Survey of Applied Mathematics Techniques

Brian Wetton [1]

June 30, 2018

[1] www.math.ubc.ca/~wetton, wetton@math.ubc.ca

Lecture 4

Series solutions, Sturm-Liouville Problems, Spectral Methods.

4.1 Series and Transform Solutions

Note 1. *If Fourier Series and Fourier Transforms are new to you, have a look at the description in an Appendix in the Lecture #1 notes.*

4.1.1 Fourier Transform

If we consider problem A1 ($x \in \mathbb{R}$), for which we derived a Green's function representation in Lecture #1

$$-u'' + u = f$$

and assume $f \in L_2$ so $u \in H^2$ so that the Fourier transform of f and u exist, we can take the transform of the equation to obtain

$$(\alpha^2 + 1)\hat{u} = \hat{f}$$

(where $\hat{f}(\alpha)$ is the Fourier Transform of f) and so

$$\hat{u} = \frac{\hat{f}}{\alpha^2 + 1}.$$

From this relationship we could have shown that $\|u\|_{H^2} \leq C\|f\|_2$ and also that the solution u is the convolution with the inverse transform of $1/(\alpha^2+1)$, which is

$$e^{-|x|}.$$

So the transform provides an alternate derivation of the Green's function.

4.1.2 Fourier Series

Consider the problem A2 (u and f periodic in $x \in [0,1]$):

$$-u'' + u = f$$

Here we can write the Fourier Series solution which has the same form as the transform solution in the previous section, but with the wave-numbers α discrete (integers)

$$\hat{u} = \frac{\hat{f}}{4\pi^2\alpha^2 + 1}.$$

The a priori bound $\|u\|_{H^2} \leq C\|f\|_2$ can come from this relationship and it is also possible to identify the Green's function as the inverse transform of $1/(4\pi^2\alpha^2+1)$ in this case.

4.1.3 Implications for time dependent problems

We have seen the heat equation in Lecture #3:

$$u_t = u_{xx} - u \tag{4.1}$$

We will consider u to be 1-periodic in x at for all $t \geq 0$. Initial conditions $u(x,0) = u_0(x) \in L_2$ are given. We look for solutions in the form

$$u(x,t) = e^{\lambda t}\phi(x)$$

which leads to

$$\phi'' - \phi = \lambda\phi.$$

That is ϕ is an eigenfunction of the operator $\frac{d^2}{dx^2} - I$ with 1-periodic boundary conditions. From the previous section, we know that the Fourier Series vectors are eigenfunctions ϕ of the operator with eigenvalues $\lambda = -(4\pi^2\alpha^2 + 1)$ and so

$$u(x,t) = \sum_{\alpha} c_\alpha e^{-(4\pi^2\alpha^2+1)t} e^{2\pi i\alpha x} \tag{4.2}$$

solves (4.1) for any choice of coefficients c_α. Matching initial conditions gives $c_\alpha = \hat{u_0}_\alpha$, the Fourier Series coefficients of the initial data. That the set of Fourier vectors is a basis for L_2 (that their span is *complete in* L_2) makes (4.2) the general solution of (4.1).

Considering the form (4.2) the interest from applications will be in the largest (least negative) eigenvalue or largest few eigenvalues of the spatial operator, as they are the ones that will persist the longest in time.

4.2 Sturm-Liouville Problems

Consider now the more general boundary eigenvalue problem

$$Lu := -(a(x)u')' + b(x)u = \lambda c(x)u \tag{4.3}$$

with $u(0) = 0$ and $u(1) = 0$, and the functions $a(x) > 0$ and $c(x) > 0$ for all x and continuous (although continuity can be relaxed as show in Problem 1 below). In addition, more general linear boundary conditions can be given, and the theory below can be extended to cases where $a = 0$ and/or $c = 0$ at isolated points.

4.2.1 Summary of Sturm-Liouville theory

The eigenvalue problem (4.3) has the following properties:

1. There are an infinite number of eigenvalues $\{\lambda_n\}$ bounded below with

$$\lim_{n\to\infty} \lambda_n = \infty.$$

   We order the sequence so that $\{\lambda_n\}$ is increasing.

2. The eigen-space associated to each λ_n is one dimensional, spanned by $u_n(x)$ (the space can be two-dimensional with periodic conditions).

3. The eigenfunctions are orthogonal in the sense that

$$(u_n, u_m) := \int_0^1 c(x)u_n(x)u_m(x)dx = 0$$

   for $n \neq m$. We can consider now $\{u_n\}$ to be normalized in the norm that is induced by the inner product above, that is

$$\|u\|^2 := \int_0^1 c(x)|u(x)|^2 dx = 1$$

4. Let $V_n = (\text{span}\{u_1, u_2, \ldots u_n\})^\perp$, then

$$\lambda_{n+1} = \min_{u\in V_n, \|u\|=1} (u, Lu)$$

   and the minimizer is u_{n+1}.

5. $\{u_n\}$ is complete in L_2.

6. u_n has $n-1$ roots in the open interval $(0, 1)$.

All of the orthogonal basis functions that you know (Fourier Series, Sine Series, Cosine Series, Bessel Series, Chebyshev polynomials, ...) come from Sturm-Liouville problems.

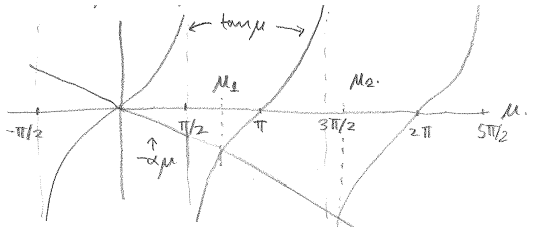

Figure 4.1: Intersection points that give the μ values. In these notes, the parameter α was taken to be 1.

4.2.2 Easy example

Consider $u(x)$ that solves

$$u_t = u_{xx}$$

with $u(0) = 0$ and $u(1) + u'(1) = 0$. As motivated above, we look for solutions in the form

$$u(x,t) = e^{-\lambda t}u(x)$$

and arrive at the problem in the form (4.3)

$$-u'' = \lambda u$$

but with boundary conditions $u(0) = 0$ and $u(1) + u'(1) = 0$, for which the theory in Section 4.2.1 still applies. We could have arrived at this problem more formally using separation of variables. There are no non-trivial solutions when $\lambda \leq 0$. This can be seen by direct computation or by showing that $(u, Lu) > 0$ (be careful with boundary term when you integrate by parts) unless $u \equiv 0$ and looking at property #4 in the previous section. Taking $\lambda = -\mu^2$ we find the general solution of the ODE as

$$u = A\cos\mu x + B\cos\mu x.$$

With $u(0) = 0$ we have $A = 0$ and with $u(1) + u'(1) = 0$ we have

$$\cos\mu + \mu\sin\mu = 0$$

to get a nontrivial solution. We can rewrite

$$\tan\mu = -\mu$$

and the intersections that give the μ values are shown in Figure 4.1 and can be approximated using Newton's method. The eigenvalues are $\lambda_n = -\mu^2$ and the eigenfunctions (not normalized) are $\sin\mu x$ and are orthogonal in the usual L_2 inner product.

4.2.3 More complicated example

Find the eigenfunctions of the negative Laplace operator on the unit disk with homogeneous Dirichlet conditions.

$$-\Delta u = -\frac{1}{r}(ru_r)_r - \frac{1}{r^2}u_{\theta\theta}$$

with the negative sign to give positive eigenvalues. Let us look for eigenvalues in the form

$$u(r,\theta) = u(r)e^{in\theta}.$$

This form could be derived formally using separation of variables. With this form, we have the eigenvalue problem

$$-\frac{1}{r}(ru')' + \frac{n^2}{r^2}u = \lambda u. \tag{4.4}$$

To proceed, we write $\lambda = \mu^2$ and rearrange the equation to

$$r^2u'' + ru' + (\mu^2r^2 - n^2)u = 0.$$

We make the change of variables $s = \mu r$ (μ still to be determined) and after some work obtain the following, where the dots denote differentiation with respect to s:

$$s^2\ddot{u} + s\dot{u} + (s^2 - n^2)u = 0$$

(the Bessel equation) with solutions $J_n(s)$ and $Y_n(s)$. Since $Y_n(s)$ is unbounded as $s \to 0_+$, we have $u = J_n(s)$. Since $u(r = 1) = u(s = \mu) = 0$, we see that μ must be a root $\mu > 0$ of J_n. The low order roots can be found in Tables online. We label these roots in increasing order for fixed n as $\mu_{m,n}$. We now have eigenfunctions of $-\Delta$ as

$$u_{m,n}J_n(\mu_{m,n}r)e^{in\theta}$$

with eigenvalues $-\mu_{m,n}^2$. For the orthogonality condition, we have to be a bit more careful. To put (4.4) in the form (4.3) we have to multiply by r,

$$-(ru')' + \frac{n^2}{r}u = \lambda ru$$

and so the eigenfunctions will be orthogonal in the inner product

$$(u,v) = \int_0^{2\pi}\int_0^1 ruv \text{ dr d}\theta$$

4.3 Spectral Methods

We consider our 1D elliptic model problem again:

$$Au := -u_{xx} + u = f \tag{4.5}$$

with $x \in [0,1]$ and u periodic. The Fourier series coefficients for 1-periodic functions is given by

$$\hat{u}_\alpha = \mathcal{F}(u) = \int_0^1 u(x)\, e^{-2\pi i \alpha x}\, dx \tag{4.6}$$

with inverse

$$u(x) = \sum_{\alpha=-\infty}^{\infty} \hat{u}_\alpha e^{2\pi i \alpha x}. \tag{4.7}$$

Note that $\{e^{2\pi i \alpha x}\}$ is an orthonormal set in L_2 norm and spans L_2 and

$$\|u\|^2 = \sum_{\alpha=-\infty}^{\infty} |\hat{u}_\alpha|^2.$$

Also, these functions are eigenfunctions of A, **i.e.**

$$A\, e^{2\pi i \alpha x} = (4\pi^2\alpha^2 + 1) e^{2\pi i \alpha x}.$$

That A has a complete set of orthogonal functions is not such a surprise - since A is symmetric and positive definite (A is unbounded which makes the proof of this a little more difficult).

An expression for the solution to (4.5) can be found easily by decomposing f into spectral representation, **i.e.** we can find $\mathcal{F}(f)$, then $\hat{u}_\alpha = \hat{f}_\alpha/(4\pi^2\alpha^2+1)$ and $u(x)$ can be found by the inverse transform.

4.3.1 Spectral methods

The idea of spectral methods is to use a finite number of terms in the expansion (4.7). We will use an approximation

$$U(x) = \sum_{-N/2}^{N/2} \hat{u}_\alpha e^{2\pi i \alpha x}.$$

where we use

$$\hat{u}_\alpha = \frac{\hat{f}_\alpha}{4\pi^2\alpha^2 + 1}$$

with the exact $\hat{f}_\alpha$ for now. Note that this approximation is a continuous function, not just at discrete points like the Finite Difference approximation. The Finite Element approximation we will discuss in a future lecture also has this property.

We can ask how close U is to the exact u in some convenient norm. For this problem, all norms are convenient. We'll use the L_2 norm here and notice that

$$\|U - u\|^2 = \sum_{|\alpha|>N/2} |\hat{u}_\alpha|^2 = \sum_{|\alpha|>N/2} |\hat{f}_\alpha|^2/(4\pi^2\alpha^2+1)^2 \tag{4.8}$$

since the coefficients for small α are the same. We assume that $f \in C^\infty$ (periodic) and derive the following simple estimate:

$$\begin{aligned} |\hat{f}_\alpha| &= |\int_0^1 f(x) e^{-2\pi\alpha x}\, dx| \\ &= |\int_0^1 \frac{f'(x)}{-2\pi i\alpha} e^{-2\pi\alpha x}\, dx| \\ &\leq B/|\alpha| \end{aligned}$$

where $B = \|f'\|_\infty/(2\pi)$. We repeat integration by parts to get

$$|\hat{f}_\alpha| \leq B_j/|\alpha|^j \tag{4.9}$$

where $B_j = \|f^{(j)}\|_\infty/(2\pi)^p$. This shows that if f is smooth, then $|\hat{f}_\alpha|$ decays rapidly with $|\alpha|$. Returning to the error equation (4.8) we see that

$$\begin{aligned} \|U-u\|^2 &= \sum_{|\alpha|>N/2} |\hat{f}_\alpha|^2/(4\pi^2\alpha^2+1)^2 \\ &\leq \left(\frac{B_j}{(N/2)^j}\right)^2 \sum_{|\alpha|>N/2} \frac{1}{(4\pi^2\alpha^2+1)^2} \\ &\leq (C_j/N^j)^2 \end{aligned}$$

for all j where C_j depends only on the size of the derivatives of f up to order j (note that this estimate is not sharp). Thus we have

$$\|U-u\| \leq C_j/N^j \tag{4.10}$$

for all j. Recall the estimate for a second order FE approximation:

$$\|U_h - u\| \leq Ch^2.$$

For higher order (q) methods, we will see higher order convergence

$$\|U_h - u\| \leq Ch^q$$

but q is still fixed. Considering (4.10) we see that spectral methods are asymptotically more accurate that any FE or FD method since they converge faster than any power of $h = 1/N$. This type of convergence (4.10) is called spectral convergence. Remember that this estimate only holds for smooth data f.

4.3.2 Aliasing and Pseudo-Spectral Methods

The only thing we don't like about the spectral method described above is the potentially laborious calculation to accurately compute $\hat{f}_\alpha$ from the integrals (4.6) and the evaluation of $U(x)$ at desired points from the summation (4.7) (recall this summation has only finite terms for U). The way to get around this

problem is to use the inverse FFT to evaluate the approximation on a uniform grid with N grid points - this procedure is fast and exact. We can also use the FFT to approximate the fourier coefficients $\hat{f}_\alpha$ from a vector of values $F_i = f(ih)$ on N grid points. The details are presented below.

Recall that the DFT is given by

$$\hat{U}_\alpha = \frac{1}{N} \sum_{k=0}^{N-1} U_k e^{-2\pi i k\alpha/N}.$$

and the inverse is given by

$$U_k = \sum_{\alpha=0}^{N-1} \hat{U}_\alpha e^{2\pi i k\alpha/N}. \tag{4.11}$$

Note that different scalings from the original scalings have been used here. We will first deal with some technical details. Note that the sum above goes from 0 to $N-1$ but the spectral method provides us with $\hat{u}_\alpha$ for $|\alpha| \leq N/2$ (this makes more sense because the smaller $|\alpha|$ are the significant ones). However, if we evaluate the functions $e^{2\pi i\alpha x}$ and $e^{2\pi i(\alpha+N)x}$ only at the points $x_k = k/N$ on the grid, we cannot distinguish them. Therefore, we can easily relabel the summation in (4.11) to go from $\alpha = -N/2+1, \ldots N/2$ where

$$\hat{U}_\alpha \sim \hat{u}_\alpha \quad \text{for } \alpha = -N/2+1, \ldots N/2-1$$

and

$$\hat{U}_{N/2} \sim \hat{u}_{-N/2} + \hat{u}_{N/2}. \tag{4.12}$$

No information has been lost in (4.12) because these two modes are indistinguishable on the grid. Whatever we do with the $N/2$ coefficient is not that important since we expect it to be very small (remember the fast decay for large α). However, it is sometimes useful to consider this term in the form (4.12), **i.e.** so that spectral evaluations of first derivatives of real f stay real. An alternative is just to set it to zero. Using the above values of $\hat{U}$, the inverse FFT can be used to evaluate the function U on the grid points exactly.

To approximate $\hat{f}_\alpha$ for $|\alpha| \leq N/2$ we follow the same plan. We get the vector F by evaluating f on the grid points and then compute $\hat{F}$ by the inverse FFT and identify

$$\hat{F}_\alpha \sim \hat{f}_\alpha \quad \text{for } \alpha = -N/2+1, \ldots N/2-1$$

and

$$\hat{F}_{N/2} \sim \hat{f}_{-N/2} + \hat{f}_{N/2}. \tag{4.13}$$

It is easy to show that in fact

$$\hat{F}_\alpha = \sum_{l=-\infty}^{\infty} \hat{f}_{\alpha+Nl}. \tag{4.14}$$

This is not surprising since we cannot distinguish between the α, $\alpha+N$, $\alpha+2N$, etc. modes on the grid. The effect in (4.14), where high frequency information

is seen as low frequency information, is called aliasing. Using the decay of the coefficients (4.9) it can be shown that

$$|\hat{F}_\alpha - \hat{f}_\alpha| \leq C_j/N^j \tag{4.15}$$

for $|\alpha| \leq N/2 - 1$ and all j. A similar bound can be made for the $N/2$ term. With the $\hat{F}$ values we can compute

$$\hat{U}_\alpha = \hat{F}_\alpha/\kappa_\alpha$$

for $\alpha = -N/2 + 1, \ldots N/2$ where κ_α is the corresponding eigenvalue $4\pi^2\alpha^2 + 1$. Note that this approximation is consistent with (4.12) and (4.13) since $\kappa_{-N/2} = \kappa_{N/2}$. Now, the values of U can be evaluated on the grid as described above. This is a fast ($O(N \log N)$) method called the pseudo-spectral method which is also spectrally accurate as shown below.

We consider pointwise errors rather than L_2 errors this time.

$$\begin{aligned} |u(x) - U(x)| &= |\sum_{\alpha=-\infty}^{\infty} \frac{\hat{f}_\alpha}{\kappa_\alpha} - \sum_{\alpha=-N/2}^{N/2} \frac{\hat{F}_\alpha}{\kappa_\alpha}| \\ &\leq \sum_{\alpha=-N/2}^{N/2} |\hat{f}_\alpha - \hat{F}_\alpha| + \sum_{|\alpha|>N/2} |\hat{f}_\alpha|. \end{aligned}$$

The second term decays spectrally as before and so does the first using (4.15). Thus we have

$$|u(x) - U(x)| \leq C_j/N^j$$

for all j and all x. We write $\|u - U\|_\infty \leq C_j/N^j$. As before, the estimates above are not sharp.

4.3.3 Problems with Variable Coefficients

We now apply a spectral method to the 1D analogue of M3:

$$-u_{xx} + b(x)u = f$$

with $b(x) = (2 + \cos(\cos(2\pi x)))$ and u 1-periodic. This is a symmetric, positive definite problem which has a complete set of eigenfunctions as before. However, we do not want to use these functions as our basis because we don't know them and we won't have a fast transform technique to do the conversion from a grid representation to a spectral one. Therefore, we will continue to use the Fourier basis from before.

Note 2 (Terminology)**.** *A spectral method does not mean we always use a spectral representation of the problem. It means we are using a representation that will give us spectral accuracy.*

We first develop infinite conditions that the exact solution u must satisfy. By taking the transform of the original problem we get

$$4\pi^2\alpha^2\hat{u}_\alpha + \widehat{bu}_\alpha = \hat{f}_\alpha.$$

However,

$$\widehat{bu}_\alpha = \sum_{n=-\infty}^{\infty} \hat{b}_n \hat{u}_{\alpha-n}$$

so the FT does not diagonalize the problem effectively (this should not be expected since the fourier terms are not eigenfunctions).

A true spectral method (a Galerkin method) can be derived by assuming $\hat{u}_\alpha = 0$ for $|\alpha| > N/2$ and then projecting the resulting equations on to the corresponding finite dimensional space (just like the approach we will take with the Finite Element Method). The resulting equations for $\hat{U}_\alpha$ for $\alpha \leq N/2$ are

$$(-4\pi^2\alpha^2 I + B)\hat{U}_\alpha = \hat{f}_\alpha$$

where the matrix B is full (with the $\hat{b}$ terms from the truncated convolution). Solving this system directly will be slow, because full matrix methods must be used. Although solving the method iteratively by the CG method is possible (the matrix is symmetric and positive definite), it will be slow because multiplication by B will be slow.

We can speed up this process, however, by evaluating $B\hat{U}$ approximately (in the pseudo-spectral sense which will introduce aliasing). Recall that $B\hat{U}$ approximates $\widehat{bu}$. Therefore, we could compute U_i by inverse FFT, multiply *pointwise* by b_i and then compute the inverse transform. This approximation is equivalent to replacing the matrix B by

$$\mathcal{F}\tilde{B}\mathcal{F}^{-1}$$

where $\tilde{B}$ is a diagonal matrix representing the pointwise multiplication. The resulting system is

$$A\hat{U} := (\Xi + \mathcal{F}\tilde{B}\mathcal{F}^{-1})\hat{U} = \mathcal{F}F \tag{4.16}$$

where Ξ is a diagonal matrix with entries $4\pi^2\alpha^2$ and it is assumed that $\hat{f}$ is approximated pseudospectrally. Note that A is a symmetric (since $\mathcal{F}^{-1} = \mathcal{F}^*$ where $*$ denotes complex transpose) and positive definite since $\tilde{B}$ and Ξ have positive entries. It is also possible to evaluate $A\hat{U}$ quickly so a Conjugate Gradient (CG) method can be applied. The CG method and its analysis is described in an appendix to these lecture notes. The number of iterations to reach a given tolerance is bounded by the square root of the condition number. The condition number is $O(N^2)$ for this matrix, and so $O(N)$ iterations are needed. In this case, an effective preconditioner can be used, the pseudospectral approximation of the constant coefficient problem with $b(x)$ replaced by 2. With this Preconditioned Conjugate Gradient (PCG) method, iterations are independent of N. MATLAB code of an implementation is provided.

Note 3 (Don't need A sparse for CG)**.** *This example shows that a matrix does not have to be sparse to make CG an efficient technique. In fact, our matrix A was dense, but we could still evaluate it quickly. For another interesting example of this phenomenon, see [6].*

Note 4 (Notation)**.** *Since almost all practical methods involve pseudo-spectral evaluation of coefficients and interpolation, the "pseudo" is being dropped from these methods and they are usually called simply spectral methods.*

4.3.4 Be Careful with Boundaries

What about domains with boundaries? We consider a 1D analogue of M1

$$-u_{xx} = f$$

with $u(0) = u(1) = 0$. The eigenfunctions of this problem are $\{\sin(n\pi x)\}$ with corresponding eigenvalues $n^2\pi^2$. Since these correspond to the basis of the fast sine transform, it seems to be a suitable basis for a spectral method. We consider computing the problem with $f \equiv 1$ (smooth), giving $u(x) = -x^2/2 + x/2$ (also smooth). However, the coefficients in the sine series for f

$$\hat{f}_n = \sqrt{2}\int_0^1 f(x)\sin(n\pi x)dx$$

decay only like n^{-1}. The corresponding solution u has sine coefficients $\hat{u}$ that decay only like n^{-3}. Thus, the "spectral" approximation of this problem described above will have errors of size $h^3 = 1/N^3$, **i.e.** errors no better than a third order FD or FE method.

In this case, the eigenfunctions are a *poor* choice for the basis functions for a method. A more appropriate choice would be the Chebyshev polynomials (using fast interpolation on an irregular grid of points). The details are given in the references.

4.4 Lecture #4 Problems

Problem 1. *A scaled heat conduction problem has the form*

$$u_t = (\kappa(x)u_x)_x$$

with $u(0) = u(1) = 0$. The scaled thermal conductivity comes from two different materials joined at $x = 1/2$,

$$\kappa = \begin{cases} 1 & \text{if } x < 1/2 \\ 2 & \text{if } x > 1/2 \end{cases}$$

Find the largest (least negative) eigenvalue of the RHS operator in the equation above, and show that it is intermediate between the eigenvalues of the two pure materials ($\kappa \equiv 1$ and $\kappa \equiv 2$). As part of the problem, you will need to specify the junction conditions for u at $x = 1/2$. Go back and look at the Lecture #3 notes for help with this.

Problem 2. *Prove as many of properties of Sturm-Liouville eigenvalue problems on the list in Section 4.2.1 as you can.*

Problem 3. *Implement a spectral solver for the nonlinear boundary value problem in the Lecture #3 notes. Use Newton iterations and preconditioned conjugate gradient solves for each iteration.*

Appendix: Conjugate Gradient Method

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

$$F, \mathbf{A}F, \mathbf{A}^2F, \ldots \tag{4.17}$$

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:

$$\begin{aligned} (U,V) &= \sum_i^M U_i \cdot V_i \quad \text{euclidean} \\ (U,V)_A &= (U, \mathbf{A}V) \end{aligned}$$

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

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 $\{d_1,\ldots d_k\}$ spans S_k (this could be done **i.e.** by applying the Gramm Schmidt process to (4.17) 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

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

or

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

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

$$r_k \quad \perp \quad S_k \tag{4.18}$$

$$r_k \quad \perp_A \quad S_{k-1}. \tag{4.19}$$

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 (4.19) we know that $\gamma_i = 0$ for $i = 1, \ldots k-2$. □

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 (4.18), $r_{N^*} \in S_{N^*}^{\perp}$. Therefore, $r_{N^*} = 0$ and so $U = U_{N^*}$. □

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

Note 5 (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

$$\|U - U_k\|_A^2 = \|r_k\|_{A^{-1}}^2 = \min_{h_k \in AS_k} \|F - h_k\|_{A^{-1}}^2. \tag{4.20}$$

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 (4.20), 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

$$\max_{|x|\leq 1} |T_k(x)| = 1 \tag{4.21}$$

and

$$T_k(x_i) = (-1)^i \text{ for } \; x_i = cos(i\pi/k), i = 0, 1, \ldots k. \tag{4.22}$$

From formula 1 we get the bound

$$T_k(\frac{a+1}{a-1}) > \frac{1}{2}\left(\frac{\sqrt{a}+1}{\sqrt{a}-1}\right). \tag{4.23}$$

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*

$$\|U - U_k\|_A \leq 2\gamma^k \|U\|_A \tag{4.24}$$

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

$$p(\epsilon) \leq \frac{1}{2}\sqrt{a}\ln(2/\epsilon) + 1. \tag{4.25}$$

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

$$C_k = T_k[(\lambda_N + \lambda_1)/(\lambda_N + \lambda_1)]^{-1} \tag{4.26}$$

which occurs $k+1$ times (with alternating sign) at the mapped points x_i from (4.22). 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 (4.26) along with (4.23) proves (4.24).

We now turn to a proof of (4.25). 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 (4.25) 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 6 (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.*

Bibliography

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

[2] Canuto, C. et. al., *Spectral Methods in Fluid Dynamics*.

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

[4] Gottlieb, D. and Orszag, S., *Numerical Analysis of Spectral Methods*.

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

[6] Rokhlin, V., "Rapid Solution of Integral Equations of Classical Potential Theory," JCP **60**, 187-207 (1983).

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

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

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