Vectors — and an Application to Least Squares

This brief review of vectors assumes you have seen the basic properties of vectors previously.

We can write a point in $\mathbb{R}^n$ as $X = (x_1, \dots, x_n)$. This point is often called a **vector**. Frequently it is useful to think of it as an arrow pointing from the origin to the point. Thus, in the plane $\mathbb{R}^2$, $X = (1, -2)$ can be thought of as an arrow from the origin to the point $(1, -2)$.

Algebraic Properties

Alg-1. ADDITION: If $Y = (y_1, \dots, y_n)$, then $X + Y = (x_1 + y_1, \dots, x_n + y_n)$.
Example: In $\mathbb{R}^4$, $(1, 2, -2, 0) + (-1, 2, 3, 4) = (0, 4, 1, 4)$.

Alg-2. MULTIPLICATION BY A CONSTANT: $cX = (cx_1, \dots, cx_n)$.
Example: In $\mathbb{R}^4$, if $X = (1, 2, -2, 0)$, then $-3X = (-3, -6, 6, 0)$.

Alg-3. DISTRIBUTIVE PROPERTY: $c(X + Y) = cX + cY$. This is obvious if one writes it out using components. For instance, in $\mathbb{R}^2$:

$$c(X + Y) = c(x_1 + y_1, x_2 + y_2) = (cx_1 + cy_1, cx_2 + cy_2) = (cx_1, cx_2) + (cy_1, cy_2) = cX + cY.$$

Length and Inner Product

NIP-1. $\|X\| := \sqrt{x_1^2 + \cdots + x_n^2}$ is the *distance* from X to the origin. We will also refer to $\|X\|$ as the *length* or *norm* of X. Similarly $\|X - Y\|$ is the *distance between X and Y*. Note that $\|X\| = 0$ if and only if $X = 0$, and also that for any constant c we have $\|CB\| = |c|\|X\|$. Thus, $\|-2X\| = \|2X\| = 2\|X\|$.

LIP-2. The *inner product* of vectors X and Y in $\mathbb{R}^n$ is, by definition,

$$\langle X, Y \rangle := x_1 y_1 + x_2 y_2 + \cdots + x_n y_n. \tag{1}$$

This is also called the *dot product* and written $X \cdot Y$. The inner product of two vectors is a number, *not* another vector. In particular, we have the vital identity $\|X\|^2 = \langle X, X \rangle$ relating the inner product and norm. For added clarity, it is sometimes useful to write the inner product in $\mathbb{R}^n$ as $\langle X, Y \rangle_{\mathbb{R}^n}$.
Example: In $\mathbb{R}^4$, if $X = (1, 2, -2, 0)$ and $Y = (-1, 2, 3, 4)$, then $\langle X, Y \rangle = (1)(-1) + (2)(2) + (-2)(3) + (0)(4) = -3$.

HIP-3. ALGEBRAIC PROPERTIES OF THE INNER PRODUCT. The following are obvious from the above definition of $\langle X, Y \rangle$:

i). $\langle X, X \rangle \ge 0$, with $\langle X, X \rangle = 0$ if (and only if) $X = 0$,
ii). $\langle X + Y, W \rangle = \langle X, W \rangle + \langle Y, W \rangle$,

iii). $\langle cX, Y\rangle = c\langle X, Y\rangle$,
iv). $\langle Y, X\rangle = \langle X, Y\rangle$.

These four properties can be viewed as the *axioms* for an inner product of real vectors.

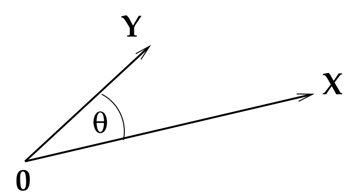
REMARK: If one works with vectors $Z := (z_1, z_2, \ldots, z_n)$, having *complex numbers* z_j as elements, then the definition of the inner product must be modified since, for a complex number $z := x + iy$ we have $|z|^2 = x^2 + y^2 = z\bar{z}$, where $\bar{z} := x - iy$ is the *complex conjugate* of z. Using this we define the *Hermitian inner product* by

$$\langle W, Z\rangle := w_1\bar{z}_1 + w_2\bar{z}_2 + \cdots + w_n\bar{z}_n. \tag{2}$$

(note: many people put the complex conjugate on the first term, w_j, instead of the z_j). The purpose is to insure that the fundamental property $\|Z\|^2 = \langle Z, Z\rangle \geq 0$ still holds. Note, however, that the symmetry property $\langle Y, X\rangle = \langle X, Y\rangle$ is now *replaced* by $\langle Z, W\rangle = \overline{\langle W, Z\rangle}$, and hence, as the following proof shows, $\langle W, cZ\rangle = \bar{c}\langle W, Z\rangle$:
PROOF: $\langle W, cZ\rangle = \overline{\langle cZ, W\rangle} = \langle \bar{c}\bar{Z}, \overline{W}\rangle = \bar{c}\overline{\langle Z, W\rangle} = \bar{c}\langle W, Z\rangle)$.
For complex vectors or matrices one *always* uses a Hermitian inner prodect.

IP-4. GEOMETRIC INTERPRETATION: The definition (1) of the inner product is easy to compute. However, it is not at all obvious that the inner product is useful – until one interprets it geometrically:

$$\langle X, Y\rangle = \|X\|\|Y\|\cos\theta, \tag{3}$$



where θ is the angle between X and Y. Since $\cos(-\theta) = \cos\theta$, the sense in which we measure the angle does not matter.

To prove (3), we can restrict our attention to the two dimensional plane containing X and Y. Thus, we need consider only vectors in $\mathbb{R}^2$. Assume we are not in the trivial case where X or Y are zero. Let α and β be the angles that $X = (x_1, x_2)$ and $Y = (y_1, y_2)$ make with the horizontal axis, so $\theta = \beta - \alpha$. Then

$$x_1 = \|X\|\cos\alpha \qquad \text{and} \qquad x_2 = \|Y\|\sin\alpha.$$

Similarly, $y_1 = \|Y\|\cos\beta$ and $y_2 = \|Y\|\sin\beta$. Therefore

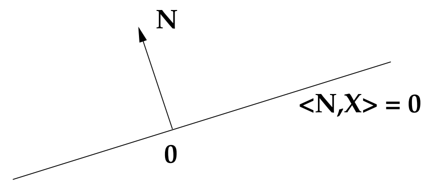
$$\begin{aligned}\langle X, Y\rangle &= x_1y_1 + x_2y_2 = \|X\|\|Y\|(\cos\alpha\cos\beta + \sin\alpha\sin\beta)\\ &= \|X\|\|Y\|\cos(\beta - \alpha) = \|X\|\|Y\|\cos\theta.\end{aligned}$$

This is what we wanted. Alternatively, the equivalence of (1) and (3) can be seen as just a restatement of the law of cosines from trigonometry.

IP-5. GEOMETRIC CONSEQUENCE: X and Y are perpendicular if and only if $\langle X, Y\rangle = 0$, since this means the angle θ between them is 90 degrees so $\cos\theta = 0$. We often use the word *orthogonal* as a synonym for *perpendicular*.

Example: The vectors $X=(1,2,4)$ and $(0,-2,1)$ are orthogonal, since $\langle X,Y\rangle = 0-4+4=0$.

Example: The straight line $-x+3y=0$ through the origin can be written as $\langle N,X\rangle = 0$, where $N=(-1,3)$ and $X=(x,y)$ is a point on the line. Thus we can interpret this line as being the points perpendicular to the vector N. The line $-x+3y=7$ is parallel to the line $-x+3y=0$, except that it does not pass through the origin. This same vector N is perpendicular to it. If X_0 is a point on the line $\langle N,X\rangle = c$, so $\langle N,X_0\rangle = c$, then we can rewrite its equation as $\langle N, X-X_0\rangle = 0$, showing analytically that N is perpendicular to $X-X_0$.



Many formulas involving $\|X\|$ are simplest if one rewrites them immediately in terms of the inner product. The following example uses this approach.

Example: [PYTHAGOREAN THEOREM] If X and Y are orthogonal vectors, then the Pythagorean law holds:

$$\|X+Y\|^2 = \|X\|^2+\|Y\|^2.$$

Since X and Y are orthogonal, then $\langle X,Y\rangle = \langle Y,X\rangle = 0$, so, as asserted

$$\begin{aligned}\|X+Y\|^2 &= \langle X+Y, X+Y\rangle \\ &= \langle X,X\rangle + \langle X,Y\rangle + \langle Y,X\rangle + \langle Y,Y\rangle \\ &= \|X\|^2+\|Y\|^2.\end{aligned}$$

since if a vector Z is orthogonal to all other vectors, in particular, it is orthogonal to itself. Thus $\|Z\|^2 = \langle Z,Z\rangle = 0$ so $Z=0$.

REMARK: Observe that the zero vector is orthogonal to all vectors. It is the *only* such vector since if $\langle Z,V\rangle = 0$ for *all* vectors V, then $Z=0$. To prove this, since we can pick any vector for V, this is true in particular if $V=Z$. But then $\|Z\|^2=\langle Z,Z\rangle = 0$ so the only possibility is $Z=0$.

IP-6. MATRICES AND THE INNER PRODUCT: If A is a $k\times n$ matrix (k rows, n columns so $A:\mathbb{R}^n\to\mathbb{R}^k$), we want to compute $\langle AX,Y\rangle_{\mathbb{R}^k}$ for vectors $X\in\mathbb{R}^n$ and $Y\in\mathbb{R}^k$ in order to introduce the concept of the *adjoint* of a matrix.

Let $e_1=(1,0,0,\ldots,0)$, ..., $e_n=(0,0,\ldots,0,1)$, be the usual standard basis vectors in $\mathbb{R}^n$ and $\varepsilon_1=(1,0,0,\ldots,0),\ldots,\varepsilon_k:=(0,\ldots,0,1)$ be the usual basis vectors in $\mathbb{R}^k$. Recall that in matrix notation, we usually think of vectors as *column vectors*. If $A=(a_{ij})$, it is easy to see that Ae_1 is the first column of A, Ae_2 the second column of $A$and so on. For instance

$$Ae_2 = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \ldots & a_{kn}\end{pmatrix}\begin{pmatrix}0\\1\\ \vdots\\0\end{pmatrix} = \begin{pmatrix}a_{12}\\a_{22}\\ \vdots \\ a_{k2}\end{pmatrix}. \tag{4}$$

In words, the image of e_2 is the second column of A, just as asserted.
Using this observation it is clear that $\langle Ae_2, \varepsilon_1\rangle_{\mathbb{R}^k} = a_{12}$. Similarly,

$$\langle Ae_i, \varepsilon_j\rangle_{\mathbb{R}^k} = a_{ji}. \tag{5}$$

We use this to define the *adjoint* of the matrix A, written A^*. It is defined by requiring that

$$\langle AX, Y\rangle = \langle X, A^*Y\rangle \quad \text{or, more formally,} \quad \langle AX, Y\rangle_{\mathbb{R}^k} = \langle X, A^*Y\rangle_{\mathbb{R}^n}. \tag{6}$$

for all vectors $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^k$.
The formula (6) looks abstract but is easy to use – although at this stage it is not at all evident that it is useful. For the moment, write $B = A^*$, so (6) says $\langle AX, Y\rangle = \langle X, BY\rangle$. Say the elements of B are b_{ij}. We would like to compute the b_{ij}'s in terms of the known elements a_{ij} of A. From (4) applied to B, we know that $B\varepsilon_1$ is the first column of B. Thus $\langle e_2, B\varepsilon_1\rangle = b_{21}$. But the definition we have $\langle X, BY\rangle = \langle X, Y\rangle$ so

$$b_{21} = \langle e_2, B\varepsilon_1\rangle = \langle Ae_2, \varepsilon_1\rangle = a_{12}.$$

In the same way, $b_{ij} = a_{ji}$ for all $i = 1, 2, \ldots n,\ j = 1, 2, \ldots k$. In other words, the first row of $B = A^*$ is simply the first column of A, etc. Thus we interchange the rows and columns of A to get A^*. For this reason A^* is often called the *transpose* of A and written A^T.

EXAMPLE

$$\text{if} \quad A := \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}, \quad \text{then} \quad A^* = A^T = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{pmatrix}. \tag{7}$$

A square matrix A is called *self adjoint* or *symmetric* if $A = A^*$. It is called *skew-adjoint* or *anti-symmetric* if $A = -A^*$. An obvious property is that $A^{**} = (A^*)^* = A$.
As an example, let's obtain the property $(AB)^* = B^*A^*$. We begin using the definition (6) applied to AB:

$$\langle (AB)^*X, Y\rangle = \langle X, (AB)Y\rangle. \tag{8}$$

But $(AB)Y = A(BY)$ so

$$\langle X, (AB)Y\rangle = \langle X, A(BY)\rangle = \langle A^*X, BY\rangle = \langle B^*(A^*X), Y\rangle = \langle (B^*A^*)X, Y\rangle. \tag{9}$$

Comparing (8) and (9) we find that $(AB)^* = B^*A^*$.

One consequence is that A^*A is a symmetric matrix, even if A is not a square matrix, because $(A^*A)^* = A^*A^{**} = A^*A$. In particular A^*A is a square matrix. Similarly AA^* is a symmetric matrix. For many applications it is useful to notice that $\langle A^*AX, X\rangle = \langle AX, AX\rangle = \|AX\|^2 \geq 0$ for *all* X.

REMARK: If, as is usual, we think of a vector $X := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ as a column vector, then we can treat it as a $1 \times n$ matrix and observe the inner product $\langle X, Y \rangle = X^T Y$, which is often useful. Also $\langle X, AY \rangle = X^T AY$ so computing inner products is now under the umbrella of matrix multiplication. This observation is quite valuable in computations.

Derivatives of Vectors

D-1. If $X(t) = (x_1(t), \dots, x_n(t))$ describes a curve in $\mathbb{R}^n$, then its *derivative* is

$$X'(t) = \frac{dX(t)}{dt} = (x_1'(t), \dots, x_n'(t)).$$

One can think of this as the *velocity vector*. It is tangent to the curve.

Example: If $X(t) = (2\cos t, 2\sin t)$, then this curve is a circle of radius 2, traversed counter-clockwise. Its velocity is $X'(t) = (-2\sin t, 2\cos t)$ and its *speed* $\|X'(t)\| = 2$. For instance, $X'(0) = (0, 2)$ is the tangent vector at $X(0) = (2, 0)$. The curve $Y(t) = (2\cos 3t, 2\sin 3t)$ also describes the motion of a particle around a circle of radius 2, but in this case the speed is $\|Y'(t)\| = 6$

D-2. DERIVATIVE OF THE INNER PRODUCT: If $X(t)$ and $Y(t)$ are two curves, then

$$\frac{d}{dt}\langle X(t), Y(t) \rangle = \langle \frac{dX(t)}{dt}, Y(t) \rangle + \langle X(t), \frac{dY(t)}{dt} \rangle. \tag{10}$$

or, more briefly, $\langle X, Y \rangle' = \langle X', Y \rangle + \langle X, Y' \rangle$.
To prove this one simply uses the rule for the derivative of a product of functions. Thus

$$\begin{aligned} \frac{d}{dt}\langle X(t), Y(t) \rangle &= \frac{d}{dt}(x_1 y_1 + x_2 y_2 + \cdots) \\ &= (x_1' y_1 + x_1 y_1') + (x_2' y_2 + x_2 y_2') + \cdots \\ &= (x_1' y_1 + x_2' y_2 + \cdots) + (x_1 y_1' + x_2 y_2' + \cdots) \\ &= \langle X', Y \rangle + \langle X, Y' \rangle. \end{aligned}$$

Example:

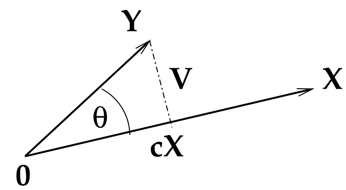
$$\frac{d}{dt}\|X(t)\|^2 = \frac{d}{dt}\langle X(t), X(t) \rangle = 2\langle X(t), X'(t) \rangle. \tag{11}$$

As a special case, if a particle moves at a constant distance c from the origin, $\|X(t)\| = c$, then $0 = dc^2/dt = d\|X(t)\|^2/dt = 2\langle X(t), X'(t) \rangle$. In particular, if a particle moves on a circle or a sphere, then the position vector $X(t)$ is always perpendicular to the velocity $X'(t)$. This also shows that the tangent to a circle, $X'(t)$, is perpendicular to the radius vector, $X(t)$.

Orthogonal Projections

Proj-1. ORTHOGONAL PROJECTION ONTO A LINE: Let X and Y be given vectors. We would like to write Y in the form $Y = cX + V$, where V is perpendicular to X. Then the vector cX is the **orthogonal projection** of Y in the line determined by the vector X.

How can we find the constant c and the vector V? We use the only fact we know: that V is supposed to be perpendicular to X. Thus we take the inner product of $Y = cX + V$ with X and conclude that $\langle X, Y\rangle = c\langle X, X\rangle$, that is



$$c = \frac{\langle X, Y\rangle}{\|X\|^2}.$$

Now that we know c, we can simply define V by the obvious formula $V = Y - cX$.
At first this may seem circular. To convince your self that this works, let $X = (1,1)$, and $Y = (2,3)$. Then compute c and V and draw a sketch showing X, Y, cX, and V.
Since $cX \perp V$, we can use the Pythagorean Theorem to conclude that

$$\|Y\|^2 = c^2\|X\|^2 + \|V\|^2 \geq c^2\|X\|^2.$$

From this, using the explicit value of c found above we conclude that

$$\|Y\|^2 \geq \left(\frac{\langle X, Y\rangle}{\|X\|^2}\right)^2 \|X\|^2.$$

and obtain the *Schwarz inequality*

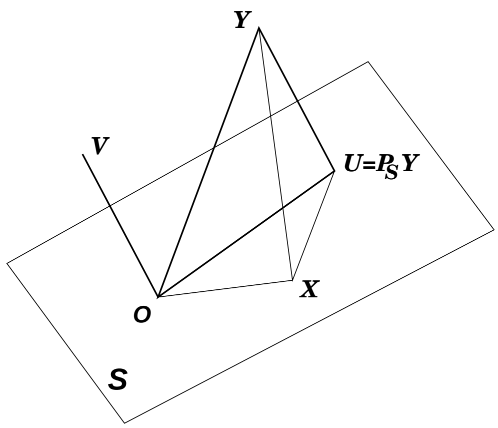
$$|\langle X, Y\rangle| \leq \|X\|\|Y\|. \tag{12}$$

Notice that this was done without trigonometry. It used only the properties of the inner product.

Proj-2. ORTHOGONAL PROJECTION INTO A SUBSPACE. If a linear space has an inner product and S is a subspace of it, we can discuss the orthogonal projection of a vector into that subspace. Given a vector Y, if we can write

$$Y = U + V,$$

where U is in S and V is perpendicular to S, then we call U the projection of Y into S and V the projection of Y perpendicular to S. The notation $U = P_S Y$, $V = P_S^{\perp} Y$ is frequently used for this projection U.



By the Pythagorean theorem

$$\|Y\|^2 = \|U\|^2 + \|V\|^2, \qquad (U = P_S Y,\ V = P_S^{\perp} Y).$$

It is easy to show that *the projection* $P_S Y$ *is closer to* Y *than any other point in* S. In other words,

$$\|Y - P_S Y\| \le \|Y - X\| \qquad \text{for all } X \text{ in } S.$$

To see this, given any $X \in S$ write $Y - X = (Y - P_S Y) + (P_S Y - X)$ and observe that $Y - P_S Y$ is perpendicular to S while $P_S Y$ and X, and hence $P_S Y - X$ are in S. Thus by the Pythagorean Theorem

$$\|Y - X\|^2 = \|Y - P_S Y\|^2 + \|P_S Y - X\|^2 \ge \|Y - P_S Y\|^2.$$

This is what we asserted.

Problems on Vectors

1. a) For which values of the constant a and b are the vectors $U = (1 + a, -2b, 4)$ and $V = (2, 1, -1)$ perpendicular?

   b) For which values of the constant a, and b is the above vector U, perpendicular to both V and the vector $W = (1, 1, 0)$?

2. Let $X = (3, 4, 0)$ and $Y = (1, -, 1)$.

   a) Write the vector Y in the form $Y = cX + V$, where V is orthogonal to X. Thus, you need to find the constant c and the vector V.

   b) Compute $\|X\|$, $\|Y\|$, and $\|V\|$ and verify the Pythagorean relation

$$\|Y\|^2 = \|cX\|^2 + \|V\|^2.$$

3. [CONVERSE OF THE PYTHAGOREAN THEOREM] If X and Y are real vectors with the property that the Pythagorean law holds: $\|X\|^2+\|Y\|^2=\|X+Y\|^2$, then X and Y are orthogonal.

4. If a vector X is written as $X=aU+bV$, where U and V are non-zero orthogonal vectors, show that $a=\langle X,U\rangle/\|U\|^2$ and $b=\langle X,V\rangle/\|V\|^2$.

5. The origin and the vectors X, Y, and $X+Y$ define a parallelogram whose diagonals have length $X+Y$ and $X-Y$. Prove the *parallelogram law*

$$\|X+Y\|^2+\|X-Y\|^2=2\|X\|^2+2\|Y\|^2;$$

   This states that in a parallelogram, the sum of the squares of the lengths of the diagonals equals the sum of the squares of the four sides.

6. a) Find the distance from the point $(2,-1)$ to the straight line $3x-4y=0$.
   b) Find the distance from the straight line $3x-4y=10$ to the origin.
   c) Find the distance from the straight line $ax+by=c$ to the origin.
   d) Find the distance between the parallel lines $ax+by=c$ and $ax+by=\gamma$.
   e) Find the distance from the plane $ax+by+cz=d$ to the origin.

7. The equation of a straight line in $\mathbb{R}^3$ can be written as $X(t)=X_0+tV$, $-\infty<t<\infty$, where X_0 is a point on the line and V is a vector along the line (in a physical setting, V might be the *velocity* vector).
   a) Find the distance from this line to the origin.
   b) If $Y(s)=Y_0+sW$, $-\infty<s<\infty$, is another straight line, find the distance between these straight lines.

8. Let P_1, P_2, …, P_k be points in $\mathbb{R}^n$. For $X\in\mathbb{R}^n$ let

$$Q(X):=\|X-P_1\|^2+\|X-P_2\|^2+\cdots\|X-P_k\|^2.$$

   Determine the point X that minimizes $Q(X)$.

9. a) If X and Y are real vectors, show that

$$\langle X,Y\rangle=\frac{1}{4}\left(\|X+Y\|^2-\|X-Y\|^2\right).$$

   This formula is the simplest way to recover properties of the inner product from the norm.

b) As an application, show that if a square matrix R has the property that it preserves length, so $\|RX\| = \|X\|$ for every vector X, then it preserves the inner product, that is, $\langle RX, RY\rangle = \langle X, Y\rangle$ for all vectors X and Y.

10. If one uses the complex inner product (2), show that the elements A^* are the transpose conjugate, $A^* = (\overline{a}_{\ell k})$, of the elements of $A = (a_{k\ell})$.

11. a) If a certain matrix C satisfies $\langle X, CY\rangle = 0$ for *all* vectors X and Y, show that $C = 0$.

    b) If the matrices A and B satisfy $\langle X, AY\rangle = \langle X, BY\rangle$ for all vectors X and Y, show that $A = B$.

12. a) Give an example of a 3×3 anti-symmetric matrix.

    b) If A is any anti-symmetric matrix, show that $\langle X, AX\rangle = 0$ for all vectors X.

13. Say $X(t)$ is a solution of the differential equation $\dfrac{dX}{dt} = AX$, where A is an *anti-symmetric* matrix. Show that $\|X(t)\| = \text{constant}$.

Application to the Method of Least Squares

THE PROBLEM. Say you have done an experiment and obtained the data points $(-1,1)$, $(0,-1)$, $(1,-1)$, and $(2,3)$. Based on some other evidence you believe this data should fit a curve of the form $y = a + bx^2$. If you substitute your data (x_j, y_j) into this equation you find

$$\begin{aligned} a+b(-1)^2 &= 1 \\ a+b(0)^2 &= -1 \\ a+b(1)^2 &= -1 \\ a+b(2)^2 &= 3 \end{aligned} \tag{13}$$

This system of equations is *over determined* since there are more equations (four) than unknowns (two: a and b). As is the case with almost all overdetermined systems, it is unlikely they can be solved exactly.

We rewrite these equations in the matrix form $AV = W$, where

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 4 \end{pmatrix}, \quad V = \begin{pmatrix} a \\ b \end{pmatrix}, \quad \text{and} \quad W = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 3 \end{pmatrix}$$

We refer to A as the *data matrix* and W as the *observation vector*.

Instead of the probably hopeless task of solving $AV = W$, we instead seek a vector V that minimizes the error (actually, the square of the error).

$$Q(V) := \|AV - W\|^2.$$

If we are fortunate and find an exact solution of $AV = W$, so much the better since then $Q(V) = 0$. We will find this error minimizing solution in two different ways, one using calculus, another using projections.

Summary. The general problem we are facing is:

> **Given:** A data matrix A and an observation vector W,
>
> **To find:** The "best solution" of $AV = W$. For us, "best" means minimizing the error $Q(V) = \|AV - W\|^2$.

SOLUTION USING CALCULUS. One approach is to use calculus to find the minimum by taking the first derivative and setting it to zero. We will do this here only using calculus of one variable (so we won't use partial derivatives, although using these gives an entirely equivalent approach).

Say V (this is what we want to compute) gives the minimum, so $Q(X) \ge Q(V)$ for all X. We pick an arbitrary vector Z and use the special family of vectors $X(t) = V + tZ$. Let

$$f(t) := Q(X(t)) = \|AX(t) - W\|^2.$$

Since $Q(X(t)) \ge Q(V) = Q(X(0))$ we know that $f(t) \ge f(0)$ so f has its minimum at $t = 0$. Thus $f'(0) = 0$. We compute this. From (11)

$$f'(t) = 2\langle AX(t) - W, AX'(t)\rangle = 2\langle AX(t) - W, AZ\rangle.$$

In particular,

$$0 = f'(0) = 2\langle AV - W, AZ\rangle.$$

We use (6) to rewrite this as $\langle A^*(AV - W), Z\rangle = 0$ (historically, this was one of the first places where the adjoint of a matrix was used). But now since Z can be *any* vector, by the REMARK at the end of property **Ip-5** above, we see that the desired V must satisfy

$$A^*(AV - W) = 0,$$

that is,

$$\boxed{A^*AV = A^*W}. \tag{14}$$

These are the desired equations to compute V. As observed above, the matrix A^*A is always a square matrix. The fundamental equation (14) is called the *normal equation*.

Example: We apply this idea to (13). Since

$$A^* = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 4 \end{pmatrix},$$

then

$$A^*A = \begin{pmatrix} 4 & 6 \\ 6 & 18 \end{pmatrix} \qquad \text{and} \qquad A^*W = \begin{pmatrix} 2 \\ 12 \end{pmatrix}.$$

The normal equations $A^*AV = A^*W$ are then

$$\begin{aligned} 4a + \ \ 6b &= 2 \\ 6a + 18b &= 12. \end{aligned}$$

Their solution is $a = -1$, $b = 1$. Thus the desired curve $y = a + bx^2$ that best fits your data points is $y = -1 + x^2$.

SOLUTION USING PROJECTIONS. As above, given a matrix A and a vector W we want V that minimizes the error:

$$Q(V) = \|AV - W\|^2.$$

Thus, we want to pick V so that the vector $U := AV$ is as close as possible to W. Notice that U must be in the image of A. From the discussion of projections (see **Proj-2** above), we want to let U be the orthogonal projection of W into the image of A.

How can we compute this? Notice that $AV - W$ will then be perpendicular to the image of A. In other words, $AV - W$ will be perpendicular to all vectors of the form AZ for any vector Z. Thus by (6) above

$$0 = \langle AZ, AV - W \rangle = \langle Z, A^*(AV - W) \rangle.$$

But now since the right side holds for *all* vectors Z we can apply the REMARK at the end of **Ip-5** above to conclude that

$$A^*AV = A^*W. \tag{15}$$

These again are the **normal equations** for V and are what we sought. Of course they are identical to those obtained above using calculus. Although this may seem abstract, it is easy to compute this explicitly.

Example: Here is a standard example using the normal equations. Say we are given n experimental data points (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$ and want to find the straight line $y = a + bx$ that fits this data best. How should be proceed? Ideally we want to pick the coefficients a and b so that

$$\begin{aligned} a + bx_1 &= y_1 \\ a + bx_2 &= y_3 \\ &\cdots \\ a + bx_n &= y_n. \end{aligned}$$

These are n equations for the two unknowns a, b. If $n > 2$ it is unlikely that we can solve them exactly. We write the above equations in matrix notation as $AV = Y$, that is,

$$AV = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot\cdot \\ y_n \end{pmatrix} = Y.$$

Next we want the normal equations $A^*AV = A^*Y$. Now

$$A^*A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{pmatrix}.$$

The computation of A^*Y is equally straightforward so the normal equations are two equations in two unknowns:

$$\begin{pmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum y_j \\ \sum x_j y_j \end{pmatrix}. \tag{16}$$

These can be solved using high school algebra. The solution is:

$$y - \overline{y} = m(x - \overline{x}), \tag{17}$$

where

$$\overline{x} = \frac{1}{n} \sum_{1 \le j \le n} x_j, \qquad \overline{y} = \frac{1}{n} \sum_{1 \le j \le n} y_j, \quad \text{and} \quad m = \frac{\sum (x_j - \overline{x})(y_j - \overline{y})}{\sum (x_j - \overline{x})^2}.$$

Notice that the straight line (17) passes through $(\overline{x}, \overline{y})$. The equations (16) are particularly simple to solve if $\overline{x} = 0$ and $\overline{y} = 0$. The general case is reduced to this special case by the natural substitution $\hat{x}_j = x_j - \overline{x}, \quad \hat{y}_j = y_j - \overline{y}$. I used this to get (17).

In these and related computations it is useful to introduce the data as vectors:

$$x = (x_1, x_2, \ldots, x_n) \quad \text{and} \quad y = (y_1, y_2, \ldots, y_n)$$

and, in occasionally confusing notation, identify the average $\overline{x}$ with the vector $\overline{x} = (\overline{x}, \ldots, \overline{x})$ having n equal components $\overline{x}$. We also use the "data inner product" and "data norm"

$$\ll x, y \gg = x_1 y_1 + x_2 y_2 + \ldots x_n y_n \qquad |x|^2 = \ll x, x \gg .$$

In statistics, $\ll x - \overline{x}, y - \overline{y} \gg$ is called the *covariance of x and y* and write $\operatorname{Cov}(x, y)$. Using this notation the slope of the above line is $m = \ll x - \overline{x}, y - \overline{y} \gg / |x - \overline{x}|^2$. Of special importance is the *correlation coefficient*

$$r(x, y) = \frac{\ll x - \overline{x}, y - \overline{y} \gg}{|x - \overline{x}||y - \overline{y}|}.$$

This measures how closely the data points (x_j, y_j) fit the straight line. The Schwarz inequality asserts that $|r(x,y)| \le 1$. If $r(x,y) = +1$ the data lies along a straight line with positive slope, while if $r(x,y) = -1$ the data lies along a straight line with negative slope. If $r(x,y) = 0$ the data forms a cloud and does not really seem to lie along any straight line. See most statistics books for a more adequate discussion along with useful examples.

Identical methods can be used to find, for instance, the cubic polynomial $y = a + bx + cx^2 + dx^3$ that best fits some data, or the plane $z = a + bx + cy$ that best fits given data. The technique of least squares is widely used in all area where one has experimental data. The key feature is that the equations be *linear* in the unknown coefficients a, b, etc. However, even if the equations are not linear in the unknown coefficients a, b, etc., frequently one can find an equivalent problem to which the techniques apply. The following example illustrates this.

Example: Say we are given n experimental data points (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) and seek an exponential curve $y = ae^{bx}$ that best fits this data. Ideally we want to pick the coefficients a and b so that

$$\begin{aligned} ae^{bx_1} &= \; y_1 \\ ae^{bx_2} &= \; y_2 \\ &\cdots \\ ae^{bx_n} &= \; y_n. \end{aligned}$$

These are n equations for the two unknowns a, b. However, they are nonlinear in b so the method of least squares does not directly apply. To get around this we take the (natural) logarithm of each of these equations and obtain

$$\begin{aligned} \alpha + bx_1 &= \; \ln y_1 \\ \alpha + bx_2 &= \; \ln y_2 \\ &\cdots \\ \alpha + bx_n &= \; \ln y_n, \end{aligned}$$

where $\alpha = \ln a$. These modified equations are *linear* in the unknowns α and b, so we can apply the method of least squares. After we know α, we can recover a simply from $a = e^{\alpha}$.

REMARK. Say one wants to fit data to the related curve $y = ae^{bx} + c$. I don't know any way to do this using least squares, where one eventually solves a linear system of equations (the normal equations). For this problem it seems that one must solve a *nonlinear* system of equations, which is much more difficult.

Example: This is similar to the previous example. Say we are given n experimental data points (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) and seek a curve of the form $y = \dfrac{ax}{1 + bx^2}$ that best fits this data. Ideally we want to pick the coefficients a and b so that

$$\begin{aligned}
\frac{ax_1}{1+bx_1^2} &= y_1 \\
\frac{ax_2}{1+bx_2^2} &= y_2 \\
&\cdots \\
\frac{ax_n}{1+bx_n^2} &= y_n.
\end{aligned}$$

These are n equations for the two unknowns a, b. However, they are nonlinear in b so the method of least squares does not apply directly. To get around this we rewrite the curve as $y(1+bx^2) = ax$, that is, $ax - bx^2y = y$. This equation is now *linear* in the unknown coefficients a and b. We want to pick these to solve the equations

$$\begin{aligned}
ax_1 - bx_1^2y_1 &= y_1 \\
ax_2 - bx_2^2y_2 &= y_2 \\
\cdots & \quad \cdots \\
ax_2 - bx_n^2y_n &= y_n.
\end{aligned}$$

with the least error. These are linear equations of the form $AV = W$, where the data matrix is

$$A = \begin{pmatrix} x_1 & -x_1^2y_1 \\ x_2 & -x_2^2y_2 \\ \cdots & \cdots \\ x_n & -x_n^2y_b \end{pmatrix}$$

so we solve the normal equations $A^*AV = A^*W$ as before.

Problems Using Least Squares

1. Use the Method of Least Squares to find the straight line $y = ax + b$ that best fits the following data given by the following four points (x_j, y_j), $j = 1, \ldots, 4$:

$$(-2,4), \qquad (-1,3), \qquad (0,1), \qquad (2,0).$$

   Ideally, you'd like to pick the coefficients a and b so that the four equations $ax_j + b = y_j$, $j = 1, \ldots, 4$ are all satisfied. Since this probably can't be done, one uses least squares to find the best possible a and b.

2. Find a curve of the form $y = a + bx + cx^2$ that best fits the following data

x	-2	-1	0	1	2	3	4
y	4	1.1	-0.5	1.0	4.3	8.1	17.5

3. Find a plane of the form $z = ax + by + c$ that best fits the following data

x	0	1	0	1	0
y	0	1	1	0	-1
z	1.1	2	-0.1	3	2.2

4. The water level in the North Sea is mainly determined by the so-called M2 tide, whose period is about 12 hours. The height $H(t)$ thus roughly has the form

$$H(t) = c + a\sin(2\pi t/12) + b\cos(2\pi t/12),$$

where time t is measured in hours (note $\sin(2\pi t/12$ and $\cos(2\pi t/12)$ are periodic with period 12 hours). Say one has the following measurements:

t (hours)	0	2	4	6	8	10
$H(t)$ (meters)	1.0	1.6	1.4	0.6	0.2	0.8

Use the method of least squares with these measurements to find the constants a, b, and c in $H(t)$ for this data.

5. a). Some experimental data (x_i, y_i) is believed to fit a curve of the form

$$y = \frac{1+x}{a+bx^2},$$

where the parameters a and b are to be determined from the data. The method of least squares does not apply directly to this since the parameters a and b do not appear linearly. Show how to find a modified equation to which the method of least squares does apply.

b). Repeat part a) for the curve $y = \dfrac{1}{a+bx}$.

c). Repeat part a) for the curve $y = \dfrac{x}{a+bx}$.

d). Repeat part a) for the curve $y = ax^b$.

e). Repeat part a) for the *logistic curve* $y = \dfrac{L}{1+e^{a-bx}}$. Here the constant L is assumed to be known. [If $b > 0$, then y converges to L as x increases. Thus the value of L can often be estimated simply by eye-balling a plot of the data for large x.]

f). Repeat part a) for the curve $y = 1 - e^{-ax^b}$.

g) Repeat part a) for the curve $y = \dfrac{a+mx}{b+x}$ assuming the constant m is known. [One might find m from the data since y tends to m for x large.]

h). Repeat part a) for the curve $y = \dfrac{a}{1+b\sin x}$

6. The comet Tentax, discovered only in 1968, moves within the solar system. The following are observations of its position (r,θ) in a polar coordinate system with center at the sun:

r	2.70	2.00	1.61	1.20	1.02
θ	48	67	83	108	126

(here θ is an angle measured in degrees).

By Kepler's first law the comet should move in a plane orbit whose shape is either an ellipse, hyperbola, or parabola (this assumes the gravitational influence of the planets is neglected). Thus the polar coordinates (r,θ) satisfy

$$r = \frac{p}{1-e\cos\theta}$$

where p and the eccentricity e are parameters describing the orbit. Use the data to estimate p and e by the method of least squares. Hint: Make some (simple) preliminary manipulation so the parameters p and e appear *linearly*; then apply the method of least squares.

7. **Plotting graphs**This problem concerns the straight line in the plane that passes through the two points $(4,0)$ and $(0,2)$ (draw a sketch). This will be useful for the next problem.

   a) If the horizontal axis is x and the vertical axis y, what is the equation for y as a function of x?

b) If the horizontal axis is $\log x$ and the vertical axis y, what is the equation for y as a function of x?

c) If the horizontal axis is x and the vertical axis $\log y$, what is the equation for y as a function of x?

d) If the horizontal axis is $\log x$ and the vertical axis $\log y$, what is the equation for y as a function of x?

8. For each of the seven closest planets, Kepler, using data from Bruno, knew the distance r from the planet to the sun (in million km)and the time T it takes to orbit the sun (the length in earth days of a year on that planet).

	Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus
r	60	110	150	230	780	1430	2870
T	90	225	365	690	4330	10750	30650

Kepler sought a formula relating r and T. It took him a long time; he did not have logarithms. Guided by the idea of using graphs as in the previous problem, you can do this fairly easily.

Make four experimental graphs of this data (as in the previous problem just above). The goal is to hope one of these four curves looks roughly like a straight line. If it does, then use least squares to find the "best" straight line – and then the desired formula for the relation between r and T.

[Since the data is only approximate and since we anticipate a "simple" answer, you may find it appropriate to use your numerical results to lead you to a simpler formula.]

9. Let $A : \mathbb{R}^n \to \mathbb{R}^k$ be a linear map. If A is not one-to-one, but the equation $Ax = y$ has some solution, then it has many. Is there a "best" possible answer? What can one say? Think about this before reading the next paragraph.

If there is some solution of $Ax = y$, show there is exactly one solution x_1 of the form $x_1 = A^* w$ for some w, so $AA^* w = y$. Moreover of all the solutions x of $Ax = y$, show that x_1 is closest to the origin (in the Euclidean distance). [REMARK: This situation is related to the case where where A is not onto, so there may not be a solution — but the method of least squares gives an "best" approximation to a solution.]

10. Let $P_1, P_2, \ldots, P_k$ be k points (think of them as *data*) in $\mathbb{R}^3$ and let $\mathcal{S}$ be the plane

$$\mathcal{S} := \left\{ X \in \mathbb{R}^3 : \langle X, N \rangle = c \right\},$$

where $N \neq 0$ is a unit vector normal to the plane and c is a real constant.

This problem outlines how to find the plane that *best approximates the data points* in the sense that it minimizes the function

$$Q(N,c) := \sum_{j=1}^{k} \text{distance}\,(P_j,\mathcal{S})^2.$$

Determining this plane means finding N and c.

a) Show that for a given point P, then

$$\text{distance}\,(P,\mathcal{S}) = |\langle P - X\,, N\rangle| = |\langle P, N\rangle - c\,|\,,$$

where X is any point in $\mathcal{S}$

b) First do the special case where the center of mass $\overline{P} := \frac{1}{k}\sum_{j=1}^{k} P_j$ is at the origin, so $\overline{P} = 0$. Show that for any P, then $\langle P, N\rangle^2 = \langle N, PP^* N\rangle$. Here view P as a column vector so PP^* is a 3×3 matrix.

Use this to observe that the desired plane $\mathcal{S}$ is determined by letting N be an eigenvector of the matrix

$$A := \sum_{j=1}^{k} P_j P_j^T$$

corresponding to it's lowest eigenvalue. What is c in this case?

c) Reduce the general case to the previous case by letting $V_j = P_j - \overline{P}$.

d) Find the equation of the line $ax + by = c$ that, in the above sense, best fits the data points $(-1,3)$, $(0,1)$, $(1,-1)$, $(2,-3)$.

e) Let $P_j := (p_{j1},\ldots,p_{j3})$, $j = 1,\ldots,k$ be the coordinates of the j^{th} data point and $Z_\ell := (p_{1\ell},\ldots,p_{k\ell})$, $\ell = 1,\ldots,3$ be the vector of ℓ^{th} coordinates. If a_{ij} is the ij element of A, show that $a_{ij} = \langle Z_i, Z_j\rangle$. Note that this exhibits A as a *Gram matrix* .

f) Generalize to where P_1, P_2,..., P_k are k points in $\mathbb{R}^n$.