Approximation

Atkinson Chapter 4, Dahlquist & Bjork Section 4.5, Trefethen's book

Topics marked with * are not on the exam

1 In approximation theory we want to find a function $p(x)$ that is 'close' to another function $f(x)$. We can define closeness using any metric or norm, e.g.

$$\|f(x)-p(x)\|_2^2 = \int (f(x)-p(x))^2 \mathrm{d}x \text{ or}$$
$$\|f(x)-p(x)\|_\infty = \sup |f(x)-p(x)| \text{ or}$$
$$\|f(x)-p(x)\|_1 = \int |f(x)-p(x)| \mathrm{d}x.$$

In order for these norms to make sense we need to restrict the functions f and p to suitable function spaces.

The *polynomial* approximation problem takes the form: Find a polynomial of degree at most n that minimizes the norm of the error. Naturally we will consider (i) whether a solution exists and is unique, (ii) whether the approximation converges as $n \to \infty$. In our section on approximation (loosely following Atkinson, Chapter 4), we will first focus on approximation in the infinity norm, then in the 2 norm and related norms.

2 Existence for optimal polynomial approximation. Theorem (no reference): For every $n \geq 0$ and $f \in C([a,b])$ there is a polynomial of degree $\leq n$ that minimizes $\|f(x)-p(x)\|$ where $\|\cdot\|$ is some norm on $C([a,b])$.

Proof: To show that a minimum/minimizer exists, we want to find some compact subset of the set of polynomials of degree $\leq n$ (which is a finite-dimensional space) and show that the inf over this subset is less than the inf over everything else. Then the minimizer must exist, and must lie in the compact subset. (For this proof it doesn't matter if $C([a,b])$ is complete with respect to the norm $\|\cdot\|$ because we are always dealing with the space of real polynomials of degree $\leq n$, which is finite-dimensional and complete wrt any norm.)

The triangle inequality implies that for any p,

$$\|f\| + \|p\| \geq \|f-p\| \geq |\|f\| - \|p\|| \,.$$

Consider the subset of p st $\|p\| \leq 2\|f\|$. This is a compact set: it is a closed, bounded subset of a finite-dimensional space. Over this set a minimizer/minimum of $\|f-p\|$ exists, and using the above implies

$$\|f\| \geq \text{ minimum over the subset } \geq 0.$$

Taking the infimum over the complement of the compact set yields

$$3\|f\| \geq \text{ inf over the complement } \geq \|f\|.$$

The minimum/minimizer over all polynomials of degree $\leq n$ must therefore exist, and must lie in the set $\|p\| \leq 2\|f\|$. This proves existence of an optimal polynomial for any norm on $C([a,b])$. To consider uniqueness, construction, and behavior of the error as $n \to \infty$ we examine specific norms.

'Minimax' (L^∞) Polynomial Approximation

1 Weierstrass Approximation Theorem. Define

$$\rho_n(f) = \min \|f - p\|_\infty$$

where the min is taken over polynomials p of degree $\leq n$. We know a minimizing polynomial exists by the previous theorem, but we don't yet know whether it's unique. Before asking about uniqueness, we'll first look at whether $\rho_n(f) \to 0$ as $n \to \infty$. (Clearly the sequence $\rho_0, \rho_1, \ldots$ is non-increasing and bounded below (by 0), so it must have some limit.)

Weierstrass approximation theorem: For every $f(x) \in C([a,b])$ and $\epsilon > 0$ there is a polynomial $p(x)$ (not unique) such that

$$\|f - p\|_\infty = \max_{x \in [a,b]} |f(x) - p(x)| \leq \epsilon.$$

A constructive but not instructive proof can be found in Atkinson. It makes use of Bernstein polynomials, which are not really useful in practice, so we skip the proof. (Several other proofs exist; see Approximation Theory and Approximation Practice, Chapter 6 by Trefethen for a list of proofs.) In the above theorem, the degree of the polynomial depends on f, and on ϵ. It is not too hard to rigorously connect this theorem to the previous question; it implies that $\rho_n(f) \to 0$ as $n \to \infty$ for any $f \in C([a,b])$. (Proof, e.g. by reductio.)

How quickly does $\rho_n(f) \to 0$? Atkinson (4.11): If f has $k \geq 0$ continuous derivatives on $[a,b]$ and the k^{th} derivative satisfies

$$\sup_{x,y \in [a,b]} |f^{(k)}(x) - f^{(k)}(y)| \leq M|x - y|^\alpha$$

for some $M > 0$ and $0 < \alpha \leq 1$ [$f^{(k)}$ satisfies a Holder condition with exponent α] then there is a constant d_k independent of f and n for which

$$\rho_n(f) \leq \frac{M d_k}{n^{k+\alpha}}.$$

(No proof.) For example if f is Lipshitz continuous with Lipshitz constant K then $M = K$ and $\alpha = 1$, and the convergence is at least linear. On the other end of the spectrum, if f is $C^\infty([a,b])$ then the convergence is faster than $1/n^k$ for any $k \geq 1$.

2* So we have existence (any norm) and convergence (max norm). Before looking at uniqueness, develop some intuition through an example. Let $[a,b] = [-1,1]$, $f(x) = e^x$, and $n = 1$. Seek a solution of the form $p(x) = a_0 + a_1 x$. Drawing the picture, we see that

- The line must cross the function at two distinct points
- The max abs error ρ_1 must be achieved at exactly three points: $x = -1, x_*, 1$
- At x_* the error has a local minimum, so $f'(x_*) - p_1'(x_*) = 0$

The last two points imply the following equations

$$e^{-1} - (a_0 - a_1) = \rho_1, \;\; e - (a_0 + a_1) = \rho_1, \;\; e^{x_*} - (a_0 + a_1 x_*) = -\rho_1, \;\; e^{x_*} - a_1 = 0$$

There are four unknowns a_0, a_1, x_*, and ρ_1. This is a nonlinear system of equations, whose solution we could find using a fixed-point iteration like Newton's method. Alternatively, we can subtract the first two equations to get

$$a_1 = \frac{e - e^{-1}}{2}$$

Then find

$$x_* = \ln(a_1)$$

Then solve the remaining 2×2 linear system for a_0 and ρ_1.

Notice that for our polynomial of degree $n = 1$, the maximum absolute error is attained at $n + 1 = 3$ distinct points in the interval. This behavior is generic/universal, as the next theorem shows.

3 Chebyshev Equioscillation Theorem. Let $f \in C([a, b])$ and $n \geq 0$. Then there is a unique polynomial q of degree $\leq n$ such that

$$\rho_n(f) = \|f - q\|_\infty.$$

Furthermore, this polynomial is the only polynomial for which there are **at least** $n + 2$ distinct points

$$a \leq x_0 < x_1 < \cdots < x_{n+1} \leq b$$

for which

$$f(x_j) - q(x_j) = \sigma(-1)^j \rho_n(f)$$

where $\sigma = \pm 1$ (depending on f and n).

We will prove only the weaker statement that "if there are $n + 2$ distinct points where the error equioscillates, then the polynomial is a minimizer." (Proof from D&B)

The proof is by contradiction, so we assume that there are at least $n + 2$ points where the error equioscillates, but that the polynomial is not optimal.

- First recall that we know a minimizer exists, just not whether it's unique. So, suppose that there is a polynomial $\tilde{p}$ of appropriate degree, and consider the set of points where its error is maximized

  $$M = \{x \in [a, b] \,|\, |f(x) - \tilde{p}(x)| = \|f - \tilde{p}\|_\infty \text{ and oscillates}\}.$$

  Recall that we are assuming that there are at least $n + 2$ points in M.

- If $\tilde{p}$ is not an optimal approximation, we can add another polynomial $p \neq 0$ such that $q = \tilde{p} + p$ is a better approximation.

  $$|f(x) - (\tilde{p}(x) + p(x))| < |f(x) - \tilde{p}(x)| \, \forall x \in M.$$

- The above inequality implies that p must have the same sign as the error $f - \tilde{p}$ at every point $x \in M$, so that adding p reduces error. This implies

  $$(f(x) - \tilde{p}(x))p(x) > 0 \, \forall x \in M.$$

- There are at least $n + 2$ points in M. The polynomial p has to have at least $(n + 2) - 1$ roots (each sign change requires a root). But p has degree $\leq n$, so it can have at most n roots. Clearly this is not possible, because then p would have to have $n + 1$ roots.

4* So now we have existence, convergence, and uniqueness, and an interesting fact about the behavior of the error (equioscillation). Can we compute the optimal polynomial? There is an iterative algorithm due to Remez (1934): the 'Remez exchange algorithm.' The algorithm is given in D&B, along with references; if f is sufficiently smooth it can be shown that the algorithm converges quadratically.

5 Can we estimate $\rho_n(f)$? Well clearly, if we have any polynomial p then $\rho_n(f) \leq \|f - p\|_\infty$. This is not terribly useful. A lower bound would be more interesting. If we know $\rho_n(f) \geq 0.25$, and we have a polynomial approximation with $\|f - p\|_\infty = 0.25001$ then maybe we're close enough.

Atkinson 4.9 (de la Vallee-Poussin). Suppose that we have a polynomial q such that $f - q$ oscillates

$$f(x_j) - q(x_j) = (-1)^j e_j, \; j = 0, \ldots, n + 1$$

and all the e_j nonzero and the same sign, and

$$a \leq x_0 < \cdots < x_{n+1} \leq b$$

Then

$$\min_j |e_j| \leq \rho_n(f).$$

The proof is similar to the equioscillation proof, and works by contradiction: assume that $\min_j |e_j| > \rho_n(f)$. Assume WLOG that the e_i are positive. Denote the optimal polynomial by p, and define $r(x) = q(x) - p(x)$. This is a polynomial of degree $\leq n$. Evaluate $r(x)$ at each of the x_j:

$$r(x_0) = q(x_0) - p(x_0) = (f-p)_0 - (f-q)_0 = (f-p)_0 - e_0 < 0$$

The last inequality is because $|(f(x) - p(x))| \leq \rho_n(f)$ for every x, and $\rho_n(f)$ is (by the assumption we're going to contradict) smaller than e_0. Next

$$r(x_1) = q(x_1) - p(x_1) = (f-p)_1 - (f-q)_1 = (f-p)_1 + e_1 > 0$$

Proceeding to the last point, we see that $r(x)$ changes sign $n+1$ times, i.e. it must have at least $n+1$ zeros, which is a contradiction (because r has degree $\leq n$ by construction).

The way that this is used is to first construct an approximating polynomial q using some other means, then (if the error oscillates) to evaluate $\min_j |e_j|$.

'Least-Squares' (L^2) Polynomial Approximation

1 We now consider the standard L^2 norm and weighted L^2 norms of the form

$$\|f - p\|_w^2 := \int (f(x) - p(x))^2 w(x) \mathrm{d}x$$

where $w(x) \geq 0$ is a 'weight function' such that $\int x^k w(x)\mathrm{d}x$ is finite for every k. We need this condition in order to guarantee that the integral $\int w(x)p(x)\mathrm{d}x$ is well-defined for polynomials of any degree. We already know that the problem

> Given a function $f \in C[a,b]$, find a polynomial p of degree $\leq n$ that minimizes $\|f - p\|$

has a solution for all these norms. (Note that we're approximating functions in $C[a,b]$ using a weighted L^2 norm; we're not approximating functions in L^2, which are much harder to define.)

L^2 and weighted L^2 norms are special because they are associated with inner products

$$\|f\|_w^2 = \langle f, f\rangle_w \text{ where } \langle f, g\rangle_w = \int f(x)g(x)w(x)\mathrm{d}x.$$

It is straightforward to verify that this is an inner product (symmetric, positive, bilinear). The fact that the norm is associated with an inner product implies that the optimization problem is quadratic, which means we can compute a solution by solving a linear system for the critical point. Other norms not associated with inner products do not have this property, and therefore lead to nonlinear systems of equations/non-quadratic optimization. **Summary:** Least-squares approximation is special because you can find the optimal approximation by solving a linear system.

2* Recall the real finite-dimensional case. Every inner product on $\mathbb{R}^N$ can be written as $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = \boldsymbol{v}^T\mathbf{C}\boldsymbol{w}$ where $\mathbf{C}$ is a symmetric positive definite matrix. The matrix $\mathbf{C}$ is analogous to the weight function $w(x)$. $\|\boldsymbol{v}\|_C = \sqrt{\boldsymbol{v}^T\mathbf{C}\boldsymbol{v}}$ is a weighted L^2 norm on $\mathbb{R}^n$.

Consider the problem of finding the closest point $\boldsymbol{v}$ in a subspace to some other point $\boldsymbol{b}$, where distance is measured using an inner-product norm. Let the columns $\boldsymbol{a}_i$ of $\mathbf{A}$ be a basis for the subspace, so that any $\boldsymbol{v}$ in the subspace can be written uniquely as $\boldsymbol{v} = \mathbf{A}\boldsymbol{c}$. The (squared) 'distance' between $\boldsymbol{v}$ and $\boldsymbol{b}$ is

$$\|\boldsymbol{v} - \boldsymbol{b}\|_C^2 = \|\mathbf{A}\boldsymbol{c} - \boldsymbol{b}\|_C^2 = (\mathbf{A}\boldsymbol{c} - \boldsymbol{b})^T\mathbf{C}(\mathbf{A}\boldsymbol{c} - \boldsymbol{b})$$

$$= \boldsymbol{c}^T\mathbf{A}^T\mathbf{C}\mathbf{A}\boldsymbol{c} - 2\boldsymbol{c}^T\mathbf{A}^T\mathbf{C}\boldsymbol{b} + \boldsymbol{b}^T\mathbf{C}\boldsymbol{b}.$$

This is a non-negative quadratic function, so a minimum must exist. Taking the derivative and setting to zero to find a critical point yields

$$\mathbf{A}^T\mathbf{C}\mathbf{A}\boldsymbol{c} = \mathbf{A}^T\mathbf{C}\boldsymbol{b}.$$

The coefficient matrix is a 'Gram matrix'; it is SPD because the columns of $\mathbf{A}$ are linearly independent. So the problem has a *unique* solution of the form

$$\boldsymbol{v} = \mathbf{A}\boldsymbol{c} = \mathbf{A}\left(\mathbf{A}^T\mathbf{C}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{C}\boldsymbol{b}.$$

Clearly this involves solving a linear sytem.

The problem is easier to solve if we use an orthogonal basis. Let the columns $\boldsymbol{q}_i$ of the matrix $\mathbf{Q}$ form an orthogonal basis for the subspace. The fact that the columns of $\mathbf{Q}$ are orthogonal wrt the $\mathbf{C}$ inner product (aka $\mathbf{C}$-conjugate) means that $\mathbf{Q}^T\mathbf{C}\mathbf{Q} = \mathbf{D}$ is diagonal with positive diagonal elements. In fact, $\mathbf{D}_{ii} = \|\boldsymbol{q}_i\|_C^2$. So the above expression can be re-written as

$$\boldsymbol{v} = \mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}^T\mathbf{C}\boldsymbol{b} = \sum_i \frac{\langle \boldsymbol{q}_i, \boldsymbol{b}\rangle}{\|\boldsymbol{q}_i\|_C^2}\boldsymbol{q}_i$$

where the inner product is as defined above. Using an orthogonal basis means we still have to solve a linear system for the coefficients, but the system is diagonal. Notice that the remainder $\boldsymbol{b} - \boldsymbol{v}$ is $\mathbf{C}$-orthogonal to the subspace:

$$\langle \boldsymbol{q}_i, \boldsymbol{b} - \boldsymbol{v}\rangle = \ldots = 0.$$

In fact, the solution $\boldsymbol{v}$ is the $\mathbf{C}$-orthogonal projection of $\boldsymbol{b}$ into the subspace. One can alternatively pose the least-squares problem as finding the $\boldsymbol{c}$ that sets the residual $\boldsymbol{b} - \mathbf{A}\boldsymbol{c}$ orthogonal to the search space; this is the basis for Galerkin projection methods.

3 The polynomial case is a direct generalization of the linear algebra. Indeed the only difference is that the thing we are trying to approximate is an element of an infinite-dimensional vector space (though the approximation is within a finite-dimensional subspace). The only meaningful implication of this fact is for *convergence* as the dimension of our subspace increases to infinity (which will be addressed in due time).

Now suppose that we have a basis $\{\phi_i(x)\}_0^n$ for the space of polynomials of degree $\leq n$. Any polynomial of degree $\leq n$ can therefore be written

$$p(x) = \sum_i c_i\phi_i(x). \text{ (Analogue: } \boldsymbol{v} = \mathbf{A}\boldsymbol{c})$$

We want to solve for the coefficients c_i of the optimal polynomial. Plug this representation in to $\|f - p\|_w^2$

$$\|f(x) - p(x)\|_w^2 = \int (f(x) - p(x))^2 w(x)\mathrm{d}x = \int f(x)^2 w(x)\mathrm{d}x - 2\int f(x)p(x)w(x)\mathrm{d}x + \int p(x)^2 w(x)\mathrm{d}x$$

$$= \|f\|_w^2 - 2\sum_i c_i \int f(x)\phi_i(x)w(x)\mathrm{d}x + \sum_i\sum_j c_i c_j \int \phi_i(x)\phi_j(x)w(x)\mathrm{d}x$$

Define the following vector and matrix:

$$\boldsymbol{f}_i = \int f(x)\phi_i(x)w(x)\mathrm{d}x = \langle f, \phi_i\rangle, \quad \mathbf{M}_{ij} = \int \phi_i(x)\phi_j(x)w(x)\mathrm{d}x.$$

With these definitions we can write

$$\|f(x) - p(x)\|_w^2 = \|f\|_w^2 - 2\boldsymbol{c}^T\boldsymbol{f} + \boldsymbol{c}^T\mathbf{M}\boldsymbol{c}.$$

This is directly analogous to the linear algebra expression with

$$\boldsymbol{f} = \mathbf{A}^T\mathbf{C}\boldsymbol{b}, \quad \mathbf{M} = \mathbf{A}^T\mathbf{C}\mathbf{A}.$$

We now want to find the minimizer $\boldsymbol{c}$ of this quadratic function. So we take the derivative, set it equal to zero, and get

$$\mathbf{M}\boldsymbol{c} = \boldsymbol{f}.$$

The matrix $\mathbf{M}$ is a Gram matrix, so it is symmetric positive definite. For example, if the basis functions are $\phi_i(x) = x^i$ and we're using the standard, unweighted L^2 inner product on $[0, 1]$, then the matrix $\mathbf{M}$ is the so-called Hilbert matrix

$$\mathbf{M}_{ij} = \frac{1}{i + j + 1}.$$

Although it's SPD, it is horrifically badly conditioned for large n.

If you have an *orthogonal* basis, then the matrix $\mathbf{M}$ is *diagonal*. Even if a diagonal matrix is horrifically badly conditioned, it can still be inverted without worry about roundoff errors. If you have an orthogonal basis then the linear system for the coefficients is

$$\mathbf{D}\boldsymbol{c} = \boldsymbol{f}$$

where the diagonal entries of $\mathbf{D}$ are $\int \phi_i(x)^2 w(x)\mathrm{d}x = \|\phi_i\|_w^2$, so the coefficients are just

$$c_i = \frac{\langle f, \phi_i \rangle}{\|\phi_i\|_w^2}.$$

Comment: Notice that to compute the optimal polynomial approximation, you have to be able to compute the integrals associated with the coefficients of the solution formula. Sometimes you can do this analytically. Otherwise you have to do a numerical approximation to the integrals, which we discuss in the section on quadrature (Atkinson, Chapter 5).

4 Now, consider whether the approximation converges as $n \to \infty$. First note that for any polynomial p of degree $\leq n$ we have

$$\int_a^b (f(x) - p(x))^2 w(x)\mathrm{d}x \leq \|f - p\|_\infty^2 \int_a^b w(x)\mathrm{d}x.$$

The inequality only assumes that w is integrable on $[a, b]$, and that f is bounded. (If the interval is infinite then it doesn't necessarily work.) Now consider the minimum of each expression over all polynomials p of degree $\leq n$

$$\text{Smallest possible weighted } L^2 \text{ approximation error of degree } n \leq \rho_n(f) \left(\int_a^b w(x)\mathrm{d}x \right)^{1/2}.$$

We know that for any continuous function $f(x)$, $\rho_n(f) \to 0$ as $n \to \infty$, so the weighted L^2 norm of the error will also go to zero. As a cautionary tale, note that the sequence of polynomials x^k converges to 0 over $[0, 1]$ in the L^2 norm, but it does not converge pointwise. In general the question of whether the weighted L^2 approximation converges uniformly (and therefore pointwise) depends on the weight function and on the function being approximated.

5 Two named relations. Bessel's inequality: Plugging the value of c_i back into the above and noting that the expression is non-negative yields ($d_{ii} = \|\phi_i\|_w^2$, $f_i = \langle \phi_i, f \rangle$)

$$0 \leq \|f\|_w^2 - 2\sum_i \frac{f_i^2}{d_{ii}} + \sum_i \frac{f_i^2}{d_{ii}} \Rightarrow \sum_{i=0}^{n} \frac{f_i^2}{d_{ii}} = \|p\|_w^2 \leq \|f\|_w^2.$$

This means that the error is non-increasing (no surprise). Together with the fact that the error $\to 0$ implies Parseval's equality

$$\sum_{i=0}^{\infty} \frac{f_i^2}{d_{ii}} = \|f\|_w^2.$$

6 We've seen existence, uniqueness, construction, and a bit of convergence. In the construction it's convenient to have an orthogonal basis for the space of polynomials of degree $\leq n$, orthogonal with respect to the inner product/weight function being used. Consider the standard L^2 norm ($w(x) = 1$) on $[-1, 1]$. We need a basis for the vector space of polynomials of degree $\leq n$. The standard one is monomials $\{1, x, \cdots, x^n\}$, but this basis is not orthogonal with respect to the L^2 inner product. But we can use Gram-Schmidt to construct such a basis

$$\phi_0(x) = 1$$
$$\phi_1(x) = x - \frac{\int_{-1}^{1}(1)(x)\mathrm{d}x}{\int_{-1}^{1}(1)^2}(1) = x$$
$$\phi_2(x) = x^2 - \frac{\int_{-1}^{1}(1)(x^2)\mathrm{d}x}{\int_{-1}^{1}(1)^2}(1) - \frac{\int_{-1}^{1}(x)(x^2)\mathrm{d}x}{\int_{-1}^{1}(x)^2}(x) = x^2 - \frac{1}{3}$$

etc. The above polynomials are not ortho*normal*, but follow the common convention to write whatever basis you're using so that the coefficient of the highest-order term is 1 ('monic').

It's clear that you can always construct an orthonormal basis by applying Gram-Schmidt to the monomial basis. This is not always the best way to construct the basis functions, but it works. It also just constructs one of an infinite number of orthogonal bases for polynomials with a given weight function; if you start with a different basis and apply Gram-Schmidt, you'll get a different orthonormal basis. Applying Gram-Schmidt to the monomial basis always produces an orthogonal basis that has increasing degree, i.e. $\phi_k(x)$ is a polynomial of degree k.

Some bases that result from applying Gram-Schmidt to monomials are

- $[-1, 1]$, $w(x) = 1$: Legendre
- $[-1, 1]$, $w(x) = 1/\sqrt{1 - x^2}$: Chebyshev (singular weight function)
- $[-1, 1]$, $w(x) = (1 - x^2)^{m-1/2}$, $m > -1/2$: Gegenbauer (aka 'ultraspherical' polynomials; appear in spherical harmonics; contains Chebyshev and Legendre)
- $[-1, 1]$, $w(x) = (1-x)^\alpha(1+x)^\beta$, $\alpha, \beta > -1$: Jacobi (contains Gegenbauer; draw some weight functions to show that you can penalize errors in different parts of the interval)
- $[0, \infty)$, $w(x) = e^{-x}$: Laguerre (often the basis is written as 'Laguerre functions $e^{-x/2}\phi_k(x)$ where $\phi_k(x)$ is the Laguerre polynomial)
- $[-\infty, \infty]$, $w(x) = e^{-x^2}$: Hermite H (for physicists; if weight is $e^{-x^2/2}$ polynomials are $He(x)$ for probabilists; often written as 'parabolic cylinder functions')

It's important to note:

- Once you pick a weight function and a polynomial degree, the optimal polynomial is unique. The optimal polynomial for one weight function is different from the optimal polynomial for a different weight function.
- The optimal polynomial can be represented in *any* basis (of which there are an infinite number). So if you pick $w(x) = 1$, a convenient/orthogonal basis in which to compute the solution is the Legendre polynomials. But you could write the optimal polynomial in any basis you want, e.g. monomials or Chebyshev or Gegenbauer, or whatever.

7* The term 'orthogonal polynomials' (referring to orthogonal bases) is quite broad, but almost always refers to a basis that has *increasing degree* i.e. $\phi_k(x)$ is a polynomial of degree k (which is what results from applying Gram-Schmidt to monomials). Such bases share a wide range of useful properties regardless of the weight function. For example, they all have a three-term recurrence (Atkinson 4.5).

$$\phi_{k+1}(x) = (x - \beta_k)\phi_k(x) - \gamma_{k-1}^2\phi_{k-1}(x)$$

where

$$\beta_k = \frac{\langle x\phi_k(x), \phi_k(x)\rangle}{\|\phi_k(x)\|_w^2}, \quad \gamma_{k-1}^2 = \frac{\|\phi_k\|_w^2}{\|\phi_{k-1}\|_w^2}. \text{ 'Darboux's formulas'}$$

The initial values are $\phi_{-1} = 0$ and $\phi_0 = 1$. Proof by induction (This proof is based on the one in D&B, but their proof mangles the induction):

- $\phi_1(x) = x - \beta_0$. We simply need to check that it's of degree 1 (obvious) and orthogonal to $\phi_0(x) = 1$. By construction,

  $$\langle\phi_0, \phi_1\rangle = \langle\phi_0, x - \beta_0\rangle = \langle\phi_0, (x - \beta_0)\phi_0\rangle = \langle x\phi_0, \phi_0\rangle - \beta_0\langle\phi_0, \phi_0\rangle = 0.$$

- Now, by construction ϕ_{k+1} has degree $k + 1$. We need to show that it's orthogonal to all previous ϕ_j, and we can assume that the previous ones are an orthogonal basis.

  Using the definition of ϕ_{k+1}, take the inner product with ϕ_j

  $$\langle\phi_j, \phi_{k+1}\rangle = \langle\phi_j, x\phi_k\rangle - \beta_k\langle\phi_j, \phi_k\rangle - \gamma_{k-1}^2\langle\phi_j, \phi_{k-1}\rangle.$$

  Now notice that $\langle xf, g\rangle = \langle f, xg\rangle$ for any functions f and g (for this kind of inner product), so

  $$\langle\phi_j, \phi_{k+1}\rangle = \langle x\phi_j, \phi_k\rangle - \beta_k\langle\phi_j, \phi_k\rangle - \gamma_{k-1}^2\langle\phi_j, \phi_{k-1}\rangle.$$

  Now $x\phi_j$ is a polynomial of degree $j + 1$. Since ϕ_k is orthogonal to any polynomial of lower degree, we have

  $$\langle\phi_j, \phi_{k+1}\rangle = 0 \text{ for } j \le k - 2.$$

  We still need to show that ϕ_{k+1} is orthogonal to ϕ_k and ϕ_{k-1}. Plugging in $j = k$

  $$\langle\phi_k, \phi_{k+1}\rangle = \langle x\phi_k, \phi_k\rangle - \beta_k\langle\phi_k, \phi_k\rangle - \gamma_{k-1}^2\langle\phi_k, \phi_{k-1}\rangle = 0$$

  Next use the definition of β_k and the fact that ϕ_k is orthogonal to ϕ_{k-1} to find $\langle\phi_k, \phi_{k+1}\rangle = 0$. Now plugging in $j = k - 1$

  $$\langle\phi_{k-1}, \phi_{k+1}\rangle = \langle x\phi_{k-1}, \phi_k\rangle - \gamma_{k-1}^2\langle\phi_{k-1}, \phi_{k-1}\rangle.$$

  This would only be zero if $\langle x\phi_{k-1}, \phi_k\rangle = \|\phi_k\|_w^2$. Fortunately that is true, as we now show. Consider (by definition)

  $$\phi_k = (x - \beta_{k-1})\phi_{k-1} - \gamma_{k-2}^2\phi_{k-2}.$$

  Take the inner product with ϕ_k and use orthogonality of ϕ_k to all previous ones (which we're allowed to assume by the induction hypothesis) to arrive at

  $$\|\phi_k\|_w^2 = \langle\phi_k, x\phi_{k-1}\rangle.$$

QED!

It's often possible to find explicit expressions for the coefficients β_k and γ_k^2. For example Legendre:

$$P_{k+1}(x) = \frac{2k+1}{k+1}xP_k(x) - \frac{k}{k+1}P_{k-1}(x)$$

Generally you can just look these up, e.g. in the DLMF. The three-term recursion can be used to construct the basis functions *and* to compute their values at specific points x.

8* Just like with interpolation, once you *find* the polynomial you want, you still need to *evaluate* it.

$$p(x) = a_0 + a_1 x + \ldots + a_n x^n$$

$$= a_0 + x\left[a_1 + x[a_2 + \ldots + x[a_{n-1} + x a_n]\ldots]\right].$$

Evaluating a polynomial with this nested approach is called 'Horner's method.' It's faster than the naive method of evaluating each term $a_k x^k$ separately and then adding them up.

The monomial basis is a convenient basis for *evaluating* a polynomial because of Horner's algorithm: you can evaluate a polynomial of degree n with $\mathcal{O}(n)$ flops. But the monomial basis is highly inconvenient for *finding* a least-squares approximation, because the associated linear system for the coefficients is typically very badly conditioned. So what are you supposed to do? Find the coefficients in one basis, then convert to coefficients in another basis? It typically costs $\mathcal{O}(n^2)$ to convert from one set of basis coefficients to another, so this would be a very-costly way to evaluate the polynomial.

It turns out that there is a fast algorithm to evaluate a polynomial when you have the coefficients in a basis that has a three-term recurrence, and all triangular orthogonal bases have three-term recurrences. Clenshaw's algorithm (not in Atkinson; see D&B). Suppose you need to evaluate a sum of the form

$$S = \sum_{i=0}^{n} c_k p_k$$

where the c_k are known. Suppose that you have a 3-term recurrence of the form (a_k depends on x, but the notation doesn't reflect this for the sake of simplicity)

$$p_{k+1} = a_k p_k + b_k p_{k-1}, \;\; p_{-1} = 0.$$

Write this in matrix form

$$\begin{bmatrix} 1 & & & & \\ -a_0 & 1 & & & \\ -b_1 & -a_1 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -b_{n-1} & -a_{n-1} & 1 \end{bmatrix} \begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ p_n \end{pmatrix} = p_0 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \Leftrightarrow \mathbf{L}\boldsymbol{p} = p_0 \boldsymbol{e}_1.$$

Write the polynomial as

$$p(x) = \sum_i c_k p_k(x) = S = \boldsymbol{c}^T \boldsymbol{p} = p_0 \boldsymbol{c}^T \mathbf{L}^{-1} \boldsymbol{e}_1.$$

Swap the order around a bit (since S is scalar)

$$S = p_0 \boldsymbol{e}_1^T (\mathbf{L}^{-T} \boldsymbol{c}).$$

So to compute S, backsolve an upper-triangular system with bandwidth 2, keep only the first element of the solution, and multiply by p_0. The backsolve can be done in $\mathcal{O}(n)$ floating-point computations. It seems that the only benefit to taking the transpose is to reduce storage cost; both methods are $\mathcal{O}(n)$ floating-point computations.

9 Another property that will be used in the quadrature section is the following (Atkinson 4.4).

> Let $\{\phi_n(x)\}$ be a family of orthogonal polynomials with degree $\phi_n(x) = n$. Then $\phi_n(x)$ has exactly n simple roots in the interval $[a, b]$.

We will prove this by contradiction; our tools will be (i) the fundamental theorem of algebra, and (ii) that $\phi_n(x)$ is orthogonal to all polynomials of lower degree. First define some notation: let $x_1, \ldots, x_m$ be the roots of $\phi_n(x)$ that lie in $[a, b]$ *and* for which $\phi_n(x)$ changes sign (i.e. odd order roots). We will prove that $m = n$, but at this point m could be anything including 0. Now define

$$B(x) = (x - x_1) \cdots (x - x_m).$$

This is a polynomial of degree m. Now consider

$$B(x)\phi_n(x).$$

This is a polynomial of degree $m+n$, but more importantly, it has a single sign on $[a,b]$. This is because $B(x)$ was constructed to change sign at every place where $\phi_n(x)$ changes sign. Now consider

$$\langle B(x), \phi_n(x)\rangle = \int_a^b B(x)\phi_n(x)w(x)\mathrm{d}x.$$

If $m < n$ then this has to be zero, because $\phi_n(x)$ is orthogonal to every polynomial of lower degree. On the other hand, the integrand is positive (except possibly on a set of measure zero). So the only way that the inner product can be positive is when the $m = n$, i.e. $\phi_n(x)$ has exactly n simple roots in $[a,b]$.

10* There are lots of other properties of orthogonal polynomials. They are eigenfunction solutions of second-order boundary-value problems (the most general setting is Sturm-Liouville theory (in Applied Analysis)). They can be computed using a 'Rodrigues formula'

$$\phi_n(x) = \frac{1}{w(x)}\frac{\mathrm{d}^n}{\mathrm{d}x^n}[w(x)(g(x))^n]$$

where $g(x)$ is a polynomial independent of n. Etc. You can look up these properties in any book on orthogonal polynomials.

Fourier & Chebyshev Least-squares Approximations

1 The most important family of orthogonal polynomials is the Chebyshev polynomials. To understand them it is best to first understand Fourier series, so we will discuss Fourier approximation. We are seeking to find a trig polynomial

$$p(x) = a_0 + \sum_{j=1}^{n}[a_j\cos(jx) + b_j\sin(jx)]$$

that minimizes the unweighted L^2 error for some function $f \in C[-\pi,\pi]$. Notice that the subspace we're seeking a solution in is finite-dimensional, so a solution must exist (by the same argument as for polynomials). Furthermore, notice that our basis for the space of trig polynomials is already orthogonal in the unweighted L^2 inner product:

$$\int_{-\pi}^{\pi}\sin(jx)\cos(kx)\mathrm{d}x = 0 \text{ for any } j,k, \quad \int_{-\pi}^{\pi}\sin(jx)\sin(kx)\mathrm{d}x = 0 \text{ unless } j = k$$

$$\int_{-\pi}^{\pi}\cos(jx)\cos(kx)\mathrm{d}x = 0 \text{ unless } j = k$$

(If $j = k$ we have π.) The solution formula therefore follows immediately from the exact same arguments used for the polynomial case:

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\mathrm{d}x = \frac{\langle 1, f\rangle}{\|1\|}$$

$$a_j = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(jx)\mathrm{d}x = \frac{\langle \cos(jx), f\rangle}{\|\cos(jx)\|}$$

$$b_j = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(jx)\mathrm{d}x = \frac{\langle \sin(jx), f\rangle}{\|\sin(jx)\|}$$

These are just the usual L^2 approximation coefficients for a orthogonal basis. The solution is unique, just like the polynomial problem.

2 We've seen existence & uniqueness for Fourier approximation. Now what about convergence? In general this is a delicate subject, which you will see more of in PDEs and/or Applied Analysis. We first quote the Riesz-Fisher theorem:

If $f(x) \in L^2([-\pi, \pi])$ then $\lim_{n\to\infty} \|f - p\|_2 = 0$.

So the 2-norm of the error goes to 0 for a huge class of functions including discontinuous functions.

Another theorem (D&B 4.6.2)

If $f(x)$ is of bounded variation and has a finite number of discontinuities then $p_n(x) \to f(x)$ at every $x \in [-\pi, \pi)$ except points of discontinuity, where $p(x)$ converges to the average of the values on either side of the discontinuity.

Now suppose that f and all its derivatives up to order $p \geq 1$ are continuous and periodic on $[-\pi, \pi)$. Then $p(x)$ converges pointwise to $f(x)$ according to the theorem above. How quickly does it converge? We can write

$$f(x) = a_0 + \sum_{j=1}^{\infty} [a_j \cos(jx) + b_j \sin(jx)]$$

so

$$f(x) - p_n(x) = \sum_{j=n+1}^{\infty} [a_j \cos(jx) + b_j \sin(jx)].$$

$$\|f(x) - p_n(x)\|_\infty \leq \sum_{n+1}^{\infty} (|a_j| + |b_j|).$$

Consider the following:

$$a_j = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(jx) \mathrm{d}x = -\frac{1}{j\pi} [f(\pi) \sin(j\pi) + f(-\pi) \sin(j\pi)] + \frac{1}{j\pi} \int_{-\pi}^{\pi} f'(x) \sin(jx) \mathrm{d}x$$

$$a_j = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(jx) \mathrm{d}x = \frac{1}{j\pi} \int_{-\pi}^{\pi} f'(x) \sin(jx) \mathrm{d}x$$

I used integration by parts and $\sin(j\pi) = 0$. Applying the same idea to b_j yields

$$b_j = \frac{1}{j\pi} [f(\pi) \cos(j\pi) - f(-\pi) \cos(j\pi)] - \frac{1}{j\pi} \int_{-\pi}^{\pi} f'(x) \cos(jx) \mathrm{d}x.$$

If we assume that f is periodic then $f(\pi) = f(-\pi)$ and we get

$$b_j = -\frac{1}{j\pi} \int_{-\pi}^{\pi} f'(x) \cos(jx) \mathrm{d}x.$$

If f is periodic and $f' \in C[-\pi, \pi]$ we can bound the coefficients:

$$|b_j|, |a_j| \leq \frac{\max_x |f'(x)|}{j\pi}.$$

If f' is both continuous and periodic then we can integrate by parts again to get

$$|b_j|, |a_j| \leq \frac{\max_x |f''(x)|}{j^2\pi}.$$

(We are clearly also assuming that $\|f''\|_\infty$ is finite.) If the function is periodic and has p continuous derivatives, then continuing this process of integration by parts yields bounds of the form

$$|b_j|, |a_j| \leq \frac{\max_x |f^{(p+1)}(x)|}{j^{p+1}\pi}.$$

(We are clearly also assuming that $\|f^{(p+1)}\|_\infty$ is finite.) Returning to the approximation error,

$$\|f(x) - p_n(x)\|_\infty \leq \frac{2\|f^{(p+1)}\|_\infty}{\pi} \sum_{n+1}^{\infty} j^{-p-1} \sim n^{-p}.$$

So the rate of convergence depends on the smoothness of f. If $f \in C^\infty[-\pi, \pi]$ and periodic, then the coefficients decay faster than any power of j.

3 Before returning to Chebyshev polynomials, we need some background. An 'even' function satisfies $f(-x) = f(x)$ for every x; cosine is even. An 'odd' function satisfies $f(-x) = -f(x)$; sine is odd. Any function can be written as a sum of even and odd parts:

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

If you multiply an even function by an odd function the product is odd. Any other product is even. The derivative of an odd function is even, and vice versa. The integral of an odd function over $[-a, a]$ is zero for any a. Thus, if f is even, then $b_j = 0$ for every j. If f is odd, then $a_j = 0$ for every $j \geq 0$.

4 Chebyshev. Consider the functions

$$T_n(x) = \cos(n \arccos(x)).$$

We will show that T_n is a polynomial of degree n. To do this, we will show that

$$\cos(n\theta) = \sum_{j=0}^{n} c_j \cos(\theta)^j.$$

If we let $\theta = \arccos(x)$ then we will have

$$\cos(n \arccos(x)) = \sum_j c_j x^j.$$

First note that we can write Euler's formula in 2 different ways:

$$e^{in\theta} = (\cos(\theta) + i\sin(\theta))^n = \cos(n\theta) + i\sin(n\theta).$$

Expand the middle part

$$\cos(\theta)^n + \begin{pmatrix} n \\ 1 \end{pmatrix} \cos(\theta)^{n-1}(i\sin(\theta)) + \begin{pmatrix} n \\ 2 \end{pmatrix} \cos(\theta)^{n-2}(i\sin(\theta))^2 + \ldots + \begin{pmatrix} n \\ n-1 \end{pmatrix} \cos(\theta)(i\sin(\theta))^{n-1} + (i\sin(\theta))^n.$$

$$\left(\text{Binomial expansion. Recall that } \begin{pmatrix} n \\ k \end{pmatrix} = \frac{n!}{k!(n-k)!}.\right)$$

The real part of this expression equals $\cos(n\theta)$. Notice that the real terms all have *even* powers of $\sin(\theta)$ which can be turned into expressions of cosine via $\sin^2 = 1 - \cos^2$. This proves that $\cos(n\theta)$ can be written as a polynomial in $\cos(\theta)$. Now consider the highest power of $\cos(\theta)$. Every term has combined cosine and sine powers equal to n; when you convert sine squared to cosine squared you don't raise the exponent, so the highest power of $\cos(\theta)$ in the expression for $\cos(n\theta)$ is n. This proves that $T_n(x)$ is a polynomial of degree $\leq n$: technically we also need to show that the leading coefficient is nonzero. This is relatively easy to show. Only the even powers of $i\sin(\theta)$ are real and need to be converted to $\cos^2$, so consider

$$(i\sin(\theta))^{2p} = (-1)^p(1 - \cos^2(\theta))^p = (-1 + \cos^2(\theta))^p.$$

After converting all the $\sin^{2p}(\theta)$ to cosines, the coefficients of $\cos^n(\theta)$ are all positive, so the degree of the trig polynomial is exactly n.

5 Now we want to show that the polynomials $T_n(x)$ are orthogonal with respect to the weight function $w(x) = 1/\sqrt{1-x^2}$ on the interval $[-1, 1]$.

$$\int_{-1}^{1} \frac{T_j(x)T_k(x)}{\sqrt{1-x^2}}\mathrm{d}x = ??$$

As usual with Chebyshev polynomials, it's easiest to address the problem after a change of variable. Let $x = \cos(\theta)$, so $\mathrm{d}x = -\sin(\theta)\mathrm{d}\theta$ and $T_n(x) = \cos(n \arccos(x)) = \cos(n \arccos(\cos(\theta))) = \cos(n\theta)$. As x moves from -1 to 1, θ moves from $-\pi$ to 0. Also, $1 - x^2 = 1 - \cos(\theta)^2 = \sin^2(\theta)$. So

$$\int_{-1}^{1} \frac{T_j(x)T_k(x)}{\sqrt{1-x^2}}\mathrm{d}x = -\int_{-\pi}^{0} \sin(\theta)\frac{\cos(j\theta)\cos(k\theta)}{|\sin(\theta)|}\mathrm{d}\theta = \int_{-\pi}^{0} \cos(j\theta)\cos(k\theta)\mathrm{d}\theta = 0 \text{ unless } j = k.$$

If $j = k$ we have $\|T_j(x)\|^2 = \pi/2$. The penultimate equality uses that $\sin(\theta)$ is negative over the interval in question, and the final equality uses the fact that $\cos(j\theta)\cos(k\theta)$ is even so the integral from $-\pi$ to 0 is just half the integral from $-\pi$ to π, which is zero. We have now established that $T_n(x)$ are polynomials of degree n, and are orthogonal under the weighted inner product.

6 Using the previously-derived expressions for the optimal polynomial, we have that

$$p(x) = \sum_j c_j T_j(x), \quad c_j = \frac{\langle T_j, f\rangle}{\|T_j\|^2}.$$

Recall that $\|T_j\|^2 = \pi/2$. Consider the coefficients a bit more carefully:

$$c_j = \frac{2}{\pi}\langle T_j(x), f(x)\rangle = \frac{2}{\pi}\int_{-1}^{1} \frac{T_j(x)f(x)}{\sqrt{1-x^2}}\mathrm{d}x.$$

Let's make the same change of variable:

$$\frac{2}{\pi}\int_{-1}^{1} \frac{T_j(x)f(x)}{\sqrt{1-x^2}}\mathrm{d}x = \frac{2}{\pi}\int_{-\pi}^{0} \cos(j\theta)f(\cos(\theta))\mathrm{d}\theta.$$

To evaluate the above integral we only need to know f on $[-1, 1]$, i.e. $\theta \in [-\pi, 0]$. Make an even extension of f so that

$$\tilde{f}(\theta) = \left\{ \begin{array}{ll} f(\cos(\theta)) & \theta \in [-\pi, 0] \\ f(\cos(-\theta)) & \theta \in [0, \pi] \end{array} \right\} = f(\cos(\theta))$$

Then the above integral is just

$$c_j = \frac{2}{\pi}\int_{-1}^{1} \frac{T_j(x)f(x)}{\sqrt{1-x^2}}\mathrm{d}x = \frac{2}{\pi}\int_{-\pi}^{0} \cos(j\theta)f(\cos(\theta))\mathrm{d}\theta = \frac{1}{\pi}\int_{-\pi}^{\pi} \cos(j\theta)\tilde{f}(\theta)\mathrm{d}\theta.$$

You should recognize that this is the Fourier coefficient a_j for the even function $\tilde{f}$.

The Chebyshev Series of f is the Fourier Series of $\tilde{f}$.

We know that the Fourier series converges quickly in the ∞ norm for smooth periodic functions. Let's see how the smoothness of f relates to the smoothness of $\tilde{f}$. First, note that $\tilde{f}$ is periodic. Then take the derivative

$$\frac{\mathrm{d}\tilde{f}}{\mathrm{d}\theta} = -\sin(\theta)f'(\cos(\theta)).$$

This is an odd periodic function; as long as f' is continuous on $[-1, 1]$, then $\tilde{f}'(\theta)$ is continuous on $[-\pi, \pi]$. The same holds for all higher derivatives, so

> If f and all its derivatives up to & including p are continuous and $f^{(p+1)}$ is well-behaved then the Chebyshev series converges at least as $\mathcal{O}(n^{-p})$ in the ∞ norm.

This shows that if you get to pick your weight function, then Chebyshev series are a very good choice.