LECTURE 16: *Introduction to Least Squares Approximation*

2.4 *Least squares approximation*

The minimax criterion is an intuitive objective for approximating a function. However, in many cases it is more appealing (for both computation and for the given application) to find an approximation to f that *minimizes the integral of the square of the error*.

> Given $f \in C[a,b]$, find $P_* \in \mathcal{P}_n$ such that
>
> $$\Big(\int_a^b (f(x)-P_*(x))^2\,\mathrm{d}x\Big)^{1/2} = \min_{p\in\mathcal{P}_n}\Big(\int_a^b (f(x)-p(x))^2\,\mathrm{d}x\Big)^{1/2}.$$

This is an example of a *least squares problem*.

2.4.1 *Inner products for function spaces*

To facilitate the development of least squares approximation theory, we introduce a formal structure for $C[a,b]$. First, recognize that $C[a,b]$ is a *linear space*: any linear combination of continuous functions on $[a,b]$ must itself be continuous on $[a,b]$.

Definition 2.2. The *inner product* of the functions $f,g \in C[a,b]$ is

$$\langle f,g\rangle = \int_a^b f(x)g(x)\,\mathrm{d}x.$$

The inner product satisfies the following basic axioms:

- $\langle \alpha f + g, h\rangle = \alpha\langle f,h\rangle + \langle g,h\rangle$ for all $f,g,h \in C[a,b]$ and all $\alpha \in \mathbb{R}$;
- $\langle f,g\rangle = \langle g,f\rangle$ for all $f,g \in C[a,b]$;
- $\langle f,f\rangle \ge 0$ for all $f \in C[a,b]$.

For simplicity we are assuming that f and g are real-valued. To handle complex-valued functions, one generalizes the inner product to

$$\langle f,g\rangle = \int_a^b f(x)\overline{g(x)}\,\mathrm{d}x,$$

which then gives $\langle f,g\rangle = \overline{\langle g,f\rangle}$.

With this inner product we associate the norm

$$\|f\|_2 := \langle f,f\rangle^{1/2} = \Big(\int_a^b f(x)^2\,\mathrm{d}x\Big)^{1/2}.$$

This is often called the 'L^2 norm,' where the superscript '2' in L^2 refers to the fact that the integrand involves the *square* of the function f; the L stands for *Lebesgue*, coming from the fact that this inner product can be generalized from $C[a,b]$ to the set of all functions that are *square-integrable*, in the sense of Lebesgue integration. By restricting our attention to continuous functions, we dodge the measure-theoretic complexities.

The Lebesgue theory gives a more robust definition of the integral than the conventional Riemann approach. With such notions one can extend least squares approximation beyond $C[a,b]$, to more exotic function spaces.

2.4.2 *Least squares minimization via calculus*

We are now ready to solve the least squares problem. We shall call the optimal polynomial $P_* \in \mathcal{P}_n$, i.e.,

$$\|f - P_*\|_2 = \min_{p \in \mathcal{P}_n} \|f - p\|_2.$$

We can solve this minimization problem using basic calculus. Consider this example for $n = 1$, where we optimize the error over polynomials of the form $p(x) = c_0 + c_1 x$. The polynomial that minimizes $\|f - p\|_2$ will also minimize its square, $\|f - p\|_2^2$. For any given $p \in \mathcal{P}_1$, define the error function

$$\begin{aligned}
E(c_0, c_1) := \|f(x) - (c_0 + c_1 x)\|_{L^2}^2 &= \int_a^b (f(x) - c_0 - c_1 x)^2 \,\mathrm{d}x \\
&= \int_a^b \Big(f(x)^2 - 2f(x)(c_0 + c_1 x) + (c_0^2 + 2c_0 c_1 x + c_1^2 x^2)\Big) \,\mathrm{d}x \\
&= \int_a^b f(x)^2 \,\mathrm{d}x - 2c_0 \int_a^b f(x) \,\mathrm{d}x - 2c_1 \int_a^b x f(x) \,\mathrm{d}x \\
&\qquad + c_0^2 (b - a) + c_0 c_1 (b^2 - a^2) + \tfrac{1}{3} c_1^2 (b^3 - a^3).
\end{aligned}$$

To find the optimal polynomial, P_*, optimize E over c_0 and c_1, i.e., find the values of c_0 and c_1 for which

$$\frac{\partial E}{\partial c_0} = \frac{\partial E}{\partial c_1} = 0.$$

First, compute

$$\begin{aligned}
\frac{\partial E}{\partial c_0} &= -2 \int_a^b f(x) \,\mathrm{d}x + 2c_0 (b - a) + c_1 (b^2 - a^2) \\
\frac{\partial E}{\partial c_1} &= -2 \int_a^b x f(x) \,\mathrm{d}x + c_0 (b^2 - a^2) + c_1 \tfrac{2}{3} (b^3 - a^3).
\end{aligned}$$

Setting these partial derivatives equal to zero yields

$$\begin{aligned}
2c_0 (b - a) + c_1 (b^2 - a^2) &= 2 \int_a^b f(x) \,\mathrm{d}x \\
c_0 (b^2 - a^2) + c_1 \tfrac{2}{3} (b^3 - a^3) &= 2 \int_a^b x f(x) \,\mathrm{d}x.
\end{aligned}$$

These equations, linear in the unknowns c_0 and c_1, can be written in the matrix form

$$\begin{bmatrix} 2(b-a) & b^2 - a^2 \\ b^2 - a^2 & \frac{2}{3}(b^3 - a^3) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} 2 \int_a^b f(x) \,\mathrm{d}x \\ 2 \int_a^b x f(x) \,\mathrm{d}x \end{bmatrix}.$$

When $b \neq a$ this system always has a unique solution. The resulting c_0 and c_1 are the coefficients for the monomial-basis expansion of the least squares approximation $P_* \in \mathcal{P}_1$ to f on $[a, b]$.

Example 2.4 ($f(x) = \mathrm{e}^x$)**.** Apply this result to $f(x) = \mathrm{e}^x$ for $x \in [0, 1]$. Since

$$\int_0^1 \mathrm{e}^x \,\mathrm{d}x = \mathrm{e} - 1, \qquad \int_0^1 x\mathrm{e}^x \,\mathrm{d}x = \big[\mathrm{e}^x(x-1)\big]_{x=0}^{1} = 1,$$

we must solve the system

$$\begin{bmatrix} 2 & 1 \\ 1 & \frac{2}{3} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} 2\mathrm{e} - 2 \\ 2 \end{bmatrix}.$$

The desired solution is

$$c_0 = 4\mathrm{e} - 10, \qquad c_1 = 18 - 6\mathrm{e}.$$

Figure 2.7 compares f to this least squares approximation P_* and the minimax approximation p_* computed earlier.

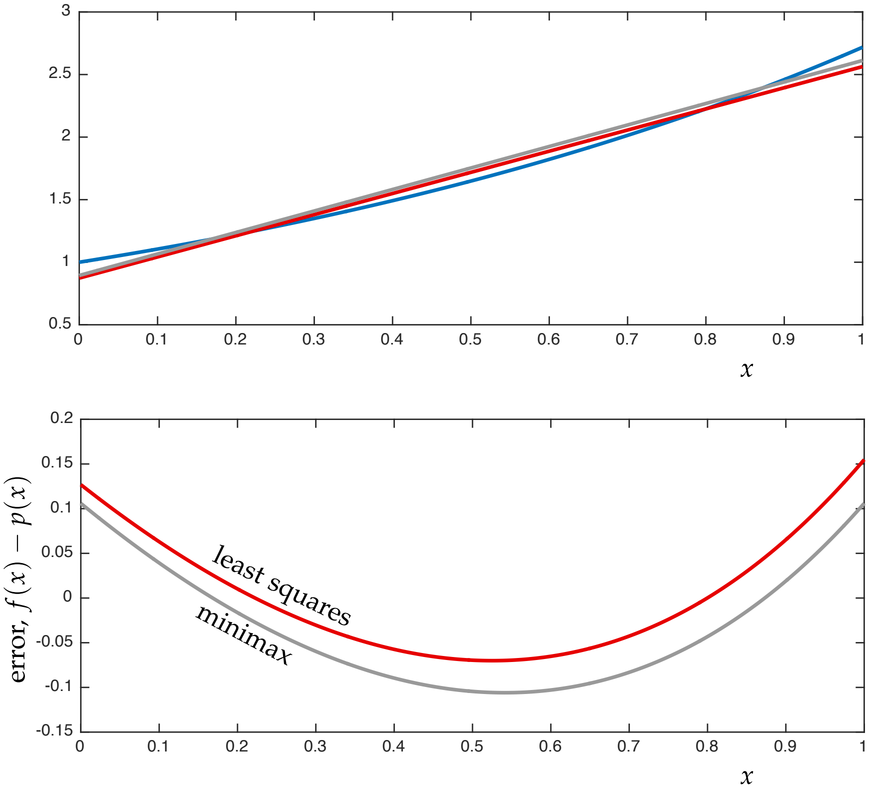

Figure 2.7: Top: Approximation of $f(x) = \mathrm{e}^x$ (blue) over $x \in [0, 1]$ via least squares (P_*, shown in red) and minimax (p_*, shown as a gray line).

Bottom: Error curves for least squares, $f - P_*$ (red), and minimax, $f - p_*$ (gray) approximation. While the curves have similar shape, note that the red curve does not attain its maximum deviation from f at $n + 2 = 3$ points, while the gray one does.

We can see from the plots in Figure 2.7 that the approximation looks decent to the eye, but the error is not terribly small. We can decrease that error by increasing the degree of the approximating polynomial. Just as we used a 2-by-2 linear system to find the best linear approximation, a general $(n+1)$-by-$(n+1)$ linear system can be constructed to yield the degree-n least squares approximation.

In fact, $\|f - P_*\|_2 = 0.06277\ldots$. This is indeed smaller than the 2-norm error of the minimax approximation p_*: $\|f - p_*\|_2 = 0.07228\ldots$.

2.4.3 *General polynomial bases*

Note that we performed the above minimization in the monomial basis: $p(x) = c_0 + c_1 x$ is a linear combination of 1 and x. Our experience with interpolation suggests that different choices for the basis may yield approximation algorithms with superior numerical properties. Thus, we develop the form of the approximating polynomial in an arbitrary basis.

Suppose $\{\phi_k\}_{k=0}^n$ is a basis for $\mathcal{P}_n$. Any $p \in \mathcal{P}_n$ can be written as

$$p(x) = \sum_{k=0}^{n} c_k \phi_k(x).$$

The error expression takes the form

$$\begin{aligned}E(c_0,\ldots,c_n) := \|f(x) - p(x)\|_{L^2}^2 = \int_a^b \Big(f(x) - \sum_{k=0}^{n} c_k\phi_k(x)\Big)^2\,\mathrm{d}x \\ = \langle f,f\rangle - 2\sum_{k=0}^{n} c_k\langle f,\phi_k\rangle + \sum_{k=0}^{n}\sum_{\ell=0}^{n} c_k c_\ell \langle \phi_k,\phi_\ell\rangle.\end{aligned}$$

To minimize E, we seek critical values of $\mathbf{c} = [c_0,\ldots,c_{n+1}]^T \in \mathbb{R}^{n+1}$, i.e., we want coefficients where the gradient of E with respect to $\mathbf{c}$ is zero: $\nabla_{\mathbf{c}}\, E = \mathbf{0}$. To compute this gradient, evaluate $\partial E/\partial c_j$ for $j = 0,\ldots,n$:

$$\begin{aligned}\frac{\partial E}{\partial c_j} &= \frac{\partial}{\partial c_j}\langle f,f\rangle - \frac{\partial}{\partial c_j}\Big(2\sum_{k=0}^{n} c_k\langle f,\phi_k\rangle\Big) + \frac{\partial}{\partial c_j}\Big(\sum_{k=0}^{n}\sum_{\ell=0}^{n} c_k c_\ell\langle\phi_k,\phi_\ell\rangle\Big)\\ &= 0 - 2\langle f,\phi_j\rangle + \frac{\partial}{\partial c_j}\Big(c_j^2\langle\phi_j,\phi_j\rangle + \sum_{\substack{k=0\\k\neq j}}^{n} c_k c_j\langle\phi_k,\phi_j\rangle + \sum_{\substack{\ell=0\\\ell\neq j}}^{n} c_j c_\ell\langle\phi_j,\phi_\ell\rangle + \sum_{\substack{k=0\\k\neq j}}^{n}\sum_{\substack{\ell=0\\\ell\neq j}}^{n} c_k c_\ell\langle\phi_k,\phi_\ell\rangle\Big)\end{aligned}$$

In this last line, we have broken the double sum on the previous line into four parts: one that contains c_j^2, two that contain c_j ($c_k c_j$ for $k \neq j$; $c_j c_\ell$ for $\ell \neq j$), and one (the double sum) that does not involve c_j at all. This decomposition makes it easier to compute the derivative:

$$\begin{aligned}&\frac{\partial}{\partial c_j}\Big(c_j^2\langle\phi_j,\phi_j\rangle + \sum_{\substack{k=0\\k\neq j}}^{n} c_k c_j\langle\phi_k,\phi_j\rangle + \sum_{\substack{\ell=0\\\ell\neq j}}^{n} c_j c_\ell\langle\phi_j,\phi_\ell\rangle + \sum_{\substack{k=0\\k\neq j}}^{n}\sum_{\substack{\ell=0\\\ell\neq j}}^{n} c_k c_\ell\langle\phi_k,\phi_\ell\rangle\Big)\\ &\qquad = 2c_j\langle\phi_j,\phi_j\rangle + \sum_{\substack{k=0\\k\neq j}}^{n} c_k\langle\phi_k,\phi_j\rangle + \sum_{\substack{\ell=0\\\ell\neq j}}^{n} c_\ell\langle\phi_j,\phi_\ell\rangle + 0\\ &\qquad = 2c_j\langle\phi_j,\phi_j\rangle + 2\sum_{\substack{k=0\\k\neq j}}^{n} c_k\langle\phi_k,\phi_j\rangle.\end{aligned}$$

These terms contribute to $\partial E/\partial c_j$ to give

$$\frac{\partial E}{\partial c_j} = -2\langle f, \phi_j\rangle + 2\sum_{k=0}^{n} c_k\langle \phi_k, \phi_j\rangle. \tag{2.12}$$

To minimize E, set $\partial E/\partial c_j = 0$ for $j = 0, \ldots, n$, which gives the $n+1$ equations

$$\sum_{k=0}^{n} c_k\langle \phi_k, \phi_j\rangle = \langle f, \phi_j\rangle, \quad j = 0, \ldots, n, \tag{2.13}$$

in the $n+1$ unknowns $c_0, \ldots, c_n$. Since these equations are linear in the unknowns, write them in matrix form:

$$\begin{bmatrix} \langle \phi_0, \phi_0\rangle & \langle \phi_0, \phi_1\rangle & \cdots & \langle \phi_0, \phi_n\rangle \\ \langle \phi_1, \phi_0\rangle & \langle \phi_1, \phi_1\rangle & & \vdots \\ \vdots & & \ddots & \vdots \\ \langle \phi_n, \phi_0\rangle & \langle \phi_n, \phi_1\rangle & \cdots & \langle \phi_n, \phi_n\rangle \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \langle f, \phi_0\rangle \\ \langle f, \phi_1\rangle \\ \vdots \\ \langle f, \phi_n\rangle \end{bmatrix},$$

which we denote $\mathbf{Gc} = \mathbf{b}$. The matrix $\mathbf{G}$ is called the *Gram matrix*. Using this matrix-vector notation, we can accumulate the partial derivatives formulas (2.12) for E into the gradient

$$\nabla_{\mathbf{c}} E = 2\big(\mathbf{Gc} - \mathbf{b}\big).$$

Since $\mathbf{c}$ is a critical point if and only if $\nabla_{\mathbf{c}} E(\mathbf{c}) = \mathbf{0}$, we must ask:

- How many critical points are there? Equivalently, how many $\mathbf{c}$ solve $\mathbf{Gc} = \mathbf{b}$?
- If $\mathbf{c}$ is a critical point, is it a (local or even global) minimum?

We will answer the first question by showing that $\mathbf{G}$ is invertible, and hence E has a unique critical point. To answer the second question, we must inspect the Hessian

$$\nabla^2_{\mathbf{c}} E = \nabla_{\mathbf{c}}(\nabla_{\mathbf{c}} E) = 2\mathbf{G}.$$

The critical point $\mathbf{c}$ is local minimum if and only if the Hessian is *symmetric positive definite*.

A matrix $\mathbf{G}$ is *positive definite* provided $\mathbf{z}^*\mathbf{Gz} > 0$ for all $\mathbf{z} \neq \mathbf{0}$.

The symmetry of the inner product implies $\langle \phi_j, \phi_k\rangle = \langle \phi_k, \phi_j\rangle$, and hence $\mathbf{G}$ is symmetric. (In this case, symmetry also follows from the equivalence of mixed partial deritivates.) The following theorem confirms that $\mathbf{G}$ is indeed positive definite.