Least Squares and the Normal Equations

The systems of equations $A\mathbf{x} = \mathbf{b}$ that we've considered thus far have been *consistent*, meaning they have at least one solution. We now consider methods for *inconsistent* systems, where the goal is to produce the vector $\hat{\mathbf{x}}$ such that $A\hat{\mathbf{x}}$ is as close to $\mathbf{b}$ as possible. Specifically, we mean "close" as measured by the 2-norm, and the goal is to minimize $||\mathbf{b} - A\hat{\mathbf{x}}||_2$. The solution $\hat{\mathbf{x}}$ is called a *least squares solution*, while the problem of finding it is called the *least squares problem*.

Two notes/reminders before we get into the linear algebra theory. One is that we are not assuming that A is square; in fact, in most applications A has more rows that columns (is $m \times n$ with $m > n$).

Second, we do know how to compute the 2-norm of a vector, as the square root of the sum of the squares of the elements. However, it is often convenient to instead compute the *squared error*, which is the sum of the squares of the elements (just omit the square root). The inputs which create the minimum value are the same in either case, so the choice will not change the answer.

Example 1. Compute the squared error and the 2-norm of the vector $\mathbf{u} = \begin{bmatrix} 5 \\ 3 \\ -1 \end{bmatrix}$.

SE $= 5^2 + 3^2 + (-1)^2 = 25 + 9 + 1 = 35$, and $||\mathbf{u}||_2 = \sqrt{35}$.

Another reminder: the 2-norm definition for vectors does NOT apply directly to matrices. You will not be asked to compute a matrix 2-norm unless by computer, as the definition requires linear algebra theory beyond the scope of this course.

Linear Algebra Review

I suggest for graphics that you check out this online linear algebra textbook. Their explanations and examples are also nice: Linear Algebra, Least Squares Section.

A key concept in the least squares problem is that of *orthogonality*.

Definition 1. Two vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v} = 0$. A set of vectors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \ldots\}$ is orthogonal if every pair of vectors within the set is orthogonal.

Example 2. Determine if the set $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is orthogonal for

$$\mathbf{u}_1 = \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix}, \mathbf{u}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \mathbf{u}_3 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}.$$

$\mathbf{u}_1 \cdot \mathbf{u}_2 = 1(0) + (-2)(0) + 0(1) = 0$, so $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthogonal.

$\mathbf{u}_1 \cdot \mathbf{u}_3 = 1(2) + (-2)(1) + 0(1) = 2 - 2 = 0$, so $\mathbf{u}_1$ and $\mathbf{u}_3$ are orthogonal.

$\mathbf{u}_2 \cdot \mathbf{u}_3 = 0(2) + (0)(1) + 1(1) = 1$, so $\mathbf{u}_2$ and $\mathbf{u}_3$ are not orthogonal and this is not an orthogonal set.

Theorem 2. *Any orthogonal set of vectors is necessarily linearly independent.*

Thus an orthogonal set of vectors in $\mathbb{R}^n$ is a basis of some *subspace* of $\mathbb{R}^n$.

Definition 3. A subspace of $\mathbb{R}^n$ is a subset V of $\mathbb{R}^n$ satisfying three properties:

1. Nonemptiness (There is at least one vector in V).
2. Closure under addition (For all $\mathbf{u}, \mathbf{v} \in V, \mathbf{u} + \mathbf{v} \in V$).
3. Closure under scalar multiplication (For all $c \in \mathbb{R}$ and $\mathbf{u} \in V$, $c\mathbf{u} \in V$).

Question 4. What vector has to be in all subspaces of $\mathbb{R}^n$?

The zero vector. Let $\mathbf{u} \in V$ and $c = 0$. Then the last property says that $0 \cdot \mathbf{u} = \mathbf{0}$ must be in V.

Definition 5. Let A be an $m \times n$ matrix. The linear combinations of the columns of A create a subspace of $\mathbb{R}^m$ called the *column space*.

$$\text{col}(A) = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = A\mathbf{x} \text{ for some } \mathbf{x} \in \mathbb{R}^n\}$$

Informally, the column space of A is the set of all possible outputs of A times a vector. In the least squares problem, we are seeking answers for $\mathbf{b}$ that are not in $\text{col}(A)$.

We can, however, decompose $\mathbf{b}$ into a vector that is in $\text{col}(A)$ plus another vector. Using an arbitrary $\mathbf{x}$, we write

$$\mathbf{b} = A\mathbf{x} + (\mathbf{b} - A\mathbf{x}).$$

The part $A\mathbf{x}$ is in $\text{col}(A)$ by definition. Our goal, in solving the least square problem, is to find an $\mathbf{x}$ where the second vector, $\mathbf{b} - A\mathbf{x}$, is orthogonal to $A\mathbf{x}$, as that $\mathbf{x}$ will be the closest ($\hat{\mathbf{x}}$) we can get to $\mathbf{b}$ (see pictures). When we find the closest $\hat{\mathbf{x}}$, we call the vector $A\hat{\mathbf{x}}$ the *orthogonal projection of* $\boldsymbol{b}$ *onto col(A)*

Every subspace V has an *orthogonal complement* $V^\perp$ that is also a subspace.

Definition 6. Let V be a subspace of $\mathbb{R}^n$. The orthogonal complement $V^\perp$ is the subset of $\mathbb{R}^n$ containing all vectors $\mathbf{u} \in \mathbb{R}^n$ such that for all $\mathbf{v} \in V$, $\mathbf{u} \cdot \mathbf{v} = 0$. The orthogonal complement is also a subspace of V.

Every matrix A also generates a subspace of $\mathbb{R}^n$ called the *null space*.

Definition 7. The null space of A is the set of vectors $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = \mathbf{0}$.

$$\text{nul}(A) = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \mathbf{0}\}$$

Example 3. State two vectors in the null space of $A = \begin{bmatrix} 1 & 1 & -1 \\ 2 & 0 & -1 \end{bmatrix}$. Then find two vectors null space of A^T.

Two vectors in nul(A) are $\begin{bmatrix}0\\0\\0\end{bmatrix}$ and $\begin{bmatrix}1\\1\\2\end{bmatrix}$. However, there is only one unique vector in nul(A^T), $\begin{bmatrix}0\\0\end{bmatrix}$.

Theorem 8. *For every matrix A, col(A)$^\perp$ =nul(A^T).*

So now we know we are looking for $\mathbf{x}$ such that $\mathbf{b} - A\mathbf{x}$ is in nul(A^T). That is,

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

Rearranging gives the first strategy for finding the least squares solution $\hat{\mathbf{x}}$: Solve the *normal equations*

$$A^T A\mathbf{x} = A^T\mathbf{b}.$$

Normal Equations Examples

Example 4. First, check if the vector $\mathbf{x}_1 = \begin{bmatrix}1\\1\end{bmatrix}$ is the least squares solution to

$$\begin{bmatrix}1 & 2\\2 & 4\\1 & 1\end{bmatrix}\mathbf{x} = \begin{bmatrix}3\\4\\2\end{bmatrix}$$

using orthogonality. If $\mathbf{x}_1$ is not the solution, compute the solution using the normal equations. Compute the squared error.

If $\mathbf{x}_1$ is the solution, then $A\mathbf{x}_1 \cdot (\mathbf{b} - A\mathbf{x}_1)$ should be 0. So we compute each vector:

$A\mathbf{x}_1 = \begin{bmatrix}1 & 2\\2 & 4\\1 & 1\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}3\\6\\2\end{bmatrix}$, which certainly doesn't seem like a poor approximation to $\begin{bmatrix}3\\4\\2\end{bmatrix}$.

Then $\mathbf{b} - A\mathbf{x}_1 = \begin{bmatrix}3\\4\\2\end{bmatrix} - \begin{bmatrix}3\\6\\2\end{bmatrix} = \begin{bmatrix}0\\-2\\0\end{bmatrix}$. Checking orthogonality: $\begin{bmatrix}3\\6\\2\end{bmatrix} \cdot \begin{bmatrix}0\\-2\\0\end{bmatrix} = -12$, not zero, so $\begin{bmatrix}1\\1\end{bmatrix}$ is not the least squares solution.

Solving the normal equations:

$A^T A = \begin{bmatrix}1 & 2 & 1\\2 & 4 & 1\end{bmatrix}\begin{bmatrix}1 & 2\\2 & 4\\1 & 1\end{bmatrix} = \begin{bmatrix}6 & 11\\11 & 21\end{bmatrix}$, and $\begin{bmatrix}1 & 2 & 1\\2 & 4 & 1\end{bmatrix}\begin{bmatrix}3\\4\\2\end{bmatrix} = \begin{bmatrix}13\\24\end{bmatrix}$ so we have

$$\begin{bmatrix}6 & 11 & 13\\11 & 21 & 24\end{bmatrix} \sim \begin{bmatrix}6 & 11 & 13\\0 & \frac{5}{6} & \frac{1}{6}\end{bmatrix} \sim \begin{bmatrix}6 & 11 & 13\\0 & 1 & \frac{1}{5}\end{bmatrix} \sim \begin{bmatrix}6 & 0 & \frac{54}{5}\\0 & 1 & \frac{1}{5}\end{bmatrix} \sim \begin{bmatrix}1 & 0 & \frac{9}{5}\\0 & 1 & \frac{1}{5}\end{bmatrix}$$

This means that $\hat{\mathbf{x}} = \begin{bmatrix} \frac{9}{5} \\ \frac{1}{5} \end{bmatrix}$, and $A\hat{\mathbf{x}} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{9}{5} \\ \frac{1}{5} \end{bmatrix} = \begin{bmatrix} \frac{11}{5} \\ \frac{22}{5} \\ 2 \end{bmatrix}$ is the orthogonal projection of $\mathbf{b}$ onto col(A).

We can also check orthogonality. First computing $\mathbf{b} - A\hat{\mathbf{x}}$, we have $\begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix} - \begin{bmatrix} \frac{11}{5} \\ \frac{22}{5} \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{4}{5} \\ \frac{-2}{5} \\ 0 \end{bmatrix}$.

Then $\begin{bmatrix} \frac{11}{5} \\ \frac{22}{5} \\ 2 \end{bmatrix} \cdot \begin{bmatrix} \frac{4}{5} \\ \frac{-2}{5} \\ 0 \end{bmatrix} = \frac{11}{5}(\frac{4}{5}) + \frac{22}{5}(\frac{-2}{5}) + 0 = 0.$

The squared error for $\mathbf{x}_1$ is $0^2 + (-2)^2 + 0^2 = 4$, while the squared error for $\hat{\mathbf{x}}$ is $(\frac{4}{5})^2 + (\frac{-2}{5})^2 + 0^2 = \frac{20}{25}$.

Example 5. Suppose you are trying to find the spring constant for a particular spring. You've conducted three measurements and found that a force of 29N stretches the spring 1m, a force of 31N compresses the spring 1m, and a force of 62N stretches the spring 2m. Estimate the spring constant.

Since Hooke's Law says that $F = kx$, we have three equations for one unknown:

$$\begin{aligned} 29 &= k(1) \\ -31 &= k(-1) \\ 62 &= k(2) \end{aligned}$$

Written as a system, we have $A = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 29 \\ -31 \\ 62 \end{bmatrix}$. Solving using the normal equations, we have $A^TA = 1 + 1 + 4 = 6$ and $A^T\mathbf{b} = 29 + 31 + 124 = 184$. So the system becomes $6k = 184$, and $k = 31$.

Further theory: Normal Equations

Least squares solutions always exist. Theorem 9 follows from the fact that every vector has an orthogonal projection. See linear algebra texts for more details.

Theorem 9. *The system $A^TA\boldsymbol{x} = A^T\boldsymbol{b}$ is consistent.*

However, least squares solutions do not have to be unique.

Theorem 10. *Let A be an $m \times n$ matrix and let $\boldsymbol{b}$ be a vector in $\mathbb{R}^m$. The following are logically equivalent:*

1. *$A\boldsymbol{x} = \boldsymbol{b}$ has a unique least-squares solution.*
2. *The columns of A are linearly independent.*

3. A^TA is invertible.

For instance, if $n > m$, the columns of A must be dependent, and that immediately leads to infinitely many least squares solutions.

When A^TA is invertible, we define the *pseudoinverse* to be

$$A^+ = (A^TA)^{-1}A^T.$$

Then we can say that $\hat{\mathbf{x}} = A^+\mathbf{b}$.

Question 11. You might recall from linear algebra that $(AB)^{-1} = B^{-1}A^{-1}$. Can we write the pseudoinverse as $A^{-1}(A^T)^{-1}A^T$?

No, not usually. The property for AB requires A and B to be invertible, which also requires square matrices, whereas we are primarily interested in A's with more rows than columns. However, when A is square and invertible, the above rewriting is true and can be simplified to just A^{-1}.

Once formed, the normal equations can be solved using any of the decomposition techniques learned earlier in the course.

Question 12. Is A^TA symmetric?

Yes, it is. To see this, use the definition of symmetric and take its transpose: $(A^TA)^T = A^T(A^T)^T = A^TA$.

If A^TA is also positive definite, the normal equations can then be solved relatively efficiently using Cholesky.

However, this also means that solving the normal equations has the same limitations as solving systems of equations. What does the condition number of A^TA look like, relative to A?

Theorem 13. *In the 2-norm, cond(A^TA) = (cond(A))^2.*

Ouch. The implication of Theorem 13 is that solving the normal equations is really only feasible for small or particularly well-conditioned matrices A.

Example 6. Suppose you are trying to determine the coefficients of a degree six polynomial: $ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$. You might start by choosing an interval, say, $[2, 4]$, and testing the output values there at a regular interval.

x	$f(x)$
2	127.000
2.25	232.743
2.5	406.234
2.75	679.087
3	1093.000
3.25	1701.718
3.5	2573.172
3.75	3791.792
4	5461.000

Letting $x_1 = 2$, $x_2 = 2.25$, and so on, we have a system of the form:

$$\begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^6 \\ 1 & x_2 & x_2^2 & x_2^3 & \dots & x_2^6 \\ \vdots & & & & & \\ 1 & x_9 & x_9^2 & x_9^3 & \dots & x_9^6 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ \vdots \\ g \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_9) \end{bmatrix}.$$

The coefficient matrix A is called a *Van der Monde matrix*. This A has a condition number of about 1.8×10^8. How many correct decimal places would you expect in an answer computed from the normal equations in double precision?

Answer: basically none. $(10^8)^2 = 10^{16}$. Recalling that machine epsilon is about 10^{-16}, we now expect $10^{-16+16} = 10^0$ meaning 0 digits of accuracy.

In actuality, the answer should be all one's, and the computed solution is roughly $\begin{bmatrix} 0.992 \\ 1.016 \\ .986 \\ 1.006 \\ .998 \\ 1.000 \\ 1.000 \end{bmatrix}$, so we still get a digit or so. But this is the limit of how large a Van der Monde matrix can be and still get an answer, and 9×6 is hardly large.

So what do you do if you have an ill-conditioned problem? Try to avoid it. In the next topic, we turn again to orthogonality and projections to get another algorithm for computing the least squares solutions, this time without forming the normal equations.

Final Note on the 2-norm

One of the biggest qualitative advantages to the 2-norm is that it punishes particularly large errors. Meaning, the 2-norm prefers solutions with multiple small errors over a few large

errors. Minimization over other norms is done in some settings as seems appropriate to the application and the types of solutions sought.