EECS 16B Designing Information Devices and Systems II
Spring 2021

Note 10A: Orthonormalization

1 Speeding up Least Squares

In 16A, you were introduced to least squares to approximate solutions to systems of equations. An example context you saw was radio transmissions from devices. We have a huge number n of devices that could potentially be transmitting its own unique signal. The m-long signal $\vec{y}$ we receive is the sum of the signals from each active device.

Specifically, let $A \in \mathbb{R}^{m \times n}$ be a matrix with $m > n$ (a tall matrix), where the kth column represents the signal that is sent by device k:

$$A = \begin{bmatrix} | & | & & | \\ \vec{S}_1 & \vec{S}_2 & \cdots & \vec{S}_n \\ | & | & & | \end{bmatrix}$$

Then, let $\vec{x} \in \mathbb{R}^n$ be a vector where the kth entry represents the strength of the kth device. Our received signal is

$$\vec{y} = A\vec{x} = \begin{bmatrix} | & | & & | \\ \vec{S}_1 & \vec{S}_2 & \cdots & \vec{S}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x[1] \\ x[2] \\ \vdots \\ x[n] \end{bmatrix} = \sum_{k=1}^{n} x[k]\vec{S}_k$$

This is the "noise-free" case. More realistically, $\vec{y}$ is only approximately given by the above, but there is another (presumably small) noise component as well. To find the $x[i]$ that best estimates which signal strengths were used, we need to use least squares to project $\vec{y}$ onto the columns of A.

$$\vec{x} = (A^\top A)^{-1} A^\top \vec{y}.$$

Now assume that often times, new devices and signals $\vec{S}_i$ are added to our database. To solve the least squares problem again, we will need to recalculate the entire equation above, including the inverse. However, matrix inversion is computationally expensive — for a $n \times n$ matrix, inversion takes O(n^3). We might have to do inversion at many time steps as more signals S_i are added to our database, so is there a way to reuse our prior computation? (Note that this may not be a very realistic issue, and that the methods we learn below are much more general and useful for other purposes.)

2 Orthogonal Vectors and Projection

Recall from 16A that two vectors $\vec{v}, \vec{w}$ are orthogonal if they are 90° apart. Remember that an equivalent definition is that they are orthogonal if and only if

$$\langle \vec{v}, \vec{w} \rangle = \vec{v}^\top \vec{w} = \vec{w}^\top \vec{v} = 0 \tag{1}$$

Recall that the orthogonal projection of a vector $\vec{y}$ on to any other nonzero vector $\vec{b}$ is

$$\vec{y}_{\vec{b}} = \frac{\vec{y}^\top \vec{b}}{\left\|\vec{b}\right\|^2}\vec{b} \tag{2}$$

Also recall that least squares is just an orthogonal projection of a vector onto an entire subspace of vectors, so

$$\vec{y}_A = A\hat{x} = A(A^\top A)^{-1}A^\top \vec{y} \tag{3}$$

In this section, we will show that if the columns of A are mutually orthogonal to each other, the projection of $\vec{y}$ onto span(A) is the sum of the projection of $\vec{y}$ onto each column of A individually. Let's take a look at the case where $j = 2$ and the songs are mutually orthogonal. Suppose the songs found so far are $\vec{S}_1$ and $\vec{S}_2$, i.e., $A_2 = \begin{bmatrix} | & | \\ \vec{S}_1 & \vec{S}_2 \\ | & | \end{bmatrix}$. Then, the projection of $\vec{y}$ onto A_2 is

$$\vec{y}_{A_2} = A_2\left(A_2^\top A_2\right)^{-1}A_2^\top \vec{y}. \tag{4}$$

Let's first compute the term $\left(A_2^\top A_2\right)^{-1}$:

$$A_2^\top A_2 = \begin{bmatrix} — & \vec{S}_1{}^\top & — \\ — & \vec{S}_2{}^\top & — \end{bmatrix}\begin{bmatrix} | & | \\ \vec{S}_1 & \vec{S}_2 \\ | & | \end{bmatrix} \tag{5}$$

$$= \begin{bmatrix} \vec{S}_1{}^\top \vec{S}_1 & \vec{S}_1{}^\top \vec{S}_2 \\ \vec{S}_2{}^\top \vec{S}_1 & \vec{S}_2{}^\top \vec{S}_2 \end{bmatrix} \tag{6}$$

$$= \begin{bmatrix} \|\vec{S}_1\|^2 & 0 \\ 0 & \|\vec{S}_2\|^2 \end{bmatrix}. \tag{7}$$

Thus, we have a diagonal matrix and so

$$\left(A_2^\top A_2\right)^{-1} = \begin{bmatrix} \frac{1}{\left\|\vec{S}_1\right\|^2} & 0 \\ 0 & \frac{1}{\left\|\vec{S}_2\right\|^2} \end{bmatrix}. \tag{8}$$

Then, substituting this matrix into the original expression, the projection of $\vec{y}$ onto span(A_2) is

$$\vec{y}_{A_2} = A_2 \left(A_2^\top A_2\right)^{-1} A_2^\top \vec{y} \tag{9}$$

$$= \begin{bmatrix} | & | \\ \vec{S}_1 & \vec{S}_2 \\ | & | \end{bmatrix} \begin{bmatrix} \frac{1}{\left\|\vec{S}_1\right\|^2} & 0 \\ 0 & \frac{1}{\left\|\vec{S}_2\right\|^2} \end{bmatrix} \begin{bmatrix} — & \vec{S}_1^{\,\top} & — \\ — & \vec{S}_2^{\,\top} & — \end{bmatrix} \vec{y} \tag{10}$$

$$= \begin{bmatrix} | & | \\ \vec{S}_1 & \vec{S}_2 \\ | & | \end{bmatrix} \begin{bmatrix} \frac{1}{\left\|\vec{S}_1\right\|^2} & 0 \\ 0 & \frac{1}{\left\|\vec{S}_2\right\|^2} \end{bmatrix} \begin{bmatrix} \vec{S}_1^{\,\top}\vec{y} \\ \vec{S}_2^{\,\top}\vec{y} \end{bmatrix} \tag{11}$$

$$= \begin{bmatrix} | & | \\ \vec{S}_1 & \vec{S}_2 \\ | & | \end{bmatrix} \begin{bmatrix} \frac{\vec{S}_1^{\,\top}\vec{y}}{\left\|\vec{S}_1\right\|^2} \\ \frac{\vec{S}_2^{\,\top}\vec{y}}{\left\|\vec{S}_2\right\|^2} \end{bmatrix} \tag{12}$$

$$= \left(\frac{\vec{S}_1^{\,\top}\vec{y}}{\left\|\vec{S}_1\right\|^2}\right)\vec{S}_1 + \left(\frac{\vec{S}_2^{\,\top}\vec{y}}{\left\|\vec{S}_2\right\|^2}\right)\vec{S}_2. \tag{13}$$

Observe that the first term in the sum above is the projection of $\vec{y}$ onto $\vec{S}_1$ and the second term is the projection of $\vec{y}$ onto $\vec{S}_2$. Generalizing this pattern, we can guess that the projection of $\vec{y}$ onto span(A_n) where A_n has mutually orthogonal columns is

$$\vec{y}_{A_n} = \left(\frac{\vec{S}_1^{\,\top}\vec{y}}{\left\|\vec{S}_1\right\|^2}\right)\vec{S}_1 + \left(\frac{\vec{S}_2^{\,\top}\vec{y}}{\left\|\vec{S}_2\right\|^2}\right)\vec{S}_2 + \cdots + \left(\frac{\vec{S}_n^{\,\top}\vec{y}}{\left\|\vec{S}_n\right\|^2}\right)\vec{S}_n. \tag{14}$$

Furthermore, observe that if $\vec{S}_1, \ldots, \vec{S}_n$ are unit vectors (i.e., they all have length 1), then the above would further reduce to

$$\vec{y}_{A_n} = \left(\vec{S}_1^{\,\top}\vec{y}\right)\vec{S}_1 + \left(\vec{S}_2^{\,\top}\vec{y}\right)\vec{S}_2 + \cdots + \left(\vec{S}_n^{\,\top}\vec{y}\right)\vec{S}_n \tag{15}$$

$$= \begin{bmatrix} | & & | \\ \vec{S}_1 & \cdots & \vec{S}_n \\ | & & | \end{bmatrix} \begin{bmatrix} — & \vec{S}_1^{\,\top} & — \\ & \vdots & \\ — & \vec{S}_n^{\,\top} & — \end{bmatrix} \vec{y} = A_n A_n^\top \vec{y} \tag{16}$$

Definition: A set of vectors $\{\vec{v}_1, \ldots, \vec{v}_n\}$ is **orthonormal** if all the vectors are mutually orthogonal to each other (i.e. $\vec{v}_i^\top \vec{v}_j = 0$ if $i \neq j$) and all are of unit length (i.e. $\|\vec{v}_i\| = 1 = \vec{v}_i^\top \vec{v}_i$). Thus above, A_n has orthonormal columns. We now will generalize what we did earlier by showing that for any matrix Q with n

orthonormal columns, $Q^\top Q = I_n$.

$$Q^\top Q = \begin{bmatrix} — & \vec{q}_1^\top & — \\ & \vdots & \\ — & \vec{q}_n^\top & — \end{bmatrix} \begin{bmatrix} | & & | \\ \vec{q}_1 & \dots & \vec{q}_n \\ | & & | \end{bmatrix} \tag{17}$$

$$= \begin{bmatrix} \vec{q}_1^\top \vec{q}_1 & \vec{q}_1^\top \vec{q}_2 & \dots & \vec{q}_1^\top \vec{q}_n \\ \vec{q}_2^\top \vec{q}_1 & \ddots & & \vec{q}_2^\top \vec{q}_n \\ \vdots & & \ddots & \vdots \\ \vec{q}_n^\top q_1 & \vec{q}_n^\top \vec{q}_2 & \dots & \vec{q}_n^\top \vec{q}_n \end{bmatrix} \tag{18}$$

$$= \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = I_n \tag{19}$$

Remember that we are using the property of orthonormal vectors that $\vec{q}_i^\top \vec{q}_j = 0$ when $i \neq j$ and $\vec{q}_i^\top \vec{q}_i = ||\vec{q}_i||^2 = 1$. Using this, notice that the least-squares estimate with orthonormal vectors simplifies to $\vec{y}_{A_n} = A_n(A_n^\top A_n)^{-1} A_n^\top \vec{y} = A_n(I)^{-1} A_n^\top \vec{y} = A_n A_n^\top \vec{y}$. By direct algebraic manipulation, we formally validated our generalization in equation 16.

Note that the above proof is general and applies to non-square matrices. However, we will specially refer to square matrices whose columns are orthonormal as orthonormal [1] matrices. If a square matrix Q is orthonormal, we can additionally say that $Q^\top Q = QQ^\top = I \implies Q^\top = Q^{-1}$.

Thus, we can see that having orthonormal vectors $\vec{S}_i$ will make least squares must faster and only consist of 1 matrix multiplication. But now you may ask how can we even ensure that $\vec{S}_i$ are orthonormal?

3 Orthonormalization

We want to take a sequence of vectors $\vec{S}_1, \vec{S}_2, \dots \vec{S}_k$ and construct a new sequence of vectors $\vec{q}_1, \vec{q}_2, \dots, \vec{q}_k$ that are orthonormal (i.e. $\vec{q}_i^\top \vec{q}_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$) and satisfy the property that the subspaces span$(\vec{S}_1, \vec{S}_2, \dots \vec{S}_\ell)$ spanned by the original set of vectors are always the same as the subspaces span$(\vec{q}_1, \vec{q}_2, \dots, \vec{q}_\ell)$ spanned by the new set of vectors.

This might seem hard but we will start at the beginning and proceed systematically. We first let $\vec{q}_1 = \frac{\vec{S}_1}{\|\vec{S}_1\|}$ to make it normal, and it will clearly have the same span as $\vec{S}_1$. We will then leverage what we know about projections and least-squares from 16A. We know that the residual vector after a projection is always orthogonal [2] to the subspace being projected upon. So to ensure that the new vector $\vec{q}_k$ is orthogonal, we can just remove all parts of $\vec{S}_k$ that lie in the span of our previous vectors. From the previous section and equation 16, since the $\vec{q}_i$ are orthonormal, this is equivalent to subtracting each individual projection onto

[1] In a bit of confusing notation, in math literature you will often see such matrices called orthogonal even when they want to explicitly require that each column is normalized to have unit norm. We will try to use "orthonormal" to avoid this confusion.

[2] Recall that this is how we actually derived the least-squares formula!

all of our previous $\vec{q}_i$ vectors. Consequently, we can recursively define

$$\vec{q}_k = \frac{\vec{S}_k - \sum_{\ell=1}^{k-1} \vec{q}_\ell (\vec{q}_\ell^\top \vec{S}_k)}{\|\vec{S}_k - \sum_{\ell=1}^{k-1} \vec{q}_\ell (\vec{q}_\ell^\top \vec{S}_k)\|}. \tag{20}$$

This has unit norm by construction and it is orthogonal to all the previous $\vec{q}_\ell$ because it removes all the projections of the new vector $\vec{S}_k$ onto the subspace spanned by them. This collection also preserves the same span because every new vector $\vec{q}_k$ is just a linear combination of the original vectors. It turns out that this very natural iterative process that we "discovered for ourselves" has a name: **Gram-Schmidt Orthonormalization**.

Said more slowly and using generic language, Gram Schmidt is a procedure that takes a list[3] of linearly independent vectors $\{\vec{S}_1, \ldots, \vec{S}_n\}$ and generates an orthonormal list of vectors $\{\vec{q}_1, \ldots, \vec{q}_n\}$ that span the same subspaces as the original list. Concretely, $\{\vec{q}_1, \ldots, \vec{q}_n\}$ satisfy the following:

$$\{\vec{q}_1, \ldots, \vec{q}_n\} \text{ is an orthonormal set of vectors} \tag{21}$$

$$\text{span}(\{\vec{S}_1, \ldots, \vec{S}_k\}) = \text{span}(\{\vec{q}_1, \ldots, \vec{q}_k\}) \quad \forall 1 \leq k \leq n \tag{22}$$

Proof of Orthonormality (21):

We first start with showing each vector is normal, or unit length. This is true by construction since we are dividing a vector by the norm of that vector, so the result must have norm 1. In other words, for all vectors $\vec{v}$,

$$\left|\left| \frac{\vec{v}}{\|\vec{v}\|} \right|\right| = \frac{1}{\|\vec{v}\|} \|\vec{v}\| = 1$$

Now onto orthogonality. We will use an inductive approach by first assuming that the first $k-1$ vectors are already orthonormal. We will then show that our new constructed vector $\vec{q}_k$ will be orthogonal to all previous ones, so $\vec{q}_p^T \vec{q}_k = 0$ for all $p = 1, \ldots, k-1$. Note that constant factors don't affect orthogonality, so for simplicity we will combine the norm division for both $\vec{q}_k$ and $\vec{q}_p$ below into one constant A.

$$\vec{q}_p^T \vec{q}_k = A \vec{q}_p^T (\vec{S}_k - \sum_{\ell=1}^{k-1} \vec{q}_\ell (\vec{q}_\ell^\top \vec{S}_k)) \tag{23}$$

$$= A(\vec{q}_p^T \vec{S}_k - \sum_{\ell=1}^{k-1} \vec{q}_p^T \vec{q}_\ell (\vec{q}_\ell^T \vec{S}_k)) \tag{24}$$

Since we assumed the first $k-1$ vectors are all orthonormal, the $\vec{q}_p^T \vec{q}_\ell$ will cause the only nonzero in the summation to occur when $\ell = p$. Then,

$$\vec{q}_p^T \vec{q}_k = A(\vec{q}_p^T \vec{S}_k - \vec{q}_p^T \vec{q}_p (\vec{q}_p^T \vec{S}_k)) \tag{25}$$

$$= A(\vec{q}_p^T \vec{S}_k - \vec{q}_p^T \vec{S}_k) = 0 \tag{26}$$

where we use the fact that $\vec{q}_p^T \vec{q}_p = 1$ since we assumed it is orthonormal. Thus, for any k we know that the next vector we construct will be orthogonal to all of our previous vectors. If this idea of induction is confusing, don't worry as CS 70 will cover it in much more detail.

[3] The fact that these are lists and not sets matters. The vectors are ordered. We don't just want the overall spans to be the same, we want the spans to be the same as we walk down the lists together.

Proof of Equivalent Span (22):

Again, we will use an inductive approach by assuming that the first $k-1$ vectors of S and Q both span the same space. This is true for our base case since $\vec{q}_1 = \vec{S}_1/\|\vec{S}_1\|$. Now all we need to show is that $\vec{S}_k$ can be written as a linear combination of $\{\vec{q}_1, \ldots, \vec{q}_k\}$. By construction of $\vec{q}_k$, that is exactly true with

$$\vec{S}_k = \left\| \vec{S}_k - \sum_{\ell=1}^{k-1} (\vec{q}_\ell^T \vec{S}_k)\vec{q}_\ell \right\| \vec{q}_k + \sum_{\ell=1}^{k-1} (\vec{q}_\ell^T \vec{S}_k)\vec{q}_\ell$$

Thus, for all k, the spans will be equivalent.

3.1 Example for three vectors

The above might have been a bit fast, so let's walk through the reasoning for the case of three vectors to make sure it is clear.

Consider three vectors $\{\vec{S}_1, \vec{S}_2, \vec{S}_3\}$ that are linearly independent of each other.

- **Step 1:** Find unit vector $\vec{q}_1$ such that span($\{\vec{q}_1\}$) = span($\{\vec{S}_1\}$).
  Since span($\{\vec{S}_1\}$) is a one dimensional vector space, we can simply scale $\{\vec{S}_1\}$ so that it is unit norm:

$$\vec{q}_1 = \frac{\vec{S}_1}{\left\|\vec{S}_1\right\|}. \tag{27}$$

- **Step 2:** Given $\vec{q}_1$ from the previous step, find $\vec{q}_2$ such that span($\{\vec{q}_1, \vec{q}_2\}$) = span($\{\vec{S}_1, \vec{S}_2\}$) and orthogonal to $\vec{q}_1$. We know that $\vec{S}_2-$ (the projection of $\vec{S}_2$ on $\vec{q}_1$) would be orthogonal to $\vec{q}_1$ from 16A. So first, we can find the error or residual

$$\vec{e}_2 = \vec{S}_2 - \left(\vec{q}_1^\top \vec{S}_2\right) \vec{q}_1, \tag{28}$$

  which is orthogonal to $\vec{q}_1$. Then, we can normalize to get $\vec{q}_2 = \frac{\vec{e}_2}{\|\vec{e}_2\|}$. Note that these operations preserve the span because $\vec{q}_1$ and $\vec{q}_2$ are just linear combinations of $\vec{S}_1$ and $\vec{S}_2$ and vice-versa.

- **Step 3:** Now given $\vec{q}_1$ and $\vec{q}_2$ in the previous steps, we would like to find $\vec{q}_3$ such that span($\{\vec{q}_1, \vec{q}_2, \vec{q}_3\}$) = span($\{\vec{S}_1, \vec{S}_2, \vec{S}_3\}$). We know that the projection of $\vec{S}_3$ onto the subspace spanned by $\vec{q}_1, \vec{q}_2$ is

$$\left(\vec{q}_2^\top \vec{S}_3\right) \vec{q}_2 + \left(\vec{q}_1^\top \vec{S}_3\right) \vec{q}_1. \tag{29}$$

  Consequently, we know that the error/residual

$$\vec{e}_3 = \vec{S}_3 - \left[\left(\vec{q}_2^\top \vec{S}_3\right) \vec{q}_2 + \left(\vec{q}_1^\top \vec{S}_3\right) \vec{q}_1\right]. \tag{30}$$

  is orthogonal to both $\vec{q}_1$ and $\vec{q}_2$. Normalizing, we have $\vec{q}_3 = \frac{\vec{e}_3}{\|\vec{e}_3\|}$.

The procedure given earlier is just a continuation of this pattern.

Inputs

- A list of linearly independent vectors $\{\vec{S}_1, \ldots, \vec{S}_n\}$.

Outputs

- An orthonormal list of vectors $\{\vec{q}_1, \ldots, \vec{q}_n\}$, where span($\{\vec{S}_1, \ldots, \vec{S}_k\}$) = span($\{\vec{q}_1, \ldots, \vec{q}_k\}$) for all $1 \leq k \leq n$.

Gram Schmidt Procedure

- compute $\vec{q}_1$: $\vec{q}_1 = \frac{\vec{S}_1}{\left\|\vec{S}_1\right\|}$
- for $(i = 2 \ldots n)$:
  (a) Compute the vector $\vec{e}_i$, such that span($\{\vec{q}_1, \ldots, \vec{q}_{i-1}, \vec{e}_i\}$) = span($\{\vec{S}_1, \ldots, \vec{S}_i\}$):

$$\vec{e}_i = \vec{S}_i - \sum_{j=1}^{i-1} \left(\vec{q}_j^{\top} \vec{S}_i\right) \vec{q}_j \tag{31}$$

  (b) Normalize to compute $\vec{q}_i$ itself: $\vec{q}_i = \frac{\vec{e}_i}{\|\vec{e}_i\|}$.

Contributors:

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