The Fundamental Theorem of Linear Algebra

Gilbert Strang

This paper is about a theorem and the pictures that go with it. The theorem describes the action of an m by n matrix. The matrix A produces a linear transformation from R^n to R^m—but this picture by itself is too large. The "truth" about $Ax = b$ is expressed in terms of four subspaces (two of R^n and two of R^m). The pictures aim to illustrate the action of A on those subspaces, in a way that students won't forget.

The first step is to see Ax as a *combination of the columns of* A. Until then the multiplication Ax is just numbers. This step raises the viewpoint to subspaces. We see Ax in the ***column space.*** Solving $Ax = b$ means finding all combinations of the columns that produce b in the column space:

$$\underbrace{\begin{bmatrix} | & | & \cdots & | \end{bmatrix}}_{\text{Columns of } A} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1(\text{column } 1) + \cdots + x_n(\text{column } n) = b.$$

The column space is the range $R(A)$, a subspace of R^m. This abstraction, from entries in A or x or b to the picture based on subspaces, is absolutely essential. Note how subspaces enter *for a purpose.* We could invent vector spaces and construct bases at random. That misses the purpose. Virtually all algorithms and all applications of linear algebra are understood by moving to subspaces.

The key algorithm is elimination. Multiples of rows are subtracted from other rows (and rows are exchanged). There is no change in the ***row space.*** This subspace contains all combinations of the rows of A, which are the columns of A^T. The row space of A is the column space $R(A^T)$.

The other subspace of R^n is the ***nullspace*** $N(A)$. It contains all solutions to $Ax = 0$. Those solutions are not changed by elimination, whose purpose is to compute them. A by-product of elimination is to display the dimensions of these subspaces, which is the first part of the theorem.

The *Fundamental Theorem of Linear Algebra* has as many as four parts. Its presentation often stops with Part 1, but the reader is urged to include Part 2. (That is the only part we will prove—it is too valuable to miss. This is also as far as we go in teaching.) The last two parts, at the end of this paper, sharpen the first two. The complete picture shows the action of A on the four subspaces with the right bases. Those bases come from the singular value decomposition.

The Fundamental Theorem begins with

Part 1. ***The dimensions of the subspaces.***

Part 2. ***The orthogonality of the subspaces.***

The dimensions obey the most important laws of linear algebra:

$$\dim R(A) = \dim R(A^T) \quad \text{and} \quad \dim R(A) + \dim N(A) = n.$$

When the row space has dimension r, the nullspace has dimension $n - r$. Elimination identifies r pivot variables and $n - r$ free variables. Those variables correspond, in the echelon form, to columns with pivots and columns without pivots. They give the dimension count r and $n - r$. Students see this for the echelon matrix and believe it for A.

The *orthogonality* of those spaces is also essential, and very easy. Every x in the nullspace is perpendicular to every row of the matrix, exactly because $Ax = 0$:

$$Ax = \begin{bmatrix} \text{—row} & 1\text{—} \\ \text{—row} & 2\text{—} \\ \text{—row} & m\text{—} \end{bmatrix} x = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

The first zero is the dot product of x with row 1. The last zero is the dot product with row m. One at a time, the rows are perpendicular to any x in the nullspace. So x is perpendicular to all combinations of the rows.

The nullspace $N(A)$ is orthogonal to the row space $R(A^T)$.

What is the fourth subspace? If the matrix A leads to $R(A)$ and $N(A)$, then its transpose must lead to $R(A^T)$ and $N(A^T)$. The fourth subspace is $N(A^T)$, ***the nullspace of*** A^T. We need it! The theory of linear algebra is bound up in the connections between row spaces and column spaces. If $R(A^T)$ is orthogonal to $N(A)$, then—*just by transposing*—the column space $R(A)$ is orthogonal to the "left nullspace" $N(A^T)$. Look at $A^T y = 0$:

$$A^T y = \begin{bmatrix} \text{column 1 of } A \\ \vdots \\ \text{column } n \text{ of } A \end{bmatrix} y = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Since y is orthogonal to each column (producing each zero), y is orthogonal to the whole column space. The point is that A^T is just as good a matrix as A. Nothing is new, except A^T is n by m. Therefore the left nullspace has dimension $m - r$.

$A^T y = 0$ means the same as $y^T A = 0^T$. With the vector on the left, $y^T A$ is a combination of the *rows* of A. Contrast that with $Ax =$ combination of the columns.

The First Picture: Linear Equations

Figure 1 shows how A takes x into the column space. The nullspace goes to the zero vector. Nothing goes elsewhere in the left nullspace—which is waiting its turn.

With b in the column space, $Ax = b$ can be solved. There is a *particular* solution x_r in the row space. The *homogeneous* solutions x_n form the nullspace. The general solution is $x_r + x_n$. The particularity of x_r is that it is orthogonal to every x_n.

May I add a personal note about this figure? Many readers of *Linear Algebra and Its Applications* [4] have seen it as fundamental. It captures so much about $Ax = b$. Some letters suggested other ways to draw the orthogonal subspaces—artistically this is the hardest part. The four subspaces (and very possibly the figure itself) are of course not original. But as a key to the teaching of linear algebra, this illustration is a gold mine.

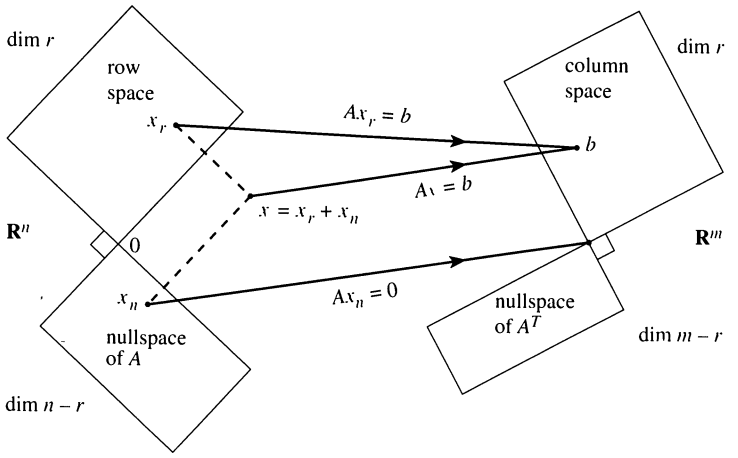


Figure 1. The action of A: Row space to column space, nullspace to zero.

Other writers made a further suggestion. They proposed a lower level textbook, recognizing that the range of students who need linear algebra (and the variety of preparation) is enormous. That new book contains Figures 1 and 2—also Figure 0, to show the dimensions first. The explanation is much more gradual than in this paper—but every course has to study subspaces! We should teach the important ones.

The Second Figure: Least Squares Equations

If b is not in the column space, $Ax = b$ cannot be solved. In practice we still have to come up with a "solution." It is extremely common to have more equations than unknowns—more output data than input controls, more measurements than parameters to describe them. The data may lie close to a straight line $b = C + Dt$. A parabola $C + Dt + Et^2$ would come closer. Whether we use polynomials or sines and cosines or exponentials, the problem is still linear in the coefficients C, D, E:

$$\begin{matrix} C + Dt_1 = b_1 \\ \vdots \\ C + Dt_m = b_m \end{matrix} \qquad \text{or} \qquad \begin{matrix} C + Dt_1 + Et_1^2 = b_1 \\ \vdots \\ C + Dt_m + Et_m^2 = b_m \end{matrix} .$$

There are $n = 2$ or $n = 3$ unknowns, and m is larger. There is no $x = (C, D)$ or $x = (C, D, E)$ that satisfies all m equations. $Ax = b$ has a solution only when the points lie exactly on a line or a parabola—then b is in the column space of the m by 2 or m by 3 matrix A.

The solution is to make the error $b - Ax$ as small as possible. Since Ax can never leave the column space, choose the closest point to b in that subspace. This point is the projection p. Then the error vector $e = b - p$ has minimal length.

To repeat: The best combination $p = A\bar{x}$ is the projection of b onto the column space. The error e is perpendicular to that subspace. Therefore $e = b - A\bar{x}$ is in the left nullspace:

$$A^T(b - A\bar{x}) = 0 \quad \text{or} \quad A^TA\bar{x} = A^Tb.$$

Calculus reaches the same linear equations by minimizing the quadratic $\|b - Ax\|^2$. The chain rule just multiplies both sides of $Ax = b$ by A^T.

The "normal equations" are $A^TA\bar{x} = A^Tb$. They illustrate what is almost invariably true—applications that start with a rectangular A end up computing with the square symmetric matrix A^TA. This matrix is invertible provided A has *independent columns*. We make that assumption: The nullspace of A contains only $x = 0$. (Then $A^TAx = 0$ implies $x^TA^TAx = 0$ which implies $Ax = 0$ which forces $x = 0$, so A^TA is invertible.) The picture for least squares shows the action over on the right side—the splitting of b into $p + e$.

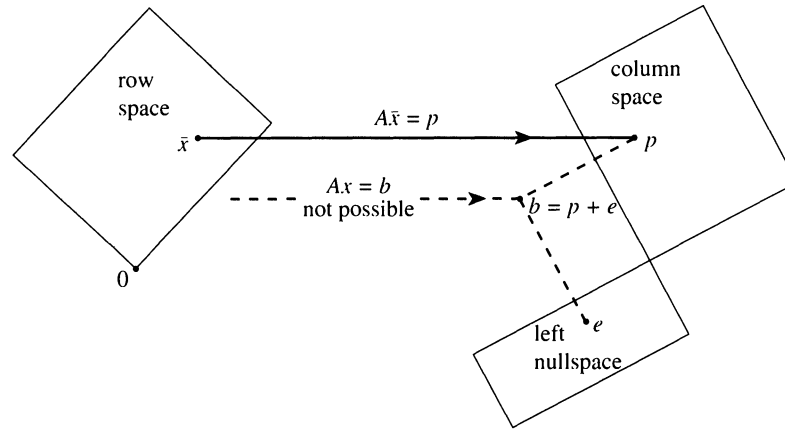


Figure 2. Least squares: $\bar{x}$ minimizes $\|b - Ax\|^2$ by solving $A^TA\bar{x} = A^Tb$.

The Third Figure: Orthogonal Bases

Up to this point, nothing was said about *bases for the four subspaces*. Those bases can be constructed from an echelon form—the output from elimination. This construction is simple, but the bases are not perfect. A really good choice, in fact a "canonical choice" that is close to unique, would achieve much more. To complete the Fundamental Theorem, we make two requirements:

Part 3. ***The basis vectors are orthonormal.***
Part 4. ***The matrix with respect to these bases is diagonal.***

If $v_1, \ldots, v_r$ is the basis for the row space and $u_1, \ldots, u_r$ is the basis for the column space, then $Av_i = \sigma_i u_i$. That gives a diagonal matrix Σ. We can further ensure that $\sigma_i > 0$.

Orthonormal bases are no problem—the Gram-Schmidt process is available. But a diagonal form involves eigenvalues. In this case they are the eigenvalues of A^TA and AA^T. Those matrices are symmetric and positive semidefinite, so they have nonnegative eigenvalues and orthonormal eigenvectors (which are the bases!). Starting from $A^TAv_i = \sigma_i^2 v_i$, here are the key steps:

$$v_i^TA^TAv_i = \sigma_i^2 v_i^Tv_i \quad \text{so that} \quad \|Av_i\| = \sigma_i$$

$$AA^TAv_i = \sigma_i^2 Av_i \quad \text{so that} \quad u_i = Av_i/\sigma_i \text{ is a unit eigenvector of } AA^T.$$

All these matrices have rank r. The r positive eigenvalues σ_i^2 give the diagonal entries σ_i of Σ.

The whole construction is called the *singular value decomposition* (*SVD*). It amounts to a factorization of the original matrix A into $U\Sigma V^T$, where

1. U is an m by m orthogonal matrix. Its columns $u_1, \ldots, u_r, \ldots, u_m$ are basis vectors for the column space and left nullspace.
2. Σ is an m by n diagonal matrix. Its nonzero entries are $\sigma_1 > 0, \ldots, \sigma_r > 0$.
3. V is an n by n orthogonal matrix. Its columns $v_1, \ldots, v_r, \ldots, v_n$ are basis vectors for the row space and nullspace.

The equations $Av_i = \sigma_i u_i$ mean that $AV = U\Sigma$. Then multiplication by V^T gives $A = U\Sigma V^T$.

When A itself is symmetric, its eigenvectors u_i make it diagonal: $A = U\Lambda U^T$. The singular value decomposition extends this spectral theorem to matrices that are not symmetric and not square. The eigenvalues are in Λ, the singular values are in Σ. The factorization $A = U\Sigma V^T$ joins $A = LU$ (elimination) and $A = QR$ (orthogonalization) as a beautifully direct statement of a central theorem in linear algebra.

The history of the *SVD* is cloudy, beginning with Beltrami and Jordan in the 1870's, but its importance is clear. For a very quick history and proof, and much more about its uses, please see [**1**]. "The most recurring theme in the book is the practical and theoretical value of this matrix decomposition." The *SVD* in linear algebra corresponds to the Cartan decomposition in Lie theory [**3**]. This is one more case, if further convincing is necessary, in which mathematics gets the properties right—and the applications follow.

Example

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} = \frac{\begin{bmatrix} 1 & -3 \\ 3 & 1 \end{bmatrix}}{\sqrt{10}} \begin{bmatrix} \sqrt{50} & 0 \\ 0 & 0 \end{bmatrix} \frac{\begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}}{\sqrt{5}} = U\Sigma V^T.$$

All four subspaces are 1-dimensional. The columns of A are multiples of $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ in U. The rows are multiples of $[1 \quad 2]$ in V^T. Both A^TA and AA^T have eigenvalues 50 and 0. So the only singular value is $\sigma_1 = \sqrt{50}$.

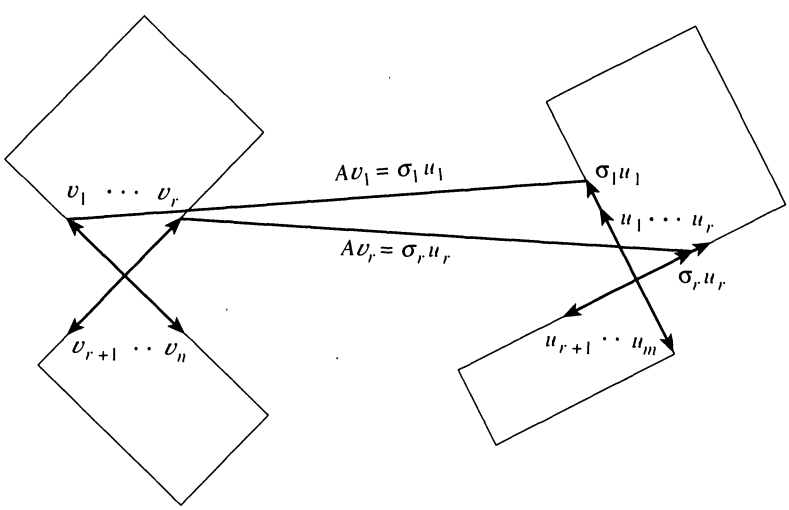


Figure 3. Orthonormal bases that diagonalize A.

The *SVD* expresses A as a combination of r rank-one matrices:

$$A = U\Sigma V^T = u_1\sigma_1 v_1^T + \cdots + u_r\sigma_r v_r^T \qquad \left(\text{here } A = \begin{bmatrix} 1 \\ 3 \end{bmatrix}\begin{bmatrix} 1 & 2 \end{bmatrix}\right).$$

The Fourth Figure: The Pseudoinverse

The *SVD* leads directly to the "*pseudoinverse*" of A. This is needed, just as the least squares solution $\bar{x}$ was needed, to invert A and solve $Ax = b$ when those steps are strictly speaking impossible. The pseudoinverse A^+ agrees with A^{-1} when A is invertible. The least squares solution of minimum length (having no nullspace component) is $x^+ = A^+b$. It coincides with $\bar{x}$ when A has full column rank $r = n$—then A^TA is invertible and Figure 4 becomes Figure 2.

A^+ takes the column space back to the row space [4]. On these spaces of equal dimension r, the matrix A is invertible and A^+ inverts it. On the left nullspace, A^+ is zero. I hope you will feel, after looking at Figure 4, that this is the one natural best definition of an inverse. Despite those good adjectives, the *SVD* and A^+ is too much for an introductory linear algebra course. It belongs in a second course. Still the picture with the four subspaces is absolutely intuitive.

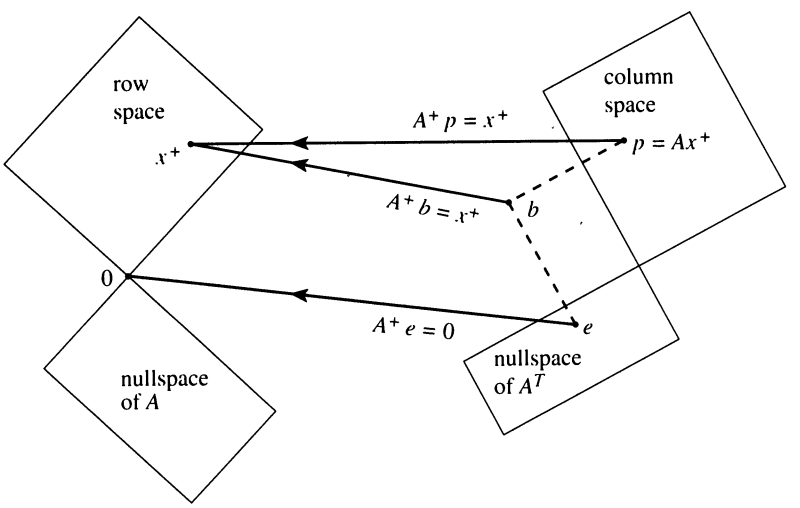


Figure 4. The inverse of A (where possible) is the pseudoinverse A^+.

The *SVD* gives an easy formula for A^+, because it chooses the right bases. Since $Av_i = \sigma_i u_i$, the inverse has to be $A^+u_i = v_i/\sigma_i$. Thus the pseudoinverse of Σ contains the reciprocals $1/\sigma_i$. The orthogonal matrices U and V^T are inverted by U^T and V. All together, *the pseudoinverse of* $A = U\Sigma V^T$ *is* $A^+ = V\Sigma^+U^T$.

Example (continued)

$$A^+ = \frac{\begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}}{\sqrt{5}} \begin{bmatrix} 1/\sqrt{50} & 0 \\ 0 & 0 \end{bmatrix} \frac{\begin{bmatrix} 1 & 3 \\ -3 & 1 \end{bmatrix}}{\sqrt{10}} = \frac{1}{50}\begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix}.$$

Always A^+A is the identity matrix on the row space, and zero on the nullspace:

$$A^+A = \frac{1}{50}\begin{bmatrix} 10 & 20 \\ 20 & 40 \end{bmatrix} = \text{projection onto the line through } \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Similarly AA^+ is the identity on the column space, and zero on the left nullspace:

$$AA^+ = \frac{1}{50}\begin{bmatrix} 5 & 15 \\ 15 & 45 \end{bmatrix} = \text{projection onto the line through } \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

A Summary of the Key Ideas

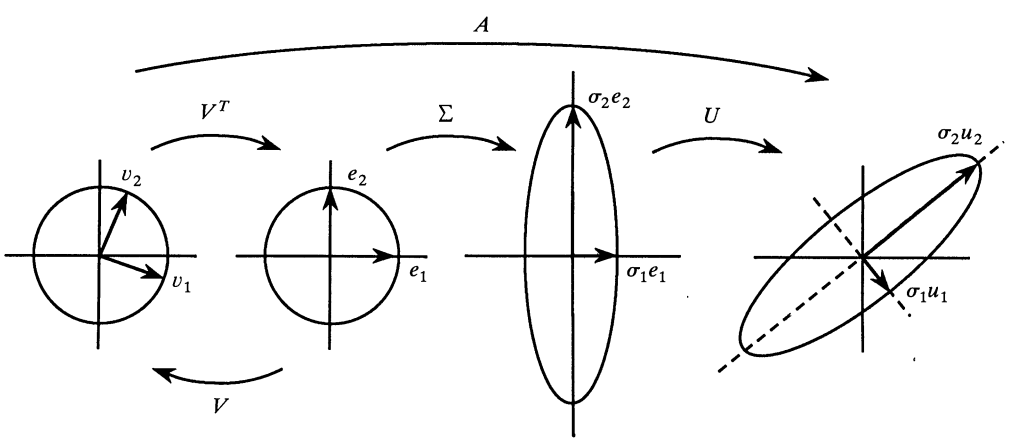
From its r-dimensional row space to its r-dimensional column space, A yields an invertible linear transformation.

Proof: Suppose x and x' are in the row space, and Ax equals Ax' in the column space. Then $x - x'$ is in both the row space and nullspace. It is perpendicular to itself. Therefore $x = x'$ and the transformation is one-to-one.

The SVD chooses good bases for those subspaces. Compare with the Jordan form for a real square matrix. There we are choosing the *same basis* for both domain and range—our hands are tied. The best we can do is $SAS^{-1} = J$ or $SA = JS$. In general J is not real. If real, then in general it is not diagonal. If diagonal, then in general S is not orthogonal. By choosing *two bases*, not one, every matrix does as well as a symmetric matrix. The bases are orthonormal and A is diagonalized.

Some applications permit two bases and others don't. For powers A^n we need S^{-1} to cancel S. Only a similarity is allowed (one basis). In a differential equation $u' = Au$, we can make one change of variable $u = Sv$. Then $v' = S^{-1}ASv$. But for $Ax = b$, the domain and range are philosophically "not the same space." The row and column spaces are isomorphic, but their bases can be different. And for least squares the *SVD* is perfect.

This figure by Tom Hern and Cliff Long [2] shows the diagonalization of A. Basis vectors go to basis vectors (principal axes). A circle goes to an ellipse. The matrix is factored into $U\Sigma V^T$. Behind the scenes are *two* symmetric matrices A^TA and AA^T. So we reach two orthogonal matrices U and V.



We close by summarizing the action of A and A^T and A^+:

$$Av_i = \sigma_i u_i \qquad A^T u_i = \sigma_i v_i \qquad A^+ u_i = v_i/\sigma_i \qquad 1 \leqslant i \leqslant r.$$

The nullspaces go to zero. Linearity does the rest.

The support of the National Science Foundation (DMS 90-06220) is gratefully acknowledged.

REFERENCES

1. Gene Golub and Charles Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press (1989).
2. Thomas Hern and Cliff Long, Viewing some concepts and applications in linear algebra, *Visualization in Teaching and Learning Mathematics*, MAA Notes 19 (1991) 173–190.
3. Roger Howe, Very basic Lie theory, *American Mathematical Monthly*, 90 (1983) 600–623.
4. Gilbert Strang, *Linear Algebra and Its Applications*, 3rd ed., Harcourt Brace Jovanovich (1988).
5. Gilbert Strang, *Introduction to Linear Algebra*, Wellesley-Cambridge Press (1993).

Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139
gs@math.mit.edu

An Identity of Daubechies

The generalization of an identity of Daubechies using a probabilistic interpretation by D. Zeilberger [100 (1993) 487], has already appeared in SIAM Review Problem 85-10 (June, 1985) in a slightly more general context. In addition to a similar probabilistic derivation there is also a direct algebraic proof. Incidentally, problem 10223 [99 (1992) 462] is the same as the identity of Daubechies and a slight generalization of this identity has appeared previously as problem 183, Crux Math. 3(1977) 69–70 and came from a list of problems considered for the Canadian Mathematical Olympiad. There was an inductive solution of the latter by Mark Kleinman, a high school student at the time and one of the top students in the U.S.A.M.O. and the I.M.O.

M. S. Klamkin
Department of Mathematics
University of Alberta
Edmonton, Alberta
CANADA T6G 2G1