TLDR

Quick summary of today's notes. Lecture starts on next page.

- The *inner product* of two vectors $u, v \in \mathbb{R}^n$ is $u \bullet v = u_1v_1 + u_2v_2 + \cdots + u_nv_n \in \mathbb{R}$.

  A set of nonzero vectors $v_1, v_2, \ldots, v_p \in \mathbb{R}^n$ is *orthogonal* if $v_i \bullet v_j = 0$ for all $i \neq j$.

  Any such set is automatically linearly independent and therefore a basis for a subspace.

  An orthogonal basis is *orthonormal* if it consists entirely of unit vectors.

  Any subspace $V \subset \mathbb{R}^n$ has at least one orthogonal basis $u_1, u_2, \ldots, u_p$.

- The *Gram-Schmidt process* is an algorithm that takes a basis $x_1, x_2, \ldots, x_p$ for a subspace of $\mathbb{R}^n$ as input, and produces an orthogonal basis $v_1, v_2, \ldots, v_p$ of the same subspace as output.

  The orthogonal basis $v_1, v_2, \ldots, v_p$ is defined from the input basis $x_1, x_2, \ldots, x_p$ by these formulas:

  $$v_1 = x_1.$$
  $$v_2 = x_2 - \tfrac{x_2 \bullet v_1}{v_1 \bullet v_1} v_1.$$
  $$v_3 = x_3 - \tfrac{x_3 \bullet v_1}{v_1 \bullet v_1} v_1 - \tfrac{x_3 \bullet v_2}{v_2 \bullet v_2} v_2.$$
  $$\vdots$$
  $$v_p = x_p - \tfrac{x_p \bullet v_1}{v_1 \bullet v_1} v_1 - \tfrac{x_p \bullet v_2}{v_2 \bullet v_2} v_2 - \cdots - \tfrac{x_p \bullet v_{p-1}}{v_{p-1} \bullet v_{p-1}} v_{p-1}.$$

- Suppose A is an $m \times n$ matrix and $b \in \mathbb{R}^n$.

  It may happen that the linear system $Ax = b$ has no solutions $x \in \mathbb{R}^n$.

  When this occurs, one wants to find an approximate solution.

  Define the *length* of $v \in \mathbb{R}^n$ by $\|v\| = \sqrt{v \bullet v}$.

  A good approximate solution to $Ax = b$ is a vector $s \in \mathbb{R}^n$ that makes $\|As - b\|$ as small as possible.

  We call such a vector $s \in \mathbb{R}^n$ a *least-squares solution* to $Ax = b$.

- Essential fact: the least-squares solutions to $Ax = b$ are just the ordinary solutions to $A^TAx = A^Tb$.

  So to find least-squares solutions, just form the matrix $\left[\begin{array}{cc} A^TA & A^Tb \end{array}\right]$ and row reduce as usual.

- If $Ax = b$ is a consistent linear system then it has the same solutions as $A^TAx = A^Tb$.

  So if $Ax = b$ is consistent then least-squares solutions are the same as ordinary/exact solutions.

- However, the system $A^TAx = A^Tb$ is always consistent even if $Ax = b$ is not.

  So there is **always** at least one least-squares solution, even if $Ax = b$ has no exact solutions.

1 Last time: orthonormal vectors, projections, orthogonal bases

Vectors $u_1, u_2, \ldots, u_p$ are *orthonormal* if each vector is a unit vector and any two vectors 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 $x, y \in \mathbb{R}^n$.
(If $m = n$ then $U^T U = I_n$ means the same thing as $U^{-1} = U^T$.)

Example. The columns of $U = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}$ are orthonormal as $U^T U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
But in this case U is not invertible since U is not a square matrix.

Let $W \subset \mathbb{R}^n$ be any subspace. Recall that $W^\perp = \{w \in \mathbb{R}^n : w \bullet v = 0 \text{ for all } v \in W\}$.
We showed last time that $W \cap W^\perp = \{0\}$ and $\dim W + \dim W^\perp = n$.

Theorem. Let $y \in \mathbb{R}^n$. 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.
To compute $\text{proj}_W(y)$, use this fact: 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 works for any choice of orthogonal basis for W, but it requires you to find such a basis.

Properties of orthogonal projections

We have $\text{proj}_W(y) = y$ if and only if $y \in W$, and $\text{proj}_W(y) = 0$ if and only if $y \in W^\perp$.
Recall $\|v\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$. Then $\|y - \text{proj}_W(y)\| < \|y - v\|$ for all $v \in W$ with $v \neq \text{proj}_W(y)$.

Every subspace of $\mathbb{R}^n$ has an orthogonal basis, which we can find using the *Gram-Schmidt process*.

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:

$$v_1 = x_1.$$

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

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

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

$$\vdots$$

$$v_p = x_p - \frac{x_p \bullet v_1}{v_1 \bullet v_1} - \frac{x_p \bullet v_2}{v_2 \bullet v_2} - \frac{x_p \bullet v_3}{v_3 \bullet v_3} - \cdots - \frac{x_p \bullet v_{p-1}}{v_{p-1} \bullet v_{p-1}} v_{p-1}.$$

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

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

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

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

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

$v_1 = e_1 - e_2$.

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

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

$\vdots$

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

2 Least-squares problems

Many linear systems $Ax = b$ that arise in applications are *overdetermined* (meaning they have more equations than variables, or equivalently that the matrix 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 an approximation for what we expect these measurements to be in terms of certain inputs $x \in \mathbb{R}^n$.

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.

There are many 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

$$\|u - v\| = \sqrt{(u - v) \bullet (u - v)} = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_n - v_n)^2}.$$

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 $s \in \mathbb{R}^n$ such that $\|b - As\| \leq \|b - Ax\|$ for all $x \in \mathbb{R}^n$.

In other words, a least-squares solution to $Ax = b$ is a vector $s \in \mathbb{R}^n$ that minimizes $\|b - As\|$.

A vector that minimizes $\|b - As\|$ will also minimize $\|b - As\|^2$, which is the sum of the squares of the entries in the vector $b - As$. 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 an important application 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 $s \in \mathbb{R}^n$ to $Ax = b$ is a vector such that $\|As - b\|$ is as small as possible.

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

As mentioned earlier, if $v \in \operatorname{Col} A$ minimizes $\|v - b\|$ then we must have $v = \operatorname{proj}_{\operatorname{Col} A}(b)$. Therefore:

Lemma. The least-squares solutions to $Ax = b$ are precisely those $s \in \mathbb{R}^n$ such that $As = \operatorname{proj}_{\operatorname{Col} A}(b)$.

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. Let $\widehat{b} = \operatorname{proj}_{\operatorname{Col} A}(b)$. 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 $s \in \mathbb{R}^n$ satisfies $As = \widehat{b}$ then $A^TAs = A^T\widehat{b} = A^Tb$.

Conversely, if $s \in \mathbb{R}^n$ satisfies $A^TAs = A^Tb$, then $A^T(As - b) = 0$ so $As - b \in \operatorname{Nul} A^T = (\operatorname{Col} A)^\perp$.

In this case, it follows by the uniqueness of orthogonal projections that $As = \operatorname{proj}_{\operatorname{Col} A}(b) = \widehat{b}$.

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 $A^TAx = A^Tb$.

We claimed that the linear system $A^TAx = A^Tb$ is always consistent. This holds since $\widehat{b} \in \operatorname{Col} A$ so, by definition there must exist some $s \in \mathbb{R}^n$ such that $As = \widehat{b}$, and it then holds that $A^TAs = A^T\widehat{b} = A^Tb$. □

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

$$s = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

which is also the unique least-squares solution to $Ax = b$. Note that $As \neq b$ as

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

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

$$As = \begin{bmatrix} 2 \\ 4 \\ -8 \end{bmatrix}$$

is the point in the plane spanned by the columns of A in $\mathbb{R}^3$ that 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 $\boxed{s = (A^TA)^{-1}A^Tb}$, 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 s by solving the system $A^TAx = A^Tb$ via row reduction.

Proof. Suppose $s \in \mathbb{R}^n$ is a least-squares solution to $Ax = b$.

If $v \in \operatorname{Nul} A$ then then $s + v$ is also a least-squares solution since $\|As - b\| = \|A(s + v) - b\|$.

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 $\operatorname{Nul}(A^TA) = \operatorname{Nul} A = \{0\}$.

In this case, A^TA is invertible since it is a square matrix, so if (b) holds then (c) holds.

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 easier to compute a least-squares solution to $Ax = b$ if the columns of A are orthogonal:

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

$$\operatorname{proj}_{\operatorname{Col} A}(b) = \frac{\begin{bmatrix} -1 \\ 2 \\ 1 \\ 6 \end{bmatrix} \bullet \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}}{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \bullet \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{\begin{bmatrix} -1 \\ 2 \\ 1 \\ 6 \end{bmatrix} \bullet \begin{bmatrix} -6 \\ -2 \\ 1 \\ 7 \end{bmatrix}}{\begin{bmatrix} -6 \\ -2 \\ 1 \\ 7 \end{bmatrix} \bullet \begin{bmatrix} -6 \\ -2 \\ 1 \\ 7 \end{bmatrix}} \begin{bmatrix} -6 \\ -2 \\ 1 \\ 7 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} -6 \\ -2 \\ 1 \\ 7 \end{bmatrix}.$$

We deduce that $s = \begin{bmatrix} 2 \\ 1/2 \end{bmatrix}$ is a least-squares solution to $Ax = b$ since $As = \operatorname{proj}_{\operatorname{Col} A}(b)$.

3 Vocabulary

Keywords from today's lecture:

1. *Overdetermined* linear system.

   A linear system with more equations than variables.

2. *Least-squares solution* to a linear system $Ax = b$.

   Assume A is an $m \times n$ matrix and $b \in \mathbb{R}^m$. Two equivalent definitions:

   - A vector $s \in \mathbb{R}^n$ is a least-squares solution to $Ax = b$ if $\|b - As\| \leq \|b - Ax\|$ for all $x \in \mathbb{R}^n$.
   - A vector $s \in \mathbb{R}^n$ is a least-squares solution to $Ax = b$ if s is an exact solution to the (always consistent) linear system $A^T Ax = A^T b$.

   To find a least-squares solution to $Ax = b$, use row reduction to find a solution to $A^T Ax = A^T b$