

The Four Subspaces Theorem

These lecture notes discuss the “four subspaces theorem” from linear algebra. Gilbert Strang calls this theorem the fundamental theorem of linear algebra.

Let $A \in \mathbb{R}^{m \times n}$. There are four subspaces naturally associated with A : $N(A)$, $R(A)$, $N(A^T)$, and $R(A^T)$. The four subspace theorem states that

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

Here is a proof. Let a_i be the i th column of A , so

$$A = [a_1 \quad \cdots \quad a_n].$$

Then

$$\begin{aligned} x &\in N(A^T) \\ \iff A^T x &= 0 \\ \iff x &\text{ is orthogonal to each column of } A \\ \iff x &\text{ is orthogonal to the span of the columns of } A \\ \iff x &\in R(A)^\perp. \end{aligned}$$

This shows that $N(A^T) = R(A)^\perp$. Applying this result to the matrix A^T shows that $N(A) = R(A^T)^\perp$, or equivalently $N(A)^\perp = R(A^T)$.

(If A has complex entries, a similar proof shows that $N(A)^\perp = R(A^*)$ and $R(A)^\perp = N(A^*)$.)

The following picture illustrates the four subspaces theorem in the case where $A \in \mathbb{R}^{2 \times 2}$. In this picture, we assume $N(A)$ has dimension 1.

Let T be the linear transformation defined by $T(x) = Ax$. Unfortunately, T may be neither 1-1 or onto. However, when T is restricted to the $R(A^T)$, we obtain a 1-1, onto mapping from $R(A^T)$ to $R(A)$. To see that the restriction of T to $R(A^T)$ is 1-1, suppose that $x_1, x_2 \in R(A^T)$ and $T(x_1) = T(x_2)$. Then $A(x_1 - x_2) = 0$ which shows that $x_1 - x_2 \in N(A)$. But $x_1 - x_2 \in R(A^T)$. It follows that $x_1 - x_2 = 0$.

The fact that “row rank equals column rank” is often seen as unintuitive. But in these case where A has real entries, it’s not just that $R(A)$ and $R(A^T)$ have the same dimension. In this case, these two subspaces are *isomorphic*, and A itself provides the isomorphism.

The four subspaces picture allows us to give a conceptual definition of the pseudoinverse of A . When trying to solve $Ax = b$, two things might go wrong:

1. b might not be in the range of A , so $Ax = b$ has no solution. In this case, a natural thing to do is replace b with \hat{b} , the projection of b onto $R(A)$. So instead of solving $Ax = b$, we solve $Ax = \hat{b}$.
2. The null space of A might be non-trivial, which means that the system $Ax = \hat{b}$ is underdetermined. So which solution should we pick? While there may be many solutions to $Ax = \hat{b}$, there is a *unique* vector $\hat{x} \in R(A^T)$ that satisfies $A\hat{x} = \hat{b}$. The vector \hat{x} is the solution to $Ax = b$ which has least 2-norm.

The pseudoinverse of A , by definition, takes b as input and returns \hat{x} as output.

1 Lagrange multipliers

We'll now use the four subspaces theorem to derive some optimality conditions for optimization problems with equality constraints. Our goal is only to give an *intuitive* derivation of the optimality conditions, not a rigorous proof. We want to show how the optimality conditions can be discovered.

Warm-up problem: consider the problem of minimizing a differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, with no constraints. From calculus, we know that a *necessary* condition for x^* to be a local minimizer of f is $\nabla f(x^*) = 0$. (This condition is only necessary, and not sufficient.) One way to prove this is to note that if x^* is a local minimizer of f , then the directional $D_u f(x^*)$ is nonnegative for any vector $u \in \mathbb{R}^n$. If $D_u f(x^*)$ were negative for some vector u , then by moving a little bit in the direction of u , the value of f would decrease and x^* would not be a local minimizer. Now using the formula for the directional derivative, we see that

$$\langle \nabla f(x^*), u \rangle \geq 0 \quad \text{for all } u \in \mathbb{R}^n.$$

From this, it follows that $\langle \nabla f(x^*), u \rangle = 0$ for all $u \in \mathbb{R}^n$. This implies that $\nabla f(x^*) = 0$.

Now consider the problem

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && Ax = b. \end{aligned} \tag{1}$$

Here $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. Let x^* be a local minimizer for this problem.

Suppose that $Au = 0$ for some vector $u \in \mathbb{R}^n$. Then, when we move away from x^* in the direction of u , *we remain feasible*. Hence, $D_u f(x^*) \geq 0$. Otherwise, by moving a little bit in the direction of u , the value of f would decrease, and x^* would not be a local minimizer. Using the formula for the directional derivative, we see that if $Au = 0$, then $\langle \nabla f(x^*), u \rangle \geq 0$. This implies that

$$\text{If } Au = 0, \text{ then } \langle \nabla f(x^*), u \rangle = 0.$$

In other words, u is in the orthogonal complement of $N(A)$! Now, the four subspaces theorem tells us that $u \in R(A^T)$. Thus, there exists a vector $\lambda \in \mathbb{R}^m$ such that

$$\nabla f(x^*) = A^T \lambda. \quad (2)$$

This equation, together with the feasibility condition $Ax^* = b$, is our *necessary* condition for x^* to be optimal for problem (1).

Next consider the problem Now consider the problem

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && g(x) = 0. \end{aligned} \quad (3)$$

Here $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are continuously differentiable functions. Let x^* be a local minimizer for this problem.

This time the constraint is nonlinear. How do we handle that?

Note that when x is near x^* , the constraint function is *approximately* linear:

$$g(x) \approx \underbrace{g(x^*)}_0 + g'(x^*)(x - x^*). \quad (4)$$

If we had exact equality rather than approximate equality in (4), then the constraints would be truly linear, and we could use our previous result (for the case of linear constraints) to conclude that

$$\nabla f(x^*) = g'(x^*)^T \lambda \quad (5)$$

for some $\lambda \in \mathbb{R}^m$. This gives us reason to *hope* that (5) holds when x^* is a local minimizer for the problem (3) with nonlinear constraints.

(In order to reach this conclusion, it turns out that we need to make an extra assumption, for example that $g'(x^*)$ has full row rank. This assumption is called a “constraint qualification” in optimization.)

We could say more about how to make this argument rigorous, but here we just want to focus on the intuition and the linear algebra. More details and a rigorous proof can be found in chapter 12 of Numerical Optimization by Nocedal and Wright.