Linear Algebra Recap

Ang Zi Qing

April 23, 2020

This write up serves as a quick review of Linear Algebra and aims to trigger my memory of Gilbert Strang's 18.06 Linear Algebra course when its ideas are needed for research purposes or possibly for future endeavours. That being said, go through the following key questions and see if they ring any bells. If it does not work, either return to the videos or read the book and then populate this note. I intend to continuously update this note hence it is highly possible that I might go beyond Strang's syllabus in the future. This version is at my best understanding of Linear Algebra at the time of writing.

1. What is the role of $\mathbf{A}$ in $\mathbf{Ax} = \mathbf{b}$?
2. What is the difference between dot product (inner product) and cross product (outer product) and what do they mean?
3. What does it mean by the linear combination of columns of $\mathbf{A}$?
4. What are the all the combinations of columns of $\mathbf{A}$, when the columns are dependent/independent of each other?
5. What does it mean when matrix $\mathbf{A}$ is singular?
6. What do we get from elimination?
7. What happens when elimination is a "temporary"/"complete" failure?
8. What is back substitution?
9. Recall elimination and permutation matrix, what does both the matrices do together?
10. Recall getting $\mathbf{U}$ back to $\mathbf{A}$, as in inverse, when does $\mathbf{A}^{-1}$ exist? If it does exist, how would you find it?
11. What is the significance of checking the invertibility of $\mathbf{A}$?
12. Can a rectangular matrix be invertible?
13. What is the difference of $\mathbf{Ax}$ and $\mathbf{x}^T\mathbf{A}^T$? What is the difference of $\mathbf{x}^T\mathbf{y}$ and $\mathbf{x}\mathbf{y}^T$?
14. What is the inverse and transpose of a permutation matrix?
15. How does symmetry produces a special relationship between $\mathbf{L}$ and $\mathbf{U}$?
16. Recall factorization into $\mathbf{A} = \mathbf{LU}$, what is $\mathbf{L}$? What is in the diagonal of $\mathbf{L}$ and $\mathbf{U}$? How is $\mathbf{LDU}$ different from $\mathbf{LU}$? What happens when there is no row exchange? What is the factorization with row exchanges?
17. What does $\mathbf{Ax} = \mathbf{0}$ tell us about $\mathbf{A}$? What happens if only $\mathbf{x} = \mathbf{0}$? What happens otherwise? What does this tells us about $\mathbf{A}^T\mathbf{A}$?

18. Connect a vector space and subspace, what is the relationship between subspace and vector space?
19. Explain the following: independence, span, basis, rank, dimension.
20. Explain the four fundamental subspaces.
21. Explain orthogonal vectors and subspaces.
22. Explain projections onto subspaces.
23. When matrix $\mathbf{A}$ is square and independent, we can solve for $\mathbf{x}$ in $\mathbf{Ax} = \mathbf{b}$. How to solve for $\mathbf{x}$ when $\mathbf{A}$ is rectangular?
24. Recall the connection between orthogonality, projection matrices and least squares. [Linear Regression].
25. What are orthogonal matrices, what do they do?
26. Recall the ten properties of determinants.
27. Connect determinant and volume.
28. What are eigenvalues, eigenvectors and eigenspace?
29. What is in the nullspace of $\mathbf{A} = \lambda\mathbf{I}$? What is this space?
30. Recall differential equations and the matrix exponential e^{At}.
31. Recall Markov matrices and the Fourier Series.
32. What does it mean when $\mathbf{Q}^{-1} = \mathbf{Q}^T$?
33. Why do we look for positive definite symmetric matrices?
34. Why are the eigenvalues of symmetric matrices real? What to do if a vector or a matrix is complex?
35. What does $\mathbf{A} = \mathbf{Q\Lambda Q}^T$ say about symmetric matrices? What is the Spectral Theorem, $\mathbf{Q\Lambda Q}^{-1} = \mathbf{Q\Lambda Q}^T$?
36. What happens when a matrix has repeating eigenvalues?
37. What happens when the vectors and matrices are complex? What do you do when the numbers become complex numbers?[Hermition]
38. How to tell if a matrix is positive definite and what does it mean? Why do we look for positive definite matrices? [Minima of matrix]
39. Why do we look for the minima (completing the square to get positive values)? What happens if the matrix is not positive definite?
40. Why do similar matrices share the same eigenvalues? What do similar matrices have in common? What makes them different?
41. What makes $\mathbf{A} = \mathbf{Q\Lambda Q}^T = \mathbf{S\Lambda S}^{-1}$?
42. What does it mean to have an SVD? How to find the three components of an SVD of a matrix $\mathbf{A}$ such that $\mathbf{A} = \mathbf{U\Sigma V}^T$?
43. In real application, the first vector of $\mathbf{V}$ and $\mathbf{U}$, as well as the biggest singular value of the decomposition of a matrix $\mathbf{A}$ that are the most important, why? Why these components make the best combination to find correlation between desired variables and hence carry the most information?

1 What is the role of A in $\mathbf{Ax} = \mathbf{b}$?

Considering the two linear equations with two unknowns,

$$2x - y = 0 \tag{1}$$

$$-x + 2y = 3 \tag{2}$$

we can express these equations in matrix form $\mathbf{Ax} = \mathbf{b}$ such that

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \end{bmatrix} \tag{3}$$

where the matrix $\mathbf{A}$ of size $m \times n$ is the coefficient matrix, the column vector $\mathbf{x}$ with m entries is the vector of unknowns and the column vector $\mathbf{b}$ with n entries is the result of matrix $\mathbf{A}$ acting on vector $\mathbf{x}$ (Recall the difference between the row picture and the column picture). Comparing to the equations, the coefficient matrix $\mathbf{A}$ consists of all the coefficients of the variables x and y in the set of linear equations. A right combination of these coefficients by finding some values for x and y would give us the vector $\mathbf{b}$.

The matrix $\mathbf{A}$ plays a crucial role in determining whether $\mathbf{Ax} = \mathbf{b}$ has a solution for $\mathbf{b}$ since a solution for the set of linear equations exists only if $\mathbf{b}$ is a linear combination of the columns of $\mathbf{A}$. To check if $\mathbf{Ax} = \mathbf{b}$ has a solution for every $\mathbf{b}$, we look for a pivot in every row of $\mathbf{A}$. If $m = n$ and $\mathbf{A}$ has a pivot in every row,

- all the combinations fill the entire space, that is, every column vector $\mathbf{b}$ is a linear combination of column vectors of $\mathbf{A}$
- columns of $\mathbf{A}$ are independent
- columns of $\mathbf{A}$ spans the column space $\mathbb{R}^m$, where $\mathbb{R}^m$ also contains all the combinations of columns of $\mathbf{A}$
- the nullspace has only the zero vector, that is $\mathbf{x} = \mathbf{0}$ (trivial solution) is the only solution to get $\mathbf{b} = \mathbf{0}$.

Solving $\mathbf{Ax} = \mathbf{b}$ is the same as solving the system described by the augmented matrix. We will eventually see that finding the pivots of an invertible $n \times n$ matrix $\mathbf{A}$ leads us to determining bases of finite-dimensional vector spaces, where a chosen basis (a sequence of independent vectors that span the space) become columns of scalars. If we know the number of pivots of $\mathbf{A}$, then the rank and hence the dimension of the column space of $\mathbf{A}$ will subsequently be known, which also gives us the number of vectors in any basis as well as the vectors that span the space. Different choices of basis gives different columns, hence also give different matrices, which corresponds to linear transformations. Ultimately the matrix becomes a tool for us to find the determinant, trace and eigenvalues of the linear transformation, which is a structure-preserving map.

Of course, not all $\mathbf{Ax} = \mathbf{b}$ can be solved. Linear algebra tells us what do to and what it means when we can solve and when we can't. The big picture of what $\mathbf{A}$ does apppears five times in Strang's textbook (Figures 3.5, 4.2, 4.3, 4.7 and 7.6 in the book), six times if we include the one on the book cover, each showing the role of $\mathbf{A}$ at increasing level of understanding of Linear Algebra.

- Level 1 of Fundamental Theorem of Linear Algebra (Figure 3.5): Finds dimensions of the four subspaces for $\mathbf{A}$ and $\mathbf{R}$ such that $\mathbf{R} = rref(\mathbf{A})$.
- Level 2 of Fundamental Theorem of Linear Algebra (Figure 4.2): Describes how the four subspaces fit together (two in $\mathbf{R}^n$ and two in $\mathbf{R}^m$) and the orthogonality of the subspaces. such that $C(\mathbf{A}^T) \perp N(\mathbf{A})$ and $C(\mathbf{A}) \perp N(\mathbf{A}^T)$. Each pair of subspaces is not only orthogonal, but also orthogonal complements to each other.

- "Splitting" of $\mathbf{x}$ (Figure 4.3): The "complements" allow $\mathbf{x}$ to split into a row space component ($\mathbf{x}_r$) and a nullspace component $(x)_n$ so that $\mathbf{x} = \mathbf{x}_r + \mathbf{x}_n$. Then when $\mathbf{A}$ acts on $\mathbf{x}$, the nullspace component goes to zero and the row space component goes to the column space, that is, multiplying $\mathbf{A}$ makes every vector go to the column space and there may be many solutions to $\mathbf{Ax} = \mathbf{b}$.

- Least squares - "Splitting" of $\mathbf{b}$ (Figure 4.7): If there is no solution to $\mathbf{Ax} = \mathbf{b}$, split $\mathbf{b}$ so $\mathbf{b} = \mathbf{p} + \mathbf{e}$. Then instead of solving $\mathbf{Ax} = \mathbf{b}$, we solve $\mathbf{A\hat{x}} = \mathbf{p}$ so we get the best $\hat{\mathbf{x}}$ and $\mathbf{e} = \mathbf{b} - \mathbf{p}$ is unavoidable. The normal equation $\mathbf{A}^T\mathbf{A\hat{x}} = \mathbf{A}^T\mathbf{b}$ fully determines the best $\hat{\mathbf{x}}$ and minimizes the error $\mathbf{e}$.

- Complete picture (Figure 7.6): Every $\mathbf{x}$ splits into $\mathbf{x}_r + \mathbf{r}_n$ and every $\mathbf{b}$ splits into $\mathbf{p} + \mathbf{e}$. We cannot do anything with $\mathbf{e}$ and we do not want $\mathbf{x}_n$ from the nullspace, which leaves $\mathbf{A\hat{x}} = \mathbf{p}$.

2 What does it mean when matrix A is singular?

Consider the following set of linear equations (1) and (2) and the matrix form (3), we can also express in terms of the column picture

$$x \begin{bmatrix} 2 \\ -1 \end{bmatrix} + y \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \end{bmatrix}. \tag{4}$$

It is clear that the combination of one of vector $\begin{bmatrix} 2 & -1 \end{bmatrix}^T$ and two of vector $\begin{bmatrix} -1 & 2 \end{bmatrix}^T$ gives us the RHS $\begin{bmatrix} 0 & 3 \end{bmatrix}^T$. This demonstrates that $\mathbf{Ax}$ is a combination of columns of $\mathbf{A}$, and the vector $\mathbf{x} = \begin{bmatrix} 1 & 2 \end{bmatrix}^T$.

Consider another set of linear equations,

$$x + 2y = 3 \tag{5}$$
$$2x + 4y = 7, \tag{6}$$

expressing these equations in matrix form gives

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \end{bmatrix} \tag{7}$$

and in the column picture,

$$x \begin{bmatrix} 1 \\ 2 \end{bmatrix} + y \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \end{bmatrix}. \tag{8}$$

This is a situation where it is impossible to produce some $\mathbf{b}$ out of the columns of $\mathbf{A}$ or their combinations. Notice that the columns of $\mathbf{A}$ are not independent since

$$2 \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix},$$

which means that the columns lie in the same plane, and their combinations will also lie in the same plane. Nevertheless, we still can get some $\mathbf{b}$ from the plane, for example, when $x = 1$ and $y = 1$,

$$\mathbf{b} = 1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 1 \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix},$$

but most other $\mathbf{b}$ would be out of the plane and unreachable. The only $\mathbf{b}$s that we can get would be the ones on that plane only. We call this kind of cases SINGULAR CASES such that:

- at least one of the column vectors of $\mathbf{A}$ is a linear combination of the rest

- the columns of $\mathbf{A}$ do not span the entire space
- there will not be a solution for every $\mathbf{b}$
- matrix $\mathbf{A}$ will not be invertible, since not all columns of $\mathbf{A}$ has a pivot. Some pivots are zero in row reduced form after elimination operation, which is the case of complete failure for elimination. [Temporary failures can be resolved by exchanging rows].
- There exist a vector $\mathbf{x}$ other than $\mathbf{x} = \mathbf{0}$ such that $\mathbf{Ax} = \mathbf{0}$, that is, the nullspace of $\mathbf{A}$ has other vectors other than the zero vector.

3 What is the significance of checking the invertibility of A?

Finding the inverse $\mathbf{A}^{-1}$ is equivalent to solving the system of linear equations. Suppose $\mathbf{E}_{ij}$ is the elimination matrix to fix position (i, j) in order to produce zeros below or above the pivots. To obtain a row echelon form for a 3×3 matrix, for example, without row exchanges,

$$\mathbf{E}_{32}\mathbf{E}_{31}\mathbf{E}_{21}\mathbf{A} = \mathbf{U}, \tag{9}$$

where $\mathbf{U}$ is the row echelon form of $\mathbf{A}$ and is an upper triangular matrix. To reverse the elimination operation,

$$\begin{aligned}\mathbf{A} &= \mathbf{E}_{21}^{-1}\mathbf{E}_{31}^{-1}\mathbf{E}_{32}^{-1}\mathbf{U} &(10)\\ &= \mathbf{LU}, &(11)\end{aligned}$$

where $\mathbf{L}$ is a lower triangular matrix and connects $\mathbf{A}$ to $\mathbf{U}$. From here we can observe that $\mathbf{L}$ is the product of inverses of the elimination matrices. Let $\mathbf{E} = \mathbf{E}_{32}\mathbf{E}_{31}\mathbf{E}_{21}$. In Gauss-Jordan elimination,

$$\mathbf{E}(\mathbf{AI}) = (\mathbf{IA}^{-1}), \tag{12}$$

then $\mathbf{EA} = \mathbf{I}$ tells us $\mathbf{E} = \mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}$ envelops the steps of elimination to reach $\mathbf{U}$. Therefore, if an inverse of $\mathbf{A}$ exists, a square matrix $\mathbf{A}$ has a full set of pivots and will always have a two sided inverse such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{AA}^{-1}. \tag{13}$$

If $\mathbf{A}^{-1}$ does not exist, then there exists a vector $\mathbf{x}$, where $\mathbf{x} \neq \mathbf{0}$ such that $\mathbf{Ax} = \mathbf{0}$, for example

$$\mathbf{Ax} = \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Since multiplying $\mathbf{A}$ by some other matrix gives a resulting matrix with columns being multiples of columns of $\mathbf{A}$, there is no way to obtain the identity matrix, because the identity matrix is not a combination of those columns since they lie on the same line in every combination. If $\mathbf{A}^{-1}$ is multiplied on both sides,

$$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\mathbf{0}, \tag{14}$$

then $\mathbf{x} = \mathbf{0}$, which is not true. Hence if $\mathbf{A}^{-1}$ exists, then matrix $\mathbf{A}$ is invertible, columns of $\mathbf{A}$ are independent, and there exist a solution for every $\mathbf{b}$.

4 Can a rectangular matrix be invertible?

A matrix $\mathbf{A}$ is invertible if and only if its column vectors are independent. Not all square matrices can fulfill this condition, hence not all square matrices are invertible, for example, equation (7). In the case of rectangular matrices, not all of them have dependent column vectors. If the column vectors of a matrix $\mathbf{A}$ are independent, then either the matrix is square (i.e. $m = n$), or there are fewer columns than rows (i.e. $m > n$). When $m < n$, the columns cannot be linear independent and hence this rectangular matrix cannot be invertible. So a rectangular matrix $\mathbf{A}$ might have a left inverse or a right inverse, but cannot have a two sided inverse, since a rectangular matrix is not one-to-one and onto.

5 How to solve for a rectangular matrix?

The general solution for both square and rectangular matrix $\mathbf{Ax} = \mathbf{b}$ exists if the matrix $\mathbf{A}$ has a full column rank $r = n$ and/or full row rank $r = m$ such that

$$\mathbf{x} = \mathbf{x_{particular}} + \mathbf{x_{special}} \tag{15}$$

For a square and invertible matrix $\mathbf{A}$, its rank $r = m = n$ and $\mathbf{Ax} = \mathbf{b}$ has only one solution such that $\mathbf{x} = \mathbf{x_{particular}} + \mathbf{0}$ where $\mathbf{x_{particular}}$ gets the scalar values from $\mathbf{b}$ of the augmented matrix after reduced row echelon form is obtained from elimination. There is no special solution since there are no free variables in the reduced row echelon form.

For a short and wide matrix $\mathbf{A}$ that has more columns than rows, it has a full row rank $r = m$ and $r \leq n$. Recall the 4 subspaces, a full rank matrix $\mathbf{A}$ means that the rows are independent but the columns are not, so the left nullspace has only the zero vector but the nullspace has other nonzero vectors in addition to the zero vector. So the particular solution $\mathbf{x_{particular}}$ can be obtained from $\mathbf{b}$ of the augmented matrix after reduced row echelon form is obtained from elimination. The particular solution will have free variable(s), which gives special solution(s) to $\mathbf{Ax} = \mathbf{0}$. Take an example with $m = 2$ and $n = 3$, that is, two equations and three unknowns with rank $r = m = 2$. The two planes in xyz space intersects in a line of solutions (can be obtained from elimination) and the particular solution is one point on the line. The special solution(s) gives the nullspace vector(s), which adding to the particular solution will move along the solution line, hence this kind of matrix has infinite solutions. (See Figure 3.3 of Strang's book)

For a tall and thin matrix $\mathbf{A}$ that has more rows than columns, it has a full column rank $r = n$ and $r \leq m$. For a solution to exists, zero rows in reduced row echelon form must also be zero in $\mathbf{b}$ of the augmented matrix after elimination. Row reduction puts $\mathbf{I}$ at the top of the augmented matrix when $\mathbf{A}$ is reduced to $\mathbf{R}$ with rank n. All columns of $\mathbf{A}$ are pivot columns and there are no free variables or special solutions. Since all columns are pivot columns, the nullspace contains only the zero vector. Therefore, this kind of matrix has only one or no solution. We will see later that a matrix with full column rank has a nice property such that multiplying $\mathbf{A}^T$ to both sides of $\mathbf{Ax} = \mathbf{b}$ so that $\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\hat{\mathbf{b}}$ where $\mathbf{A}^T\mathbf{A}$ is square and invertible.

If $r < n$ and $r < m$, the system has zero or infinitely many solutions.

While answering this question, we can also see the significance of knowing the rank of $\mathbf{A}$. Strang brings up three levels of understanding for the definition of rank:

- Level 1: Rank tells us number of pivots.
- Level 2: Rank tells us number of independent rows and columns.
- Level 3: Rank tells us the dimensions of the column and the row spaces, which combining with the information of m and n tells us also the dimensions of the nullspace and the left-nullspace.

6 What is the meaning behind least squares? (projection and orthogonality)

To solve $\mathbf{Ax} = \mathbf{b}$ is to express $\mathbf{b}$ as a combination of columns of $\mathbf{A}$. Therefore the system $\mathbf{Ax} = \mathbf{b}$ is solvable if and only if $\mathbf{b}$ is in the column space of $\mathbf{A}$, $C(\mathbf{A})$, that is, $\mathbf{b}$ is a combination of the columns. Suppose $\mathbf{Ax} = \mathbf{b}$ has no solution, we can choose a vector $\mathbf{p}$ in the column space that is closest to $\mathbf{b}$ and solve $\mathbf{A\hat{x}} = \mathbf{p}$ instead such that $\mathbf{\hat{x}}$ is the best possible combination of the columns of $\mathbf{A}$, that is, we project $\mathbf{b}$ into the column space $C(\mathbf{A})$.

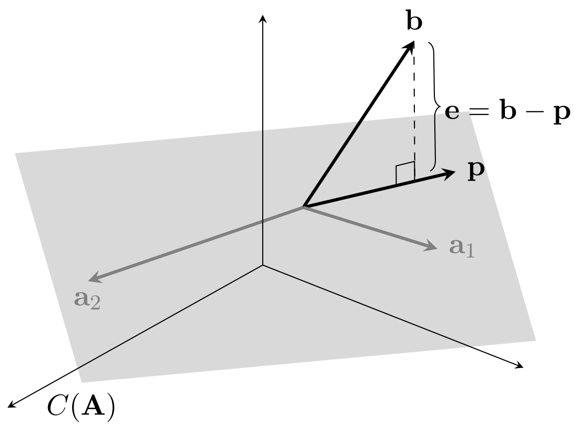

Figure 1: Projection of $\mathbf{b}$ onto the column space of $\mathbf{A}$, $C(\mathbf{A})$.

Consider

$$\mathbf{A} = \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2\end{bmatrix},$$

then

$$\mathbf{p} = \hat{x}_1\mathbf{a}_1 + \hat{x}_2\mathbf{a}_2 = \mathbf{A\hat{x}}.$$

From Figure 1, we observe that

$$\mathbf{e} = \mathbf{b} - \mathbf{p} \perp \mathbf{a}_1$$
$$\mathbf{e} = \mathbf{b} - \mathbf{p} \perp \mathbf{a}_2,$$

so

$$\mathbf{a}_1^T(\mathbf{b} - \mathbf{A\hat{x}}) = 0$$
$$\mathbf{a}_2^T(\mathbf{b} - \mathbf{A\hat{x}}) = 0.$$

Expressing in matrix form, we have

$$\begin{bmatrix}\mathbf{a}_1^T \\ \mathbf{a}_2^T\end{bmatrix}(\mathbf{b} - \mathbf{A\hat{x}}) = \begin{bmatrix}0 \\ 0\end{bmatrix}.$$

Thus we have $\mathbf{A}^T(\mathbf{b} - \mathbf{A\hat{x}}) = \mathbf{0}$, which tells us that $\mathbf{e}$ is in the nullspace of $\mathbf{A}^T$. Recalling the orthogonality of the four subspaces, this also tells us that $\mathbf{e}$ is perpendicular to $C(\mathbf{A})$. Rearraging the terms, we have

$$\mathbf{A}^T\mathbf{A\hat{x}} = \mathbf{A}^T\mathbf{b}, \tag{16}$$

$$\mathbf{\hat{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}. \tag{17}$$

So

$$\mathbf{p} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}, \tag{18}$$

and the projection matrix $\mathbf{P}$ that is multiplying $\mathbf{b}$ to give the projection is

$$\mathbf{P} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T. \tag{19}$$

The projection matrix has two properties:

- $\mathbf{P}$ is symmetric such that $\mathbf{P}^T = \mathbf{P}$
- $\mathbf{P}$ is idempotent such that $\mathbf{P}^2 = \mathbf{P}$, which means a second projection does not change the first projection.

If $\mathbf{A}$ is a square and invertible matrix, $C(\mathbf{A})$ is the whole of $\mathbb{R}^n$, then the projection matrix that projects back into the whole space is the identity matrix $\mathbf{I}$. Since $\mathbf{b}$ is already in the column space, $\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{I}$. On the other hand, if $\mathbf{A}$ is not a square matrix, then there is no inverse $\mathbf{A}^{-1}$ and $\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{I}$ is not allowed. So we have to stay with Equation (18) to project into a subspace. When $\mathbf{A}$ has independent columns, $\mathbf{A}^T\mathbf{A}$ is square, symmetric, and invertible (see page 212 of Strang's book for proof).

Going back to the focus of this part, when there is no solution for $\mathbf{Ax} = \mathbf{b}$, usually when there are more equations than unknowns, then the n columns span a small part of the m -dimensional space and $\mathbf{b}$ lies beyond $C(\mathbf{A})$. In this case, we multiply $\mathbf{A}^T$ to both sides of $\mathbf{Ax} = \mathbf{b}$ so that now have the normal equation, that is Equation (16). As shown above, this projects $\mathbf{b}$ to $C(\mathbf{A})$ and the error $\mathbf{e}$ is unavoidable unless all measurements are perfect (that is, $\mathbf{b}$ is in $C(\mathbf{A})$). The normal equation makes the error $\mathbf{e}$ as small as possible, so that $\hat{\mathbf{x}}$ is a least square solution. Strang's book explains least square in three different perspectives: by Geometry, by Algebra and by Calculus. But the name "least squares" probably came from the Calculus' perspective. Suppose the error function is

$$E = ||\mathbf{Ax} - \mathbf{b}||^2, \tag{20}$$

then the partial derivatives of Equation (20) are zero when $\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b}$. The equations from Calculus are the same as the normal enquations from Linear Algebra, that is, the derivatives of Equation (20) give the n equations $\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b}$. The method of least squares is popular since the derivative of a square is linear.

What if columns of $\mathbf{A}$ are not independent? Which $\hat{\mathbf{x}}$ is best? Projecting $\mathbf{b}$ to $\mathbf{p}$, an equation with no solution will become an equation with infinitely many solutions. Using SVD, the "pseudoinverse" of $\mathbf{A}$ will choose the shortest solution to $\mathbf{A}\hat{\mathbf{x}} = \mathbf{p}$, that is arbitrarily choosing the nullspace component of the solution $\mathbf{x}^+$ to be zero. When $\mathbf{A}$ has independent columns, nullspace only contains the zero vector and the pseudoinverse is the usual left inverse $\mathbf{L} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$. We will look into this later when we discuss about SVD.

7 Another kind of factorization: A = QR (Orthogonal matrices and Gram-Schmidt)

Previously we have seen that $\mathbf{A}$ can be factorized into $\mathbf{LU}$ in Equation (10), such that the factorization expresses the elimination process since $\mathbf{L}$ that connects $\mathbf{A}$ to $\mathbf{U}$ contains the inverses of elimination matrices. Now we look at another factorization $\mathbf{A} = \mathbf{QR}$, where $\mathbf{Q}$ consists of orthonormal column vectors.

The vectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$ are orthonormal if

$$\mathbf{q}_i^T\mathbf{q}_j = \begin{cases} 0 & \text{when } i \neq j \text{ (orthogonal vectors)} \\ 1 & \text{when } i = j \text{ (unit vectors: } ||\mathbf{q}_i|| = 1) \end{cases} \tag{21}$$

Since $\mathbf{q}_i$ are independent from each other,

$$\mathbf{Q} = \begin{bmatrix}\mathbf{q_1} & \dots & \mathbf{q_n}\end{bmatrix} \tag{22}$$

and

$$\mathbf{Q}^T\mathbf{Q} = \begin{bmatrix}\mathbf{q}_1^T \\ \\ \vdots \\ \\ \mathbf{q}_n^T\end{bmatrix} \begin{bmatrix}\mathbf{q}_1 & \dots & \mathbf{q}_n\end{bmatrix} = \begin{bmatrix}1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1\end{bmatrix} \tag{23}$$

Note that $\mathbf{Q}$ does not have to be square, but it is always true that $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$. If $\mathbf{Q}$ is rectangular, $\mathbf{Q}^T$ is only an inverse from the left. But if $\mathbf{Q}$ is square, then it is also true that $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$, so $\mathbf{Q}^T$ is the two-sided inverse of $\mathbf{Q}$ such that $\mathbf{Q}^T = \mathbf{Q}^{-1}$, and in this case $\mathbf{Q}$ is called an orthogonal matrix. Multiplication by any orthogonal matrix $\mathbf{Q}$ preserves the lengths and angles (i.e. rotation, permutation and reflection). $\mathbf{Q}$ also preserves dot products since $(\mathbf{Q}\mathbf{x})^T(\mathbf{Q}\mathbf{y}) = \mathbf{x}^T\mathbf{Q}^T\mathbf{Q}\mathbf{y} = \mathbf{x}^T\mathbf{y}$, hence the property $\mathbf{Q}^{\mathbf{Q}} = \mathbf{I}$ is very convinient in calculations. Then here comes the question: What calculations have been made easy by having orthogonal matrices?

Consider a projection onto a column space, and the basis vectors are orthonormal, so $\mathbf{A}^T\mathbf{A} = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}$. Then we have the projection matrix

$$\mathbf{P} = \mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T = \mathbf{Q}\mathbf{Q}^T \tag{24}$$

Then substituting A with Q in the normal equation, we have

$$\mathbf{Q}^T\mathbf{Q}\hat{\mathbf{x}} = \mathbf{Q}^T\mathbf{b} \tag{25}$$

$$\mathbf{I}\cdot\hat{\mathbf{x}} = \mathbf{Q}^T\mathbf{b} \tag{26}$$

$$\hat{x}_i = \hat{q}_i^T\mathbf{b} \tag{27}$$

so we observe that there is no inversion involved and $\hat{x}_i$ is just a dot product. Also, if $\mathbf{Q}$ is square, the projection matrix is the identity matrix, which means the subspace is the whole space and the projection of $\mathbf{b}$ onto the whole space is $\mathbf{b}$ itself. The main takeaway from this part is that: orthogonal is good. But we often only have matrix $\mathbf{A}$ that may not have orthonormal columns. So in order to work with orthonormal vectors, we need to convert columns of $\mathbf{A}$ into orthonormal vectors. While there are better ways of conversion such as by means of reflection or rotation, Strang has chosen to explain it via Gram-Schmidt so that the factorization $\mathbf{A} = \mathbf{Q}\mathbf{R}$ can be understood. After setting the first orthonormal vector of $\mathbf{Q}$, Gram-Schmidt constructs the rest of the orthonormal vectors by projecting the n -th column of $\mathbf{A}$ into a space spanned by the previous columns of $\mathbf{A}$ and finds the new orthogonal vector by deducting the projection from the n -th column of $\mathbf{A}$ and finally normalizes the orthogonal vector. The non-involvement of later vectors became the key point of Gram-Schmidt such that $\mathbf{A}$ and $\mathbf{Q}$ is connected by a triangular matrix $\mathbf{R}$:

$$\mathbf{A} = \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n\end{bmatrix} \tag{28}$$

$$= \begin{bmatrix}\mathbf{q}_1 & \mathbf{q}_2 & \dots & \mathbf{q}_n\end{bmatrix} \begin{bmatrix}\mathbf{q}_1^T\mathbf{a}_1 & \mathbf{q}_1^T\mathbf{a}_2 & \dots & \mathbf{q}_1^T\mathbf{a}_n \\ & \mathbf{q}_2^T\mathbf{a}_2 & \dots & \mathbf{q}_1^T\mathbf{a}_n \\ & & \ddots & \vdots \\ & & & \mathbf{q}_n^T\mathbf{a}_n\end{bmatrix} \tag{29}$$

$$= \mathbf{Q}\mathbf{R}. \tag{30}$$

Multiplyting $\mathbf{Q}^T$ to both sides, we get $\mathbf{R} = \mathbf{Q}^T\mathbf{A}$, which is an upper triangular since the later $\mathbf{q}$'s are orthogonal to earlier $\mathbf{a}$'s. Any m by n matrix $\mathbf{A}$ with independent columns can be factored into $\mathbf{A} = \mathbf{QR}$ and $\mathbf{A}$ has the same column space as $\mathbf{Q}$. Using this result in least squares, we have $\mathbf{A}^T\mathbf{A} = (\mathbf{QR})^T\mathbf{QR} = \mathbf{R}^T\mathbf{Q}^T\mathbf{QR} = \mathbf{R}^T\mathbf{R}$. So the least squares equation simplifies to $\mathbf{R}^T\mathbf{R}\hat{\mathbf{x}} = \mathbf{R}^T\mathbf{Q}^T\mathbf{b}$, so $\mathbf{R}\hat{\mathbf{x}} = \mathbf{Q}^T\mathbf{b}$. Then instead of solving the unsolvable $\mathbf{Ax} = \mathbf{b}$, we solve for $\hat{\mathbf{x}} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{b}$.

What if $\mathbf{A}$ has dependent columns?

8 A hint of determinants and what we have so far before we proceed...

The determinant is a single number that tells us something about the matrix, for example, whether the matrix is invertible since determinant equals zero when there is no inverse. Using Strang's words, the determinant shows how $\mathbf{A}^{-1}$ changes as $\mathbf{A}$ changes. Determinants can be found in 3 ways, which will not be elaborated here:

1. The pivot formula, such that $det(\mathbf{A}) = \pm$(product of pivots).
2. The big formula, such that $det(\mathbf{A}) = \sum_{n!} det(\mathbf{P})(a_{1\alpha}a_{2\beta}\dots a_{n\omega})$.
3. The cofactor formula, such that $det(\mathbf{A}) = a_{i1}C_{i1} + a_{i2}C_{i2} + \dots + a_{in}C_{in}$ where $C_{ij} = (-1)^{i+j}det(M_{ij})$.

The determinant is defined by $det(\mathbf{I}) = 1$, sign reversal after row exchanges and linearity in each row, and from here we can also find the 10 properties of determinants. Among two remarkable properties are:

- $det(\mathbf{AB}) = det(\mathbf{A})det(\mathbf{B})$
- $det(\mathbf{A}^T) = det(\mathbf{A})$

Applications of determinants include but not exhaustive:

- Cramer's Rule, which solves $\mathbf{Ax} = \mathbf{b}$ and find $\mathbf{A}^{-1}$ by algebra using determinants.
- When edges of a box are rows of $\mathbf{A}$, then volume of box is $|det(\mathbf{A})|$.
- For n special numbers λ (eigenvalues, which we will look into in the coming section), $det|\mathbf{A} - \lambda\mathbf{I}| = 0$.

Before we proceed, let's summarize what we have so far. Consider $\mathbf{A}$ with independent columns. If $\mathbf{A}$ is square, then we can factorize $\mathbf{A}$ into

$$\mathbf{A} = \mathbf{LU} = \mathbf{QR}$$

and can easily solve for $\mathbf{x}$ in $\mathbf{Ax} = \mathbf{b}$. If $\mathbf{A}$ is long and thin, then we project $\mathbf{b}$ into $C(\mathbf{A})$ by multiplying $\mathbf{A}^T$ to both sides of $\mathbf{Ax} = \mathbf{b}$ and solve for $\hat{\mathbf{x}}$. This entire part with $\mathbf{Ax} = \mathbf{b}$ all the while is about balance, equilibrium and steady state. We will now move on to the next part that involves change, that is the involvement of time.

9 What is in the nullspace of $\mathbf{A} - \lambda\mathbf{I}$? What is this space? [Eigenvalues and Eigenvectors]

When time enters the picture, we will have difference equations to represent changes at different time steps and differential equations to represent changes in continuous time. These equations are not solved by elimination. The motivation of taking in the new ideas involving eigenvalues and eigenvectors is to avoid the complications presented by matrix $\mathbf{A}$. Almost all vectors change direction when they are multiplied by $\mathbf{A}$. If a solution vector stays in the direction of a fixed vector $\mathbf{x}$, then we only need to find the number that is changing with time that multiplies $\mathbf{x}$. That is

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \tag{31}$$

and the number λ is the eigenvalue of $\mathbf{A}$, which tells us whether the eigenvector $\mathbf{x}$ is stretch or shrunk or reversed or left unchanged when multiplied by $\mathbf{A}$. Eigenvalues and eigenvectors give a great way to understand the powers of a matrix. Suppose $\mathbf{A}$ contains information of change at one time step, then we will have $\mathbf{A}^{100}$ at the 100-th time step. If we know the eigenvalue of $\mathbf{A}$, that is λ, then instead of multiplying the matrix $\mathbf{A}$ a hundred times, we can set the number λ to the power of 100 and multiply it to the vector $\mathbf{x}$. When $\mathbf{A}$ is squared, eigenvectors stay the same and eigenvalues are squared. Continuing the pattern a hundred times, the eigenvectors stay in their own directions. But how do we find the eigenvectors and eigenvalues? This is where determinants come into action.

Consider an $n \times n$ matrix $\mathbf{A}$. The search of eigenvalues and eigenvectors begins by solving Equation (31),

$$\begin{aligned} \mathbf{A}\mathbf{x} &= \lambda\mathbf{x} \\ \mathbf{A}\mathbf{x} - \lambda\mathbf{x} &= 0 \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} &= 0. \end{aligned}$$

From here, we can observe that the eigenvectors make up the nullspace of $\mathbf{A} - \lambda\mathbf{I}$. If $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ has non-zero solution, then $\mathbf{A} - \lambda\mathbf{I}$ is not invertible and hence $det(\mathbf{A} - \lambda\mathbf{I}) = 0$. By solving for this determinant, the resulting n roots are the n eigenvalues of $\mathbf{A}$, and these eigenvalues make $\mathbf{A} - \lambda\mathbf{I}$ singular. Then for each λ, solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ to find an eigenvector $\mathbf{x}$. Note that λ is an eigenvalue of $\mathbf{A}$ if and only if $\mathbf{A} - \lambda\mathbf{I}$ is singular. When $\mathbf{A}$ is singular, $\mathbf{A}\mathbf{x} = 0\mathbf{x}$ has solutions and $\lambda = 0$ is one of the eigenvalues. So the solutions are the eigenvectors for $\lambda = 0$. But $det(\mathbf{A} - \lambda\mathbf{I}) = 0$ is the way to find all eigenvalues and eigenvectors. If $\mathbf{A}$ is singular, the eigenvectors for $\lambda = 0$ fill the nullspace, since $\mathbf{A}\mathbf{x} = 0\mathbf{x} = 0$. If $\mathbf{A}$ is invertible, then zero is not an eigenvalue, we shift $\mathbf{A}$ by a multiple of $\mathbf{I}$ to make it singular. So we can say that $\mathbf{A} - \lambda\mathbf{x}$ is a space that consist of vectors that will not change directions when multiplied by a λ. This is the eigenspace, where the space is generated by eigenvectors that correspond to the same eigenvalue, that is, the space of all vectors that are linear combinations of those eigenvectors. It is crucial to take note here that, repeated eigenvalues for a matrix is possible, and may very well be the source of trouble. This may result in not being able to obtain a full set of eigenvectors, that is, we don't have a basis and we cannot write every $\mathbf{v}$ as a combination of eigenvectors. That would mean that we cannot diagonalize a matrix without n independent eigenvectors, which we will talk about later.

It seems appropriate to bring in determinant and trace at this point to relate to eigenvalues. While they don't help to make the calculation for λ easier, they can help in checking the values if the computation went wrong. Apparently the product of the n eigenvalues equals the determinant of $\mathbf{A}$, and the sum of the n eigenvalues equals the sum of the n diagonal entries, which is the trace. Then Strang left a question for his readers to ponder on the best matrices for finding eigenvalues - triangular matries: "Why do the eigenvalues of a triangular matrix lie along its diagonal?". Elimination does not preserve λ, so if the diagonal entries of $\mathbf{U}$, that is the pivots are the eigenvalues of $\mathbf{U}$, they are not eigenvalues of $\mathbf{A}$. Another point to take note here is that given an eigenvalue λ for $\mathbf{A}$ and an eigenvalue β for $\mathbf{B}$, eigenvalue for $\mathbf{AB}$ is not necessarily $\lambda \times \beta$. This is because $\mathbf{A}$ and $\mathbf{B}$ might not necessarily share the same eigenvector $\mathbf{x}$. For the same reason, eigenvalues of $\mathbf{A} + \mathbf{B}$ is not necessarily $\lambda + \beta$. If all n eigenvalues are shared between $\mathbf{A}$ and $\mathbf{B}$, then we can multiply their eigenvalues. So $\mathbf{A}$ and $\mathbf{B}$ share the same n independent eigenvectors if and only if $\mathbf{AB} = \mathbf{BA}$.

Going back to the $\mathbf{A}^{100}$ example, we want to have a clearer picture of what $\mathbf{A}$ does to $\mathbf{x}$. Now we know that a $n \times n$ $\mathbf{A}$ has n eigenvalues λ_i and each λ_i has an eigenspace consisting of the corresponding eigenvectors. Suppose for each n eigenvalue, we choose an eigenvector, then each column vector of $\mathbf{A}$ is a combination of eigenvectors, such that when the column vector of $\mathbf{A}$ is multiplied by $\mathbf{A}$ again, each eigenvector is multiplied by its eigenvalue. So when we do $\mathbf{A}^{100}$, we times λ_1^{100} to the first eigenvector, λ_2^{100} to the second eigenvector and so on. Hence, if $\lambda_i = 1$, the eigenvector remains the same even after the 100th multiplication, which we call the "steady state". If $0 < \lambda_i < 1$, then the eigenvector will gradually decay with each multiplication. Therefore when $\mathbf{A}$ has steady state eigenvalues and decaying eigenvalues, at higher power of $\mathbf{A}$ the eigenvector that corresponds to a decaying eigenvector diminishes and since the column vector of $\mathbf{A}$ is a combination of the eigenvectors, the column vector will approach a steady state.

Here we list a few cases with their possible eigenvalues (the full list is given by Strang in his textbook, page 363):

- A projection matrix has eigenvalues $\lambda = 1$ and $\lambda = 0$, such that a projection keeps the column space and destroys the nullspace.
- A reflection matrix has $\lambda = 1$ and $\lambda = -1$.
- A Markov matrix has $\lambda = 1$ as one of the eigenvalues.
- A singular matrix has $\lambda = 1$ as one of the eigenvalues.

Special properties of a matrix lead to special eigenvalues and eigenvectors. It is possible that a matrix with real values has complex eigenvalues and eigenvectors, for example the rotation matrix $\mathbf{Q} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ that has eigenvalues $\lambda_1 = i$, $\lambda_2 = -i$, trace equals zero and $det(\mathbf{Q}) = \lambda_1\lambda_2 = 1$. The eigenvalues of $\mathbf{Q}$ tells us that $\mathbf{Q}$ is an orthogonal matrix so each $|\lambda| = 1$ and $\mathbf{Q}$ is a skew-symmetric matrix so each λ is pure imaginary. Strang gave the following summary:

- A symmetric matrix ($\mathbf{S}^T = \mathbf{S}$) can be compared to a real number.
- A skew-symmetric matrix ($\mathbf{A}^T = -\mathbf{A}$) can be compared to an imaginary number.
- An orthogonal matrix ($\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$) corresponds to a complex number with $|\lambda| = 1$.

The eigenvalues for these cases are perpendicular. We will discuss further in the coming sections.

To be continued...

10 Symmetric matrices and Positive definiteness

11 Singular Value Decomposition (SVD)

12 Principal Component Analysis (PCA)

13 Linear Transformations