1 Last time: orthonormal vectors, projections, orthogonal bases

Vectors $u_1, u_2, \ldots, u_p$ are *orthonormal* if each u_i is a unit vector and any two vectors u_i and u_j (with $i \neq j$) are orthogonal. In other words, if $u_i \bullet u_j = 0$ when $i \neq j$ and $u_i \bullet u_i = 1$ for all $i = 1, 2, \ldots, p$.

The standard basis $e_1, e_2, \ldots, e_n$ of $\mathbb{R}^n$ consists of orthonormal vectors.

If $v_1, v_2, \ldots, v_p$ are orthogonal and all nonzero then $\frac{1}{\|v_1\|}v_1, \frac{1}{\|v_2\|}v_2, \ldots, \frac{1}{\|v_p\|}v_p$ are orthonormal.

Theorem. Let U be an $m \times n$ matrix. The columns of U are orthonormal vectors if and only if $U^T U = I_n$. If this happens then $(Ux) \bullet (Uy) = x \bullet y$ for all $x, y \in \mathbb{R}^n$.

Let $V \subset \mathbb{R}^n$ be any subspace.

Recall that $V^\perp = \{w \in \mathbb{R}^n : w \bullet v = 0 \text{ for all } v \in V\}$.

We showed last time that $V \cap V^\perp = \{0\}$ and $\dim V + \dim V^\perp = n$.

Theorem (Orthogonal projections)**.** If $W \subset \mathbb{R}^n$ is a subspace and $y \in \mathbb{R}^n$ then there is a unique vector $\text{proj}_W(y) \in W$ such that $y - \text{proj}_W(y) \in W^\perp$.

We call $\text{proj}_W(y)$ the *orthogonal projection* of y onto W.

If $u_1, u_2, \ldots, u_p$ is any orthogonal basis of W then

$$\text{proj}_W(y) = \frac{y \bullet u_1}{u_1 \bullet u_1}u_1 + \frac{y \bullet u_2}{u_2 \bullet u_2}u_2 + \cdots + \frac{y \bullet u_p}{u_p \bullet u_p}u_p.$$

This formula does not depend on the choice of orthogonal basis for W: choose another basis, apply the same formula, and you'll end up with the same vector $\text{proj}_W(y) \in W$.

Properties of orthogonal projections

We have $\text{proj}_W(y) = y$ if and only if $y \in W$.

We have $\text{proj}_W(y) = 0$ if and only if $y \in W^\perp$.

It holds that $\|y - \text{proj}_W(y)\| < \|y - v\|$ for all $v \in W$ with $v \neq \text{proj}_W(y)$.

Theorem. Every nonzero subspace of $\mathbb{R}^n$ has at least one orthogonal basis.

(In fact, any nonzero subspace has infinitely many orthogonal bases.)

The *Gram-Schmidt process* is an important algorithm which takes a basis for a subspace $W \subset \mathbb{R}^n$ as input and produces an orthogonal basis for W as output.

Gram-Schmidt process.

Let $W \subset \mathbb{R}^n$ be a subspace.

Suppose $x_1, x_2, \ldots, x_p$ is a basis for W.

Define $v_1, v_2, \ldots, v_p \in W$ inductively by the following formulas:

$$v_1 = x_1.$$

$$v_2 = x_2 - \frac{x_2 \bullet v_1}{v_1 \bullet v_1}v_1.$$

$$v_3 = x_3 - \frac{x_3 \bullet v_1}{v_1 \bullet v_1} v_1 - \frac{x_3 \bullet v_2}{v_2 \bullet v_2} v_2.$$

$$v_4 = x_4 - \frac{x_4 \bullet v_1}{v_1 \bullet v_1} v_1 - \frac{x_4 \bullet v_2}{v_2 \bullet v_2} v_2 - \frac{x_4 \bullet v_3}{v_3 \bullet v_3} v_3.$$

$$\vdots$$

$$v_p = x_p - \frac{x_p \bullet v_1}{v_1 \bullet v_1} - \frac{x_p \bullet v_2}{v_2 \bullet v_2} - \cdots - \frac{x_p \bullet v_{p-1}}{v_{p-1} \bullet v_{p-1}} v_{p-1}.$$

For each $i = 1, 2, \ldots, p$, the vectors $v_1, v_2, \ldots, v_i$ are an orthogonal basis for the subspace

$$\mathbb{R}\text{-span}\{x_1, x_2, \ldots, x_i\} = \mathbb{R}\text{-span}\{v_1, v_2, \ldots, v_i\} \subset W$$

and consequently v_{i+1} is just x_{i+1} minus the orthogonal projection of v_{i+1} onto this subspace.

The full list of vectors $v_1, v_2, \ldots, v_p$ is an orthogonal basis for W.

Example. Let $W = \text{Nul}\left(\begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}\right) = \{w \in \mathbb{R}^4 : w_1 + w_2 + w_3 + w_4 = 0\}$.

A basis for W is given by $x_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}$, $x_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix}$, $x_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}$.

To find an orthogonal basis, we let

$$v_1 = x_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}.$$

$$v_2 = x_2 - \frac{x_2 \bullet v_1}{v_1 \bullet v_1} v_1 = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix} - \frac{0 - 1 + 0 + 0}{1 + 1 + 0 + 0} \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \\ -1 \\ 0 \end{bmatrix}.$$

$$v_3 = x_3 - \underbrace{\frac{x_3 \bullet v_1}{v_1 \bullet v_1}}_{=0} v_1 - \frac{x_3 \bullet v_2}{v_2 \bullet v_2} v_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} + \frac{2}{3} \begin{bmatrix} 1/2 \\ 1/2 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \\ -1 \end{bmatrix}.$$

Thus $\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1/2 \\ 1/2 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \\ -1 \end{bmatrix}$ are an orthogonal basis for W.

The rescaled vectors $\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -2 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ -3 \end{bmatrix}$ are also an orthogonal basis for W.

In general, the Gram-Schmidt process applied to the basis

$$x_1 = e_1 - e_2, \quad x_2 = e_2 - e_3, \quad x_3 = e_3 - e_4, \quad \dots, \quad x_{n-1} = e_{n-1} - e_n$$

of $\operatorname{Nul}\left(\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \end{bmatrix}\right)$ will produce the orthogonal basis

$v_1 = e_1 - e_2.$

$v_2 = \frac{1}{2}e_1 + \frac{1}{2}e_2 - e_3.$

$v_3 = \frac{1}{3}e_1 + \frac{1}{3}e_2 + \frac{1}{3}e_3 - e_4.$

$\vdots$

$v_{n-1} = \frac{1}{n-1}e_1 + \frac{1}{n-1}e_2 + \cdots + \frac{1}{n-1}e_{n-1} - e_n.$

A nicer orthogonal basis is provided by rescaling:

$$e_1 - e_2, \quad e_1 + e_2 - 2e_3, \quad e_1 + e_2 + e_3 - 3e_4, \quad \dots \quad e_1 + e_2 + \cdots + e_{n-1} - (n-1)e_n.$$

We discussed one other relevant result last time:

Theorem (QR-factorisation)**.** Let A be an $m \times n$ matrix with linearly independent columns. Then $A = QR$ where Q is an $m \times n$ matrix whose columns are an orthonormal basis for $\operatorname{Col} A$ and R is an $n \times n$ upper-triangular matrix with positive entries on the diagonal.

One calls the decomposition $A = QR$ a *QR-factorisation* of A.

2 Least-squares problems

Many linear systems $Ax = b$ that arise in the real world are *overdetermined* (meaning they have more equations than variables, or equivalently that A has more rows than columns) and often inconsistent (meaning they have no exact solution $x \in \mathbb{R}^n$).

For example, $b \in \mathbb{R}^m$ might be a vector of measurements and each row of Ax might provide a linear approximation to what we expect these measurements to be in terms of certain inputs x.

Because measurements are noisy and because our linear approximations are inexact, there may be no input vector $x \in \mathbb{R}^n$ such that $Ax = b$. When no exact solution is available, the next best thing to provide is an input vector $x \in \mathbb{R}^n$ such that Ax is as "close" to the vector $b \in \mathbb{R}^m$ as possible.

In general, there are many reasonable ways to quantify how close two vectors are to each other. One of the most common is the distance function we have already seen: define the *distance* between vectors $u, v \in \mathbb{R}^n$ to be $\|u - v\| = \sqrt{(u-v) \bullet (u-v)}$. Two vectors are close if their distance in this sense is small. The distance function $\|\cdot\|$ is called the *Euclidean distance* or L^2*-distance*. In two and three dimensions, this distance corresponds to the usual way that we measure distance between points in space.

Definition. If A is an $m \times n$ matrix and $b \in \mathbb{R}^m$, then a *least-squares solution* to the linear system $Ax = b$ is a vector $\widehat{x} \in \mathbb{R}^n$ such that $\|b - A\widehat{x}\| \leq \|b - Ax\|$ for all $x \in \mathbb{R}^n$.

In other words, a least-squares solution to $Ax = b$ is a vector $\widehat{x} \in \mathbb{R}^n$ that minimises $\|b - A\widehat{x}\|$.

A vector that minimises $\|b - A\widehat{x}\|$ will also minimise $\|b - A\widehat{x}\|^2$, which is the sum of the squares of the entries in the vector $b - A\widehat{x}$. This accounts for the name "least-squares."

Least-squares problems (that is, problems requiring us to find a least-squares solution to some linear system) arise all over the place in engineering and statistics. Being able to solve such problems is maybe one of the most important applications of the material covered in this course. Our goal today is to describe the general solution to the least-squares problem. Here are the keys points:

- If $Ax = b$ is a consistent linear system then every least-squares solution will be an exact solution.
- There may be more than one least-squares solution to a given linear system $Ax = b$.
- However, in contrast to exact solutions, there is **always** at least one least-squares solution.

The last fact is not immediately obvious from the definition of a least-squares solution.

Solving least-squares problems in general.

Fix an $m \times n$ matrix A and a vector $b \in \mathbb{R}^m$

A least-squares solution $\widehat{x} \in \mathbb{R}^n$ to $Ax = b$ is a vector such that $\|A\widehat{x} - b\|$ is as small as possible.

If $\widehat{x} \in \mathbb{R}^n$ then we necessarily have $A\widehat{x} \in \operatorname{Col} A$.

Suppose $\widehat{b} \in \operatorname{Col} A$ minimises the distance $\|\widehat{b} - b\|$. From results last time, $\widehat{b}$ must then be the projection

$$\widehat{b} = \operatorname{proj}_{\operatorname{Col} A}(b).$$

We conclude that:

Lemma. The least-squares solutions to $Ax = b$ are precisely those $\widehat{x} \in \mathbb{R}^n$ such that $A\widehat{x} = \widehat{b}$.

Using this lemma, we can prove something even more explicit:

Theorem. The set of least-squares solutions to $Ax = b$ is the set of exact solutions to the linear system $A^TAx = A^Tb$. This new linear system is always consistent so its set of solutions is nonempty.

Proof. Since $b - \widehat{b} \in (\operatorname{Col} A)^\perp = \operatorname{Nul} A^T$, we have $A^T(b - \widehat{b}) = 0$ and $A^T\widehat{b} = A^Tb$.

Thus, if $\widehat{x} \in \mathbb{R}^n$ satisfies $A\widehat{x} = \widehat{b}$ then $A^TAx = A^T\widehat{b} = A^Tb$.

Conversely, if $\widehat{x} \in \mathbb{R}^n$ satisfies $A^TA\widehat{x} = A^Tb$, then $A^T(A\widehat{x} - b) = 0$ so $A\widehat{x} - b \in \operatorname{Nul} A^T = (\operatorname{Col} A)^\perp$. In this case, it follows by the uniqueness of orthogonal projections that $A\widehat{x} = \operatorname{proj}_{\operatorname{Col} A}(b) = \widehat{b}$.

This shows that the set of exact solutions to $Ax = \widehat{b}$, which is precisely the set of least-squares solutions to $Ax = b$, is the same as the set of exact solutions to $A^TAx = A^Tb$.

The last thing we need to show is that the linear system $A^TAx = A^Tb$ is always consistent. This holds since $\widehat{b} \in \operatorname{Col} A$ so, by the definition of the column space, there must exist some $\widehat{x} \in \mathbb{R}^n$ such that $A\widehat{x} = \widehat{b}$, and it then holds that $A^TA\widehat{x} = A^T\widehat{b} = A^Tb$. □

Example. Here is a simple, somewhat contrived example.

Let $A = \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix}$ and $b = \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix}$.

To find a least-squares solution to $Ax = b$, we compute

$$A^TA = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix} \qquad \text{and} \qquad A^Tb = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix} = \begin{bmatrix} 19 \\ 11 \end{bmatrix}.$$

The least-squared solutions we want are the exact solutions to $A^TAx = A^Tb$. Solve this by row reducing:

$$\begin{bmatrix} 17 & 1 & 19 \\ 1 & 5 & 11 \end{bmatrix} \sim \begin{bmatrix} 1 & 5 & 11 \\ 17 & 1 & 19 \end{bmatrix} \sim \begin{bmatrix} 1 & 5 & 11 \\ 0 & -84 & -168 \end{bmatrix} \sim \begin{bmatrix} 1 & 5 & 11 \\ 0 & 1 & 2 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix}.$$

In this case $A^TAx = A^Tb$ has a unique solution

$$\widehat{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

which is also the unique least-squares solution to $Ax = b$. Note that $A\widehat{x} \neq b$ as

$$\|A\widehat{x} - b\| = \left\| \begin{bmatrix} 4 \\ 4 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix} \right\| = \left\| \begin{bmatrix} 2 \\ 4 \\ -8 \end{bmatrix} \right\| = \sqrt{4 + 16 + 64} = \sqrt{84}.$$

Geometrically, we interpret the least-squares solution as meaning that

$$A\widehat{x} = \begin{bmatrix} 2 \\ 4 \\ -8 \end{bmatrix}$$

is the point in the plane in $\mathbb{R}^3$ spanned by the columns of A which is closest to b.

A linear system $Ax = b$ has a unique solution for every $b \in \mathbb{R}^m$ if and only if the matrix A is invertible. The following theorem describes, analogously, when $Ax = b$ has a unique least-squares solution.

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

(a) $Ax = b$ has a unique least-squares solution for each $b \in \mathbb{R}^m$.

(b) The columns of A are linearly independent.

(c) A^TA is invertible.

When these properties hold, the unique least-squares solution to $Ax = b$ is the vector

$$\widehat{x} = (A^TA)^{-1}A^Tb$$

(which is the unique exact solution to $A^TAx = A^Tb$).

Remark. In practice, the product $(A^TA)^{-1}A^Tb$ is never computed directly for a large linear system — it is more efficient to find $\widehat{x}$ by solving the system $A^TAx = A^Tb$ via row reduction.

Proof. If $\widehat{x} \in \mathbb{R}^n$ is a least-squares solution to $Ax = b$, then $\widehat{x} + v$ is also a least-squares solution for any $v \in \operatorname{Nul} A$ since $\|A\widehat{x} - b\| = \|A(\widehat{x} + v) - b\|$. Therefore if (a) holds then we must have $\operatorname{Nul} A = \{0\}$ so (b) must also hold.

If $v \in \mathbb{R}^n$ then $A^TAv = 0$ if and only if $Av \in \operatorname{Col} A \cap \operatorname{Nul} A^T = \operatorname{Col} A \cap (\operatorname{Col} A)^\perp = \{0\}$. Therefore $\operatorname{Nul}(A^TA) = \operatorname{Nul}(A)$. Hence if (b) holds then $\operatorname{Nul}(A^TA) = \operatorname{Nul} A = \{0\}$, which means that A^TA is invertible since this matrix if square.

Finally, if (c) holds then the linear system $A^TAx = A^Tb$ has a unique solution so (a) holds by the previous theorem. The chain of implications (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (a) shows that the properties are equivalent. $\square$

It is often a lot easier to compute a least-squares solution to $Ax = b$ if the columns of A are orthogonal. The following example illustrates this.

Example. Suppose $A = \begin{bmatrix} 1 & -6 \\ 1 & -2 \\ 1 & 1 \\ 1 & 7 \end{bmatrix}$ and $b = \begin{bmatrix} -1 \\ 2 \\ 1 \\ 6 \end{bmatrix}$. The columns of A are orthogonal.

The orthogonal projection of b onto $\operatorname{Col} A$ is therefore

$$\begin{aligned}
\widehat{b} &= \operatorname{proj}_{\operatorname{Col} A}(b) \\
&= \frac{\begin{bmatrix} -1 \\ 2 \\ 1 \\ 6 \end{bmatrix} \bullet \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}}{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \bullet \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{\begin{bmatrix} -1 \\ 2 \\ 1 \\ 6 \end{bmatrix} \bullet \begin{bmatrix} -6 \\ -2 \\ 1 \\ 7 \end{bmatrix}}{\begin{bmatrix} -6 \\ -2 \\ 1 \\ 7 \end{bmatrix} \bullet \begin{bmatrix} -6 \\ -2 \\ 1 \\ 7 \end{bmatrix}} \begin{bmatrix} -6 \\ -2 \\ 1 \\ 7 \end{bmatrix} \\
&= \frac{8}{4} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{45}{90} \begin{bmatrix} -6 \\ -2 \\ 1 \\ 7 \end{bmatrix} \\
&= 2 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} -6 \\ -2 \\ 1 \\ 7 \end{bmatrix} = 2a_1 + \frac{1}{2} a_2
\end{aligned}$$

where $A = \begin{bmatrix} a_1 & a_2 \end{bmatrix}$. Without any further work, we deduce that $\widehat{x} = \begin{bmatrix} 2 \\ 1/2 \end{bmatrix}$ is a least-squares solution to $Ax = b$ since $A\widehat{x} = 2a_1 + \frac{1}{2}a_2 = \widehat{b}$.

As our final result today, we mention one application of QR-factorisations. Recall the definition from earlier in this lecture (or the end of last lecture).

Theorem. Suppose A is an $m \times n$ matrix with linearly independent columns and $A = QR$ is a QR-factorisation. Then for each $b \in \mathbb{R}^n$, the system $Ax = b$ has a unique least-squares solution given by

$$\widehat{x} = R^{-1}Q^T b.$$

Proof. The columns of Q are an orthonormal basis for $\operatorname{Col} A$. Therefore if $\widehat{x} = R^{-1}Q^T b$ then

$$A\widehat{x} = QR\widehat{x} = QRR^{-1}Q^T b = QQ^T b = \operatorname{proj}_{\operatorname{Col} A}(b)$$

where the last equality holds by a result in the previous lecture. □