - \vec{x}\|^2 && (73)
\end{aligned}$$

where we used the fact that $\vec{z} - \vec{y} \in S$ since S is a subspace, then invoked (ii). Thus

$$\|\vec{z} - \vec{x}\|^2 > \|\vec{y} - \vec{x}\|^2 \qquad \text{for all } \vec{z} \in S \setminus \{\vec{y}\} \tag{74}$$

which implies that $\vec{y} = \text{proj}_S(\vec{x})$ and thus (i) holds.

□

A.5 Proof of Theorem 21

Proof of Theorem 21. Let $\{\vec{a}_1, \dots, \vec{a}_\ell\} \subseteq \mathbb{R}^n$ be a set of vectors and $\{\vec{q}_1, \dots, \vec{q}_l\} := \text{GRAMSCHMIDT}(\{\vec{a}_1, \dots, \vec{a}_\ell\})$. For notation's sake, let $\vec{z}_1 := \vec{a}_1$.

The usual flow for a proof of correctness is to show that each line, or block of lines, in the algorithm does what we want it to do. Given that the general flow of the algorithm is to build from $\vec{a}_i$'s to $\vec{z}_i$'s to $\vec{p}_i$'s to $\vec{q}_i$'s, we will start with proving facts about the $\vec{z}_i$'s.

(a) We want to show some claims about the $\vec{z}_i$'s:

(I) $\{\vec{z}_1, \dots, \vec{z}_\ell\}$ is an orthogonal set.

(II) For each i we have $\text{Span}(\vec{z}_1, \dots, \vec{z}_i) = \text{Span}(\vec{a}_1, \dots, \vec{a}_i)$.

(III) $\vec{z}_i = \vec{0}_n$ if and only if $\vec{a}_i \in \text{Span}(\vec{a}_1, \dots, \vec{a}_{i-1})$.

To do this, we first write the $\vec{z}_i$ in terms of $\vec{a}_i$ and other $\vec{z}_k$.

$$\vec{z}_i = \vec{a}_i - \sum_{k=1}^{i-1} \langle \vec{a}_i,\, \vec{p}_i \rangle \vec{p}_i \tag{75}$$

$$= \vec{a}_i - \sum_{\substack{k=1 \\ \vec{p}_k \neq \vec{0}_n}}^{i-1} \langle \vec{a}_i, \vec{p}_k \rangle \vec{p}_k - \sum_{\substack{k=1 \\ \vec{p}_k = \vec{0}_n}}^{i-1} \langle \vec{a}_i, \vec{p}_k \rangle \underbrace{\vec{p}_k}_{=\vec{0}_n} \tag{76}$$

$$= \vec{a}_i - \sum_{\substack{k=1 \\ \vec{p}_k \neq \vec{0}_n}}^{i-1} \langle \vec{a}_i, \vec{p}_k \rangle \vec{p}_k \tag{77}$$

$$= \vec{a}_i - \sum_{\substack{k=1 \\ \vec{z}_k \neq \vec{0}_n}}^{i-1} \left\langle \vec{a}_i, \frac{\vec{z}_k}{\|\vec{z}_k\|} \right\rangle \frac{\vec{z}_k}{\|\vec{z}_k\|} \tag{78}$$

$$= \vec{a}_i - \sum_{\substack{k=1 \\ \vec{z}_k \neq \vec{0}_n}}^{i-1} \frac{\langle \vec{a}_i, \vec{z}_k \rangle}{\|\vec{z}_k\|^2} \vec{z}_k \tag{79}$$

$$= \vec{a}_i - \sum_{\substack{k=1 \\ \vec{z}_k \neq \vec{0}_n}}^{i-1} \frac{\langle \vec{a}_i, \vec{z}_k \rangle}{\langle \vec{z}_k, \vec{z}_k \rangle} \vec{z}_k. \tag{80}$$

This is an expression for $\vec{z}_i$ in terms of $\vec{z}_k$ for $k < i$, as well as $\vec{a}_i$. Note that we summed over only k such that $\vec{z}_k \neq \vec{0}_n$; this is to avoid division by 0 errors.

Now we are ready to prove our claims.

(I) We show that $\{\vec{z}_1, \ldots, \vec{z}_\ell\}$ is an orthogonal set. We do this by showing that for each i, $\vec{z}_i$ is orthogonal to all the $\vec{z}_k$ for $k < i$. We will show this for each i starting at $i = 1$ and proceeding iteratively.[2]

We start with $i = 1$, where there are no $\vec{z}_k$ for $k < i$, and so we are done by definition – we have shown that $\vec{z}_i$ is orthogonal to all the $\vec{z}_k$ coming before it (because there are no such $\vec{z}_k$ and the statement becomes tautological).

For our general case, we show that $\vec{z}_i$ is orthogonal to all $\vec{z}_k$ before it, i.e., for $k < i$. If $\vec{z}_k = \vec{0}_n$ then

$$\langle \vec{z}_i, \vec{z}_k \rangle = \left\langle \vec{z}_i, \vec{0}_n \right\rangle = 0 \tag{81}$$

so $\vec{z}_i$ is orthogonal to $\vec{z}_k$. We now just need to show that $\vec{z}_i$ is orthogonal to all $\vec{z}_k \neq \vec{0}_n$ for $k < i$. These are exactly the terms appearing in the sum in Equation (80). Indeed, we can take the inner product and use Equation (80) to get

$$\langle \vec{z}_i, \vec{z}_k \rangle = \left\langle \vec{a}_i - \sum_{\substack{j=1 \\ \vec{z}_j \neq \vec{0}_n}}^{i-1} \frac{\langle \vec{a}_i, \vec{z}_j \rangle}{\langle \vec{z}_j, \vec{z}_j \rangle} \vec{z}_j, \vec{z}_k \right\rangle \tag{82}$$

$$= \langle \vec{a}_i, \vec{z}_k \rangle - \left\langle \sum_{\substack{j=1 \\ \vec{z}_j \neq \vec{0}_n}}^{i-1} \frac{\langle \vec{a}_i, \vec{z}_j \rangle}{\langle \vec{z}_j, \vec{z}_j \rangle} \vec{z}_j, \vec{z}_k \right\rangle \tag{83}$$

$$= \langle \vec{a}_i, \vec{z}_k \rangle - \sum_{\substack{j=1 \\ \vec{z}_j \neq \vec{0}_n}}^{i-1} \frac{\langle \vec{a}_i, \vec{z}_j \rangle}{\langle \vec{z}_j, \vec{z}_j \rangle} \langle \vec{z}_j, \vec{z}_k \rangle \tag{84}$$

[2] If we had CS70 or an equivalent class as a prerequisite, this would be essentially an induction proof. But again, knowing the formalism for induction isn't needed for reading and understanding this.

Here we used bilinearity of the inner product to separate out terms and move scalars to the outside of the inner product. So now we have a bunch of inner products of the form $\langle \vec{z}_j, \vec{z}_k \rangle$ for $1 \leq j, k < i$. Since we already processed steps $1, \dots, i-1$ by the time we get to step i in our iteration, we already know that $\vec{z}_j$ is orthogonal to all the $\vec{z}_m$ for $m < j$, and similarly that $\vec{z}_k$ is orthogonal to all the $\vec{z}_m$ for $m < k$. Thus, if one of j or k is less than the other, then they are orthogonal; otherwise, $j = k$.

Anyways, now that we know $\langle \vec{z}_j, \vec{z}_k \rangle = 0$ unless $j = k$, we can write

$$\begin{aligned}
\langle \vec{z}_i, \vec{z}_k \rangle &= \langle \vec{a}_i, \vec{z}_k \rangle - \sum_{\substack{j=1 \\ \vec{z}_j \neq \vec{0}_n}}^{i-1} \frac{\langle \vec{a}_i, \vec{z}_j \rangle}{\langle \vec{z}_j, \vec{z}_j \rangle} \langle \vec{z}_j, \vec{z}_k \rangle && (85) \\
&= \langle \vec{a}_i, \vec{z}_k \rangle - \frac{\langle \vec{a}_i, \vec{z}_k \rangle}{\langle \vec{z}_k, \vec{z}_k \rangle} \langle \vec{z}_k, \vec{z}_k \rangle - \sum_{\substack{j=1 \\ \vec{z}_j \neq \vec{0}_n \\ j \neq k}}^{i-1} \frac{\langle \vec{a}_i, \vec{z}_j \rangle}{\langle \vec{z}_j, \vec{z}_j \rangle} \underbrace{\langle \vec{z}_j, \vec{z}_k \rangle}_{=0} && (86) \\
&= \langle \vec{a}_i, \vec{z}_k \rangle - \frac{\langle \vec{a}_i, \vec{z}_k \rangle}{\langle \vec{z}_k, \vec{z}_k \rangle} \langle \vec{z}_k, \vec{z}_k \rangle && (87) \\
&= \langle \vec{a}_i, \vec{z}_k \rangle - \langle \vec{a}_i, \vec{z}_k \rangle && (88) \\
&= 0 && (89)
\end{aligned}$$

and thus $\vec{z}_i$ is orthogonal to $\vec{z}_k$. This shows that $\vec{z}_i$ is orthogonal to $\vec{z}_1, \dots, \vec{z}_{i-1}$ for each i. Thus we have that $\{\vec{z}_1, \dots, \vec{z}_i\}$ is an orthogonal set for every i, and so $\{\vec{z}_1, \dots, \vec{z}_\ell\}$ is an orthogonal set.

(II) We show that $\operatorname{Span}(\vec{z}_1, \dots, \vec{z}_i) = \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_i)$ for each i. This is equivalent to showing that $\vec{z}_i$ can be written as a linear combination of $\vec{a}_1, \dots, \vec{a}_i$, and $\vec{a}_i$ can be written as a linear combination of $\vec{z}_1, \dots, \vec{z}_i$, for each i.

We first claim that $\vec{a}_i$ can be written as a linear combination of $\vec{z}_1, \dots, \vec{z}_i$, for each i. Indeed, by rearranging Equation (80), we have

$$\vec{a}_i = \vec{z}_i + \sum_{\substack{k=1 \\ \vec{z}_k \neq \vec{0}_n}}^{i-1} \frac{\langle \vec{a}_i, \vec{z}_k \rangle}{\langle \vec{z}_k, \vec{z}_k \rangle} \vec{z}_k. \tag{90}$$

This is a linear combination of $\vec{z}_1, \dots, \vec{z}_i$ whose value is $\vec{a}_i$, and so we are done.

Now we claim that $\vec{z}_i$ can be written as a linear combination of $\vec{a}_1, \dots, \vec{a}_i$. We show this for every $1 \leq i \leq n$, starting at $i = 1$ and proceeding iteratively.

We start with $i = 1$, in which case $\vec{z}_1 = \vec{a}_1$ is the linear combination we want.

The general case starts again with Equation (80), giving us:

$$\vec{z}_i = \vec{a}_i - \sum_{\substack{k=1 \\ \vec{z}_k \neq \vec{0}_n}}^{i-1} \frac{\langle \vec{a}_i, \vec{z}_k \rangle}{\langle \vec{z}_k, \vec{z}_k \rangle} \vec{z}_k. \tag{91}$$

Since $k < i$, we have already processed $\vec{z}_k$ in our iteration, and thus can write $\vec{z}_k = \sum_{j=1}^{k} \alpha_{jk} \vec{a}_k$, i.e., generate a linear combination of $\vec{a}_1, \dots, \vec{a}_k$ whose value is $\vec{z}_k$. This gets us

$$\vec{z}_i = \vec{a}_i - \sum_{\substack{k=1 \\ \vec{z}_k \neq \vec{0}_n}}^{i-1} \frac{\langle \vec{a}_i, \vec{z}_k \rangle}{\langle \vec{z}_k, \vec{z}_k \rangle} \vec{z}_k \tag{92}$$

$$= \vec{a}_i - \sum_{\substack{k=1 \\ \vec{z}_k \neq \vec{0}_n}}^{i-1} \frac{\langle \vec{a}_i,\, \vec{z}_k \rangle}{\langle \vec{z}_k,\, \vec{z}_k \rangle} \sum_{j=1}^{k} \alpha_{jk} \vec{a}_j \tag{93}$$

$$= \vec{a}_i - \sum_{\substack{k=1 \\ \vec{z}_k \neq \vec{0}_n}}^{i-1} \sum_{j=1}^{k} \alpha_{jk} \frac{\langle \vec{a}_i,\, \vec{z}_k \rangle}{\langle \vec{z}_k,\, \vec{z}_k \rangle} \vec{a}_j \tag{94}$$

which is a linear combination of $\vec{a}_1, \dots, \vec{a}_i$ (remember that the term $\frac{\langle \vec{a}_i,\, \vec{z}_k \rangle}{\langle \vec{z}_k,\, \vec{z}_k \rangle}$ is just a scalar). Thus we are done and the spans are equal.

(III) Finally, we want to show that $\vec{z}_i = \vec{0}_n$ if and only if $\vec{a}_i \in \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_{i-1})$. Indeed, we start with Equation (80), use Theorem 18 to write the sum as a projection, and then simplify with the fact that adding zero vectors to a set maintains the span of the set.

$$\vec{z}_i = \vec{a}_i - \sum_{\substack{k=1 \\ \vec{z}_k \neq \vec{0}_n}}^{i-1} \frac{\langle \vec{a}_i,\, \vec{z}_k \rangle}{\langle \vec{z}_k,\, \vec{z}_k \rangle} \vec{z}_k \tag{95}$$

$$= \vec{a}_i - \operatorname{proj}_{\operatorname{Span}\left(\vec{z}_k \colon 1 \leq k < i,\, \vec{z}_k \neq \vec{0}_n\right)}(\vec{a}_i) \tag{96}$$

$$= \vec{a}_i - \operatorname{proj}_{\operatorname{Span}(\vec{z}_1, \dots, \vec{z}_{i-1})}(\vec{a}_i). \tag{97}$$

We can now use (II) to simplify even more, since $\operatorname{Span}(\vec{z}_1, \dots, \vec{z}_{i-1}) = \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_{i-1})$.

$$\vec{z}_i = \vec{a}_i - \operatorname{proj}_{\operatorname{Span}(\vec{z}_1, \dots, \vec{z}_{i-1})}(\vec{a}_i) \tag{98}$$

$$= \vec{a}_i - \operatorname{proj}_{\operatorname{Span}(\vec{a}_1, \dots, \vec{a}_{i-1})}(\vec{a}_i). \tag{99}$$

Thus $\vec{z}_i = \vec{0}_n$ if and only if $\vec{a}_i = \operatorname{proj}_{\operatorname{Span}(\vec{a}_1, \dots, \vec{a}_{i-1})}(\vec{a}_i)$, i.e., if and only if $\vec{a}_i \in \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_{i-1})$.

Phew! That was quite a lot. What just transpired is one of the single most involved proofs in this course. But it's necessary to completely understand Gram-Schmidt, which is an extremely critical sub-routine in many concepts we will cover.

If you've come this far, take a breather, and then dive right back into it. We promise, the rest of the proof is *a lot* simpler.

(b) We now want to show some claims about the $\vec{p}_i$'s:

(I) $\{\vec{p}_1, \dots, \vec{p}_\ell\}$ is an orthogonal set.

(II) For every i, either $\vec{p}_i = \vec{0}_n$ or $\|\vec{p}_i\| = 1$.

(III) For each i we have $\operatorname{Span}(\vec{p}_1, \dots, \vec{p}_i) = \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_i)$.

(IV) $\vec{p}_i = \vec{0}_n$ if and only if $\vec{a}_i \in \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_{i-1})$.

We can show them rather quickly, as a result of our previous claims about the $\vec{z}_i$'s. The crucial observation to make here is that, for each i, $\vec{p}_i$ is a scalar multiple of $\vec{z}_i$, say $\vec{p}_i = \alpha_i \vec{z}_i$. If $\vec{z}_i = \vec{0}$ we may take $\alpha_i = 1$; otherwise we can take $\alpha_i = \frac{1}{\|\vec{z}_i\|}$. Thus in either case $\alpha_i \neq 0$. With this framework, we now give the proofs.

(I) Let $1 \leq k < i \leq \ell$. Then

$$\langle \vec{p}_i,\, \vec{p}_k \rangle = \langle \alpha_i \vec{z}_i,\, \alpha_k \vec{z}_k \rangle \tag{100}$$

$$= \alpha_i \alpha_k \langle \vec{z}_i, \vec{z}_k \rangle \tag{101}$$

$$= 0 \tag{102}$$

since $\{\vec{z}_1, \dots, \vec{z}_\ell\}$ is an orthogonal set. Thus $\{\vec{p}_1, \dots, \vec{p}_\ell\}$ is an orthogonal set.

(II) Take i such that $\vec{z}_i = \vec{0}_n$. Then by the algorithm, $\vec{p}_i = \vec{0}_n$, so

$$\|\vec{p}_i\| = \left\|\vec{0}_n\right\| = 0. \tag{103}$$

Now take i such that $\vec{z}_i \neq \vec{0}_n$. Then

$$\|\vec{p}_i\| = \|\alpha_i \vec{z}_i\| = \alpha_i \|\vec{z}_i\| = \frac{\|\vec{z}_i\|}{\|\vec{z}_i\|} = 1. \tag{104}$$

(III) For each i we have $\operatorname{Span}(\vec{z}_1, \dots, \vec{z}_i) = \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_i)$. Since $\vec{p}_i = \alpha_i \vec{z}_i$ where $\alpha_i \neq 0$, we know that $\operatorname{Span}(\vec{z}_1, \dots, \vec{z}_i) = \operatorname{Span}(\vec{p}_1, \dots, \vec{p}_i)$. Thus $\operatorname{Span}(\vec{a}_1, \dots, \vec{a}_i) = \operatorname{Span}(\vec{p}_1, \dots, \vec{p}_i)$.

(IV) We have that $\vec{p}_i = \alpha_i \vec{z}_i$ with $\alpha_i \neq 0$, so $\vec{p}_i = \vec{0}_n$ if and only if $\vec{z}_i = \vec{0}_n$. We have already shown that this is true if and only if $\vec{a}_i \in \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_{i-1})$, so we obtain the conclusion.

Nearly there! Once we have the facts proven above, we can prove the required original claims made in Theorem 21. Suppose for now that $\{\vec{a}_1, \dots, \vec{a}_\ell\}$ is a general set of vectors.

(i) To show that $\{\vec{q}_1, \dots, \vec{q}_l\}$ is an orthonormal set, we use the fact that $\{\vec{p}_1, \dots, \vec{p}_\ell\}$ is an orthogonal set where each vector is either the zero vector or has unit norm. Since $\{\vec{q}_1, \dots, \vec{q}_l\}$ is $\{\vec{p}_1, \dots, \vec{p}_\ell\}$ with all zero vectors removed, we have that $\{\vec{q}_1, \dots, \vec{q}_l\}$ is an orthogonal set where all vectors have unit norm, and is hence an orthonormal set.

(ii) To show that $\operatorname{Span}(\vec{q}_1, \dots, \vec{q}_l) = \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_\ell)$, note that we already know $\operatorname{Span}(\vec{p}_1, \dots, \vec{p}_\ell) = \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_\ell)$. Since $\{\vec{q}_1, \dots, \vec{q}_l\}$ is $\{\vec{p}_1, \dots, \vec{p}_\ell\}$ with all zero vectors removed, and zero vectors don't contribute to any span, we have that $\operatorname{Span}(\vec{q}_1, \dots, \vec{q}_l) = \operatorname{Span}(\vec{p}_1, \dots, \vec{p}_\ell)$. Putting this together, we have that $\operatorname{Span}(\vec{q}_1, \dots, \vec{q}_l) = \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_\ell)$ as desired.

Now suppose that $\{\vec{a}_1, \dots, \vec{a}_\ell\}$ is a linearly independent set of vectors.

(iii) To show that $l = \ell$, we show that no $\vec{p}_i$ is a zero vector that gets discarded. Indeed, we know that $\vec{p}_i = \vec{0}_n$ if and only if $\vec{a}_i = \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_{i-1})$. But since $\{\vec{a}_1, \dots, \vec{a}_\ell\}$ is a linearly independent set, this never happens, and thus $\vec{p}_i \neq \vec{0}_n$ for all i. Thus no discards happen, so $l = \ell$.

(iv) Since $l = \ell$, we have $\vec{p}_i = \vec{q}_i$ for all i. Since we know $\operatorname{Span}(\vec{p}_1, \dots, \vec{p}_i) = \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_i)$ for each i, we also have that $\operatorname{Span}(\vec{q}_1, \dots, \vec{q}_i) = \operatorname{Span}(\vec{a}_1, \dots, \vec{a}_i)$ for each i, as desired.

□

A.6 Proof of Theorem 25

Proof of Theorem 25. By Theorem 21, we can let $\{\vec{q}_1, \dots, \vec{q}_\ell\}$ be the output of Gram-Schmidt when run on the columns of A (the theorem invocation is to ensure that we receive ℓ vectors back). we have

$$\vec{a}_i = \operatorname{proj}_{\operatorname{Span}(\vec{a}_1, \dots, \vec{a}_i)}(\vec{a}_i) \tag{105}$$

$$= \operatorname{proj}_{\operatorname{Span}(\vec{q}_1, \dots, \vec{q}_i)}(\vec{a}_i) \tag{106}$$

$$= \sum_{k=1}^{i} \langle \vec{a}_i, \vec{q}_k \rangle \vec{q}_k \tag{107}$$

$$= \begin{bmatrix} \vec{q}_1 & \cdots & \vec{q}_i \end{bmatrix} \begin{bmatrix} \langle \vec{a}_i, \vec{q}_1 \rangle \\ \vdots \\ \langle \vec{a}_i, \vec{q}_i \rangle \end{bmatrix} \tag{108}$$

$$= \begin{bmatrix} \vec{q}_1 & \cdots & \vec{q}_i & \vec{q}_{i+1} & \cdots & \vec{q}_\ell \end{bmatrix} \begin{bmatrix} \langle \vec{a}_i, \vec{q}_1 \rangle \\ \vdots \\ \langle \vec{a}_i, \vec{q}_i \rangle \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{109}$$

Putting these columns together, we get

$$A = \begin{bmatrix} \vec{a}_1 & \cdots & \vec{a}_\ell \end{bmatrix} \tag{110}$$

$$= \begin{bmatrix} \vec{q}_1 & \cdots & \vec{q}_\ell \end{bmatrix} \begin{bmatrix} \langle \vec{a}_1, \vec{q}_1 \rangle & \langle \vec{a}_2, \vec{q}_1 \rangle & \cdots & \langle \vec{a}_\ell, \vec{q}_1 \rangle \\ 0 & \langle \vec{a}_2, \vec{q}_2 \rangle & \cdots & \langle \vec{a}_\ell, \vec{q}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle \vec{a}_\ell, \vec{q}_\ell \rangle \end{bmatrix} \tag{111}$$

which is the desired decomposition. □

Contributors:

- Druv Pai.
- Ashwin Vangipuram.
- Anant Sahai.
- Jennifer Shih.
- Rachel Hochman.
- Vasuki Narasimha Swamy.
- Steven Cao.