Math 521: Finite Element Methods
Spring, 2014

Instructor: Brian Wetton
Office: MATX 1107
E-mail: wetton@math.ubc.ca

Webpage: www.math.ubc.ca/~wetton/

Course Description:
Over the last few decades, finite element methods have emerged as the numerical solution methods of choice for large classes of partial differential equations arising in fluid dynamics, solid mechanics, and electromagnetics. This course is an introduction to the mathematical theory of finite element methods. We will introduce finite element discretizations for equations of various types and discuss how these discrete problems can be solved efficiently. We will address mathematical questions related to the concepts of consistency, stability, convergence, and error estimation. We will further look at applications in fluid mechanics. Implementation will be done using MATLAB or optionally with the freeware package DEAL-II.

The outline of the course is as follows:

- Finite element methods for boundary-value problems
- Nonlinear and time-dependent problems
- A-priori and a-posteriori error estimation
- Numerical Linear Algebra topics
- Mixed finite element methods for incompressible fluid flow problems

Text:
There will be no prescribed text, but there will be lecture notes available for most parts of the course material. Some optional references will be listed.

Prerequisites:
Some undergraduate level training in at least one of: partial differential equations, analysis, or numerical analysis.

Assessment:
There will be several challenging homework assignments involving both analysis and computation. In addition, students can choose whether to do a course project or have an oral final exam. Assignments are worth 60% of the final grade, the project or oral exam 40%.

Introduction to Scientific Computation: Finite Difference Methods

Brian Wetton *

September 19, 2014

1 Motivation

Many problems in Science, Engineering, and Finance involve the solution of differential equations (DE). Often these problems cannot be solved analytically but must be approximated numerically. This approximation must be done to a certain precision that depends on the application. While the differential equations do not exactly describe the real system they model (there is a modelling error) it is important to minimize the errors from the numerical approximation to be able to confirm whether the underlying mathematical model is valid. This is shown graphically in Figure 1.

2 A First DE Problem

Problem 1 *Find $u(x)$, $x \in [0,1]$ with u and its derivatives 1-periodic that satisfies*

$$-u'' + au = f(x)$$

at every x where a is a given positive constant and $f(x)$ is a given C_2 (periodic) function.

Here, C_2 (periodic) is the set of functions on the unit interval which have continuous and 1-periodic derivatives up to second order. In general, we write C_n for the set of functions that have continuous derivatives up to order n. These functions have a norm

$$\|u\|_{C_n} = \max_{0 \le j \le n} max_x |u^{(j)}(x)|$$

where $u^{(j)}$ denotes the j'th derivative of u. We will have other norms for functions, so we will label the norm we are using unless it is completely clear.

We will use the following theorem

*wetton@math.ubc.ca

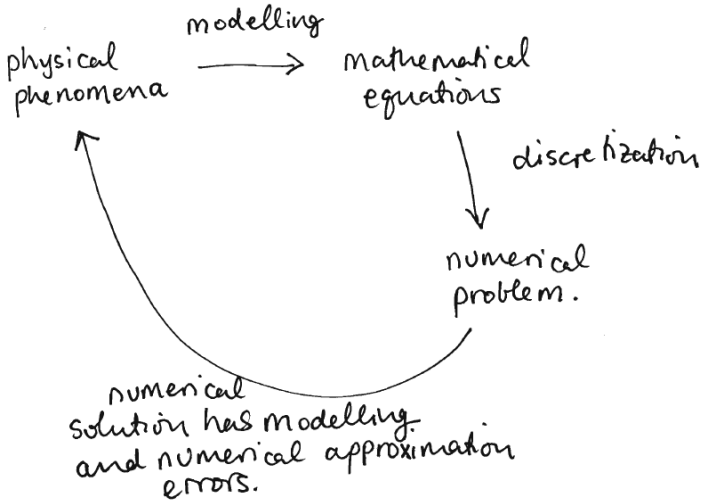


Figure 1: Stages in computational modelling. By reducing numerical errors, a clear picture of the model accuracy can be obtained.

Theorem 1 *Problem 1 has a ! (unique) solution with $u \in C_4$ with the bound*

$$\|u\|_{C_4} \leq K\|f\|_{C_2}$$

for all given $f \in C_2$ for some K that depends on a but not f.

The inequality in the Theorem is known as an a-priori bound. For a given f we don't know the solution u but we know there will be a solution, there will only be one solution, and it will have four continuous derivatives with size limited by the derivatives up to second order of f.

3 Finite Difference Discretization

3.1 Discretization

Determining the solution $u(x)$ of Problem 1 requires finding an infinite number of unknowns (the values of u at every point x in an interval). To proceed computationally, we need to deal with only a finite number of unknowns (discretization). Let's look first at a simple, Finite Difference (FD) discretization. Let U_i, $i = 1, 2, \ldots N$ approximate $u(x)$ at the ends of subintervals with length $h = 1/N$. That is

$$U_i \approx u(ih).$$

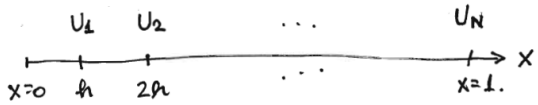


Figure 2: Uniform grid in spatial discretization.

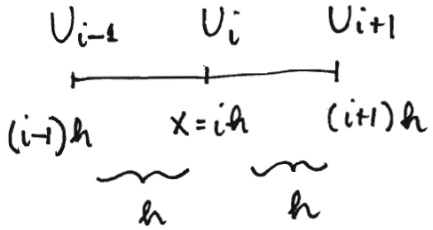


Figure 3: Values used in the finite difference approximation of the second derivative.

This is shown in Figure 2. Using the periodicity of the problem, we would have

$$U_0 = U_N \text{ and } U_{N+1} = U_1 \tag{1}$$

Convention: In these notes, I will always use lower case letters for exact solutions and upper case letters for numerically computed, approximate values.

3.2 Approximating Derivatives

Let us assume for the moment that we knew the exact solution, at least at grid points. We can use the values u_{i-1}, u_i and u_{i+1} to approximate $u''(ih)$ with the following formula:

$$D_2 u_i := \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = u''(ih) + \frac{h^2}{12} u''''(\theta) \tag{2}$$

for some $\theta \in [(i-1)h, (i+1)h]$. The geometry of this linear combination of values is shown in Figure 3. The geometry and the weights of the linear combination is called the *stencil* of the discrete approximation.

This can be derived in a number of ways, starting with Taylor Series. Two approaches are described in section 3.5 below. The last term on the right is an error term (we wanted $u''(ih)$ but got that extra term as well). Since u'''' is bounded, we can guarantee answers as accurate as we want by taking $h \to 0$

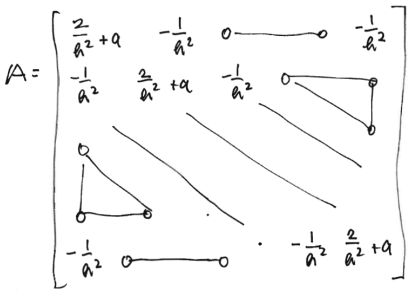


Figure 4: Structure of the matrix $\mathcal{A}$.

(that is, by refining the grid). The term h^2 in the error makes this a *second order approximation.*

3.3 Discrete Equations

Using the equation that u solves in Problem 1 we can write

$$-D_2 u_i + a u_i = f(ih) + \frac{h^2}{12} u''''(\theta) \tag{3}$$

The last term above is again an error term. We call it the *truncation error*, the residual when we put the exact solution into a discrete equation. The truncation error goes to zero as $h \to 0$. We say that the discrete equation is *consistent.* By ignoring a small (as $h \to 0$) residual, we could specify our discrete scheme for the approximate values U_i:

$$-D_2 U_i + a U_i = F_i \quad \text{for } i = 1, 2, \ldots N \tag{4}$$

where $F_i = f(ih)$ are given values. Note that (4) is a linear system with N equations for N unknowns. The system can be written in vector form

$$\mathcal{A}\mathbf{U} = \mathbf{F} \tag{5}$$

where $\mathcal{A}$ is an $N \times N$ matrix with form shown in Figure 4. We will show below that $\mathcal{A}$ is invertible, so the discrete scheme (4) has a solution $\mathbf{U}$ for every given right hand side $\mathbf{F}$ for any h.

The matrix $\mathcal{A}$ has mostly zeros. As $N \to \infty$ ($h \to 0$) the fraction of non-zeros decreases. We call such matrices *sparse.* However, $\mathcal{A}^{-1}$ is not sparse. It has no zero entries. This implies that all entries of $\mathbf{F}$ affect solution values at all locations. This is characteristic of elliptic problems.

3.4 The Next Step: Test on a Known Problem

We have a scheme for our problem. The very next thing we should do is test out the scheme on as simple a problem as we can that has a known solution. One trick is to *pick* the exact solution u, put it into the DE, and whatever residual there is call it $f(x)$. We can then use the values of f on the grid ($\mathbf{F}$) in the discrete scheme and compute $\mathbf{U}$. We can then compare $\mathbf{U}$ to the exact u at grid points to see how accurate the scheme is at various grid resolutions. For example, we could choose

$$u(x) = e^{\cos x}$$

and find

$$f(x) = (a - \sin^2 x + \cos x)e^{\cos x}$$

where we have taken the periodic interval of Problem 1 to be $[0, 2\pi]$ instead of $[0, 1]$ to make these expressions a little simpler. Choosing good test examples is somewhat of an art. You want them to represent the type of problem you are interested in, but much easier to compute (possibly lower dimensional) with an exact solution you know. In some cases, it is not possible to find an exact solution to a representative problem. In well established fields, there are often benchmark problems with high accuracy solutions you can use to test new schemes. As you are picking test problems, make sure that the solutions you use do not have zero values for the derivatives that compose the truncation error. The scheme will behave anomalously (better than usual) on such problems.

In your test, you should see how errors behave as the grid is refined (by factors of 2 for example, $h = 1/10, 1/20, 1/40, \ldots$). There are three things to look for:

1. As $h \to 0$, do the computed values U tend to the exact u values? That is, do the errors tend to zero? If so, we say that the scheme converges. The computational test is not a proof of convergence, but is strong evidence for it.

2. Is there any odd behaviour in the errors $U - u$? Odd behaviour could be an indication of a program error or it might be just a characteristic of the scheme. Odd error behaviours can be called *numerical artifacts*.

3. If the test case is (roughly) similar to problems you are actually interested in, can you achieve the desired accuracy with an N (h) value that leads to a computation that takes an acceptable length of time? If not, you should spend some time exploring ways to make the computation more efficient or more powerful computer architectures.

Testing the method on the test example above with exact solution with the parameter $a = 1$ gives the results shown in Table 1. MATLAB code for this computation is provided. Plots of the errors as functions of x for $N = 40$ and $N = 80$ are shown in Figure 5. Note that the error in the computed solution goes down by a factor of approximately 4 when N is doubled (h is halved). This matches with our discussion above, where truncation error goes down by a

N	$E_N = \max_i \lvert U_i - u(ih)\rvert$	E_N/E_{2N}
10	6.71e-2	4.24
20	1.58e-2	4.05
40	3.90e-3	4.01
80	9.73e-3	

Table 1: Errors in the finite difference method applied to the example in Section 3.4.

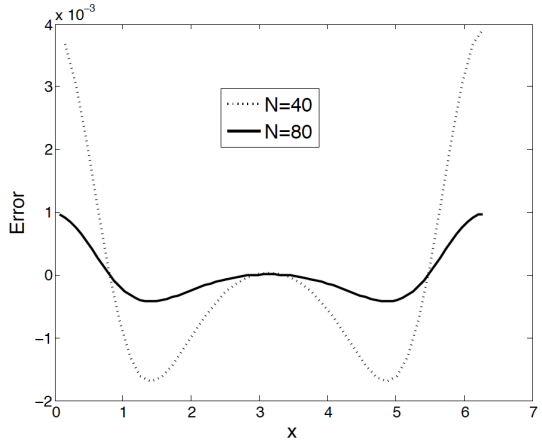


Figure 5: Spatial structure of the errors in the finite difference method applied to the example in Section 3.4.

factor of 4 when N is doubled. The exact relationship between truncation error and solution error is derived in Section 3.7 where we prove convergence of the method. Note that we have chosen to measure the error in the maximum norm

$$\|\mathbf{U}\|_\infty := \max_i |U_i|$$

in this case. Other norms could have been used and for some types of problems, other norms are more appropriate.

3.5 Derivation of FD formulae

In this section, we prove the result (2). We begin with Taylor's Polynomial Approximation Theorem:

Theorem 2 *If $f(x)$ is C_{n+1} in a neighbourhood of a then if x is in that neigh-*

bourhood,

$$f(x) = f(a)+f'(a)(x-a)+\frac{1}{2}f''(a)(x-a)^2+\cdots\frac{1}{n!}f^{(n)}(a)(x-a)^n+\frac{1}{(n+1)!}f^{(n+1)}(\theta)(x-a)^{n+1}$$

for some $\theta \in (a, x)$.

Here, the value of $f(x)$ is approximated by information at $x = a$. The first $n+1$ terms in the expression above are the n'th order Taylor Polynomial approximation of f based at $x = a$ and the last term is a remainder term, the error in the approximation. In case you have never seen the proof of this theorem, I will show the $n = 1$ case (linear approximation) in Section 3.5.1 below. We can use the Theorem to verify the properties of D_2:

$$\begin{aligned}
D_2u_i &= \frac{u_{i-1}-2u_i+u_{i+1}}{h^2} \quad (\text{let } a = ih) \\
&= \frac{1}{h^2}\left(u(a-h)-2u(a)+u(a-h)\right) \quad (\text{use the Theorem}) \\
&= \frac{1}{h^2}\Bigg(u(a)-hu'(a)+\frac{h^2}{2}u''(a)-\frac{h^3}{6}u'''(a)+\frac{h^4}{24}u''''(\theta_1) \\
&\quad -2u(a)+u(a)+hu'(a)+\frac{h^2}{2}u''(a)+\frac{h^3}{6}u'''(a)+\frac{h^4}{24}u''''(\theta_2)\Bigg)
\end{aligned}$$

where $\theta_1 \in ((i-1)h, ih)$ and $\theta_2 \in (ih, (i+1)h)$. Continuing with the expression above, we have the desired expression

$$\begin{aligned}
D_2u_i &= u''(a)+\frac{h^2}{12}\left(u''''(\theta_1)+u''''(\theta_2)\right)/2 \\
&= u''(a)+\frac{h^2}{12}u''''(\theta)
\end{aligned}$$

for some $\theta \in (\theta_1, \theta_2) \subset ((i-1)h, (i+1)h)$ where in this last step, we have used the intermediate value theorem.

This shows that the D_2 stencil has the desired properties. We could have found the coefficient values in the stencil starting with the same machinery. Taking the Ansatz

$$D_2u_i = \alpha u_{i-1}+\beta u_i+\gamma u_{i+1}$$

and wanting D_2u_i to be $u''(ih)$ with a small truncation error leads to the following requirements for the coefficients when u_{i-1} and u_{i+1} are expanded in Taylor polynomials as above:

$$\begin{aligned}
O(1) \text{ terms:} &\quad \alpha+\beta+\gamma = 0 \\
O(h) \text{ terms:} &\quad -h\alpha+h\gamma = 0 \\
O(h^2) \text{ terms:} &\quad h^2\alpha/2+h^2\gamma/2 = 1.
\end{aligned}$$

This is a linear system that can be solved for $\alpha = 1/h^2$, $\beta = -2/h^2$, $\gamma = 1/h^2$. Note that the $O(h^3)$ term also cancels with these parameters. It is typical for centred difference approximations of even derivatives that they have order of accuracy one order higher than "expected".

3.5.1 Proof of Theorem 2, $n = 1$ case

We will use Rollé's Theorem from first year calculus:

Theorem 3 *If f is differentiable in (a, b) and continuous in $[a, b]$ and $f(a) = 0$ and $f(b) = 0$ then*

$$f'(\theta) = 0$$

for some $\theta \in (a, b)$.

Consider the $n = 1$ (linear) Taylor Approximation of Theorem 2 at a specific point $x = b$. Then consider the function

$$q(x) = f(x) - L(x) - \frac{f(b) - L(b)}{(b-a)^2}(x-a)^2$$

where $L(x) = f(a) + f'(a)(x - a)$ is the linear approximation. We want to investigate $f(b) - L(b)$, the error of the linear approximation at $x = b$. This is a constant that appears in the $q(x)$ function above. Note that $q(a) = 0$ and $q(b) = 0$ so using Rollé we know that $q'(\theta_1) = 0$ for some $\theta_1 \in (a, b)$. Also, $q'(a) = 0$ so using Rollé again we have $q''(\theta) = 0$ for $\theta \in (a, \theta_1) \subset (a, b)$. We can compute

$$q''(x) = f''(x) - 2\frac{f(b) - f(a)}{(b-a)^2}$$

so $q''(\theta) = 0$ gives

$$f(b) - L(b) = \frac{f''(\theta)}{2}(b-a)^2,$$

the desired result.

3.6 Direct Solution of Sparse Linear Systems

Consider the structure of the nonzero entries of the matrix $\mathcal{A}$ in the discrete problem (5) shown in figure 4. A matrix such that

$$|A_{ii}| > \sum_{j \neq i} |A_{ij}|$$

for every row i is said to be strictly diagonally dominant. Our matrix $\mathcal{A}$ has this property. It can be shown that Gaussian elimination can be applied to such matrices stably (that is, without significant growth of floating point round-off errors) without pivoting. It can be seen that Gaussian elimination and back substitution can be done for the system (5) with a finite number of operations per row independent of the number of rows. The structure of the LU decomposition of $\mathcal{A}$ is shown in figure 6. The total operation count to find the solution is $O(N)$, that is the operations are bounded by a constant times N. Thus a direct solver applied to this problem taking into account the sparsity of $\mathcal{A}$ has optional complexity.

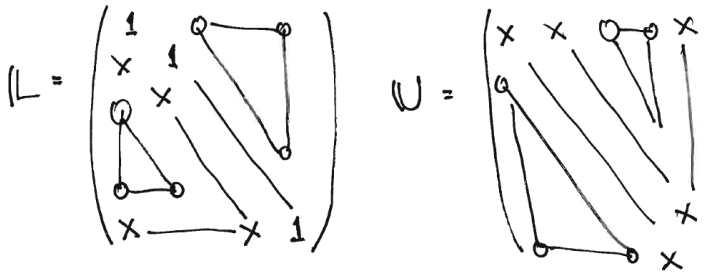


Figure 6: Structure of the LU decomposition of the matrix A.

Note: This is close to the ideal example for sparse numerical linear algebra. The matrix A is (almost) narrowly banded, is diagonally dominant, is symmetric and positive definite. We will see that the situation for discretizations in higher dimensional problems is not as ideal. Still, it is possible now to solve 3D problems of *modest* size with direct solvers on basic computers.

3.7 Convergence Proof

We came up with our discrete scheme (4) by neglecting a small truncation error in relationships the exact solution at the grid points satisfy (3):

$$\begin{aligned} \mathcal{A}\mathbf{U} &= \mathbf{F} \\ \mathcal{A}\mathbf{u} &= \mathbf{F} + \tau \end{aligned}$$

where $\tau_i = h^2 u^{(4)}(\theta_i)/12$ are the truncation errors at every grid point. As discussed above, the truncation errors go to zero as $h \to 0$ (the definition of a consistent scheme). The vector of errors in the computed solutions at grid points is

$$\mathbf{E} = \mathbf{U} - \mathbf{u}.$$

For this linear problem, we can take the difference of the two equations above to obtain

$$\mathcal{A}\mathbf{E} = \tau \quad \text{or} \quad \mathbf{E} = \mathcal{A}^{-1}\tau. \tag{6}$$

Note: We haven't shown yet that A is invertible (but our numerical implementation suggests it is for all h) and in practice we would never compute the full matrix $\mathcal{A}^{-1}$. This is just a representation for the theory.

Considering (6) we see that with τ small, $\mathbf{E}$ will be small as long as multiplying by $\mathcal{A}^{-1}$ does not increase its size by more than a constant (independent of h). Formally, we want to have the following property:

$$\|\mathcal{A}^{-1}\mathbf{b}\|_\infty \leq C\|\mathbf{b}\|_\infty \tag{7}$$

for all $\mathbf{b}$ with C independent of h. Here,

$$\|\mathbf{b}\|_\infty := \max_{i=1,\ldots N} |b_i|$$

is the maximum norm of vectors. For fixed h we can define

$$\|\mathcal{A}^{-1}\|_\infty := \max_{\mathbf{b}\neq\mathbf{0}} \frac{\|\mathcal{A}^{-1}\mathbf{b}\|_\infty}{\|\mathbf{b}\|_\infty}.$$

This is known as an induced matrix norm. The value $\|\mathcal{A}^{-1}\|_\infty$ is the smallest value C for fixed N such that (7) holds for all $\mathbf{b}$. The property we are looking for is then that

$$\sup_h \|\mathcal{A}^{-1}\|_\infty$$

is finite. This property defines *maximum norm stability* of the scheme. Showing stability of numerical schemes in general is quite difficult, but easy for this particular scheme. Consider the left hand equation of (6) where we look back to (4) to see the details of the matrix $\mathcal{A}$:

$$\begin{aligned} -D_2E_i + aE_i &= \tau_i \\ -\frac{1}{h^2}E_{i-1} + (\frac{2}{h^2} + a)E_i - \frac{1}{h^2}E_{i+1} &= \tau_i. \end{aligned} \tag{8}$$

Suppose that the $\max_i |E_i|$ is attained at an index j and that $E_j > 0$. Thus, $E_{j-1} \leq E_j$ and $E_{j+1} \leq E_j$ and thus

$$2E_j - E_{j+1} - E_{j-1} \geq 0$$

When this used in (8) we find that

$$aE_j \leq \tau_j \quad \text{or} \quad E_j \leq \frac{1}{a}|\tau_j|.$$

Since $\|\mathbf{E}\|_\infty = E_j$ we have

$$\|\mathbf{E}\|_\infty \leq \frac{1}{a}\|\tau\|_\infty.$$

Since $\mathbf{E} = \mathcal{A}^{-1}\tau$ we have shown the maximum norm stability of the scheme. If $\|\mathbf{E}\|_\infty$ is attained at an index j where $E_j < 0$, a similar argument applies.

Recall that $\tau_i = \frac{h^2}{12}u^{(4)}(\theta_i)$ for some $\theta_i \in ((i-1)h, (i+1)h)$ so

$$\|\tau\|_\infty \leq \frac{\|u\|_{C_4}}{12}h^2.$$

using Theorem 1 we have

$$\|\tau\|_\infty \leq \frac{K\|f\|_{C_2}}{12}h^2.$$

Using the stability result we derived,

$$\|\mathbf{E}\|_\infty \leq \frac{K\|f\|_{C_2}}{12a}h^2. \tag{9}$$

This proves the convergence of the scheme since as $h \to 0$, $\|\mathbf{E}\|_\infty \to 0$. With $\|\mathbf{E}\|_\infty < Ch^2$ we say that the convergence is *second order*.

The analysis leading to the convergence result above is an example of the Lax Equivalence Theorem for linear problems, often stated informally as

```
Consistency + Stability = Convergence
```

Note that in (9) that if $\|f\|_{C_2}$ is large (*i.e.* f is highly oscillatory) then h must be quite small to make the errors small. This makes sense: to resolve oscillatory behaviour, a fine grid is needed.

3.8 von Neumann Analysis

Consider again our discretization of Problem 1,

$$\mathcal{A}\mathbf{U} = \mathbf{F}, \quad \mathcal{A} = -D_2 + aI.$$

It is possible to show the stability of the scheme in other norms. Here, we will consider the l_2 norm, also known as the energy norm or the (scaled) Euclidean norm

$$\|\mathbf{b}\|_2 := \sqrt{h \sum_{i=1}^{N} |b_i|^2}. \tag{10}$$

Note that the scaling by h is such that if $b_i = C$ for all i, then $\|\mathbf{b}\|_2 = C$ for every N (h) since $N = 1/h$. Also, this makes the norm analogous to the continuous energy norm for the space of functions L_2:

$$\|f\|_{L_2} = \sqrt{\int_0^1 |f(x)|^2 dx}.$$

The discrete l_2 norm (10) can be seen as an approximation (trapezoidal rule) of the continuum norm.

3.8.1 Discrete Fourier vectors

To proceed, we introduce the complex valued Discrete Fourier (DF) vectors $\mathbf{f}_\alpha$:

$$f_{\alpha,j} = e^{2\pi i \alpha j h}.$$

To clarify the labelling, $\mathbf{f}_\alpha$ is a vector with N complex components for every $\alpha = 0, \ldots N-1$. The N components are indexed by j above and the i is the pure imaginary unit. The set $\{\mathbf{f}_\alpha\}$ is a basis of $\mathbf{C}_N$, orthonormal in the inner product that corresponds to the l_2 norm:

$$\begin{aligned} (\mathbf{a}, \mathbf{b}) &:= h \sum_{j=1}^{N} a_j b_j^* \\ \text{so } (\mathbf{a}, \mathbf{a}) &= \|\mathbf{a}\|_2^2 \end{aligned}$$

Since the DF vectors are a basis, we can write any vector as a linear combination of these vectors, that is

$$\mathbf{a} = \hat{a}_0\mathbf{f}_0 + \hat{a}_1\mathbf{f}_1 + \ldots + \hat{a}_{N-1}\mathbf{f}_{N-1}$$

for ! coefficients $\hat{\mathbf{a}}$, the scaled DF transform of $\mathbf{a}$. We are pursuing some theoretical properties here, but there are practical situations where $\hat{\mathbf{a}}$ is desired and it can be computed efficiently using the Fast Fourier Transform algorithm. Since $\{\mathbf{f}_\alpha\}$ is an orthonormal basis,

$$\|\mathbf{a}\|_2 = \|\hat{\mathbf{a}}\|_2. \tag{11}$$

3.8.2 Application to discretizations

It can be shown that the DF vectors are always the complete set of eigenvectors of any linear, constant coefficient, periodic, finite difference discretization on a uniform grid. As an example, this will be shown explicitly for our discretization of Problem 1.

$$\begin{aligned} D_2 f_{\alpha,j} &= \frac{1}{h^2}(f_{\alpha,j-1} - 2f_{\alpha,j} + f_{\alpha,j+1}) \\ &= \frac{1}{h^2}e^{2\pi i\alpha jh}\left(e^{-2\pi i\alpha h} - 2 + e^{2\pi i\alpha h}\right) \\ &= \frac{2}{h^2}(\cos(2\pi\alpha h) - 1)f_{\alpha,j}. \end{aligned}$$

Thus

$$D_2\mathbf{f}_\alpha = \frac{2}{h^2}(\cos(2\pi\alpha h) - 1)\mathbf{f}_\alpha$$

and

$$\mathcal{A}\mathbf{f}_\alpha = \lambda_\alpha\mathbf{f}_\alpha$$

with $\lambda_\alpha = \frac{2}{h^2}(1 - \cos(2\pi\alpha h)) + a$. The set of the eigenvalues $\{\lambda_\alpha\}$ of $\mathcal{A}$ corresponding to the DF vectors is called the *symbol* of $\mathcal{A}$. Consider again our discretization:

$$\mathcal{A}\mathbf{U} = \mathbf{F}.$$

We can write it in terms of the the DF components

$$\hat{U}_\alpha = \frac{1}{\lambda_\alpha}\hat{F}_\alpha$$

where we have gained considerable insight from diagonalizing the problem. The eigenvalues λ_α are all positive and $\geq a$ so

$$\begin{aligned} & |\hat{U}_\alpha| \leq \frac{1}{a}|\hat{F}_\alpha| \quad \text{for every } \alpha \\ \Rightarrow \quad & \|\hat{\mathbf{U}}\|_2 \leq \frac{1}{a}\|\hat{\mathbf{F}}\|_2 \\ \Rightarrow \quad & \|\mathbf{U}\|_{\mathbf{2}} \leq \frac{\mathbf{1}}{\mathbf{a}}\|\mathbf{F}\|_{\mathbf{2}} \end{aligned}$$

using (11). This shows the l_2 norm stability of the scheme. Applying the same result to $\mathcal{A}\mathbf{E} = \tau$ shows the l_2 convergence of the scheme,

$$\|\mathbf{E}\|_2 \leq Ch^2$$

using the bounds on the truncation error τ from the previous section. Note that a (sub-optimal) maximum norm convergence result can be derived from this, since

$$\begin{aligned} & h|E_i|^2 \leq h\sum_{j=1}^{N}|E_j|^2 := \|\mathbf{E}\|_2^2 \leq (Ch^2)^2 \text{ for each } i \\ \Rightarrow \quad & |E_i|^2 \leq C^2h^3 \text{ for each } i \\ \Rightarrow \quad & \|\mathbf{E}\|_\infty \leq Ch^{3/2} \end{aligned}$$

So from von Neumann analysis we can show that scheme does converge in maximum norm, but at the non-optimal rate of 3/2. We know from our numerical test that second order accuracy is maximum norm is observed and confirmed that in our first, maximum norm, stability analysis.

Remark: Sometimes there is a gap between what can be proved and the actual behaviour of the method.

3.9 Implementing Boundary Conditions

Consider Problem 1 in the interval [0,1], but with local boundary conditions specified at the ends $x = 0$ and $x = 1$ rather than periodic conditions. Possible conditions for this problem at $x = 0$ are

$$u(0) = a, \quad a \text{ given (Dirichlet)} \tag{12}$$

$$u'(0) = a \quad \text{(Neumann)} \tag{13}$$

$$u'(0) - \alpha u(0) = a, \quad \alpha > 0 \text{ given (Robin).} \tag{14}$$

Similar conditions can be given at $x = 1$. For (14) at $x = 1$, $\alpha < 0$ for physically stable models. In this setting with a uniform grid with spacing $h = 1/N$ we will in general have $N + 1$ discrete unknowns $U_0, U_1, \ldots U_N$ at $x = 0, h, \ldots 1$. If we have Dirichlet conditions at both interval ends, we can apply $U_0 = a$ directly, and the same for U_N. The value of U_0 only appears in the stencil at grid point 1:

$$D_2U_1 = \frac{U_0 - 2U_1 + U_2}{h^2} = \frac{-2U_1 + U_2}{h^2} + \frac{a}{h^2}.$$

In the implementation, the first two terms of the right expression above become part of the matrix $\mathcal{A}$ and the last term contributes to the right hand side vector. In this case, the end point values have been eliminated and the system to be solved is of size $N - 1$.

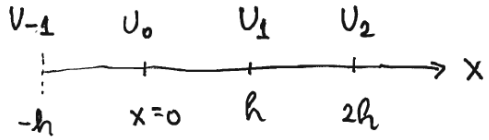


Figure 7: Grid points near the $x = 0$ boundary.

3.9.1 First derivative approximations

To proceed to the other conditions (Neumann and Robin) we need to discuss finite difference approximations of the first derivative.

$$
\begin{aligned}
D_+U_j &:= \frac{U_{j+1} - U_j}{h} = u'(jh) + \frac{h}{2}u''(jh) + \dots \text{ (forward differencing, first order)} \\
D_-U_j &:= \frac{U_j - U_{j-1}}{h} = u'(jh) - \frac{h}{2}u''(jh) + \dots \text{ (backward differencing, first order)} \\
D_1U_j &:= \frac{U_{j+1} - U_{j-1}}{2h} = u'(jh) + \frac{h^2}{6}u'''(jh) + \dots \text{ (centred differencing, second order)} \\
\tilde{D}_+U_j &:= \frac{-\frac{3}{2}U_j + 2U_{j+1} - \frac{1}{2}U_{j+2}}{h} = u'(jh) + \frac{h^2}{3}u'''(jh) + \dots \text{ (second order forward differencing)}
\end{aligned}
$$

Notice that although both D_1 and $\tilde{D}_+$ are second order accurate, $\tilde{D}_+$ has a larger error constant. Note also that $D_1 = \frac{1}{2}(D_+ + D_-)$ and that this combination cancels the first order error terms.

3.9.2 Implementing Neumann conditions

Consider now implementing the Neumann condition (13). There are several approaches. Since implementing boundary conditions is often a source of confusion, I will go through some of the options in detail. In what follows, refer to Figure 7 for the numbering of the unknowns.

U_0 **equation for Neumann condition:** In this scenario, U_0 remains an unknown and we use an approximation of the Neumann condition for its corresponding discrete equation. Using first order differencing

$$D_+U_0 := \frac{U_1 - U_0}{h} = a$$

leads to approximate values that are only first order accurate at *all* grid points. We can easily maintain global second order accuracy by using

$$\tilde{D}_+U_0 = a \tag{15}$$

instead.

U_0 eliminated using one sided differencing: We notice again that U_0 only appears in the U_1 equation and so we can use (15) to eliminate it:

$$\begin{aligned} D_2U_1 &:= \frac{U_0 - 2U_1 + U_2}{h^2} \\ &= \frac{\frac{2}{3}(2U_1 - \frac{1}{2}U_2 - ha) - 2U_1 + U_2}{h^2} \\ &= \frac{2(U_2 - U_1)}{3h^2} - \frac{2a}{3h} \end{aligned} \tag{16}$$

However, (16) can cause some confusion since the truncation error in (15) is second order but (16) is only first order accurate. Analytically, (16) is the right way to view the system since the corresponding $(N+1)\times(N+1)$ matrices are max-norm stable.

Introduce the ghost point U_{-1}: You can also introduce the ghost point U_{-1} as shown in Figure 7. Consider U_{-1} to approximate the solution extended from the interior to $x = -h$ using a Taylor polynomial of high enough order. The Neumann condition $u'(0) = a$ can then be approximated by

$$D_1U := \frac{U_1 - U_{-1}}{2h} = a.$$

This equation and U_{-1} can be added to the system or U_{-1} can be eliminated using this condition as done above. This approximation has a smaller truncation error constant than the second order one-sided approach and so is preferred.

Robin conditions can be implemented in a similar manner.

3.9.3 Implementing boundary conditions for staggered grid discretizations

We can consider approximate values at subinterval centres rather than subinterval ends. For second order methods, this is equivalent to considering unknowns that are the integral average of the unknown function over the subinterval (the basis of finite volume methods). Values on this grid near the $x = 0$ boundary are shown in Figure 8. Here, a ghost value is needed even for a Dirichlet condition, which is then approximated to second order using linear interpolation (averaging):

$$\frac{U_{1/2} + U_{-1/2}}{2} = a \quad \text{approximates } u(0) = a$$

and short centred differencing is used for a Neumann condition

$$\frac{U_{1/2} - U_{-1/2}}{h} = a \quad \text{approximates } u'(0) = a.$$

Note that this approximation has a dominant truncation error term of $\frac{h^2}{24}u'''(0)$. This is the most accurate second order way to approximate Neumann conditions

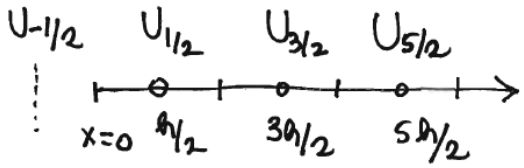


Figure 8: Staggered grid points near the $x = 0$ boundary.

in this finite difference framework. As discussed above, these equations can be incorporated into a matrix of the discretization $\mathcal{A}$, or the ghost value $U_{-1/2}$ can be eliminated.

3.10 Asymptotic Error Analysis

Our convergence proof for the discretization of the periodic problem showed that

$$\|\mathbf{E}\|_\infty \leq Ch^2$$

where $\mathbf{E} = \mathbf{U} - \mathbf{u}$. Computationally, we saw that in fact

$$E_i = c(ih)h^2 + o(h^2)$$

with an underlying smooth function $c(x)$ independent of h. Above, $o(h^2)$ (notice the lower case o) is a quantity smaller than any constant times h^2 as $h \to 0$. For smooth data f this result is not hard to show and the remainder term $o(h^2)$ is $O(h^4)$. Let us see what $c(x)$ would have to be to have the property

$$U_i = u(ih) + c(ih)h^2 + \ldots \tag{17}$$

This is known as an asymptotic error expansion. Plug this into the discrete equations $-D_2U_i + aU_i = F_i$ and use the remainder terms for the approximation for D_2 we derived in Section 3.5:

$$-u'' - \frac{h^2}{12}u'''' + O(h^4) - h^2c'' + au + ah^2c = f \tag{18}$$

where these terms are all evaluated at $x = ih$. If (18) is to hold for all h, then the coefficients of powers of h must match:

$$O(1): \qquad -u'' + au = f \quad \text{with } u \text{ periodic} \tag{19}$$

$$O(h^2): \qquad -c'' + cu = \frac{1}{12}u'''' \quad \text{with } c \text{ periodic.} \tag{20}$$

Equation (19) is satisfied by the exact solution. We don't have to actually solve (20) but we do know that this problem has a smooth solution $c(x)$. Note that

$c(x)h^2$ is the dominant error term (17) and from (20) we see that $c(x)$ is the response of the system to a RHS that is the truncation error. This makes sense.

All of this may seem like formal so far, but now consider

$$\tilde{\mathbf{E}} = \mathbf{U} - (\mathbf{u} + h^2\mathbf{c}).$$

We have

$$\mathcal{A}\tilde{\mathbf{E}} = O(h^4)$$

using (19) and (20). Using the stability result for $\mathcal{A}$ from Section 3.7 we have

$$\tilde{\mathbf{E}} = O(h^4).$$

This shows that (17) is accurate to fourth order, the desired result.

There are several consequences of the result

$$\mathbf{U} = \mathbf{u} + h^2\mathbf{c} + O(h^4)$$

with $c(x)$ smooth, independent of h:

Richardson Extrapolation: This justifies Richardson extrapolation

$$\tilde{\mathbf{U}}_h := \frac{4}{3}\mathbf{U}_{h/2} - \frac{1}{3}\mathbf{U}_h = \mathbf{u} + O(h^4)$$

where in the expression above $\mathbf{U}_h$ is a coarse grid computation and $\mathbf{U}_{h/2}$ are values from a fine grid (refined by a factor of 2) computation taken at coarse grid points. Note that this is not an efficient way to make a fourth order method. In addition, it is not reliable since not all schemes have regular errors like this one.

Discrete Smoothness: This result also shows that derivative approximations converge with full order. Consider our basic estimate

$$\mathbf{U} = \mathbf{u} + O(h^2).$$

If we were interested in values of the first derivative of the solution, we would compute

$$D_1\mathbf{U} = D_1\mathbf{u} + O(h) = \mathbf{u}' + O(h)$$

where the order of h is lost because we divide the $O(h^2)$ by h when we apply D_1 and we have not taken into account the structure of the $O(h^2)$ term. However, if we know the asymptotic error result applies

$$\mathbf{U} = \mathbf{u} + h^2\mathbf{c} + O(h^4)$$

with $c(x)$ smooth then we see that

$$D_1\mathbf{U} = D_1\mathbf{u} + h^2 D_1\mathbf{c} + O(h^3) = \mathbf{u}' + h^2(\mathbf{c}' + \frac{1}{6}\mathbf{u}''') + O(h^3) = \mathbf{u}' + O(h^2)$$

and we see convergence order preserved for derivatives. In Finite Element Method (FEM) literature, this "unexpected" increase in convergence order is sometimes called "superconvergence" (although this term also has other meanings).

Discrete Embedding: An asymptotic error expansion can also overcome deficiencies in the stability analysis using weak norms. Consider the conversion of the l_2 estimate to the maximum norm estimate considered in Section 3.8. Following that argument with

$$\tilde{\mathbf{E}} := \mathbf{U} - (\mathbf{u} + h^2\mathbf{c}) = O(h^4)$$

gives

$$\begin{aligned} \|\tilde{\mathbf{E}}\|_\infty &\leq Ch^{7/2} \\ \|\mathbf{E} - h^2\mathbf{c}\|_\infty &\leq Ch^{7/2} \\ \|\mathbf{E}\|_\infty &= O(h^2) \end{aligned}$$

where in the last step we have used the triangle inequality.

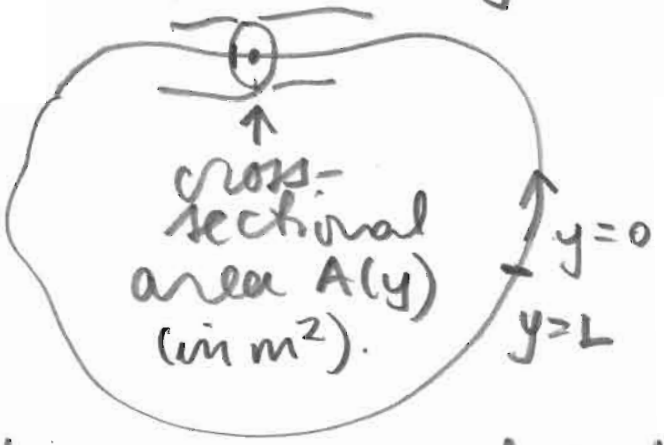
The existence of regular asymptotic error behaviour (and so all the results above) relies on the problem having smooth solutions and being computed on a regular (structured) grid. This shows one of the advantages of using structured meshes. In addition, discretizations on structured meshes are more efficiently implemented, especially on specific computational architectures like Graphical Processing Units (GPUs). However, structured meshes restrict local adaptivity and do not apply in a straightforward way to general problem geometries in higher dimensions.

Scaling & Nondimensionalization
Example.

Brian Wetton, wetton@math.ubc.ca

January 7, 2013.

Consider heat conduction in a metal rod of length L (in m) bent around with its two ends joined.



Assume that the aspect ratio of the rod is large enough that the temperature $V(y,t)$ (in °C) can be assumed to be constant in the cross-section.

Local heat balance gives the following

$$A\rho c\, V_t = (KAV_y)_y + f(y,t) - g(V,V_0,y) \qquad (1)$$

where:

ρ: density of the rod material kg/m³

c: heat capacity J/kg/°C

K: thermal conductivity J/s/m/°C

F: given applied heating per unit length
J/s/m.

g: heat loss per unit length to ambient
J/s/m. This has a functional form that must be fit to experiments. We allow a dependence on y since this could be affected by the cross-section shape.
clearly, $g(V_0, V_0, y) = 0$ for every y.

Note: Every termin (1) has the same units, J/s/m. Checking for unit consistency is a useful step in modelling.

Now let's make some assumptions: the external heating $f(y)$ does not depend on t and we consider the system after a long enough time that transients have died away, so $V(y)$. Also assume that A is constant and g does not depend on y. The final assumption at this stage is that a linear Taylor approx of $g(V, V_0) \approx M(V - V_0)$
is sufficient (M has units of $\frac{J}{sm^\circ C}$)

M is determined experimentally or through asymptotics of the surface heat transfer to the ambient medium.

Now (1) becomes

$$-KA\frac{d^2V}{dy^2} + M(V-V_0) = f(y). \qquad (2)$$

Scale spatial variables

$$y = Lx, \qquad x \in [0,1], \text{ dimensionless.}$$

and shift V, $w = V - V_0$, so (2) becomes.

$$-\frac{KA}{L^2}\frac{d^2W}{dx^2} + MW = f(xL). \qquad (3)$$

Let F (in J/s/m) be a representative size of f, so that

$$\frac{f(xL)}{F} := \hat{f}(x)$$

is dimensionless and has unit size, $O(1)$.

Scale w,

$$w = Vu, \quad u \text{ dimensionless}$$

with V in °C to be determined. (3) becomes

$$-\frac{KAV}{L^2}u'' + MVu = F\hat{f}(x) \Rightarrow$$

$$-u'' + \frac{ML^2}{KA} u = \frac{FL^2}{KAV} \hat{f}(x)$$

So if we choose the temperature scale to be

$$V = \frac{FL^2}{KA}$$

dimensionless parameter.

and let $a = ML^2/KA$, we arrive at our scaled problem

$$-u'' + au = \hat{f}(x).$$

in which all quantities are dimensionless.

Math 521, Spring 2013
Notes, Part III

Let's get ready to describe the Finite Element Method. First, we have to describe weak solutions and do a little bit of functional analysis.

Problem 1 Solve for $u(x)$ given $a>0$ and $f(x)$ that satisfies

$$-u'' + a u = f \qquad (1)$$

at all x, in the 1-periodic setting.

We know if $f \in C_2$ then $u \in C_4$. Multiply (1) by any C_∞ (periodic) function $\varphi(x)$ and integrate from 0 to 1.

$$-\int_0^1 u'' \varphi \, dx + a\int_0^1 u\varphi \, dx = \int_0^1 f \varphi \, dx \qquad (2)$$

↓ integrate by parts

$$\int_0^1 u'\varphi' \, dx - u'\varphi\Big|_0^1 + a\int_0^1 u\varphi \, dx = \int_0^1 f\varphi \, dx$$

0, periodicity

$$\int_0^1 u'\varphi' \, dx + a\int_0^1 u\varphi \, dx = \int_0^1 f\varphi \, dx \qquad (3)$$

If (3) is satisfied for all $\varphi \in C_\infty$ (periodic) then u is called a weak solution of (1).

Note that for (3) to make sense, f can be discontinuous (piecewise continuous functions are a class interesting to applications) and u does not have to be C_2.

We'll need to introduce some function spaces, and do a little functional analysis. Consider $f, g \in C_\infty$ (periodic), and introduce the inner product

$$(f,g) = \int_0^1 f(x)g(x)dx.$$

← if we consider complex valued functions: $\int_0^1 f(x)g^*(x)dx$

This has all the properties of an inner (or dot) product on the vector space C_∞.

$$(f,f) = \int_0^1 (f(x))^2 dx := \|f\|^2$$

has the properties of a norm, including

$$|(f,g)| \leq \|f\|\,\|g\|. \quad \text{Cauchy-Schwarz inequality}$$

$$\|f+g\| \leq \|f\| + \|g\|. \quad \text{triangle inequality.}$$

Now, the space C_∞ with the norm

$$\|f\| = \sqrt{\int_0^1 |f(x)|^2 dx}$$

is not a great fit. It is like doing arithmetic with only the rational numbers.

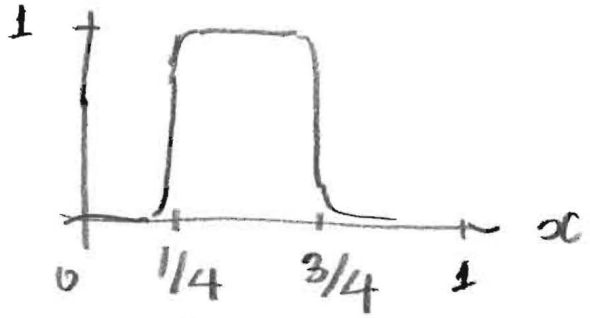
We can have sequences of rational numbers $\{x_n\}$ that should converge in the sense that they are Cauchy sequences

$$\lim_{n,m\to\infty} |x_n - x_n| = 0$$

but they approach an irrational number in the limit, which is not in the original set of numbers. One way to think about the set of real numbers is to start with the rationals and add all possible limits.

In the same way sequences of C^∞ functions $\{f_n\}$ can have limits in $\|\cdot\|$ that are not C^∞ functions.

Consider $f_n(x)$ of the following form:



width of transition layers is $\frac{1}{n}$.

Then $|f_n(x) - f_m(x)| < 1$ on two intervals of length $\max\left(\frac{1}{n}, \frac{1}{m}\right)$ so

$$\|f_n - f_m\|^2 < 2 \max\left(\frac{1}{n}, \frac{1}{m}\right)$$

As $\lim_{n,m\to\infty} \|f_n - f_m\| = 0$

but $\lim_{n\to\infty} f_n(x)$ has discontinuities, is not a C_∞ function.

The function space L_2 is C_∞ functions and all possible limits (a technical idea called metric completion). This definition is equivalent to the set of Lebesgue integrable functions with finite values of $\int_0^1 |f|^2 dx$

(see also pages 4B & 4C)

Piecewise smooth functions are in L_2. Also some functions with (mild) singularities.

$$f(x) = \frac{1}{\sqrt[3]{x}}$$

$$[f(x)]^2 = x^{-2/3}$$

So $\int_0^1 [f(x)]^2 = 3x^{1/3}\Big|_0^1 = 3.$

Similarly $f(x) = \frac{1}{x^p}$ is in L_2 for any $p < 1/2$.

Now, we'll have other function spaces, so we have to label the norms and inner

→ p.5A

There is also a connection to Fourier Series.

$f(x)$ 1-periodic

$$\Rightarrow \quad \hat{f}_n = \int_0^1 f(x) e^{-2\pi i n x} dx$$

$$\Rightarrow \quad f(x) = \sum_{n=-\infty}^{\infty} \hat{f}_n e^{2\pi i n x} \qquad (4)$$

↑

Complex form of the Fourier Series.

If $f(x)$ were continuous, the convergence would be pointwise.

Note that (4) can be written as

$$\hat{f}_n = (f, \varphi_n)_{L_2}$$

with $\varphi_n(x) = \frac{1}{\sqrt{2\pi}} e^{2\pi i n x}$ and $\{\varphi_n\}$ are an orthonormal set, since

$$\int_0^1 \varphi_n(x) \varphi_m^*(x) dx$$

$$= \int_0^1 e^{2\pi i (n-m) x} dx$$

1 periodic.

$$= \begin{cases} 1 & \text{if } n = m \\ \dfrac{1}{2\pi i (n-m)} e^{2\pi i (n-m) x} \Big|_0^1 = 0 \end{cases}$$

The result (4) shows that $\{\psi_n\}$ are a basis for a class of functions (not trivial to show in this infinite dimension setting). It shows that the dimension of L_2 is countably infinite.
With a little more work it also follows that

$$\int_0^1 |F(x)|^2 dx = \sum_{n=-\infty}^{\infty} |\hat{f}_n|^2$$

So we can think of L_2 as coming from sequences $\{\hat{f}_n\}$ that are square summable.

However, in this case the expression in (4) does not converge at all points (but at "almost all" points).

Note that if we consider

$$S_N(x) = \sum_{n=-N}^{N} \hat{f}_n e^{2\pi i n x} \leftarrow \text{a } C^\infty \text{ function}$$

Then $\|f - S_N\|_{L_2}^2 = \sum_{|n|>N} |\hat{f}_n|^2 \to 0$ as $N \to \infty$

so this is a way to generate the sequence of C^∞ functions that have $f(x) \in L_2$ as a limit.

products when there is any possible confusion.

$$\|f\|_{L_2}, \quad (f,g)_{L_2}$$

or just $\|f\|_2$, $(f,g)_2$.

The next space we'll look at is H^1 with norm

$$\|f\|_{H^1}^2 = \int_0^1 \{|f|^2 + |f'|^2\}\, dx \qquad (f,g)_{H_1} = (f,g)_{L_2} + (f',g')_{L_2}$$

$$= \|f\|_{L_2}^2 + \|f'\|_{L_2}^2.$$

Again, functions in H^1 are C^∞ functions and their limits in this norm. Note that neither of the previous examples are in this H^1 space (see page 5B)

We can similarly define H^n with

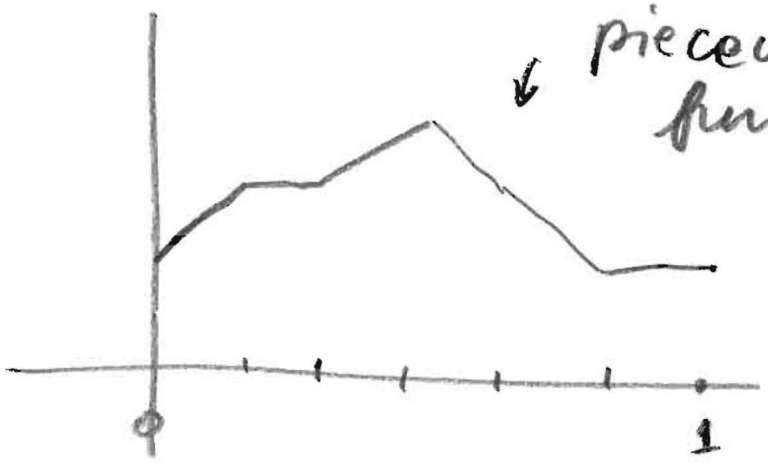
$$\|f\|_{H^n}^2 = \sum_{j=0}^{n} \|f^{(j)}\|_{L_2}^2, \quad \leftarrow \text{ so } H^0 = L_2$$

so square integrable derivatives up to order n. Some analysis naturally involves fractional and negative n...

The final space we'll consider is related to the weak formulation of problem 1. We'll call the space A with norm page 6 →

$$\|f\|_A^2 = \int_0^1 |f'|^2 dx + a\int_0^1 |f|^2 dx = \|f'\|_{L^2}^2 + a\|f\|$$

Functions that are in H_1 are functions that are continuous but also piecewise C_1. In particular, continuous and piecewise linear functions are in H_1. These will be the functions used in our first FEM approximations.



In terms of Fourier series,

$$\|f\|_{H^1}^2 = \sum_{n=-\infty}^{\infty} (1 + 4\pi^2 n^2) |\hat{f}_n|^2$$

so for $\|f\|_{H^1}$ to be finite, $\hat{f}_n$ must go to zero as $n \to \infty$ relatively fast. The decay rate of $\hat{f}_n$ as $n \to \infty$ is related to the smoothness [number of derivatives] of f.

$$(f,g)_A = \int_0^1 f'g'\,dx + a\int_0^1 fg\,dx.$$

Note that the space A and the space H^1 are identical. The norms are different but equivalent,

$$\|f\|_A^2 < \max(1,a)\,\|f\|_{H^1}^2$$

and $$\|f\|_{H^1}^2 < \max\left(1,\frac{1}{a}\right)\|f\|_A$$

So convergence in A is equivalent to convergence in H^1. So we can rewrite (3) as

$$(u,\varphi)_A = (f,\varphi)_{L_2} \quad \text{for all } \varphi \in C_0^\infty.$$

We can take limits of φ in the A norm (which will also converge in L_2 since $\|f\|_{L_2} < \|f\|_{H^1} < \sqrt{\max\left(1,\frac{1}{a}\right)}\,\|f\|_A$) and see this must be true for all $\varphi \in H_1$.

$$(u,\varphi)_A = (f,\varphi)_{L_2} \quad \text{for all } \varphi \in H_1 = A. \quad (5)$$

So our problem is reduced to representing an inner product in one space over all functions by the inner product in another space.

We'll turn to the existence proof of this weak formulation next, as a warm up exercise for the FEM analysis.

Definitions

(i) A Hilbert space is a vector space with an inner product that is complete in the norm from this inner product (complete means all Cauchy sequences tend to elements of the space).

L_2, H^1, and A are Hilbert spaces.

(ii) A Hilbert space is said to be separable if it has a dense, countable subset. This is equivalent to having a basis with a countable number of elements.

L_2, H^1 and A are separable.

(iii) A linear functional on a Hilbert space X is a map $L: X \to \mathbb{R}$ such that L is linear and bounded.

[linear] $L(c_1 g_1 + c_2 g_2) = c_1 L(g_1) + c_2 L(g_2)$
for all c_1 & $c_2 \in \mathbb{R}$ and all g_1 & $g_2 \in X$.

[bounded] $|L(g)| \leq C \|g\|$
for all $g \in X$ for some fixed C.

Consider a fixed $f \in X$, then

$$L_f(g) = (f,g)_X \qquad (6)$$

is an example of a linear functional. It is linear since the inner product is linear in each of its arguments separately (bilinear). It is bounded since (Cauchy - Schwartz)

$$|L_f(g)| = |(f,g)| \leq \|f\| \, \|g\|$$

so we can use $\|f\|$ as the C in the definition on the last page. It is shown in the theorem below that all linear functionals on X can be written in the form (6)

Theorem (Riesz Representation) If X is a separable Hilbert Space and L is a linear functional, then there exists a ! f so that

$$L(g) = (f,g) \quad \text{for all } g \in X.$$

[In words, L is equivalent to (can be represented by) the inner product with a certain f].

Note: The theorem is true even in the case where X is not separable, but the following proof relies on a countably infinite o.n. basis $\{\varphi_n\}$ of X.

Proof: Consider

$$c_i = L(\varphi_i) \qquad [c_i \text{ real numbers}]$$

and $$f_N = \sum_{i=1}^{N} c_i \varphi_i \qquad [f_N \in X] \qquad (7)$$

In finite (N) dimensional spaces X, the story is done, f_N is the desired element of X, because we can write any $g \in X$ in this basis,

$$g = \sum_{i=1}^{N} (g, \varphi_i)\, \varphi_i$$

and then $$L(\varphi) = \sum_{i=1}^{N} (g, \varphi_i)\, c_i \qquad (8)$$

but $$(f_N, g) = \left(\sum_{i=1}^{N} c_i \varphi_i,\ \sum_{i=1}^{N} (g, \varphi_i)\, \varphi_i \right)$$

$$= \sum_{i=1}^{N} c_i\, (g, \varphi_i) \quad \text{same form as (8)} \ \checkmark$$

using the ortho-normality of $\{\varphi_n\}$. This is just a case of the general result that all linear transformations from $\mathbb{R}^N$ to $\mathbb{R}^m$ can be represented by matrix multiplication.

In the infinite dimensional setting, the argument is the same once we show that

the sum (4) converges in X. First notice that

$$L(f_N) = \sum_{i=1}^{N} c_i^2 = \|f_N\|^2 = \|f_N\| \cdot \|f_N\|$$

since L is bounded with constant K, we see that $\|f_N\|$ must be bounded by K, independent of K, so $\lim_{N\to\infty} \sum_{i=1}^{N} c_i^2$ exists. Therefore,

$$\lim_{n,m\to\infty} \sum_{i=n}^{m} c_i^2 \to 0$$

and so $\|f_n - f_m\|^2 = \sum_{i=n}^{m} c_i^2 \to 0$.

Therefore, the sum (7) converges in L_2. Now, the proof proceeds like the finite dimensional case.

see page 11 for uniqueness

Application to (3), written in form (5).

$$(u, \varphi)_A = (f, \varphi)_{L_2} \quad \text{for all } \varphi \in H_1.$$

Note that the RHS is a linear functional on A, since

$$|(f, \varphi)| \overset{\text{Cauchy Schwarz}}{\leq} \|f\|_{L_2} \|\varphi\|_{L_2} \leq \max\left(1, \tfrac{1}{a}\right) \|f\|_{L_2} \|\varphi\|_A$$

so using the representation theorem,

$(f, \varphi)_{L_2} = (u, \varphi)_A$ for some $u \in A$.

That u is the weak solution to problem 1!
That solution u is unique by the representation theorem.

---

uniqueness in the Riesz Representation Theorem.

Suppose $L(g) = (f_1, g) = (f_2, g)$ for all g.

Then $(f_1 - f_2, g) = 0$ for all g.

Take $g = f_1 - f_2 \Rightarrow$

$$(f_1 - f_2, f_1 - f_2) = \|f_1 - f_2\|^2 = 0$$

so $f_1 = f_2$, uniqueness ✓

One last topic in these notes that are an introduction to functional analysis. We know that L_2 functions are not necessarily continuous, and H^1 functions do not necessarily have continuous derivatives. However, it can be shown that H^1 functions (in 1D) are continuous. This is known as a Sobolev space embedding result. Sobolev spaces are Hilbert Spaces (and generalizations

of functions and their derivatives. I'll prove, part of this embedding result below.

Consider $u \in C^\infty$ [periodic]. Note that if we can prove the result using only qualities of u that use the H^1 norm, we can pass to the limit to any $u \in H^1$.

$$u_{ave} = \int_0^1 u(x)\,dx = (u, 1)_{L_2} \leq \|u\|_{L_2}$$

↑ the function with values 1 at every x

By the intermediate value theorem, u_{ave} is attained by u at some point x_*

$$u(x_*) = u_{ave}.$$

Now consider $u(x)$ at all other points x, using the following formula.

$$u(x) = u_{ave} + \int_{x_*}^{x} u'(s)\,ds.$$

$$|u(x)| \leq |u_{ave}| + \int_0^1 |u'(x)|\,dx$$

$$\leq \|u\|_{L_2} + (|u'(x)|, 1)_{L_2}$$

$$\leq \|u\|_{L_2} + \|u'\|_{L_2} = \|u\|_{H^1}$$

This shows that $\|u\|_\infty \leq \|u\|_{H^1}$.
Showing continuity takes a bit more technical effort.

Math 521, Notes IV
Spring, 2013.

Begin with the weak form of our usual periodic problem:

$$-u'' + au = f \qquad \text{(strong)}$$

$$(u', \varphi') + a(u, \varphi) = (f, \varphi) \qquad \text{(weak)}$$
$$\text{for all } \varphi \in H^1$$

where the inner products are L_2 unless specified. The weak form can be rewritten as

$$(u, \varphi)_A = (f, \varphi) \qquad \forall\, \varphi \in H^1 \qquad (1)$$

where $(u, \varphi)_A := \int_0^1 u' \varphi' + a\int_0^1 u\varphi$. The space A is the same as H^1_0, although $\|u\|_A \neq \|u\|_{H^1}$ in general unless $a = 1$. However, the norms are equivalent.

Consider an N-dimensional subspace S of H^1. We can make a numerical method by finding $U \in S$ such that

$$(U, \varphi)_A = (f, \varphi) \qquad \forall\, \varphi \in S \qquad (2)$$

This is known as a Galerkin method, characterized by the space that the numerical solution is in is the same as the space of test functions φ.

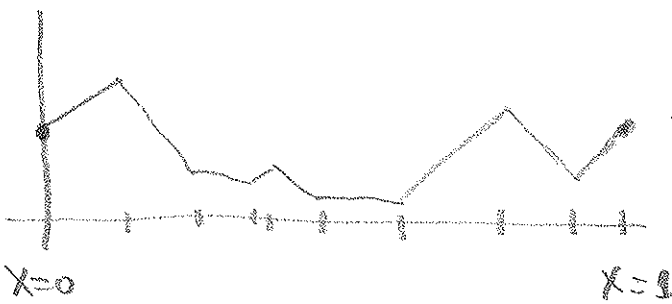
Defn If $S \subset H^1$ we say that this is a conforming method.

We could take a basis $\{\varphi_i\}_{i=1}^N$ of S and so write

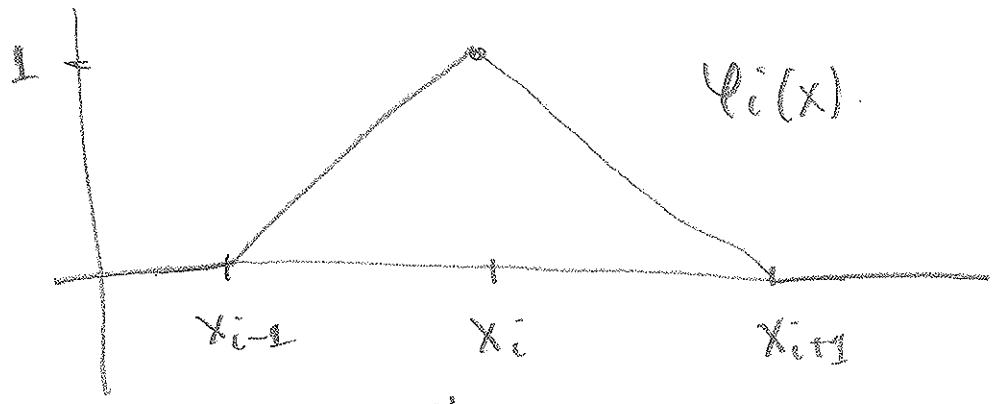
$$U = \sum_{i=1}^{N} a_i \varphi_i(x) \qquad (3)$$

with a_i to be determined. It would then be sufficient to satisfy (2) for $\varphi \in \{\varphi_i\}_{i=1}^N$. This generates N linear equations for the N a_i values. We will show below that this system is always solvable.

Example A simple choice for S is piecewise linear, continuous functions on subintervals between grid points $\{x_i\}_{i=1}^N$. We're still considering the periodic case.



A basis for this space consists of the functions $\varphi_i(x)$ that take value 1 at x_i and zero at all other grid points.



Now $U(x) = \sum_{i=1}^{N} a_i \varphi_i(x)$. In this case, a_i are the values of $U(x_i)$ so it makes sense to label these coefficients as U_i. It is not always the case that a_i correspond to function values for all finite element spaces S.

Returning to the general case (3), we insert this form into (2) and take the φ_i as test functions, leading to the following linear system for $\underline{a}$:

$$A\,\underline{a} = \underline{F}$$

where $A = K + aM$, $F_i = \int_0^1 \varphi_i(x) f(x)\,dx$,

$K_{ij} = \int_0^1 \varphi_i' \varphi_j'\,dx$ Stiffness matrix

$M_{ij} = \int_0^1 \varphi_i \varphi_j\,dx$ mass matrix.

Note that F_i will have to be evaluated numerically. This is discussed in section 6 of the notes.

Example Consider the piecewise linear elements of the previous example on a uniform grid with spacing h.

$$K_{ij} = \begin{cases} 2\int_0^h \left(\frac{1}{h}\right)^2 dx = \frac{2}{h} & \text{if } j=i \\ -\int_0^h \left(\frac{1}{h}\right)^2 dx = -\frac{1}{h} & \text{if } j=i-1 \text{ or } j=i+1 \\ 0 \quad \text{otherwise} \end{cases}$$

$$M_{ij} = \begin{cases} 2\int_0^h \left(1-\frac{x}{h}\right)^2 dx = \frac{2h}{3} & \text{if } j=i \\ \int_0^h \frac{x}{h}\left(1-\frac{x}{h}\right) dx = \frac{h}{6} & \text{if } j=i-1 \text{ or } j=i+1 \\ 0 \quad \text{otherwise} \end{cases}$$

$$F_i = \int_{x_i-h}^{x_i} \left(\frac{x-x_i-h}{h}\right) f(x)\,dx + \int_{x_i}^{x_i+h} \left(1-\frac{(x-x_i)}{h}\right) f(x)\,dx$$

$$= h\,f(x_i) + O(h^3).$$

can be shown by Taylor series.

So in this simple case, the resulting method looks like a finite difference method of second order.

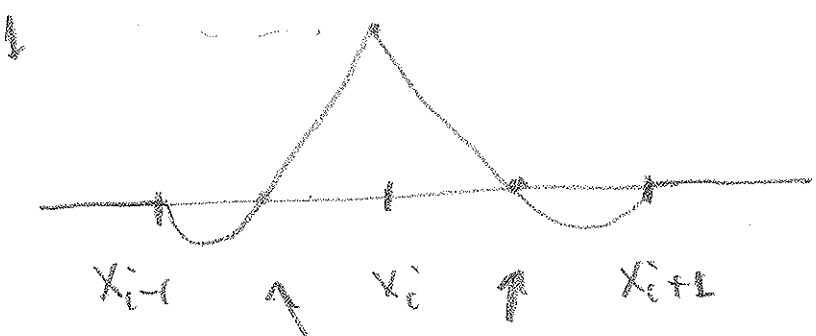
$$-\frac{1}{h}U_{i-1} + \frac{2}{h}U_i - \frac{1}{h}U_{i+1}$$

$$+ a\left(\frac{h}{6}U_{i-1} + \frac{2h}{3}U_i + \frac{h}{6}\right) \approx h\,F(x_i).$$

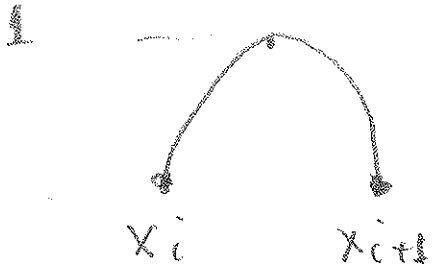
$$= h\,u(x_i) + O(h^3)$$

Other elements:

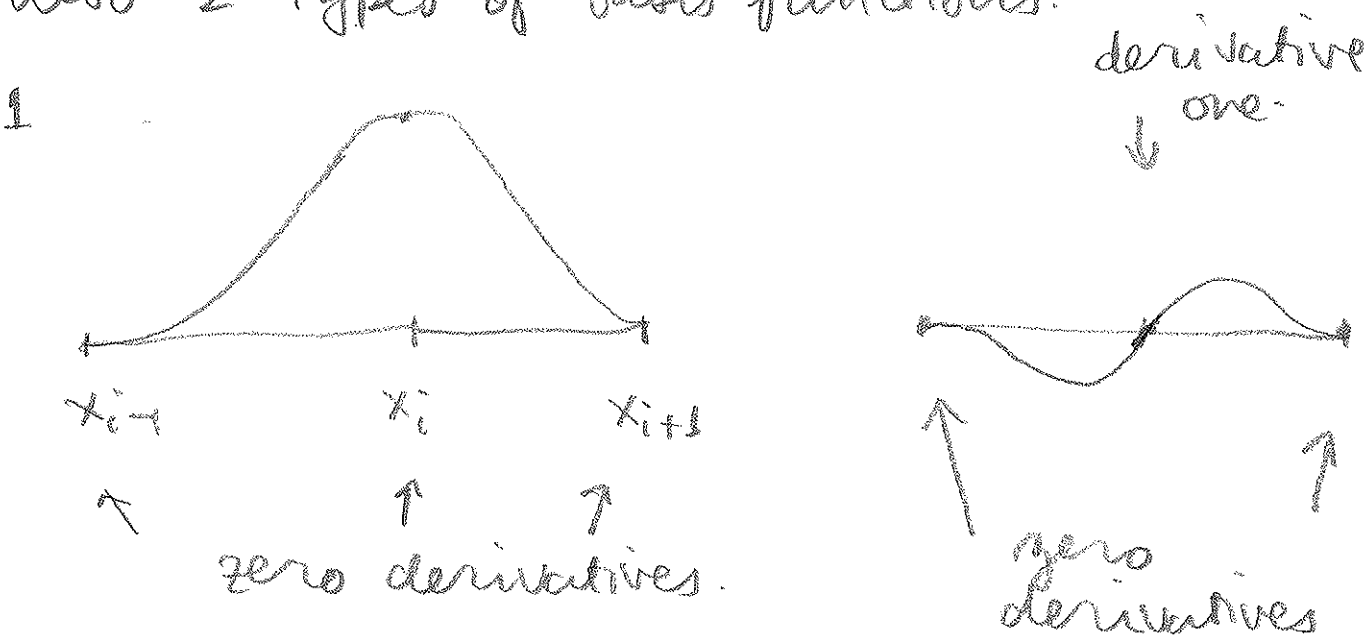
(i) Piecewise quadratic, continuous on N subintervals. This is a 2N dimensional subspace with additional degrees of freedom for each subinterval. Basis functions come in 2 types:



zero at centres of adjacent subintervals.



(ii) Piecewise cubic, C_1 on N subintervals.
This is also a 2N dimensional subspace.
Also 2 types of basis functions.



The unknown coefficients of the solution in this basis are the solution values and derivative values at subinterval end points.

Let's consider now the error analysis of the method. We'll have some general remarks on the behaviour for general S, then turn to the case of piecewise linear functions.

$$(U, \varphi)_A = (F, \varphi) \qquad \forall \varphi \in S \qquad (4)$$

defines the method.

$$(u, \varphi)_A = (F, \varphi) \qquad \forall \varphi \in H^1 \qquad (5)$$

defines the exact solution.

General theory for this problem shows that

$$\|u\|_{H^2} \leq C \|F\|_{L_2}.$$

Note: In what follows, we will use C for absolute constants and K for constants that depend on $\|F\|_{L_2}$.

Let $E(x) = U(x) - u(x)$, subtract (5) from (4) to obtain.

$$(E, \varphi)_A = 0 \qquad \forall\ \varphi \in S \qquad (6)$$

where we have used the fact that S is conforming ($S \subset H^1$).

Consider the size of $V(x) - u(x)$ for any element $V \in S$ in the A-norm.

$$\|V - u\|_A^2 = \|V - U + \underbrace{U - u}_{E}\|_A^2$$

$$= (V - U + E,\ V - U + E)_A \qquad \text{but } (E, V-U) = 0 \text{ by (6)}$$

$$= \|E\|_A^2 + \|V - U\|_A^2.$$

Thus

$$\|E\|_A \leq \|u - V\|_A \qquad \text{for all } V \in S. \quad (7)$$

In words, this states that the computed U is the closest element in S to the exact solution u.

In what follows, we restrict the discussion to piecewise linear approximation.
To proceed, let $V(x) \in S$ be the linear interpolation of the exact solution $u(x)$, so

$$W = u - V$$

satisfies $W(x_j) = 0$ for all j. Note, to define a linear interpolation, the solution $u(x)$ must have pointwise values. In 1D, this is true for $u \in H^1$ as discussed at the end of the last set of notes.

Consider $W(x)$ on each subinterval $[x_j, x_{j+1}]$. $W(x_j) = W(x_{j+1}) = 0$ so by Rolle, $W'(\theta) = 0$ for $\theta \in (x_j, x_{j+1})$.

$V'' = 0$ since it is linear on subintervals

$$W'(x) = \int_\theta^x W''(s)\,ds = \int_\theta^x u''(s)\,ds$$

$$|W'(x)| \le \int_{x_j}^{x_{j+1}} |u''(s)|\,ds.$$

($x_j \uparrow$)

imagine a 1 here and use C-S.

$$|W'(x)|^2 \le h_j \int_{x_j}^{x_{j+1}} |u''(s)|^2\,ds.$$

$$\int_{x_j}^{x_{j+1}} |W'(x)|^2 \le h_j^2 \int_{x_j}^{x_{j+1}} |u''(s)|^2\,ds. \qquad (8)$$

summing over subintervals,

$$\int_0^1 |W'(x)|^2 \le \sum_{j=1}^{N} h_j^2 \int_{x_j}^{x_{j+1}} |u''(s)|^2\,ds. \qquad (9)$$

letting $h = \max_j h_j$ we have

$$\int_0^1 |w'(x)|^2 \leq h^2 \int_0^1 |u''|^2 \leq K h^2. \qquad (10)$$

Considering (9) we see that we can make the RHS smaller by taking h_j smaller where u'' is large. This is the idea behind adaptive gridding.

In addition, using similar arguments.

$$w(x) = \int_{x_j}^{x} w'(s)\, ds \qquad x \in [x_j, x_{j+1}]$$

$$|w(x)|^2 \leq h_j \int_{x_j}^{x_{j+1}} |w'(x)|^2 dx \quad \text{using (8).}$$

$$\int_{x_j}^{x_{j+1}} |w(x)|^2 \leq h_j^4 \int_{x_j}^{x_{j+1}} |u''(s)|\, ds.$$

and then $\int_0^1 |w|^2 \leq K h^4$.

Combining with (10) we have that

$$\|w\|_A^2 = \|w'\|_{L^2}^2 + a \|w\|_{L^2}^2$$

$$\leq K (h^2 + a h^4) \quad \text{different } K \text{ restrict } h \leq 1.$$

$$\leq K h^2$$

Thus

$$\|W\|_A \leq Kh.$$

Using (7) we see that

$$\|E\|_A \leq \|W\|_A \leq Kh.$$

Thus, we see convergence of first order in the A-norm.

Note: This is all we can expect, since $V'(x)$ is constant on subintervals we will always have an $O(h)$ error in this norm. However, we can show second order convergence in L_2 norm, as seen below.

Let v solve

$$-v'' + av = E$$

(Don't need to actually solve for v, it's just a theoretical tool).

Theory $\Rightarrow$

$$\|v\|_{H^2} \leq C\|E\|_{L_2}$$

$$(v, \psi)_A = (E, \psi)_{L_2}.$$

Consider $\psi = E$, recalling (6), so

$$\|E\|_{L_2}^2 = (E, E)_{L_2} = (E, v - \tilde{v})_A$$

$\tilde{v}$: interpolate v in S as before.

So $\|E\|_{L_2}^2 \leq \|E\|_A \cdot Ch\,\|v\|_{H_2} \leq Ch\|E\|_A\,\|E\|_{L_2}$

Dividing $\|E\|_{L_2}$ on both sides and using $\|E\|_A \leq Kh$ as derived previously, we have

$$\|E\|_{L_2} \leq Kh^2.$$

In summary, U converges to u in $H^1(A)$ with first order, in L_2 with second order when a piecewise linear approximation is used.

The Conjugate Gradient Method

Brian Wetton

November 13, 2010

Introduction to Numerical Linear Algebra

In these notes, we will consider the solution of the symmetric, positive definite linear systems arising from a discretization of an elliptic problem. We will consider this in the general form

$$\mathbf{A}_h U = F$$

where h is considered to be a grid spacing parameter in the discretization. Although this seems like a very different problem from what we considered earlier (analysing the error from the discretization) it is still in the area of Numerical Analysis. A good choice of solver is vital if we want to obtain numerical results in a reasonable time.

The two extremes in numerical solution techniques we've seen so far are fast transform methods (extremely fast but very specialized) and full, direct solution (extremely slow but very general). Let's consider now something in between (somewhat general, not too slow).

Conjugate Gradient Method

In our discretization and many others, it is possible to multiply a vector by $\mathbf{A}_h$ quickly. Let us run with this idea. With our $\mathbf{A}$ it would be fast to construct the vectors

$$F, \mathbf{A}F, \mathbf{A}^2F, \ldots \tag{1}$$

If we denote the space span $\{F, \mathbf{A}F, \mathbf{A}^{k-1}F\}$ by S_k we might want to choose an approximation $U_k \in S_k$ to U that minimizes the error $\|U - U_k\|$. There are three "tricks" that are involved in turning this idea in to a useful algorithm. These are described below, followed by an error analysis.

How it works (three tricks)

It turns out that the "right" norm to use to minimize the error is not the Euclidean norm. We introduce several inner products on the space R^M:

$$(U, V) \;=\; \sum_i^M U_i \cdot V_i \quad \text{euclidean}$$

$$\begin{aligned}(U,V)_A &= (U,\mathbf{A}V)\\ (U,V)_{A^{-1}} &= (U,\mathbf{A}^{-1}V)\\ (U,V)_{A^2} &= (\mathbf{A}U,\mathbf{A}V) \quad \text{residual.}\end{aligned}$$

These norms all make sense for matrices $\mathbf{A}$ that are symmetric positive definite. The CG method finds the best approximation U_k (in the $\mathbf{A}$-norm) for U in the subspace S_k at each step k.

To make the minimization easy, we will construct an $(\cdot,\cdot)_A$-orthogonal ($\perp_A$) sequence $\{d_i\}$ such that $\{d_1,\dots d_k\}$ spans S_k (this could be done **i.e.** by applying the Gramm Schmidt process to (1) but can be much more efficiently as shown below). We write our approximation U_k as

$$U_k = \sum_{i=1}^{k} \alpha_i d_i$$

In order to minimize the error (in $(\cdot,\cdot)_A$ remember) between U and $U_k \in S_k$ for each k we must have

$$(U,\mathbf{A}d_j) = (U_k,\mathbf{A}d_j) \text{ for } \; j = 1,\dots,k$$

or

$$\alpha_j = \frac{(U,\mathbf{A}d_j)}{(d_j,\mathbf{A}d_j)} \text{ for } \; j = 1,\dots,k$$

using the fact that $\{d_i\}$ is $\mathbf{A}$-orthogonal. I think this is clear in itself, but it is a consequence of the following result: if U is given and $\mathcal{L}$ is a linear subspace of R^M and $V \in \mathcal{L}$ minimizes

$$\|U - V\|_?, V \in \mathcal{L}$$

then $U - V \in \mathcal{L}^{\perp_?}$.

Now comes the magic aspect of the choice of inner product:

$$\alpha_j = \frac{(U,\mathbf{A}d_j)}{(d_j,\mathbf{A}d_j)} = \frac{(\mathbf{A}U,d_j)}{(d_j,\mathbf{A}d_j)} = \frac{(F,d_j)}{(d_j,\mathbf{A}d_j)}$$

and so the coefficients in the approximations for U do not involve U (good, since we don't know it) but only F (which is known).

The algorithm is made practical by a trick to calculate the $\perp_A$ vectors without needing the previous vectors. This is accomplished by introducing the residual $r_k = AU_k - F = A(U_k - U)$. Note that $h_k := r_k + F$ is in the space $\mathbf{A}S_k$. We rewrite the original minimization problem to show that h_k minimizes the problem

$$\|\mathbf{A}^{-1}(h - F)\|_A^2, \; h \in \mathbf{A}S_k.$$

Since

$$\|\mathbf{A}^{-1}x\|_A^2 = (\mathbf{A}^{-1}x, \mathbf{A}\mathbf{A}^{-1}x) = (\mathbf{A}^{-1}x, x) = \|x\|_{A^{-1}}^2$$

our problem is also that of minimizing

$$\|(h - F)\|_{A^{-1}}^2, \; h \in \mathbf{A}S_k.$$

As before, the minimizer h_k must be such that $h_k - F = r_k$ is $\perp_{A^{-1}}$ to $\mathbf{A}S_k$. This shows easily that

$$r_k \quad \perp \quad S_k \tag{2}$$
$$r_k \quad \perp_A \quad S_{k-1}. \tag{3}$$

Theorem 1 *The vector d_k is in span $\{r_{k-1}, d_{k-1}\}$.*

Proof: Note that $S_k = \text{span}\,\{F, r_1, \ldots, r_{k-1}\}$. This follows easily from the recurrence relation $r_k = r_{k-1} + \alpha_k \mathbf{A} d_k$. Thus, we can write

$$d_k = r_{k-1} - \sum_{i=1}^{k-1} \gamma_i d_i.$$

We take $\mathbf{A}$-inner products to solve for γ_i:

$$\gamma_i = (r_{k-1}, \mathbf{A}d_i)/(d_i, \mathbf{A}d_i).$$

Now, using (3) we know that $\gamma_i = 0$ for $i = 1, \ldots k-2$. $\square$.

We change notation slightly from the theorem and write

$$d_k = r_{k-1} + \beta_k d_{k-1}$$

where

$$\beta_k = (r_{k-1}, \mathbf{A}d_{k-1})/(d_{k-1}, \mathbf{A}d_{k-1}).$$

Using the orthogonality results above many other formulas for α and β can be derived as in the algorithm described below. Begin with $U_0 = 0$ and $r_0 = F$ and compute

1. $\beta_j = (r_{j-1}, r_{j-1})/(r_{j-2}, r_{j-2})$ (except $\beta_1 = 0$)
2. $d_j = r_{j-1} + \beta_j d_{j-1}$ (except $d_1 = r_0$)
3. $\alpha_j = (r_{j-1}, r_{j-1})/(d_j, \mathbf{A}d_j)$
4. $U_j = U_{j-1} + \alpha_j d_j$
5. $r_j = r_{j-1} - \alpha_j \mathbf{A} d_j$

We now show that the above method (called the conjugate gradient method) terminates with the exact solution in a finite number of steps. (Below, N is the size of the problem $\mathbf{A}_h U = F$).

Theorem 2 *Let $N^* = \dim S_N$. The CG algorithm provides the exact solution U in N^* steps (***i.e.*** $U = U_{N^*}$).*

Proof: If $N^* = N$ then $S_N = R_N$ (R_N is the whole space of real N-vectors) and so $U_N = U$. Otherwise, consider $r_{N^*} = \mathbf{A}U_{N^*} - F$. Note that $r_{N^*} \in S_{N^*}$ since this is the biggest space generated by the action of $\mathbf{A}$ on F. Also, by (2), $r_{N^*} \in S_{N^*}^{\perp}$. Therefore, $r_{N^*} = 0$ and so $U = U_{N^*}$. $\square$

Note that only one matrix multiply is needed per iteration. We make the following remark:

Note 1 (You Don't Need The Exact Solution of the Discrete Problem) *Our discrete solution U is only an approximation of the exact solution. Therefore, we can terminate the CG method before we have the exact solution to the discrete problem and not care as long as our answer is "accurate enough". However, it is often helpful to make the error from the solution procedure much smaller than the discretization error so as not to confuse the source of errors. When the method has been tested on a class of problems, the accuracy of the solution procedure can be relaxed to increase efficiency.*

How well it works

We know that the CG method will give the exact solution in at most N iterations, but how large is the error at each step? Recall that

$$\|U - U_k\|_A^2 = \|r_k\|_{A^{-1}}^2 = \min_{h_k \in AS_k} \|F - h_k\|_{A^{-1}}^2. \tag{4}$$

Notice that all vectors of the form $F - h_k$ can be written as $P(A)F$, where P is a polynomial in Π_k which includes all polynomials of degree $\leq k$ with $P(0) = 1$. Recall that A is positive definite symmetric so it has a full set of ortho-normal (in the euclidean inner product $(\cdot, \cdot)$) eigenvectors $\{v_i\}$ with eigenvalues λ_i. We expand F in this basis

$$F = \sum_{i=1}^{N} a_i v_i$$

and note that $U = \sum_{i=1}^{N} \lambda_i^{-1} a_i v_i$ and that $\|F\|^2 = \sum a_i^2$, $\|F\|_A^2 = \sum \lambda_i a_i^2$, etc. Suppose we pick a polynomial P in Π_k such that

$$|P(\lambda_i)| \leq M, \quad \text{for all } i.$$

Then

$$P(A)F = \sum P(\lambda_i) a_i v_i$$

and so

$$\begin{aligned} \|P(A)F\|_{A-1}^2 &= \sum \lambda_i^{-1} P^2(\lambda_i) a_i^2 \\ &\leq M^2 \sum \lambda_i^{-1} a_i^2 \\ &= M^2 \|U\|_A. \end{aligned}$$

Therefore, using (4), we have

$$\|U - U_k\|_A \leq M\|U\|_A.$$

We see that the study of the performance of CG methods for finite iterations boils down to the study of polynomials on the spectral set of A (more on spectra next section). To prove the main theorem of this section, we will use some properties of the Chebyshev polynomials T_k coming from the two equivalent formulas below:

1. $T_k(x) = \frac{1}{2}[(x + \sqrt{x^2 - 1})^k + (x - \sqrt{x^2 - 1})^k]$.
2. $T_k(x) = \cos[k \cos^{-1} x]$ for $|x| \leq 1$.

From formula 2 we see that

$$\max_{|x|\leq 1} |T_k(x)| = 1 \tag{5}$$

and

$$T_k(x_i) = (-1)^i \text{ for } \; x_i = cos(i\pi/k), i = 0, 1, \ldots k. \tag{6}$$

From formula 1 we get the bound

$$T_k(\frac{a+1}{a-1}) > \frac{1}{2}\left(\frac{\sqrt{a}+1}{\sqrt{a}-1}\right). \tag{7}$$

We are now in a position to state the main theorem which gives a bound on the error after k steps using only the values of the extreme eigenvalues,**i.e.** λ_1 and λ_N with $\lambda_1 < \lambda_N$ and $\lambda_i \in [\lambda_1, \lambda_N]$ for all i.

Theorem 3 *The CG iterates satisfy*

$$\|U - U_k\|_A \leq 2\gamma^k \|U\|_A \tag{8}$$

where

$$\gamma = \frac{\sqrt{a} - 1}{\sqrt{a} + 1}$$

and $a = \lambda_M/\lambda_1$. In addition, the number of iterations $p(\epsilon)$ to reduce the initial error $\|U\|_A$ by a factor of ϵ satisfies

$$p(\epsilon) \leq \frac{1}{2}\sqrt{a}\ln(2/\epsilon) + 1. \tag{9}$$

Proof: Since we know nothing about the structure of the spectra in the interval $[\lambda_1, \lambda_N]$ we will attempt to find the polynomials in Π_k that have a minimal maximum value B_k over the interval. It turns out that the scaled Chebyshev polynomials

$$P_k(x) = \frac{T_k[(\lambda_N + \lambda_1 - 2x)/(\lambda_N - \lambda_1)]}{T_k[(\lambda_N + \lambda_1)/(\lambda_N + \lambda_1)]}$$

have this property. Note that the scaling in the argument in numerator maps the interval $[\lambda_1, \lambda_N]$ in to the interval $[-1, 1]$ and that the argument in the denominator is greater than 1 so the value cannot be zero (all the zeros of T_k are in $[-1, 1]$). The maximum value of P_k on the spectral interval is

$$C_k = T_k[(\lambda_N + \lambda_1)/(\lambda_N + \lambda_1)]^{-1} \tag{10}$$

which occurs $k + 1$ times (with alternating sign) at the mapped points x_i from (6). To show that P_k has the desired property, assume that there is a polynomial $Q_k \in \Pi_k$ with

$$\max_{x \in [\lambda_1, \lambda_N]} |Q_k(x)| = B_k < C_k.$$

Now consider $R_k = Q_k - P_k$. This has a zero at $x = 0$ and alternates sign at the $k + 1$ mapped points x_i (since $|Q_k(x)| < C_k$ at all points in the interval). Since R_k must have a zero between each sign change, it has $k + 1$ zeros, so $R \equiv 0$. This proves the optimal quality of the polynomial P_k. The bound (10) along with (7) proves (8).

We now turn to a proof of (9). Clearly, we will satisfy the condition if p satisfies

$$2\gamma^p \leq \epsilon$$

or

$$p \geq \frac{\log(2/\epsilon)}{\log(1/\gamma)}.$$

Since

$$\log[(\sqrt{a} + 1)/(\sqrt{a} - 1)] > 2/\sqrt{a}, \quad \text{for all} \quad a > 1$$

the result (9) follows. □

Clearly, this result is not the most optimal result, because it does not take into account the distribution of the eigenvalues within the interval. For instance, even if $\lambda_1 = 1$ and $\lambda_N = 1000$ (large ratio, expect poor performance from the theorem) we will get convergence in *two* iterations if they are the only eigenvalues present (with as many multiplicities as you want).

We now consider what the spectrum looks like for our first model problem. Actually, we have the answer already because the problem was diagonalized by the DFT. The eigenvalues were

$$n^2 + 1$$

going fron $n = 0$ to $n = N/2$. Thus the minimum eigenvalue is 1 and the maximum is $N^2/4 + 1$ so the ratio is $a \sim N^2$. Therefore, using the results of the theorem above, we get convergence to tolerance ϵ in $O(N \log(1/\epsilon))$ iterations.

Note 2 (Good News, Bad News) *This is a remarkable improvement over the original CG method (run to termination) that we get essentially for free. However, we could still be unhappy because the rate of convergence gets worse as the problem gets*

bigger. We would like the rate to be constant as the problem gets bigger (since we have to do more work per iteration anyway). As far as I know, this property holds only for some "nice" preconditioned CG methods (an example is given below) and Multi-Grid methods.

References

[1] Axelsson and Barker, "FE Solution of Boundary Value Problems".

[2] Glub and Van Loan, "Matrix Computations".

[3] W. Hackbush, "Multi-Grid Methods and Applications".

[4] G. Strang, "Introduction to Applied Math".

[5] R. Varga, "Matrix Iterative Analysis".

[6] van der Sluis and van der Horst, "The Rate of Convergence of Conjugate Gradients," Numer. Mathe. **48**, 543-560 (1986).

Math 521, Notes V

Application of boundary conditions in the FEM. As before, we can consider

(i) $u = 0$ Dirichlet

(ii) $u' = 0$ Neumann

(iii) $u \mp \alpha u' = 0$ ($\alpha > 0$, $-$ at $x=0$, $+$ at $x=1$)

(left end, right end)

(i) Consider the space H_0^1, completing C_{00} functions φ with $\varphi = 0$ at $x=0$ and $x=1$ in the H^1 norm. Then

$$(u, \varphi)_R = (f, \varphi)_{L_2} \qquad \forall\, \varphi \in H_0^1 \qquad (1)$$

defines the weak form of the problem including the boundary conditions. This is consistent with the strong solution since

$$\int_0^1 u'\varphi' + u'\varphi \Big|_0^1 + \alpha \int_0^1 u\varphi = \int_0^1 f\varphi. \qquad (2)$$

(zero if we consider $\varphi \in H_0^1$)

and gives a unique solution. Considering $S \subset H_0^1$ in (1) naturally leads to an approximation.

If $u(0) = \beta \neq 0$, consider a given function $b(x)$ that has $b(0) = \beta$, $b(1) = 0$. Then, consider

$u = b + v$, with $v \in H_0^1$. Then, v solves the homogeneous problem

$$(v, \varphi)_A = (f, \varphi)_{L_2} - (b, \varphi)_A \qquad \forall\ \varphi \in H_0^1.$$

Although v may depend on the choice of $b(x)$, the resulting u does not. In the FEM, if there is an element $\varphi_0(x)$ that would correspond to the value at $x=0$, we can take $b(x) = \beta\,\varphi_0(x)$. This is equivalent to introducing the element $\varphi_0(x)$ and setting its coefficient to β rather than applying (1) with φ_0 as a test function.

(ii) $u' = 0$ is applied in a weak sense as follows. Considering (2), with $u' = 0$ at $x=0$ and $x=1$ we have

$$(u, \varphi)_A = (f, \varphi)_{L_2} \qquad \forall\ \varphi \in H^1.$$

Thus, $u' = 0$ appears as a "natural" condition.

The details for a uniform grid, piecewise linear approximation are shown below. Here, there are $N+1$ unknowns at $x_0 = 0$, $x_1 = h$, ... $x_N = 1$.

$$K_{ij} = \int_0^1 \varphi_i' \varphi_j'\, dx, \qquad M_{ij} = \int_0^1 \varphi_i \varphi_j\, dx$$

In the interior ($i \neq 0, N$), we previously derived

$$K_{ij} = \begin{cases} 2/h & \leftarrow i=j \rightarrow \\ -1/h & \leftarrow i=j\pm1, \rightarrow \\ 0 & \leftarrow \text{otherwise} \rightarrow \end{cases} \quad \begin{matrix} 2h/3 \\ h/6 \\ 0 \end{matrix} \Bigg\} = M_{ij}$$

Consider $K_{00} = \int_0^1 [\varphi_0']^2 dx = \int_0^h \left(-\frac{1}{h}\right)^2 = 1/h.$

$K_{01} = -1/h.$

$M_{00} = \int_0^1 (\varphi_0)^2 dx = \int_0^h \left(1-\frac{x}{h}\right)^2 dx = h/3.$

$M_{01} = h/6.$

$F_0 = \int_0^h \left(1-\frac{x}{h}\right) f(x) \approx f(0)\, h/2.$

Thus we see that the $i=0$ equation looks like the interior equations divided by 2, if we considered introducing a ghost value U_{-1} and setting $U_{-1} = U_1$, equivalently

$$\frac{U_1 - U_{-1}}{2h} = 0,$$

which we saw in the FD case was a second order approximation of $u'=0$. Note that it was not necessary to introduce U_{-1} in the FEM,

This was just to show that the "natural" application of $u'=0$ led to something we could recognize.

If $u'(0)=\beta \neq 0$, we can implement this in the following way; considering again (2). For each element for which $\varphi_i(0) \neq 0$ (in the piecewise linear setting above this is only φ_0) we would need to consider

$$\int_0^1 U'\varphi_i' + u'\varphi_i\Big|_0^1 + a\int_0^1 u\varphi_i = \int_0^1 \varphi_i f$$

$$\sum_{j=0}^{N} U_j \varphi_j' \varphi_i' - \beta\,\varphi_i(0) + a\sum_{j=0}^{N} U_j \varphi_j \varphi_i = F_i$$

usual term — Constant, move to the RHS. — usual terms.

Here, it is assumed that if $\varphi_i(0)\neq 0$, $\varphi_i(1)=0$.

(iii) Proceed as in (2), leading to

$$(u,\varphi)_A + \alpha\left(\varphi(0)u(0) + \varphi(1)u(1)\right) = (f,\varphi)_{L_2}.$$

Note that the sign choices in (iii) are physically motivated and important in what follows.

Note that the statement above is equivalent to

$$(u,\varphi)_{\hat{A}} = (f,\varphi)_{L_2} \qquad (3)$$

where $(u,\varphi)_{\hat{A}} = (u,\varphi)_A + \alpha\left(\varphi(0)u(0) + \varphi(1)u(1)\right)$.

Note that this is a well defined inner product space because of the sign choices

$$\|u\|^2_{\hat{A}} = \|u\|^2_A + \alpha\left([u(0)]^2 + [u(1)]^2\right)$$

A FEM can be based on (3) in a straightforward way, with modified terms at the boundary due to the extra terms in the inner product.

Math 521, Spring 2013
Notes, part VI, Quadrature (Numerical Integration)

To approximate the integrals

$$F_i = \int_0^1 \varphi_i(x)\, f(x)\, dx \qquad (1)$$

$$K_{ij} = \int_0^1 \varphi_i'(x)\, \varphi_j'(x)\, dx \qquad (2)$$

$$M_{ij} = \int_0^1 \varphi_i(x)\, \varphi_j(x)\, dx \qquad (3)$$

$$\|E\|_{H_1}^2 = \int_0^1 \left\{ |U' - u'|^2 + |U - u|^2 \right\} dx \qquad (4).$$

we can consider different numerical integration techniques. Note that (4) is needed if we are investigating the convergence to a known solution $u(x)$. Also note that (2) and (3) involve integrals of piecewise polynomials, so these integrals can be done analytically.

All integrals above can be written as a sum of integrals over the grid subintervals

$$\int_{x_m}^{x_{m+1}} g(x)\, dx \qquad (5).$$

so let's consider these more basic integrals separately.

Consider the affine transformation $x \to y$ where

$$x \in [x_m, x_{m+1}] \to y \in [-1, 1].$$

$$y(x) = 2\left(\frac{x - x_m}{x_{m+1} - x_m}\right) - 1, \text{ with inverse}$$

$$x(y) = \frac{1}{2}(y+1)(x_{m+1} - x_m) + x_m.$$

We can use substitution in the integral (5) to change it to an integral over $y \in [-1, 1]$.

$$\int_{x_m}^{x_{m+1}} g(x)\,dx = \int_{-1}^{1} g(x(y))\frac{dx}{dy}\cdot dy$$

$$= \frac{h_m}{2}\int_{-1}^{1} g(x(y))\,dy \qquad (6)$$

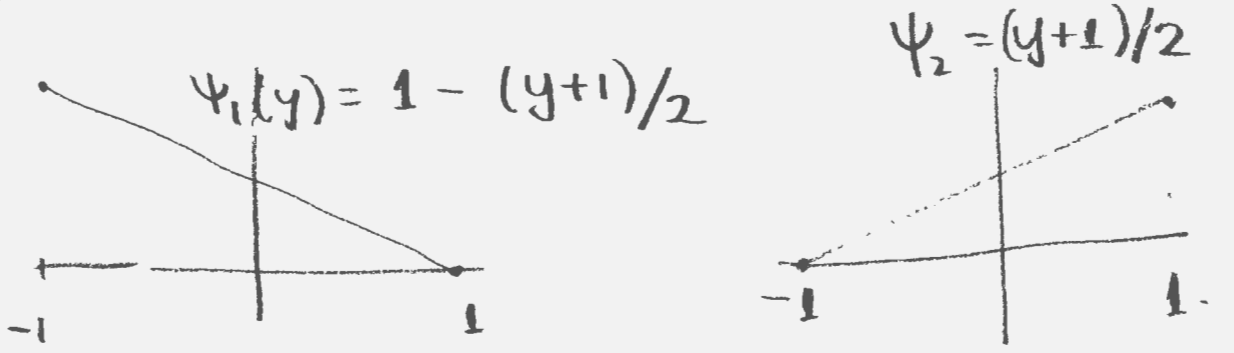
where $h_m = x_{m+1} - x_m$ is the length of the m'th grid subinterval.

The interval $[-1, 1]$ is called the reference interval. In (1) & (3), if we assume that both $\varphi_i(x)$ and $\varphi_j(x)$ are non zero in the interval $[x_m, x_{m+1}]$ (otherwise they would not contribute to the integral) then

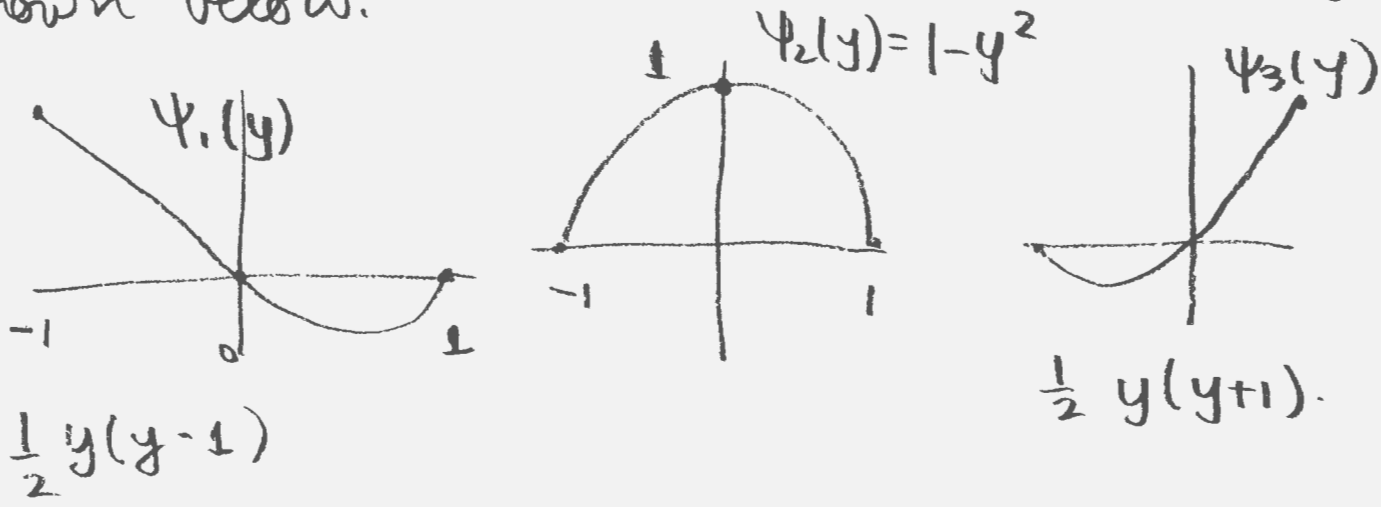
$$\varphi_i(x(y)) \text{ and } \varphi_j(x(y))$$

can only be functions in a finite set of functions $\{\psi_e(y)\}$ called reference elements.

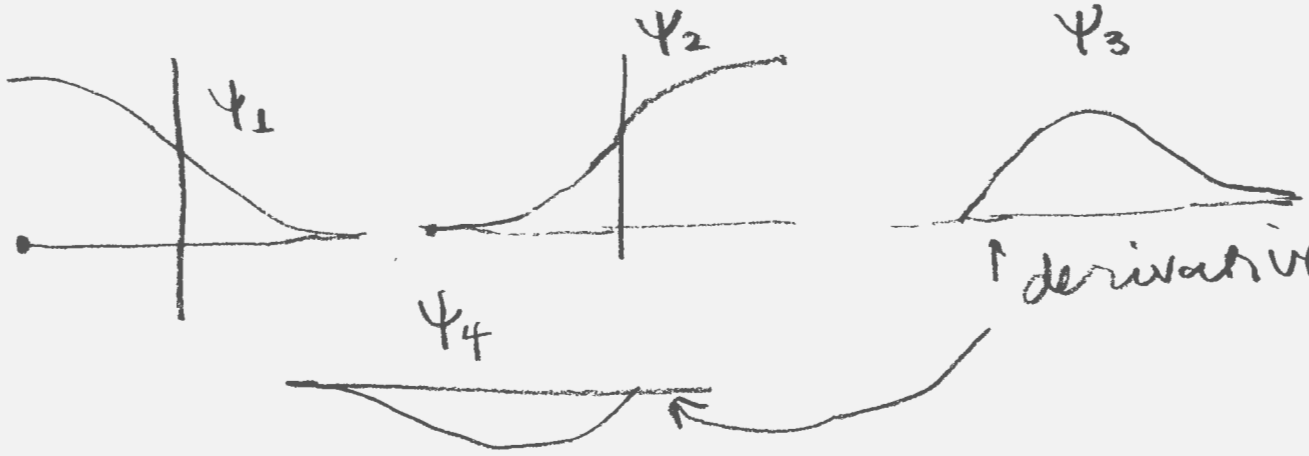
Examples: For piecewise linear elements, the reference elements are $\{\psi_1(y), \psi_2(y)\}$ shown below:



For piecewise quadratic elements discussed in class, we have $\{\psi_1(y), \psi_2(y), \psi_3(y)\}$ shown below.



For piecewise cubic C_1 elements we would have $\{\psi_1, \psi_2, \psi_3, \psi_4\}$.



These are cubic polynomials that satisfy

$$\psi_1(-1)=1,\quad \psi_1'(-1)=0,\quad \psi_1(1)=0,\quad \psi_1'(1)=0$$

for example.

$$\psi_1(y) = a+by+cy^2+dy^3 \quad \text{(linear algebra)}$$

$$= \frac{1}{2} - \frac{3}{4}y + \frac{1}{4}y^3$$

For integrals (2) and (4) where φ_i' and φ_j' appear, they can be replaced by ψ_e' in the reference interval integrals (6).

Gaussian quadrature for (6) using n points is exact for polynomials up to order $2n-1$.

$$\int_{-1}^{1} g(y)\,dy \approx \sum_{i=1}^{n} w_i\, g(y_i) \qquad (7)$$

Gauss-Legendre points and weights:

$n=1$, $\quad x_1 = 0$, $\quad w_1 = 2$ (midpoint rule)

$n=2$, $\quad x_{1,2} = \pm\frac{1}{\sqrt{3}}$, $\quad w_{1,2} = 1$.

$n=3$ $\quad x_1 = 0$ $\quad w_1 = 8/9$

$x_{2,3} = \pm\sqrt{3/5}$ $\quad w_{2,3} = 5/9$.

I will derive these in later notes, but the important properties of (7) are that it is exact for polynomials of order $2n-1$ and for other functions the error is proportional to $g^{(2n)}(\theta)$.

Integrals (2) & (3) can then be done exactly with Gaussian quadrature of high enough order. Since these integrals involve only reference elements Ψ_e in (6), only $l(l-1)$ integrals need be done this way, and then assembled in the sum with weights $hm/2$ as seen in (6).

The integral (1) cannot be done exactly. It becomes a sum of integrals of the form (6) with

$$g(x(y)) = f(x(y))\,\Psi_e(y).$$

Consider using Gaussian quadrature on this integral with n points using p'th order elements (Ψ_e are polynomials of order p). The error in (7) [leaving out the factor hm in (6)] is proportional to the $2n$'th derivative of g above, with respect to y.

$$\frac{dg}{dy} = \frac{df}{dx}\cdot\frac{dx}{dy}\Psi_e + f(x(y))\,\Psi_e'(y).$$

In the formula above $\frac{dx}{dy} = \frac{h_m}{2}$, $\frac{df}{dx}$ is the derivative of f in regular coordinates. We can continue

$$\frac{d^2g}{dy^2} = \frac{d^2f}{dx^2}\left(\frac{h_m}{2}\right)^2 \psi_e + 2\left(\frac{h_m}{2}\right)\frac{df}{dx}\psi_e' + f(x(y))\,\underbrace{\psi_e''}_{\text{for linear elements this is zero.}}$$

In general,

$$\frac{d^n g}{dy^n} = \sum_{j=0}^{2n} \binom{2n}{j} \frac{d^j f}{dx^j}\left(\frac{h_m}{2}\right)^j \psi_e^{(2n-j)}$$

↑ Binomial coefficients $\frac{n!}{j!(n-j)!}$

Since $\psi_e^{(2n-j)} \equiv 0$ if $2n-j > p$ (we assumed $\{\psi_e\}$ were p'th order polynomials), the sum above is actually

$$\frac{d^n g}{dy^n} = \sum_{j=2n-p}^{2n} \binom{2n}{j} \frac{d^j f}{dx^j}\left(\frac{h_m}{2}\right)^j \psi_e^{(2n-j)}$$

which is $O(h_m^{2n-p})$ if the RHS $f(x)$ is sufficiently smooth. In general, elements with p'th order polynomials are order $p+1$ accurate in L_2, so we would want to preserve this accuracy as shown below by taking $2n-p \geq p+1$, or

$2n \geq 2q+1$, or $n > q$. This suggests that for second order methods ($m=1$, linear elements) we should use quadrature with $n=2$ points.

Note: Let's investigate this further for the $m=1$ case. Consider a representative subinterval of length h, and

$$\int_{-h/2}^{h/2} \left(\frac{x+h/2}{h}\right) f(x)\,dx \approx \int_{-h/2}^{h/2} \left(\frac{x+h/2}{h}\right)\left(f(0) + f'(0)x + \frac{f''}{2}(0)x^2 + \cdots\right)$$

↑ representative $\ell_i(x)$

$$= \frac{h}{2} f(0) + \boxed{\frac{h^2}{12} f'(0)} + O(h^3).$$

↑ midpoint rule ✓.

This is the term that prevents second order convergence.

However, on a uniform grid, this term is combined with

$$\int_{h/2}^{3h/2} \left(\frac{3h/2 - x}{h}\right) f(x)\,dx \approx \frac{h}{2} f(h) - \boxed{\frac{h^2}{12} f'(h)} + O(h^3)$$

and the two boxed terms cancel out at highest order. The same is true when trapezoidal rule is used on regular grids.

To examine the influence of the quadrature error on the solution, consider modifying the right hand side function in each subinterval. (This is not something to implement, just to investigate how the quadrature error affect the accuracy of the solution).

$$\int_{x_m}^{x_{m+1}} \varphi_i(x) f(x)\, dx = \frac{h_m}{2} \int_{-1}^{1} \psi_\ell(y) f(x(y))\, dy.$$

$$= \frac{h_m}{2} \left(\begin{array}{c} \text{Gaussian} \\ \text{quadrature} \\ \text{approx} \end{array} G_{m,i} + \tau_{\ell,m} \right)$$

with $\tau_{\ell,m} = O(h_m^{2n-q})$, with q the order of the element functions as before.

or, $$\frac{h_m}{2} \left(G_{m,i} \right) = \int_{x_m}^{x_{m+1}} \varphi_i f(x)\, dx - \frac{h_m}{2} \tau_{\ell,m}.$$

On each interval, construct

$$r(x) = \sum_{\ell=1}^{q+1} a_\ell \psi_\ell(x(y)) \quad , a_\ell \text{ constants}$$

so that $$\int_{x_m}^{x_{m+1}} \psi_\ell(x(y))\, r(x)\, dx = \frac{h_m}{2} \tau_{\ell,m} \quad \text{for each } \ell.$$

This can be done by solving the linear system

$$B \underline{a} = \underline{\tau}$$

with $B_{ij} = \int_{-1}^{1} \psi_i(y)\,\psi_j(y)\,dy$

Since $\{\psi_\ell(y)\}$ are l.i. on $[-1,1]$, B is invertible. We have that $a_\ell = O(h_m^{2n-8})$, the same size as $\tau_{\ell,m}$.

$r(x)$ is constructed so that if we assemble

$$R(x) = \sum_{\substack{\text{intervals} \\ m}} r_m(x)$$

then solving for $U(x)$ with the FEM using Gaussian quadrature on subintervals with n points, is the same as solving for $U(x)$ with exact integration using a RHS of $f(x) - R(x)$.

Let $v(x)$ solve

$$(v, \varphi)_A = (R, \varphi)_{L_2} \qquad \forall\ \varphi \in H^1.$$

(our usual problem with RHS R). Then

$$\|v\|_{H^2} \le \|R\|_{L_2} = O(h^{2n-8})$$

Now

$$\|U - u\|_A = \|U - (u-v) - v\|_A \le \cdots$$

U with Gaussian quadrature, or equivalently the problem with RHS $f - R$

$u - v$: the exact solution to $f - R$

$$\| U - (u-V) \|_A + \|V\|_A$$

$$\downarrow \text{usual FE approximation}$$

$$\leq C_1 h^q + C_2 h^{2n-q}.$$

So far $n \geq q$ is sufficient to preserve the order of accuracy of the scheme. However, for the L_2 norm estimate,

$$\| U - (u-V) \|_{L_2} + \|V\|_{L_2}$$

$$\downarrow \text{usual} \qquad \downarrow \text{same estimate}$$

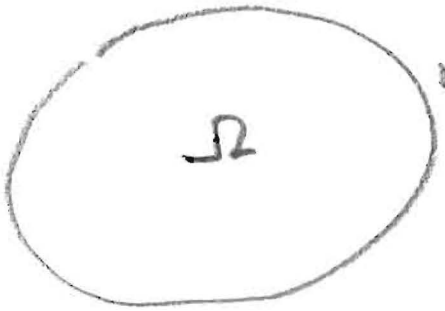
$$\leq C_1 h^{q+2} + C_2 h^{2n-q}$$

then it appears that $n > q$ is needed.

Note: I am not sure if this is real or a limitation of the analysis. I suggest we try it out computationally to see what convergence we get in L_2 with $n = q$.

Math 521, Spring 2013
Notes VII

Let us now consider 2D problems. One of the strengths of the FEM is its ability to handle complex geometries, so let's consider a problem in a domain Ω



boundary denoted $\partial\Omega$ or Γ.

We will suppose in general that the boundary of Ω is C_1, that is it can be represented by a parametrized curve

$$\partial\Omega = \{ (x(s), y(s)),\ 0 \le s \le 1 \}$$

with x, y 1-periodic and continuously differentiable describing a simple closed curve (no self intersections). Our old problem 1 has a close relative in 2D:

Problem 2 Find $u(x,y)$ for $(x,y) \in \overline{\Omega}$ such that

$$-\Delta u + a u = f(x,y) \qquad (1)$$

and $u = 0$ on $\partial\Omega$. Here, $a > 0$ and $f(x,y)$

are given.

"Δ" is called "the laplacian" and is sometimes denoted "∇^2" or "$\nabla \cdot \nabla$". It is defined as

$$\Delta u := u_{xx} + u_{yy} \quad \text{in 2D}$$

$$\Delta u := u_{xx} + u_{yy} + u_{zz} \quad \text{in 3D.}$$

As in 1D, we can define a weak version of Problem 2 by multiplying (1) by a test function $\varphi(x,y)$ [C_∞ with zero values on the boundary] and integrating by parts (Gauss identity).

$$(u, \varphi)_A = (f, \varphi)_{L_2} \quad \forall \varphi \qquad (2)$$

Uniquely determines $u \in H^2 \cap H_0^1$ for every $f \in L_2$. In 2D, the inner products and spaces are

$$(f, g)_{L_2} = \int_\Omega f g \, dx dy.$$

$$(u, \varphi)_A = (u_x, \varphi_x)_{L_2} + (u_y, \varphi_y)_{L_2} + a(u, \varphi)_{L_2}$$

and H_0^1 is the completion of C_∞ functions with zero values on the boundary in the

norm

$$\|u\|_{H^1}^2 = \|u\|_{L_2}^2 + \|u_x\|_{L_2}^2 + \|u_y\|_{L_2}^2$$

and finally the H^2 norm is given by

$$\|u\|_{H^2}^2 = \|u\|_{H^1}^2 + \|u_{xx}\|_{L_2}^2 + \|u_{yy}\|_{L_2}^2 + \|u_{xy}\|_{L_2}^2$$

Again following the approach in 1D, we can consider a finite dimensional subspace $S \subset H_0^1$ and consider an approximate solution $U \in S$ that satisfies

$$(U, \varphi)_A = (F, \varphi)_{L_2} \qquad \forall\ \varphi \in S. \tag{3}$$

If we take a basis $\{\varphi_i\}_{i=1}^N$ for S then

$$U(x,y) = \sum_{i=1}^{N} U_i\, \varphi_i(x,y)$$

and (3) is satisfied if it holds for all φ_i, leading to the linear system

$$\mathbb{A}\,\underline{U} = \underline{F} \tag{4}$$

where $\mathbb{A} = \mathbb{K} + a\mathbb{M}$,

$$K_{ij} = (\varphi_{i,x}, \varphi_{j,x})_{L_2} + (\varphi_{i,y}, \varphi_{j,y})_{L_2}$$

$$M_{ij} = (\varphi_i, \varphi_j)_{L_2}$$

and $F_i = (\varphi_i, f)_{L_2} = \int_\Omega \varphi_i(x,y) f(x,y)\, dx dy$

As in the 1D case it can be shown that (4) is always solvable and the resulting solution $U(x,y)$ is the closest element in S to the exact solution in the A-norm

$$\underbrace{\|U-u\|_A}_{\text{error}} \leq \|u-v\|_A \quad \text{for all } v \in S.$$

In the next set of notes, we will consider using the interpolation of u onto S to show convergence of the method for various elements.

Some examples of conforming element spaces, presented as a list of the following:

$\hat{K}$: reference element

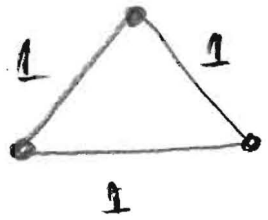
$\hat{P}$: finite dimensional set of functions considered on $\hat{K}$.

$\hat{\Sigma}$: Basis for $\hat{P}$.

F : allowable maps from $\hat{K}$ to an element K in the mesh.

Example 1 Piecewise linear, continuous elements on triangles.

$\hat{K}$: equilateral triangle with unit lengths.



$\hat{P}$: first order polynomials (P_1), span $\{1, x, y\}$

$\hat{\Sigma}$: $\{\psi_\ell\}_{\ell=1}^{3}$ chosen so that they have unit values at one of the triangle vertices.

F: first order polynomials for each component.

Note: If F is $\hat{P} \times \hat{P}$ then the element is known as an isoparametric element.

Note: If the mesh is composed of triangles with no hanging nodes, then the values along each edge will be linear between the vertex values, making the resulting approximating function continuous.

Example 2 Piecewise continuous, quadratic elements.

$\hat{K}$: unit equilateral triangle

$\hat{P}$: $P_2 = \text{span}\{1, x, y, x^2, xy, y^2\}$.

$\hat{\Sigma}$: $\{\psi_e\}_{e=1}^{6}$ chosen to have unit values at each of the six points marked on the triangle above.

F: We can consider allowing this to be an isoparametric element with some care. At all interior edges, the edges must be straight and midpoints must map to midpoints. This ensures that the map is linear restricted to that edge and so ensures the continuity of the grid function. However, we allow more general maps to boundary edges, to be able to obtain full order convergence there.

Example 3 Bilinear elements

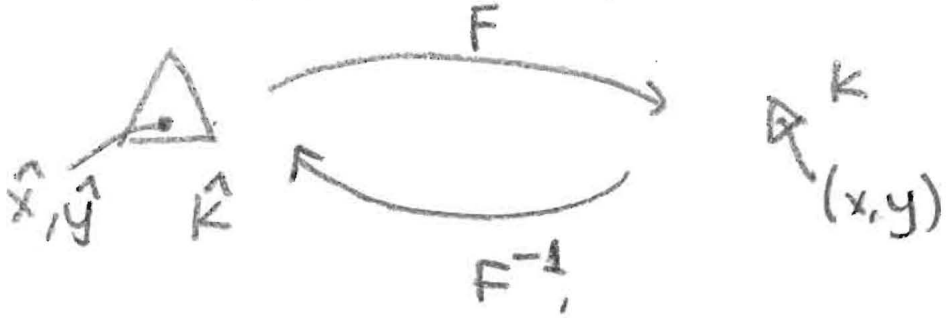
$\hat{K}$: unit square 

$\hat{P}$: span $\{1, x, y, xy\}$.

$\hat{\Sigma}$: $\{\psi_e\}_{e=1}^{4}$ with unit values marked above.

F: these elements can be taken to be isoparametric, since $\hat{P}$ restricted to edges is just linear, so this choice ensures continuity between elements.

Quadrature is done on the reference element. Consider part of the calculation of an F_i value, doing the integral on one element:



$$\int_K f(x,y)\, \psi_i(x,y)\, dx\, dy$$

$$= \int_K f(x,y)\, \psi_e\left(F^{-1}(x,y)\right) dx\, dy \qquad (4)$$

for some $\psi_e \in \hat{\Sigma}$. Now change the integral to one over the reference element

$$(4) = \int_{\hat{K}} f(F(\hat{x},\hat{y}))\, \psi_e(\hat{x},\hat{y}) \det J_F(\hat{x})\, d\hat{x}$$

where $$J_F = \begin{bmatrix} \frac{\partial F_1}{\partial \hat{x}} & \frac{\partial F_1}{\partial \hat{y}} \\ \frac{\partial F_2}{\partial \hat{x}} & \frac{\partial F_2}{\partial \hat{y}} \end{bmatrix}.$$

In the case where F is in $P_1 \times P_1$, $\det J_F$ is a constant, the ratio of the area of K to the area of $\hat{K}$. In general, this term will be a polynomial, and the order of this polynomial plus the order of polynomials in $\hat{P}$ must be taken into account when choosing the quadrature order.

Quadrature points on the square can be taken to be the tensor product of 1D quadrature points, which is convenient. However, for higher order quadrature this is not optimal.

For the equilateral triangle, evaluating at the midpoint with weight 1/2 (the area)

is second order accurate (exact for functions in P_1) and third order accuracy (exact for P_2) can be obtained with three points each with weight 1/6. (This is an assignment question).

Math 521, Spring 2013
Notes VIII

Lemma 1 Consider $u(\underline{x})$ with $\underline{x} \in \Omega$. Then if Ω satisfies a cone condition specified below and $u \in H^n$ with $n > d/2$, u can be identified with a continuous function on Ω with

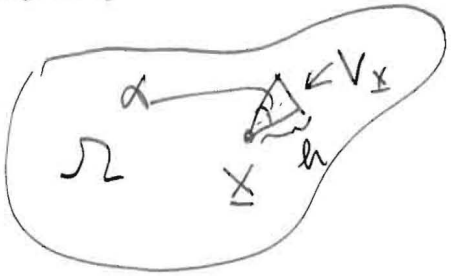
$$\|u\|_\infty \leq C \|u\|_{H^n(\Omega)}$$

with C depending only on n and Ω.

Cone condition: Ω satisfies a cone condition if there exists $\alpha, h > 0$ such that for any $\underline{x} \in \Omega$ one can fit a right spherical cone $V_{\underline{x}}$ with vertex $\underline{x}$, angular opening α and height h with $V_{\underline{x}} \subset \Omega$.

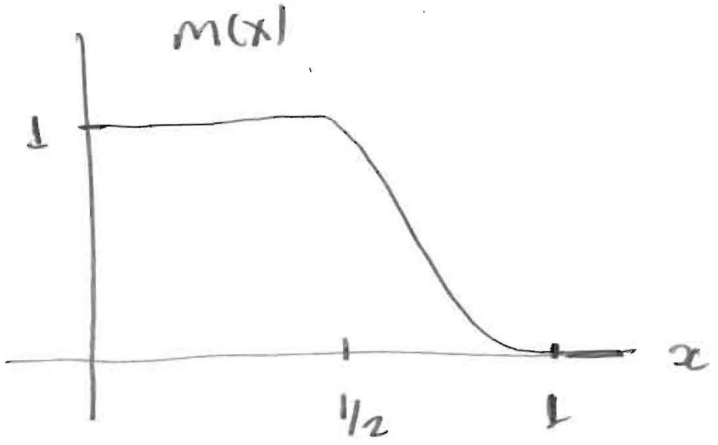
Note: If $\partial\Omega$ is C_1 then Ω satisfies a cone condition.

The picture of the cone is shown in the 2D ($d=2$) case below:



Proof: Construct a C_∞ function of one variable $m(x)$ with $x \geq 0$ with

$$m(x) = \begin{cases} 1 & 0 \leq x \leq 1/2 \\ 0 < m(x) < 1 & 1/2 < x < 1 \\ 0 & x \geq 1 \end{cases}$$



Fix $\underline{x} \in \Omega$, find the cone $V_{\underline{x}}$ guaranteed by the cone condition and parametrize it by (solid) angle θ and $0 \le r \le h$.

Now $$u(x) = -\int_0^h \frac{\partial}{\partial r}\left(m\left(\frac{r}{a}\right) u(r,\theta)\right) dr.$$

for every θ. Now integrate over the (solid) angle θ (dS_θ)

$$u(x) = \frac{-1}{|\alpha|} \int_\theta \int_0^h \frac{\partial}{\partial r}\left(m\left(\frac{r}{a}\right) u(r,\theta)\right) dr \cdot dS_\theta$$

↑ "1".

Integrate by parts $n-1$ times, using the fact that all all powers of r are zero at $r=0$ and m and all its derivatives are zero at argument 1, obtaining.

$$|u(x)| \le C \int_\theta \int_0^h r^{n-1} \frac{\partial^n}{\partial r^n}\left[m\left(\frac{r}{a}\right) u(r,\theta)\right] dr d$$

Write $r^{n-1} = r^{n-d} r^{d-1}$ and note that

$$r^{d-1}\, dr\, dS_\theta = dV$$

leading to

$$|u(x)| \le C \int_{V_x} r^{n-d} \frac{\partial^n}{\partial r^n}\left[m\left(\frac{r}{h}\right)u\right] dV. \quad (1)$$

Since $n > d/2$, $n-d > -d/2$ so

$$\int_{V_x} (r^{n-d})^2 dV = C \int_0^h (r^{n-d})^2 r^{d-1} dr$$

$$= C \int_0^h r^p dr \quad \text{with } p = 2(n-d)+d-1 > -1$$

finite

So $r^{n-d} \in L_2(V_x)$. use C-S on (1) to obtain

$$|u(x)| \le C \|u\|_{H^n(V_x)} \le C \|u\|_{H^n(\Omega)}$$

Lemma 2 If $\{v_n\}$ is a bounded sequence in $H^{n+1}(\Omega)$, then it has a convergent subsequence in $H^n(\Omega)$.

This is Rellich's lemma, and the proof is left to a part B assignment question.

Theorem Consider a reference element $\hat{K}$ and a set of reference functions $\{\psi_\ell\}$ that span polynomials of order k. assume that $\{\psi_\ell\}$ corresponds to values at points $\underline{x}_\ell$ and that $u \in H^{k+1}(R)$ with $k+1 > d/2$, ensuring with lemma 1 that u has pointwise defined values. Then

$$\|u\|_{H^{k+1}(R)} \le C \left(|u|_{H^{k+1}(R)} + \sum_\ell |u(\underline{x}_\ell)| \right)$$

where $|u|_{H^{k+1}(\hat{K})}$ denotes the L_2 norm of order $k+1$ (but not lower order) derivatives of u.

Proof: Suppose not, then there exists a sequence $\{v_n\}$ such that

$$\|v_n\|_{H^{k+1}(\hat{K})} = 1$$

and $$|v_n|_{H^{k+1}(\hat{K})} + \sum_e |v_n(x_e)| \to 0 \quad \text{as } n \to \infty. \quad (2)$$

Using Lemma 2, there is a subsequence $\{v_{n_k}\}$ that converges in $H^k(\hat{K})$. Since from (2) we see that $|v_{n_k}|_{H^{k+1}} \to 0$ we have that $\{v_{n_k}\}$ must converge in H^{k+1} to some v_*, with $\|v_*\|_{H^{k+1}} = 1$. Since $|v_*|_{H^{k+1}}(\hat{K}) = 0$ we see that $v_* \in P^k$ (polynomials of order k).

However, $\sum_e |v_n(x_e)| = 0$ together with the points x_e specifying a set of functions that include P^k shows that $v_* \equiv 0$. This contradicts $\|v_*\|_{H^{k+1}} = 1$ proving the Theorem.

Corollary: If $v \in H^{k+1}(\hat{K})$ and

$$\hat{v}(x) = \sum_e v(x_e)\,\psi_e(x) \quad \text{(the interpolation of } v \text{ using the reference functions)}$$

then

$$\| v - \tilde{v} \|_{H^{k+1}(\hat{K})} \leq C \, |v|_{H^{k+1}}$$

Proof using the Theorem, $v - \tilde{v}|_{x_e} = 0$ and $|\tilde{v}|_{H^{k+1}} = 0$

Here, it is assumed that the span of $\{\psi_e\}$ is exactly P^k.

Now consider the reference element mapped to an element in the mesh:

using the previous result, we see that

$$\| u - \hat{u} \|^2_{H^1(\hat{K})} \leq C \, |u|_{H^{k+1}(\hat{K})} \qquad (3)$$

Now assume that the elements have a uniform length scale $O(h)$, then

$$dx = d(h^d) \, d\hat{x}$$

$$\uparrow \qquad \uparrow$$

$$K \qquad \hat{K}.$$

Also, $\dfrac{\partial u}{\partial x_i} \sim \dfrac{K}{h} \dfrac{\partial u}{\partial \hat{x}_i}$

$$\frac{\partial^{k+1} u}{\partial x_\alpha} \sim \frac{K}{h^{k+1}} \frac{\partial u}{\partial \hat{x}_\alpha}. \qquad (4)$$

Note that here, we require that the triangles not have arbitrarily small angles, since then the constant K in (4) can become arbitrarily large or small.

Consider now

$$\|u-\hat{u}\|^2_{H^1(K)} = \int_K \left\{(u-\hat{u})^2 + |\nabla(u-\hat{u})|^2\right\} dx$$

(K: actual element)

$$\leq C h^d \int_{\hat{K}} \left\{(u-\hat{u})^2 + \frac{K}{h^2} |\hat{\nabla}(u-\hat{u})|^2\right\} d\tilde{x}$$

$$\leq C h^{d-2} |u|^2_{H^{k+1}(K)} \qquad \text{using (3)}$$

$$\leq C h^{2k} |u|^2_{H^{k+1}(K)}.$$

Thus we see that

$$\|u-\hat{u}\|_{H^1(K)} \leq O(h^k) |u|_{H^{k+1}(K)}.$$

Summing over all elements K in the mesh, we have

$$\|u-\hat{u}\|_{H^1(\Omega)} = O(h^k) |u|_{H^{k+1}(\Omega)}.$$

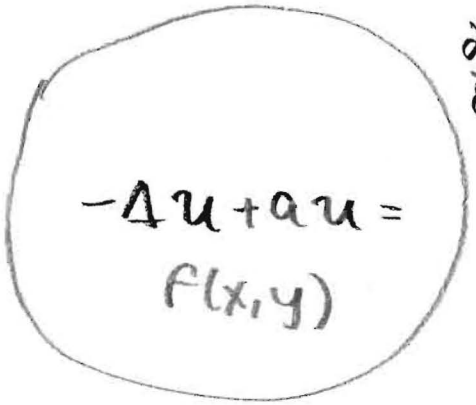
This approximation result proves the convergence of conforming finite element methods whose elements span P^k.

With linear elements (P^1) we see first order convergence in H^1_0 as before, and the same argument as in 1D can be used to show convergence of order $k+1$ in L_2.

Math 521, Spring 2013

Notes X on implementing boundary conditions in the FEM for the 2D Poisson problem

$u(x,y)$ to be determined



Note: if $a=0$ then we have the Poisson problem with Neumann conditions. We require the solvability condition

$$\int_\Omega f = 0$$

and u is only determined up to a constant in this case.

Multiply the equation by a test function $\varphi(x,y) \in H^1$ and integrate by parts.

$$\int_\Omega \nabla u \cdot \nabla \varphi - \int_\Gamma \varphi \frac{\partial u}{\partial n} ds + \int u\varphi = \int f\varphi \qquad (2)$$

so as in the 1D case, $\frac{\partial u}{\partial n} = 0$ appears as a "natural" condition, treating values on

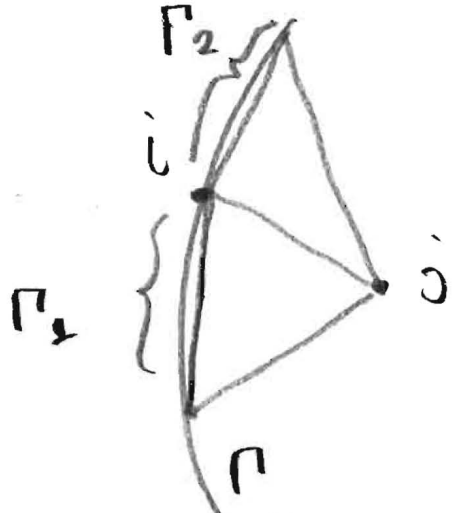
boundary nodes the same as interior nodes. If we have non-homogeneous conditions

$$\frac{\partial u}{\partial n} = g(s)$$

where s is the arc length parameter of the boundary curve Γ, $(x(s), y(s))$ then the boundary integral in (2) becomes

$$\int_{\Gamma} g(s)\, \varphi(x(s), y(s))\, ds \qquad (3)$$

When a finite element approximation is done, this adds an extra term to the RHS of every equation i where $\varphi_i(x,y)$ is nonzero on some part of the boundary. If we consider piecewise linear elements on triangles as an example,



See that $\varphi_j(x,y)$ is zero on Γ so there is no contribution from (3) in this case.

$\varphi_i(x(s), y(s))$ is not zero on the boundary on segments Γ_1 and Γ_2

↑ extend its value outside the triangle

so (3) will contribute to the RHS with terms

$$\int_{\Gamma_k} \varphi_i(x(s), y(s))\, g(s)\, ds \quad , \quad k = 1, 2.$$

these integrals can be approximated with quadrature as usual. In the literature, it is mentioned that these integrals must be done to a high precision to avoid loss of accuracy in the scheme, although I have not seen the concrete details presented in a clear way.

Math 521, Spring 2013
Notes XI

Consider the following 1D model nonlinear problem for $u(x)$, 1-periodic in x.

$$-u'' + u + u^3 = F(x) \qquad (1)$$

This nonlinear problem could represent a thermal conduction problem where the term u^3 represents a nonlinear heat loss to an ambient at zero temperature. This problem is an "easy" nonlinear problem. It can be shown that it has a strong solution for any continuous data $F(x)$. The weak formulation is found by multiplying (1) by test functions $\varphi \in H^1$ as usual: (2)

$$(u, \varphi)_{H^1} + (u^3, \varphi)_{L2} = (F, \varphi)_{L2} \qquad \forall \varphi \in H^1$$

where the usual "A" norm becomes the H^1 norm with the convenient choice of "a" = 1 in (1).

Consider now a FE approximation based on the weak formulation (2) starting with the approximation below as usual

$$u(x) \approx U(x) = \sum_{j=1}^{N_h} U_j \varphi_j(x)$$

Using (2) with $U(x)$ and test functions $\varphi_j(x)$ gives the system

$$\mathbb{A}\underline{U} + \underline{C}(\underline{U}) = \underline{F} \qquad (3)$$

where $\mathbb{A} = \mathbb{K} + \mathbb{M}$ and $F_i = \int_0^1 f(x)\,\varphi_i(x)\,dx$ approximated with quadrature as before. The nonlinear term

$$\left[\underline{C}(\underline{U})\right]_i = \int_0^1 \left(\sum_{j=1}^{N} U_j\,\varphi_j(x)\right)^3 \varphi_i(x)\,dx \qquad (4)$$

Now in general, this term can be written as a sum of terms of total power 3 of elements of $\underline{U}$ with known coefficients. The only elements that appear in the sum are the ones that overlap with $\varphi_i(x)$. This is shown explicitly for the case of piecewise linear elements below, where (4) becomes

$$\left[\underline{C}(\underline{U})\right]_i = \int_0^1 \left[U_{i-1}\,\varphi_{i-1}(x) + U_i\,\varphi_i(x) + U_{i+1}\varphi_{i+1}\right]^3 \times \varphi_i(x)\,dx.$$

The integrand above has nine terms when the cube is expanded:

$$U_{i-1}^3\,\varphi_{i-1}^3\,\varphi_i + 3\,U_{i-1}^2\,U_i\,\varphi_{i-1}^2\,\varphi_i^2 + \cdots$$

↑ U_j's are numbers, φ's are $\varphi(x)$.

So $\left[\underline{C}(\underline{U})\right]_i$ has the form

$$\sum_{\substack{0\le\alpha,\beta,\gamma\le 3 \\ \alpha+\beta+\gamma=3}} L_{\alpha\beta\gamma}\,U_{i-1}^{\alpha}\,U_i^{\beta}\,U_{i+1}^{\gamma}$$

where $L_{\alpha\beta\gamma} = \int_0^1 \varphi_{i-1}^{\alpha} \varphi_i^{\beta+1} \varphi_{i+1}^{\gamma} dx$ are constants that can be evaluated analytically or exactly with quadrature of sufficiently high order.

Now rewrite (3) as a root finding problem for a nonlinear vector system:

$$\underline{N}(\underline{U}) = \underline{0} \tag{4}$$

with $\underline{N}(\underline{U}) = A\,\underline{U} + \underline{C}(\underline{U}) - \underline{F}$

We can use Newton's method to iterate to a solution of (4), with $\underline{U}^{(m)}$ the m'th iterate. First compute the residual

$$\underline{R}^{(m)} = \underline{N}(\underline{U}^{(m)}) \quad \text{(we want } \underline{R}^{(m)} \to \underline{0}\text{)}.$$

Then compute the Jacobian (sensitivity) matrix of $\underline{N}(\underline{U})$ evaluated at the current estimate $\underline{U}^{(m)}$:

$$J_{ij} = \frac{\partial N_i}{\partial U_j} = A_{ij} + C_{ij}(\underline{U}^{(m)})$$

where $C_{ij}(\underline{U}) = \dfrac{\partial C_{ij}}{\partial U_j} =$

$$\begin{cases} \sum L_{\alpha\beta\gamma}\,\alpha\, U_{i-1}^{\alpha-1} U_i^{\beta} U_{i+1}^{\gamma} & \text{if } j = i-1 \\ \sum L_{\alpha\beta\gamma}\,\beta\, U_{i-1}^{\alpha} U_i^{\beta-1} U_{i+1}^{\gamma} & \text{if } j = i \\ \sum L_{\alpha\beta\gamma}\,\gamma\, U_{i-1}^{\alpha} U_i^{\beta} U_{i+1}^{\gamma-1} & \text{if } j = i+1 \\ 0 \text{ otherwise.} \end{cases}$$

Notice that $\mathbb{J}$ has the same sparsity structure as our original A.

Now solve

$$\mathbb{J}\,\underline{V}^{(m)} = -\,\underline{R}^{(m)}$$

and use $\underline{V}^{(m)}$ as the increment to the next iterate

$$\underline{U}^{(m+1)} = \underline{U}^{(m)} + \underline{V}^{(m)}$$

The usual numerical convergence criteria is that $\|\underline{R}^{(m)}\|$ is smaller than a certain user defined tolerance.

Note: like in 1D, Newton's method for systems finds $\underline{U}^{(m+1)}$ as the root of the system linearized around the current estimate of the root, $\underline{U}^{(m)}$.

If you have never seen Newton's method for systems before, please refer to any vector calculus textbook.

Like in 1D, Newton's method will in general only converge if $\underline{U}^{(0)}$, the initial estimate for the root, is accurate enough. To help find good initial estimates, a technique called continuation can be used. Here, a solution to $\underline{N}(\underline{U}) = 0$ is desired and the

solution to a related problem $\underline{\mathcal{E}}(\underline{U}) = \underline{0}$ can be found easily (maybe $\underline{\mathcal{E}}(\underline{U}) = \underline{0}$ is linear for example). Consider solutions $\underline{U}(\theta)$ to

$$\theta \underline{N}(\underline{U}) + (1-\theta)\,\underline{\mathcal{E}}(\underline{U}) = 0.$$

So $\underline{U}(0)$ solves the easy problem and $\underline{U}(1)$ the desired problem. Starting with $\theta = 0$, θ can be increased (adaptively) to $\theta = 1$. Solutions with only slightly changed θ values are assumed to be close, so Newton's method will always converge.

Math 521, Spring 2013
Notes XII, on time stepping

We will consider time stepping schemes first for two model time-dependent problems for a scalar $u(t)$:

$$\frac{du}{dt} = f(u,t), \quad u(0) = u_0 \qquad (1)$$

$$\frac{du}{dt} = \lambda u, \quad u(0) = 1. \qquad (2)$$

where the more general equation (1) is used to assess the accuracy of the schemes and (2) is used to assess its stability. Think of λ as being real and negative (and possibly large) which will correspond to discretizations of parabolic PDE's.

Consider approximating $u(t)$ on a regular grid in time with interval size $k = \Delta t$.

$$U^n \approx u(nk).$$

The simplest method for approximating (1) is the explicit Forward Euler method

$$U^{n+1} = U^n + k f(U^n, nk). \qquad (3)$$

This is first order accurate in the sense that if (3) is used with N time steps of size $k = T/N$ to compute out to time T, the error, $|u(T) - U^N| = O(k) = O(1/N)$.

Note that if the exact solution is put into (3)

$$u((n+1)k) - u(nk) - k f(u(nk), nk) = \frac{k^2}{2}\ddot{u}(nk) + O(k^3) = k\underbrace{\left(\frac{k}{2}\ddot{u} + O(k^2)\right)}_{T} \quad (4)$$

so the local error is $O(k^2)$. Thus, the error after one time step is $O(k^2)$, the error after $O(\frac{1}{k})$ time steps is $O(k)$ which makes sense. The local error is often written as kT, where T is the truncation error.

Theorem 1 (ODE theory). If $f(u,t) \in C^n$ then $u(t)$ exists, is unique and in C^{n+1} in a neighborhood of $t=0$.

Theorem 2 (Dahlquist) Any "reasonable" one-step time stepping scheme converges to the exact solution where it is defined with an order of convergence equal to the order of the truncation error of the scheme.

Notes: a one-step scheme is one in which the solution at $t=nk$ determines the solution at $t=(n+1)k$.

"reasonable" schemes are ones which are consistent and have a basic stability

result.

Since all of the schemes presented below are "reasonable" this theorem does not help us choose a scheme suitable to our problem.

Consider FE (3) applied to (2),

$$U_0 = 1, \quad U^{n+1} = (1 + k\lambda) U^n,$$

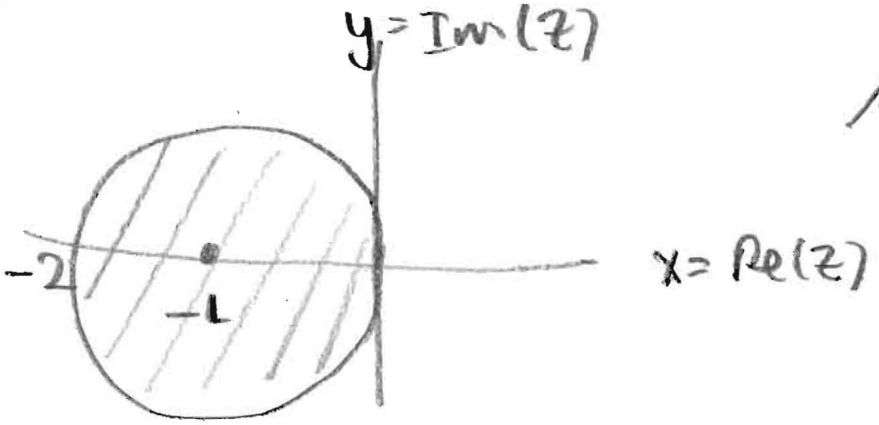
so $U^n = (1+z)^n$ with $\boxed{z = k\lambda}$

The scheme has a so-called growth factor of $G(z) = 1 + z$.

We can consider the set of z such that $|G(z)| \leq 1$. This is known as the stability region of the method. For FE,

$$|G(z)| = |1 + z| \leq 1$$

is a circle of radius 1 in the complex plane centered at $z = -1$:



stability region for forward euler.

So if λ is real and negative, the solution should decay, but the FE solution will only decay if

$$|z| = |k\lambda| < 2,$$

$$k < \frac{2}{|\lambda|} \tag{5}$$

Note that this does not violate Dahlquist's theorem since as $k \to 0$, (5) is eventually satisfied.

FE (3) is called an explicit method since U^{n+1} can be calculated explicitly in terms of U^n. A second order explicit method called Improved Euler is given below

$$U^* = U^n + k\,F(U^n, nk)$$

$$U^{n+1} = U^n + \frac{k}{2}\left(F(U^n, nk) + F(U^*, (n+1)k)\right)$$

This is a two-stage method, but still one-step since U^n does determine U^{n+1}. It is one of a family of second order Runge-Kutta methods. Writing it compactly as (6)

$$U^{n+1} = U^n + \frac{k}{2}\left(F(U^n, nk) + F\left(U^n + kF(U^n, nk), (n+1)k\right)\right)$$

makes it easy to determine the truncation error.

$$\tau = \frac{1}{k}\left[U^{n+1} - U^n - \frac{k}{2}\left(f + f(U^n + kf, (n+1)k)\right)\right]$$

($f(U^n, nk)$ and $f(U^n + kf, (n+1)k)$)

expanding in a Taylor series about $t = nk$.

$$\tau = \frac{1}{k}\left[k\dot{u} + \frac{k^2}{2}\ddot{u} + \frac{k^3}{6}\dddot{u} - \frac{k}{2}\dot{u} + O(k^4)\right.$$

$$\left. - \frac{k}{2}\left(\dot{u} + k(f_u f + f_t) + \frac{k^2}{2}\left(f_{uu}f^2 + f_{tt} + f_{ut}f\right)\right)\right]$$

where $f = \dot{u}$ is used and all quantities are evaluated at $t = nk$.

Starting with $\dot{u} = f(u,t)$, $\ddot{u} = f_u \dot{u} + f_t = f_u f + f_t$

so it can be seen that the truncation error above is second order, with dominant term

$$\tau = k^2\left(\frac{1}{6}\ddot{u} - \frac{1}{4}\left(f_{uu}f^2 + f_{tt} + f_{ut}f\right) + O(k^3)\right.$$

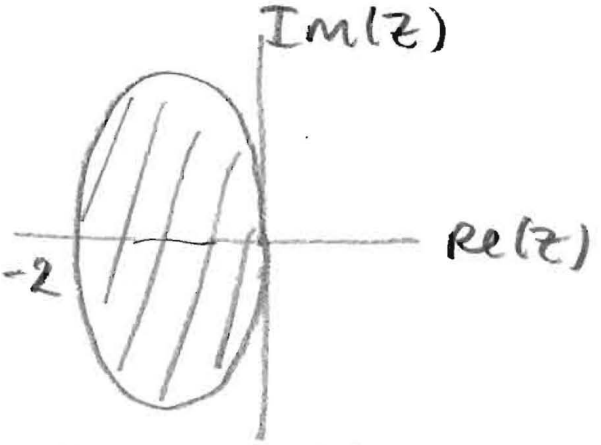
using $\dddot{u} = f_{uu}f^2 + 2f_{ut}f + (f_u)^2 f + f_u f_t + f_{tt}$

it can be seen that the dominant term in the truncation error is not a simple time derivative of u, but a different mix of f and its partial derivatives. This has implications for error estimation as we shall see in later discussion.

We can consider the stability of the improved euler scheme by considering the form (6) with $f(u,t) = \lambda u$, obtaining

$$U^{n+1} = \underbrace{\left(1 + z + \frac{z^2}{2}\right)}_{G(z)} U^n$$

The stability region $\{z : |G(z)| < 1\}$ is shown below



As with forward euler, it is seen that this second order explicit scheme is not suitable for problems with λ large in magnitude and negative (stiff problems)

Theorem The stability region of every explicit scheme is bounded.

Thus, no explicit scheme is suitable for stiff problems.

Let us consider what behaviour we would want in a time-stepping scheme, besides accuracy.

Consider (2) with $\mathrm{Re}(\lambda) < 0$. In this case, the solution should decay. If we wanted to have this property preserved by the time-stepping scheme, then $|G(z)| < 1$ for every $z = k\lambda$ with $\mathrm{Re}(z) < 0$. In other words, the stability region for the method should contain the left half plane.

Definition A time-stepping scheme is called L-stable if $|G(z)| < 1$ for all z with $\mathrm{Re}(z) < 0$.

Theorem No explicit scheme is L-stable.

Thus, we turn to the consideration of implicit schemes. The simplest scheme is the first order Backward Euler scheme:

$$U^{n+1} = U^n + kf(U^{n+1}, (n+1)k).$$

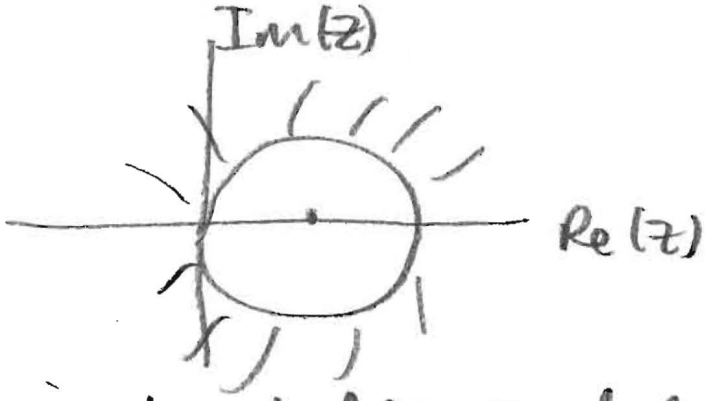
It is called an implicit method because U^{n+1} is determined by an implicit formula. A linear or nonlinear system must be solved for U^{n+1} at each time step. It has truncation error

$$T = -\frac{k}{2}\ddot{u}(nk) + O(k^2) \qquad (7)$$

and growth factor

$$G(z) = \frac{1}{1-z} \qquad (8)$$

and so has a stability region that is outside the unit circle centered at $z=1$.



thus, it is L-stable and suitable for stiff problems.

Let us now consider three second order implicit methods, each with some advantages and some disadvantages.

(I) Implicit midpoint rule

$$U^{n+1} = U^n + k\left(F\left((U^{n+1}+U^n)/2, (n+\tfrac{1}{2})k\right)\right)$$

Trapezoidal Rule

$$U^{n+1} = U^n + \frac{k}{2}\left(F(U^n, nk) + F(U^{n+1}, (n+1)k)\right)$$

These are different schemes but become identical when applied to linear problem In addition, they have the same dominant error term and the same growth factor $G(z)$ and hence the same stability region.

$$G(z) = \frac{1+z/2}{1-z/2} \qquad (9)$$

Here, the stability region is exactly the left half plane so the method is L-stable. Despite this fact, and the fact that the method has a small error coefficient, this method should be used on stiff problems with caution. Consider

$$G(z) \approx e^z$$

as $Re(z) \to -\infty$, $e^z \to 0$. a scheme that matches this property is called A-stable.

Definition a scheme is A-stable if

$$G(z) \to 0 \text{ as } Re(z) \to -\infty.$$

Important Note: In some literature, the "A" and the "L" definitions are reversed! There were two schools that could not agree on notation. I like "L" for left-half plane and "A" for asymptotic.

Considering (9), $\lim_{Re(z)\to-\infty} \frac{1+z/2}{1-z/2} = -1.$

Thus, these schemes are not A-stable. Thus, some components of the solution

that should practically disappear after one time step instead just oscillate in sign.

Backward euler is A-stable as can be seen from (8).

(II) Second order A-stable & L-stable Diagonally Implicit Runge-Kutta (DIRK) method.

This is a two stage method. $\alpha = 1 - \frac{\sqrt{2}}{2}$

$$U^* = U^n + k\alpha f(U^*, t_n + \alpha k)$$

$$U^{n+1} = U_n + k\left(\alpha f(U^{n+1}, t_n + k) + (1-\alpha) f(U^*, t_n + \alpha k)\right)$$

The terms "diagonal" applies since each stage is only implicit in the value at that stage.

This is a one-step scheme with good properties, but is little used compared with the next scheme.

(III) Second order Backward Differentiation Formula (BDF-2). This is a multi-step method, involving the values of two previous steps, U^{n-1} and U^n, to compute

U^{n+1}:

$$U^{n+1} = \frac{4}{3}U^n - \frac{1}{3}U^{n-1} + \frac{2k}{3}f(U^{n+1}, (n+1)k).$$

It is based on the second order one sided difference formula

$$\ddot{u}((n+1)k) \approx \frac{\frac{3}{2}U^{n+1} - 2U^n + \frac{1}{2}U^{n-1}}{k}$$

It has truncation error

$$\frac{2}{9}k^2 \dddot{u}((n+1)k) + O(k^3).$$

and so, like BE, its dominant error is a simple time derivative of u. To analyze the stability, consider the method applied to (2),

$$U^{n+1} = \frac{4}{3}U^n - \frac{1}{3}U^{n-1} + \frac{2}{3}z U^{n+1}$$

or

$$\left(1 - \frac{2z}{3}\right)U^{n+1} - \frac{4}{3}U^n + \frac{1}{3}U^{n-1} = 0.$$

This is a second order constant coefficient difference equation with solution

$$U^n = A G_1^n + B G_2^n$$

these n's are powers

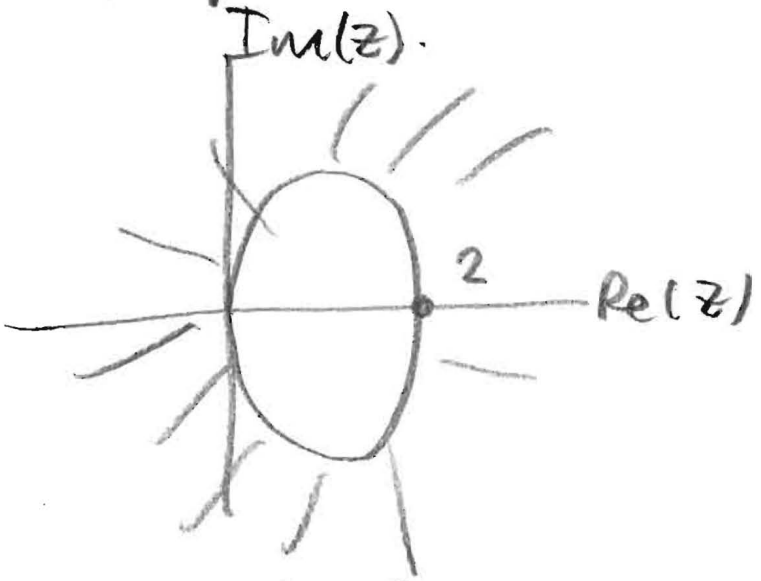
for some constants A & B and $G_1(z), G_2(z)$ roots of

$$\left(1-\frac{2z}{3}\right)G^2 - \frac{4}{3}G + \frac{1}{3} = 0. \qquad (10)$$

The stability region in this multi-step method case is

$$\{z: \ |G_1(z)| < 1 \text{ and } |G_2(z)| < 1\}.$$

It can be shown that the stability region is shaped as shown below



So BDF-2 is L-stable. Since

$$G_{1,2} = \frac{2/3 \pm \sqrt{\frac{4}{9} - \frac{1}{3}\left(1-\frac{2z}{3}\right)}}{(1-2z/3)}$$

it is also clear that it is A-stable ($|G_{1,2}(z)| = O\left(\frac{1}{|z|^{1/2}}\right)$ in this limit). BDF-2 is less accurate than DIRK-2 per time time step, but more accurate

per implicit solve. Thus, it is used much more often in practice than DIRK-2. However, it is nontrivial to implement adaptive time step strategies, since the formula assumes that U^{n-1}, U^n and U^{n+1} have the same spacing in time.

Considering (10) in the limit as $z \to 0$ (fixed λ, $k \to 0$), we have

$$G_1 \approx 1, \quad G_2 \approx 1/3.$$

The term G_2^n appears as an initial layer that accounts for the error from the initialization procedure for U^1. Since this is only one step, BE can be used to compute U^1 and second order accuracy overall is maintained.

One final method to consider is the third order, 2 stage, Radau-IIA method. It is both L-stable and A stable.

$$U^* = U^n + k\left(\frac{5}{12} f(U^*, t_n + \frac{k}{3}) + \frac{1}{12} f(U^{n+1}, t_{n+1}\right)$$

$$U^{n+1} = U^n + k\left(\frac{3}{4} f(U^*, t_n + \frac{k}{3}) + \frac{1}{4} f(U^{n+1}, t_{n+1})\right)$$

Note that this scheme is implicit in U^* and U^{n+1} simultaneously, and that the Jacobian is not symmetric.

The coefficients of one-step methods can be represented in Butcher Tables. The general s stage method (s function evaluations) is given by

$$U^{n+1} = U^n + \sum_{i=1}^{s} b_i K_i$$

$$K_i = k f\left(t_n + c_i k, U^n + \sum_{j=1}^{s} a_{ij} K_j\right)$$

This information can be entered into a table:

$$\begin{array}{c|ccc} c_1 & a_{11} & a_{12} \cdots & a_{1s} \\ c_2 & a_{21} & & \\ \vdots & \vdots & & \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & b_2 \cdots & b_s \end{array} \quad \Big\} \; /\!/A$$

An explicit method has zeros in the diagonal of A and above. A diagonally implicit method has nonzeros on the diagonal but zeros above it. The tables for some of the schemes so far are shown below:

Improved Euler

0	0	0
1	1	0
	1/2	1/2

DIRK-2

α	α	0
1	$1-\alpha$	α
	$1-\alpha$	α

$$\alpha = 1 - \frac{\sqrt{2}}{2}$$

Radau II-A

1/3	5/12	-1/12
1	3/4	1/4
	3/4	1/4

We can also discuss adaptive time stepping with a particular example. Consider applying BE to a problem and wanting to make the error at each time step smaller than a user-defined tolerance, δ. The local error is (7)

$$kT \approx -\frac{k^2}{2}\ddot{u}(nk)$$

to highest order. Recall that the local error for FE is (top of p. 2)

$$+\frac{k^2}{2}\ddot{u}(nk)$$

So if we compute the FE solution U^{FE} as well as the BE solution U^{n+1}, the local error in U^{n+1} is approximately

$$e \approx \frac{1}{2} |U^{n+1} - U^{FE}|.$$

Note that U^{FE} is cheap to compute (an explicit method) and may be useful to give a good initial guess to iterative solvers for the implicit problem for U^{n+1}.

If $e > \delta$ we would fail the time step and recompute U^{n+1} with a reduced time step k. If $e < \delta$ we accept the time step, and compute a new time step so that $e < \delta$ will be satisfied again. Since

$$e \approx C k_{old}^2,$$

and we want $e < \delta$, we could take k_{new} with

$$C k_{new}^2 = \frac{e}{k_{old}^2} k_{new}^2 < \delta$$

In practice, the formula

$$k_{new} = \theta \sqrt{\frac{\delta}{e}} k_{old}.$$

with $\theta < 1$ a "safety factor", $\theta = 0.8$ is my favourite.

On a failed step for which e is not that much bigger than δ, this formula can also be used. If it is a "bad" failure, k is typically drastically reduced (i.e. factor of 2). Note that single step methods have more complicated error structure and generic error assessment using (11) higher order approximation

---

Consider a model time-dependent problem for $u(x,t)$, 1-periodic in x.

$$u_t = u_{xx} - a u + f(x,t)$$

with $u(x,0) = u_0(x)$ given.

Consider discretizing in space only to begin, with a FEM.

$$U(x,t) = \sum_{j=1}^{N} U_j(t)\, \varphi_j(x).$$

a weak formulation of (11) is obtained by multiplying by test functions $\varphi(x)$. Similarly, the FEM is obtained by using the form above and testing with $\varphi_j(x)$. The following results

$$M \dot{\underline{U}} = -K \underline{U} - aM \underline{U} + \underline{F}(t) \qquad (12)$$

with $F_j(t) = \int_0^1 f(x,t)\, \varphi_j(x)\, dx$, and M and

K the usual mass and stiffness matrices

$$M_{ij} = \int_0^1 \varphi_i(x)\,\varphi_j(x)\,dx$$

$$K_{ij} = \int_0^1 \varphi_i'(x)\,\varphi_j'(x)\,dx$$

(12) is a semi-discretization or method of lines. It is an ODE system to which the time stepping methods of the previous sections can be applied. Eigenvalues of the system are

$$\lambda \in \left[-\frac{c}{h^2}, \ldots -a\right].$$

with a constant depending on the exact FEM used. This makes (12) a stiff problem and explicit methods require

$$k < \tilde{C} h^2$$

for stability, which is often prohibitively small. In addition, the mass matrix in front of $\dot{\underline{U}}$ naturally leads us to implicit methods. BE computations would be

$$\underbrace{(M + kK + akM)}_{\text{positive definite}}\,\underline{U}^{n+1} = M U^n + k\,\underline{F}(t_{n+1})$$

and other implicit methods have similar structure.

Notes XIII on incompressible flow
Math 521, Spring 2013.

The equations of incompressible flow (and the equations of elastic deformation of incompressible materials) have some differences in structure to what we have seen before. We'll consider the simplest case, that of 2D Stokes flow. Here the unknowns are the two components of fluid velocity

$$\underline{u}(\underline{x}) = (u(x,y), v(x,y))$$

and the scalar pressure $p(x,y)$.

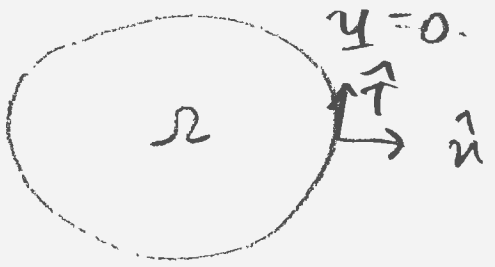
$$\left.\begin{aligned} -(u_{xx} + u_{yy}) + p_x &= f_1(x,y) \\ -(v_{xx} + v_{yy}) + p_y &= f_2(x,y) \end{aligned}\right\} \quad (1)$$

$$u_x + v_y = 0.$$

The first two equations express a force balance and the last expresses the incompressibility of the fluid. The equations can be written in vector form

$$\left.\begin{aligned} -\Delta \underline{u} + \nabla p &= \underline{f} \\ \nabla \cdot \underline{u} &= 0. \end{aligned}\right\} \quad (2)$$

The equations are well posed in a bounded domain Ω when homogeneous Dirichlet conditions are given on the boundary.



In this application $\underline{u}\cdot\hat{n}=0$ is known as the "no flow" condition and $\underline{u}\cdot\hat{\tau}$ is the "no slip" condition. Non-homogeneous Dirichlet conditions and other boundary conditions are of interest but not considered in these notes.

Note that there are no boundary conditions for the pressure (none are needed). Note also that since p only appears in the term ∇p it is arbitrary up to a constant. You can fix the constant by requiring

$$\int_\Omega p = 0 \tag{3}$$

for example. p can be considered as a Lagrange multiplier for the incompressibility constraint.

---

aside: Consider the simple problem for $u(x)$, 1-periodic,

$$-u'' = f(x). \tag{4}$$

Integrating (4) and using the periodicity of u', we see that $\int_0^1 f(x)\,dx = 0$ for there to

be a solution. There is a solution $u(x)$ to the problem in this case, unique up to a constant. This is an example of the Fredholm alternative in this infinite dimensional setting. The solution can be made unique if we require

$$\int_0^1 u(x)\,dx = 0,$$

analagously to (3). Consider now a finite difference approximation of (4) on a usual regular grid with spacing h.

$$-D_2 \underline{U} = \underline{F}. \tag{5}$$

Using a summation by parts formula it can be shown that

$$\sum_i F_i = 0$$

for there to be a solution. If $\int_0^1 F(x)\,dx = 0$ this is not exactly satisfied, but there is only an $O(h^2)$ residual. By taking the projection onto the zero mean subspace that is replacing $\underline{F}$ in (5) by

$$\tilde{\underline{F}} = \underline{F} - \left(\frac{\sum_{i=1}^{N} F_i}{N}\right) \tag{6}$$

average of $\underline{F}$ values.

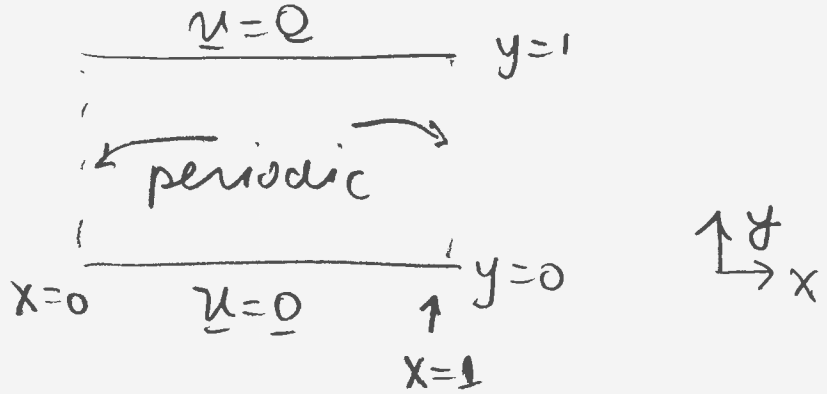
we can guarantee a solution exists. Since (5) does not have a solution, the system is not full rank, but rank $N-1$ with the

1D nullspace of the constant vectors. Using (6) however, we can remove any row of (5) and replace it with

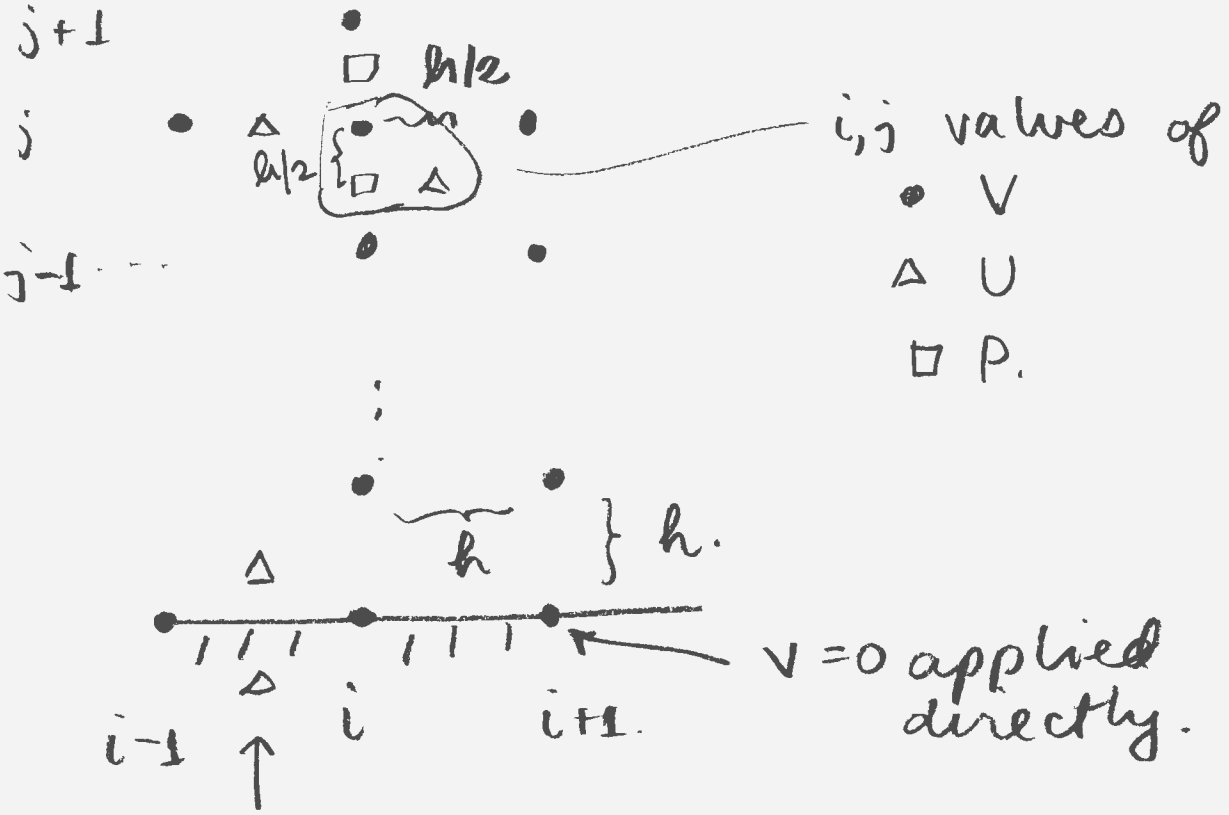
$$\sum_{i=1}^{N} U_i = 0$$

which approximates the normalization condition $\int_0^1 u(x)\,dx = 0$. Any equation of (5) can be replaced and the resulting solution $\underline{U}$ is the same.

Consider now a finite difference approximation of (2) on the so-called Marker-and-Cell grid shown below. We consider the domain as the unit square with solutions periodic in x and homogeneous Dirichlet conditions on the top and bottom sides $y=0$ and $y=1$. This is the simplest geometry to consider that has boundaries.



In the Marker-and-Cell grid (MAC) the quantities u, v and p are given at locations staggered from each other.



v requires ghost point approximation of $u=0$ boundary conditions.

Considering (1), this grid allows all first derivatives to be calculated using short centered differences.

$$-\tilde{\Delta}_h U_{ij} + D_{+,x} P_{ij} = F_{1,ij}$$

with modified stencil at the boundary due to ghost points

$$-\Delta_h V_{ij} + D_{+,y} P_{ij} = F_{2,ij}$$

$$D_{-,x} U_{ij} + D_{-,y} V_{ij} = 0.$$

This can be summarized as

$$\begin{pmatrix} -\tilde{\Delta}_h & & D_{+,x} \\ & -\Delta_h & D_{+y} \\ D_{-,x} & D_{-y} & 0 \end{pmatrix} \begin{pmatrix} \underline{U} \\ \underline{V} \\ \underline{P} \end{pmatrix} = \begin{pmatrix} \underline{F_1} \\ \underline{F_2} \\ \underline{0} \end{pmatrix}$$

or using a vector form for the velocities,

$$\begin{pmatrix} -\Delta_h & G \\ G^T & 0 \end{pmatrix} \begin{pmatrix} \underline{U} \\ \underline{P} \end{pmatrix} = \begin{pmatrix} \underline{F} \\ \underline{0} \end{pmatrix} \qquad (7)$$

as seen in the representation above, it can be shown that the system is symmetric. Unfortunately, it is not positive definite: it has both positive and negative eigenvalues. Thus, the conjugate gradient method cannot be applied directly. There are many approaches to the fast, numerical solution of (7). One of them is called the pressure equation (PE) approach shown below. Formally, the first block equation in (7) can be written

$$-\Delta_h \underline{U} + G\underline{P} = \underline{F}.$$

so $$\underline{U} = \Delta_h^{-1}(G\underline{P} - \underline{F}) \qquad (8)$$

then the second block of (7) reads

$$G^T \Delta_h^{-1} G\underline{P} = G^T \Delta_h^{-1} \underline{F}. \qquad (9)$$

The matrix $G^T \Delta_h^{-1} G$ is symmetric and positive semi-definite (there is always a null space for the pressure). It has an $O(1)$ condition number, so (9) can be solved with a CG approach in a few iterations. At every iteration "multiplication" by Δ_h^{-1} is done with a solve with matrix Δ_h, which can be done with an inner CG iteration for example.

Consider now a weak formulation of (2). We would take the inner product of the first equation with a test function $\underline{\varphi}(\underline{x})$ in $H_0^1 \times H_0^1$ and integrate, similarly the divergence free condition is tested with a function $\psi \in L_2 \backslash R$ (L_2 with mean 0).

$$\left.\begin{aligned} &\int \underbrace{\nabla \underline{\varphi} : \nabla \underline{u}}_{\nabla \varphi_1 \cdot \nabla u + \nabla \varphi_2 \cdot \nabla v} - \int (\nabla \cdot \underline{\varphi}) p = \int \underline{f} \cdot \underline{\varphi} \\ &\int (\nabla \cdot \underline{u}) \psi = 0 \end{aligned}\right\} \quad (10)$$

Note that $\int (\nabla \cdot \underline{u}) \psi = 0$ for $\psi \in L_2 \backslash R$ only makes $\nabla \cdot \underline{u}$ be a constant. However, since

$$\int_\Omega \nabla \cdot \underline{u} = \int_\Gamma \underline{u} \cdot \hat{n} \, ds = 0$$

Green's identity

this constant must be zero by the boundary conditions for $\underline{u}$.

Consider $\mathbb{J} = \{ \underline{u} \in H_0^1 \times H_0^1 \mid \nabla \cdot \underline{u} = 0 \}$. Such $\underline{u}$ are already divergence free. If we take $\underline{\varphi} \in \mathbb{J}$ also,

$$\int \nabla \underline{\varphi} : \nabla \underline{u} = \int \underline{f} : \underline{\varphi} \qquad \forall\, \underline{\varphi} \in \mathbb{J} \tag{11}$$

This makes the Stokes equations look like the Poisson problem in a different space. Our same argument can be used to show that this problem has a ! solution $\underline{u} \in \mathbb{J}$. If (10) has a solution $(\underline{u}, p)$ then the $\underline{u}$ that solves (11) must be the same. We would also need (12)

$$(\nabla \cdot \underline{\varphi}, p) = (\nabla \underline{\varphi} : \nabla \underline{u}) - (\underline{f}, \underline{\varphi}) \quad \forall\, \underline{\varphi} \in H_0^1 \times H$$

It is already true for $\underline{\varphi} \in \mathbb{J}$ by assumption. Note that

$$\|\underline{u}\|_{H_0^1} \leq C \|\underline{f}\|_{H^{-1}}$$

from (11), so from (12) we can determine that

$$|(\nabla \cdot \underline{\varphi}, p)| \leq \|\underline{\varphi}\|_{H_0^1} \|f\|_{H^{-1}}. \tag{13}$$

To proceed, we need the following result known as the "inf-sup" or Babuska-Brezzi (BB) condition. It can be shown for suitable domains Ω.

Defn The domain Ω satisfies the BB condition if for every $p \in L_2 \backslash R$ there exists

a nonzero $\underline{\varphi} \in H_0^1 \times H_0^1$ such that

$$|(p, \nabla \cdot \underline{\varphi})| \geq \alpha \|\varphi\|_{H_0^1} \|p\|_{L_2 \backslash R}.$$

with α independent of p. This can be stated

$$\inf_{p \in L_2 \backslash R} \sup_{\varphi \in H_0^1 \times H_0^1} \frac{(p, \nabla \cdot \underline{\varphi})}{\|p\|_{L_2 \backslash R} \|\varphi\|_{H_0^1}} = \alpha > 0.$$

If this condition is satisfied, then by choosing the $\underline{\varphi}$ guaranteed by the theorem in (13) we can show that

$$\|p\|_{L_2 \backslash R} \leq \frac{1}{\alpha} \|F\|_{H^{-1}}$$

This shows that solutions of (12) if they exist are unique and bounded by the data. There is still some work to show that solutions exist in this infinite dimensional setting.

We can turn now to a FEM approximation of (10). Pick subspaces S^N of $H_0^1 \times H_0^1$ and B^M of $L_2 \backslash R$ for $\underline{U}(x,y)$ and $p(x,y)$ and take corresponding subspaces for the test functions

$$K\underline{U} - GP = \underline{F} \qquad K_{ij} = \int \nabla \underline{\varphi}_i : \nabla \underline{\varphi}_j$$

$$G^T \underline{U} = 0 \qquad G_{ij} = \int (\nabla \cdot \varphi_i) \psi_k.$$

We could consider

$$J_h = \{ \underline{U} \in S^N : G^T \underline{U} = \underline{0} \}$$

as an approximation of J, but for all low order elements, J_h does not have a local basis.

Analagously to the continuous BB condition, there is one necessary in the discrete setting; necessary for solvability and convergence:

$$\inf_{q \in B^M} \sup_{\underline{\varphi} \in S^N} \frac{(q, \nabla \cdot \underline{\varphi})}{\|q\|_{L_2 \backslash \mathbb{R}} \|\underline{\varphi}\|_{H_0^1}} \geq \alpha$$

with α independent of h in a suitably good grid.

Rather than show the proofs that some elements satisfy this condition and <u>some do not</u>, let me show you an example of an S^N and B^M pair that fail this criteria, and the consequences. This is a simple example, but an extreme case.

Consider a triangulated 2D domain with P_1 approximation of $\underline{U}$ (piecewise linear, conforming in $H^1 v$) and P_0 approximation of p (piecewise constant, fine for $L_2 v$). Let t be the number of triangles in the mesh, v_I the number of interior nodes and v_B the number of boundary vertices. Consider the

dimension of J, using the geometrical result, Euler's relationship

$$t = 2V_I + V_B - 2.$$

$$\dim(J) \leq \underbrace{2V_I}_{\text{velocity unknown (2 components) at each interior node.}} - \underbrace{(t-1)}_{\text{rank of } G^T \text{ is the number of degrees of freedom for } p \text{ - which has a value in each triangle but must integrate to zero.}} = 3 - V_B.$$

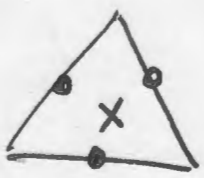
So as long as $V_B \geq 3$ (always true), $J = \{\underline{0}\}$. Thus, this discretization always computes $\underline{U} \equiv \underline{0}$. This is an extreme example of the failure of the BB condition.

Some 2D elements that do satisfy the BB condition are listed below.

$\underline{U}$ P

diagram showing V unknowns as dots, P as crosses.

nonconforming P_1 P_0



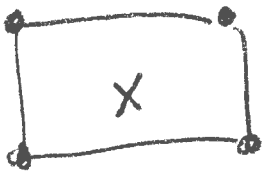
Note that such elements are not conforming (approximation spaces not subsets of H_0^1) and

additional techniques are needed to show their convergence.

U P diagram.

non conforming

$\widetilde{Q}_1 = \text{span}\{1, x, y, x^2 - y^2\}$ → P_0

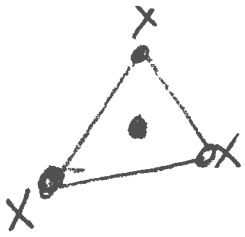


conforming

P_1 + "cubic bubble" → P_1

↑

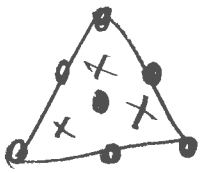
cubic function with value 1 at centre, zero along all edges.



"MINI element"

Crouzeix-Raviart P_1 non conforming

P_2 + bubble



Consider the Navier-Stokes equations in 2D: $\underline{u} = (u, v)$, parameter $Re > 0$

$$\left.\begin{aligned} -\frac{1}{Re}\Delta u + uu_x + vu_y + p_x &= F(x,y) \\ -\frac{1}{Re}\Delta v + uv_x + vv_y + p_y &= g(x,y) \end{aligned}\right\} (1)$$

$$u_x + v_y = 0.$$

In vector form, these are written

$$-\frac{1}{Re}\Delta \underline{u} + \underline{u}\cdot\nabla\underline{u} + \nabla p = \underline{f}$$

$$\nabla\cdot\underline{u} = 0.$$

with notation $\underline{u}\cdot\nabla\underline{u} = (\underline{u}\cdot\nabla)\underline{u}$

$$= \left(u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y}\right)\underline{u}.$$

Note that using the divergence condition, the nonlinear terms can be written in divergence (conservation) form.

$$\left.\begin{aligned} &uu_x + vu_y \\ &= (u^2)_x + (uv)_y \\ &uv_x + vv_y \\ &= (uv)_x + (v^2)_y. \end{aligned}\right\} (2)$$

These terms are called convection terms and represent the transport of momentum with the flow.

The constant Re (Reynolds number) is a dimensionless parameter that remains after scaling (like the "a" in problem #1). The Δ terms represent viscous terms (momentum diffusion). So Re represents the relative size of convection compared to diffusive momentum transfer.

In 2D, it is known that smooth solutions exist and are unique, also for the time dependent problem.

In 3D, the existence and regularity of solutions is an open problem, one of the Millennium Problems for which there is a $1 Million prize.

For large Re, the flow becomes increasingly complex - turbulence begins at moderate Re and so the steady assumption is not valid in general. Spatial scales of $O(1/Re)$ occur in scaled length in turbulent flow. Often these are too fine to resolve numerically and empirical turbulence models are used to describe a macroscopic average flow.

assuming steady, moderate Re flow that can be resolved on a suitable mesh, approximating (1) or with the convection terms in conservation form (2) is done in a straight-forward way. The term

$$(u^2)_x + (uv)_y$$

is tested with a function $\varphi(\underline{x})$:

$$-\int_\Omega \{\varphi_x (u^2) + \varphi_y (v^2)\}$$

This is evaluated with quadrature with given discrete nodal values for u & v in a FE discretization, and the Jacobian can be found, and Newton's method used. Note: the resulting matrix is not symmetric.

Often an approximate Newton method is used, based on delaying the update of one of the quadratic factors.

$$[uu_x]^{(n)} \approx u^{(n-1)} u_x^{(n)}.$$

↑ iterate (n)

Considering (1) with $(\underline{u}\cdot\nabla)$ coming from a known $\underline{u}$ is called the Oseen equation.

Fluid flow is well understood and computational methods have been extensively used. Tricks that improve performance by even a few % are well received in the literature.

Let's look briefly at time-dependent incompressible flows: $\underline{u}(\underline{x}, t)$.

$$\left.\begin{aligned} \underline{u}_t + (\underline{u}\cdot\nabla)\underline{u} &= \frac{1}{Re}\Delta\underline{u} - \nabla p + \underline{f} \\ \nabla\cdot\underline{u} &= 0. \end{aligned}\right\} \quad (3)$$

From here, you can see what the convective terms are doing. Consider a particle moving with the flow on trajectories $\underline{x}(t)$ ← Lagrangian coordinates $\underline{x}(0)$.

$$\frac{d\underline{x}}{dt} = \underline{u}(\underline{x}(t), t).$$

acceleration in the first coordinate

$$\frac{d^2x}{dt^2} = \frac{d}{dt}\left(u(x(t), y(t), t)\right)$$

$$= \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} + \frac{\partial u}{\partial y}\frac{dy}{dt}$$

$$= \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y}.$$

Because of this, $\frac{\partial}{\partial t} + \underline{u} \cdot \nabla$ is called the material derivative.

We can discretize (3) implicitly and get a similar problem to the steady one (1) at each time step.

Variants:

- $(\underline{u} \cdot \nabla)\underline{u}$ can be handled explicitly (an IMEX method).
- $(\underline{u} \cdot \nabla)\underline{u}$ giving a linear
  ↑ explicitly ↑ implicitly
  solve for implicit time stepping.
- The incompressibility condition can be handled in a separate, "projection" step. I'll discuss this later in the week.

Often, fluid computations are under-resolved. Recall that spatial structures of size $O(\frac{1}{Re})$ can occur. The spatial discretizations described above (equivalent to centered finite difference methods) can lead to oscillations in the solutions

The way to overcome such oscillations are:

- locally refine the mesh.
- introduce a turbulence model and then describe the macroscopic average of the flow.
- borrow shock-capturing methods (flux limiting) techniques from compressible, inviscid flow.

Consider an evolution equation for $u(\underline{x}, t)$ in conservation (divergence) form.

$$u_t + \nabla \cdot \underline{F}(u) = 0 \quad \leftarrow \text{possibly a source term } g(\underline{x}, t, u) \qquad (1)$$

$\uparrow$ could also depend on the derivatives of $\underline{u}$.

For any volume V and its enclosing surface S, we can integrate over V and use the divergence theorem. (2)

$$\frac{d}{dt}\int_V u \, dV + \int_S \underline{F} \cdot \hat{n} \, dS = \int_V g \, dV.$$

$\int_V u \, dV$: Total amount of quantity u in V.

$\int_S \underline{F} \cdot \hat{n} \, dS$: Total flux out of V.

$\int_V g \, dV$: source term.

So this is a PDE that represents conservation of quantity u.

Finite volume methods are based on a discretization of (2). It is the integral form of (1), not quite the weak formulation but similar.

for a finite set of V's.

Problem #2 (well, the time dependent version whose number I have forgotten) is in this form, with

$\underline{F} = -\nabla u$, $g = -au + f(x)$.

The idea is that the mesh or grid defines control volumes. The integral of u on these volumes are the unknowns - like the first term in (2) suggests.

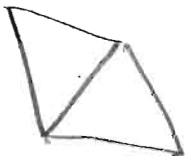
The boundaries of V (S) are simple (triangular pieces of planes in the simplest case) and there is an explicit formula for $\hat{n}$ on these pieces - constant in the planar boundary case.

To evaluate the second term in (2), $\underline{F}$ must be approximated at quadrature points on the pieces of S. This can be done by matching Taylor series using nearby volumes - an example is given below. This allows a lot of flexibility in the choice of mesh and the way that the fluxes are approximated. For example, upwinding or flux or slope limited fluxes can be computed for hyperbolic conservation laws.

The third term of (2) can be evaluated by quadrature.

Note that $\int_{S_i} \underline{F} \cdot \hat{n} \, ds$ is approximated just once, and the same value is used on both volumes or either

side of S_i, but with $\hat{n}$ reversed in sign.

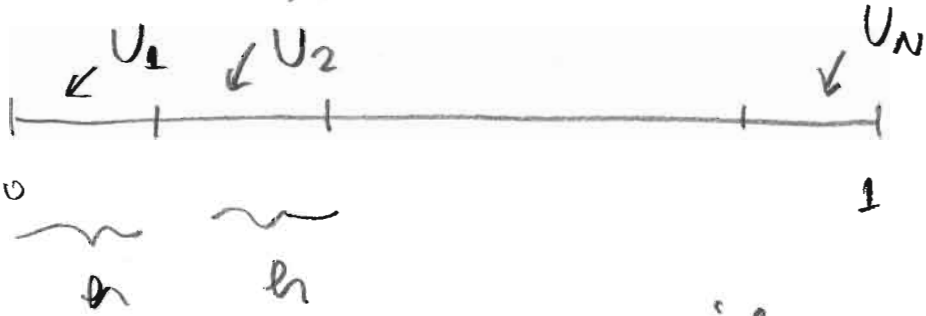


Thus, the approximation of (2) on elementary volumes is also consistent for any collection of these volumes.

Meshes can be rectangular or tetrahedrons or any mix. Higher order discretizations can be done using least squares solution (weighted) using more nearby points.
Let's look at a simple case - a uniform grid solution of problem #1.

$$-(u_x)_x + au = F(x)$$

$$(u_x)_x = au - F(x).$$



$h = 1/N$.

U_j approximates $\int_{(j-1)h}^{jh} u(x)\,dx$.

$$= h\,u(jh) - \frac{h^2}{2}u'(jh) + \frac{h^3}{3}u''(jh) + \cdots$$

$$U_{j+1} = h\,u(jh) + \frac{h^2}{2}u'(jh) + \frac{h^3}{3}u''(jh) + \cdots$$

aU_j is the correct cell average.

$$F_j = \int_{(j-1)h}^{jh} f(x)\,dx \approx h\, f((j-\tfrac{1}{2})h) + \frac{h^3}{24} f''((j-\tfrac{1}{2})h) + O(h^5)$$

can be approximated by midpoint rule.

$$\int_{(j-1)h}^{jh} (u_x)_x\,dx = u_x\Big|_{(j-1)h}^{jh}$$

these must be approximated. But

$$u_x\Big|_{jh} \approx \frac{U_{j+1} - U_j}{h^2} + O(h^2).$$

So $$\int_{(j-1)h}^{jh} (u_x)_x\,dx \approx \frac{U_{j+1} - U_j}{h^2} - \frac{U_j - U_{j-1}}{h^2} + O(h^2)$$

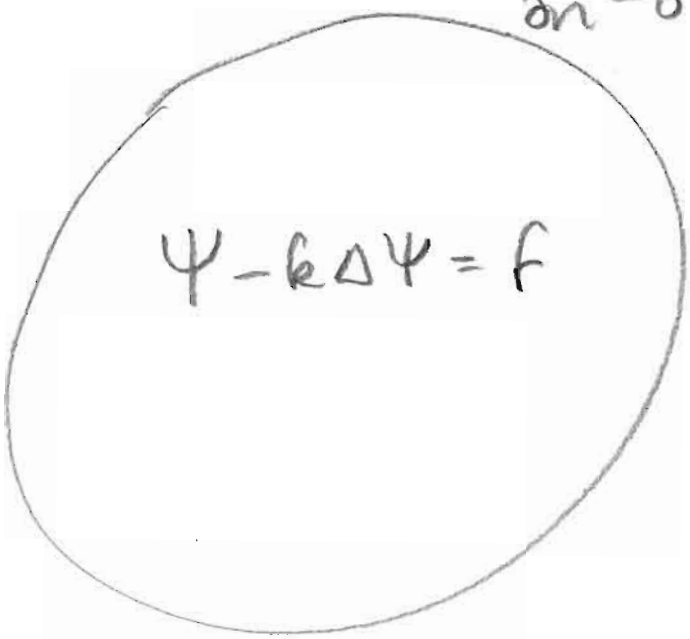
Now (2) needs on the interval $[(j-1)h, jh]$

$$\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} = aU_j - F_j$$

This is exactly our FD discretization on a staggered grid, with U_j & F_j scaled by h.

FV meshes can be constructed in a variety of ways, including Voronoi meshes.

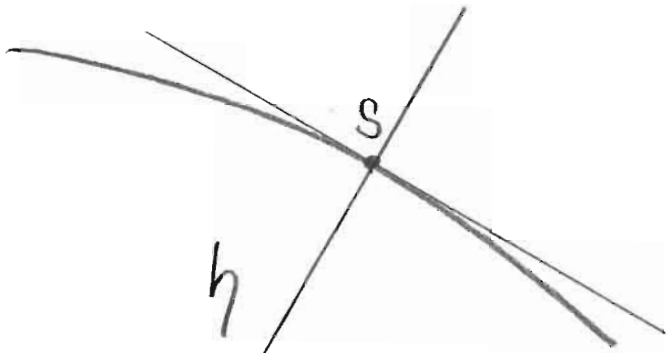
aside



k small (singular perturbation)

then $\psi = f$ if $k = 0$, but this does not satisfy the boundary conditions.

Consider s & η boundary fitted coordinates



Only defined near the boundary, but that is all we need.

$$\psi \approx f + \sqrt{k}\,\frac{\partial f}{\partial n}(s)\, e^{-\eta/\sqrt{k}}$$

$$(\mathbb{I} - k\Delta)\,\psi \approx f + \left(1 - k\frac{\partial^2}{\partial \eta^2}\right)\frac{\partial f}{\partial n}(s)\, e^{-\eta/\sqrt{k}}$$

$$\approx f \quad \text{to highest order}$$

$$-\underbrace{\frac{\partial}{\partial \eta}}_{\text{normal derivative}}\left(F + \sqrt{k}\,\frac{\partial f}{\partial n}(s)\, e^{-\eta/\sqrt{k}}\right)\Big|_{\eta=0} \quad \text{(boundary)}$$

$$= \frac{\partial f}{\partial n} - \frac{\partial f}{\partial n} = 0 \quad \checkmark$$

$$\left.\begin{aligned} \underline{u}_t &= \Delta \underline{u} - \nabla p + \underline{f} \\ \nabla \cdot \underline{u} &= 0 \end{aligned}\right\} \quad (1)$$

recall $A = \{ \underline{w} : \nabla \cdot \underline{w} = 0 \text{ and } \underline{w} \cdot \hat{n} = 0 \}$.

$B = \{ \underline{w} : w = \nabla \psi \text{ for some } \psi \}$.

Let $\mathbb{P}$ be the projection onto B.

$\mathbb{P}\underline{w} = \nabla \psi$, where ψ solves

$$\Delta \psi = \nabla \cdot \underline{w}$$

$$\frac{\partial \psi}{\partial n} = \underline{w} \cdot \hat{n}$$

Take ψ so that $\int_\Omega \psi = 0$.

$\mathbb{Q} = \mathbb{I} - \mathbb{P}$ projection onto A.

So (1) reads

$$\underline{u}_t = \mathbb{Q}\,\Delta \underline{u} + \underline{f} \qquad \underline{f} \in A \quad \text{(WLOG can take)}$$

$$p = \mathbb{P}\,\Delta \underline{u} \quad \text{(abuse of notation)}$$

Consistent BE

$$\left.\begin{aligned} \underline{U}^{n+1} &= \underline{U}^n + k\underline{f}^{n+1} + kQ\Delta\underline{U}^{n+1} \\ p^{n+1} &= \mathcal{P}\Delta\underline{U}^{n+1} \end{aligned}\right\} \text{(*)}$$

(Coupled problem for $\underline{U}^{n+1}$ and p^{n+1}). This has smooth error expansions.

Implicit projection method

$$\tilde{U} = U^n + k\Delta\tilde{U} + k\underline{f}^{n+1} \quad \text{"easy" problem for } \tilde{U}$$

$$U^{n+1} = Q\tilde{U} \quad \text{"easy" problem for } \Psi.$$

Note: $\underline{\tilde{U}}|_{\partial\Omega} = 0$, both components.

$U^{n+1}\cdot\hat{n} = 0$, but $U^{n+1}\cdot\hat{\tau} \neq 0$.

Convenient to analyze this in terms of $\tilde{U}$

$$\tilde{U}^{n+1} = Q\tilde{U}^n + k\Delta\tilde{U}^{n+1} + k\underline{f}^{n+1} \qquad (2)$$

$$\tilde{U}^{n+1} = U^{n+1} + \nabla\Psi^{n+1} \qquad (3)$$

Consider the projection of (2) onto A & B:

$$A: \quad \underline{U}^{n+1} = \underline{U}^n + Q\Delta\underline{U}^{n+1} + k\underline{f}^{n+1}$$

$\mathcal{B}$ commutes with Δ.

[looks the same as (*) ✓].

$$\Psi^{n+1} = k\mathcal{P}\Delta U^{n+1} + k\Delta\Psi^{n+1}$$

So $(I - k\Delta)\psi^{n+1} = k \underbrace{\mathcal{P}\Delta U^{n+1}}_{p.}$ (4)

From (3), $\frac{\partial \psi}{\partial n} = 0$, but from (4) to highest order $\psi^{n+1} = k p^{n+1}$.

Using the aside on the first pages,

$$\psi^{n+1} = k p^{n+1} - k^{3/2} \frac{\partial p^{n+1}}{\partial n} e^{-h/\sqrt{k}}.$$

Identifying $p^{n+1} = \frac{\psi^{n+1}}{k}$ we see that the computed pressure has an $O(\sqrt{k})$ boundary layer of width $O(\sqrt{k})$.

How does this affect the values of $\underline{U}$ and $\hat{\underline{U}}$?

$\hat{U}_T = U_T - k \frac{\partial p}{\partial s}$ has an $O(k^{3/2})$ boundary layer.

$\hat{U}_n = U_n - k \frac{\partial p}{\partial n}$ has an $O(k)$ boundary layer.

U_T has nonzero boundary values at $O(k^{3/2})$.

MATH 521, Spring 2014
Assignment I - due Monday, January 27

- Do all of part A and one of part B.
- For students planning to do a course project, the next assignment will have a part C involving a proposal for the project. Come and see me to discuss your project before then to get early feedback.

Recall Problem #1 in the lecture notes for $u(x)$, $1-$periodic in x:

$$-u'' + au = f$$

with given $a > 0$ and $f(x)$, $1-$periodic in x. This problem is considered in some of the questions below.

A1. Suppose the $N \times N$ matrix $\mathbf{A}$ is strictly diagonally dominant, *i.e.* all rows satisfy

$$|A_{ii}| > \sum_{j \neq i} |A_{ij}|.$$

(a) Show that $\mathbf{A}$ is invertible.

(b) If $A_{ii} > 0$ for all i show that all eigenvalues of $\mathbf{A}$ are positive.

A2. Derive a fourth order accurate approximation of the second derivative based on a wide stencil using centred differencing based on the values

$$U_{i+2}, U_{i+1}, U_i, U_{i-1}, U_{i-2}.$$

Apply this to a discretization of Problem #1 above. Use a test problem to verify that the method converges with fourth order accuracy. Show the l_2-stability (energy stability) of the method using von Neumann analysis.

A3. Consider the wide approximation above applied to a modified Problem #1 with Dirichlet boundary conditions $u(0) = 0$ and $u(1) = 0$ replacing the periodicity condition. An additional numerical boundary condition must be given at each end (for U_{-1} and U_{N+1}) because of the wide stencil. Experiment numerically with different conditions until you find some that give fourth order convergence.

A4. Derive a fourth order accurate finite difference method for Problem #1 above using a short stencil, *i.e.* using only U_{i+1}, U_i, U_{i-1}. Implement the scheme on a test problem and verify fourth order convergence. *Hint:* Use the second derivative of the equation to help eliminate the second order error term from the standard approximation.

A5. Consider the second order finite difference approximation of Problem #1 derived in class, with matrix $\mathbf{A}$. Find an explicit formula for the entries of $\mathbf{A}^{-1}$. *Note:* You should check your analytical answer against MATLAB computations for small N.

Part B: do one question in this part

B1: Consider Problem #1 stated above. Show that

$$\|u\|_{C_2} \leq K\|f\|_{C_0}$$

for all f with $K(a)$. The choice of K does not have to be optimal, but make it as small as you can. Extend your result to show that

$$\|u\|_{C_{n+2}} \leq K\|f\|_{C_n}$$

for integer $n \geq 0$. Here $K(a, n)$. *Hint:* One approach is to use Greens function techniques, *i.e.* find a function $G(x)$, $1-$periodic, piecewise C^1 (with discontinuity in G' at $x = 0$ and periodic images) such that

$$u(x) = \int_0^1 G(x - \xi) f(\xi) d\xi$$

and use this representation to derive the desired property.

B2: Show that the fourth order wide method in question #A2 above is stable in maximum norm.

B3: Perform an asymptotic error analysis of the scheme you propose in question #A3 above.

B4: Download and install the freeware FEM package, Deal-II, from

`http://www.dealii.org`

Follow the tutorial programs to step-4. Pick this question if you want to become familiar with this FEM package and do the part-B questions on it in later assignments.

MATH 521, Spring 2014
Assignment II - due Wednesday, February 12

- Do all of part A and one of part B.
- For students planning to do a course project, part C involves a proposal for the project. Come and see me to discuss your project if you want early feedback.

Recall Problem #1 in the lecture notes for $u(x)$, $1-$periodic in x:

$$-u'' + au = f$$

with given $a > 0$ and $f(x)$, $1-$periodic in x. This problem is considered in some of the questions below.

Part A: do all questions in this part

A1. Code up the finite element method with basic linear elements for the problem #1 above. Your code should be able to handle any given discrete grid. You can print out the section of your code that fills in the matrix $\mathbf{A}$ as the answer to this question (normally I do not want to see code).

A2. Use your code in A1 above to do the following computation: start with a regular grid of N points with spacing $h = 1/N$. Then move the points randomly with a uniform distribution on the interval $[-h/3, h/3]$. This should give points randomly between $h/3$ and $5h/3$ apart. Compute errors to a known solution on several resolutions with $h \to 0$, making the grids random in this way at every resolution. Report the test example you picked and the errors you observed. *Note:* this should give you the flavour of what convergence on an unstructured grid is like.

A3. Consider the case of elements that are piecewise quadratic on subintervals and continuous across subinterval boundaries. Consider an additional discrete value at the centre of subintervals. Basis functions for this case come in two types. For interval interior points, the basis functions are quadratic and nonzero only in the interval, with value 1 at the centre and 0 at the ends. For interval end points, the basis functions are piecewise quadratic in both adjacent intervals, with value 1 at the point in consideration and zero at all other points. There are derivative discontinuities at subinterval ends. Consider applying this scheme to problem 1 above on a uniform grid. What are the entries of the resulting matrix $\mathbf{A}$? *Note:* a package like Maple capable of symbolic integration might be helpful here.

A4. (optional, but this opens up some part B question options later in the course). Access a version of MATLAB that has the PDE Toolbox. Read the documentation for the command "pdedemo1" and execute it. Print out one of the plots from this assignment and hand it in as the answer to this question. *Note:* the Mathematics IT people have asked me not to

use the MATLAB version on the machine, "pascal", for this course. The toolbox is available on the undergraduate network. If you don't yet have access to this network, let me know and I will arrange it.

Part B: do one question in this part

B1: We saw that if $f \in L_2$ then

$$(f, \phi)_{L_2}$$

was a bounded linear functional for $\phi \in H^1$. Starting with $f \in C_\infty$, we can complete a space X in the usual way in the following norm:

$$\|f\|_X := \sup \frac{(f, \phi)_{L_2}}{\|\phi\|_{H^1}}$$

where the supremum excludes $\phi = \mathbf{0}$. Show that this has the properties of a norm. Determine if there is a corresponding inner product. Characterize the elements of the complete space. The space is known as H^{-1} if you want to look for some results in the literature.

B2: Prove analytically the convergence of the scheme with the elements in question A3.

B3: Implement the scheme with elements in question A3 on problem 1 on a uniform grid. What order of convergence do you observe? It should be higher order convergence than expected. Explain why this occurs.

B4. Consider the case of elements in 1D that are piecewise cubic and C_1 across subinterval boundaries. Write a basis for this approximating space. Consider C_1 elements in 2D using a triangular mesh. What is the minimum polynomial degree needed for such elements? (use a counting argument).

B5: If you downloaded the freeware FEM package, Deal-II, from

`http://www.dealii.org`

you can continue to work becoming familiar with this package. There is some flexibility here. You could modify some of the demo programs to handle a different domain shape or consider some different terms in the PDE. Report on what you did and attach some plots of the output. Please come discuss this with me if it would help.

Part C: project proposal

There is an option to have an oral final exam or do a project for the course. If you intend to do a project, please hand in a proposal with this assignment.

Format: Make a 2 page (strict maximum) proposal of your planned project, which can either be a longer proof or a computation of an applied problem. In your proposal, state clearly the theorem you want to prove or give some details

of the application you are interested in. In either case, give a rough outline of your approach. Identify clearly three stages for the project: the first stage of a simple warm-up problem (plan zero), the second (plan A) which you are quite sure you can do, and then (plan B) the more difficult problems you will tackle if part A goes well. Your proposal is worth 5/40 of the marks for the project. Based on your proposal, I can verify that your proposed plan A results would give you a good mark if completed. As I mentioned at the beginning of the term, I would be happy if your project overlapped with your research work.

MATH 521, Spring 2014
Assignment III - due Monday, March 3

Part A: do all questions in this part

A1. Adapt your code form Assignment #2, question #A1 to handle the boundary conditions $u(0) - u'(0) = 1$ and $u'(1) = 2$. Test your code on an example for which you have an exact solution.

A2. Find the dimension of P_3 (polynomials of degree 3) in 2 dimensions (2D). Describe an element on triangles that is conforming (continuous between elements if linear mappings from the reference element are used) and spans P_3.

A3. Find a third order quadrature method on the unit equilateral triangle. That is, find n points (x_i, y_i) and weights w_i such that

$$\int_\Omega p(x,y) = \sum_{i=1}^{n} w_i p(x_i, y_i)$$

for all polynomials $p(x)$ of degree 2. Your weights should all be positive.

A4. Consider quadrature rules in 1D. There are certain advantages to evaluating the integrand at end points ($y = -1$ and $y = 1$ in the reference interval). For $n = 2$ points (only the end points) we get Trapezoidal Rule (second order accurate) and for $n = 3$ we have Simpson's Rule (fourth order). Find the two interior points and the weights for all points for the $n = 4$ quadrature rule that includes the end points. It will be sixth order accurate.

Part B: do one question in this part

B1: Prove (one case of) Rellich's Lemma. That is, if $\{\phi_n\}$ is a bounded sequence in H^m then it has a convergent subsequence in H^n if $n < m$.

B2: Compute the solution to $\Delta u = e^y + x^2$ on the unit disk using MATLAB's PDE Toolbox. Convince yourself (and me!) that your result has error less than 0.001 in maximum norm.

B3: Do question #B2 with the Deal-II package instead of in MATLAB. If you are following other interests, let me know and we can negotiate this question.

B4: Write your own code to solve #B2 with the element of your choice.

B4. Find a fourth order quadrature method on the unit equilateral triangle with positive weights. Describe briefly how you would find higher order quadrature formulae on the triangle.

MATH 521, Spring 2014
Assignment IV - due Monday, March 24

Part A: do all questions in this part

A1. Solve the nonlinear problem

$$-u'' + u + u^3 = \sin 2\pi x$$

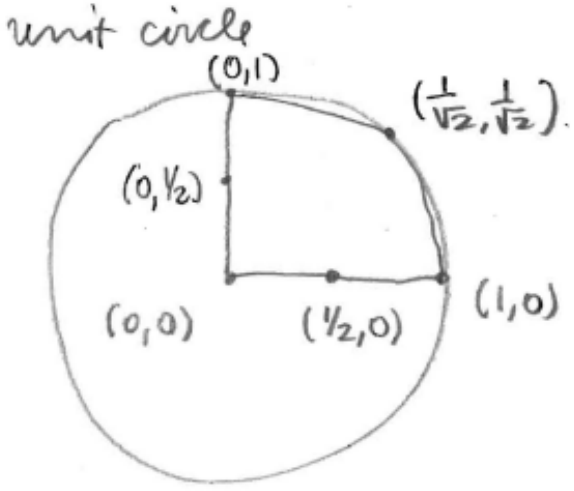
for $u(x)$, 1-periodic in x. Use a finite difference method on a uniform grid for the discretization and Newton's method to handle the nonlinear solve.

A2. Describe how you would discretize the general nonlinear problem

$$-u'' + u + f(u) = \sin 2\pi x$$

with a finite element method and generate the Jacobian matrix for your discretization. You can assume that the function $f(u)$ and its derivative are given to you.

A3. Consider the isoperimetric P_2 element on the reference element that is the unit equilateral triangle. Find the quadratic mapping that maps the reference triangle to the mesh element shown below. Verify that the map that you find is linear when restricted to the two interior edges.



Part B: do one question in this part

B1: Prove existence and uniqueness of solutions to the nonlinear problem

$$-u'' + u + u^3 = \sin 2\pi x$$

for $u(x)$, 1-periodic in x.

B2: Using Deal-II or MATLAB's PDE Toolbox, compute the solution to

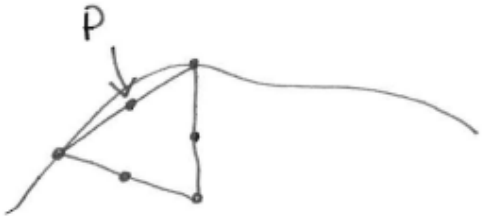
$$-\Delta u + u_x = 1$$

with homogeneous Dirichlet boundary conditions in a 2D domain that is not a circle or polygon. Alternate Deal-II problems can be negotiated.

B3: Consider the use of the P_2 element in 2D with reference element the unit length equilateral triangle with only linear maps used from the reference element to the elements in the mesh. Consider the application of the method to the problem

$$-\Delta u + u = f(x, y)$$

with homogeneous boundary data given for u. At the boundary, the boundary condition $u = 0$ is only an approximation at element edge midpoints as shown below (point P). What is the size of the interpolation error in the H^1 norm introduced by this approximation? You can assume that the domain boundary is smooth.



MATH 521, Spring 2014
Assignment IV - due Thursday, April 17

- Projects are also due April 17 and we'll have oral exams that afternoon.
- Note that there is no part B to this assignment.

Part A: do all questions in this part

A1. Show that the matrix coming from the Marker and Cell (MAC) grid for incompressible flow is symmetric.

A2. Find the eigenvalues of the MAC grid. How do the minimum and maximum eigenvalues (in magnitude, consider positive and negative eigenvalues separately) scale with discretization h? I intend that you use numerical experiments (i.e. MATLAB `eig` routine) to answer this, although pencil and paper estimates are fine as well.

A3. Show that the DIRK-2 method in the class notes is indeed second order and A-stable.

A4. Find the stability region of third order Radua-IIA. It will be possible to look up the shape of the region, but I would like you to work out how the boundary can be calculated.