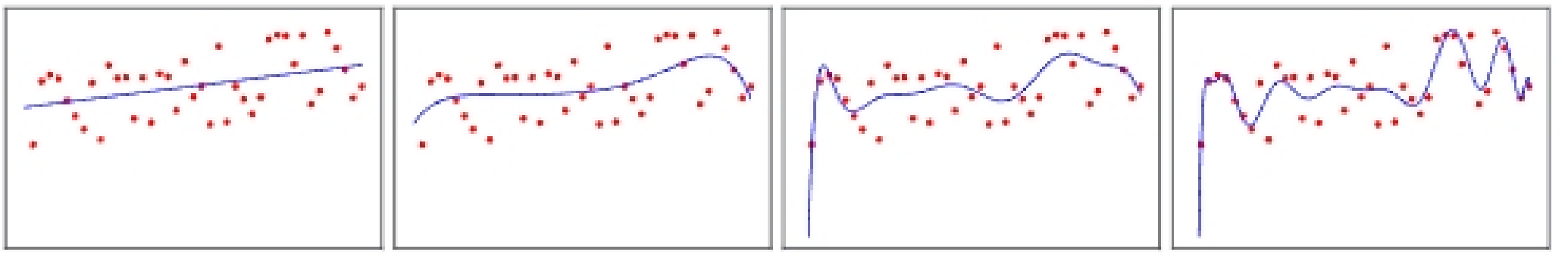
Section 5.4 : 2,10,22,34,40,16*,18*

GOAL. The best possible "solution" of an inconsistent linear systems $Ax = b$ is called the **least square solution**. It is the orthogonal projection of b onto the image $\mathrm{im}(A)$ of A. What we know about the kernel and the image of linear transformations helps to understand this situation and leads to an explicit formulas for the least square fit. Why do we care about non-consistent systems? Often we have to solve linear systems of equations with more constraints than variables. An example is when we try to find the best polynomial which passes through a set of points. This problem is called **data fitting**. If we wanted to accommodate all data, the degree of the polynomial would become too large. The fit would look too wiggly. Taking a smaller degree polynomial will not only be more convenient but also give a better picture. Especially important is **regression**, the fitting of data with lines.



The above pictures show 30 data points which are fitted best with polynomials of degree 1, 6, 11 and 16. The first linear fit maybe tells most about the trend of the data.

THE ORTHOGONAL COMPLEMENT OF $\mathrm{im}(A)$. Because a vector is in the kernel of A^T if and only if it is orthogonal to the rows of A^T and so to the columns of A, the kernel of A^T is the orthogonal complement of $\mathrm{im}(A)$: $\boxed{(\mathrm{im}(A))^\perp = \ker(A^T)}$

EXAMPLES.

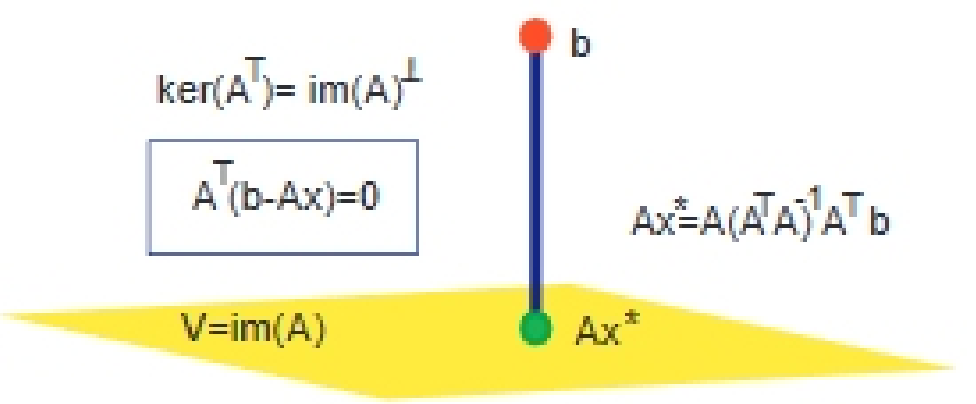
1) $A = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$. The kernel V of $A^T = \begin{bmatrix} a & b & c \end{bmatrix}$ consists of all vectors satisfying $ax + by + cz = 0$. V is a plane. The orthogonal complement is the image of A which is spanned by the normal vector $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ to the plane.

2) $A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$. The image of A is spanned by $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ the kernel of A^T is spanned by $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

ORTHOGONAL PROJECTION. If $\vec{b}$ is a vector and V is a linear subspace, then $\mathrm{proj}_V(\vec{b})$ is the vector closest to $\vec{b}$ on V: given any other vector $\vec{v}$ on V, one can form the triangle $\vec{b}, \vec{v}, \mathrm{proj}_V(\vec{b})$ which has a right angle at $\mathrm{proj}_V(\vec{b})$ and invoke Pythagoras.

THE KERNEL OF A^TA. For any $m \times n$ matrix $\boxed{\ker(A) = \ker(A^TA)}$ Proof. $\subset$ is clear. On the other hand $A^TAv = 0$ means that Av is in the kernel of A^T. But since the image of A is orthogonal to the kernel of A^T, we have $A\vec{v} = 0$, which means $\vec{v}$ is in the kernel of A.

LEAST SQUARE SOLUTION. The **least square solution** of $A\vec{x} = \vec{b}$ is the vector $\vec{x}^*$ such that $A\vec{x}^*$ is closest to $\vec{b}$ from all other vectors $A\vec{x}$. In other words, $A\vec{x}^* = \mathrm{proj}_V(\vec{b})$, where $V = \mathrm{im}(V)$. Because $\vec{b} - A\vec{x}^*$ is in $V^\perp = \mathrm{im}(A)^\perp = \ker(A^T)$, we have $A^T(\vec{b} - A\vec{x}^*) = 0$. The last equation means that $\vec{x}^*$ is a solution of $\boxed{A^TA\vec{x} = A^T\vec{b}, \text{ the } \textbf{normal equation} \text{ of } A\vec{x} = \vec{b}}$. If the kernel of A is trivial, then the kernel of A^TA is trivial and A^TA can be inverted. Therefore $\boxed{\vec{x}^* = (A^TA)^{-1}A^T\vec{b} \text{ is the least square solution.}}$



WHY LEAST SQUARES? If $\vec{x}^*$ is the least square solution of $A\vec{x} = \vec{b}$ then $||A\vec{x}^* - \vec{b}|| \leq ||A\vec{x} - \vec{b}||$ for all $\vec{x}$. Proof. $A^T(A\vec{x}^* - \vec{b}) = 0$ means that $A\vec{x}^* - \vec{b}$ is in the kernel of A^T which is orthogonal to $V = \text{im}(A)$. That is $\text{proj}_V(\vec{b}) = A\vec{x}^*$ which is the closest point to $\vec{b}$ on V.

ORTHOGONAL PROJECTION If $\vec{v}_1, \ldots, \vec{v}_n$ is a basis in V which is not necessarily orthonormal, then the orthogonal projection is $\boxed{\vec{x} \mapsto A(A^TA)^{-1}A^T(\vec{x})}$ where $A = [\vec{v}_1, \ldots, \vec{v}_n]$.
Proof. $\vec{x} = (A^TA)^{-1}A^T\vec{b}$ is the least square solution of $A\vec{x} = \vec{b}$. Therefore $A\vec{x} = A(A^TA)^{-1}A^T\vec{b}$ is the vector in $\text{im}(A)$ closest to $\vec{b}$.

Special case: If $\vec{w}_1, \ldots, \vec{w}_n$ is an orthonormal basis in V, we had seen earlier that AA^T with $A = [\vec{w}_1, \ldots, \vec{w}_n]$ is the orthogonal projection onto V (this was just rewriting $A\vec{x} = (\vec{w}_1 \cdot \vec{x})\vec{w}_1 + \cdots + (\vec{w}_n \cdot \vec{x})\vec{w}_n$ in matrix form.) This follows from the above formula because $A^TA = I$ in that case.

EXAMPLE Let $A = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$. The orthogonal projection onto $V = \text{im}(A)$ is $\vec{b} \mapsto A(A^TA)^{-1}A^T\vec{b}$. We have

$A^TA = \begin{bmatrix} 5 & 0 \\ 2 & 1 \end{bmatrix}$ and $A(A^TA)^{-1}A^T = \begin{bmatrix} 1/5 & 2/5 & 0 \\ 2/5 & 4/5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

For example, the projection of $\vec{b} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ is $\vec{x}^* = \begin{bmatrix} 2/5 \\ 4/5 \\ 0 \end{bmatrix}$ and the distance to $\vec{b}$ is $1/\sqrt{5}$. The point $\vec{x}^*$ is the point on V which is closest to $\vec{b}$.

Remember the formula for the distance of $\vec{b}$ to a plane V with normal vector $\vec{n}$? It was $d = |\vec{n} \cdot \vec{b}|/||\vec{n}||$. In our case, we can take $\vec{n} = [-2, 1, 0]$ and get the distance $1/\sqrt{5}$. Let's check: the distance of $\vec{x}^*$ and $\vec{b}$ is $||(2/5, -1/5, 0)|| = 1/\sqrt{5}$.

EXAMPLE. Let $A = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \end{bmatrix}$. Problem: find the matrix of the orthogonal projection onto the image of A.
The image of A is a one-dimensional line spanned by the vector $\vec{v} = (1, 2, 0, 1)$. We calculate $A^TA = 6$. Then

$$A(A^TA)^{-1}A^T = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 & 1 \end{bmatrix} /6 = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 2 & 4 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix} /6$$

DATA FIT. Find a quadratic polynomial $p(t) = at^2 + bt + c$ which best fits the four data points $(-1, 8), (0, 8), (1, 4), (2, 16)$.

$$A = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \end{bmatrix} \quad \vec{b} = \begin{bmatrix} 8 \\ 8 \\ 4 \\ 16 \end{bmatrix}^T. \quad A^TA = \begin{bmatrix} 18 & 8 & 6 \\ 8 & 6 & 2 \\ 6 & 2 & 4 \end{bmatrix} \text{ and } \vec{x}^* = (A^TA)^{-1}A^T\vec{b} = \begin{bmatrix} 3 \\ -1 \\ 5 \end{bmatrix}.$$

Software packages like Mathematica have already built in the facility to fit numerical data:

```
DataPoints = {{-1, 8}, {0, 8}, {1, 4}, {2, 16}}
f=Function[y, Fit[DataPoints, {1, x, x^2}, x] /. x->y];
Show[{ListPlot[DataPoints], Plot[f[t], {t, -1, 2}]}];

Series[f[x], {x, 0, 2}]
```

The series expansion of f showed that indeed, $f(t) = 5 - t + 3t^2$ is indeed best quadratic fit. Actually, Mathematica does the same to find the fit then what we do: **"Solving" an inconsistent system of linear equations as best as possible.**

PROBLEM: Prove $\text{im}(A) = \text{im}(AA^T)$.

SOLUTION. The image of AA^T is contained in the image of A because we can write $\vec{v} = AA^T\vec{x}$ as $\vec{v} = A\vec{y}$ with $\vec{y} = A^T\vec{x}$. On the other hand, if $\vec{v}$ is in the image of A, then $\vec{v} = A\vec{x}$. If $\vec{x} = \vec{y} + \vec{z}$, where $\vec{y}$ in the kernel of A and $\vec{z}$ orthogonal to the kernel of A, then $A\vec{x} = A\vec{z}$. Because $\vec{z}$ is orthogonal to the kernel of A, it is in the image of A^T. Therefore, $\vec{z} = A^T\vec{u}$ and $\vec{v} = A\vec{z} = AA^T\vec{u}$ is in the image of AA^T.