The Fundamental Theorem of Linear Algebra

Alexey Grigorev
Technische Universität Berlin
grigorev@campus.tu-berlin.de

1. INTRODUCTION

In this report we discuss a paper "The Fundamental Theorem of Linear Algebra" by Gilbert Strang [3]. This paper is about the four subspaces of a matrix and the actions of the matrix are illustrated visually with pictures. The paper describes the "Strang's diagram", a diagram that shows actions of A, an $m \times n$ matrix, as linear transformations from the space $\mathbb{R}^m$ to $\mathbb{R}^n$. The diagram helps to understand the fundamental concepts of Linear Algebra in terms of the four subspaces by visually illustrating the actions of A on all these subspaces.

The goal of this paper is to present these concepts "in a way that students won't forget". The problem that the author faced is that students have difficulties understanding Linear Algebra. He proposes to solve this problem with the aforementioned diagrams.

There are four parts of the Fundamental Theorem of Linear Algebra: **part 1**, the dimensions of the subspaces; **part 2**, the orthogonality of the subspaces; **part 3**, the basis vectors are orthogonal; **part 4**, the matrix with respect to these bases is orthogonal. In this report, we discuss **part 1** and **part 2** only, and describe two diagrams: the solutions to a system of linear equations $A\mathbf{x} = \mathbf{b}$ and the Least Squares equations. We believe that it should give sufficient understanding to proceed with **part 3** and **part 4**, described in the paper. Additionally, in this report we elaborate some proofs from the paper and illustrate the concepts with examples.

1.1 Notation

In this report we use the following notation: Greek lower case letter $\alpha, \beta,$ ... are used for scalars, bold letters $\mathbf{b}, \mathbf{x},$ ... – vectors, lowercase indexed letters $x_1, x_2,$ ... – components of vectors, capital letters A – matrices, indexed bold letters $\mathbf{a}_1, \mathbf{a}_2,$ ... , $\mathbf{r}_1, \mathbf{r}_2,$ ... – columns or rows of a matrix. $\mathbf{0}$ is a vector of appropriate dimensionality with 0 in each component.

2. FUNDAMENTAL SUBSPACES AND DIMENSIONALITY

A *vector space* over real numbers $\mathbb{R}$ is a set where addition and scalar multiplication operations are defined in such a way that certain axioms, such as associativity, commutativity and distributivity are satisfied [2]. A *subspace* is a subset of some vector space such that the subset is closed under addition and scalar multiplication, i.e. given some subspace S, if $\mathbf{x}, \mathbf{y} \in S$ then for any $\alpha, \beta \in \mathbb{R}$, $(\alpha\mathbf{x} + \beta\mathbf{y}) \in S$.

For the matrix $A \in \mathbb{R}^{m \times n}$ there are four fundamental subspaces [1]:

- $C(A)$: the *column space* of A, it contains all linear combinations of the columns of A
- $C(A^T)$: the *row space* of A, it contains all linear combinations of the rows of A (or, columns of A^T)
- $N(A)$: the *nullspace* of A, it contains all solutions to the system $A\mathbf{x} = \mathbf{0}$
- $N(A^T)$: the *left nullspace* of A, it contains all solutions to the system $A^T\mathbf{y} = \mathbf{0}$.

All of them are subspaces because they are closed under addition and scalar multiplication.

2.1 Dimensionality

These subspaces have the following dimensions: $\dim C(A) = \dim C(A^T) = r$, where r is the rank of A; $\dim C(A^T) + \dim N(A) = n$, i.e. $\dim N(A) = n - r$. Also, $\dim C(A) + \dim N(A^T) = m$, i.e. $\dim N(A) = m - r$.

After applying Gaussian elimination for A with rank r, in the result we get r independent rows and the rest $m-r$ rows are all set to $\mathbf{0}$. Because of this, only r columns have non-zero entries in the pivot position, and thus $\dim C(A^T) = \dim C(A) = r$. The rest $n - r$ columns have no pivots and correspond to free variables. The basis of $N(A)$ is formed by $n - r$ "special" solutions to $A\mathbf{x} = \mathbf{0}$: we take free variables $x_{r+1}, x_{r+2},$ ... , x_n and assign them some values, making it possible to solve the system for remaining $x_1, x_2,$ ... , x_r variables. It is possible to choose only $n-r$ linearly independent solutions, and, hence, $\dim N(A) = n - r$. The same is true for A^T, thus, it's true for $N(A^T)$.

2.2 Orthogonality

Two vectors are *orthogonal* if their dot product produces 0. If all vectors of one subspace are orthogonal to all vectors of another subspace, these subspaces are called *orthogonal*.

PROPOSITION 1. *The row space $C(A^T)$ and the nullspace $N(A)$ of A are orthogonal. The column space $C(A)$ and the left nullspace $N(A^T)$ are also orthogonal.*

PROOF. Consider an $m \times n$ matrix A. Let $\mathbf{r}_1, \ \dots \ , \mathbf{r}_m$ be the rows of A. The row space $C(A^T)$ is formed by all linear combinations of rows, i.e. it is $\alpha_1\mathbf{r}_1 + \ \dots \ + \alpha_m\mathbf{r}_m$ for all possible choices of $\alpha_1, \ \dots \ , \alpha_m$. The nullspace $C(A)$ is formed by all the solutions $\mathbf{x}$ to the system $A\mathbf{x} = \mathbf{0}$.

Let us take any vector $\mathbf{r} \in C(A^T)$. Because $\mathbf{r} \in C(A^T)$, it can be expressed as $\mathbf{r} = \alpha_1\mathbf{r}_1 + \ \dots \ + \alpha_m\mathbf{r}_m$. We also can take any vector $\mathbf{n} \in N(A)$, and because $\mathbf{n} \in N(A)$, we know that $A\mathbf{n} = \mathbf{0}$. By the matrix-vector multiplication rule, the ith component of $A\mathbf{n}$ is $\mathbf{r}_i^T\mathbf{n}$ and since $A\mathbf{n} = \mathbf{0}$, $\mathbf{r}_i^T\mathbf{n} = 0$ for all $i = 1 \ .. \ m$.

Now let us consider $\mathbf{r}^T\mathbf{n}$: $\mathbf{r}^T\mathbf{n} = \left(\alpha_1\mathbf{r}_1 + \ \dots \ + \alpha_m\mathbf{r}_m\right)^T\mathbf{n} = \alpha_1\mathbf{r}_1^T\mathbf{n} + \ \dots \ + \alpha_m\mathbf{r}_m^T\mathbf{n} = \alpha_1 0 + \ \dots \ + \alpha_m 0 = 0$. Thus, $\mathbf{r}$ and $\mathbf{n}$ are orthogonal, and since they are chosen arbitrarily, it holds for all $\mathbf{r} \in C(A^T)$ and $\mathbf{n} \in N(A)$.

The same is true for $C(A)$ and $N(A^T)$. To show this, it is enough to transpose the matrix A. □

If two spaces are orthogonal and they together span the entire space, they are called *orthogonal compliments*. $C(A^T)$ and $N(A)$ are orthogonal compliments as well as $C(A)$ and $N(A^T)$. We can illustrate this with a picture (see fig. 1): the row space $C(A^T)$ and the nullspace $N(A)$ are orthogonal and meet only in the origin. They together span the space $\mathbb{R}^n$. $C(A)$ and $N(A^T)$ are also orthogonal and they together span $\mathbb{R}^m$.

3. SOLUTION TO $A\mathbf{x} = \mathbf{b}$

3.1 The row space solution

Now we consider a system $A\mathbf{x} = \mathbf{b}$. The general solution is $\mathbf{x} = \mathbf{x}_p + \mathbf{x}_n$, where $\mathbf{x}_p$ is some solution to $A\mathbf{x} = \mathbf{b}$, and $\mathbf{x}_n$ is the homogenous solution to $A\mathbf{x} = \mathbf{0}$, because $A\mathbf{x} = A(\mathbf{x}_p + \mathbf{x}_n) = A\mathbf{x}_p + A\mathbf{x}_n = \mathbf{b} + \mathbf{0} = \mathbf{b}$.

Since $C(A^T)$ and $N(A)$ are orthogonal compliments, they span the entire space $\mathbb{R}^n$ and every $\mathbf{x} \in \mathbb{R}^n$ can be expressed as $\mathbf{x} = \mathbf{x}_r + \mathbf{x}_n$ such that $\mathbf{x}_r \in C(A^T)$ and $\mathbf{x}_n \in N(A)$.

PROPOSITION 2. *$\mathbf{x}_r$ is unique.*

PROOF. Suppose there's another solution $\mathbf{x}_r' \in C(A^T)$. Since $C(A^T)$ is a subspace, it's close under subtraction, so $(\mathbf{x}_r - \mathbf{x}_r') \in C(A^T)$. Let's multiply the difference by A: $A(\mathbf{x}_r - \mathbf{x}_r') = A\mathbf{x}_r - A\mathbf{x}_r' = \mathbf{b} - \mathbf{b} = \mathbf{0}$. So $(\mathbf{x}_r - \mathbf{x}_r') \in N(A)$ and $(\mathbf{x}_r - \mathbf{x}_r') \in C(A^T)$. Since $C(A^T)$ and $N(A)$ are orthogonal compliments, the only place where they meet is in $\mathbf{0}$, so $\mathbf{x}_r - \mathbf{x}_r' = \mathbf{0}$ or $\mathbf{x}_r = \mathbf{x}_r'$. In other words, $\mathbf{x}_r$ is indeed unique. □

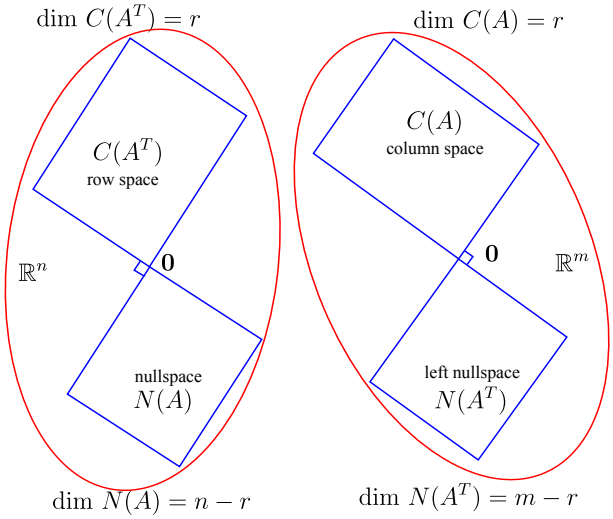

Figure 1: The row space $C(A^T)$ and the null space $N(A)$ are orthogonal compliments in $\mathbb{R}^n$. The columns space $C(A)$ and the left nullspace $N(A^T)$ are orthogonal compliments in $\mathbb{R}^m$.

As we mentioned earlier, there are many possible choices of $\mathbf{x}_p$. Among all these choices of $\mathbf{x}_p$ there's once special choice $\mathbf{x}_r$ – the *row space solution* to the system. It's special because it belongs to the row space, and it's unique. So any solution to $A\mathbf{x} = \mathbf{b}$ can be written as a combination $\mathbf{x} = \mathbf{x}_r + \mathbf{x}_n$.

3.2 Existence

A solution to $A\mathbf{x} = \mathbf{b}$ exists only if $\mathbf{b} \in C(A)$, i.e. when $\mathbf{b}$ is a linear combination of columns of A.

Let $\mathbf{a}_i$ be the columns of A, i.e. $A = \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix}$. Then for the solution to exist, there must exist $(x_1, \ \dots \ , x_n)$ s.t. $x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \ \dots \ + x_n\mathbf{a}_n = \mathbf{b}$. If these $(x_1, \ \dots \ , x_n)$ exist, they form a solution $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. Note that $x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \ \dots \ + x_n\mathbf{a}_n = \mathbf{b}$ is the same as writing $A\mathbf{x} = \mathbf{b}$.

We can illustrate this with a diagram (see fig. 2): $\mathbf{b} \in C(A)$, so there is a solution $\mathbf{x}$ to the system. The solution $\mathbf{x}$ can be expressed as $\mathbf{x}_r + \mathbf{x}_n$ s.t. $\mathbf{x}_r \in C(A^T)$ and $\mathbf{x}_n \in N(A)$; both $A\mathbf{x}_r = \mathbf{b}$ and $A\mathbf{x} = \mathbf{b}$, and it is shown with arrows to $\mathbf{b}$.

3.3 Example

Consider a system with $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$.

Let us first find the column space $C(A)$: it is formed by linear combinations $\alpha_1 \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \alpha_3 \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$ for all possible choices of $(\alpha_1, \alpha_2, \alpha_3)$.

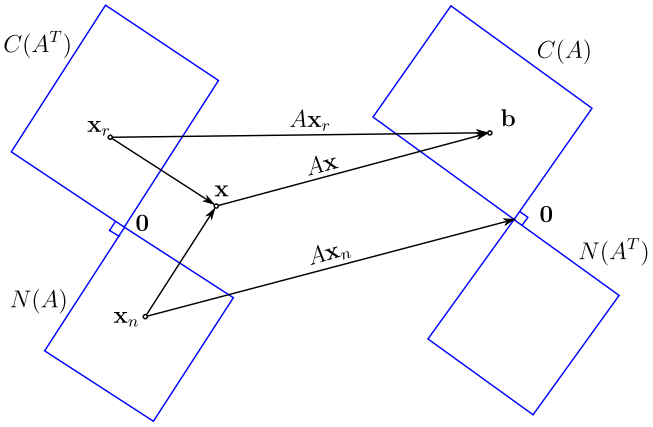

Figure 2: The general solution $\mathbf{x}$ to the system $A\mathbf{x} = \mathbf{b}$ consists of two components $\mathbf{x}_r \in C(A^T)$ and $\mathbf{x}_n \in N(A)$. The solution to the system exists because $\mathbf{b} \in C(A)$.

To check if the system $A\mathbf{x} = \mathbf{b}$ has a solution, we need to show that $\mathbf{b} \in C(A)$. In other words, we need to show that it is possible to find such $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ that $x_1 \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} + x_2 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + x_3 \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$. In this example $x_1 = -1, x_2 = 1$ and $x_3 = 0$: $-1 \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} + 1 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + 0 \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$. Note that here we not only established that $\mathbf{b} \in C(A)$, but also found a solution $\mathbf{x}_p = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$. This solution is not necessarily the row space solution, i.e. it may not belong to the row space $C(A^T)$.

The nullspace $N(A)$ contains all the solutions to $A\mathbf{x} = \mathbf{0}$. To find them, we use Gaussian Elimination and transform A to the Row-Reduced Echelon Form: $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix}$. There are two pivot variables x_1 and x_2: they have 1 at the pivot position, and there is one free variable x_3 that doesn't have a pivot: it has 0 on this position. The free variable can take any value, for example, we can assign $x_3 = 1$. Then we have $x_n = \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}$, we solve the system and obtain $x_1 = 1, x_2 = -2$, so the solution is $x_n = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$. There is nothing special about the choice $x_3 = 1$, so instead we can choose $x_3 = \alpha$, and obtain $x_n = \begin{bmatrix} \alpha \\ -2\alpha \\ \alpha \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$. All possible choices of α form the nullspace $N(A)$.

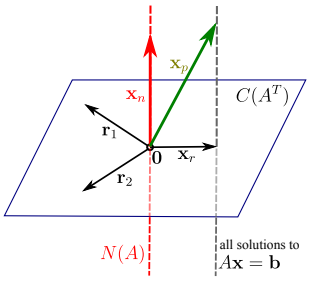

Figure 3: The basis of $C(A^T)$ is two rows of A, $N(A)$ is orthogonal to $C(A^T)$, and the set of all solutions $\mathbf{x}$ is just shifted $N(A)$. The row space solution $\mathbf{x}_r$ belongs to $C(A^T)$ and it's a solution to the system.

The complete solution $\mathbf{x}$ is a sum of some solution and the homogenous solutions: $\mathbf{x} = \mathbf{x}_p + \mathbf{x}_n = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} + \alpha \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$. Note that the set of all solutions $\mathbf{x}$ is just a shifted nullspace $N(A)$ (see fig. 3).

Because the rank of A is two, the row space $C(A^T)$ contains only two linearly independent vectors, so it is a plane in $\mathbb{R}^3$ formed by two rows $\mathbf{r}_1$ and $\mathbf{r}_2$. The row space solution $\mathbf{x}_r$ also belongs to $C(A^T)$, but the solution $\mathbf{x}_p = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$ does not (see fig. 3). If we want to find $\mathbf{x}_r$, at first we need to recognize that it's a projection of $\mathbf{x}_p$ onto $C(A^T)$. We will see how to find this projection in the next section.

4. THE LEAST SQUARES

Sometimes there is no solution to the system $A\mathbf{x} = \mathbf{b}$.

For example, consider a system $\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$. The column space of A is $C(A) = \alpha_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. But for this $\mathbf{b}$ it is not possible to find such α_1, α_2 that would produce $\mathbf{b}$: it is not a combination of columns of A, thus $\mathbf{b} \notin C(A)$ and therefore there is no solution to $A\mathbf{x} = \mathbf{b}$. What if we still need to find some solution to this system, not necessarily exact, but as good as possible?

4.1 Projection on column space

So the goal is to find some approximation $\hat{\mathbf{x}}$ such that $A\hat{\mathbf{x}} \in C(A)$. To do it, we need $\mathbf{p}$ to be as close as possible to the original $\mathbf{b}$. Such $\mathbf{p}$ is called a *projection* of $\mathbf{b}$ onto $C(A)$. Let $\mathbf{e} = \mathbf{b} - \mathbf{p}$ be the *projection error*. The projection error need to be as small as possible (see fig. 4).

PROPOSITION 3. *The projection error* $\mathbf{e}$ *is minimal, when it's perpendicular to* $C(A)$.

PROOF. Let $\mathbf{e} = \mathbf{b} - \mathbf{p}$ be perpendicular to $C(A)$ and consider another vector $\mathbf{p}' \in C(A)$, $\mathbf{p}' \neq \mathbf{p}$ such that $\mathbf{e}' = \mathbf{b} - \mathbf{p}'$

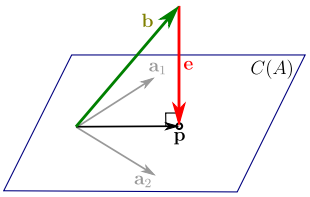

Figure 4: $\mathbf{b} \notin C(A)$, so there's no solution to $A\mathbf{x} = b$. $\mathbf{p} \in C(A)$ is a projection of $\mathbf{b}$ onto $C(A)$, so there exists a solution to $A\hat{\mathbf{x}} = \mathbf{p}$.

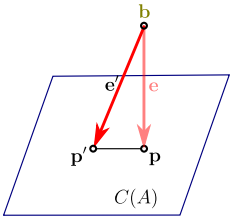

Figure 5: The projection error $\mathbf{e}$ is smallest when it's perpendicular to $C(A)$.

is not perpendicular to $C(A)$. Then, by the Pythagoras theorem, $\|\mathbf{e}'\|^2 = \|\mathbf{e}\|^2 + \|\mathbf{p} - \mathbf{p}'\|^2 > \|\mathbf{e}\|^2$, so $\|\mathbf{e}'\| > \|\mathbf{e}\|$ for any $\mathbf{e}' \neq \mathbf{e}$ (see fig. 5). Thus, $\mathbf{e}$ is smallest when it is perpendicular to $C(A)$. □

We need to find such $\hat{\mathbf{x}}$ that $\mathbf{e}$ is smallest. $\mathbf{e}$ is smallest when it's orthogonal to $C(A)$, i.e. to all vectors on $C(A)$: $\mathbf{a}_1^T\mathbf{e} = \mathbf{0}$ and $\mathbf{a}_2^T\mathbf{e} = \mathbf{0}$. We can write the same as $A^T\mathbf{e} = \mathbf{0}$. Since $\mathbf{e} = \mathbf{b} - \mathbf{p} = \mathbf{b} - A\hat{\mathbf{x}}$ and $A^T\mathbf{e} = \mathbf{0}$, we have $A^T(\mathbf{b} - A\hat{\mathbf{x}}) = \mathbf{0}$ or $A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$. This is called the *Normal Equation* and it minimizes the error $\mathbf{e}$. The *Least Squares solution* is $\hat{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$.

There is another way to arrive at the same solution using calculus. Suppose we want to minimize the sum of squared errors $\|\mathbf{e}\|^2$. So the goal is to find such $\mathbf{x}$ that minimizes $\|\mathbf{e}\|^2 = \|\mathbf{b} - A\mathbf{x}\|^2$. First, expand it as $\|\mathbf{b} - A\mathbf{x}\|^2 = (\mathbf{b} - A\mathbf{x})^T(\mathbf{b} - A\mathbf{x}) = \mathbf{b}^T\mathbf{b} - 2\mathbf{x}^TA^T\mathbf{b} + \mathbf{x}^TA^TA\mathbf{x}$. Now by taking the derivative w.r.t. $\mathbf{x}$ and we obtain $-2A^T\mathbf{b} + 2A^TA\hat{\mathbf{x}} = \mathbf{0}$ or $A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$. We come to the same conclusion, and because we minimized the squared error, this technique is called "Least Squares".

There is one additional condition: A^TA is invertible only if A has independent columns. A has independent columns when $N(A) = \{\ \mathbf{0}\ \}$, so it is enough to show that A^TA and A have the same nullspaces.

PROPOSITION 4. $N(A) \equiv N(A^TA)$

PROOF. In this proof, we show that both $N(A^TA) \subseteq N(A)$ and $N(A) \subseteq N(A^TA)$ hold at the same time, and hence $N(A) \equiv N(A^TA)$.

First, we prove that if $A^TA\mathbf{x} = \mathbf{0}$ then $A\mathbf{x} = \mathbf{0}$, i.e. $N(A^TA) \subseteq N(A)$. Suppose $\mathbf{x}$ is a solution to $A^TA\mathbf{x} = \mathbf{0}$. By multiplying it by $\mathbf{x}^T$ we get $\mathbf{x}^TA^TA\mathbf{x} = \mathbf{0}$. A dot product of vector

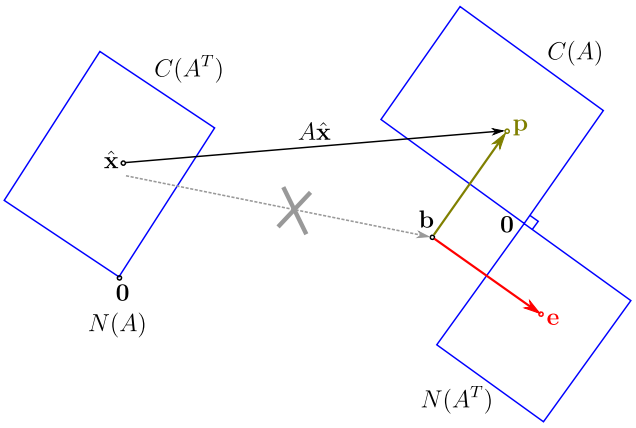

Figure 6: There is no solution to $A\mathbf{x} = \mathbf{b}$ because $\mathbf{b} \notin C(A)$, but the projection $\mathbf{p} \in C(A)$ and there's a solution to $A\hat{\mathbf{x}} = \mathbf{p}$. The projection error $\mathbf{e} \in N(A^T)$ and $N(A)$ is empty: it contains only 0.

with itself is a squared L_2 norm, so we have $\|A\mathbf{x}\|^2 = \mathbf{0}$. A vector can have length 0 only if it is a zero vector, so $A\mathbf{x} = \mathbf{0}$. Thus, $\mathbf{x}$ is a solution to $A\mathbf{x} = \mathbf{0}$ as well.

Next, we show that if $A\mathbf{x} = \mathbf{0}$ then $A^TA\mathbf{x} = \mathbf{0}$, i.e. $N(A) \subseteq N(A^TA)$. If $\mathbf{x}$ is a solution to $A\mathbf{x} = \mathbf{0}$, then by multiplying it by A^T on the left we get $A^TA\mathbf{x} = \mathbf{0}$.

Since $N(A^TA) \subseteq N(A)$ and $N(A) \subseteq N(A^TA)$, we conclude that $N(A) \equiv N(A^TA)$. □

This technique can be illustrated by the diagram as well (see fig. 6): $\mathbf{b}$ is not in $C(A)$, so we can't solve the system, but we can project $\mathbf{b}$ onto $C(A)$ to get $\mathbf{p}$ and then solve $A\hat{\mathbf{x}} = \mathbf{p}$. Note that $\mathbf{b} = \mathbf{p} + \mathbf{e}$, and $\mathbf{e} \in N(A^T)$. This is because to obtain the Normal Equation we solve $A^T\mathbf{e} = \mathbf{0}$, and the left nullspace $N(A^T)$ contains all the solutions to $A^T\mathbf{y} = \mathbf{0}$.

4.2 Example

Consider a system with $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$, and $\mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$. To solve $A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$, we first calculate $A^TA = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$. Then, we calculate $A^T\mathbf{b} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Now we solve the system $\begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}\hat{\mathbf{x}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and the solution is $\hat{\mathbf{x}} = \begin{bmatrix} 4/3 \\ -1/2 \end{bmatrix}$.

What if A did not have independent columns? Suppose $A = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{bmatrix}$. Then $A^TA = \begin{bmatrix} 3 & 6 \\ 6 & 12 \end{bmatrix}$. This matrix is singular, i.e. it doesn't have the inverse, and thus we cannot solve the system $A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$.

4.3 Application: OLS Regression

The Least Squares method is commonly used in Statistics and Machine Learning to find a best fit line for a given data data set. This method is called *OLS Regression* (*Ordinary Least Squares Linear Regression*) or just *Linear Regression*.

Linear Regression problem:

Given a dataset $\mathcal{D} = \{ (\mathbf{x}_i, y_i) \}$ of n pairs $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ we train a model that can predict y for new unseen data points $\mathbf{x}$ as good as possible. To do this, we fit a line $y = w_0 + w_1 x_1 + w_2 x_2 + \ \ldots \ + w_d x_d$. This is the *best fit line*, w_0 is the *intercept* coefficient, and $w_1, \ \ldots \ , w_n$ are the *slope* coefficients. Let $x_0 = 1$, so we can write $y = w_0 x_0 + w_1 x_1 + w_2 x_2 + \ \ldots \ + w_d x_d = \mathbf{w}^T \mathbf{x}$. Let $\mathbf{X} = \begin{bmatrix} - \ \mathbf{x}_1 \ - \\ - \ \mathbf{x}_2 \ - \\ \vdots \\ - \ \mathbf{x}_n \ - \end{bmatrix}$. This $\mathbf{X}$ is called the *data matrix* and its rows are formed by $\mathbf{x}_i$. $\mathbf{X}$ is a $n \times (d+1)$ matrix. Also let $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \in \mathbb{R}^{d+1}$ and $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$.

We need to solve the system $\mathbf{X}\mathbf{w} = \mathbf{y}$, but usually there is no solution, so we use the Normal Equation, and the Least Squares solution to this problem is given by $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Example. Consider a dataset $\mathcal{D} = \{(1,1),(2,0),(3,0)\}$. We add $w_0 = 1$ to each observation and have $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$. Let $\mathbf{X} = \begin{bmatrix} - \ \mathbf{x}_1 \ - \\ - \ \mathbf{x}_2 \ - \\ - \ \mathbf{x}_3 \ - \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$. Then $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 4/3 \\ -1/2 \end{bmatrix}$ (see fig. 7).

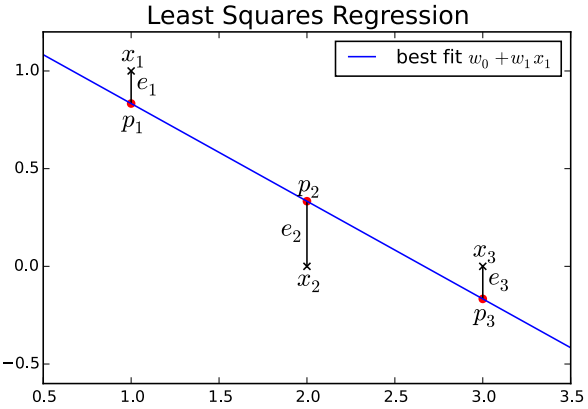

Figure 7: The best fit line with $w_0 = 4/3$ and $w_1 = -1/2$. x_1, x_2, x_3 are the data points, p_1, p_2, p_3 are OLS predictions, and e_1, e_2, e_3 are prediction errors.

5. CONCLUSION

The paper presents the "Strang's diagram" for helping students to understand Linear Algebra better. The purpose of this paper is educational, there is no novelty (the pictures had been presented earlier in the author's textbook [2]) and no research. Also, the claim that the pictures illustrate actions of the matrix "in a way they [the students] won't forget" is not supported by any statistical evaluation. And finally, the reader should already be familiar with concepts of Linear Algebra to understand the paper, and there are no supporting examples.

However, the presented diagram is indeed an effective tool for illustrating the four fundamental subspaces and their relation to the important concepts like orthogonality, solution existence, projections, all in one place in a concise form. It also helps to think of many Linear Algebra problems, such as solving the system $A\mathbf{x} = \mathbf{b}$ or finding the Least Squares solution, in terms of the subspaces and it is very beneficial for understanding these problems.

Lastly, the paper summarizes the author's textbook [2] and therefore reading the paper is a good way of refreshing the key concepts of Linear Algebra.

6. REFERENCES

[1] G. Strang. The four fundamental subspaces: 4 lines.

[2] G. Strang. *Linear Algebra and Its Applications*. Brooks Cole, 1988.

[3] G. Strang. The fundamental theorem of linear algebra. *American Mathematical Monthly*, pages 848–855, 1993.