Math 54: Lecture 7/30/19

Jeffrey Kuan

July 30, 2019

Least Squares

Earlier in this class, we considered the problem of finding a solution to the system of equations $A\mathbf{x} = \mathbf{b}$. We called such a system **consistent** if there was at least one solution, and we called such a system **inconsistent** if it did not have a solution. We learned how to find all solutions of a consistent solution, and in the case of an inconsistent solution, we simply said there was no solution.

But what if we wanted to say more in the case of an inconsistent system $A\mathbf{x} = \mathbf{b}$? In this case, there is no vector $\mathbf{x}$ that gives $A\mathbf{x} = \mathbf{b}$, but can we maybe find some vector $\hat{\mathbf{x}}$ that gives the best possible approximation to a solution to $A\mathbf{x} = \mathbf{b}$, an "almost solution" of sorts?

This is the idea behind least squares. Let A be an m by n matrix. We will say that $\hat{\mathbf{x}}$ is a **least squares solution** or least squares best approximation to the system $A\mathbf{x} = \mathbf{b}$ if

$$||A\hat{\mathbf{x}} - \mathbf{b}|| \leq ||A\mathbf{x} - \mathbf{b}|| \text{ for all } \mathbf{x} \in \mathbb{R}^n$$

We call $||A\hat{\mathbf{x}} - \mathbf{b}||$ the least squares distance. Intuitively, this means that $\hat{\mathbf{x}}$ makes $A\hat{\mathbf{x}}$ as close as possible to $\mathbf{b}$, so that even if $A\hat{\mathbf{x}}$ is not equal to $\mathbf{b}$, the least squares solution has $A\hat{\mathbf{x}}$ as close to $\mathbf{b}$ as possible. This shortest distance is exactly the least squares distance.

Here are some important remarks about the least square solution.

- If $A\mathbf{x} = \mathbf{b}$ is a consistent system, then the set of least square solutions is just the set of solutions, since the left hand side in the above inequality can be as small as possible (equal to 0 in this case) precisely when $\hat{\mathbf{x}}$ is a solution to the consistent system.

- Note that there may be many possible least squares approximations to $A\mathbf{x} = \mathbf{b}$. This is precisely the case when the homogeneous $A\mathbf{x} = 0$ has infinitely many solutions, because if $\mathbf{x}_h$ is a solution to the homogeneous equation, then $A\mathbf{x} = A(\mathbf{x} + \mathbf{x}_h)$.

The definition of the least squares solution above says that to find a least square solution, we must minimize the distance between the fixed vector $\mathbf{b}$ and the subspace of all possible $A\mathbf{x}$, which as we already discussed earlier in the class is exactly the **column space of** A. So in particular, we want to minimize the distance from $\mathbf{b}$ to $\text{Col}(A)$. We can do this by using orthogonal projection.

In particular, we should set

$$A\hat{\mathbf{x}} = \text{proj}_{\text{Col}(A)}(\mathbf{b}) \tag{1}$$

Then, the set of all $\hat{x}$ that satisfy the above system is the set of least squares solutions to $A\mathbf{x} = \mathbf{b}$. We will call equation (1) the **associated least squares system for the system** $A\mathbf{x} = \mathbf{b}$. Solving the system of equations (1) gives all least squares solutions.

Note that by the definition of orthogonal projection, this equation also implies that

$$A\hat{\mathbf{x}} - \mathbf{b} \text{ is orthogonal to } \text{Col}(A)$$

In particular, this means the dot product of each column of A with $A\hat{\mathbf{x}} - \mathbf{b}$ is zero, or equivalently, the dot product of each row of A^t with $A\hat{\mathbf{x}} - \mathbf{b}$ is zero. So in particular, $\hat{\mathbf{x}}$ being a least squares solution to $A\mathbf{x} = \mathbf{b}$ is equivalent to

$$A^t(A\hat{\mathbf{x}} - \mathbf{b}) = 0$$

So the set of least squares solutions $\hat{\mathbf{x}}$ to $A\mathbf{x} = \mathbf{b}$ is also the solutions to the system

$$A^tA\mathbf{x} = A^t\mathbf{b} \tag{2}$$

The system of equations in (2) is called the **normal system of equations for the system** $A\mathbf{x} = \mathbf{b}$. It is easier to use the normal system rather than the associated system to calculate least squares solutions, since the normal equations involve just matrix multiplication with A^t while the associated system requires calculating an orthogonal projection. However, both would still give the same answer.

As a final note, it is interesting to ask when the least squares solution to $A\mathbf{x} = \mathbf{b}$ is unique. If such a least squares solution is unique, it means that the associated system (1) and the normal system (2) will have unique solutions. Using (2), this means that A^tA is invertible (we can talk about invertibility here since A^tA is a square matrix, though A might not necessarily be a square matrix). Thus, we discover the following important fact: $A\mathbf{x} = \mathbf{b}$ has a unique least squares solution precisely when the square matrix A^tA is invertible.

Problem 1

Find a least squares solution to the system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 1 \\ 0 & 2 & 1 & 0 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$$

What is the least squares distance?

Problem 2

Find a least squares solution to the system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 2 \\ 2 & -1 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$$

What is the least squares distance?

Spectral Theorem for Symmetric Matrices

Now that we know the ideas of orthonormality and inner products, we can now consider one of the most fundamental theorems in linear algebra, the spectral theorem for symmetric matrices. Recall that diagonalization is useful because it allows us to compute powers of a matrix easily, and it gives us a nice basis of eigenvectors in which it is easy to understand the action of a diagonalizable linear operator T. But it is a surprising fact that every symmetric matrix multiplication operator on $\mathbb{R}^n$ is diagonalizable, and more specifically, it is orthogonally diagonalizable.

A symmetric operator on $\mathbb{R}^n$ is a linear operator $T : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\langle Tv, w\rangle = \langle v, Tw\rangle$$

for every vector $v, w \in \mathbb{R}^n$. As you showed (or will show) on Problem Set 8, such a transformation can really be thought of as a matrix multiplication map where the matrix A that we are multiplying by is symmetric. So we will focus on symmetric matrix multiplication maps, given by

$$T(v) = Av$$

for $v \in \mathbb{R}^n$, where A is a symmetric n by n matrix. In matrix form, we have the above identity in the following form.

$$\langle Av, w\rangle = \langle v, Aw\rangle$$

for a symmetric n by n matrix A and $v, w \in \mathbb{R}^n$.

We will first show the following fundamental fact: Any two eigenvectors of a symmetric matrix A with different eigenvalues are orthogonal. To see this, if v and w are eigenvectors for distinct eigenvalues λ_1 and λ_2, then

$$\langle Av, w\rangle = \lambda_1\langle v, w\rangle = \langle v, Aw\rangle = \lambda_2\langle v, w\rangle$$

so since $\lambda_1 \neq \lambda_2$ and $\lambda_1\langle v, w\rangle = \lambda_2\langle v, w\rangle$, we have that $\langle v, w\rangle = 0$. So v and w are orthogonal.

This means that the eigenspaces W_{λ_1} and W_{λ_2} for distinct eigenvalues are orthogonal, and from our discussion of diagonalization, we know they are also linearly independent. In particular, two vectors from distinct eigenspaces are always orthogonal.

Therefore, if we calculate all of the eigenspaces for an n by n symmetric matrix A, we have that all of the eigenspaces are mutually orthogonal subspaces in $\mathbb{R}^n$. The last question that is left to answer is: Do all of the eigenspaces of the symmetric n by n matrix A span $\mathbb{R}^n$? Or equivalently, is A diagonalizable?

A fact is that every symmetric n by n matrix A is diagonalizable (so all geometric multiplicities are equal to the algebraic multiplicities for each eigenvalue), with n real eigenvalues (counted with multiplicity). This leads to the spectral theorem for symmetric matrices.

Spectral Theorem for Symmetric Matrices: Let A be an n by n symmetric matrix.

- A is diagonalizable, and has n real eigenvalues (counted with multiplicity, either algebraic or geometric since they are equal for a diagonalizable matrix).

- The eigenspaces of A for distinct eigenvalues are mutually orthogonal and span $\mathbb{R}^n$.
- There exists an orthonormal basis $v_1, v_2, ..., v_n$ of eigenvectors of A for $\mathbb{R}^n$.
- A can be diagonalized orthogonally as $A = ODO^t$, where D is a diagonal matrix of the eigenvalues and O is an orthogonal matrix.

Since the eigenspaces of A are mutually orthogonal, this means that eigenvectors for different eigenvalues are orthogonal and linearly independent from each other. So if we find an orthonormal basis for each eigenspace and combine all of these orthonormal bases together, then we get an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of A.

Recall that by diagonalization, we can write $A = PDP^{-1}$, where P is a matrix where the columns of P are the eigenvectors in the basis of eigenvectors. In this case, there is an orthonormal basis of eigenvectors which will then go into the columns of P. So the columns of P form an orthonormal basis for $\mathbb{R}^n$, and thus P by definition is an orthogonal matrix. Thus, $P^{-1} = P^t$. So if we denote the matrix P instead by O (since it is an orthogonal matrix), we have that $A = ODO^t$, where O is the change of basis matrix from the orthonormal basis for $\mathbb{R}^n$ to the standard basis.