6.4 The Gram-Schmidt Process

This section introduces an algorithm for producing an orthogonal basis for any nonzero subspace of $\mathbb{R}^n$ given any basis. The idea is to start with an ordered basis $\{\bar{x}_1, \ldots, \bar{x}_p\}$ for a nonzero subspace W of $\mathbb{R}^n$. We then create a new basis out of the old basis iteratively, where each new vector is orthogonal to all the preceding vectors (hence also linearly independent from all the preceding vectors). When this process terminates, we are left with an orthogonal basis for W.

Theorem (The Gram-Schmidt Process)
Given a basis $\{\bar{x}_1, \ldots, \bar{x}_p\}$ for a nonzero subspace W of $\mathbb{R}^n$, define

$$\begin{aligned}
\bar{v}_1 &= \bar{x}_1 \\
\bar{v}_2 &= \bar{x}_2 - \frac{\bar{x}_2 \cdot \bar{v}_1}{\bar{v}_1 \cdot \bar{v}_1}\bar{v}_1 \\
\bar{v}_3 &= \bar{x}_3 - \frac{\bar{x}_3 \cdot \bar{v}_1}{\bar{v}_1 \cdot \bar{v}_1}\bar{v}_1 - \frac{\bar{x}_3 \cdot \bar{v}_2}{\bar{v}_2 \cdot \bar{v}_2}\bar{v}_2 \\
&\vdots \\
\bar{v}_p &= \bar{x}_p - \frac{\bar{x}_p \cdot \bar{v}_1}{\bar{v}_1 \cdot \bar{v}_1}\bar{v}_1 - \frac{\bar{x}_p \cdot \bar{v}_2}{\bar{v}_2 \cdot \bar{v}_2}\bar{v}_2 - \cdots - \frac{\bar{x}_p \cdot \bar{v}_{p-1}}{\bar{v}_{p-1} \cdot \bar{v}_{p-1}}\bar{v}_{p-1}.
\end{aligned}$$

Then $\{\bar{v}_1, \ldots, \bar{v}_p\}$ is an orthogonal basis for W satisfying

$$\text{Span}\{\bar{v}_1, \ldots, \bar{v}_k\} = \text{Span}\{\bar{x}_1, \ldots, \bar{x}_k\} \qquad \text{for all } 1 \le k \le p.$$

At each step $2 \le k \le p$ in the construction of the basis $\{\bar{v}_1, \ldots, \bar{v}_p\}$, we are removing the $\bar{v}_1, \ldots, \bar{v}_{k-1}$ components from $\bar{x}_k$ in order to get $\bar{v}_k$. This is what ensures that $\bar{v}_k$ is orthogonal to all of the preceding vectors $\bar{v}_1, \ldots, \bar{v}_{k-1}$. Note also that at each step $2 \le k \le p$, we have that each of the vectors $\bar{x}_1, \ldots, \bar{x}_{k-1}$ is a linear combination of $\bar{v}_1, \ldots, \bar{v}_{k-1}$ (and vice-versa), so the spans of the respective sets coincide. Once we have an orthogonal basis $\{\bar{v}_1, \ldots, \bar{v}_k\}$ of W, it is easy to find an ortho*normal* basis by normalizing each vector; i.e.

$$\left\{ \frac{\bar{v}_1}{\|\bar{v}_1\|}, \quad \frac{\bar{v}_2}{\|\bar{v}_2\|}, \quad \ldots, \quad \frac{\bar{v}_p}{\|\bar{v}_p\|} \right\}$$

is an orthonormal basis of W.

Example Let

$$\bar{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \bar{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \bar{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

Then $\{\bar{x}_1, \bar{x}_2, \bar{x}_3\}$ is easily seen to be linearly independent, hence is a basis for $W = \text{Span}\{\bar{x}_1, \bar{x}_2, \bar{x}_3\}$. However, it is easy to see that this basis is not orthogonal, since, for example, $\bar{x}_1 \cdot \bar{x}_2 = 3$. Our task is to find an orthogonal basis for W. After that, we will normalize our orthogonal basis to get an orthonormal basis. It works like this:

$$\begin{aligned}
\bar{v}_1 &= \bar{x}_1 \\
\bar{v}_2 &= \bar{x}_2 - \frac{\bar{x}_2 \cdot \bar{v}_1}{\bar{v}_1 \cdot \bar{v}_1}\bar{v}_1 \\
&= \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} - \frac{3}{4}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} -\frac{3}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix}.
\end{aligned}$$

Now we find $\bar{v}_3$, as follows:

$$\begin{aligned}
\bar{v}_3 &= \bar{x}_3 - \frac{\bar{x}_3 \cdot \bar{v}_1}{\bar{v}_1 \cdot \bar{v}_1}\bar{v}_1 - \frac{\bar{x}_3 \cdot \bar{v}_2}{\bar{v}_2 \cdot \bar{v}_2}\bar{v}_2 \\
&= \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} - \frac{2}{4}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - \frac{\frac{2}{4}}{\frac{12}{16}}\begin{bmatrix} -\frac{3}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - \frac{2}{3}\begin{bmatrix} -\frac{3}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix} \\
&= \begin{bmatrix} -\frac{1}{2} + \frac{1}{2} \\ -\frac{1}{2} - \frac{1}{6} \\ \frac{1}{2} - \frac{1}{6} \\ \frac{1}{2} - \frac{1}{6} \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ -\frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}.
\end{aligned}$$

Hence

$$\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -\frac{3}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix}, \begin{bmatrix} 0 \\ -\frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix} \right\}$$

is an orthogonal basis for W. To find an orthonormal basis, we normalize each vector:

$$\begin{aligned}
\|\bar{v}_1\| &= \sqrt{4} = 2 \\
\|\bar{v}_2\| &= \sqrt{\frac{12}{16}} = \frac{\sqrt{3}}{2} \\
\|\bar{v}_2\| &= \sqrt{\frac{6}{9}} = \sqrt{\frac{2}{3}}.
\end{aligned}$$

Hence an orthonormal basis for W is

$$\left\{ \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \frac{2}{\sqrt{3}}\begin{bmatrix} -\frac{3}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix}, \sqrt{\frac{3}{2}}\begin{bmatrix} 0 \\ -\frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix} \right\}.$$

6.5 Least-Squares Problems

In applications it may be the case that a solution to some linear system does not exist, but we can get by with a "solution" that is close enough. So if we are considering a matrix equation $A\bar{x} = \bar{b}$, it makes sense to study how we can get as close as possible to $\bar{b}$ using linear combinations of the columns of A. How should we make this formal? Notice that making $A\bar{x}$ close to $\bar{b}$ is the same as minimizing the distance between $A\bar{x}$ and $\bar{b}$. In other words, what we seek is some vector, let's call it $\hat{x}$ (read x "hat"), such that $\|A\hat{x} - \bar{b}\|$ is as small as possible. This idea leads to the following definition.

Definition If A is an $m \times n$ matrix and $\bar{b} \in \mathbb{R}^m$, a **least-squares solution** of the equation $A\bar{x} = \bar{b}$ is a vector $\hat{x} \in \mathbb{R}^m$ such that

$$\|A\hat{x} - \bar{b}\| \leq \|A\bar{x} - \bar{b}\|$$

for all $\bar{x} \in \mathbb{R}^m$.

So how do we actually find such an element $\hat{x}$? Recall the *Best Approximation Theorem* from section 6.3 which said that if we project $\bar{b}$ onto some subspace H of $\mathbb{R}^m$, then the result is the vector $\hat{b} \in H$ that is closest to $\bar{b}$. In our special case, we are considering the column space of A, so we compute

$$\hat{b} = \mathrm{proj}_{\mathrm{Col}A}\bar{b}.$$

Now, since $\hat{b}$ is in Col$A$, we can find a solution to $A\bar{x} = \hat{b}$, and this solution is $\hat{x}$.

Let's do a few manipulations. If $\hat{x}$ satisfies $A\hat{x} = \hat{b}$, then we know that $\bar{b} - \hat{b}$ is orthogonal to Col$A$. In other words, $\bar{b} - A\hat{x}$ is orthogonal to *every column* of A. In particular

$$\bar{a}_i^T(\bar{b} - A\hat{x}) = 0$$

for each column $\bar{a}_i$ of A (remember that the expression on the left hand side is just the dot product of $\bar{a}_i$ with the vector $\bar{b} - A\hat{x}$). Running over all columns of $\bar{a}_i$, we get

$$A^T(\bar{b} - A\hat{x}) = 0,$$

where 0 now represents the zero vector of $\mathbb{R}^n$. Rearranging, we see that

$$A^TA\hat{x} = A^T\bar{b}.$$

This leads to the following theorem whose proof we omit. (The proof is short, and you can read it in the book if you are interested.)

Theorem The set of least-squares solutions of $A\bar{x} = \bar{b}$ is the nonempty set of solutions of $A^TA\hat{x} = A^T\bar{b}$. (We call the equations implied by the formula the **normal equations**.)

Example Find a least-squares solution of $A\bar{x} = \bar{b}$, where

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 0 & 1 \end{bmatrix}, \qquad \bar{b} = \begin{bmatrix} 3 \\ 3 \\ 2 \end{bmatrix}.$$

By the theorem, we are seeking the solution(s) of $A^TA\hat{x} = A^T\bar{b}$. What we need for this computation are A^TA and $A^T\bar{b}$. Observe that

$$\begin{aligned} A^TA &= \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 4 \\ 4 & 6 \end{bmatrix}, \\ A^T\bar{b} &= \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 9 \\ 11 \end{bmatrix}. \end{aligned}$$

Also note that A^TA is invertible, and by the inverse formula for a 2×2 matrix we have

$$(A^TA)^{-1} = \frac{1}{14}\begin{bmatrix} 6 & -4 \\ -4 & 5 \end{bmatrix}.$$

Since A^TA is invertible, we can solve (uniquely!) for $\hat{x}$. Specifically,

$$\hat{x} = (A^TA)^{-1}A^T\bar{b} = \frac{1}{14}\begin{bmatrix} 6 & -4 \\ -4 & 5 \end{bmatrix}\begin{bmatrix} 9 \\ 11 \end{bmatrix} = \frac{1}{14}\begin{bmatrix} 10 \\ 19 \end{bmatrix}.$$

How can we know, in general, if the solution to $A^TA\hat{x} = A^T\bar{b}$ will be unique? The following theorem gives several characterizations (we just saw (c) in action).

Theorem Let A be an $m \times n$ matrix. The following statements are equivalent:

(a) The equation $A\bar{x} = \bar{b}$ has a *unique* least-squares solution for each $\bar{b} \in \mathbb{R}^m$.
(b) The columns of A are linearly independent.
(c) The matrix A^TA is invertible.

If these statements are true, we have an explicit expression for $\hat{x}$ given by

$$\hat{x} = (A^TA)^{-1}A^T\bar{b}.$$

Note We call the matrix $(A^TA)^{-1}A^T$ in the above theorem the *pseudo-inverse* of A. Notice, in particular, that if A is invertible, then

$$(A^TA)^{-1}A^T = A^{-1}(A^T)^{-1}A^T = A^{-1}.$$

In other words, the expression $(A^TA)^{-1}A^T$ reduces to A^{-1} in the special case that A is invertible. If A is not invertible, it behaves similar to an inverse of A in the sense that it helps us "solve" the equation $A\bar{x} = \bar{b}$.

Definition When a least-squares solution $\hat{x}$ is used to find $A\hat{x}$ as an approximation to $\bar{b}$, we call the quantity $\|A\hat{x} - \bar{b}\|$ the **least-squares** error of the approximation.

Example What is the least-squares error in the previous example? First note that

$$\bar{b} = \begin{bmatrix} 3 \\ 3 \\ 2 \end{bmatrix}, \qquad A\hat{x} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 0 & 1 \end{bmatrix}\frac{1}{14}\begin{bmatrix} 10 \\ 19 \end{bmatrix} = \frac{1}{14}\begin{bmatrix} 48 \\ 39 \\ 19 \end{bmatrix}.$$

Then

$$\begin{aligned} \bar{b} - A\hat{x} &= \frac{1}{14}\begin{bmatrix} 42 \\ 42 \\ 28 \end{bmatrix} - \frac{1}{14}\begin{bmatrix} 48 \\ 39 \\ 19 \end{bmatrix} \\ &= \frac{1}{14}\begin{bmatrix} -6 \\ 3 \\ 9 \end{bmatrix}. \end{aligned}$$

Hence

$$\begin{aligned} \|\bar{b} - A\hat{x}\| &= \frac{1}{14}\sqrt{(-6)^2 + 3^2 + 9^2} \\ &= \frac{1}{14}\sqrt{36 + 9 + 81} \\ &= \frac{1}{14}\sqrt{126} \\ &\approx .8018. \end{aligned}$$

In other words, for any $\bar{x} \in \mathbb{R}^2$, the distance between $\bar{b}$ and $A\bar{x}$ is *at least* $\frac{\sqrt{126}}{14}$.

Note We can see from the example above why "least-squares" is an appropriate name for such an approximation: we are minimizing the sum of squares that appears under the square root!