## 4.3 Least Squares Approximations

It often happens that $A\boldsymbol{x} = \boldsymbol{b}$ has no solution. The usual reason is: *too many equations*. The matrix has more rows than columns. There are more equations than unknowns ($m$ is greater than $n$). The $n$ columns span a small part of $m$-dimensional space. Unless all measurements are perfect, $\boldsymbol{b}$ is outside that column space. Elimination reaches an impossible equation and stops. But we can't stop just because measurements include noise.

To repeat: We cannot always get the error $\boldsymbol{e} = \boldsymbol{b} - A\boldsymbol{x}$ down to zero. When $\boldsymbol{e}$ is zero, $\boldsymbol{x}$ is an exact solution to $A\boldsymbol{x} = \boldsymbol{b}$. *When the length of $\boldsymbol{e}$ is as small as possible, $\widehat{\boldsymbol{x}}$ is a least squares solution.* Our goal in this section is to compute $\widehat{\boldsymbol{x}}$ and use it. These are real problems and they need an answer.

The previous section emphasized $\boldsymbol{p}$ (the projection). This section emphasizes $\widehat{\boldsymbol{x}}$ (the least squares solution). They are connected by $\boldsymbol{p} = A\widehat{\boldsymbol{x}}$. The fundamental equation is still $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$. Here is a short unofficial way to reach this equation:

**When $A\boldsymbol{x} = \boldsymbol{b}$ has no solution, multiply by $A^{\mathrm{T}}$ and solve** $\boxed{A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}.}$

**Example 1** A crucial application of least squares is fitting a straight line to $m$ points. Start with three points: *Find the closest line to the points* $(0, 6)$, $(1, 0)$, *and* $(2, 0)$.

No straight line $b = C + Dt$ goes through those three points. We are asking for two numbers $C$ and $D$ that satisfy three equations. Here are the equations at $t = 0, 1, 2$ to match the given values $b = 6, 0, 0$:

| | | |
|---|---|---|
| $\boldsymbol{t = 0}$ | The first point is on the line $b = C + Dt$ if | $C + D \cdot 0 = 6$ |
| $\boldsymbol{t = 1}$ | The second point is on the line $b = C + Dt$ if | $C + D \cdot 1 = 0$ |
| $\boldsymbol{t = 2}$ | The third point is on the line $b = C + Dt$ if | $C + D \cdot 2 = 0.$ |

This 3 by 2 system has *no solution*: $\boldsymbol{b} = (6, 0, 0)$ is not a combination of the columns $(1, 1, 1)$ and $(0, 1, 2)$. Read off $A$, $\boldsymbol{x}$, and $\boldsymbol{b}$ from those equations:

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} C \\ D \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} \quad A\boldsymbol{x} = \boldsymbol{b} \text{ is } \textit{not} \text{ solvable.}$$

The same numbers were in Example 3 in the last section. We computed $\widehat{\boldsymbol{x}} = (5, -3)$. Those numbers are the best $C$ and $D$, so $5 - 3t$ will be the best line for the 3 points. We must connect projections to least squares, by explaining why $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$.

In practical problems, there could easily be $m = 100$ points instead of $m = 3$. They don't exactly match any straight line $C + Dt$. Our numbers $6, 0, 0$ exaggerate the error so you can see $e_1$, $e_2$, and $e_3$ in Figure 4.6.

### Minimizing the Error

How do we make the error $\boldsymbol{e} = \boldsymbol{b} - A\boldsymbol{x}$ as small as possible? This is an important question with a beautiful answer. The best $\boldsymbol{x}$ (called $\widehat{\boldsymbol{x}}$) can be found by geometry or algebra or calculus: 90° angle or project using $P$ or set the derivative of the error to zero.

**By geometry** Every $Ax$ lies in the plane of the columns $(1,1,1)$ and $(0,1,2)$. In that plane, we look for the point closest to $\boldsymbol{b}$. *The nearest point is the projection* $\boldsymbol{p}$.

The best choice for $A\widehat{\boldsymbol{x}}$ is $\boldsymbol{p}$. The smallest possible error is $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p}$. The three points at heights $(p_1, p_2, p_3)$ *do lie on a line*, because $\boldsymbol{p}$ is in the column space. In fitting a straight line, $\widehat{\boldsymbol{x}}$ gives the best choice for $(C, D)$.

**By algebra** Every vector $\boldsymbol{b}$ splits into two parts. The part in the column space is $\boldsymbol{p}$. The perpendicular part in the nullspace of $A^{\mathrm{T}}$ is $\boldsymbol{e}$. There is an equation we cannot solve $(A\boldsymbol{x} = \boldsymbol{b})$. There is an equation $A\widehat{\boldsymbol{x}} = \boldsymbol{p}$ we do solve (by removing $\boldsymbol{e}$):

$$A\boldsymbol{x} = \boldsymbol{b} = \boldsymbol{p} + \boldsymbol{e} \quad \text{is impossible;} \qquad A\widehat{\boldsymbol{x}} = \boldsymbol{p} \quad \text{is solvable.} \tag{1}$$

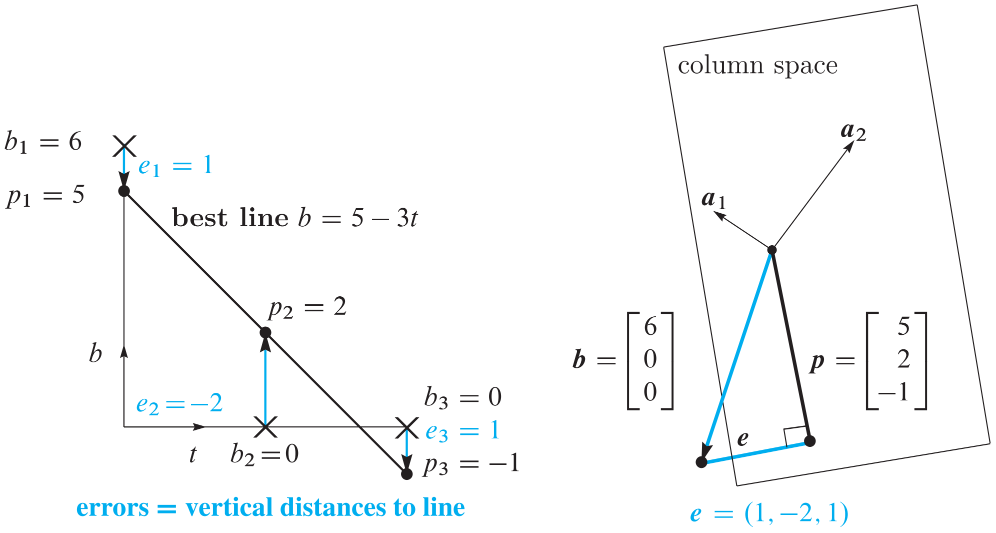
The solution to $A\widehat{\boldsymbol{x}} = \boldsymbol{p}$ leaves the least possible error (which is $\boldsymbol{e}$):

$$\textbf{Squared length for any } \boldsymbol{x} \qquad \|A\boldsymbol{x} - \boldsymbol{b}\|^2 = \|A\boldsymbol{x} - \boldsymbol{p}\|^2 + \|\boldsymbol{e}\|^2. \tag{2}$$

This is the law $c^2 = a^2 + b^2$ for a right triangle. The vector $A\boldsymbol{x} - \boldsymbol{p}$ in the column space is perpendicular to $\boldsymbol{e}$ in the left nullspace. We reduce $A\boldsymbol{x} - \boldsymbol{p}$ to zero by choosing $\boldsymbol{x}$ to be $\widehat{\boldsymbol{x}}$. That leaves the smallest possible error $\boldsymbol{e} = (e_1, e_2, e_3)$.

Notice what "smallest" means. The *squared length* of $A\boldsymbol{x} - \boldsymbol{b}$ is minimized:

***The least squares solution $\widehat{x}$ makes $E = \|Ax - b\|^2$ as small as possible.***



$b_1 = 6$
$e_1 = 1$
$p_1 = 5$
best line $b = 5 - 3t$
$p_2 = 2$
$b$
$e_2 = -2$
$b_3 = 0$
$e_3 = 1$
$t$
$b_2 = 0$
$p_3 = -1$
errors = vertical distances to line
column space
$\boldsymbol{a}_2$
$\boldsymbol{a}_1$
$\boldsymbol{b} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$
$\boldsymbol{p} = \begin{bmatrix} 5 \\ 2 \\ -1 \end{bmatrix}$
$\boldsymbol{e}$
$\boldsymbol{e} = (1, -2, 1)$

Figure 4.6: **Best line and projection: Two pictures, same problem.** The line has heights $\boldsymbol{p} = (5, 2, -1)$ with errors $\boldsymbol{e} = (1, -2, 1)$. The equations $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$ give $\widehat{\boldsymbol{x}} = (5, -3)$. The best line is $b = 5 - 3t$ and the projection is $\boldsymbol{p} = 5\boldsymbol{a}_1 - 3\boldsymbol{a}_2$.

Figure 4.6a shows the closest line. It misses by distances $e_1, e_2, e_3 = 1, -2, 1$. *Those are vertical distances*. The least squares line minimizes $E = e_1^2 + e_2^2 + e_3^2$.

Figure 4.6b shows the same problem in 3-dimensional space ($\boldsymbol{b}\ \boldsymbol{p}\ \boldsymbol{e}$ space). The vector $\boldsymbol{b}$ is not in the column space of $A$. That is why we could not solve $A\boldsymbol{x} = \boldsymbol{b}$. No line goes through the three points. The smallest possible error is the perpendicular vector $\boldsymbol{e}$. This is $\boldsymbol{e} = \boldsymbol{b} - A\widehat{\boldsymbol{x}}$, the vector of errors $(1, -2, 1)$ in the three equations. Those are the distances from the best line. Behind both figures is the fundamental equation $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$.

Notice that the errors $1, -2, 1$ add to zero. The error $\boldsymbol{e} = (e_1, e_2, e_3)$ is perpendicular to the first column $(1, 1, 1)$ in $A$. The dot product gives $e_1 + e_2 + e_3 = 0$.

**By calculus** Most functions are minimized by calculus! The graph bottoms out and the derivative in every direction is zero. Here the error function $E$ to be minimized is a *sum of squares* $e_1^2 + e_2^2 + e_3^2$ (the square of the error in each equation):

$$E = \|A\boldsymbol{x} - \boldsymbol{b}\|^2 = (C + D \cdot 0 - 6)^2 + (C + D \cdot 1)^2 + (C + D \cdot 2)^2. \tag{3}$$

The unknowns are $C$ and $D$. With two unknowns there are *two derivatives*—both zero at the minimum. They are "partial derivatives" because $\partial E/\partial C$ treats $D$ as constant and $\partial E/\partial D$ treats $C$ as constant:

$$\begin{aligned}\partial E/\partial C &= 2(C + D \cdot 0 - 6) \quad + 2(C + D \cdot 1) \quad + 2(C + D \cdot 2) \quad = 0\\ \partial E/\partial D &= 2(C + D \cdot 0 - 6)(\mathbf{0}) + 2(C + D \cdot 1)(\mathbf{1}) + 2(C + D \cdot 2)(\mathbf{2}) = 0.\end{aligned}$$

$\partial E/\partial D$ contains the extra factors $\mathbf{0}, \mathbf{1}, \mathbf{2}$ from the chain rule. (The last derivative from $(C + 2\mathrm{D})^2$ was 2 times $C + 2\mathrm{D}$ times that extra 2.) In the $C$ derivative the corresponding factors are $1, 1, 1$, because $C$ is always multiplied by 1. It is no accident that 1, 1, 1 and 0, 1, 2 are the columns of $A$.

Now cancel 2 from every term and collect all $C$'s and all $D$'s:

$$\begin{array}{ll}\text{The } C \text{ derivative is zero:} & 3C + 3D = 6\\ \text{The } D \text{ derivative is zero:} & 3C + 5D = 0\end{array} \qquad \textbf{This matrix} \begin{bmatrix} 3 & 3\\ 3 & 5\end{bmatrix} \textbf{ is } \boldsymbol{A^{\mathrm{T}}A} \tag{4}$$

***These equations are identical with*** $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$. The best $C$ and $D$ are the components of $\widehat{\boldsymbol{x}}$. The equations from calculus are the same as the "normal equations" from linear algebra. These are the key equations of least squares:

***The partial derivatives of*** $\|A\boldsymbol{x} - \boldsymbol{b}\|^2$ ***are zero when*** $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$.

The solution is $C = 5$ and $D = -3$. Therefore $b = 5 - 3t$ is the best line—it comes closest to the three points. At $t = 0, 1, 2$ this line goes through $\boldsymbol{p} = 5, 2, -1$. It could not go through $\boldsymbol{b} = 6, 0, 0$. The errors are $1, -2, 1$. This is the vector $\boldsymbol{e}$!

## The Big Picture

The key figure of this book shows the four subspaces and the true action of a matrix. The vector $\boldsymbol{x}$ on the left side of Figure 4.3 went to $\boldsymbol{b} = A\boldsymbol{x}$ on the right side. In that figure $\boldsymbol{x}$ was split into $\boldsymbol{x}_r + \boldsymbol{x}_n$. There were many solutions to $A\boldsymbol{x} = \boldsymbol{b}$.

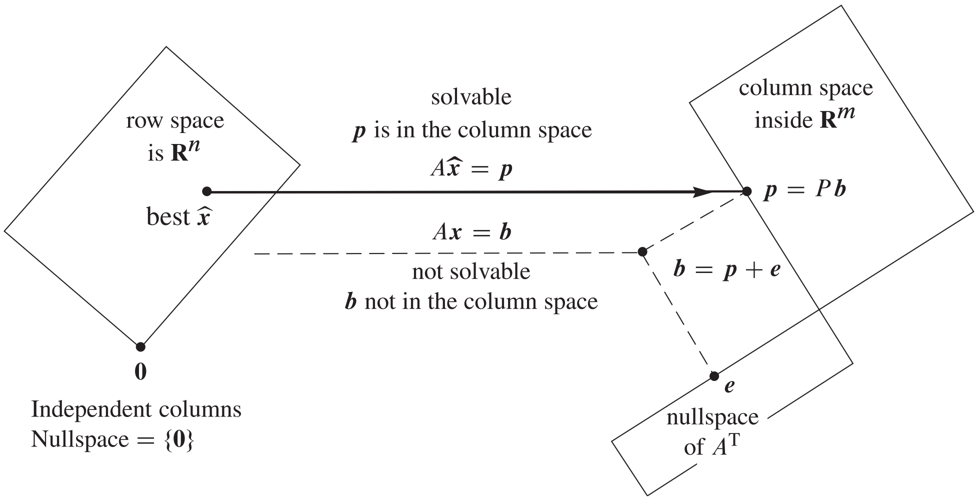

row space
is $\mathbf{R}^n$
best $\widehat{\boldsymbol{x}}$
solvable
$\boldsymbol{p}$ is in the column space
$A\widehat{\boldsymbol{x}} = \boldsymbol{p}$
$A\boldsymbol{x} = \boldsymbol{b}$
not solvable
$\boldsymbol{b}$ not in the column space
column space
inside $\mathbf{R}^m$
$\boldsymbol{p} = P\boldsymbol{b}$
$\boldsymbol{b} = \boldsymbol{p} + \boldsymbol{e}$
$\boldsymbol{e}$
nullspace
of $A^{\mathrm{T}}$
$\mathbf{0}$
Independent columns
Nullspace = $\{\mathbf{0}\}$

Figure 4.7: The projection $\boldsymbol{p} = A\widehat{\boldsymbol{x}}$ is closest to $\boldsymbol{b}$, so $\widehat{\boldsymbol{x}}$ minimizes $E = \|\boldsymbol{b} - A\boldsymbol{x}\|^2$.

In this section the situation is just the opposite. There are *no* solutions to $A\boldsymbol{x} = \boldsymbol{b}$. *Instead of splitting up $\boldsymbol{x}$ we are splitting up $\boldsymbol{b}$.* Figure 4.3 shows the big picture for least squares. Instead of $A\boldsymbol{x} = \boldsymbol{b}$ we solve $A\widehat{\boldsymbol{x}} = \boldsymbol{p}$. The error $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p}$ is unavoidable.

Notice how the nullspace $\boldsymbol{N}(A)$ is very small—just one point. With independent columns, the only solution to $A\boldsymbol{x} = \mathbf{0}$ is $\boldsymbol{x} = \mathbf{0}$. Then $A^{\mathrm{T}}A$ is invertible. The equation $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$ fully determines the best vector $\widehat{\boldsymbol{x}}$. The error has $A^{\mathrm{T}}\boldsymbol{e} = \mathbf{0}$.

Chapter 7 will have the complete picture—all four subspaces included. Every $\boldsymbol{x}$ splits into $\boldsymbol{x}_r + \boldsymbol{x}_n$, and every $\boldsymbol{b}$ splits into $\boldsymbol{p} + \boldsymbol{e}$. The best solution is $\widehat{\boldsymbol{x}}_r$ in the row space. We can't help $\boldsymbol{e}$ and we don't want $\boldsymbol{x}_n$—this leaves $A\widehat{\boldsymbol{x}} = \boldsymbol{p}$.

## Fitting a Straight Line

Fitting a line is the clearest application of least squares. It starts with $m > 2$ points, hopefully near a straight line. At times $t_1, \ldots, t_m$ those $m$ points are at heights $b_1, \ldots, b_m$. The best line $C + Dt$ misses the points by vertical distances $e_1, \ldots, e_m$. No line is perfect, and the least squares line minimizes $E = e_1^2 + \cdots + e_m^2$.

The first example in this section had three points in Figure 4.6. Now we allow $m$ points (and $m$ can be large). The two components of $\widehat{\boldsymbol{x}}$ are still $C$ and $D$.

A line goes through the $m$ points when we exactly solve $A\boldsymbol{x} = \boldsymbol{b}$. Generally we can't do it. Two unknowns $C$ and $D$ determine a line, so $A$ has only $n = 2$ columns. To fit the $m$ points, we are trying to solve $m$ equations (and we only want two!):

$$A\boldsymbol{x} = \boldsymbol{b} \quad \text{is} \quad \begin{aligned} C + Dt_1 &= b_1 \\ C + Dt_2 &= b_2 \\ &\vdots \\ C + Dt_m &= b_m \end{aligned} \quad \text{with} \quad A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix}. \tag{5}$$

The column space is so thin that almost certainly $\boldsymbol{b}$ is outside of it. When $\boldsymbol{b}$ happens to lie in the column space, the points happen to lie on a line. In that case $\boldsymbol{b} = \boldsymbol{p}$. Then $A\boldsymbol{x} = \boldsymbol{b}$ is solvable and the errors are $\boldsymbol{e} = (0, \ldots, 0)$.

> *The closest line $C + Dt$ has heights $p_1, \ldots, p_m$ with errors $e_1, \ldots, e_m$.*
>
> *Solve $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b$ for $\widehat{x} = (C, D)$. The errors are $e_i = b_i - C - Dt_i$.*

Fitting points by a straight line is so important that we give the two equations $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$, once and for all. The two columns of $A$ are independent (unless all times $t_i$ are the same). So we turn to least squares and solve $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$.

$$\textbf{Dot-product matrix} \quad A^{\mathrm{T}}A = \begin{bmatrix} 1 & \cdots & 1 \\ t_1 & \cdots & t_m \end{bmatrix} \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} = \begin{bmatrix} m & \sum t_i \\ \sum t_i & \sum t_i^2 \end{bmatrix}. \tag{6}$$

On the right side of the normal equation is the 2 by 1 vector $A^{\mathrm{T}}\boldsymbol{b}$:

$$A^{\mathrm{T}}\boldsymbol{b} = \begin{bmatrix} 1 & \cdots & 1 \\ t_1 & \cdots & t_m \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} \sum b_i \\ \sum t_i b_i \end{bmatrix}. \tag{7}$$

In a specific problem, these numbers are given. The best $\widehat{\boldsymbol{x}} = (C, D)$ is in equation (9).

The line $C + Dt$ minimizes $e_1^2 + \cdots + e_m^2 = \|A\boldsymbol{x} - \boldsymbol{b}\|^2$ when $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$:

$$\begin{bmatrix} m & \sum t_i \\ \sum t_i & \sum t_i^2 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} \sum b_i \\ \sum t_i b_i \end{bmatrix}. \tag{8}$$

The vertical errors at the $m$ points on the line are the components of $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p}$. This error vector (the *residual*) $\boldsymbol{b} - A\widehat{\boldsymbol{x}}$ is perpendicular to the columns of $A$ (geometry). The error is in the nullspace of $A^{\mathrm{T}}$ (linear algebra). The best $\widehat{\boldsymbol{x}} = (C, D)$ minimizes the total error $E$, the sum of squares:

$$E(\boldsymbol{x}) = \|A\boldsymbol{x} - \boldsymbol{b}\|^2 = (C + Dt_1 - b_1)^2 + \cdots + (C + Dt_m - b_m)^2.$$

When calculus sets the derivatives $\partial E/\partial C$ and $\partial E/\partial D$ to zero, it produces $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$.

Other least squares problems have more than two unknowns. Fitting by the best parabola has $n = 3$ coefficients $C, D, E$ (see below). In general we are fitting $m$ data points by $n$ parameters $x_1, \ldots, x_n$. The matrix $A$ has $n$ columns and $n < m$. The derivatives of $\|A\boldsymbol{x} - \boldsymbol{b}\|^2$ give the $n$ equations $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$. **The derivative of a square is linear.** This is why the method of least squares is so popular.

**Example 2** $A$ has *orthogonal columns* when the measurement times $t_i$ add to zero.

Suppose $b = 1, 2, 4$ at times $t = -2, 0, 2$. Those times add to zero. The columns of $A$ have *zero dot product*:

$$\begin{aligned} C + D(-2) &= 1 \\ C + \phantom{-}D(0) &= 2 \\ C + \phantom{-}D(2) &= 4 \end{aligned} \qquad \text{or} \qquad A\boldsymbol{x} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}.$$

Look at the zeros in $A^{\mathrm{T}}A$:

$$A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b} \qquad \text{is} \qquad \begin{bmatrix} 3 & 0 \\ 0 & 8 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 7 \\ 6 \end{bmatrix}.$$

*Main point*: Now $A^{\mathrm{T}}A$ *is diagonal*. We can solve separately for $C = \frac{7}{3}$ and $D = \frac{6}{8}$. The zeros in $A^{\mathrm{T}}A$ are dot products of perpendicular columns in $A$. The diagonal matrix $A^{\mathrm{T}}A$, with entries $m = 3$ and $t_1^2 + t_2^2 + t_3^2 = 8$, is virtually as good as the identity matrix.

Orthogonal columns are so helpful that it is worth moving the time origin to produce them. To do that, subtract away the average time $\widehat{t} = (t_1 + \cdots + t_m)/m$. The shifted times $T_i = t_i - \widehat{t}$ add to $\sum T_i = m\widehat{t} - m\widehat{t} = 0$. With the columns now orthogonal, $A^{\mathrm{T}}A$ is diagonal. Its entries are $m$ and $T_1^2 + \cdots + T_m^2$. The best $C$ and $D$ have direct formulas:

$$\boldsymbol{T} \textbf{ is } \boldsymbol{t - \widehat{t}} \qquad C = \frac{b_1 + \cdots + b_m}{m} \qquad \text{and} \qquad D = \frac{b_1T_1 + \cdots + b_mT_m}{T_1^2 + \cdots + T_m^2}. \tag{9}$$

***The best line is*** $C + DT$ ***or*** $C + D(t - \widehat{t})$. The time shift that makes $A^{\mathrm{T}}A$ diagonal is an example of the Gram-Schmidt process: *orthogonalize the columns in advance*.

## Fitting by a Parabola

If we throw a ball, it would be crazy to fit the path by a straight line. A parabola $b = C + Dt + Et^2$ allows the ball to go up and come down again ($b$ is the height at time $t$). The actual path is not a perfect parabola, but the whole theory of projectiles starts with that approximation.

When Galileo dropped a stone from the Leaning Tower of Pisa, it accelerated. The distance contains a quadratic term $\frac{1}{2}gt^2$. (Galileo's point was that the stone's mass is not involved.) Without that $t^2$ term we could never send a satellite into the right orbit. But even with a nonlinear function like $t^2$, the unknowns $C, D, E$ appear linearly! Choosing the best parabola is still a problem in linear algebra.

**Problem** Fit heights $b_1, \ldots, b_m$ at times $t_1, \ldots, t_m$ by a parabola $C + Dt + Et^2$.

**Solution** With $m > 3$ points, the $m$ equations for an exact fit are generally unsolvable:

$$\begin{matrix} C + Dt_1 + Et_1^2 = b_1 \\ \vdots \\ C + Dt_m + Et_m^2 = b_m \end{matrix} \qquad \text{has the } m \text{ by 3 matrix} \qquad A = \begin{bmatrix} 1 & t_1 & t_1^2 \\ \vdots & \vdots & \vdots \\ 1 & t_m & t_m^2 \end{bmatrix}. \tag{10}$$

**Least squares** The closest parabola $C + Dt + Et^2$ chooses $\widehat{\boldsymbol{x}} = (C, D, E)$ to satisfy the three normal equations $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$.

May I ask you to convert this to a problem of projection? The column space of $A$ has dimension _____. The projection of $\boldsymbol{b}$ is $\boldsymbol{p} = A\widehat{\boldsymbol{x}}$, which combines the three columns using the coefficients $C, D, E$. The error at the first data point is $e_1 = b_1 - C - Dt_1 - Et_1^2$. The total squared error is $e_1^2 +$ _____. If you prefer to minimize by calculus, take the partial derivatives of $E$ with respect to _____, _____, _____. These three derivatives will be zero when $\widehat{\boldsymbol{x}} = (C, D, E)$ solves the 3 by 3 system of equations _____.

Section 8.5 has more least squares applications. The big one is Fourier series—approximating functions instead of vectors. The function to be minimized changes from a sum of squared errors $e_1^2 + \cdots + e_m^2$ to an integral of the squared error.

**Example 3** For a parabola $b = C + Dt + Et^2$ to go through the three heights $b = 6, 0, 0$ when $t = 0, 1, 2$, the equations are

$$\begin{aligned} C + D \cdot 0 + E \cdot 0^2 &= 6 \\ C + D \cdot 1 + E \cdot 1^2 &= 0 \\ C + D \cdot 2 + E \cdot 2^2 &= 0. \end{aligned} \tag{11}$$

This is $A\boldsymbol{x} = \boldsymbol{b}$. We can solve it exactly. Three data points give three equations and a square matrix. The solution is $\boldsymbol{x} = (C, D, E) = (6, -9, 3)$. The parabola through the three points in Figure 4.8a is $b = 6 - 9t + 3t^2$.

What does this mean for projection? The matrix has three columns, which span the whole space $\mathbf{R}^3$. The projection matrix is the identity. The projection of $\boldsymbol{b}$ is $\boldsymbol{b}$. The error is zero. We didn't need $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$, because we solved $A\boldsymbol{x} = \boldsymbol{b}$. Of course we could multiply by $A^{\mathrm{T}}$, but there is no reason to do it.

Figure 4.8 also shows a fourth point $b_4$ at time $t_4$. If that falls on the parabola, the new $A\boldsymbol{x} = \boldsymbol{b}$ (four equations) is still solvable. When the fourth point is not on the parabola, we turn to $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$. Will the least squares parabola stay the same, with all the error at the fourth point? Not likely!

The smallest error vector $(e_1, e_2, e_3, e_4)$ is perpendicular to $(1, 1, 1, 1)$, the first column of $A$. Least squares balances out the four errors, and they add to zero.

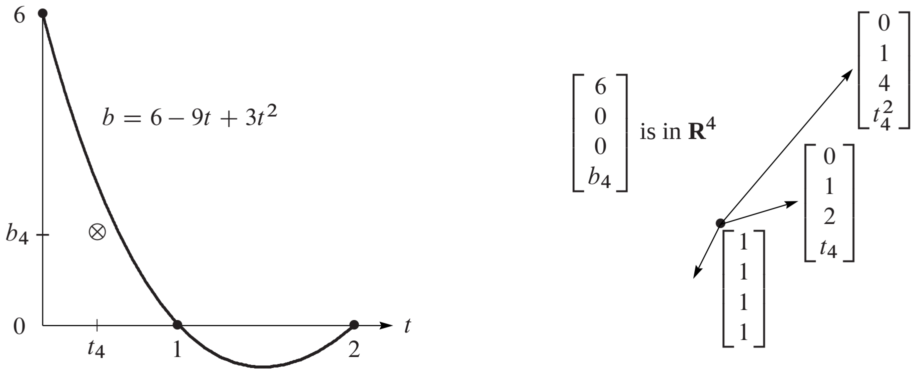

6
$b = 6 - 9t + 3t^2$
$b_4$
0
$t_4$
1
2
$t$
$\begin{bmatrix} 6 \\ 0 \\ 0 \\ b_4 \end{bmatrix}$ is in $\mathbf{R}^4$
$\begin{bmatrix} 0 \\ 1 \\ 4 \\ t_4^2 \end{bmatrix}$
$\begin{bmatrix} 0 \\ 1 \\ 2 \\ t_4 \end{bmatrix}$
$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

Figure 4.8: From Example 3: An exact fit of the parabola at $t = 0, 1, 2$ means that $\boldsymbol{p} = \boldsymbol{b}$ and $\boldsymbol{e} = \boldsymbol{0}$. The point $b_4$ off the parabola makes $m > n$ and we need least squares.

## ■ REVIEW OF THE KEY IDEAS ■

1. The least squares solution $\widehat{x}$ minimizes $E = \|Ax - \boldsymbol{b}\|^2$. This is the sum of squares of the errors in the $m$ equations $(m > n)$.

2. The best $\widehat{x}$ comes from the normal equations $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}\boldsymbol{b}$.

3. To fit $m$ points by a line $b = C + Dt$, the normal equations give $C$ and $D$.

4. The heights of the best line are $\boldsymbol{p} = (p_1, \ldots, p_m)$. The vertical distances to the data points are the errors $\boldsymbol{e} = (e_1, \ldots, e_m)$.

5. If we try to fit $m$ points by a combination of $n < m$ functions, the $m$ equations $A\boldsymbol{x} = \boldsymbol{b}$ are generally unsolvable. The $n$ equations $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}\boldsymbol{b}$ give the least squares solution—the combination with smallest MSE (mean square error).

## ■ WORKED EXAMPLES ■

**4.3 A** Start with nine measurements $b_1$ to $b_9$, *all zero*, at times $t = 1, \ldots, 9$. The tenth measurement $b_{10} = 40$ is an outlier. Find the best *horizontal line* $y = C$ to fit the ten points $(1,0), (2,0), \ldots, (9,0), (10,40)$ using three measures for the error $E$:

**(1)** Least squares $E_2 = e_1^2 + \cdots + e_{10}^2$ (then the normal equation for $C$ is linear)

**(2)** Least maximum error $E_\infty = |e_{\max}|$ **(3)** Least sum of errors $E_1 = |e_1| + \cdots + |e_{10}|$.

**Solution** **(1)** The least squares fit to $0, 0, \ldots, 0, 40$ by a horizontal line is $C = 4$:

$$A = \text{column of 1's} \quad A^{\mathrm{T}}A = 10 \quad A^{\mathrm{T}}\boldsymbol{b} = \text{sum of } b_i = 40. \quad \text{So } 10C = 40.$$

**(2)** The least maximum error requires $C = 20$, halfway between 0 and 40. $= 40 + 8C$ for $C \geqslant 0$

**(3)** The least sum requires $C = 0$ (!!). The sum of errors $9|C| + |40 - C|$ would increase if $C$ moves up from zero.

The least sum comes from the *median* measurement (the median of $0, \ldots, 0, 40$ is zero). Many statisticians feel that the least squares solution is too heavily influenced by outliers like $b_{10} = 40$, and they prefer least sum. But the equations become nonlinear.

Now find the least squares straight line $C + Dt$ through those ten points.

$$A^{\mathrm{T}}A = \begin{bmatrix} 10 & \sum t_i \\ \sum t_i & \sum t_i^2 \end{bmatrix} = \begin{bmatrix} 10 & 55 \\ 55 & 385 \end{bmatrix} \qquad A^{\mathrm{T}}\boldsymbol{b} = \begin{bmatrix} \sum b_i \\ \sum t_i b_i \end{bmatrix} = \begin{bmatrix} 40 \\ 400 \end{bmatrix}$$

Those come from equation (8). Then $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}\boldsymbol{b}$ gives $C = -8$ and $D = 24/11$.

What happens to $C$ and $D$ if you multiply the $b_i$ by 3 and then add 30 to get $\boldsymbol{b}_{\text{new}} = (30, 30, \ldots, 150)$? Linearity allows us to rescale $\boldsymbol{b} = (0, 0, \ldots, 40)$. Multiplying $\boldsymbol{b}$ by 3 will multiply $C$ and $D$ by 3. Adding 30 to all $b_i$ will add 30 to $C$.

**4.3 B** Find the parabola $C+Dt+Et^2$ that comes closest (least squares error) to the values $\boldsymbol{b} = (0,0,1,0,0)$ at the times $t = -2,-1,0,1,2$. First write down the five equations $A\boldsymbol{x} = \boldsymbol{b}$ in three unknowns $\boldsymbol{x} = (C,D,E)$ for a parabola to go through the five points. No solution because no such parabola exists. Solve $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$.

I would predict $D = 0$. Why should the best parabola be symmetric around $t = 0$? In $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$, equation 2 for $D$ should uncouple from equations 1 and 3.

**Solution** The five equations $A\boldsymbol{x} = \boldsymbol{b}$ have a rectangular "Vandermonde" matrix $A$:

$$\begin{array}{lllllll} C & + & D\,(-2) & + & E\,(-2)^2 & = & 0 \\ C & + & D\,(-1) & + & E\,(-1)^2 & = & 0 \\ C & + & D\quad (0) & + & E\quad (0)^2 & = & 1 \\ C & + & D\quad (1) & + & E\quad (1)^2 & = & 0 \\ C & + & D\quad (2) & + & E\quad (2)^2 & = & 0 \end{array} \qquad A = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \qquad A^{\mathrm{T}}A = \begin{bmatrix} 5 & \mathbf{0} & 10 \\ \mathbf{0} & 10 & \mathbf{0} \\ 10 & \mathbf{0} & 34 \end{bmatrix}$$

Those zeros in $A^{\mathrm{T}}A$ mean that column 2 of $A$ is orthogonal to columns 1 and 3. We see this directly in $A$ (the times $-2,-1,0,1,2$ are symmetric). The best $C, D, E$ in the parabola $C + Dt + Et^2$ come from $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$, and $D$ is uncoupled:

$$\begin{bmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix} \begin{bmatrix} C \\ D \\ E \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{leads to} \quad \begin{array}{l} C = 34/70 \\ D = 0 \;\text{ as predicted} \\ E = -10/70 \end{array}$$

## Problem Set 4.3

**Problems 1–11 use four data points $\boldsymbol{b} = (0,8,8,20)$ to bring out the key ideas.**

✓ **1** With $b = 0,8,8,20$ at $t = 0,1,3,4$, set up and solve the normal equations $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$. For the best straight line in Figure 4.9a, find its four heights $p_i$ and four errors $e_i$. What is the minimum value $E = e_1^2 + e_2^2 + e_3^2 + e_4^2$?

✓ **2** (Line $C + Dt$ does go through $p$'s) With $b = 0,8,8,20$ at times $t = 0,1,3,4$, write down the four equations $A\boldsymbol{x} = \boldsymbol{b}$ (unsolvable). Change the measurements to $p = 1,5,13,17$ and find an exact solution to $A\widehat{\boldsymbol{x}} = \boldsymbol{p}$.

✓ **3** Check that $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p} = (-1,3,-5,3)$ is perpendicular to both columns of $A$. What is the shortest distance $\|\boldsymbol{e}\|$ from $\boldsymbol{b}$ to the column space of $A$?

✓ **4** (By calculus) Write down $E = \|A\boldsymbol{x} - \boldsymbol{b}\|^2$ as a sum of four squares—the last one is $(C + 4D - 20)^2$. Find the derivative equations $\partial E/\partial C = 0$ and $\partial E/\partial D = 0$. Divide by 2 to obtain the normal equations $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$.

✓ **5** Find the height $C$ of the best *horizontal line* to fit $\boldsymbol{b} = (0,8,8,20)$. An exact fit would solve the unsolvable equations $C = 0$, $C = 8$, $C = 8$, $C = 20$. Find the 4 by 1 matrix $A$ in these equations and solve $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}\boldsymbol{b}$. Draw the horizontal line at height $\widehat{x} = C$ and the four errors in $\boldsymbol{e}$.

**6** Project $\boldsymbol{b} = (0, 8, 8, 20)$ onto the line through $\boldsymbol{a} = (1, 1, 1, 1)$. Find $\widehat{x} = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}/\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}$ and the projection $\boldsymbol{p} = \widehat{x}\boldsymbol{a}$. Check that $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p}$ is perpendicular to $\boldsymbol{a}$, and find the shortest distance $\|\boldsymbol{e}\|$ from $\boldsymbol{b}$ to the line through $\boldsymbol{a}$.

**7** Find the closest line $b = Dt$, *through the origin*, to the same four points. An exact fit would solve $D \cdot 0 = 0$, $D \cdot 1 = 8$, $D \cdot 3 = 8$, $D \cdot 4 = 20$. Find the 4 by 1 matrix and solve $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}\boldsymbol{b}$. Redraw Figure 4.9a showing the best line $b = Dt$ and the $e$'s.

**8** Project $\boldsymbol{b} = (0, 8, 8, 20)$ onto the line through $\boldsymbol{a} = (0, 1, 3, 4)$. Find $\widehat{x} = D$ and $\boldsymbol{p} = \widehat{x}\boldsymbol{a}$. The best $C$ in Problems 5–6 and the best $D$ in Problems 7–8 do *not* agree with the best $(C, D)$ in Problems 1–4. That is because $(1, 1, 1, 1)$ and $(0, 1, 3, 4)$ are _____ perpendicular.

**9** For the closest parabola $b = C + Dt + Et^2$ to the same four points, write down the unsolvable equations $A\boldsymbol{x} = \boldsymbol{b}$ in three unknowns $\boldsymbol{x} = (C, D, E)$. Set up the three normal equations $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$ (solution not required). In Figure 4.9a you are now fitting a parabola to 4 points—what is happening in Figure 4.9b?

**10** For the closest cubic $b = C + Dt + Et^2 + Ft^3$ to the same four points, write down the four equations $A\boldsymbol{x} = \boldsymbol{b}$. Solve them by elimination. In Figure 4.9a this cubic now goes exactly through the points. What are $\boldsymbol{p}$ and $\boldsymbol{e}$?

**11** The average of the four times is $\widehat{t} = \frac{1}{4}(0 + 1 + 3 + 4) = 2$. The average of the four $b$'s is $\widehat{b} = \frac{1}{4}(0 + 8 + 8 + 20) = 9$.

(a) Verify that the best line goes through the center point $(\widehat{t}, \widehat{b}) = (2, 9)$.

(b) Explain why $C + D\widehat{t} = \widehat{b}$ comes from the first equation in $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$.

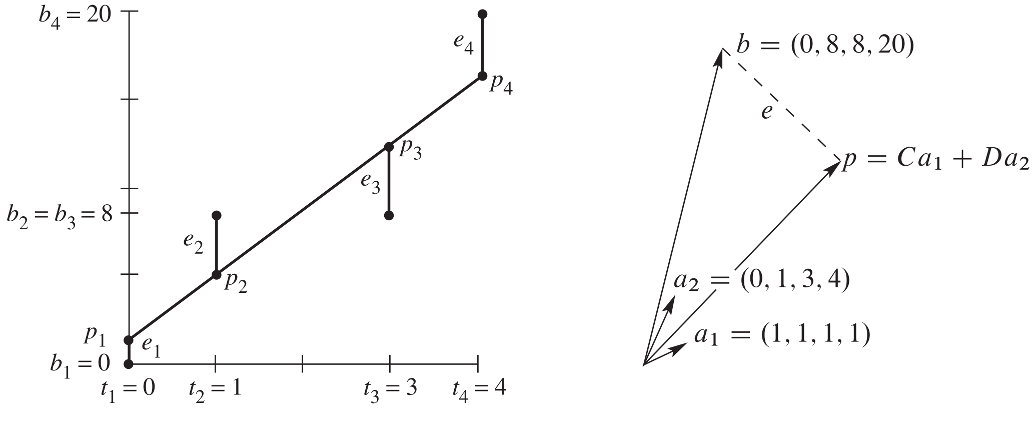


Figure 4.9: **Problems 1–11**: The closest line $C + Dt$ matches $C\boldsymbol{a}_1 + D\boldsymbol{a}_2$ in $\mathbf{R}^4$.

**Questions 12–16 introduce basic ideas of statistics—the foundation for least squares.**

**12** (Recommended) This problem projects $\boldsymbol{b} = (b_1, \ldots, b_m)$ onto the line through $\boldsymbol{a} = (1, \ldots, 1)$. We solve $m$ equations $\boldsymbol{a}x = \boldsymbol{b}$ in 1 unknown (by least squares).

(a) Solve $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}\widehat{x} = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}$ to show that $\widehat{x}$ is the *mean* (the average) of the $\boldsymbol{b}$'s.

(b) Find $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{a}\widehat{x}$ and the *variance* $\|\boldsymbol{e}\|^2$ and the *standard deviation* $\|\boldsymbol{e}\|$.

(c) The horizontal line $\widehat{\boldsymbol{b}} = 3$ is closest to $\boldsymbol{b} = (1, 2, 6)$. Check that $\boldsymbol{p} = (3, 3, 3)$ is perpendicular to $\boldsymbol{e}$ and find the 3 by 3 projection matrix $P$.

**13** First assumption behind least squares: $A\boldsymbol{x} = \boldsymbol{b} -$ ***(noise e with mean zero)***. Multiply the error vectors $\boldsymbol{e} = \boldsymbol{b} - A\boldsymbol{x}$ by $(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ to get $\widehat{\boldsymbol{x}} - \boldsymbol{x}$ on the right. The estimation errors $\widehat{\boldsymbol{x}} - \boldsymbol{x}$ also average to zero. The estimate $\widehat{\boldsymbol{x}}$ is *unbiased*.

**14** Second assumption behind least squares: The $m$ errors $\boldsymbol{e}_i$ are independent with variance $\sigma^2$, so the average of $(\boldsymbol{b} - A\boldsymbol{x})(\boldsymbol{b} - A\boldsymbol{x})^{\mathrm{T}}$ is $\sigma^2 I$. Multiply on the left by $(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ and on the right by $A(A^{\mathrm{T}}A)^{-1}$ to show that the average matrix $(\widehat{\boldsymbol{x}} - \boldsymbol{x})(\widehat{\boldsymbol{x}} - \boldsymbol{x})^{\mathrm{T}}$ is $\sigma^2(A^{\mathrm{T}}A)^{-1}$. This is the ***covariance matrix*** $P$ in section 8.6.

**15** A doctor takes 4 readings of your heart rate. The best solution to $x = b_1, \ldots, x = b_4$ is the average $\widehat{x}$ of $b_1, \ldots, b_4$. The matrix $A$ is a column of 1's. Problem 14 gives the expected error $(\widehat{x} - x)^2$ as $\sigma^2(A^{\mathrm{T}}A)^{-1} =$ _____. ***By averaging, the variance drops from*** $\sigma^2$ ***to*** $\sigma^2/4$.

**16** If you know the average $\widehat{x}_9$ of 9 numbers $b_1, \ldots, b_9$, how can you quickly find the average $\widehat{x}_{10}$ with one more number $b_{10}$? The idea of *recursive* least squares is to avoid adding 10 numbers. What number multiplies $\widehat{x}_9$ in computing $\widehat{x}_{10}$?

$$\widehat{x}_{10} = \tfrac{1}{10}b_{10} + \underline{\qquad}\,\widehat{x}_9 = \tfrac{1}{10}(b_1 + \cdots + b_{10}) \quad \text{as in Worked Example 4.2 C.}$$

**Questions 17–24 give more practice with $\widehat{\boldsymbol{x}}$ and $\boldsymbol{p}$ and $\boldsymbol{e}$.**

**17** Write down three equations for the line $b = C + Dt$ to go through $b = 7$ at $t = -1$, $b = 7$ at $t = 1$, and $b = 21$ at $t = 2$. Find the least squares solution $\widehat{\boldsymbol{x}} = (C, D)$ and draw the closest line.

**18** Find the projection $\boldsymbol{p} = A\widehat{\boldsymbol{x}}$ in Problem 17. This gives the three heights of the closest line. Show that the error vector is $\boldsymbol{e} = (2, -6, 4)$. Why is $P\boldsymbol{e} = \boldsymbol{0}$?

**19** Suppose the measurements at $t = -1, 1, 2$ are the errors $2, -6, 4$ in Problem 18. Compute $\widehat{\boldsymbol{x}}$ and the closest line to these new measurements. Explain the answer: $\boldsymbol{b} = (2, -6, 4)$ is perpendicular to _____ so the projection is $\boldsymbol{p} = \boldsymbol{0}$.

**20** Suppose the measurements at $t = -1, 1, 2$ are $\boldsymbol{b} = (5, 13, 17)$. Compute $\widehat{\boldsymbol{x}}$ and the closest line and $\boldsymbol{e}$. The error is $\boldsymbol{e} = \boldsymbol{0}$ because this $\boldsymbol{b}$ is _____.

**21** Which of the four subspaces contains the error vector $\boldsymbol{e}$? Which contains $\boldsymbol{p}$? Which contains $\widehat{\boldsymbol{x}}$? What is the nullspace of $A$?

**22** Find the best line $C + Dt$ to fit $b = 4, 2, -1, 0, 0$ at times $t = -2, -1, 0, 1, 2$.

**23** Is the error vector $\boldsymbol{e}$ orthogonal to $\boldsymbol{b}$ or $\boldsymbol{p}$ or $\boldsymbol{e}$ or $\widehat{\boldsymbol{x}}$? Show that $\|\boldsymbol{e}\|^2$ equals $\boldsymbol{e}^{\mathrm{T}}\boldsymbol{b}$ which equals $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b} - \boldsymbol{p}^{\mathrm{T}}\boldsymbol{b}$. This is the smallest total error $E$.

**24** The partial derivatives of $\|A\boldsymbol{x}\|^2$ with respect to $x_1, \ldots, x_n$ fill the vector $2A^{\mathrm{T}}A\boldsymbol{x}$. The derivatives of $2\boldsymbol{b}^{\mathrm{T}}A\boldsymbol{x}$ fill the vector $2A^{\mathrm{T}}\boldsymbol{b}$. So the derivatives of $\|A\boldsymbol{x} - \boldsymbol{b}\|^2$ are zero when _____.

## Challenge Problems

**25** *What condition on* $(t_1, b_1), (t_2, b_2), (t_3, b_3)$ *puts those three points onto a straight line*? A column space answer is: $(b_1, b_2, b_3)$ must be a combination of $(1,\ 1,\ 1)$ and $(t_1,\ t_2,\ t_3)$. Try to reach a specific equation connecting the $t$'s and $b$'s. I should have thought of this question sooner!

**26** Find the *plane* that gives the best fit to the 4 values $\boldsymbol{b} = (0, 1, 3, 4)$ at the corners $(1, 0)$ and $(0, 1)$ and $(-1, 0)$ and $(0, -1)$ of a square. The equations $C + Dx + Ey = b$ at those 4 points are $A\boldsymbol{x} = \boldsymbol{b}$ with 3 unknowns $\boldsymbol{x} = (C, D, E)$. What is $A$? At the center $(0, 0)$ of the square, show that $C + Dx + Ey =$ average of the $b$'s.

**27** (Distance between lines) The points $P = (x, x, x)$ and $Q = (y, 3y, -1)$ are on two lines in space that don't meet. Choose $x$ and $y$ to minimize the squared distance $\|P - Q\|^2$. The line connecting the closest $P$ and $Q$ is perpendicular to _____.

**28** Suppose the columns of $A$ are not independent. How could you find a matrix $B$ so that $P = B(B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}$ does give the projection onto the column space of $A$? (The usual formula will fail when $A^{\mathrm{T}}A$ is not ivertible.)

**29** Usually there will be exactly one hyperplane in $\mathbf{R}^n$ that contains the $n$ given points $\boldsymbol{x} = \mathbf{0}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{n-1}$. (Example for $n = 3$: There will be one plane containing $0, \boldsymbol{a}_1, \boldsymbol{a}_2$ unless _____.) What is the test to have exactly one plane in $\mathbf{R}^n$?