

Mat1062: Computational Methods for PDE

Mary Pugh

March 13, 2008

1 Overview of Projection methods

Throughout this course, we have discussed Fourier modes in the context of stability. Generally, we assume that a PDE or a discrete scheme has a solution of the form $u(x, t) = U(t) \exp(i\xi x)$ where ξ is a real number. Then we look for the time dependence of $U(t)$, and if it grows, we conclude that the method is unstable. An implicit assumption in this approach is that *any* initial data $u_0(x)$ may be written as a combination of these modes, so these special solutions are stable.

Now we turn this point of view into a full numerical method.

Suppose we have a PDE which is either time-dependent, $u_t = \mathcal{L}u + f$ or $u_{tt} = \mathcal{L}u = f$, or is elliptic $\mathcal{L}u = -f$, where \mathcal{L} is a linear differential operator such as $\mathcal{L}u = u_{xx}$. The idea behind projection methods is to approximate the time-dependent solution $u(x, t)$ with

$$u_N(x, t) = \sum_{j=1}^N c_j(t) \phi_j(x)$$

and the elliptic solution $u(x)$ with

$$u_N(x) = \sum_{j=1}^N c_j \phi_j(x).$$

Here u_N is in the N -dimensional space spanned by the basis functions ϕ_1, \dots, ϕ_N .

How shall we choose the basis functions $\{\phi_j\}$? “Good” properties are

- *Localization in space.* This may give us a natural interpretation of the coefficients $c_j = u(x_j)$. It makes the linear system we’ll have to solve sparse so the linear algebra is doable though not necessarily easy. Also,

if the basis functions are localised in space it makes it easier to refine the solution as needed — if something interesting is happening in a certain region of space we would add extra basis functions, choosing those that are localized where the action is.

- *Smoothness.* Our PDE depends on spatial derivatives, and it is nice if the approximating functions have at least a few. Using integration by parts we can reduce the required number, but we still need some.
- *Analytic simplicity.* They should be easy to work with.

Different methods focus on different combinations of virtues:

- **Finite element** methods take the ϕ_j to be piecewise low-order polynomials, based on a selection of node points. They are very well localized in space, but often have only barely enough smoothness. They work in strange-shaped regions in space. The errors are typically some power of the node point spacing h , or some negative power of the number of basis functions N .
- **Spectral methods** take the ϕ_j to be the eigenfunctions of the operator \mathcal{L} . For example, if $\mathcal{L} = \partial_{xx}$ then ϕ_j would be functions of the form $\sin(jx)$, $\cos(jx)$, or $\exp(ijx)$. If the operator is self-adjoint then the eigenfunctions are orthogonal which is very helpful. Also they are smooth. One the downside, they aren't localised in space and we have formulae for them only for certain domains. (There are theorems saying that they exist for more general domains but actually finding them is a different computational challenge.) For smooth problems, the error is typically *smaller than any power* of N .
- **Wavelet methods** use basis functions that are localized in space, but preserve some of the nice linear algebra properties of spectral methods. They are analytically rather complicated to work with.

The PDE acts on a space \mathcal{B} of functions. The approximation is in an N -dimensional subspace, \mathcal{B}_N , of \mathcal{B} . The main problems we have to address are 1) how to define the problem on \mathcal{B}_N given the original problem on \mathcal{B} and 2) how to show that as $N \rightarrow \infty$ our approximate solution u_N converges to the desired solution u .

- The **Galerkin** approximation is in terms of inner products (no node points). We rewrite the original problem as

$$\langle u_t, \phi_k \rangle = \langle \mathcal{L}u, \phi_k \rangle + \langle f, \phi_k \rangle \quad \text{for each } k = 1, \dots, N.$$

This gives the linear system

$$\sum_{j=1}^N c_j' \langle \phi_j, \phi_k \rangle = \sum_{j=1}^N c_j \langle \mathcal{L}\phi_j, \phi_k \rangle + \langle f, \phi_k \rangle, \quad k = 1, \dots, N$$

which must be solved to give (c_1, \dots, c_N') in terms of (c_1, \dots, c_N) . If necessary (if the basis functions don't have enough smoothness), we can use integration by parts to rewrite the term $\langle \mathcal{L}\phi_j, \phi_k \rangle$. Clearly it will be very advantageous if $\langle \phi_j, \phi_k \rangle = \delta_{jk}$ so we don't have to do any linear algebra at all.

- The **collocation** method is based on a set of N node points x_1, \dots, x_N (no inner products). We keep track of the solution by its values u_1, \dots, u_N with $u_j = u(x_j)$, rather than directly by the coefficients c_j . Thus this is conceptually like the finite difference method. It is much easier to incorporate nonlinear terms than with Galerkin.

2 Finite Element Methods

For finite element methods we need a Hilbert Space. This is a complete vector space which has an inner product and a countable basis.

To start, we will focus on the two-point boundary value problem

$$\begin{cases} -u''(x) = f(x) & \forall x \in (0, 1) \\ u(0) = 0 \\ u'(1) = 0 \end{cases} \quad (1)$$

Given a continuous f , the solution is easy to find:

$$u(x) = F(x) + mx + b$$

where $F''(x) = f(x)$ and the integration constants m and b are chosen to satisfy the boundary conditions. We say that u is a *classical solution* if u has two derivatives, if u'' is continuous on $(0, 1)$ and if the boundary value problem is satisfied. We would like to define a weaker type of solution, one which is based on a variational formulation. (Because the problem is exactly solvable all of this may feel a little pointless. But the point is: we will be studying a simple case using methods that generalise to harder problems.)

To understand how we might formulate a weak solution, assume that u is a classical solution and v is a continuous function whose first derivative

v' is continuous on $(0, 1)$. Then

$$\begin{aligned} \int_0^1 f(x)v(x) dx &= - \int_0^1 u''(x)v(x) dx \\ &= \int_0^1 u'(x)v'(x) dx - u'(1)v(1) + u'(0)v(0) \\ &= \int_0^1 u'(x)v'(x) dx + u'(0)v(0) \end{aligned}$$

Above, we used that $u'(1) = 0$. This calculation helps us in choosing what Hilbert Space we will look in for our solution. Specifically, we choose the Hilbert space

$$V = \left\{ v(x) \text{ real-valued functions on } \mathbb{R} \mid \int_0^1 v(x)^2 dx < \infty, \int_0^1 v'(x)^2 dx < \infty, v(0) = 0 \right\} \quad (2)$$

We then define a *weak solution* as: if $f \in L^2([0, 1])$ then u is a weak solution of the boundary value problem (1) if $u \in V$ and $\int u'v' = \int fv$ for all $v \in V$.

Before proceeding, we ask if the Hilbert space V makes sense. It's a subspace of $H^1([0, 1])$ which is a Hilbert space. And so it will be a Hilbert space in its own right. The only thing to worry about is whether it makes sense to specify the value of v at a point. If all we knew about a function, v , was that $\int v^2 < \infty$ then it would not make sense to ask what v equals at a particular point. However, it turns out that if $\int v'^2 < \infty$ then this forces v to be continuous on $[0, 1]$. And so it does make sense to talk about the value of v at a point. Note that the space V has one of the two boundary conditions built in to it. And so if $u \in V$ it automatically satisfies $u(0) = 0$. A natural question is whether or not we know that $u'(1) = 0$. (Note: if all we know is that $\int u'^2 < \infty$ then we can't know pointwise information like $u'(1) = 0$. But if we happen to know that u' is continuous then we would hope that $u'(1) = 0$ follows somehow.)

Now that we're satisfied with the definition of V we ask whether

$$\int_0^1 u'(x)v'(x) dx = \int_0^1 f(x)v(x) dx, \quad \forall v \in V$$

makes sense. Specifically, we need that each of these integrals is finite. This follows via the Schwartz inequality:

$$\left| \int_0^1 f(x)v(x) dx \right| \leq \sqrt{\int_0^1 f(x)^2 dx} \sqrt{\int_0^1 v(x)^2 dx} < \infty,$$

where we used that $f \in L^2([0, 1])$ and that $v \in V$. Similarly,

$$\left| \int_0^1 u'(x)v'(x) dx \right| \leq \sqrt{\int_0^1 u'(x)^2 dx} \sqrt{\int_0^1 v'(x)^2 dx} < \infty$$

where we used that $u, v \in V$.

Now that we're satisfied with our definition of weak solution, we ask the natural question: "If u is smooth and is a weak solution does this imply u is a classical solution?" The answer is "yes". (If the answer were "no" then we'd really have to question our definition of weak solution.) Before proving this, we introduce some notation:

$$a(u, v) := \int_0^1 u'(x)v'(x) dx \quad \langle f, v \rangle := \int_0^1 f(x)v(x) dx$$

Theorem Assume f and u'' are continuous on $[0, 1]$. Let the space V be as defined in (2). If $u \in V$ and $a(u, v) = \langle f, v \rangle$ for all $v \in V$ then u is a classical solution of the boundary value problem (1).

Proof Choose $v \in V$ such that v' is continuous on $[0, 1]$. Then

$$\begin{aligned} \int_0^1 u'(x)v'(x) dx &= \int_0^1 f(x)v(x) dx \\ \implies - \int_0^1 u''(x)v(x) dx + u'(1)v(1) - u'(0)v(0) &= \int_0^1 f(x)v(x) dx \\ \implies - \int_0^1 u''(x)v(x) dx + u'(1)v(1) &= \int_0^1 f(x)v(x) dx \\ \implies u'(1)v(1) &= \int_0^1 (f(x) + u''(x)) v(x) dx \end{aligned} \tag{3}$$

The identity (3) holds for any $v \in V$ that has a continuous first derivative. We now show that this implies that the continuous function $f + u''$ is identically zero on $[0, 1]$. Assume that $f + u''$ isn't identically zero on $[0, 1]$. Because $f + u''$ is continuous on $[0, 1]$ this would imply we could find an interval $(x_0, x_1) \subset [0, 1]$ on which $f + u''$ is positive. (If we can find no such interval then we can find an interval on which $f + u''$ is strictly negative.) We use this interval to construct a specific test function

$$v(x) = \begin{cases} (x - x_0)^2(x - x_1)^2 & x_0 \leq x \leq x_1 \\ 0 & \text{otherwise} \end{cases}$$

For this test function, (3) becomes

$$0 = \int_0^1 (f(x) + u''(x)) v(x) dx = \int_{x_0}^{x_1} (f(x) + u''(x)) v(x) dx > 0$$

which is impossible. (Note: if $f + u''$ were strictly negative on the interval then the above would have yielded $0 < 0$ which is again impossible.) This proves that $-u''(x) = f(x)$ at all points in $[0, 1]$. As a result, identity (3) reduces to

$$u'(1)v(1) = 0$$

for any $v \in V$ that has a continuous first derivative. For example, it holds for $v(x) = x$ resulting in $u'(1)v(1) = u'(1) = 0$. There was nothing really special about this choice — any $v \in V$ which has continuous first derivative (so that (3) applies) and has $v(1) \neq 0$ would have resulted in $u'(1) = 0$. This finishes the proof.

We now have a weak formulation that we are happy with:

$$u \in V, \quad a(u, v) = \langle f, v \rangle \quad \forall v \in V \quad (4)$$

Another way to understand this would be

$$u \in V, \quad \int_0^1 u'(x)v'(x) - f(x)v(x) dx = 0 \quad \forall v \in V$$

which is the same thing as looking at the first variation in V of the quantity

$$\mathcal{E}(u) = \int_0^1 \frac{1}{2}u_x^2(x) - f(x)u(x) dx$$

Finally, we note that $a(u, v)$ is actually an inner product on V . To be an inner product we need to check the following

1. $a(u, v) = \overline{a(v, u)} \quad \forall u, v \in V$. This holds automatically because u and v are real-valued functions.
2. $a(u, v + w) = a(u, v) + a(u, w) \quad \forall u, v, w \in V$. This holds because $(v + w)' = v' + w'$.
3. $a(\lambda u, v) = \lambda a(u, v) \quad \forall u, v \in V, \forall \lambda \in \mathbb{R}$. This holds because you can pull constants out of integrals.
4. $a(u, u) \geq 0 \quad \forall u \in V$ This holds because $u'^2(x) \geq 0$ for all x .

5. $a(u, u) = 0 \iff u = 0$. It's clear that $u = 0$ implies $a(u, u) = 0$. We can rigorously prove the other direction as well. Here is a nearly-rigorous proof of why $a(u, u) = 0$ implies $u = 0$. Fix $x \in [0, 1]$. Then

$$u(x) = u(0) + \int_0^x u'(y) dy \quad \implies \quad u(x) = \int_0^x u'(y) dy.$$

Applying the Schwartz inequality, for $x \in (0, 1]$

$$|u(x)| \leq \left| \int_0^x u'(y) dy \right| \leq \sqrt{x} \sqrt{\int_0^x (u'(y))^2 dy} \leq \sqrt{a(u, u)}.$$

And so $a(u, u) = 0$ implies $u = 0$ pointwise.

2.1 The Ritz-Galerkin Approximation Problem

We seek approximate solutions of (4). We do this via subspaces of V . That is, we will consider a family of subspaces $\cdots \subset V_n \subset V_{n+1} \subset \cdots \subset V$ and in each subspace we will find and solve an approximate problem, resulting in a solution u_n . The goal is to choose the subspaces in a smart way, one that allows us to show that u_n converges to something, call it u , as $n \rightarrow \infty$ and that the limit u is a weak solution in the sense (4).

Given a subspace $V_n \subset V$, The Ritz-Galerkin approximation problem is

$$u_n \in V_n, \quad a(u, v) = \langle f, v \rangle \quad \forall v \in V_n \quad (5)$$

If V_n is a finite-dimensional subspace then there will always be a solution u_n and it will be unique:

Theorem Assume V_n is an n -dimensional subspace of V and $f \in L^2([0, 1])$. Then the Ritz-Galerkin approximation problem (5) has a unique solution u_n .

Proof: Let $\{\phi_j\}$ be a basis of V_n . First of all, if $v \in V_n$ then

$$v = \sum_{j=1}^n V_j \phi_j \quad \implies \quad a(u_n, v) = \langle f, v \rangle \iff \sum_{j=1}^n V_j a(u_n, \phi_j) = \sum_{j=1}^n V_j \langle f, \phi_j \rangle$$

as a result, it suffices to find $u_n \in V_n$ such that

$$a(u_n, \phi_j) = \langle f, \phi_j \rangle, \quad \forall 1 \leq j \leq n.$$

We seek U_1, U_2, \dots, U_n such that

$$u_n = \sum_{j=1}^n U_j \phi_j \implies a(u_n, \phi_j) = \langle f, \phi_j \rangle \iff \sum_{i=1}^n U_i a(\phi_i, \phi_j) = \langle f, \phi_j \rangle \quad \forall 1 \leq j \leq n.$$

This is a linear algebra problem. If

$$\vec{U} = \begin{pmatrix} U_1 \\ \dots \\ U_n \end{pmatrix}, \quad \vec{F} = \begin{pmatrix} \langle f, \phi_1 \rangle \\ \dots \\ \langle f, \phi_n \rangle \end{pmatrix}, \quad K_{ij} = a(\phi_i, \phi_j)$$

then we seek a solution \vec{U} of

$$\vec{U}^T K = \vec{F}^T$$

There will be a unique solution if and only if the null space contains only the zero vector. That is, we need to show

$$\vec{V}^T K = \vec{0}^T \iff \vec{V} = \vec{0}.$$

We see this as follows:

$$\begin{aligned} \vec{V}^T K = \vec{0}^T &\implies \langle \vec{V}^T K, \vec{V} \rangle = 0 \\ &\implies a(v, v) = 0 \quad \text{where } v = \sum V_j \phi_j \\ &\implies v(x) = 0 \quad \forall x \in [0, 1] \\ &\implies \vec{V} = \vec{0}. \end{aligned}$$

In the last step, we used that $\{\phi_j\}$ is a basis and therefore the only way a linear combination of ϕ_j s can equal zero is if each coefficient equals zero.

This shows that the null space is trivial which implies there exists a unique solution u_n , as desired. This finishes the proof.

The proof is helpful in that it shows us how to solve for u_n once we have a basis for the subspace V_n .

2.1.1 A piecewise linear basis

We now consider a collection of functions which we hope can become a basis for V . Fix a set of $n + 1$ points in $[0, 1]$ such that

$$0 = x_0 < x_1 < \dots < x_{n-1} < x_n = 1.$$

We call these points “nodes”. Using these nodes, we create n piecewise linear functions ϕ_j . First, let $dx_j := x_j - x_{j-1}$ for $1 \leq j \leq n$. Then for $1 \leq j \leq n-1$

$$\phi_j(x) = \begin{cases} \frac{1}{dx_j}(x - x_{j-1}) & x_{j-1} \leq x \leq x_j \\ \frac{-1}{dx_{j+1}}(x - x_{j+1}) & x_j < x \leq x_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\phi_n(x) = \begin{cases} \frac{1}{dx_n}(x - x_{n-1}) & x_{n-1} \leq x \leq x_n \\ 0 & \text{otherwise} \end{cases}.$$

Each ϕ_j is 1 at $x = x_j$. Also, note that each ϕ_j is zero at $x = 0$ and that ϕ_n is nonzero at $x = 1$. Clearly, if a linear combination of the ϕ_j s is zero then each coefficient must be zero, proving linear independence.

$$\begin{aligned} \sum_{i=1}^n C_i \phi_i(x) &= 0 \quad \forall x \in [0, 1] \\ \implies \sum_{i=1}^n C_i \phi_i(x_j) &= C_j \phi_j(x_j) = 0 \quad \forall 1 \leq j \leq n \\ \implies C_j &= 0 \quad \forall 1 \leq j \leq n \end{aligned}$$

We’d like to show that as we add more and more nodes that the set expands to become (in the limit) a basis for V . We will address this later.

We take $V_n := \text{span}\{\phi_j\}$. Rather than starting with a subspace and finding a basis for it, we start with a set of linearly independent functions and take their span to be the subspace V_n .

To solve the Ritz-Galerkin approximation problem, we need the matrix K and the vector \vec{F} . Because the elements ϕ_j are piecewise linear, their derivatives are piecewise constant which makes it easy to compute $K_{i,j} = a(\phi_i, \phi_j)$. For $1 \leq i < n$

$$\begin{cases} K_{i,i-1} = -\frac{1}{dx_i} \\ K_{i,i} = \frac{1}{dx_i} + \frac{1}{dx_{i+1}} \\ K_{i,i+1} = -\frac{1}{dx_{i+1}} \end{cases}$$

and

$$\begin{cases} K_{n,n-1} = -\frac{1}{dx_n} \\ K_{n,n} = \frac{1}{dx_n} \end{cases}$$

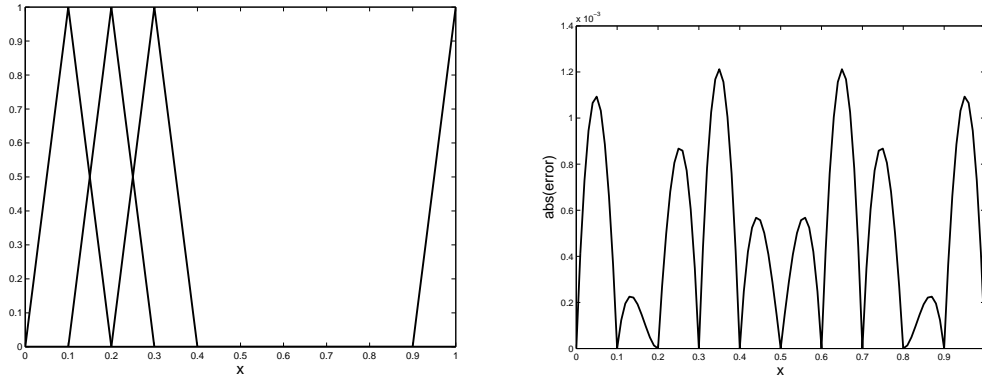


Figure 1: The eleven nodes are uniformly spaced in $[0, 1]$ with $dx = 1/10$. Left plot: the first three basis functions, ϕ_1, ϕ_2, ϕ_3 are plotted along with the last basis function, ϕ_{10} . Right plot: We take $f(x) = \cos(3\pi x)$. We solve the Ritz-Galerkin approximation problem and construct u_{10} . Plotted here is the absolute value of the error, sampled at 101 points.

2.1.2 Let's compute!

I take $f(x) = \cos(3\pi x)$, yielding the exact solution

$$u(x) = \frac{1}{9\pi^2} \cos(3\pi x) - \frac{1}{9\pi^2}.$$

I use maple to compute $\langle f, \phi_j \rangle$. Maple can do this — it's just integration by parts. This produces a long, nasty formula that I then cut and paste into my matlab program. I take the nodes to be uniformly spaced ($dx_j = 1/n$) resulting in

$$\begin{aligned} \langle f, \phi_j \rangle &= -\frac{2}{9dx\pi^2} \cos(\pi j dx) (\cos(\pi dx) - 1) \\ &\quad \cdot (1 + 2 \cos(\pi dx))^2 (-3 + 4 \cos(\pi j dx)^2) \quad \forall 1 \leq j \leq n - 1 \\ \langle f, \phi_n \rangle &= \frac{1}{9\pi^2 dx} (\cos(\pi dx) - 1) (1 + 2 \cos(\pi dx))^2 \end{aligned}$$

I solve the system $\vec{U}^T K = \vec{F}^T$ and use \vec{U} to construct

$$u_n(x) = \sum_{j=1}^n U_j \phi_j(x).$$

I find that if I evaluate the error $u - u_n$ at the nodes x_j then I get zero to machine precision. And so to evaluate the error elsewhere, I need to sample

u_n and u away from the nodes. Between each node, I sample at 9 equally spaced points. In the right plot of Figure 1, I present the absolute value of the error as a function of x . Its largest value (the L^∞ norm) is approximately $1.2e-2$.

I then test the scheme for convergence by doing seven runs. The first run has 11 nodes with uniform spacing $dx = 1/10$. The left plot of Figure 1 shows some of the basis functions. The second run has 21 nodes with uniform spacing $dx = 1/20$ and so on. The spacing decreases by a factor of two in each subsequent run. To measure the error, I sample on a uniform mesh with meshwidth $dx/10$. I compute the L^∞ norm of the error as well as the L^2 error.

number of nodes	$\ \text{err}\ _{L^\infty}$	$\ \text{err}\ _{L^2}$	ratio of L^2 errors
11	1.2e-3	6.3e-4	3.9421
21	3.1e-4	1.6e-4	3.9855
41	7.8e-5	4.0e-5	3.9964
81	2.0e-5	1.0e-5	3.9991
161	4.9e-6	2.5e-6	3.9998
321	1.2e-6	6.3e-7	3.9999
641	3.1e-7	1.7e-7	

We see that the L^2 norm of the error is decreasing by powers of 4. In fact, so is the L^∞ norm.

I now consider a non-uniform distribution of nodes. I do this by parametrizing the interval $[0, 1]$ via $x(s)$ where $x(0) = 0$ and $x(1) = 1$ and $x'(s) > 0$ (nonconstant). Specifically, I take

$$x(s) = s + \left(\frac{1}{2\pi} - \frac{1}{100} \right) \sin(2\pi s) \quad (6)$$

and then sample uniformly in s to produce nodes x_j . I chose this function because x' varies by a factor of 30. I start with 11 nodes, chosen by taking 11 equally spaced points s_j in $[0, 1]$ and applying the mapping (6). In the left plot of Figure 2, I show some of the basis functions for the case of 11 non-uniformly spaced nodes. The distance between nodes is greatest near $x = 0$ and $x = 1$ and is smallest near $x = 1/2$. In the right plot of Figure 2, I present the pointwise error. Note that the error is larger between the nodes that are further apart than between the nodes that are closer together. This is somewhat intuitive in that the function $f(x) = \cos(3\pi x)$ isn't especially different near $x = 1/2$ than near $x = 0$ and $x = 1$. You could imagine that

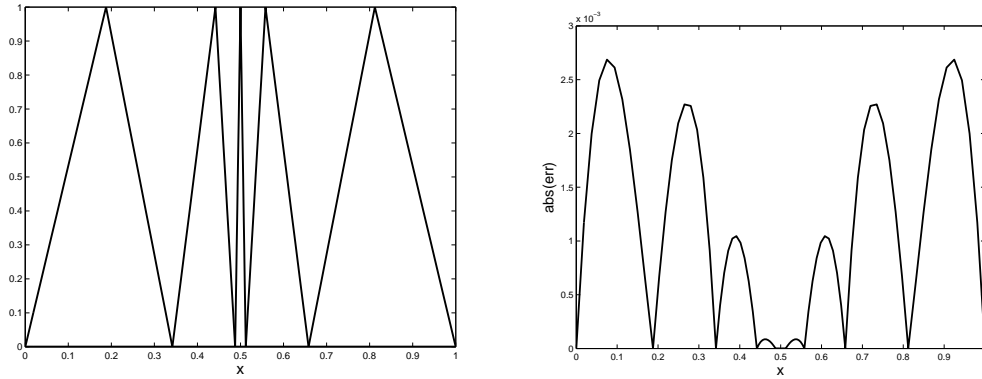


Figure 2: The eleven nodes are nonuniformly spaced in $[0, 1]$ by taking the parametrization (6) with $ds = 1/10$. Left plot: the odd-indexed basis functions, ϕ_1, \dots, ϕ_9 . Right plot: As before, we take $f(x) = \cos(3\pi x)$. We solve the Ritz-Galerkin approximation problem and construct u_{10} . Plotted here is the absolute value of the error, sampled at 101 points. Note that the error is smaller where the spatial resolution is finer.

if f had finer structure near $x = 1/2$ than near $x = 0$ and $x = 1$ then this might cause the errors near $x = 1/2$ to be comparable to, or larger than, the errors near $x = 0$ and $x = 1$.

Again, the errors decrease by a factor of 4 when I refine the mesh by factors of 2.

$\min(dx)$	$\max(dx)$	$\ err\ _{L^\infty}$	$\ err\ _{L^2}$	ratio of L^2 errors
1.2e-2	1.9e-1	2.7e-3	2.2e-3	3.6970
3.9e-3	9.6e-2	1.0e-3	5.8e-4	3.9402
1.7e-3	4.8e-2	2.8e-4	1.5e-4	3.9849
8.0e-4	2.4e-2	7.3e-5	3.7e-5	3.9962
3.9e-4	1.2e-2	1.8e-5	9.2e-6	3.9991
2.0e-4	6.1e-3	4.5e-6	2.3e-6	3.9998
9.8e-5	3.0e-3	1.1e-6	5.7e-7	