Approximation Theory

Function approximation is the task of constructing, for a given function, a simpler function so that the difference between the two functions is small and to then provide a quantifiable estimate for the size of the difference.

Why would one want to do this? Consider the evaluation of the integral

$$\int_0^1 e^{x^2} dx.$$

The antiderivative of e^{x^2} cannot be expressed in terms of simple functions, i.e., in terms of powers of x, trigonometric functions, exponential functions, ... Suppose we could find a function $p(x)$ that is "close" to e^{x^2} over the interval $[0, 1]$ and which has an easy to define antiderivative $F(x)$. Then, we can use $p(x)$ as a surrogate for $f(x)$ and then approximate the integral of $f(x)$ by the integral of $p(x)$, i.e., we have

$$\int_0^1 e^{x^2} dx \approx \int_0^1 p(x) dx = F(1) - F(0).$$

In the calculus, one learns a particular way to define a simple function $p(x)$: use the Maclaurin series, i.e., the Taylor series about the point $x = 0$, for e^{x^2} which is given by

$$e^{x^2} = 1 + x^2 + \frac{1}{2}x^4 + \frac{1}{6}x^6 + \cdots = \sum_{j=0}^{\infty} \frac{1}{j!}x^{2j}.$$

We approximate e^{x^2} by keeping only the first $n + 1$ terms in the series:

$$e^{x^2} \approx p(x) = \sum_{j=0}^{n} \frac{1}{j!}x^{2j}.$$

We then replace the integrand e^{x^2} by $\sum_{j=0}^{n} x^{2j}/j!$ to obtain the approximation

$$\int_0^1 e^{x^2}dx \approx \sum_{j=0}^{n} \int_0^1 \frac{1}{j!}x^{2j}dx = \sum_{j=0}^{n} \frac{1}{j!}\left(\frac{x^{2j+1}}{2j+1}\Big|_0^1\right) = \sum_{j=0}^{n} \frac{1}{j!(2j+1)}.$$

For $n = 2$, we then have

$$\int_0^1 e^{x^2}dx \approx 1 + \frac{1}{3} + \frac{1}{10} = \frac{43}{30} = 1.433333\ldots.$$

How big an error are we making by replacing the integral $\int_0^1 e^{x^2} dx$ by $F(1) - F(0)$? That is, how close is $\frac{43}{30}$ to the correct answer?

We do not know the exact answer but MATLAB has a very good method for approximating integrals. The MATLAB approximation is

$$\int_0^1 e^{x^2} dx \approx 1.46265.$$

Thus, the error in our approximation is about $1.46265 - 1.43333 = 0.02932$.

What about keeping fewer or more terms in the Taylor series? We have that

$$\int_0^1 e^{x^2} dx \approx \begin{cases} 1 + \frac{1}{3} = \frac{4}{3} \approx 1.3333333 & \text{for } n = 1 \\ 1 + \frac{1}{3} + \frac{1}{10} + \frac{1}{42} = \frac{51}{35} \approx 1.4571429 & \text{for } n = 3 \\ 1 + \frac{1}{3} + \frac{1}{10} + \frac{1}{42} + \frac{1}{216} = \frac{22039}{15120} \approx 1.4617726 & \text{for } n = 4 \end{cases}$$

The corresponding errors are given in the table below, from which we observe

that the approximation seems to get better as we increase n.

n	error	relative error $= \frac{\text{error}}{\text{exact answer}}$
0	0.46265	32%
1	0.12932	8.8%
2	0.02932	2.0%
3	0.00551	0.38%
4	0.00088	0.06%

□

In this example, we use MATLAB to presumably obtain something very close to the exact value of the integral, and then use that "exact" answer to determine the error in the value obtained using Talyor polynomial approximations of the integrand.

What if we do not have[1] MATLAB? Can we get some idea about the size of the error we make by approximating the integral using Taylor polynomials? We can indeed do this, but, instead of doing so for the integral of a function $f(x)$, we do it directly for the function $f(x)$ itself.

[1] Of course, for this simple problem, if we have MATLAB we would just go ahead and use it to approximate the integral and not bother with the Taylor polynomial approximation! But here we are trying to learn how functions are approximated, which includes learning something about how MATLAB does it.

From the calculus, we recall the following: given the interval $[a, b]$, a point $x_0 \in [a, b]$, and a function $f(x)$ possessing $(n + 1)$ continuous derivatives for all $x \in [a, b]$, we then have Taylor's formula

$$f(x) = T_n(x) + R_n(x), \tag{1}$$

where $T_n(x)$ denotes the Taylor polynomial of degree less than[2] or equal to n given by

$$\begin{aligned} T_n(x) &= f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \cdots + \frac{1}{n!}f^n(x_0)(x - x_0)^n \\ &= \sum_{j=0}^{n} \frac{1}{j!} f^{(j)}(x_0)(x - x_0)^j \quad \text{for all } x \in [a, b],, \end{aligned} \tag{2}$$

where $f^{(n)}(x)$ denotes the n^{th} derivative of $f(x)$. The remainder term $R_n(x)$ is given

$$R_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\theta)(x - x_0)^{n+1} \quad \text{for some } \theta \in [a, b] \text{ and for all } x \in [a, b]. \tag{3}$$

[2]We have to say "less than or equal to n" because it is possible that $f^{(n)}(x_0) = 0$ in which case $Tt_n(x)$ is of degree at most $(n - 1)$. Of course, if in *addition* $f^{(n-1)}(x_0) = 0$, then $T_n(x)$ is of degree at most $(n - 2)$, and so on, i.e., it is possible for $T_n(x)$ to be of any degree between 0 and n, or to even be identically zero for all $x \in [a, b]$.

From (1), (2), and (3), we have that

$$\left|f(x) - T_n(x)\right| = \left|R_n(x)\right| = \frac{1}{(n+1)!}|f^{(n+1)}(\theta)|\,|(x-x_0)^{n+1}|. \qquad (4)$$

Clearly, we have that

$$|f^{n+1}(\theta)| \leq \max_{x\in[a,b]} |f^{(n+1)}(x)| \quad \text{for any } \theta \in [a,b].$$

Then, (4) is now given by

$$\left|f(x) - T_n(x)\right| = \frac{1}{(n+1)!} \max_{x\in[a,b]} |f^{(n+1)}(x)|\,|(x-x_0)^{n+1}|$$

so that

$$\max_{x\in[a,b]} \left|f(x) - T_n(x)\right| = \frac{1}{(n+1)!} \max_{x\in[a,b]} |f^{(n+1)}(x)| \max_{x\in[a,b]} |(x-x_0)^{n+1}|\,. \qquad (5)$$

Because $x_0 \in [a,b]$, we have that

$$|(x-x_0)^{n+1}| \leq (b-a)^{n+1} \quad \text{for all } x \in [a,b]$$

so that

$$\max_{x\in[a,b]} |(x-x_0)^{n+1}| \leq (b-a)^{n+1}. \qquad (6)$$

The substitution of (6) into (5) then yields

$$\max_{x\in[a,b]} \left|f(x) - T_n(x)\right| \leq \frac{1}{(n+1)!} \max_{x\in[a,b]} |f^{(n+1)}(x)|(b-a)^{n+1}. \qquad (7)$$

(7) provides an *estimate* for the *error*, i.e., for the maximum value,[3] over the interval $[a, b]$, of the difference between a function $f(x)$ having at least $(n+1)$ continuous derivatives on an interval $[a, b]$ and its Taylor polynomial $T_n(x; x_0)$ of degree less than or equal to n.

Example Let $f(x) = e^x$, $a = -1$, $b = 1$, and $x_0 = 0$. We then have that the Taylor polynomial approximation of e^x given by

$$e^x \approx t_n[e^x] = \sum_{j=1}^{n} \frac{1}{j!} x^j.$$

We have $f^{(n)}(x) = e^x$ for all x so that $\max_{x\in[a,b]} |f^{(n+1)}(x)| = e$, We also have that $(b-a)^{n+1} = 2^{n+1}$. Then, from (7), we have the error estimate

$$\max_{x\in[a,b]} \left|e^x - T_n[e^x]\right| \leq e\frac{2^{n+1}}{(n+1)!}.$$

[3] As discussed below, using the maximum value of the difference in not the only way to define the error.

Because $\frac{2^{n+1}}{(n+1)!} \to 0$ as $n \to \infty$, we see that we can make the error as small as we like by letting n, the degree of the Taylor polynomial, be large enough. This desirable result does not always hold for other functions. $\square$

Best approximations

We have touched on two approaches (Taylor polynomials and interpolants) for approximating a given function $f(x)$. One may ask if either of these approaches produces an approximating function that is the best possible approximation of the given function, where now we know two ways of measuring what is "best."

For example, we ask if one can construct a polynomial $p_n^{*,2}(x)$ of degree less than or equal to n such that

$\|f(x)-p_n^{*,2}(x)\|_2 \leq \|f(x)-p(x)\|_2$ for all polynomials $p(x)$ of degree n or less

That is, $p_n^{*,2}$ is the polynomial of degree $\leq n$ which approximates $f(x)$ better than all other polynomials in $\mathcal{P}^n$ when measured in the L^2 norm.

Alternately, we could measure the difference in the L^∞ norm. Then we ask if we can construct a polynomial $p_n^{*,\infty}(x)$ of degree less than or equal to n such that

$\|f(x)-p_n^{*,\infty}(x)\|_\infty \leq \|f(x)-p(x)\|_\infty$ for all polynomials $p(x)$ of degree n or less.

For obvious reasons, polynomials such as $p_n^{*,2}(x)$ and $p_n^{*,\infty}(x)$ are referred to as best approximations of the function $f(x)$.

$p_n^{*,2}(x)$ is referred to as the best $L^2(a,b)$ or least-squares approximation of $f(x)$.

$p_n^{*,\infty}(x)$ is referred to as the best $L^\infty[a,b]$ or best uniform or best pointwise approximation of $f(x)$.

Because we have that

$$\begin{aligned}\|f(x) - p_n^{*,\infty}(x)\|_\infty &= \min_{p\in\mathcal{P}_n} \|f(x) - p(x)\|_\infty \\ &= \min_{p\in\mathcal{P}_n} \max_{x\in[a,b]} |f(x) - p_n(x)|,\end{aligned}$$

$p_n^{*,\infty}(x)$ is also referred to as the best min-max approximation of $f(x)$.

Note that, in general, $p_n^{*,2}(x) \neq p_n^{*,\infty}(x)$ so that what is "best" depends on how we choose to measure the error.

Because of the definition of best approximations, we have that the error in the Taylor polynomial or an interpolant for a function $f(x)$ is always greater or equal

than the error in the best approximation, e.g., we have that

$$\|f(x) - p_n^{*,\infty}(x)\|_\infty \leq \|f(x) - T_n(x)\|_\infty,$$

where $T_n(x)$ denotes the Taylor polynomial of degree less than or equal to n, and

$$\|f(x) - p_n^{*,\infty}(x)\|_\infty \leq \|f(x) - p_n(x)\|_\infty,$$

where $p_n(x)$ denotes the interpolating polynomial of degree less than or equal to n.

Best approximations can also be defined for discrete data, i.e., when all that is available is the number f_j corresponding to the point x_j, $j = 1, \ldots, m$.

The Weierstrass theorem

One of the most important theorems in approximation theory, indeed, in all of mathematics, is the Weierstrass approximation theorem.

Weierstrass Theorem. Given any function $f(x)$ that is continuous on the interval $[a, b]$ and given any $\varepsilon > 0$, there exists a polynomial $p(x)$ such that

$$\|f(x) - p(x)\|_\infty < \varepsilon. \qquad \square$$

Thus, no matter how small one chooses ε, there exists a polynomial whose value at every point in the interval $[a, b]$ is within ε of the value of $f(x)$ at that point. Polynomial interpolation and best approximation are motivated by the desire to construct polynomials that do well at approximating continuous functions.

General Results about Best Approximation

We collect some general results about best approximation. The context is functions belonging to a normed linear vector space V over the reals $\mathbb{R}$, not necessarily finite dimensional.

Recall that a *norm* on V is a function $\|\cdot\|$ that maps elements of V to $[0,\infty)$ having the properties

(i) $\|v\| \geq 0$ for $v \in V$ with $\|v\| = 0$ if and only if $v = 0$

(ii) $\|\lambda v\| = |\lambda|\|v\|$ for all $v \in V$ and $\lambda \in \mathbb{R}$

(iii) the triangle inequality: $\|u+v\| \leq \|u\| + \|v\|$ for all $u, v \in V$.

The *distance* between two elements $u, v \in V$ is defined by the nonnegative number $\|u - v\|$.

In the following example and throughout, $C[a,b]$ denotes the set of continuous function $f(x)$ defined on the interval $[a,b]$.

Example On $C[a,b]$, we define the $L^\infty[a,b]$ norm

$$\|f\|_\infty = \max_{a\le x\le b} |f(x)|.$$

Alternately, on $V = C[a,b]$, we define the $L^2(a,b)$ norm

$$\|f\|_2 = \left(\int_a^b \big(f(x)\big)^2 \, dx \right)^{1/2},$$

or more generally, for $1 \le p < \infty$, the $L^p(a,b)$ norm

$$\|f\|_p = \left(\int_a^b \big(f(x)\big)^p \, dx \right)^{1/p}.$$

$C[a,b]$ equipped with any of these norms define a normed linear vector space. □

Let W denote a subset of V. Then, the best approximation problem is defined as follows.

Definition. Best approximation in a normed linear space. Given any $v \in V$, find a $w^* \in W$ such that the distance between v and w^* is least among all $w \in W$, i.e., such that $\|v - w^*\| \leq \|v - w\|$ for all $w \in W$. Such a w^* is referred to as a the best approximation to $v \in V$ out of $W \subset V$. $\square$

Naturally, one wants to know if a best approximation exists and is unique. By imposing an additional condition on W, the existence of a best approximation can be proved.

Theorem. Existence of the best approximation. Let V denote a normed linear space and let W denote a finite-dimensional subspace of V. Then, given any $v \in V$, there exists a $w^* \in W$ such that

$$\|v - w^*\| \leq \|v - w\| \quad \textit{for all } w \in W. \qquad \square$$

The requirement that W be finite dimensional is essential, i.e., the result is not true in general if W is infinite dimensional.

Example Let $V = C[a,b]$ equipped with the $\|f\|_\infty$ norm. Let $W = \mathcal{P}_n$, the space of all polynomials of degree less that or equal to n, restricted to the interval $[a,b]$. Clearly, W is a finite-dimensional subspace of $C[a,b]$. Then, the theorem tells us that given any $f(x) \in C[a,b]$, there exists a $p_n^{*,\infty}(x) \in \mathcal{P}_n$ such that $\|f - p_n^{*,\infty}\|_\infty \leq \|f - p_n\|_\infty$ for all $p_n(x) \in \mathcal{P}_n$, i.e., such that $\max_{a\leq x\leq b} |f - p_n^{*,\infty}| \leq \max_{a\leq x\leq b} |f - p_n|$ for all $p_n(x) \in \mathcal{P}_n$.

Alternately, for $1 \leq p < \infty$, let $V = C[a,b]$ equipped with the $\|f\|_p$ norm. Then, the theorem tells us that given any $f(x) \in C[a,b]$, there exists a $p_n^{*,p}(x) \in \mathcal{P}_n$ such that $\|f - p_n^{*,p}\|_p \leq \|f - p_n\|_p$ for all $p_n(x) \in \mathcal{P}_n$. Of course, $p_n^{*,p}$ is different for all values of p. □

We now consider the uniqueness of the best approximations. To this end, we say that a vector space V has a strictly convex norm if the closed ball $\{v \in V \mid \|v\| \leq 1\}$ which is a convex set is strictly convex, i.e., if $u, v \in V$ are such that $\|u\| = 1$ and $\|v\| = 1$, then $\|\theta u + (1-\theta)v\| < 1$ for all $0 < \theta < 1$. We then have the following result.

Theorem. Uniqueness result for best approximations. Let V denote a normed linear space with a strictly convex norm. Let W denote a finite-dimensional subspace of V. Then, given any $v \in V$, there exists at most one best approximation to v out of W. $\square$

Example The $L^\infty[a,b]$ norm is not strictly convex. To see this consider a specific choice of $u, v \in L^\infty[0,1.$] Let

$$u(x) = x \quad \text{and} \quad v(x) = x^2 \quad \text{on } [0,1].$$

Clearly

$$\|u\|_{L^\infty[0,1]} = 1 \quad \text{and} \quad \|v\|_{L^\infty[0,1]} = 1.$$

So now we want to show that there is a constant $0 < \theta < 1$ such that the set is not strictly convex, i.e., that $\|\theta u + (1-\theta)v\| \not\leq 1$. If we take $\theta = 1/2$ then we have

$$\|\frac{1}{2}(u+v)\|_{L^\infty[0,1]} = \frac{1}{2} \max_{0 \leq x \leq 1} |(x + x^2)\| = 1.$$

Thus, the theorem does not provide any information about the uniqueness of the best uniform approximation in that norm. It turns that the approximation is unique, but that has to proved by other means.

On the other hand, it can be shown that the $L^2(a,b)$ norm is strictly convex so that the theorem tells us that the least-squares approximation is unique. In fact, this is true for any $L^p(a,b)$ norm, so long as $1 < p < \infty$. However, the $L^1(a,b)$ norm is not strictly convex; again, by other means, one can prove that the best $L^1(a,b)$ approximation is unique. $\square$

Best Approximation in Inner Product Spaces

If the space V is an *inner product space*, then a complete analysis of the best approximation problem in finite-dimensional subspaces can be given. With V denoting a vector space over $\mathbb{R}$, an inner product on V is a map[4] $(\cdot,\cdot) : V \times V \to \mathbb{R}$ having the following properties:

(i) $(v,v) \geq 0$ for all $v \in V$ and $(v,v) = 0$ if and only if $v = 0$

(ii) $(u,v) = (v,u)$ for all $u, v \in V$

(iii) $(cu + dv, w) = c(u,w) + d(v,w)$ for all $u, v, w \in V$ and all $c, d \in \mathbb{R}$.

It can then be shown that

- the mapping $\|v\| = (v,v)^{1/2}$ from V to $\mathbb{R}$ is a norm on V, referred to as the norm induced by the inner product $(\cdot,\cdot)$, or more succinctly, as an induced norm on V;

[4] The map assigns to every pair of elements $u, v \in V$ a real number.

- the induced norm is strictly convex; and
- the Cauchy-Schwarz inequality $|(u, v)| \leq \|u\|\|v\|$ holds for all $u, v \in V$.

In general, we refer to $u, v \in V$ as being orthogonal if and only if $(u, v) = 0$.

From the results of the previous section, it immediately follows that if W is a finite-dimensional subspace of V, then, for any $v \in V$, there exists a unique best approximation to V out of W with respect to the induced norm $\|v\| = (v, v)^{1/2}$. For inner product spaces, we have an important additional results about the best approximation.

Theorem. Best Approximation on an Inner Product Space Let W denote a finite-dimensional subspace of an inner product space V. Then, for any $v \in V$, $w^* \in W$ is the best approximation to v out of W with respect to the norm $\|v\| = (v, v)^{1/2}$ if and only if

$$(v - w^*, w) = 0 \quad \text{for all } w \in W, \tag{8}$$

i.e., the difference between v and w^* (or the error $v - w^*$) is orthogonal to all functions belonging to the subspace W. $\square$

Example The space $V = C[a, b]$ equipped with the inner product

$$(f, g) = \int_a^b f(x)g(x)\,dx \quad \text{for all } f, g \in C[a, b] \tag{9}$$

is an inner product space. The associated induced norm is, of course, the $L^2(a, b)$ norm

$$\|f\|_{L^2(a,b)} = \Big(\int_a^b \big(f(x)\big)^2 dx \Big)^{1/2}.$$

$L^2(a,b)$ is very special among the $L^p(a,b)$ spaces, $0 \leq p \leq \infty$, in that $L^2(a,b)$ *is the only* $L^p(a,b)$ *space that is an inner product space.* $\square$

Construction of the Best Approximation in an Inner Product Space

The characterization (8) of best approximations in inner product spaces results in a very straightforward means for the *construction* of the best approximation.

Let W denote a finite-dimensional subspace of an inner product space V and let $\{w_i\}_{j=1}^m$ denote a basis for W. Necessarily, given $v \in V$, we have that the best approximation $w^*(x)$ of out W can be written as

$$w^*(x) = \sum_{j=1}^{m} c_j w_j(x). \tag{10}$$

It then follows from (8) that

$$0 = (v - w^*, w_k) = \Big(v - \sum_{j=1}^{m} c_j w_j \,,\, w_k\Big) = (v, w_k) - \sum_{j=1}^{m} c_j (w_j, w_k) \quad \text{for } k = 1, \ldots, m$$

so that

$$\sum_{j=1}^{m} c_j(w_j, w_k) = (v, w_k) \quad \text{for } k = 1, \ldots, m. \tag{11}$$

Let the $m \times m$ matrix G have entries

$$G_{kj} = (w_j, w_k) \quad \text{for } j, k = 1, \ldots, m,$$

and the vectors $\vec{c} \in \mathbb{R}^m$ and $\vec{b} \in \mathbb{R}^m$ have entries

$$(\vec{b})_j = c_j \quad \text{for } j = 1, \ldots, m \qquad (\vec{b})_k = (v, w_k) \quad \text{for } k = 1, \ldots, m.$$

respectively. Then, clearly, () is equivalent to the linear algebraic system

$$G\vec{c} = \vec{b}. \tag{12}$$

The matrix G is referred to as the Gram matrix for the basis $\{w_i\}_{j=1}^m$ and the linear algebraic system (12) is referred as the normal equations. Because of property (ii) of the inner product, G is clearly a symmetric matrix. It is also a positive definite matrix. To see the latter, let $\vec{c} = (c_1, \ldots, c_m)^T$ denote a vector

in $\mathbb{R}^m$. Then,

$$\vec{c}^T G\vec{c} = \sum_{j,k=1}^{m} G_{jk}c_j c_k = \sum_{j,k=1}^{m} (w_k, w_j)c_j c_l = \Big(\sum_{k=1}^{m} c_k w_k\,,\, \sum_{j=1}^{m} c_j w_j\Big)$$
$$= \Big\|\sum_{j=1}^{m} c_j w_j\Big\|^2 > 0 \quad \text{for } \vec{c} \neq \vec{0}.$$

The components $c_1, \ldots, c_m$ of $\vec{c}$ are determined by solving (12) and then the best approximation is given by $w^*(x) = \sum_{j=1}^{m} c_j w_j(x)$. Because G is a positive definite matrix, the system (12) uniquely determines the coefficients $c_1, \ldots, c_m$ and therefore uniquely determines the best approximation $w^*(x) = \sum_{j=1}^{m} c_j w_j(x)$. Thus, this is an alternate, constructive means of proving the existence and uniqueness of the best approximation in inner product spaces.

We provide two examples of how best approximations in inner product spaces are constructed.

Example Let $V = C[0,1]$ and let $W = \mathcal{P}_n$, the space of polynomials of degree less than or equal to n, and chose the monomial basis $\{x^j\}_{j=0}^{n}$ for W, i.e., we

have $w_j(x) = x^j$ for $j = 0, 1, \ldots, n$. We use the inner product given in (9) and then construct the best approximation of a given function $f(x) \in C[0, 1]$. We have that

$$G_{jk} = \int_0^1 x^{j+k} dx = \frac{1}{j+k+1} \qquad \text{for } j, k = 0, \ldots, n \tag{13}$$

and

$$b_k = \int_0^1 x^k f(x) dx \qquad \text{for } k = 0, \ldots, n.$$

The best approximation is then determined by solving the linear system (12) with these choices for G and $\vec{b}$.

Unfortunately, the matrix G given in (13), known as the *Hilbert matrix*, is very ill conditioned; see the table in linear algebra notes.

This rapid growth in the condition number of the Gram matrix for the monomial basis motivates the upcoming discussion for orthogonal polynomials. □

Example Let $V = C[-\pi, \pi]$ equipped with the inner product

$$(f, g) = \int_{-\pi}^{\pi} f(\theta) g(\theta) d\theta \qquad \text{for all } f, g \in C[-\pi, \pi]. \tag{14}$$

We consider finite-dimensional subspace $W = \mathcal{S}_n$ of $C[-\pi, \pi]$ of dimension $2n+1$ spanned by the basis $\{\cos i\theta\}_{i=0}^{n} \cup \{\sin k\theta\}_{i=1}^{n}$ This subspace is referred as the space of *trigonometric polynomials* of degree less than or equal to n. We then have the basis

$$w_j(\theta) = \begin{cases} \cos j\theta & \text{for } j = 0, \ldots, n \\ \sin(j-n)\theta & \text{for } j = n+1, \ldots, 2n. \end{cases} \tag{15}$$

The Gram matrix associated with this basis is given by

$$G_{jk} = (w_k, w_j) = \int_{-\pi}^{\pi} w_j(\theta) w_k(\theta) d\theta = \begin{cases} \pi & \text{if } j = k \neq 0 \\ 2\pi & \text{if } j = k = 0 \\ 0 & \text{if } j \neq k. \end{cases}$$

Thus, G is a diagonal matrix and has inverse

$$(G^{-1})_{jk} = \begin{cases} \dfrac{1}{\pi} & \text{if } j = k \neq 0 \\ \dfrac{1}{2\pi} & \text{if } j = k = 0 \\ 0 & \text{if } j \neq k. \end{cases}$$

Therefore, the best approximation $s_n^*(\theta) \in \mathcal{S}_n$ to $f(\theta) \in C[-\pi, \pi]$ with respect

to the norm induced by the inner product (14) is given by

$$s_n^*(\theta) = \frac{a_0}{2} + \sum_{k=1}^{n} \big(a_k \cos k\theta + b_k \sin k\theta\big), \tag{16}$$

where

$$\begin{aligned} a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\cos k\theta d\theta \qquad \text{for } k = 0, \ldots, n \\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\sin k\theta d\theta \qquad \text{for } k = 1, \ldots, n. \end{aligned} \tag{17}$$

The best approximation (16) is a truncated *Fourier series* for f and the coefficients given in (17) are the *Fourier coefficients.* Likewise, the basis $\{\cos k\theta\}_{k=0}^{n} \cup \{\sin k\theta\}_{k=1}^{n}$, or equivalently (15), is referred as the *Fourier basis.* □

Least-squares Polynomial Approximation

We looked at least-squares approximation of a discrete set of data which uses the ℓ_2-norm, i.e., the Euclidean norm on $\mathbb{R}^n$. Now we consider the case where we have a function $f(x)$ and approximate it using the $L^2(\Omega)$-norm.

The conditioning problem for the Gram matrix corresponding to the monomial basis motivates consideration of alternate bases that are better conditioned. In this section we will consider bases that result in *diagonal* Gram matrices with positive diagonal entries so that, because such matrices are trivial to invert, conditioning is not an issue.

We also consider best approximations with respect to a generalization of the $L^2(a,b)$ norm. Specifically, let $\rho(x)$ denote a function such that $\rho(x) \geq 0$ for all $x \in [a,b]$ with $\rho(x) = 0$ at most at a finite number of points in $[a,b]$. Also, assume $\rho(x)$ is integrable, i.e., there exists $\widetilde{\rho} < \infty$ such that such that

$\int_a^b \rho(x)dx = \widetilde{\rho}$. Then $C[a,b]$ equipped with the inner product

$$(f,g) = \int_a^b \rho(x)f(x)g(x)\,dx \quad \text{for all } f,g \in C[a,b] \tag{18}$$

is also an inner product space. The inner product defined in (18) is referred to as the *weighted $L^2_\rho(a,b)$ inner product on $[a,b]$ with respect to the weight $\rho(x)$*. It induces the norm

$$\|f\|_{L^2_\rho(a,b)} = \Big(\int_a^b \rho(x)\big(f(x)\big)^2 dx\Big)^{1/2} \quad \text{for all } f \in C[a,b]. \tag{19}$$

From our previous discussion, by choosing $V = C[a,b]$ equipped with the norm (19) and by also choosing $W = \mathcal{P}_n$, we have the following result.[5]

[5] To economize notation, we continue to use $\|f\|_2$ instead of $\|f\|_{L^2_\rho(a,b)}$. For the same reason, we continue to use $p_n^{*,2}(x)$ to denote the best $L^2_\rho(a,b)$ norm approximation.

Theorem. Continuous Least Squares Approximation. Given $f(x) \in C[a,b]$ and a weight $\rho(x)$ defined on $[a,b]$, there exists a unique $p_n^{*,2}(x) \in \mathcal{P}_n$, referred to as the best $L^2_\rho(a,b)$ or least-squares approximation to f out of $\mathcal{P}_n$, that satisfies

$$\|f - p_n^{*,2}\|_2 = \Big(\int_a^b \big(f(x) - p_n^{*,2}(x)\big)^2 dx \Big)^{1/2} \rho(x)dx < \|f - p_n\|_2$$

for all $p_n \in \mathcal{P}_n$ such that $p_n \neq p_n^{*,2}$.

Moreover, $p_n^{*,2}(x) \in \mathcal{P}_n$ is the least-squares approximation to $f \in C[a,b]$ out of $\mathcal{P}_n$ if and only if

$$\int_a^b \big(f(x) - p_n^{*,2}(x)\big) p_n(x) \rho(x) dx \qquad \text{for all } p_n(x) \in \mathcal{P}_n. \quad \square$$

Orthogonal Polynomial Bases

In this section, we consider orthogonal polynomial bases for $\mathcal{P}_n$; as we shall see, such bases are very well suited for determining best polynomial approximation with respect to the $L^2_\rho(a,b)$ inner product and norm given in (18) and (19).

Let $\{\Phi_j(x)\}_{j=0}^n = \{\Phi_0, \Phi_1, \ldots, \Phi_n\}$ denote a basis for $\mathcal{P}_n$ so that any $p_n \in \mathcal{P}_n$ can be written as $p_n(x) = \sum_{j=0}^n \lambda_j \Phi_j(x) = \lambda_0\Phi_0 + \lambda_1\Phi_1 + \cdots + \lambda_n\Phi_n$ for some constants $\{\lambda_0, \lambda_1, \ldots, \lambda_n\}$. To make the Gram matrix have condition number 1, one can construct a basis $\{\Phi_j(x)\}_{j=0}^n$ that is orthonormal with respect to the weight $\rho(x)$, i.e., such that

$$(\Phi_j, \Phi_k) = \int_a^b \Phi_j(x)\Phi_k(x)\rho(x)dx = 0 \qquad \text{for } j,k = 0,\ldots,n \text{ with } j \neq k \tag{20}$$

$$(\Phi_j, \Phi_j) = \|\Phi_j\|_2^2 = \int_a^b \big(\Phi_j(x)\big)^2\rho(x)dx = 1 \qquad \text{for } j = 0,\ldots,n, \tag{21}$$

or, more concisely, $(\Phi_j, \Phi_k) = \delta_{jk}$. The "ortho" part of orthonormal refers to

(20), i.e., to $\Phi_j(x)$ being orthogonal to $\Phi_k(x)$ for $j \neq k$, whereas the "normal" part refers to (21), i.e., to $(\Phi_j, \Phi_j) = 1$. The set $\{\Phi_j(x)\}_{j=0}^n = \{\Phi_0, \Phi_1, \ldots, \Phi_n\}$ satisfying (20) and (21) is referred to as the orthonormal basis for $\mathcal{P}_n$ with respect to the weight ρ.

Obviously, we have that the Gram matrix corresponding to the orthonormal basis $\{\Phi_j(x)\}_{j=0}^n$ satisfies $G_{jk} = (\Phi_j, \Phi_k) = \delta_{jk}$ for $j, k = 0, \ldots, n$, i.e., the coefficient matrix G in or (12) is the identity matrix.[6] As a result, if

$$p_n^{*,2}(x) = \sum_{j=0}^{n} \lambda_j \Phi_j(x) = \lambda_0 \Phi_0 + \lambda_1 \Phi_1 + \cdots + \lambda_n \Phi_n \tag{22}$$

denotes the least-squares approximation to a given function $f(x) \in C[a, b]$ out of $\mathcal{P}_n$, then we easily obtain that

$$\lambda_j = (f, \Phi_j) = \int_a^b f(x)\Phi_j(x)\rho(x)dx \qquad \text{for } j = 0, 1, \ldots, n. \tag{23}$$

[6] Thus we see that the orthonormal polynomial basis $\{\Phi_j(x)\}_{j=0}^n$ for $\mathcal{P}_n$ plays, for least-squares approximation, a similar role to the role played by the Lagrange fundamental polynomial basis $\{\ell_j(x)\}_{j=0}^n$ for Lagrange interpolation.

Thus, the least-squares approximation out of $\mathcal{P}_n$ is given by

$$p_n^{*,2}(x) = \sum_{j=0}^{n} (f, \Phi_j)\Phi_j(x). \tag{24}$$

It is important to note that, in general, to determine the coefficients $\lambda_j = (f, \Phi_j)$ for $j = 0, \ldots, n$, the integrals in (23) have to be approximated using a quadrature rule.

If we increase the degree of the polynomial, i.e., we seek the least-squares approximation out of $\mathcal{P}_{n+1}$, we have that

$$p_{n+1}^{*,2}(x) = \sum_{i=0}^{n+1} (f, \Phi_j)\Phi_j(x) = p_n^{*,2}(x) + (f, \Phi_{n+1})\Phi_{n+1}(x) \tag{25}$$

so that all the information used to determine $p_n^{*,2}(x)$, i.e., all the coefficients $\lambda_j = (f, \Phi_j)$, $j = 0, \ldots, n$ determined for (24) are reused in (25). Thus, if (24) is known, then one only has to determine the single new coefficient $\lambda_{n+1} = (f, \Phi_{n+1})$ to determine $p_{n+1}^{*,2}(x)$.

Construction of Orthogonal Polynomials

The question remains, given a weight function $\rho(x)$, how does one *construct* the orthonormal polynomials $\{\Phi_j(x)\}_{j=0}^n$? There exists systematic processes for doing so, including the Gram-Schmidt process which results in the following algorithm. First, define the auxiliary polynomials $\{\widetilde{\Phi}_j(x)\}_{j=0}^n$ by

$$\begin{cases} \widetilde{\Phi}_0(x) = 1 \\ \widetilde{\Phi}_1(x) = x - \alpha_1 \\ \widetilde{\Phi}_j(x) = (x - \alpha_j)\widetilde{\Phi}_{j-1}(x) - \beta_j \widetilde{\Phi}_{j-2}(x) \qquad \text{for } j = 2, \ldots, n, \end{cases} \tag{26}$$

where

$$\begin{cases} \alpha_j = \dfrac{(x\widetilde{\Phi}_{j-1}, \widetilde{\Phi}_{j-1})}{\|\widetilde{\Phi}_{j-1}\|_2^2} \qquad \text{for } j = 1, \ldots, n \\ \beta_j = \dfrac{(x\widetilde{\Phi}_{j-1}, \widetilde{\Phi}_{j-2})}{\|\widetilde{\Phi}_{j-2}\|_2^2} \qquad \text{for } j = 2, \ldots, n. \end{cases} \tag{27}$$

Then, set

$$\Phi_j(x) = \frac{1}{\|\widetilde{\Phi}_j\|_2}\widetilde{\Phi}_j(x) \qquad \text{for } j = 0, \ldots, n. \tag{28}$$

One can verify that the polynomials $\{\widetilde{\Phi}_j(x)\}_{j=0}^n$ defined by (26) and (27) are orthogonal with respect to the weight $\rho(x)$ so that, due to (28), the polynomials $\{\Phi_j(x)\}_{j=0}^n$ are orthonormal. For $j = 0, \ldots, n$, $\Phi_j(x)$ is a polynomial of degree exactly j so that $\Phi_j(x) \in \mathcal{P}_n$ and also the set $\{\widetilde{\Phi}_j(x)\}_{j=0}^n$ is linearly independent. Thus, as desired, $\{\widetilde{\Phi}_j(x)\}_{j=0}^n$ *is an orthonormal polynomial basis for* $\mathcal{P}_n$.

The last equation in (26) is referred to as a *three-term recurrence relation*, i.e., $\widetilde{\Phi}_j(x)$ is determined from $\widetilde{\Phi}_{j-1}(x)$ and $\widetilde{\Phi}_{j-2}(x)$. Despite the fact that $\widetilde{\Phi}_0, \ldots, \widetilde{\Phi}_{j-3}$ do not enter into the defining equation for $\widetilde{\Phi}_j(x)$, the latter is orthogonal to all of those polynomials as well as to $\widetilde{\Phi}_{j-1}(x)$ and $\widetilde{\Phi}_{j-2}(x)$.

The coefficients α_j and β_j involve integrals, e.g., we have that

$$\alpha_j = \frac{\displaystyle\int_a^b x\big(\widetilde{\Phi}_{j-1}(x)\big)^2\rho(x)dx}{\displaystyle\int_a^b \big(\widetilde{\Phi}_{j-1}(x)\big)^2\rho(x)dx}$$

so that for general weight functions $\rho(x)$, the integrals have to be also approximated using a quadrature rule.

Legendre and Chebyshev polynomials. Orthogonal polynomials are very useful in many areas of mathematics and science. As such, they have been extensively studied and orthogonal polynomial families for many specific useful weights have been catalogued. A particularly useful family are the Jacobi polynomials for which, for the interval[7] $[-1,1]$, we have the weight[8]

$$\rho(x) = (1-x)^\alpha(1-x)^\beta \qquad \text{for } x \in [-1,1], \tag{29}$$

where $\alpha > -1$ and $\beta > -1$ are constants. We give two examples of well-known special cases of Jacobi polynomials.

[7] General intervals $[a,b]$ can be handled by mapping from the interval $[-1,1]$.

[8] For $\alpha \le -1$ or $\beta \le -1$, $\rho(x)$ is not integrable on $[-1,1]$ which violates one of our assumptions about the weight function $\rho(x)$.

Example . Legendre polynomials. The *Legendre polynomials* result for the weight $\rho(x) = 1$, i.e., $\alpha = \beta = 0$ in (29). From (26) and (27), one can show that

$$\begin{cases} \widetilde{\Phi}_0(x) = 1 \\ \widetilde{\Phi}_1(x) = x \\ \widetilde{\Phi}_j(x) = x\widetilde{\Phi}_{j-1}(x) - \dfrac{(j-1)^2}{4(j-1)^2 - 1}\widetilde{\Phi}_{j-2}(x) \qquad \text{for } j = 2, \ldots, n. \end{cases} \tag{30}$$

Thus, for example, we have

$$\widetilde{\Phi}_2(x) = x^2 - \frac{1}{3} \qquad \text{and} \qquad \widetilde{\Phi}_3(x) = x(x^2 - \frac{1}{3}) - \frac{4}{15}x = x^3 - \frac{3}{5}x.$$

In fact, each $\widetilde{\Phi}_j(x)$ has x^j the highest power term, i.e., the coefficient of the highest power term is one. Also, if j is odd, $\widetilde{\Phi}_j(x)$ only involves odd powers of x whereas if j is even, only even powers of x are involved. Thus, on the interval $[-1, 1]$, $\widetilde{\Phi}_j(x)$ is an odd function for j odd and an even function for j even.

At this point we could use (28) to transform the orthogonal basis $\{\widetilde{\Phi}_j(x)\}_{j=0}^n$ into an orthonormal one. However, historically, the Legendre polynomials are normalized so that the have unit value at $x = 1$. The resulting polynomials,

which we denote by $\{\Phi_j^{Leg}(x)\}_{j=0}^n$, are then constructed from the recurrence

$$\begin{cases} \Phi_0^{Leg}(x) = 1 \\ \Phi_1^{Leg}(x) = x \\ \Phi_j^{Leg}(x) = \dfrac{2j-1}{j} x \Phi_{j-1}^{Leg}(x) - \dfrac{(j-1)}{j} \Phi_{j-2}^{Leg}(x) \qquad \text{for } j = 2, \ldots, n. \end{cases}$$

For example, we have $\Phi_2^{Leg}(x) = \frac{3}{2}x^2 - \frac{1}{2}$, $\Phi_4^{Leg}(x) = \frac{5}{2}x^3 - \frac{3}{2}x$, and $\Phi_5^{Leg}(x) = \frac{35}{8}x^4 - \frac{30}{8}x^2 + \frac{3}{8}$. Note that we have that, if $\Phi_{j-1}^{Leg}(x) = 1$ and $\Phi_{j-2}^{Leg}(x) = 1$, then $\Phi_j^{Leg}(1) = \frac{2j-1-j+1}{j} = 1$.

The Legendre polynomials are orthogonal so that the corresponding Gram matrix is diagonal. However, see have set $\Phi_j^{Leg}(x) = 1$ instead of $\|\Phi_j^{Leg}\|_2 = 1$, they are not orthonormal so that the diagonal entries are not unity. □

Example Chebyshev polynomials. The *Chebyshev polynomials* result for the weight $\rho(x) = (1-x^2)^{1/2}$, i.e., $\alpha = \beta = -\frac{1}{2}$ in (29). The recurrence relation

for the Chebyshev polynomials is given by

$$\begin{cases} \Phi_0^{Che}(x) = 1 \\ \Phi_1^{Che}(x) = x \\ \Phi_j^{Che}(x) = 2x\Phi_{j-1}^{Che}(x) - \Phi_{j-2}^{Che}(x) \qquad \text{for } j = 2, \ldots, n. \end{cases}$$

Interestingly, the Chebyshev polynomials can also be determined from

$$\Phi_j^{Che}(x) = \cos(j \arccos x) \qquad \text{for } j = 0, 1, \ldots, n.$$

The Chebyshev polynomials are orthogonal but the are not orthonormal. In fact, we have that

$$\int_{-1}^{1} \Phi_j^{Che}(x)\Phi_k^{Che}(x)\frac{dx}{\sqrt{1-x^2}} = \begin{cases} 0 & \text{if } j \neq k \\ \pi & \text{if } j = k = 0 \\ \dfrac{\pi}{2} & \text{if } j = k > 0. \end{cases}$$

The Gram matrix G is clearly diagonal.

The polynomials $\Phi_j^{Che}(x)$ are often referred to as the *Chebyshev polynomials of the first kind* to differentiate them from the Jacobi polynomials corresponding to $\alpha = \beta = \frac{1}{2}$ which are known as the *Chebyshev polynomials of the second kind*

and for which we have the recurrence relation

$$\begin{cases} \Phi_0^{Che2}(x) = 1 \\ \Phi_1^{Che2}(x) = 2x \\ \Phi_j^{Che2}(x) = 2x\Phi_{j-1}^{Che2}(x) - \Phi_{j-2}^{Che2}(x) \qquad \text{for } j = 2, \ldots, n. \end{cases}$$

Note that the only change from the recurrence relation for $\Phi_j^{Che}(x)$ is in the choice of $\Phi_1^{Che2}(x)$. We also have that

$$\Phi_j^{Che2}(x) = \frac{\sin(j \arccos x)}{\sin(\arccos x)} \qquad \text{for } j = 0, 1, \ldots, n. \qquad \square$$

Errors in Best Polynomial Approximation

In this section, we collect some results about the error incurred when on approximates a function by a best approximation.

Convergence of the best $L^\infty[a,b]$ approximation

Suppose that $V = C[a,b]$, the space of continuous functions on the interval $[a,b]$, and $W = \mathcal{P}_n$, the space of polynomials of degree less than or equal to n on $[a,b]$. Suppose we are given a function $f(x) \in C[a,b]$ for which we compute the best $L^\infty[a,b]$ (or uniform or min-max) approximation $p_n^{*,\infty}(x)$ and suppose we want the error in that approximation to be less than some given number $\varepsilon > 0$, i.e., we want

$$\|f(x) - p_n^{*,\infty}(x)\|_{L^\infty[a,b]} < \varepsilon. \tag{31}$$

The question is: can we guarantee that this estimate for the error holds just by choosing n large enough?[9] It is easy to see that the answer is yes. Recall that the Wierstrass theorem tells that, given a $f(x) \in C[a,b]$ and an error tolerance

[9] Recall that this was not true for the interpolation of general continuous functions.

$\varepsilon > 0$, there exists a large enough n and polynomial $p_n^{Weir}(x)$ such that

$$\|f(x) - p_n^{Weir}(x)\|_{L^\infty[a,b]} < \varepsilon.$$

On the other hand, from the definition of the best approximation, we have that, for that value of n,

$$\|f(x) - p_n^{*,\infty}(x)\|_{L^\infty[a,b]} \leq \|f(x) - p_n^{Weir}(x)\|_{L^\infty[a,b]}.$$

The last two results then imply that, given a $f(x) \in C[a,b]$ and an error tolerance $\varepsilon > 0$, there exists a large enough n such that (31) holds. Thus, *the error in best* $L^\infty[a,b]$ *approximation can be made as small as we want just by choosing* n *large enough.*

Error in the least-squares approximation

We derive a formula for the error in the best approximation in an inner product space. We want to determine the error $\|f - p_n^{*,2}\|_2$. We have that

$$\|f - p_n^{*,2}\|_2^2 = (f - p_n^{*,2}, f - p_n^{*,2}) = (v, f - p_n^{*,2}) - (p_n^{*,2}, f - p_n^{*,2}).$$

From (8) with $v = f$ and $w = w^* = p_n^{*,2}$, we have that $(p_n^{*,2}, f - p_n^{*,2}) = 0$ so that

$$\|f - p_n^{*,2}\|_2^2 = (f, f - p_n^{*,2}) = (f, f) - (f, p_n^{*,2}) = \|f\|_2^2 - (f, p_n^{*,2}).$$

Then, from (10), we have that

$$\|f - p_n^{*,2}\|_2^2 = \|f\|_2^2 - \big(f, \sum_{j=0}^{n} \lambda_j \Phi_j\big) = \|f\|_2^2 - \sum_{j=0}^{n} \lambda_j (f, \Phi_j).$$

But, from (23), we have that $\lambda_j = (f, \Phi_j)$ so that the error is given by

$$\|f - p_n^{*,2}\|_2^2 = \|f\|_2^2 - \sum_{j=0}^{n} (f, \Phi_j)^2. \tag{32}$$

This relation is known as *Parseval's equality* and says that the square of the error in the least-squares approximation can be computed from the norm of the given function f and the coefficients $\lambda_j = (f, \Phi_j)$ in (22) or (24).

Note that (32) is not an estimate for the error, but actually tells us the exact error: the terms on the right-hand side are computable because we know f (it is the given function) and we know $\{\Phi_j(x)\}_{j=0}^n$ (it is the chosen basis.) Of course, those terms involve integrals (see (19) and (23)) that, in practice, have to be approximated using a quadrature rule.

Are we guaranteed that we can make the error $f - p_n^{*,2}$ as small as we want simply by choosing n large enough? From the definition of the best $L_\rho^2(a, b)$

approximation $p_n^{*,2}(x)$, we have that

$$\|f(x) - p_n^{*,2}(x)\|_2 \leq \|f(x) - p_n^{*,\infty}(x)\|_2 = \Big(\int_a^b \big(f(x) - p_n^{*,\infty}(x)\big)^2 \rho(x) dx \Big)^{1/2}, \tag{33}$$

where $p_n^{*,\infty}(x)$ denotes the best $L^\infty[a,b]$ approximation to $f(x)$. Note that because the weight function $\rho(x)$ is integrable and is positive for almost all $x \in [a,b]$, we have that $\int_a^b \rho(x)dx = \widetilde{\rho}$ for some number $\widetilde{\rho}$ such that $0 < \widetilde{\rho} < \infty$. Then, form (33), we have that

$$\|f(x) - p_n^{*,2}(x)\|_2^2 \leq \max_{a \leq x \leq b} \big|f(x) - p_n^{*,\infty}(x)\big|^2 \int_a^b \rho(x)dx \leq \widetilde{\rho} \big\|f(x) - p_n^{*,\infty}(x)\big\|_{L^\infty[a,b]}^2.$$

But we have already shown that see can make $\|f - p_n^{*,\infty}\|_{L^\infty[a,b]}$ as small as we want by choosing n large enough. Thus, the same can be said about $\|f - p_n^{*,2}\|_2$.

Uniform norm error in the least-squares approximation

We saw that good approximation when the error measured in the $L^2(a,b)$ norm does not necessarily imply good approximation when the error is measured in the $L^\infty(a,b)$ norm. So, we ask: do we know that, if we measure the error in the $L^2(a,b)$ norm, the error in the least-squares approximation gets smaller as the

degree of the polynomial n increases. However, is the same true if we measure the error in the least-squares approximation in the $L^\infty[a,b]$ norm? In general, the answer is no.

Theorem Let $f(x) \in C[-1,1]$ and let $p_n^{*,2}(x)$ denote its least-squares approximation out of $\mathcal{P}_n$ with respect to the $L^2_\rho(a,b)$ norm with Chebyshev weight $\rho(x) = (1-x^2)^{-1/2}$. Also, let $p_n^{*,\infty}(x)$ denote the best approximation with respect to the $L^\infty[a,b]$ norm. We then have that

$$\|f(x) - p_n^{*,2}(x)\|_\infty < \Big(4 + \frac{4}{\pi^2}\ln n\Big)\|f(x) - p_n^{*,\infty}(x)\|_\infty.$$

Because of the $\ln n$ term, this estimate does not guarantee that $\|f(x)-p_n^{*,2}(x)\|_\infty \to 0$ as $n \to \infty$, even though we know that $\|f(x) - p_n^{*,\infty}(x)\|_\infty$ does so. If we want to show that $\|f(x) - p_n^{*,2}(x)\|_\infty \to 0$ as $n \to \infty$, we need to show that $\frac{1}{\ln n}\|f(x) - p_n^{*,\infty}(x)\|_\infty \to 0$ as $n \to \infty$. For general continuous functions, this is not true. However, if $f(x)$ then this is true.