MATH 5330: Computational Methods of Linear Algebra

Lecture Note 8: Linear Least Squares Problem

Xianyi Zeng
Department of Mathematical Sciences, UTEP

1 From Linear System to Least Squares

In previous sections we solve the linear system $A\boldsymbol{x}=\boldsymbol{b}$ when A is square and non-singular. In the more general case, the problem is not mathematically well-posed. Let A be any $n\times n$ matrix, but $\det A=0$, then the system $A\boldsymbol{x}=\boldsymbol{b}$:

- Has no solution if $\boldsymbol{b}\notin \mathrm{col}(A)$.
- Has infinite number of solutions if $\boldsymbol{b}\in \mathrm{col}(A)$.

Here $\mathrm{col}(A)$ is the column space of A:

$$\mathrm{col}(A)=\mathrm{span}(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n)\,, \tag{1.1}$$

where $\boldsymbol{a}_i$, $i=1,\cdots,n$ are the column vectors of A: $A=[\boldsymbol{a}_1\ \boldsymbol{a}_2\ \cdots\ \boldsymbol{a}_n]$.

In a more general case let $A\in\mathbb{R}^{m\times n}$ be any matrix, then $A\boldsymbol{x}=\boldsymbol{b}$ for $\boldsymbol{b}\in\mathbb{R}^m$ has at least one solution $\boldsymbol{x}\in\mathbb{R}^n$ if and only if $\boldsymbol{b}\in\mathrm{col}(A)$. In particular, if $\boldsymbol{b}\in\mathrm{col}(A)$:

- The solution is unique if and only if $n\le m$ and $\mathrm{rank}(A)=n$, or equivalently $\dim(\mathrm{col}(A))=n$.
- There are infinite number of solutions in all other situations.

We'll briefly prove the first statement:

Proof. If $n\le m$ and $\mathrm{rank}(A)=n$, then $A^tA\in\mathbb{R}^{n\times n}$ is non-singular (see Appendix A). Let $\boldsymbol{x}$ be any vector that satisfies $A\boldsymbol{x}=\boldsymbol{b}$, then:

$$(A^tA)\boldsymbol{x}=A^t(A\boldsymbol{x})=A^t\boldsymbol{b}\,,$$

or equivalently:

$$\boldsymbol{x}=(A^tA)^{-1}A^t\boldsymbol{b}\,, \tag{1.2}$$

which demonstrates the uniqueness.

Conversely, we suppose $A\boldsymbol{x}=\boldsymbol{b}$ has a unique solution $\boldsymbol{x}_0\in\mathbb{R}^n$. Let $\boldsymbol{y}\in\ker(A)$, the null space of A, then $\boldsymbol{x}_0+\boldsymbol{y}$ is also a solution:

$$A(\boldsymbol{x}_0+\boldsymbol{y})=A\boldsymbol{x}_0+A\boldsymbol{y}=\boldsymbol{b}+0=\boldsymbol{b}\,.$$

Due to the uniqueness of $\boldsymbol{x}_0$, we know that $\ker(A)$ only contains the zero vector, i.e., $\dim(\ker(A))=0$. But from the *dimension theorem* of linear algebra we know that $\dim(\ker(A))=n-\dim(\mathrm{col}(A))$, thus:

$$n-\dim(\mathrm{col}(A))=0 \quad\Rightarrow\quad \dim(\mathrm{col}(A))=n\,,$$

and consequently A has full rank n and $m\ge n$. □

In practice, however, we do not always have a well-posed system to solve (see the example in the next section). To this end, people define instead the next least squares problem for any $A \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} ||A\boldsymbol{x} - \boldsymbol{b}||_2 \,. \tag{1.3}$$

Note that this problem always has at least one solution[1]; but the solution may not be unique. Particularly, if $\boldsymbol{x}_0$ minimize the residual then $\boldsymbol{x}_0 + \boldsymbol{y}$ also minimize the residual for all $\boldsymbol{y} \in \ker(A)$. Thus the least squares problem (1.3) is always *solvable*, but not necessarily well-posed. A well-posed extension is presented at the end of this lecture.

Theorem 1.1. $\boldsymbol{x}$ *solves the least squares problem (1.3) if and only if it solves the normal equation:*

$$A^t A\boldsymbol{x} = A^t \boldsymbol{b} \,. \tag{1.4}$$

This equation always has a solution, and the solution is unique if and only if the columns of A are linearly independent (i.e., $\dim(\mathrm{col}(A)) = \mathrm{rank}(A) = n \le m$*).*

Proof. The first statement can be obtained as a consequence of the stationary condition for minimizing the convex quadratic form $\phi(\boldsymbol{x}) = (A\boldsymbol{x} - \boldsymbol{b})^t(A\boldsymbol{x} - \boldsymbol{b})$, that is, $\nabla\phi(\boldsymbol{x}) = 0$. Here we instead use a more direct approach. Let $\boldsymbol{x}$ solve the normal equation and let $\boldsymbol{e} \in \mathbb{R}^n$ be arbitrary, then:

$$\begin{aligned} ||A(\boldsymbol{x}+\boldsymbol{e}) - \boldsymbol{b}||^2 &= (A\boldsymbol{x} - \boldsymbol{b} - A\boldsymbol{e})^t(A\boldsymbol{x} - \boldsymbol{b} - A\boldsymbol{e}) = ||A\boldsymbol{x} - \boldsymbol{b}||^2 + ||A\boldsymbol{e}||^2 - 2\boldsymbol{e}^t(A^tA\boldsymbol{x} - A^t\boldsymbol{b}) \\ &= ||A\boldsymbol{x} - \boldsymbol{b}||^2 + ||A\boldsymbol{e}||^2 \ge ||A\boldsymbol{x} - \boldsymbol{b}||^2 \,, \end{aligned}$$

and due to the arbitrary choice of $\boldsymbol{e}$, the same $\boldsymbol{x}$ solves the least squares problem. Conversely if $\boldsymbol{x}$ solves (1.3) we define $\boldsymbol{d} = A^tA\boldsymbol{x} - A^t\boldsymbol{b}$, then for all $\boldsymbol{e} \in \mathbb{R}^n$ we have:

$$0 \le ||A(\boldsymbol{x}+\boldsymbol{e}) - \boldsymbol{b}||^2 - ||A\boldsymbol{x} - \boldsymbol{b}||^2 = ||A\boldsymbol{e}||^2 - 2\boldsymbol{e}^t\boldsymbol{d} \,.$$

With the particular choice $\boldsymbol{e} = -\alpha\boldsymbol{d}$, we have for all $\alpha > 0$:

$$0 \le \alpha^2 ||A\boldsymbol{d}||^2 - 2\alpha||\boldsymbol{d}||^2 \quad \Rightarrow \quad 0 \le \alpha||A\boldsymbol{d}||^2 - 2||\boldsymbol{d}||^2 \,.$$

Letting $\alpha \to 0$, we have $2||\boldsymbol{d}||^2 \le 0$ or $\boldsymbol{d} = 0$, i.e., $\boldsymbol{x}$ solves the normal equation.

For the second part, clearly we just need to show that the normal equation always has a solution. Indeed, due to Appendix A the column spaces of A^t and A^tA are the identical. Hence for all $\boldsymbol{b} \in \mathbb{R}^m$, we have:

$$A^t\boldsymbol{b} \in \mathrm{col}(A^t) = \mathrm{col}(A^tA) \,,$$

or there exists an $\boldsymbol{x} \in \mathbb{R}^n$ such that $A^t\boldsymbol{b} = A^tA\boldsymbol{x}$. □

To this end, the problem of solving the least squares problem is equivalent to solving the normal equation, which we will discuss in more detail in the next lectures.

[1] An elegant proof can be obtained by the following result of real analysis: Any continuous function defined on a compact set $K \subset \mathbb{R}^n$ achieves its minimum and maximum on the set K.

2 Example: Polynomial Regression

An important application of the least squares problem is to find a polynomial fit of scattered data. Let our data set be:

$$\mathcal{D} = \{(x_i, y_i) : x_i\,,\, y_i \in \mathbb{R}\,,\, i = 1, \cdots, n\}\,. \tag{2.1}$$

The target is to find a function $f : \mathbb{R} \mapsto \mathbb{R}$, such that:

$$y_i \approx f(x_i)\,,\; \forall 1 \leq i \leq n\,.$$

In polynomial regression, we search for a polynomial of degree m:

$$p_m(x) = a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0\,, \tag{2.2}$$

so that the L^2-norm of the difference vector $\boldsymbol{z} = [z_i] \in \mathbb{R}^n$, $z_i = y_i - p_m(x_i)$ is minimized. Note that:

$$z_i = y_i - p_m(x_i) = y_i - \sum_{k=0}^{m} a_k x_i^k\,, \quad 1 \leq i \leq n\,,$$

we may write $\boldsymbol{z}$ in the matrix form:

$$\boldsymbol{z} = \boldsymbol{y} - V_m(\boldsymbol{x})\boldsymbol{a}\,, \tag{2.3}$$

where $\boldsymbol{a} = [a_k] \in \mathbb{R}^{m+1}$ and $V_m(\boldsymbol{x}) \in \mathbb{R}^{n \times (m+1)}$ is the Vandermonde matrix:

$$V_m(\boldsymbol{x}) = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix}_{n \times (m+1)}\,. \tag{2.4}$$

Then the problem of polynomial fitting, i.e., finding the coefficient vector $\boldsymbol{a}$, reduces to the least squares problem:

$$\boldsymbol{a} = \arg \min_{\boldsymbol{a}' \in \mathbb{R}^{m+1}} \left|\left| V_m(\boldsymbol{x})\boldsymbol{a}' - \boldsymbol{y} \right|\right|\,. \tag{2.5}$$

The existence of a solution (and hence p_m) is shown by Theorem 1.1, furthermore, if $x_i \neq x_j$ for all $i \neq j$ and $n \geq m+1$, the solution is unique. Due to the same theorem, we just need to show that if all x_i's are different from each other, $V_m(\boldsymbol{x})$ has full rank; the latter is the consequence of a special case – when $m+1 = n$ and all x_i's are different, $V_m(\boldsymbol{x})$ is non-singular. In fact, one can use induction to show that if $m+1 = n$, then:

$$\det V_m(\boldsymbol{x}) = \prod_{1 \leq i < j \leq n} (x_j - x_i) \neq 0\,. \tag{2.6}$$

Hence we deduce that when all the data points x_i are different, there is always a unique fit using the polynomial of any degree m.

3 Pseudoinverse

Lastly, we briefly discuss a well-posed problem for general matrix $A \in \mathbb{R}^{m\times n}$:

$$\boldsymbol{x} = \arg\min_{\boldsymbol{x}'\in\mathcal{S}} \left|\left|\boldsymbol{x}'\right|\right|, \quad \mathcal{S} = \{\boldsymbol{x}' \in \mathbb{R}^n : \left|\left|A\boldsymbol{x}' - \boldsymbol{b}\right|\right| = \min_{\boldsymbol{y}\in\mathbb{R}^n} ||A\boldsymbol{y} - \boldsymbol{b}||\}. \tag{3.1}$$

That is to say, we want to find the least squares solution with the smallest L^2-norm. Its existence is obtained similarly to the footnote before Theorem 1.1. To show the uniqueness, we need a lemma: if $\boldsymbol{y}_1,\boldsymbol{y}_2 \in \mathbb{R}^n$ such that $||\boldsymbol{y}_1|| = ||\boldsymbol{y}_2||$, then $||(\boldsymbol{y}_1+\boldsymbol{y}_2)/2|| \le ||\boldsymbol{y}_1||$; and the identity holds only if $\boldsymbol{y}_1 = \boldsymbol{y}_2$. This is a direct consequence of the parallelogram identity:

$$2||\boldsymbol{y}_1||^2 + 2||\boldsymbol{y}_2||^2 = ||\boldsymbol{y}_1+\boldsymbol{y}_2||^2 + ||\boldsymbol{y}_1-\boldsymbol{y}_2||^2. \tag{3.2}$$

Now if both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are solutions to (3.1), we must have $||\boldsymbol{x}_1|| = ||\boldsymbol{x}_2||$. Furthermore, $\boldsymbol{x}_1,\boldsymbol{x}_2 \in \mathcal{S}$ indicates

$$||A\boldsymbol{x}_1 - \boldsymbol{b}|| = ||A\boldsymbol{x}_2 - \boldsymbol{b}||.$$

By the previous lemma

$$||A((\boldsymbol{x}_1+\boldsymbol{x}_2)/2) - \boldsymbol{b}|| \le ||A\boldsymbol{x}_1 - \boldsymbol{b}|| \quad \Rightarrow \quad \frac{1}{2}(\boldsymbol{x}_1+\boldsymbol{x}_2) \in \mathcal{S}.$$

Invoking the lemma again, we have:

$$||(\boldsymbol{x}_1+\boldsymbol{x}_2)/2|| \le ||\boldsymbol{x}_1||,$$

and by the assumption we must have the identity hold, i.e., $\boldsymbol{x}_1 = \boldsymbol{x}_2$. We thusly conclude that the solution to (3.1) is unique.

Now we deviate ourselves from the numerical techniques and look at the analytical solution. Let $A = U\Sigma V$ be the singular value decomposition (SVD) of A, where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal in the sense that only the (i,i)-th components of Σ are (possibly) nonzero, where $1 \le i \le \min(m,n)$. In addition, denoting the (i,i)-th component of Σ by σ_i we have:

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_k \ge 0, \quad \text{where} \quad k = \min(m,n). \tag{3.3}$$

Using the SVD of A and noting the fact that pre-multiplying by an orthogonal matrix preserves the L^2-norm, (3.1) is equivalent to:

$$\boldsymbol{y} = \arg\min_{\boldsymbol{y}'\in\mathcal{S}'} \left|\left|\boldsymbol{y}'\right|\right|, \quad \mathcal{S}' = \{\boldsymbol{y}' \in \mathbb{R}^n : \left|\left|\Sigma\boldsymbol{y}' - U^t\boldsymbol{b}\right|\right| = \min_{\boldsymbol{z}\in\mathbb{R}^n} \left|\left|\Sigma\boldsymbol{z} - U^t\boldsymbol{b}\right|\right|\},$$

where the solution $\boldsymbol{y}$ is related to the solution $\boldsymbol{x}$ of (3.1) by $\boldsymbol{y} = V\boldsymbol{x}$. The members of $\mathcal{S}'$ have a surprisingly simple structure as the m equations are decoupled from each other. In particular, $\boldsymbol{y}' = [y_i'] \in \mathcal{S}'$ is given by:

$$\begin{cases} y_i' = \frac{1}{\sigma_i}[U^t\boldsymbol{b}]_i, & \text{if } \sigma_i \ne 0; \\ y_i' \text{ is arbitrary}, & \text{if } i > k \text{ or } \sigma_i = 0. \end{cases}$$

Clearly the solution is given by setting all y_i' in the second line to be zero. The final result can be written as:

$$\boldsymbol{x} = V^t\boldsymbol{y} = V^t\Sigma^+U^t\boldsymbol{b},$$

where $\Sigma^+ \in \mathbb{R}^{n\times m}$ has a similar structure as Σ^t, where every nonzero σ_i is replaced by $1/\sigma_i$. $A^+ \stackrel{\text{def}}{=\!=} V^t\Sigma^+U^t$ is known as the pseudoinverse of the matrix A.

Exercises

Exercise 1. *Show that if we want to fit the data set (2.1) such that* $x_i \neq x_j$, $\forall i \neq j$ *by a constant function, that is, a polynomial of degree zero, this function is given by the arithmetic average of all* y_i*'s.*

Exercise 2. *Verify (2.6) for the case* $m+1=n=4$*. That is, show that:*

$$\det \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ 1 & x_4 & x_4^2 & x_4^3 \end{bmatrix} = (x_4-x_3)(x_4-x_2)(x_4-x_1)(x_3-x_2)(x_3-x_1)(x_2-x_1)\,.$$

Exercise 3. *Compute the pseudoinverse of the following* 3×2 *matrix* A*:*

$$A = \begin{bmatrix} 2\sqrt{2} & 2\sqrt{2} \\ -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

You may use any software to compute the SVD decomposition of A*.*

A Show that $\mathrm{col}(A^tA) = \mathrm{col}(A)$ for any $A \in \mathbb{R}^{m\times n}$

Let $\mathcal{A}_1 = \mathrm{col}(A)$ and $\mathcal{A}_2 = \mathrm{col}(A^tA)$, then we have:

$$\mathcal{A}_2 = \{A^tA\boldsymbol{x} \,:\, \boldsymbol{x} \in \mathbb{R}^n\} \;\subseteq\; \{A^t\boldsymbol{y} \,:\, \boldsymbol{y} \in \mathbb{R}^m\} = \mathcal{A}_1\,.$$

Thus $\mathcal{A}_1^\perp \subseteq \mathcal{A}_2^\perp$. In fact, by the definition of orthogonal complement. Indeed:

$$\boldsymbol{u} \in \mathcal{A}_1^\perp \quad\Rightarrow\quad \boldsymbol{u}\cdot\boldsymbol{v} = 0\,,\,\forall\,\boldsymbol{v} \in \mathcal{A}_1 \quad\Rightarrow\quad \boldsymbol{u}\cdot\boldsymbol{v} = 0\,,\,\forall\,\boldsymbol{v} \in \mathcal{A}_2 \quad\Rightarrow\quad \boldsymbol{u} \in \mathcal{A}_2^\perp\,.$$

However, we can also show $\mathcal{A}_2^\perp \subseteq \mathcal{A}_1^\perp$ as follows:

$$\boldsymbol{u} \in \mathcal{A}_2^\perp \quad\Rightarrow\quad \boldsymbol{u}^tA^tA = 0 \quad\Rightarrow\quad ||A\boldsymbol{u}||^2 = \boldsymbol{u}^tA^tA\boldsymbol{u} = 0 \quad\Rightarrow\quad \boldsymbol{u}^tA^t = 0 \quad\Rightarrow\quad \boldsymbol{u} \in \mathcal{A}_1^\perp\,.$$

Thus $\mathcal{A}_1^\perp = \mathcal{A}_2^\perp$ and $\mathcal{A}_1 = (\mathcal{A}_1^\perp)^\perp = (\mathcal{A}_2^\perp)^\perp = \mathcal{A}_2$[2].

As a corollary, if A has full column rank (i.e., $\mathrm{rank}(A) = n$) then A^tA is non-singular.

[2] Check that for all subspace $\mathcal{V}$ of $\mathbb{R}^n$, $(\mathcal{V}^\perp)^\perp = \mathcal{V}$.