Theorem : (Orthogonal Decomposition Theorem)

Let W be a subspace of $\mathbb{R}^n$. Then, each $y \in \mathbb{R}^n$ can be written uniquely as

$$y = \hat{y} + z$$

where $\hat{y}$ lies in W and z lies in $W^{\perp}$.

$\hat{y}$ is the projection of y onto W, and can be calculated as follows:

If $\{u_1, u_2, \dots, u_p\}$ is any orthogonal basis for W, then

$$\hat{y} = \frac{y \cdot u_1}{u_1 \cdot u_1} u_1 + \frac{y \cdot u_2}{u_2 \cdot u_2} u_2 + \cdots + \frac{y \cdot u_p}{u_p \cdot u_p} u$$

and $z = y - \hat{y}$.

(last class)

Remark:
When we did projection onto a line, or along a given vector u, we did the same thing: write $y = \hat{y} + z$ where $\hat{y} \in \text{Span}\{u\}$ and z was orthogonal to u. This is just a special case of the above theorem, taking $W = \text{span}\{u\}$.

Example: Let $u_1 = (2, 5, -1)$, $u_2 = (-2, 1, 1)$.

Find the projection of $y = (1, 2, 3)$ onto the subspace $W = \text{Span}\{u_1, u_2\}$.

Solution: Observe that $u_1 \cdot u_2 = 0$.

Therefore, $\{u_1, u_2\}$ is an orthogonal basis for W. (Do you understand why?)

Thus, the projection $\hat{y}$ can be calculated as

$$\hat{y} = \frac{y \cdot u_1}{u_1 \cdot u_1} u_1 + \frac{y \cdot u_2}{u_2 \cdot u_2} u_2$$

$$= \frac{(1,2,3) \cdot (2,5,-1)}{(2,5,-1) \cdot (2,5,-1)} (2,5,-1)$$

$$+ \frac{(1,2,3) \cdot (-2,1,1)}{(-2,1,1) \cdot (-2,1,1)} (-2,1,1).$$

$$= \frac{9}{30} (2,5,-1) + \frac{3}{6} (-2,1,1)$$

$$= \left(\frac{-2}{5}, 2, \frac{1}{5}\right).$$

Remarks: 1) Try calculating the projection of $(0,5,0)$ on W. Does the answer surprise you?

The point to note is that if $y \in W$, the projection of y onto W is y itself!

2) Be sure to use an orthogonal basis while calculation of $\hat{y}$. If the given basis of W is not orthogonal, we must first find an orthogonal basis for W and then use it to calculate projections.

Now do we find an orthogonal basis? Gram-Schmidt!

Example: Find the projection of $y = (1, 2, 1)$ on $W = \text{span}\{(1, 1, 0), (2, 1, 1)\}$

Solution: $\{(1, 1, 0), (2, 1, 1)\}$ is linearly independent but not orthogonal.

Let $x_1 = (1, 1, 0)$ $x_2 = (2, 1, 1)$. Applying Gram-Schmidt:

$$v_1 = x_1 = (1, 1, 0)$$

$$v_2 = x_2 - \frac{x_2 \cdot v_1}{v_1 \cdot v_1} v_1$$

$$= (2, 1, 1) - \frac{3}{2}(1, 1, 0)$$

$= \left(\frac{1}{2}, -\frac{1}{2}, 1\right) \rightsquigarrow (1, -1, 2)$.

scale by $\times 2$

Thus, $\{(1,1,0), (1,-1,2)\}$ is an orthogonal basis for W. We can now calculate $\hat{y}$ by:

$$\hat{y} = \frac{y \cdot v_1}{v_1 \cdot v_1} v_1 + \frac{y \cdot v_2}{v_2 \cdot v_2} v_2$$

$$= \frac{3}{2} v_1 + \frac{1}{6} v_2$$

$$= \frac{3}{2}(1,1,0) + \frac{1}{6}(1,-1,2)$$

$$= \left(\frac{10}{6}, \frac{8}{6}, \frac{2}{6}\right)$$

$$= \frac{1}{3}(5,4,1).$$

QR Factorization of matrices

Theorem: If A is an $m \times n$ matrix with linearly independent columns, then A can be factored as

$$A = QR$$

where:

- Q is an $m \times n$ matrix whose columns form an o.n.b for the column space of A
- R is an $n \times n$ upper triangular matrix that has positive entries on its diagonal.

Proof See textbook. (Lay-Lay-Macdonald pg. 403).

Process:

Step 1 Apply Gram-Schmidt algorithm to the column vectors of A. Let Q be the matrix formed by orthonormalizing the column vectors.

Step 2 Since $A = QR$, to find R, observe that $Q^TQ = I$ because the columns of Q are orthonormal. Hence

$$Q^TA = Q^T(QR) = (Q^TQ)R = IR = R.$$

Thus, R is obtained by computing Q^TA.

Example Find a QR factorization of

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Soln. Let $x_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$, $x_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$ and $x_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$.

Apply Gram-Schmidt:

$$v_1 = x_1 = (1,1,1,1)$$

$$v_2 = x_2 - \frac{x_2 \cdot v_1}{v_1 \cdot v_1} v_1$$

$$= (0,1,1,1) - \frac{3}{4}(1,1,1,1)$$

$$= \left(-\frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \rightsquigarrow (-3,1,1,1)$$

$$v_3 = x_3 - \frac{x_3 \cdot v_1}{v_1 \cdot v_1} v_1 - \frac{x_3 \cdot v_2}{v_2 \cdot v_2} v_2$$

$$= (0,0,1,1) - \frac{2}{4}(1,1,1,1) - \frac{2}{12}(-3,1,1,1)$$

$$= (0,0,1,1) - \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right) - \left(-\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$$

$$= \left(0, -\frac{2}{3}, \frac{1}{3}, \frac{1}{3}\right) \rightsquigarrow (0,-2,1,1)$$

to clear denominators

Apply Gram-Schmidt to get orthogonal vectors $v_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$, $v_2 = \begin{bmatrix} -3 \\ 1 \\ 1 \\ 1 \end{bmatrix}$, $v_3 = \begin{bmatrix} 0 \\ -2 \\ 1 \\ 1 \end{bmatrix}$.

divide by the norm to make them unit vectors

We then **normalize** these vectors to obtain $u_1 = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$, $u_2 = \begin{bmatrix} -3/\sqrt{12} \\ 1/\sqrt{12} \\ 1/\sqrt{12} \\ 1/\sqrt{12} \end{bmatrix}$, $u_3 = \begin{bmatrix} 0 \\ -2/\sqrt{6} \\ 1/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix}$.

$$\therefore Q = \begin{bmatrix} 1/2 & -3/\sqrt{12} & 0 \\ 1/2 & 1/\sqrt{12} & -2/\sqrt{6} \\ 1/2 & 1/\sqrt{12} & 1/\sqrt{6} \\ 1/2 & 1/\sqrt{12} & 1/\sqrt{6} \end{bmatrix}$$

$$R = Q^T A$$

$$= \begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ -3/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} \\ 0 & -2/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 3/2 & 1 \\ 0 & 3/\sqrt{12} & 2/\sqrt{12} \\ 0 & 0 & 2/\sqrt{6} \end{bmatrix}.$$

Least squares approximation

If $A\bar{x} = \bar{b}$ has no solution, we look for an "approximate" solution, i.e, try to find x such that $\|A\bar{x} - \bar{b}\|$ is as small as possible.

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

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

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

Solution of the general least-squares problem:

Theorem: The set of least-squares solutions of $A\bar{x} = \bar{b}$ is the solution set of the linear system:

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

Example 1 Find a least-squares solution of the inconsistent system $A\bar{x} = b$ for

$$A = \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix}$$

Solution: We solve the system

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

$$A^T A = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix}$$

$$A^T b = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix} = \begin{bmatrix} 19 \\ 11 \end{bmatrix}$$

Let $\hat{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ s.t $A^T A \hat{x} = A^T b$

Then $\begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 19 \\ 11 \end{bmatrix}$

$$\therefore \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{84} \begin{bmatrix} 5 & -1 \\ -1 & 17 \end{bmatrix} \begin{bmatrix} 19 \\ 11 \end{bmatrix}$$

$$= \frac{1}{84} \begin{bmatrix} 84 \\ 168 \end{bmatrix}$$

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

The vector $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ is a least-squares solution of the given linear system.

Example 2: Find a least-squares solution of $Ax = b$ for $A = \begin{bmatrix} 1 & -6 \\ 1 & -2 \\ 1 & 1 \\ 1 & 7 \end{bmatrix}$ $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $b = \begin{bmatrix} -1 \\ 2 \\ 1 \\ 6 \end{bmatrix}$.

Solution: $A^TA =$

-1
0
$3/2$ $2+\frac{7}{2}$
$11/2$

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ -6 & -2 & 1 & 7 \end{bmatrix} \begin{bmatrix} 1 & -6 \\ 1 & -2 \\ 1 & 1 \\ 1 & 7 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 90 \end{bmatrix}$$

$$A^Tb = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -6 & -2 & 1 & 7 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \\ 1 \\ 6 \end{bmatrix} = \begin{bmatrix} 8 \\ 45 \end{bmatrix}$$

The least-squares solution is obtained by solving $A^TA\hat{x} = A^Tb$:

$$\begin{bmatrix} 4 & 0 \\ 0 & 90 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 45 \end{bmatrix}$$

$\therefore x_1 = 2 \quad x_2 = 1/2.$ $\qquad \therefore \hat{x} = \begin{bmatrix} 2 \\ 1/2 \end{bmatrix}$

Def.
The least-squares error is the quantity $\|A\hat{x} - b\|$.

In example 1:

$$A = \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix}, \quad \hat{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

So the least-squares error is

$$\|A\hat{x} - b\| = \|(4,4,3) - (2,0,11)\|$$

$$= \|(2,4,-8)\|$$

$$= \sqrt{2^2 + 4^2 + 8^2}$$

$$= \sqrt{84}.$$

64
16
8

Similarly, in example 2:

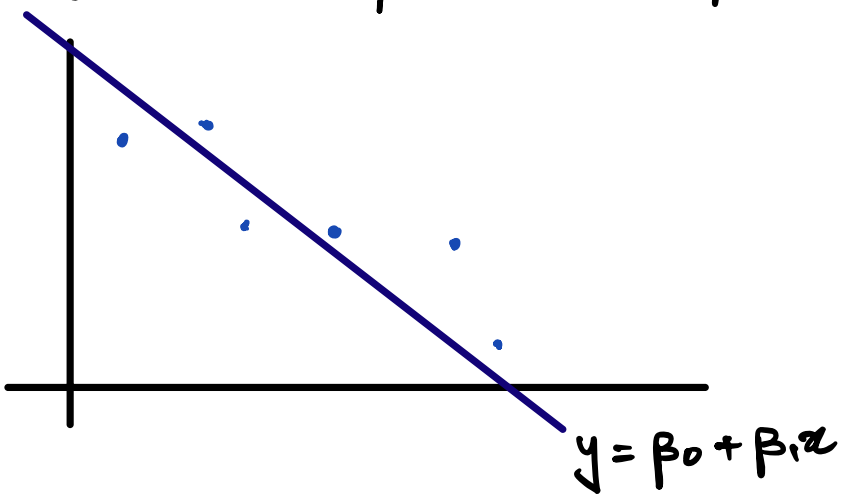
$$\|A\hat{x} - b\| = \|(-1,0,5/2,11/2) - (-1,2,1,6)\|$$

$$= \|(0,-2,3/2,-1/2)\| = \sqrt{4 + \frac{9}{4} + \frac{1}{4}} = \frac{\sqrt{26}}{2}.$$

Application to linear models

Experimental data often provides points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) that, when graphed, seem to lie close to a line. We want to determine parameters β_0

and β. that make the line $y = \beta_0 + \beta_1 x$ as close to the points as possible:



Suppose the data points are (x_j, y_j).

To each such point, there is a point

$$(x_j, \beta_0 + \beta_1 x_j)$$

lying on the line $y = \beta_0 + \beta_1 x$.

We y_j as the <u>observed value</u> of y and $\beta_0 + \beta_1 x_j$ as the <u>predicted value</u> of y.

The difference between the two is called a "residual".

If the data points were on the line, the parameters β_0 and β_1 would satisfy the equations

Predicted y-value		Observed y-value
$\beta_0 + \beta_1 x_1$	$=$	y_1
$\beta_0 + \beta_1 x_2$	$=$	y_2
$\vdots$		$\vdots$
$\beta_0 + \beta_1 x_n$	$=$	y_n

Converting this into a matrix equation, we have

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Since the data-points are almost never actually going to fall on a line, the above system is almost always going to be inconsistent.

The next best thing to do is to find an approximate solution, i.e., the β_0, β_1 such that it best fits the data.
i.e., we find the least-squares solution!

Example : Find the equation $y = \beta_0 + \beta_1 x$ of the least-squares line that best fits the data points $(2, 1), (5, 2), (7, 3)$ and $(8, 3)$.

Solution .

We want:

$$\beta_0 + \beta_1 \cdot 2 = 1$$
$$\beta_0 + \beta_1 \cdot 5 = 2$$
$$\beta_0 + \beta_1 \cdot 7 = 3$$
$$\beta_0 + \beta_1 \cdot 8 = 3$$

To obtain the least-squares line, we solve $X^T X \beta = X^T y$; where

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix}, \quad X^T y = \begin{bmatrix} 9 \\ 57 \end{bmatrix}$$

∴ We solve:

$$\begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 9 \\ 57 \end{bmatrix}$$

Solving gives $\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 2/7 \\ 5/14 \end{bmatrix}$.

Thus, the least-squares line has the equation $y = \frac{2}{7} + \frac{5}{14} x$.

Extra material for your understanding:

We learned two techniques : Gram-Schmidt process to obtain an orthogonal basis, and a method to get an approximate solution to a linear system that has no solutions.

You might be wondering:

1. Why does the Gram-Schmidt process work?

2. Why does solving $A^TA\bar{x} = A^Tb$ give the best approximate solution to an inconsistent system?

To answer the above questions, we start by recalling the foll. result:

Theorem : (Orthogonal Decomposition Theorem)

Let W be a subspace of $\mathbb{R}^n$. Then, each $y \in \mathbb{R}$ can be written uniquely as

$$y = \hat{y} + z$$

where $\hat{y}$ lies in W and z lies in $W^{\perp}$.

$\hat{y}$ is the projection of y onto W, and can be calculated as follows:

If $\{u_1, u_2, \ldots, u_p\}$ is any orthogonal basis for W, then

$$\hat{y} = \frac{y \cdot u_1}{u_1 \cdot u_1} u_1 + \frac{y \cdot u_2}{u_2 \cdot u_2} u_2 + \cdots + \frac{y \cdot u_p}{u_p \cdot u_p} u_p$$

(i.e., $\hat{y}$ is the projection of y onto W)

and $z = y - \hat{y}$.

(1) Why the Gram-Schmidt algorithm works:

We construct each vector by making sure it is orthogonal to the vectors constructed before it:

Let the given basis be $\{x_1, \ldots, x_p\}$.

Then, $v_1 = x_1$

$$v_2 = x_2 - \frac{x_2 \cdot v_1}{v_1 \cdot v_1} v_1$$

→ orth. projection of x_2 onto $\text{span}\{v_1\} = \text{span}\{x_1\}$

This is of the form $z = y - \hat{y}$ in the above theorem, with $W = \text{span}\{v_1\}$, Here we have $v_2 = x_2 - \hat{x}_2$, so $v_2 \in W^\perp$, i.e., v_2 is orthogonal to v_1.

Next,

$$v_3 = x_3 - \left(\frac{x_3 \cdot v_1}{v_1 \cdot v_1} v_1 + \frac{x_3 \cdot v_2}{v_2 \cdot v_2} \right)$$

→ orth. proj'n of x_3 onto $\text{span}\{v_1, v_2\} = \text{span}\{x_1, x_2\}$.

This is of the form $z = y - \hat{y}$ of the theorem, this time $W = \text{span}\{v_1, v_2\}$.

So $v_3 = x_3 - \hat{x}_3$, and $v_3 \in W^\perp$, so v_3 is orthogonal to BOTH v_1 and v_2.

Proceeding in this manner,

$$\vdots$$

$$v_i = x_i - \underbrace{\left(\frac{x_i \cdot v_1}{v_1 \cdot v_1} v_1 + \cdots + \frac{x_i \cdot v_{i-1}}{v_{i-1} \cdot v_{i-1}} v_{i-1} \right)}_{\text{orth proj'n of } x_i \text{ onto } \operatorname{span}\{v_1, \ldots, v_{i-1}\} = \operatorname{span}\{x_1, \ldots, x_{i-1}\}}$$

$$v_i = x_i - \hat{x}_i$$

where x_i is projected on $W = \operatorname{span}\{v_1, \ldots, v_{i-1}\}$

and so $v_i \in W^{\perp}$, which means

v_i is orthogonal to $v_1, \ldots, v_{i-1}$.

(2) Why the least-squares solution works:

Theorem: (Best approximation theorem)

Let W be a subspace of $\mathbb{R}^n$, let y be any vector in $\mathbb{R}^n$, and let $\hat{y}$ be the orthogonal projection of y onto W. Then $\hat{y}$ is the

closest point in W to y, in the sense that

$$\|y - \hat{y}\| < \|y - v\|$$

for all vectors v in W, with $v \neq \hat{y}$.

Now, in the problem of finding the best approximate solution of an inconsistent linear system:

$$Ax = b,$$

we are using the above result as follows:

Recall that solving $Ax = b$ is the same as determining whether or not b is in the column space of A, i.e., $C(A)$.

If the system is inconsistent, then $b \notin C(A)$. So letting W be the subspace $C(A)$, we are looking for

a vector "closest" to the subspace W,
and the best approximation theorem
says that this vector must be $\hat{b}$,
the projection of b onto the
column space $\mathrm{Col}(A)$

This means there does exist $\hat{x}$
solving the linear system $A\hat{x} = \hat{b}$.

How to find $\hat{x}$?

Since $\hat{b}$ is the orth. proj of b onto $\mathrm{Col}(A)$ (= W),

$b - \hat{b}$ lies in $W^{\perp}$.

Fact: The orthogonal complement of
$\mathrm{Col}(A)$ is the null space of A^T.

i.e., $A^T(b - \hat{b}) = 0$

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

$\therefore \hat{x}$ solves $A^T b = A^T A\hat{x}$.

This is why we solve the system $A^TAx = A^Tb$

to obtain the least-squares solution $\hat{x}$.