

Lecture 16: Discrete Linear Least Squares

3. Approximation Theory.

Interpolation with high-degree polynomials is not always the best way to approximate a function: we have seen an example where the polynomial diverges from the function it is meant to approximate between nodes as the polynomial degree grows. Numerical errors add further complexity. There is also a more fundamental objection: if the function or data you are modeling cannot be approximated well with low-degree polynomials, perhaps you should be using another class of functions (rational functions, trigonometric functions, or e^x and e^{-x} , etc.). Piecewise polynomial interpolation provides an alternative, provided smoothness is not a concern.

In this lecture, we consider an alternative to interpolation, *approximation by polynomials*:

Given $f \in C[a, b]$ and $m + 1$ points $\{x_j\}_{j=0}^m$ satisfying $a \leq x_0 < x_1 < \cdots < x_m \leq b$, determine some $p \in \mathcal{P}_n$ ($n \leq m$) such that

$$p(x_j) \approx f(x_j) \quad \text{for } j = 0, \dots, m.$$

Notice that this is essentially just the standard interpolation problem when $m = n$, in which case we have seen that there exists a unique $p \in \mathcal{P}_n$ such that $p(x_j) = f(x_j)$ for $0 \leq j \leq n$. However, when $m > n$, there generally will be no $p \in \mathcal{P}_n$ that delivers equality $p(x_j) = f(x_j)$ for all $j = 0, \dots, m$. We must settle for approximation, $p(x_j) \approx f(x_j)$, together with a method for quantifying this approximation. For example, we could choose p to minimize the maximum error at any grid point:

$$\min_{p \in \mathcal{P}_n} \max_{0 \leq j \leq m} |f(x_j) - p(x_j)|,$$

or the sum of the squares of the errors:

$$\min_{p \in \mathcal{P}_n} \sum_{j=0}^m |f(x_j) - p(x_j)|^2. \quad (16.1)$$

Other alternatives include ignoring the specific points x_0, \dots, x_m , and generalizing the above two measures to the entire interval of interest, $[a, b]$:

$$\min_{p \in \mathcal{P}_n} \max_{x \in [a, b]} |f(x) - p(x)| \quad \text{or} \quad \min_{p \in \mathcal{P}_n} \int_a^b |f(x) - p(x)|^2.$$

Each of these different error metrics gives rise different ‘optimal’ approximations, and inspire distinct algorithms for their construction. In this lecture, we study minimization of the second error, (16.1), the sum of the square of the errors at the grid points. Suppose we seek p in the monomial basis, $p(x) = c_0 + c_1x + c_2x^2 + \cdots + c_nx^n$. Define

$$\mathbf{A} = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \end{bmatrix}.$$

In this notation, problem (16.1) is equivalent to the matrix optimization problem

$$\min_{\mathbf{c} \in \mathbb{C}^n} \|\mathbf{f} - \mathbf{A}\mathbf{c}\|_2.$$

Because we are minimizing a sum of squares in (16.1), this is called a *least squares problem*. (Sometimes we say that the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is *overdetermined*, meaning that an excessive number of constraints prevents the system from having an exact solution \mathbf{x} . Some authors will thus describe the approximation procedure we are about to study as ‘solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ in the least-squares sense.’)

3.1. Discrete least squares problems.

We focus our attention to the general least squares problem

$$\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2,$$

where $\mathbf{A} \in \mathbb{C}^{m \times n}$, $\mathbf{b} \in \mathbb{C}^m$ with $m \geq n$.

Before tackling this problem, recall the Fundamental Theorem of Linear Algebra, stated in Lecture 2. The range of \mathbf{A} is the subspace

$$\text{Ran}(\mathbf{A}) = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathbb{C}^n\},$$

and the left null space is

$$\text{Ker}(\mathbf{A}^*) = \{\mathbf{y} \in \mathbb{C}^m : \mathbf{A}^*\mathbf{y} = 0\}.$$

The Fundamental Theorem of Linear Algebra states that

$$\mathbb{C}^m = \text{Ran}(\mathbf{A}) \oplus \text{Ker}(\mathbf{A}^*),$$

and also that $\text{Ran}(\mathbf{A}) \perp \text{Ker}(\mathbf{A}^*)$. This immediately implies that any $\mathbf{b} \in \mathbb{C}^m$ can be written uniquely as $\mathbf{b} = \mathbf{b}_R + \mathbf{b}_N$, where $\mathbf{b}_R \in \text{Ran}(\mathbf{A})$ and $\mathbf{b}_N \in \text{Ker}(\mathbf{A}^*)$ with $\mathbf{b}_R \perp \mathbf{b}_N$.

For any $\mathbf{x} \in \mathbb{C}^n$, define the *residual* vector

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}.$$

Decomposing $\mathbf{b} = \mathbf{b}_R + \mathbf{b}_N$ as described above, we have

$$\begin{aligned} \mathbf{r} &= \mathbf{b} - \mathbf{A}\mathbf{x} \\ &= \mathbf{b}_R - \mathbf{A}\mathbf{x} + \mathbf{b}_N. \end{aligned}$$

One can immediately see from the definition of $\text{Ran}(\mathbf{A})$ that $\mathbf{A}\mathbf{x} \in \text{Ran}(\mathbf{A})$. Since $\mathbf{b}_R \in \text{Ran}(\mathbf{A})$, too, and $\text{Ran}(\mathbf{A})$ is a subspace, it must be that $\mathbf{b}_R - \mathbf{A}\mathbf{x} \in \text{Ran}(\mathbf{A})$. The Fundamental Theorem of Linear Algebra ensures that

$$\begin{aligned} \|\mathbf{r}\|_2^2 &= \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \\ &= \|(\mathbf{b}_R - \mathbf{A}\mathbf{x}) + \mathbf{b}_N\|_2^2 \\ &= \|\mathbf{b}_R - \mathbf{A}\mathbf{x}\|_2^2 + \|\mathbf{b}_N\|_2^2, \end{aligned}$$

since $\mathbf{b}_R - \mathbf{A}\mathbf{x} \perp \mathbf{b}_N$.[†] Thus,

$$\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 = \min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{b}_R - \mathbf{A}\mathbf{x}\|_2^2 + \|\mathbf{b}_N\|_2^2.$$

[†]The last equality is simply the Pythagorean Theorem: The vectors $\mathbf{b}_R - \mathbf{A}\mathbf{x}$ and \mathbf{b}_N are orthogonal, so they meet at a right angle; \mathbf{r} is the hypotenuse of the right-triangle whose legs are $\mathbf{b}_R - \mathbf{A}\mathbf{x}$ and \mathbf{b}_N . This result can be easily verified by directly computing $\|\mathbf{r}\|^2 = \mathbf{r}^*\mathbf{r} = ((\mathbf{b}_R - \mathbf{A}\mathbf{x}) + \mathbf{b}_N)^*((\mathbf{b}_R - \mathbf{A}\mathbf{x}) + \mathbf{b}_N)$.

Note that \mathbf{x} appears nowhere in the last term of this sum: the $\|\mathbf{b}_N\|_2^2$ component is inaccessible regardless of the choice of \mathbf{x} . Thus, the best hope for minimizing $\|\mathbf{b} - \mathbf{Ax}\|_2$ is to minimize $\|\mathbf{b}_R - \mathbf{Ax}\|_2$. As this term is always non-negative, the optimal solution is some \mathbf{x} that makes $\mathbf{b}_R - \mathbf{Ax} = \mathbf{0}$. Since $\mathbf{b}_R \in \text{Ran}(\mathbf{A})$, the definition of $\text{Ran}(\mathbf{A})$ guarantees there must be some $\mathbf{x} \in \mathbb{C}^n$ such that $\mathbf{b}_R = \mathbf{Ax}$.

3.1.1. Solving least squares problems via the normal equations.

There are several ways to obtain this solution, \mathbf{x} . The first relies on a simple trick. If $\mathbf{b}_R = \mathbf{Ax}$, then

$$\mathbf{b} - \mathbf{Ax} = (\mathbf{b}_R - \mathbf{Ax}) + \mathbf{b}_N = \mathbf{b}_N,$$

and since $\mathbf{b}_N \in \text{Ker}(\mathbf{A}^*)$, we immediately have

$$\mathbf{A}^*(\mathbf{b} - \mathbf{Ax}) = \mathbf{A}^*\mathbf{b}_N = \mathbf{0}.$$

We rearrange this equation into

$$\mathbf{A}^*\mathbf{Ax} = \mathbf{A}^*\mathbf{b}. \quad (16.2)$$

If \mathbf{A} is full rank (i.e., $\dim(\text{Ran}(\mathbf{A})) = n$), then $\mathbf{A}^*\mathbf{A}$ will be invertible,[‡] so we can write

$$\mathbf{x} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b}.$$

The formulation (16.2) arises sufficiently often to have its own name: the *normal equations*.

Note that $\mathbf{A}^*\mathbf{A} \in \mathbb{C}^{n \times n}$, which is often a small matrix. (Recall that $m \geq n$; in many applications, $m \gg n$.) Thus, $O(n^3)$ floating point operations are needed to solve the system $(\mathbf{A}^*\mathbf{A})\mathbf{x} = (\mathbf{A}^*\mathbf{b})$. However, it will be more costly to form the matrix $\mathbf{A}^*\mathbf{A}$: this matrix-matrix multiplication requires roughly mn^2 operations. (Why not $2mn^2$?) Moreover, this process is prone to magnify rounding errors, and hence is not favored by numerical analysts. Still, for ‘well conditioned problems,’ the normal equations approach can perform well. (We know about the condition number of a square matrix \mathbf{A} with respect to the solution of $\mathbf{Ax} = \mathbf{b}$. We shall investigate the condition number of \mathbf{A} with respect to the least squares problem on the fourth problem set.)

3.1.2. Solving least squares problems via QR factorization.

We next describe a technique for solving least squares problems that is more robust to rounding errors. Recall that any matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ can be written as $\mathbf{A} = \mathbf{QR}$ for a unitary matrix $\mathbf{Q} \in \mathbb{C}^{m \times m}$ and an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{m \times n}$. Substitute this factorization into the least squares objective function:

$$\|\mathbf{b} - \mathbf{Ax}\|_2 = \|\mathbf{b} - \mathbf{QRx}\|_2.$$

If we could remove \mathbf{Q} from the right-hand side, we would be left with a very simple upper triangular problem. Recall that the induced matrix 2-norm is invariant to unitary transformations. Since \mathbf{Q} is unitary, $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$; Moreover, since \mathbf{Q} is square, this means that \mathbf{Q}^* must be the unique inverse of \mathbf{Q} : $\mathbf{Q}^{-1} = \mathbf{Q}^*$. Thus,

$$\begin{aligned} \|\mathbf{b} - \mathbf{QRx}\|_2 &= \|\mathbf{QQ}^*\mathbf{b} - \mathbf{QRx}\|_2 \\ &= \|\mathbf{Q}(\mathbf{Q}^*\mathbf{b} - \mathbf{Rx})\|_2 \\ &= \|\mathbf{Q}^*\mathbf{b} - \mathbf{Rx}\|_2. \end{aligned}$$

[‡]When \mathbf{A} is not full rank, there are infinitely many choices for \mathbf{x} that yield the same residual norm. The singular value decomposition, the subject of the next lecture, provides a mechanism for describing this set and selecting one distinguished \mathbf{x} from the infinitely many candidates.

Now partition $\mathbf{Q}^*\mathbf{b} \in \mathbb{C}^m$ into two sections:

$$\mathbf{Q}^*\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix},$$

where $\mathbf{b}_1 \in \mathbb{C}^n$ and $\mathbf{b}_2 \in \mathbb{C}^{m-n}$. The rectangular upper triangular matrix \mathbf{R} can be similarly partitioned:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix},$$

where $\mathbf{R}_1 \in \mathbb{C}^{n \times n}$, and the zero block has dimension $(m-n)$ -by- n . Thus,

$$\mathbf{Q}^*\mathbf{b} - \mathbf{R}\mathbf{x} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} - \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{b}_1 - \mathbf{R}_1\mathbf{x} \\ \mathbf{b}_2 \end{bmatrix}.$$

Just as in the derivation of the normal equations, we observe that an optimal choice for \mathbf{x} will completely annihilate part of the residual, while leaving another component untouched. (In moving from $\|\mathbf{b} - \mathbf{Q}\mathbf{R}\mathbf{x}\|_2$ to the equivalent quantity $\|\mathbf{Q}^*\mathbf{b} - \mathbf{R}\mathbf{x}\|_2$, we have effectively transformed into a coordinate system in which $\text{Ran}(\mathbf{A})$ has been mapped to vectors that are zero in their final $m-n$ components, while $\text{Ker}(\mathbf{A}^*)$ now corresponds to vectors that are zero in their first n components, for full rank \mathbf{A} .) In particular,

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 = \left\| \begin{bmatrix} \mathbf{b}_1 - \mathbf{R}_1\mathbf{x} \\ \mathbf{b}_2 \end{bmatrix} \right\|_2^2 = \|\mathbf{b}_1 - \mathbf{R}_1\mathbf{x}\|_2^2 + \|\mathbf{b}_2\|_2^2.$$

Thus, the optimal choice for \mathbf{x} is simply

$$\mathbf{x} = \mathbf{R}_1^{-1}\mathbf{b}_1 = \mathbf{R}_1^{-1}\mathbf{Q}_1^*\mathbf{b},$$

where $\mathbf{Q}_1 \in \mathbb{C}^{m \times n}$ consists of the first n columns of \mathbf{Q} .[§]

3.1.3. Example from polynomial approximation.

We revisit the problem posed at the beginning of this lecture: given the set of distinct points $\{x_0, x_1, \dots, x_m\} \subset [a, b]$, find the polynomial $p \in \mathcal{P}_n$ that minimizes

$$\min_{p \in \mathcal{P}_n} \sum_{j=0}^m |f(x_j) - p(x_j)|^2$$

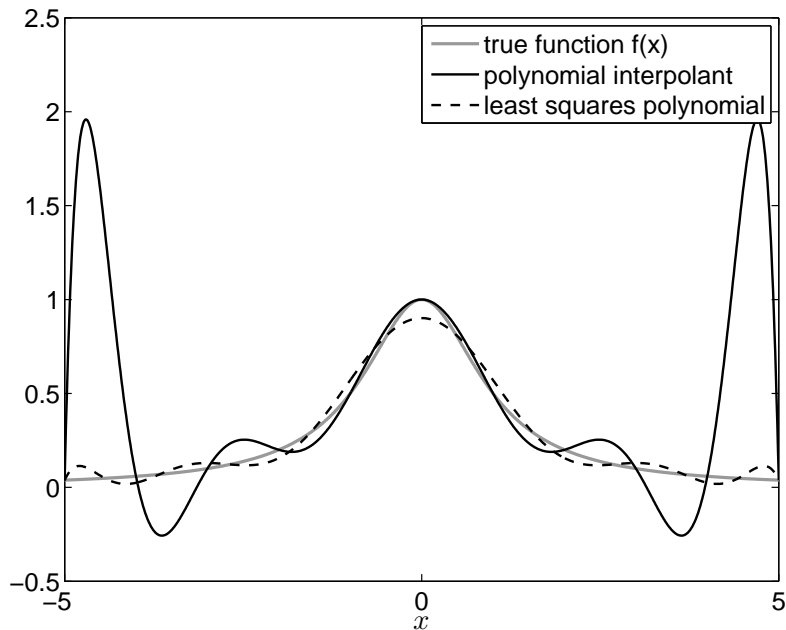
for some $f \in C[a, b]$. When $m > n$, this problem leads to an overdetermined linear system for the polynomial coefficients, one solved by MATLAB's `polyfit` command.

The coefficient matrix \mathbf{A} is full-rank provided the points $\{x_j\}$ are distinct, and the resulting least squares problem can be readily solved using the techniques just described. To see this approach in action, we revisit the troublesome Runge function,

$$f(x) = \frac{1}{1+x^2}$$

for $x \in [-5, 5]$. Recall that, for this example, the polynomial interpolants at uniformly spaced points did not converge as the degree of the polynomial increased. The figure below compares the $n = 10$ interpolant at uniformly spaced points with the least squares polynomial of degree $n = 10$ that approximates f at 21 uniformly spaced points ($m = 20$).

[§]Note that \mathbf{R}_1 is invertible if and only if \mathbf{A} is full rank. The QR factorization of a rank-deficient $\mathbf{A} \in \mathbb{C}^{m \times n}$ with $m \geq n$ must have a zero on the main diagonal. Can you explain why?



Now compare the overall error for the degree- n interpolant at uniformly spaced points with the error for the degree- n least squares polynomial based on uniformly spaced points with $m = 10n$. (This error is estimated by sampling $|f(x) - p_n(x)|$ at many points.) The least squares approach has a clear advantage here. Though simple and effective, the choice of $m = 10n$ approximation points is ad hoc. In upcoming lectures we study a more elegant approach that approximates f throughout the entire interval $[a, b]$ without the need to distinguish certain approximation points.

