Least Squares Approach to the Rayleigh-Ritz Method

Chad Magers
Mississippi State University
jcm1@ra.msstate.edu

Faculty Advisor: Michael Pearson
Pearson@math.msstate.edu

Abstract

In this paper the author will attempt to present the Rayleigh-Ritz method of approximating solutions to Sturm-Louiville BVPs using linear algebra. Another approach to this problem involves using the calculus of variations. However, the author feels more insight can be obtained by using a least squares approach by defining a norm using the differential equation itself.

1 Introduction

As is often the case, finding explicit solutions to boundary value problems for ordinary differential equations can be an arduous task. Therefore we would like techniques that give us good approximations to the actual solutions. There are finite-difference techniques that give good approximations as well as others, such as various “shooting” methods. In this paper however we will discuss the Rayleigh-Ritz Method of approximating solutions to certain boundary value problems. In the Rayleigh-Ritz method what we do is reformulate the problem into one of choosing, from the set of all sufficiently differentiable functions satisfying the boundary conditions, the function that approximates the actual solution as closely as possible in a certain norm that will be defined later. We will also be interested in investigating what happens when we pick, as our set of all sufficiently differentiable functions satisfying the boundary conditions, the set with basis functions consisting of the eigenfunctions of the corresponding differential operator, which in our case will be the regular Sturm-Louiville operator.
2 The Least Squares approximation and Fourier Coefficients

To begin our discussion of the least squares approximation we note that our setting will be a vector space, $V$, with a norm, $|| \cdot ||$, defined on it. Recall that a norm, $|| \cdot || : V \rightarrow \mathbb{R}$, has the following properties

$$
||u|| \geq 0 \quad \text{for all } u \in V
$$

$$
||\alpha u|| = |\alpha| ||u|| \quad \text{for all } v \in V \text{ and any scalar } \alpha
$$

$$
||u + w|| \leq ||v|| + ||w|| \quad \text{for all } v, w \in V
$$

A vector space, $V$, with a norm defined on it is called a normed linear space. Here in our discussion we will be working with normed linear spaces where the norm arises from an inner product. Such spaces are called inner product spaces. We recall that an inner product, $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$, satisfies

$$
\langle u, u \rangle \geq 0, \quad \text{with equality iff } v = 0
$$

$$
\langle u, v \rangle = \langle v, u \rangle
$$

$$
\langle \alpha u, v \rangle = \alpha \langle u, v \rangle
$$

$$
\langle u + w, v \rangle = \langle u, v \rangle + \langle w, v \rangle
$$

Now, in many problems, solutions cannot be obtained explicitly, thus we look for approximations to those solutions. In terms of the norm what we try to do in least squares approximation is minimize the error between a vector $u \in V$ and its approximation $v$. We usually take $v$ to be an element of a subspace, $S$, of $V$ and we assume that $u$ is not in $S$. Our approximation $v$ takes on the form

$$
v = \sum_{i=1}^{n} c_i \phi_i
$$

where $\{c_i\}_{i=1}^{n}$ are constants and $\{\phi_i\}_{i=1}^{n}$ are linearly independent vectors in $V$ whose span is $S$, i.e., $\{\phi_i\}_{i=1}^{n}$ is a basis for $S$. The error of our approximation is represented by

$$
||u - v||^2.
$$

Therefore what we want to do is find constants $\{c_i\}_{i=1}^{n}$ that will minimize

$$
||u - \sum_{i=1}^{n} c_i \phi_i||^2.
$$

So, let us assume that there exist $(c_1, c_2, \ldots, c_n)$ in $\mathbb{R}^n$ such that $v = \sum_{i=1}^{n} c_i \phi_i$ does indeed give us a minimum for the error

$$
||u - v||^2 \quad \text{or} \quad ||u - \sum_{i=1}^{n} c_i \phi_i||^2.
$$

Now using the properties of the inner product we find that

$$
||u - \sum_{i=1}^{n} c_i \phi_i||^2 = \left\langle u - \sum_{i=1}^{n} c_i \phi_i, u - \sum_{i=1}^{n} c_i \phi_i \right\rangle
$$

$$
= \langle u, u \rangle - 2 \sum_{i=1}^{n} c_i \langle u, \phi_i \rangle + \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \langle \phi_i, \phi_j \rangle.
$$


Since \( u \) and the \( \phi_i \)’s are fixed, we can define a function \( J : \mathbb{R}^n \to \mathbb{R} \) by

\[
J(c_1, c_2, \ldots, c_n) = \langle u, u \rangle - 2 \sum_{i=1}^{n} c_i \langle u, \phi_i \rangle + \sum_{i=1}^{n} c_i \sum_{j=1}^{n} c_j \langle \phi_i, \phi_j \rangle.
\]  

(2.1)

A necessary condition for \( J \) to have a minimum is that

\[
\frac{\partial J}{\partial c_k} = 0 \quad \text{for} \quad k = 1, 2, \ldots, n.
\]

(2.2)

Therefore, (2.2) implies that

\[-2c_k \langle u, \phi_k \rangle + 2 \sum_{j=1}^{n} c_j \langle \phi_k, \phi_j \rangle = 0 \quad \text{for} \quad k = 1, 2, \ldots, n.\]

Now we have a system of \( n \) equations in \( n \) unknowns, which can be written as

\[A \vec{c} = \vec{b}\]

where

\[a_{ij} = \langle \phi_i, \phi_j \rangle\]

and

\[b_j = \langle u, \phi_j \rangle.\]

Our linear system is

\[
\begin{pmatrix}
\langle \phi_1, \phi_1 \rangle & \langle \phi_1, \phi_2 \rangle & \cdots & \langle \phi_1, \phi_n \rangle \\
\langle \phi_2, \phi_1 \rangle & \langle \phi_2, \phi_2 \rangle & \cdots & \langle \phi_2, \phi_n \rangle \\
\vdots & \vdots & \ddots & \vdots \\
\langle \phi_n, \phi_1 \rangle & \langle \phi_n, \phi_2 \rangle & \cdots & \langle \phi_n, \phi_n \rangle
\end{pmatrix}
\begin{pmatrix}
c_1 \\
c_2 \\
\vdots \\
c_n
\end{pmatrix}
= 
\begin{pmatrix}
\langle u, \phi_1 \rangle \\
\langle u, \phi_2 \rangle \\
\vdots \\
\langle u, \phi_n \rangle
\end{pmatrix}.
\]

(2.3)

Since the \( \phi_i \)’s are independent the row and column vectors are linearly independent which implies that \( A \) is non-singular. This in turn implies that we will have a unique solution for the \( c_i \)’s.

The simplest solution that (2.3) can have is when the \( \phi_i \)’s are orthogonal, i.e.,

\[\langle \phi_i, \phi_j \rangle = 0 \quad \text{for} \quad i \neq j.\]

Our linear system then becomes

\[
\begin{pmatrix}
\langle \phi_1, \phi_1 \rangle & 0 & \cdots & 0 \\
0 & \langle \phi_2, \phi_2 \rangle & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \langle \phi_n, \phi_n \rangle
\end{pmatrix}
\begin{pmatrix}
c_1 \\
c_2 \\
\vdots \\
c_n
\end{pmatrix}
= 
\begin{pmatrix}
\langle u, \phi_1 \rangle \\
\langle u, \phi_2 \rangle \\
\vdots \\
\langle u, \phi_n \rangle
\end{pmatrix}.
\]

(2.4)

It is clear now that

\[c_i = \frac{\langle u, \phi_i \rangle}{\langle \phi_i, \phi_i \rangle}.\]

Also, since every basis of a finite dimensional vector space can be written as an orthogonal basis by the Gram-Schmidt process, the linear system (2.3) can always be reduced to the linear system (2.4). The \( c_i \)’s that we recover from this system of equations are called the Fourier coefficients of \( u \) with respect to \( \{\phi_i\}_{i=1}^{n} \).
3 Rayleigh-Ritz

Let \( C^2[a, b] = \left\{ \phi(x) \mid \phi \text{ is twice continuously differentiable on the interval } [a, b] \right\} \) and let \( D_0^2[a, b] = \left\{ \phi(x) \in C^2[a, b] \mid \phi(a) = \phi(b) = 0 \right\} \).

Also, let us define an inner product on \( C^2[a, b] \) by
\[
\langle f, g \rangle = \int_a^b f(x) g(x) \, dx.
\]

It isn’t difficult to show that \( \langle f, g \rangle \) is in fact an inner product.

Now, suppose we have a linear two-point boundary value problem described as
\[
- \frac{d}{dx} \left( p(x) \frac{du}{dx} \right) + q(x)u(x) = f(x) \quad \text{for } a \leq x \leq b,
\]
\[u(a) = u(b) = 0.\]  \hspace{1cm} (3.1)

We assume that \( p \in C^1[a, b], \) \( q, f \in C[a, b], \) and \( p(x) > 0, \) \( q(x) > 0 \) on \([a, b].\) This is the Sturm-Louiville problem.

Now, define \( L : D_0^2[a, b] \to C[a, b] \) by
\[
Lu = - \frac{d}{dx} \left( p(x) \frac{du}{dx} \right) + q(x)u(x). \]  \hspace{1cm} (3.2)

Therefore, the Sturm-Louiville problem can be restated as given \( f \in C[a, b], \) find \( u \in D_0^2[a, b] \) s.t. \( Lu = f.\) Clearly \( L \) is linear and it is easily shown by integration by parts that \( L \) is self-adjoint on \( D_0^2[a, b], \) i.e., \( \langle Lu, v \rangle = \langle u, Lv \rangle \) \( \forall u, v \in D_0^2[a, b]. \) \( L \) is referred to as the Sturm-Louiville operator.

Before we proceed further we need to show that \( L \) is positive definite, i.e., \( \langle Lu, u \rangle > 0 \) for \( u \) not identically zero. Let \( u \in D_0^2[a, b]. \) Then,
\[
\langle Lu, u \rangle = \int_a^b \left( \frac{d}{dx} \left( p(x) \frac{du}{dx} \right) + q(x)u(x) \right) u(x) \, dx
\]
\[
= \int_a^b \left( p(x) \left( \frac{du}{dx} \right)^2 + q(x)(u(x))^2 \right) \, dx \quad \text{by integration by parts}
\]
\[
\geq 0 \quad \text{since } u \text{ is a continuous function.}
\]

Furthermore, \( \langle Lu, u \rangle = 0 \) precisely when \( u \equiv 0 \) on \([a, b].\)

The eigenvalue problem for the Sturm-Louiville problem is stated as
\[
L \phi = \lambda w(x) \phi
\]
where \( \lambda \) is an eigenvalue of \( L \) with corresponding eigenvector \( \phi \) with respect to a weight function, \( w(x) > 0, \) that is dependent on \( L. \) Let \( (\lambda_i, \phi_i) \) be an eigenvalue-eigenfunction pair with respect to the weight function, \( w(x) > 0, \) for \( L \) for \( i = 1, 2, \ldots \). Now, since \( L \) is self-adjoint we know that all of its eigenvalues are real.
Let us note that when $\lambda_i \neq \lambda_j$ then $\phi_i$ is orthogonal to $\phi_j$ with respect to $w(x)$, i.e., $\langle w \phi_i, \phi_j \rangle = 0$ or $\langle \phi_i, w \phi_j \rangle = 0$, since

$$(\lambda_i - \lambda_j) \langle w \phi_i, \phi_j \rangle = \langle \lambda_i w \phi_i, \phi_j \rangle - \langle \phi_i, \lambda_j w \phi_j \rangle$$ since $\lambda_j$ is real

$$= \langle L\phi_i, \phi_j \rangle - \langle \phi_i, L\phi_j \rangle$$

$$= \langle L\phi_i, \phi_j \rangle - \langle L\phi_i, \phi_j \rangle$$ since $L$ is self-adjoint

$$= 0.$$

Therefore, when $i \neq j$, $\langle w \phi_i, \phi_j \rangle = 0$, which is what we wanted to show.

Now, since $L$ is self-adjoint and positive, we can define a new inner product on $D_0^2[a, b]$ by

$$[f, g] = \langle Lf, g \rangle = \int_a^b \left( -\frac{d}{dx} \left( p(x) \frac{df}{dx} \right) + q(x) f(x) \right) g(x) \, dx. \quad (3.3)$$

This inner product induces a norm in the space $D_0^2[a, b]$ via the equation

$$||u|| = [u, u]^{\frac{1}{2}}.$$

In the spirit of the previous section we want to approximate the solution, $u(x)$, of (3.1). Therefore in the Rayleigh-Ritz method we choose a relatively simple set of linearly independent functions in $D_0^2[a, b]$. Then clearly the span of these functions forms a subspace, $S$, of $D_0^2[a, b]$ with the functions serving as a basis of $S$.

Now that we have an inner product space at our disposal we can use the results we obtained in section 2. Using those results we arrive at the linear system

$$(\begin{bmatrix} \phi_1, \phi_1 & \phi_1, \phi_2 & \cdots & \phi_1, \phi_n \\ \phi_2, \phi_1 & \phi_2, \phi_2 & \cdots & \phi_2, \phi_n \\ \vdots & \vdots & \ddots & \vdots \\ \phi_n, \phi_1 & \phi_n, \phi_2 & \cdots & \phi_n, \phi_n \end{bmatrix}) \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} [u, \phi_1] \\ [u, \phi_2] \\ \vdots \\ [u, \phi_n] \end{bmatrix}. \quad (3.4)$$

Glancing at the system we notice that to solve the system in this form we must have a knowledge of the exact solution $u$, which in many cases we do not know. We get around this by noting that

$$[u, \phi_j] = \langle Lu, \phi_j \rangle = \langle f, \phi_j \rangle$$

and likewise

$$[\phi_i, \phi_j] = \langle L\phi_i, \phi_j \rangle.$$

Therefore our system becomes

$$(\begin{bmatrix} \langle L\phi_1, \phi_1 \rangle & \langle L\phi_1, \phi_2 \rangle & \cdots & \langle L\phi_1, \phi_n \rangle \\ \langle L\phi_2, \phi_1 \rangle & \langle L\phi_2, \phi_2 \rangle & \cdots & \langle L\phi_2, \phi_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle L\phi_n, \phi_1 \rangle & \langle L\phi_n, \phi_2 \rangle & \cdots & \langle L\phi_n, \phi_n \rangle \end{bmatrix}) \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \langle f, \phi_1 \rangle \\ \langle f, \phi_2 \rangle \\ \vdots \\ \langle f, \phi_n \rangle \end{bmatrix}. \quad (3.5)$$

The simplest solution that (3.5) can have is when your basis functions are eigenvectors of $L$, then since

$$\langle w \phi_i, \phi_j \rangle = 0 \text{ for } i \neq j$$
our linear system becomes

\[
\begin{bmatrix}
\lambda_1 \langle w \phi_1, \phi_1 \rangle & 0 & \ldots & 0 \\
0 & \lambda_2 \langle w \phi_2, \phi_2 \rangle & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \lambda_n \langle w \phi_n, \phi_n \rangle
\end{bmatrix}
\begin{bmatrix}
c_1 \\
c_2 \\
\vdots \\
c_n
\end{bmatrix}
= \begin{bmatrix}
\langle f, \phi_1 \rangle \\
\langle f, \phi_2 \rangle \\
\vdots \\
\langle f, \phi_n \rangle
\end{bmatrix}.
\] (3.6)

It is clear now that

\[ c_i = \frac{\langle f, \phi_i \rangle}{\lambda_i \langle w \phi_i, \phi_i \rangle}, \]

where again these are the Fourier coefficients of \( u \) w.r.t \( \{\phi_i\}_{i=1}^n \).

Our approximation in this case would be

\[ v(x) = \sum_{i=1}^n \frac{\langle f, \phi_i \rangle}{\lambda_i \langle w \phi_i, \phi_i \rangle} \phi_i(x). \]

Now, define a new inner product on \( D_0^2[a, b] \) by

\[ \langle f, g \rangle = \langle f \cdot w g \rangle = \langle w f, g \rangle. \]

We will now show that the Fourier coefficients we arrive at in the inner product space w.r.t \( \langle \cdot, \cdot \rangle \) are identical to the Fourier coefficients we found in the inner product space w.r.t \( [\cdot, \cdot] \).

Using the least squares approach that we use in Section 2, if \( u(x) \) is the exact solution to (3.1) and

\[ v(x) = \sum_{j=1}^n v_j \phi_j(x) \quad \text{where} \quad L \phi_j = \lambda_j w(x) \phi_j(x) \]

minimizes the error

\[ \langle (u - v, u - v) \rangle, \]

then the \( v_j \)'s must have the form

\[ w_j = \frac{\langle u, \phi_j \rangle}{\langle \phi_j, \phi_j \rangle}, \]

which are the Fourier coefficients of \( u(x) \) with respect to \( \{\phi_i\}_{i=1}^n \). But

\[
\frac{\langle u, \phi_j \rangle}{\langle \phi_j, \phi_j \rangle} = \frac{\langle u, \frac{1}{\lambda_j} L \phi_j \rangle}{\langle \phi_j, \phi_j \rangle} = \frac{\langle L u, \phi_j \rangle}{\langle \phi_j, \phi_j \rangle} = \frac{\langle f, \phi_j \rangle}{\langle \phi_j, \phi_j \rangle},
\]

since \( L \) is self-adjoint w.r.t \( \langle \cdot, \cdot \rangle \).

Therefore we have shown that the coefficients for the best approximation to the actual solution to \( Lu = f \) with respect to the norm \( || \cdot || \) are in fact the Fourier coefficients for the actual solution with respect to \( \langle \cdot, \cdot \rangle \). It must be noted that this is a formal treatment only, because one might ask what assumptions on \( u \) need to be made before we can assume that it can be written as an infinite series. Therefore, we discuss the subject of convergence of these infinite series which we call Fourier series, next.


4 Convergence of Fourier Series

So far we have talked about approximations that are expanded in a finite series of orthogonal eigenfunctions for a Sturm-Liouville operator. We would like to know if continuous functions can be represented by an infinite series of the type we have studied so far. For the rest of this section we will be concerned with functions that are continuous and periodic with period 2π. First we denote the space \( \mathbb{T} = \mathbb{R}/2\pi \mathbb{Z} \). Then, \( f : \mathbb{T} \to \mathbb{C} \) such that \( f(t) = f(t + 2k\pi), \ k \in \mathbb{N} \). Let us now suppose that \( f(t) \) can be expressed as an infinite series of the form

\[
 f(t) = \sum_{r=-\infty}^{\infty} \hat{f}(r) e^{irt} \tag{4.1}
\]

To get an idea of how we want to define \( \hat{f}(r) \) we multiply both sides of (4.1) by \( e^{-ikt} \) and integrate from 0 to 2π. This gives us

\[
 \int_{0}^{2\pi} f(t) e^{-ikt} \, dt = \int_{0}^{2\pi} \sum_{r=-\infty}^{\infty} \hat{f}(r) e^{irt} e^{-ikt} \, dt \\
 = \sum_{r=-\infty}^{\infty} \hat{f}(r) \int_{0}^{2\pi} e^{irt} e^{-ikt} \, dt \\
 = 2\pi \hat{f}(k) \quad \text{since} \quad \int_{0}^{2\pi} e^{irt} e^{-ikt} \, dt = \begin{cases} 0 & \text{if } r \neq k \\ 2\pi & \text{if } r = k \end{cases}.
\]

This implies that

\[
 \hat{f}(k) = (2\pi)^{-1} \int_{0}^{2\pi} f(t) e^{-ikt} \, dt. \tag{4.2}
\]

We note that this derivation is completely formal since we do not know for sure if we can interchange summation and integration. Nevertheless we define \( \hat{f}(r) \) in (4.2) to be the Fourier coefficients of the function \( f : \mathbb{T} \to \mathbb{C} \). We also note that (4.2) can be expressed as

\[
 \hat{f}(k) = \frac{\langle f, e^{-ikt} \rangle}{\langle e^{-ikt}, e^{-ikt} \rangle}.
\]

Now, let

\[
 S_n(f, t) = \sum_{r=-n}^{n} \hat{f}(r)e^{irt}. \tag{4.3}
\]

Normally we would look at the sequence \( \{S_n(f, t)\}_{n=0}^{\infty} \) and try to show \( S_n \to f \). Fourier claimed that this was always true and in his book *Théorie Analytique de la Chaleur* showed how formulae of the kind \( \sum_{r=-\infty}^{\infty} \hat{f}(r)e^{irt} \) could be used to solve certain linear partial differential equations. After several mathematicians had produced more or less misleading proofs of convergence, Dirichlet took up the problem and in a paper which set up new standards of rigour and clarity in analysis he was able to prove convergence under quite general conditions. He showed that if \( f \) is continuous and has a bounded continuous derivative except, possibly, at a finite number of points then \( S_n(f, t) \to f(t) \) as \( n \to \infty \) at all points \( t \) where \( f \) is continuous. It turned out that the conditions on \( f \) could not be relaxed indefinitely. The question was then raised that asked, if \( f \) is a continuous function from \( \mathbb{T} \) to \( \mathbb{C} \) then, given the Fourier coefficients of \( f \) as defined above, can we find \( f(t) \) for \( t \in \mathbb{T} \)? Fejér (then aged only 19) showed that the answer was yes. He started from the observation that if a sequence \( s_0, s_1, \ldots \) is not terribly well behaved, its behavior may be improved by considering averages \( s_0, (s_0 + s_1)/2, (s_0 + s_1 + s_2)/3, \ldots \) The first person to study this
phenomena was Cesàro about ten years before Fejér’s discovery. We will be looking at “Cesàro” sums. These are sums of the form

$$\sigma_n(f, t) = \frac{1}{n+1} \sum_{j=0}^{n} S_j(f, t)$$

We can see from definition that we are looking at averages of the $S_n$’s when we consider these “Cesaro” sums. So, what does $\sigma_n(f, t)$ look like? First, for the following calculation let $\hat{g}(r) = \hat{f}(r)e^{irt}$.

$$\sigma_n(f, t) = \frac{1}{n+1} \sum_{j=0}^{n} S_j(f, t)$$

$$= \frac{1}{n+1} \sum_{j=0}^{n} \sum_{r=-j}^{j} \hat{g}(r)$$

$$= \frac{1}{n+1} \left[ \hat{g}(0) + (\hat{g}(-1) + \hat{g}(0) + \hat{g}(1)) + \ldots + (\hat{g}(-n) + \hat{g}(1-n) + \ldots + \hat{g}(-1) + \hat{g}(0) + \hat{g}(1)) + \ldots + \hat{g}(n-1) + \hat{g}(n) \right]$$

$$= \frac{1}{n+1} \left[ (n+1)\hat{g}(0) + n(\hat{g}(-1) + \hat{g}(1)) + (n-1)(\hat{g}(-2) + \hat{g}(2)) + \ldots + 2(\hat{g}(1-n) + \hat{g}(n-1)) + (\hat{g}(-n) + \hat{g}(n)) \right]$$

$$= \frac{1}{n+1} \sum_{r=-n}^{n} (n+1-r)(\hat{g}(-r) + \hat{g}(r)) + \hat{g}(0)$$

$$= \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1} \hat{g}(r)$$

$$= \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1} \hat{f}(r) e^{irt}$$

Therefore, we have

$$\sigma_n(f, t) = \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1} \hat{f}(r) e^{irt}$$

We can simplify (4.5) even further by noting that

$$\sum_{r=-n}^{n} \frac{n+1-|r|}{n+1} \hat{f}(r) e^{irt} = \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1} \int_{0}^{2\pi} f(s) e^{-irt} ds e^{irt}$$

$$= \int_{0}^{2\pi} f(s) \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1} e^{irt} ds$$

$$= \int_{0}^{2\pi} f(s) K_n(t-s) ds$$

where $K_n(t-s) = \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1} e^{irt}$

$$= \int_{t-2\pi}^{t} f(x-t) K_n(x) dx$$

$$= \int_{0}^{2\pi} f(x-t) K_n(x) dx$$

since $f$ and $K_n$ are periodic.

Finally, we have

$$\sigma_n(f, t) = \int_{0}^{2\pi} f(x-t) K_n(x) dx$$

(4.6)
where

\[ K_n(x) = \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1} e^{irx} \]

This form of \( \sigma_n(f,t) \) suggests that we take a closer look at the structure of \( K_n(x) \), and we will do just that. First, notice that

\[ K_n(x) = \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1} e^{irx} = \frac{1}{n+1} \sum_{j=0}^{n} \sum_{r=-j}^{j} e^{irx}, \]

then expanding and rearranging terms we find that when \( x \neq 0 \),

\[
K_n(x) = \sum_{r=-n}^{n} \frac{n+1-|r|}{n+1} e^{irx} \\
= \frac{1}{n+1} \left( \sum_{k=0}^{n} e^{i(k-\frac{n}{2})x} \right)^2 \\
= \frac{1}{n+1} \left( e^{-i(\frac{n}{2})x} \left( \frac{1 - e^{i(n+1)x}}{1 - e^{ix}} \right) \right)^2 \\
= \frac{1}{n+1} \left( e^{-i(\frac{n+1}{2})x} - e^{i(\frac{n+1}{2})x} \right)^2 \\
= \frac{1}{n+1} \left( \frac{\sin(\frac{n+1}{2}x)}{\sin(\frac{x}{2})} \right)^2
\]

and by direct calculation we see that

\[ K_n(0) = n + 1. \]

At this point we notice three important properties of \( K_n(x) \).

1. \( K_n(x) \geq 0 \) for \( s \in \mathbb{T} \)
2. \( \forall \delta > 0, \ K_n(x) \to 0 \) uniformly for \( \delta \leq |x| \leq \pi \)
3. \( (2\pi)^{-1} \int_{\mathbb{T}} K_n(x) \, dx = 1 \)

The proof of (1.) is obvious.
To prove (2.) we note that for \( \delta > 0 \) and \( \delta \leq |x| \leq 2\pi \),
\[
|K_n(x)| \leq \frac{1}{n+1} \left( \frac{1}{\sin(\frac{x}{2})} \right)^2 \\
\leq \frac{1}{n+1} \left( \frac{1}{\sin(\frac{\pi}{2})} \right)^2 \to 0 \text{ as } n \to \infty
\]

The proof of (3.) is a straightforward calculation.
Now we have everything we need to prove

**Fejér’s Theorem 1.** If \( f : \mathbb{T} \to \mathbb{C} \) is periodic with period \( 2\pi \) and continuous then

\[ \sigma_n(f,t) \to f(t) \quad \text{uniformly on } \mathbb{T}. \]
Proof: First, since $f$ is continuous on $\mathbb{T}$ it is uniformly continuous and bounded. Therefore \( \exists M \in \mathbb{R} \ni |f(x)| \leq M \) for all $x \in \mathbb{T}$.

Also, let $\epsilon > 0$ be given. Then $\exists \delta > 0 \ni |x| < \delta \Rightarrow |f(t - x) - f(t)| < \frac{\epsilon}{2}$ for all $x \in \mathbb{T}$. From property (2.) we have that $\exists N \in \mathbb{N} \ni n \geq N \Rightarrow |K_n(x)| < \frac{1}{4M\pi}$ for $\delta \leq |x| \leq \pi$. Therefore, for $n \geq N$ we have

\[
|\sigma_n(f, t) - f(t)| = \left| (2\pi)^{-1} \int_{\mathbb{T}} f(t - x) K_n(x) \, dx - f(t) \right|
\]

\[
= \left| (2\pi)^{-1} \int_{\mathbb{T}} f(t - x) K_n(x) \, dx - (2\pi)^{-1} \int_{\mathbb{T}} f(t) K_n(x) \, dx \right| \text{ from property (3.)}
\]

\[
\leq (2\pi)^{-1} \int_{\mathbb{T}} |f(t - x) - f(t)| K_n(x) \, dx
\]

\[
= (2\pi)^{-1} \int_{-\pi}^{\pi} |f(t - x) - f(t)| K_n(x) \, dx
\]

\[
= (2\pi)^{-1} \int_{-\pi}^{-\delta} |f(t - x) - f(t)| K_n(x) \, dx + (2\pi)^{-1} \int_{-\delta}^{\delta} |f(t - x) - f(t)| K_n(x) \, dx + (2\pi)^{-1} \int_{\delta}^{\pi} |f(t - x) - f(t)| K_n(x) \, dx
\]

\[
< (\pi - \delta) \frac{M \epsilon}{4M\pi} + \frac{\epsilon}{2} (2\pi)^{-1} \int_{-\delta}^{\delta} K_n(x) \, dx + (\pi - \delta) \frac{M \epsilon}{4M\pi}
\]

\[
\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} (2\pi)^{-1} \int_{-\pi}^{\pi} K_n(x) \, dx
\]

\[
= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \quad \text{for all } t \in \mathbb{T} \quad \blacksquare
\]
5 Applications of Fejér’s Theorem

A nice consequence of Fejér’s theorem is

**Theorem 2.** If \( f : \mathbb{T} \to \mathbb{C} \) is periodic with period \( 2\pi \) and continuous and \( \epsilon > 0 \) then there exists a trigonometric polynomial \( P(t) \), such that

\[
\sup_{t \in T} |P(t) - f(t)| < \epsilon.
\]

**Proof:** First we notice that \( P(t) = \sum_{r=-n}^{n} \hat{f}(r)e^{irt} \) is a trigonometric polynomial. Therefore, theorem 2 follows by Fejér’s theorem. \( \square \)

What this tells us is that the trigonometric polynomials are dense in the continuous functions (w.r.t the uniform metric). Right now we state a theorem of Weierstrass which is

**Theorem 3 (Weierstrass).** If \( f : [a, b] \to \mathbb{C} \) is continuous and \( \epsilon > 0 \) we can find a polynomial \( P \) such that

\[
\sup_{t \in [a,b]} |P(t) - f(t)| < \epsilon.
\]

Since we can always define a change of variables that maps the interval \([a, b]\) into the interval \([0, 1]\), we will prove the following simpler theorem of Weierstrass which is

**Theorem 4.** If \( f : [0, 1] \to \mathbb{C} \) is continuous and \( \epsilon > 0 \) we can find a polynomial \( P \) such that

\[
\sup_{t \in [0,1]} |P(t) - f(t)| < \epsilon.
\]

**Proof:** Let \( f : [0, 1] \to \mathbb{C} \) be continuous and let \( \epsilon > 0 \). In order to be able to use Theorem 2 we define \( g : \mathbb{T} \to \mathbb{C} \) by

\[
g(t) = f(|t|) \quad \text{for } |t| \leq 1,
\]

\[
g(t) = f(1) \quad \text{for } |t| > 1.
\]

Now, since \( g \) is continuous on \( \mathbb{T} \), by Theorem 2 we can find a trigonometric polynomial, \( \sum_{r=-n}^{n} a_r e^{irt} \), such that

\[
\left| g(t) - \sum_{r=-n}^{n} a_r e^{irt} \right| < \frac{\epsilon}{2}.
\]

Furthermore, from advanced calculus we know that \( \sum_{k=0}^{n} (is)^k/k! \) converges uniformly to \( e^{is} \) on any bounded interval \([-R, R]\). Therefore, for any \(-n \leq r \leq n\) we can find \( m(r) \) such that

\[
\sum_{k=0}^{m(r)} \frac{(irt)^k}{k!} - e^{irt} \leq \frac{\epsilon}{2(2n+1)|a_r|+1} \quad \text{for all } |t| \leq 1.
\]

Define

\[
P(t) = \sum_{r=-n}^{n} a_r \sum_{k=0}^{m(r)} \frac{(irt)^k}{k!}.
\]
Clearly, $P(t)$ is a polynomial and

$$|P(t) - f(t)| = |P(t) - g(t)| \leq \left| P(t) - \sum_{r=-n}^{n} a_r e^{irt} \right| + \left| \sum_{r=-n}^{n} a_r e^{irt} - g(t) \right|$$

$$< \sum_{r=-n}^{n} |a_r| \left| \sum_{k=0}^{m(r)} (irt)^k / k! - e^{irt} \right| + \frac{\epsilon}{2}$$

$$\leq \sum_{r=-n}^{n} |a_r| \frac{\epsilon}{2(2n+1)|a_r| + 1} + \frac{\epsilon}{2}$$

$$< \sum_{r=-n}^{n} \frac{\epsilon}{2(2n+1)} + \frac{\epsilon}{2}$$

$$= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \text{ for all } 0 \leq t \leq 1 \quad \blacksquare$$

We have now shown that the polynomials are also dense in the continuous functions (w.r.t the uniform metric).
6 Convergence of Rayleigh-Ritz

We conclude the paper with a final remark about the convergence of the Rayleigh-Ritz method as we have presented it here. Earlier we showed that if we choose, as our basis functions, eigenfunctions of the Sturm-Liouville operator, then the coefficients we obtain through least squares are equivalent to the fourier coefficients of the exact solution. This is a reassuring observation since we know the general form for the eigenfunctions of the Sturm-Liouville problem

\[-\frac{d}{dx} \left( p(x) \frac{du}{dx} \right) + q(x)u(x) = f(x) \quad \text{for } a \leq x \leq b,\]

\[u(a) = u(b) = 0.\]

Moreover, we know that we obtain an infinite set of linearly independent trigonometric eigenfunctions which we have shown to be dense in the continuous functions with respect to the uniform metric. Therefore when we choose as our basis functions, the eigenfunctions of the operator, then we can expect our approximation to converge to the actual solution as we increase the number of eigenfunctions used in our approximation.

References

