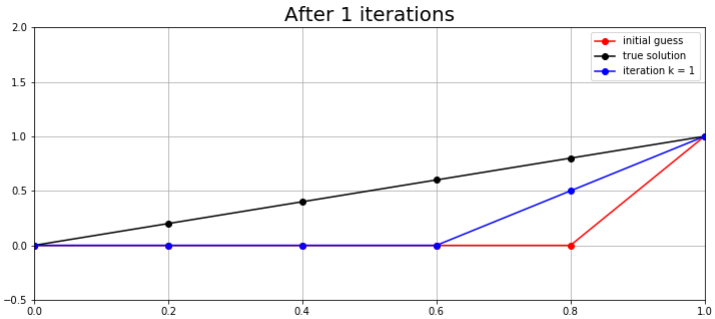
AMath 585
Homework #6
Due Thursday, March 12, 2020

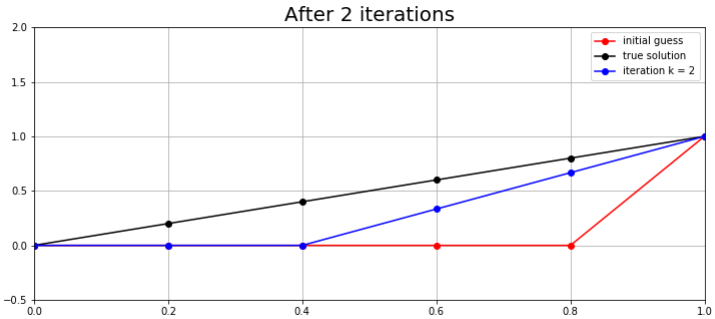
Homework is due to Canvas by 11:00pm PDT on the due date.

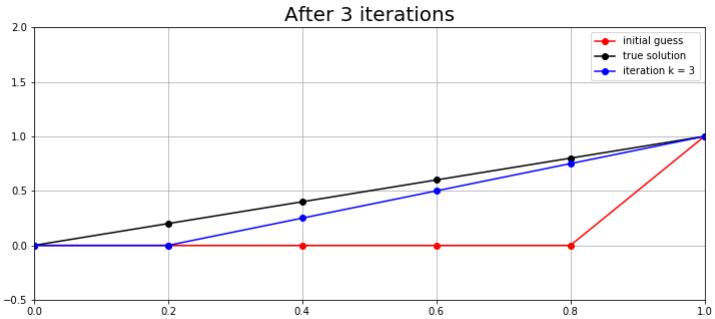
To submit, see `https://canvas.uw.edu/courses/1352870/assignments/5284853`

Problem 1. Consider the BVP $u''(x) = 0$ on $0 \le x \le 1$ with Dirichlet boundary conditions $u(0) = 0$ and $u(1) = 1$. The exact solution is $u(x) = x$.

Discretize with the standard centered approximation using m equally spaced interior points. If we apply the Conjugate-Gradient method with initial data $u_i^{[0]} = 0$ for $i = 1, 2, \ldots, m$ then we see the sort of behavior that is illustrated in the plots below for the case $m = 4$. For $k < m$ the approximate solution is always piecewise linear and has $u_i^{[k]} = 0$ for $i \le m - k$. After m iterations, $u^{[m]}$ is equal to the exact solution.







(a) For the case $m = 3$, work through the C-G algorithm by hand to explicitly calculate the vectors $r^{[k]}$, $b^{[k]}$, and $u^{[k]} \in \mathbb{R}^3$ in each iteration. This should help you see why the behavior seen in the plots makes sense.

(b) To show this behavior is seen for general m, show by induction that each residual $r^{[k]}$ is a unit vector (all zeros except in one element). Hint: Use the fact that we know that all the residuals generated in C-G are pairwise orthogonal to one another, and that the only elements that can change from one iteration to the next are those in which the search direction $b^{[k]}$ has nonzero components, which can also be determined in general.

(c) Explain how the result of (b) implies the behavior seen in the plots.

Solution:

a)

We will consider the BVP $u''(x) = 0$ on $0 \leq x \leq 1$ with Dirichlet boundary conditions $u(0) = 0$ and $u(1) = 1$. Then, we discretize with the standard centered approximation using m equally spaced interior points.

For the case $m = 3$ we will work through the C-G algorithm and calculate vectors $r^{[k]}, b^{[k]}$ and $u^{[k]}$.

Notice: $m = 3 \implies h = \frac{1}{m+1} = \frac{1}{4} = 0.25 \implies h^2 = \frac{1}{16}$

Recall (for $m = 3$):

$$A = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{bmatrix} = 16\begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{bmatrix}$$

Also:

$$f = \begin{bmatrix} 0 \\ 0 \\ -\frac{1}{h^2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -16 \end{bmatrix}$$

Initial guess:

$$u^{[0]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Then,

$$r^{[0]} = f - Au^{[0]} = \begin{bmatrix} 0 \\ 0 \\ -16 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -16 \end{bmatrix}$$

Also,

$$b^{[0]} = r^{[0]} = \begin{bmatrix} 0 \\ 0 \\ -16 \end{bmatrix}$$

Then,

$$w^{[0]} = Ab^{[0]} = \begin{bmatrix} 0 \\ -256 \\ 512 \end{bmatrix}$$

Then,

$$\alpha^{[0]} = \frac{r^{[0]T} r^{[0]}}{b^{[0]T} w^{[0]}} = \frac{256}{-16 \times 512} = -\frac{16}{512} = -\frac{1}{32}$$

.

Thus,

$$u^{[1]} = u^{[0]} + \alpha^{[0]} b^{[0]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - \frac{1}{32} \begin{bmatrix} 0 \\ 0 \\ -16 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{2} \end{bmatrix}$$

Then:

$$r^{[1]} = r^{[0]} - \alpha^{[0]} w^{[0]} = \begin{bmatrix} 0 \\ 0 \\ -16 \end{bmatrix} + \frac{1}{32} \begin{bmatrix} 0 \\ -256 \\ 512 \end{bmatrix} = \begin{bmatrix} 0 \\ -8 \\ 0 \end{bmatrix}$$

$$\beta^{[0]} = \frac{r^{[1]T} r^{[1]}}{r^{[0]T} r^{[0]}} = \frac{64}{256} = \frac{1}{4}$$

So,

$$b^{[1]} = r^{[1]} + \beta^{[0]} b^{[0]} = \begin{bmatrix} 0 \\ -8 \\ 0 \end{bmatrix} + \frac{1}{4} \begin{bmatrix} 0 \\ 0 \\ -16 \end{bmatrix} = \begin{bmatrix} 0 \\ -8 \\ -4 \end{bmatrix}$$

.

In the second iteration:

$$w^{[1]} = Ab^{[1]} = 16 \begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} 0 \\ -8 \\ -4 \end{bmatrix} = \begin{bmatrix} -128 \\ 192 \\ 0 \end{bmatrix}$$

$$\alpha^{[1]} = \frac{r^{[1]T} r^{[1]}}{b^{[1]T} w^{[1]}} = -\frac{64}{1536} = -\frac{1}{24}$$

$$u^{[2]} = u^{[1]} + \alpha^{[1]} b^{[1]} = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{2} \end{bmatrix} - \frac{1}{24} \begin{bmatrix} 0 \\ -8 \\ -4 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}$$

$$r^{[2]} = r^{[1]} - \alpha^{[1]} w^{[1]} = \begin{bmatrix} 0 \\ -8 \\ 0 \end{bmatrix} + \frac{1}{24} \begin{bmatrix} -128 \\ 192 \\ 0 \end{bmatrix} = \begin{bmatrix} -\frac{16}{3} \\ 0 \\ 0 \end{bmatrix}$$

$$\beta^{[1]} = \frac{r^{[2]T} r^{[2]}}{r^{[1]T} r^{[1]}} = \frac{\frac{256}{9}}{64} = \frac{4}{9}$$

So,

$$b^{[2]} = r^{[2]} + \beta^{[1]} b^{[1]} = \begin{bmatrix} -\frac{16}{3} \\ 0 \\ 0 \end{bmatrix} + \frac{4}{9} \begin{bmatrix} 0 \\ -8 \\ -4 \end{bmatrix} = \begin{bmatrix} -\frac{16}{3} \\ -\frac{32}{9} \\ -\frac{16}{9} \end{bmatrix}$$

.

In the third iteration:

$$w^{[2]} = Ab^{[2]} = 16 \begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} -\frac{16}{3} \\ -\frac{32}{9} \\ -\frac{16}{9} \end{bmatrix} = 16 \begin{bmatrix} \frac{32}{3} - \frac{32}{9} \\ -\frac{16}{3} + \frac{64}{9} - \frac{16}{9} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1024}{9} \\ 0 \\ 0 \end{bmatrix}$$

$$\alpha^{[2]} = \frac{r^{[2]T} r^{[2]}}{b^{[2]T} w^{[2]}} = -\frac{\frac{256}{9}}{\frac{16384}{27}} = -\frac{256}{9} \times \frac{27}{16384} = -\frac{3}{64}$$

$$u^{[3]} = u^{[2]} + \alpha^{[2]} b^{[2]} = \begin{bmatrix} 0 \\ \frac{1}{3} \\ \frac{2}{3} \end{bmatrix} - \frac{3}{64} \begin{bmatrix} -\frac{16}{3} \\ -\frac{32}{9} \\ -\frac{16}{9} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{bmatrix}$$

$$r^{[3]} = r^{[2]} - \alpha^{[2]} w^{[2]} = \begin{bmatrix} -\frac{16}{3} \\ 0 \\ 0 \end{bmatrix} + \frac{3}{64} \begin{bmatrix} \frac{1024}{9} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

We stop here because $\|r^{[3]}\| = 0$ which will be less than any tolerance we pick.

b)

We will show that $r^{[k]}$ is an unit vector for general m.

Base case: $k = 0$

$r^{[0]} = f - Au^{[0]}$

Our initial guess is $u_i^{[0]} = 0$ for $i = 1, 2, \ldots, m$. So, $Au^{[0]} = \mathbf{0}$.

Therefore:

$$r^{[0]} = f$$

Recall:

$$f = \begin{bmatrix} f(x_1) - \frac{u(0)}{h^2} \\ f(x_2) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) - \frac{u(1)}{h^2} \end{bmatrix}$$

However, $f(x) = 0$ so it will be zero at all grid points. Therefore, $f(x_i) = 0$ for $i = 1, 2, \ldots, m$. Also, $u(0) = 0$ and so $f(x_1) - \frac{u(0)}{h^2} = 0$. Given $u(1) = 1$,

$$r^{[0]} = f = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -\frac{1}{h^2} \end{bmatrix}$$

Therefore, $r^{[0]}$ is a unit vector, that is, all zeros except in one element and so the base case holds.

Now, suppose for $0 \leq j \leq k$, $r^{[j]}$ is an unit vector. We will show that $r^{[k+1]}$ will also be an unit vector.

Recall: $r^{[k+1]} = r^{[k]} - \alpha^{[k]} w^{[k]}$ where $w^{[k]} = Ap^{[k]}$.

We will consider what is going on with $p^{[k]}$

Notice: $p^{[k]}$ is just a linear combination of $r^{[k]}$, that is, the residual vector at that step , and the previous search direction $p^{[k-1]}$.

$p^{[0]} = r^{[0]}$ and so it will also be an unit vector with a nonzero value in the mth entry.

Then, $Ap^{[0]}$ will be a vector with nonzero values in the last two entries of the vector. $Ap^{[0]}$ has the following form:

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ a \\ b \end{bmatrix}$$

for some constants a and b.

This makes sense because A is just the second different operator on the vector that is zero everywhere except the last value. Therefore, there is a kink at the $(m-1)$th entry where the function being discretized rises from 0 to the nonzero value. The second derivative approximation at the last point will also be non zero because it would be of the form $-2U_i$ and U_i is nonzero from the fact that $p^{[0]}$ was nonzero at the last entry.

Then, $r^{[1]}$ is just a linear combination of $Ap^{[0]}$ and $r^{[0]}$ and so it should look like the following:

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ c \\ d \end{bmatrix}$$

for some constants c and d.

However, we know that the residuals are pairwise orthogonal to one another. Therefore, $\frac{1}{h^2} \times d = 0$. Since, $\frac{1}{h^2} \neq 0$, $d = 0$. If $c = 0$, we would be done.

Therefore,

$$r^{[1]} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ c \\ 0 \end{bmatrix}$$

Therefore, $r^{[1]}$ will be an unit vector with a nonzero value in the $(m-1)$th entry.

Now, we move to $p^{[1]}$ again. This will be a linear combination of $r^{[1]}$ and the previous search direction $p^{[0]}$. Therefore, $p^{[1]}$ will be of the form:

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ c \\ e \end{bmatrix}$$

for some constants c and e.

The mth and $(m-1)$th entry will be nonzero. Then, $Ap^{[1]}$ will be a vector with zeroes everywhere except the last three entries. Again, we can use the same argument about the second difference operator as above. Since, $(m-1)$th entry is nonzero there will be a kink for the function that is being discretized by $p^{[1]}$. Therefore, $(m-2)$th entry will be nonzero because the second derivative would not be zero at that point and so on. The $(m-1)$th and mth entry might be zero depending on values of c and e but we do not really worry about that. We only care about the fact that at least the $(m-2)$th entry will be nonzero.

Then, $r^{[2]}$ is a linear combination of $r^{[1]}$ and $Ap^{[1]}$. So, $r^{[2]}$ will have nonzero values in the last three entries. It would look like:

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ g \\ h \\ i \end{bmatrix}$$

.

However, $r^{[2]}$ will be orthogonal to $r^{[1]}$ and $r^{[0]}$. This would mean that $h \times c = 0 \implies h = 0$. Also, $i \times \frac{1}{h^2} = 0 \implies i = 0$. Therefore, $r^{[2]}$ will be an unit vector with a non zero value in the $(m-2)$th entry. Again, if $g = 0$ we would be done.

Now, $p^{[2]}$ would be a linear combination of $r^{[2]}$ and $p^{[1]}$. So, $p^{[2]}$ would have nonzero values in the $(m-2)$th, $(m-1)$th and mth entry.

We can notice a pattern here. $p^{[k]}$ will have nonzero values from the $(m-k)$th entry to the mth entry. The 1st to the $(m-k-1)$th entry will be zero.

Then, $Ap^{[k]}$ will definitely have a non zero entry at the $(m-k-1)$th entry because there will be a kink at this value for the function being discretized by $p^{[k]}$. This is using the same argument by viewing A as the second difference operator. So, only one new entry will get a nonzero value.

Then, $r^{[k+1]}$ is going to be a linear combination of $Ap^{[k]}$ and $r^{[k]}$.Also, looking at the discussion above $r^{[k]}$ will be a unit vector with the nonzero value being at the $(m-k)$th entry. So, $r^{[k+1]}$ will have non zero values from $(m-k-1)$th entry to mth entry.

However, $r^{[k+1]}$ has to be orthogonal to all of the previous residual vectors. For each $0 \leq j \leq k$.$r^{[j]}$ is going to have a non zero value at the $(m-j)$th entry. Therefore, [$(m-j)$th entry of $r^{[j]}$] $\times$ [$(m-j)$th entry of $r^{[k+1]}$] $= 0$. Since $(m-j)$th entry of $r^{[j]}$ is non zero we have that $(m-j)$th entry of $r^{[k+1]}$ must be zero for $0 \leq j \leq k$. Therefore, the only non zero entry left in $r^{[k+1]}$ is at the $(m-k-1)$th entry. If this entry was zero, we would be done. So, $r^{[k+1]}$ is an unit vector with zeros everywhere except at the $(m-k-1)$th entry.

Thus, we have shown that $r^{[k]}$ is a unit vector by induction for general m.

c)

The result of (b) implies the behavior seen in the plots because the residual vectors are just unit vectors. This means that the only one entry is nonzero. Also, the search direction starts off with one non zero entry and then gains an additional entry at the point before the nonzero entry at each iteraiton. This should imply a piecewise linear function with one kink. This is because if the second derivative is taken of a linear function it should be zero. The only place it is nonzero is where there is a kink. This can be seen from the operator as well. If three points are collinear, then the result will be zero. If u was going to be linear at each iteration, then the residual vectors would not be of this form. The one nonzero entry implies a kink and therefore a piecewise linear function. Also, at each iteration the solution is a linear combination of the previous solution and the search direction. Since, the search direction gains a non zero entry at each iteration, so does the solution. Therefore, the kink position moves up in the solution, that is, it moves to the left. It happens sooner in the grid of x. Recall, that our initial guess was a vector of all zeros. So the linear combination of that and the first seach direction would provide one non zero entry at the last entry. So, the kink happens later on in the grid. The solution at second iteration still has mostly zeros but it adds a nonzero value in the second last position and so the kink occurs in the previous grid point.

Recall: $r_k = f - Au_k$ where subscript k denotes the iteration number.

f just has a nonzero entry at the last entry. We know that r_k has a nonzero in the $(m-k)$th entry. This means that, the 1st to the $(m-k-1)$th entry of Au_k must be zero since those entries of f are zero. This means that the second difference operator applied on the solution will yield some non zero values. This implies that the solution must be piecewise linear, because, the start of the switch from zero to nonzero in the next entry implies a kink.

Problem 2. Consider a linear system $Au = f$ in which the matrix A is **not** symmetric positive definite, so C-G cannot be applied directly.

(a) Show that if A is nonsingular then the matrix $B = A^T A$ is symmetric positive definite.

(b) So one approach to solving $Au = f$ is multiply both sides by A^T to get $Bu = A^T f$ and then solve this system with C-G. The problem with this approach is that the condition number increases. Show that the 2-norm condition number of B is the square of the 2-norm condition number of A.

(c) On page 93 it is noted that applying C-G to the two-dimensional Poisson problem $Au = f$ on an m by m grid (with second order centered differencing) requires $O(m^3)$ work to converge to a fixed tolerance. Suppose we multiplied both sides by A^T as described above (even though not necessary here since A is already SPD) and solved the resulting system (which is still SPD) by C-G. What order of work would now be required to reach a fixed tolerance?

(d) Given that the global error for this discretization is $O(h^2) = O(1/m^2)$ for smooth solutions, it makes more sense to look at the work required to get the error in the C-G solution down to this level. How does this change the work estimates given above, both for solving $Au = f$ and $A^T Au = A^T f$?

Note: there are better approaches for nonsymmetric matrices than the approach described above that do not magnify the condition number, see Section 4.4 and other references.

Solution:

a)

We will show that if A is nonsingular then the matrix $B = A^T A$ is symmetric positive definite.

First, we will show it is symmetric.

Notice: $(A^T A)^T = A^T (A^T)^T$

It is easy to see that $(A^T)^T = A$.

Therefore:

$$(A^T A)^T = A^T A$$

So, $B = A^T A$ is symmetric.

Now, we want to show it is positive definite. Let $\mathbf{x} \neq 0$. We will show that

$$\mathbf{x}^T B\mathbf{x} > 0$$

Notice:

$$\mathbf{x}^T B\mathbf{x} = \mathbf{x}^T A^T A\mathbf{x} = (A\mathbf{x})^T A\mathbf{x} = (\|A\mathbf{x}\|_2)^2 \geq 0$$

.

So, $(\|A\mathbf{x}\|_2)^2 = 0 \iff \|A\mathbf{x}\|_2 = 0 \iff A\mathbf{x} = 0$. However, since $\mathbf{x} \neq 0$ and given that A is nonsingular, we have that $A\mathbf{x} \neq \mathbf{0}$.

Therefore, $(\|A\mathbf{x}\|_2)^2 \neq 0 \implies (\|A\mathbf{x}\|_2)^2 > 0 \implies \mathbf{x}^T B\mathbf{x} > 0$. So, B is positive definite.

Therefore, B is symmetric positive definite.

b)

We will show that the 2-norm condition number of B is the square of the 2-norm condition number of A.

Let $A = U\Sigma V$ where we have taken the SVD of A.

Recall: Σ is going to be a diagonal matrix with the diagonal entries being the singular values (σ_is).

Then,

$$\kappa_2(A) = \frac{\max\limits_i \sigma_i}{\min\limits_i \sigma_i}$$

.

Notice: $A^TA = (U\Sigma V)^T(U\Sigma V) = V^T\Sigma^T U^T U\Sigma V = V^T(\Sigma^T\Sigma)V$ since $U^TU = I$.

This is the SVD of A^TA and the singular values are now the diagonal entries of the diagonal matrix $\Sigma^T\Sigma$. The diagonal entries of $\Sigma^T\Sigma$ is just going to be the square of the diagonal entries of Σ, (σ_i)s. Therefore, the singular values of A^TA are the square of the singular values of A.

Therefore, the maximum singular value of A^TA is just going to be the maximum singular value of A squared. The same will hold for the minimum singular value.

Thus,

$$\kappa_2(B) = \kappa_2(A^TA) = \frac{(\max\limits_i \sigma_i)^2}{(\min\limits_i \sigma_i)^2} = (\frac{\max\limits_i \sigma_i}{\min\limits_i \sigma_i})^2 = \kappa_2(A)^2.$$

Note: This could also be solved using the fact that $\|A\|_2 = \sqrt{\rho(A^TA)}$.

c)

Applying C-G to the two-dimensional Poisson problem $Au = f$ on an m by m grid (with second order centered differencing) requires $\mathcal{O}(m^3)$ work to converge to a fixed tolerance.

Now, suppose we multiply both sides by A^T to get $A^TAu = A^Tf \implies Bu = f$ for $B = A^TA$.

We have shown above that the $\kappa_2(B) = \kappa_2(A)^2$.

The book tells us that the maximum number of iterations required $k = \mathcal{O}\sqrt{\kappa}$. The standard second order discretization of the Poisson problem on a grid with m points gives a matrix with $\kappa = \mathcal{O}(\frac{1}{h^2})$ with $h = \frac{1}{m+1}$.

Since,

$$\kappa_2(B) = \kappa_2(A)^2 = \mathcal{O}(\frac{1}{h^4})$$

.

The bound suggests that Conjugate Gradient using B would require $\mathcal{O}(\sqrt{m^4}) = \mathcal{O}(m^2)$ iterations. We also require m^2 work per iteration to compute Ap_{k-1} and so the total work required is $\mathcal{O}(m^4)$ to reach a fixed tolerance.

d) We want to bring the error down to $\mathcal{O}(h^2) = \mathcal{O}(\frac{1}{m^2})$. We know that the error bound is provided by:

$$2e^{-\frac{2k}{\sqrt{\kappa}}}$$

where k is the number of iterations.

Suppose we want to bring the error down to $\epsilon = Ch^2$. Then, $k = (\log(C) + 2\log(h))\sqrt{\kappa}$.

When we use the the matrix A, $\kappa = \mathcal{O}(\frac{1}{h^2}) \implies \sqrt{\kappa} = \mathcal{O}(\frac{1}{h})$.

Since, $h = \frac{1}{m+1}$ we get the following number of iterations:

$$k = (\log(C) + 2\log(\frac{1}{m}))\mathcal{O}(m) = \mathcal{O}(\log(m)m)$$

since, $\log(\frac{1}{m}) = \log(1) - \log(m)$.

We need m^2 work per iteration and so using A the total work required to bring the error down to ϵ is $\mathcal{O}(m^3 \log(m))$.

When we use B, the number of iterations required becomes,

$$k = \mathcal{O}(\log(m)m^2)$$

and the work required becomes,

$$\mathcal{O}(\log(m)m^4)$$

since we still require m^2 work per iteration.

Problem 3. As in the previous homework, consider the one-dimensional BVP

$$\frac{d}{dx}\left(\kappa(x)u'(x)\right) = 0$$

on $0 \le x \le 1$ with Dirichlet boundary conditions $u(0) = 0$ and $u(1) = 1$, again discretizing this problem using the system (2.71) in the text. (Or negate it if you prefer, to make it positive definite.)

Consider the piecewise constant diffusivity

$$\kappa(x) = \begin{cases} \epsilon & \text{if } x < 0.5, \\ 1 & \text{if } x > 0.5. \end{cases}$$

where $\epsilon > 0$.

(a) Generalizing what you did in HW5, determine the exact solution, in terms of the parameter ϵ.

(c) Implement the conjugate gradient method for this problem. For the convergence test require $\|r^{[k]}\|_2 < 10^{-14}$. Allow more than m iterations, if necessary.

Make semilogy plots of the max-norm of the error and the 2-norm of the residual as a function of iteration k for the case $m = 19$ with $\epsilon = 0.1$. Also try $\epsilon = 10^{-3}$. You should observe that more than m iterations are required to get good results. Comment on the behavior of the iterates in each case.

(d) Implement the preconditioned C-G algorithm (PCG) using the diagonal preconditioner and observe that this greatly improves the convergence behavior.

Note: Make sure you do this in a way for which M is symmetric positive definite and not negative definite, as discussed in the notebook `PCG.ipynb` and video that goes with it. This also contains corrections to some typos in the PCG algorithm written on page 95.

The notebook `DarcyFlow.ipynb` provides an implementation of the PCG algorithm for the two dimensional version of this problem that may be useful to follow.

Solution:

a) Consider the one-dimensional BVP

$$\frac{d}{dx}\left(\kappa(x)u'(x)\right) = 0$$

on $0 \le x \le 1$ with Dirichlet boundary conditions $u(0) = 0$ and $u(1) = 1$, for the piecewise constant diffusivity

$$\kappa(x) = \begin{cases} \epsilon & \text{if } x < 0.5, \\ 1 & \text{if } x > 0.5. \end{cases}$$

where $\epsilon > 0$.

We want to find an exact solution in terms of the parameter ϵ.

Notice:

$$\frac{d}{dx}\left(\kappa(x)u'(x)\right) = 0 \implies \kappa(x)u'(x) = c$$

for some constant c.

If $x < 0.5$, the we have $\epsilon u'(x) = c$.

Dividing by $\kappa = \epsilon$ and integrating:

$$\int_0^x u'(t)\ dt = \int_0^x \frac{c}{\epsilon}\ dt$$

$$\implies u(x) - u(0) = \frac{c}{\epsilon}x$$

We know that $u(0) = 0$ and so $u(x) = \frac{c}{\epsilon}x$ when $x < 0.5$.

Now suppose $x > 0.5$.

Then,

$$u(x) = \int_0^{0.5} \frac{c}{\epsilon}\ dt + \int_{0.5}^x c\ dt$$

$$\implies u(x) = \frac{0.5c}{\epsilon} + cx - 0.5c$$

We know that $u(1) = 1$ and so $\frac{0.5c}{\epsilon} + c - 0.5c = 1 \implies c(\frac{0.5}{\epsilon} + 0.5) = 1$

$$\frac{0.5}{\epsilon} + 0.5 = \frac{1}{2\epsilon} + \frac{1}{2} = \frac{1+\epsilon}{2\epsilon}$$

Therefore,

$$c(\frac{0.5}{\epsilon} + 0.5) = 1 \implies c = \frac{1}{\frac{0.5}{\epsilon} + 0.5} = \frac{1}{\frac{1+\epsilon}{2\epsilon}} = \frac{2\epsilon}{1+\epsilon}$$

So,

$$u(x) = \begin{cases} \frac{2}{(1+\epsilon)}x & \text{if } x < 0.5, \\ \frac{1}{(1+\epsilon)} + \frac{2\epsilon x}{1+\epsilon} - \frac{\epsilon}{1+\epsilon} & \text{if } x > 0.5. \end{cases}$$

$$\implies u(x) = \begin{cases} \frac{2}{(1+\epsilon)}x & \text{if } x < 0.5, \\ \frac{1}{(1+\epsilon)}(2\epsilon x + 1 - \epsilon) & \text{if } x > 0.5. \end{cases}$$

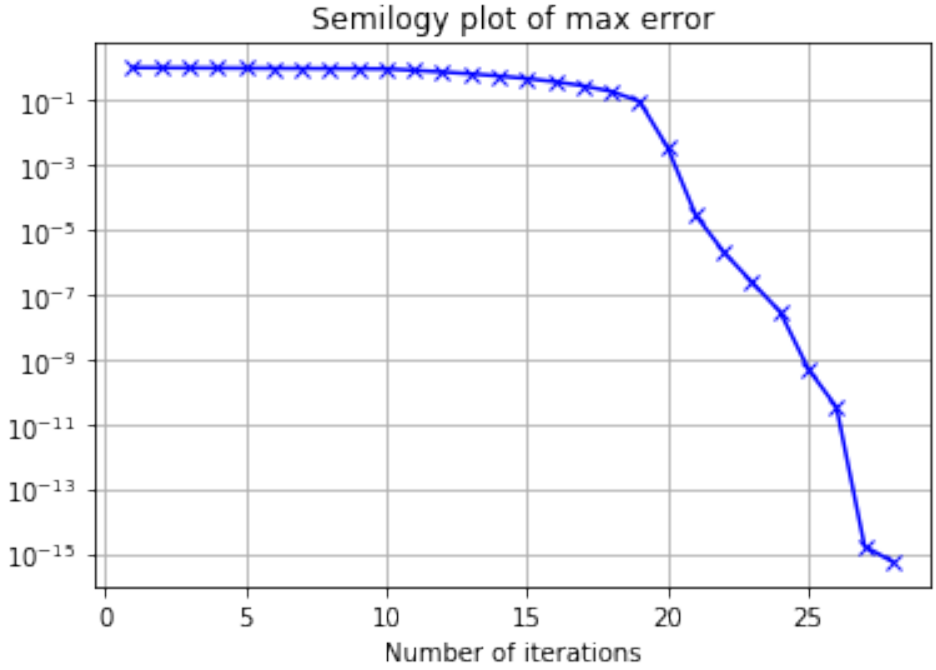
b)

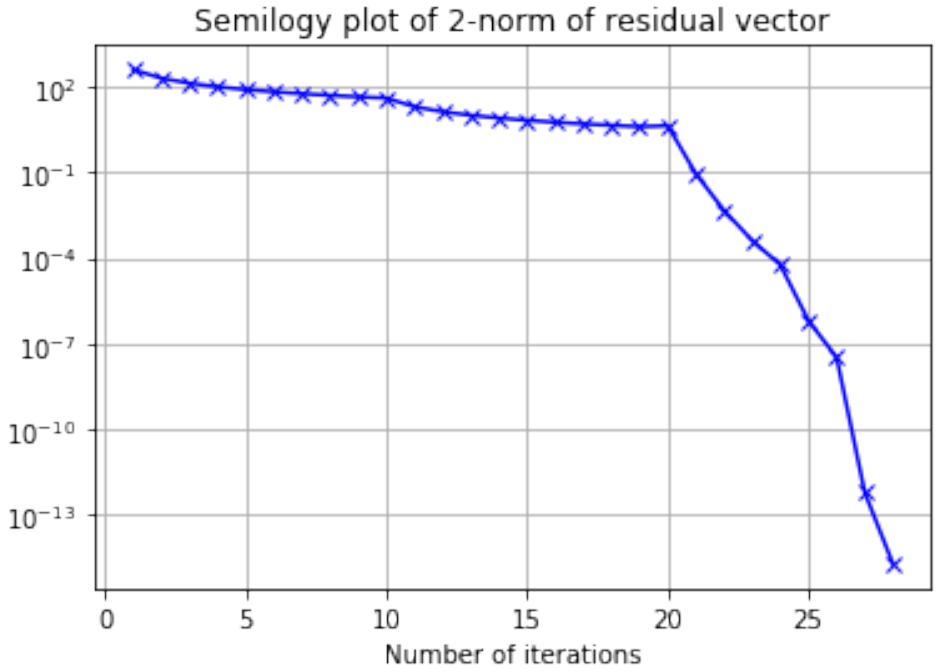
Let the tolerance level be $\|r^{[k]}\|_2 < 10^{-14}$, $m = 19$ and $\epsilon = 0.1$.

We produce semilogy plots of the max-norm of the error and the 2-norm of the residual as a function of iteration k.

We require 27 iterations to reach the desired level of tolerance. We expect C-G should converge in m iterations but it takes longer than 19 iterations for it to converge to the desired level. We can see that the max norm of the error is constant for the first 15 or so iterations and starts to really decrease after 17 iterations. The 2-norm of the residual vector also decreases more after about 20 iterations. Initially, the right side of the graph seems to do better. The left side seems to be stuck at the intial guess for the first few iterations as the right piecewise part moves closer to the right piecewise part of

the actual solution. So, the solution seems to do a better job for the higher value of κ at the beginning.



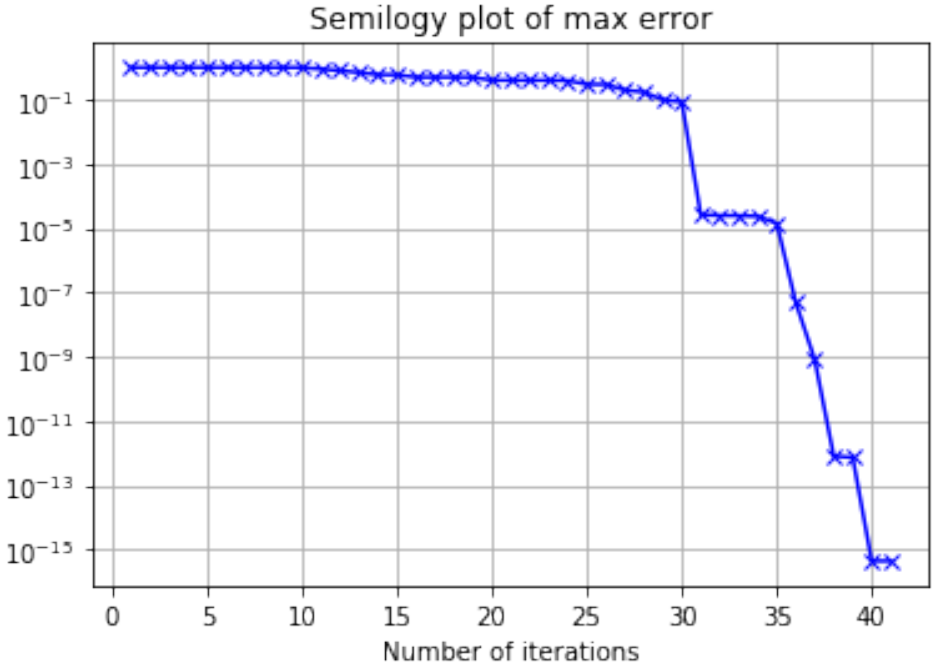


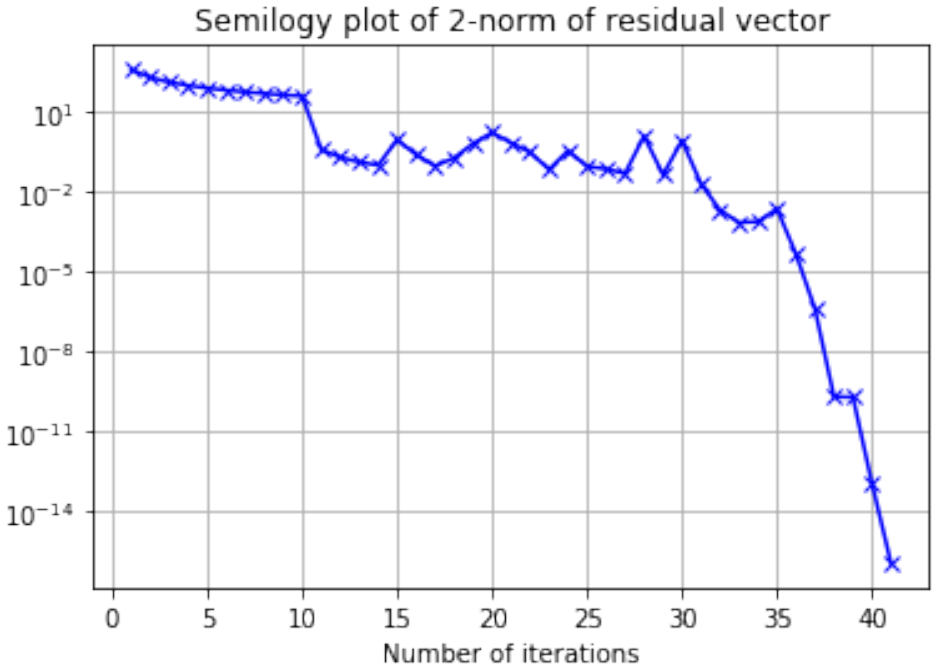
Let $\epsilon = 10^{-3}$.

We produce semilogy plots of the max-norm of the error and the 2-norm of the residual as a function of iteration k.

Again, we require more than m iterations for the solution to converge to reach the fixed tolerance level. We take 40 iterations to converge for this particular value of ϵ. This gives an indication that ϵ is responsible for the change because that is the only thing we changed in the system. We see that the max norm of the error does not decrease as much for the first 30 iterations after which it starts

decreasing steeply. However, for a few sequence of iterations the error does not decrease as much or seems to remain constant. For example, during iteration 31-35 the error does not seem to change much. The 2-norm of the residual error is even more erratic. It decreases and increases erratically until about 35 iterations after it starts decreasing steeply. Again, we see a similar behavior between the left and right piecewise piece. The right piece seems to do a better job in the beginning and the left piece seems to be stuck in the intial guess for a few iterations. Even after the left piece starts moving up to the actual solution, the right piece does a better job and achieves plotting accuracy before the left piece. In the previous case, the left piece caught up with the right piece but in this case the convergence of the left piece is much slower. Even after the right piece achieves plotting accuracy, the left piece takes a while to achieve the same level of accuracy. So, higher value of κ seems to do better.



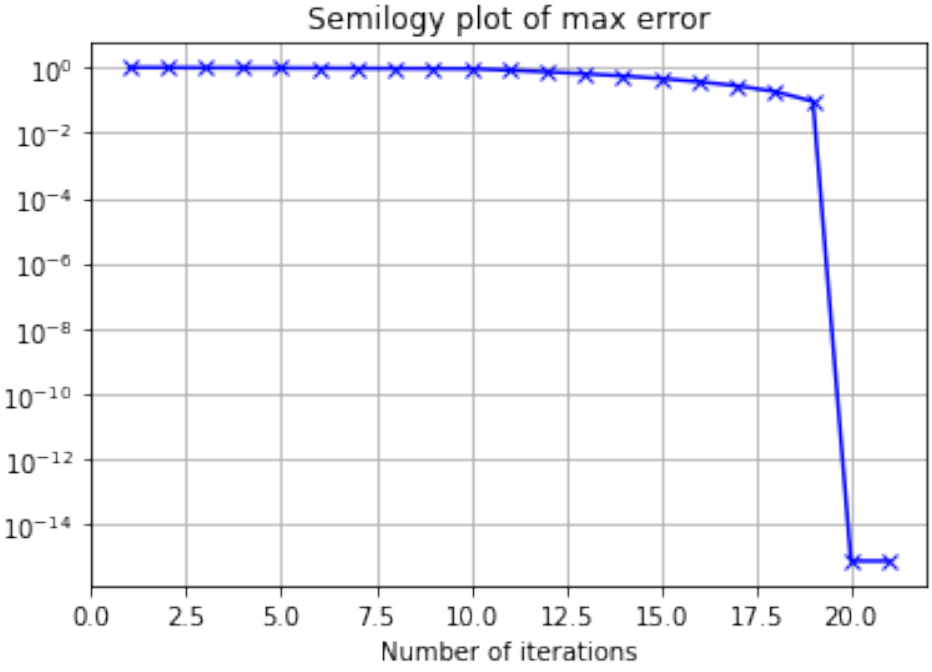


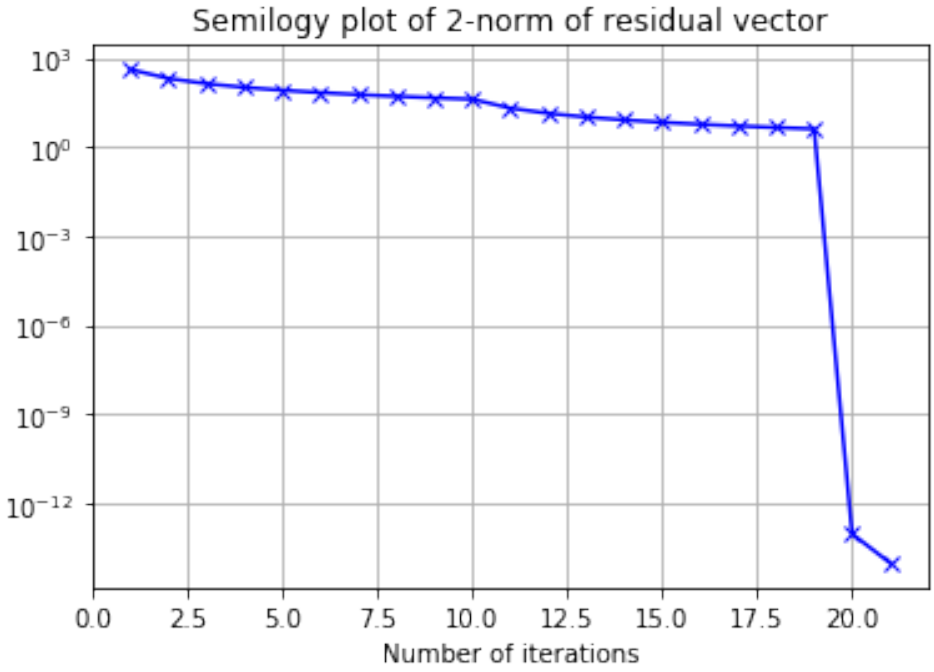
c)

We use C-G algorithm using the diagonal preconditioner and observe that this greatly improves the convergence behavior. Let $\epsilon = 0.1$.

We produce semilogy plots of the max-norm of the error and the 2-norm of the residual as a function of iteration k.

We see that it only takes about 20 iterations to converge to the desired level of tolerance. $m = 19$ and so we converge in about m iterations. The interesting behavior we notice is that both the max norm of the error and the 2-norm of the residual does not decrease much for the first 18 iterations after which it drops rapidly to the desired level of tolerance. We also notice the same behavior on the right and left piece of the piecewise function. The right side seems to do better initially.

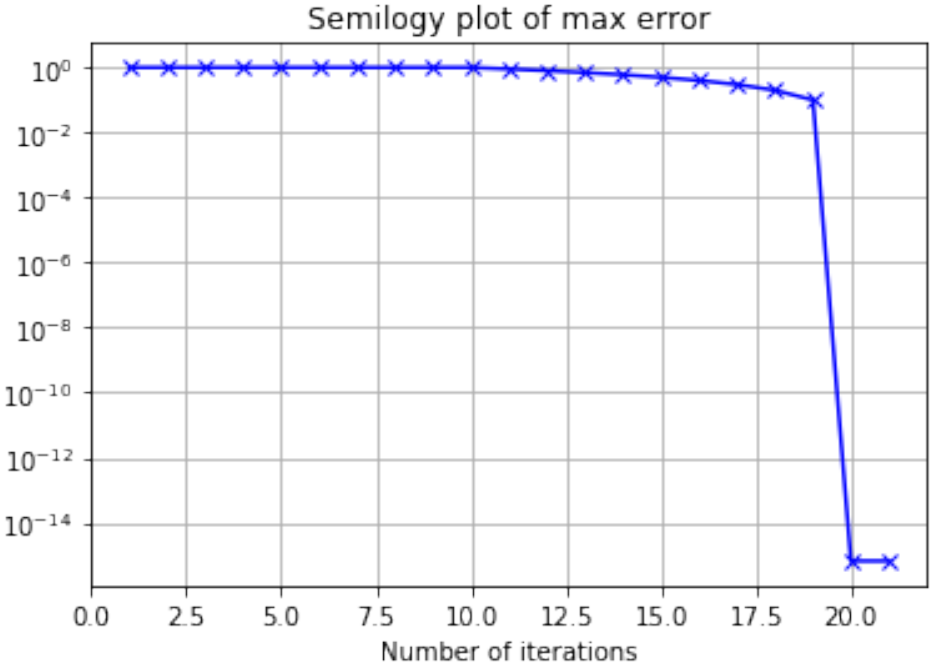


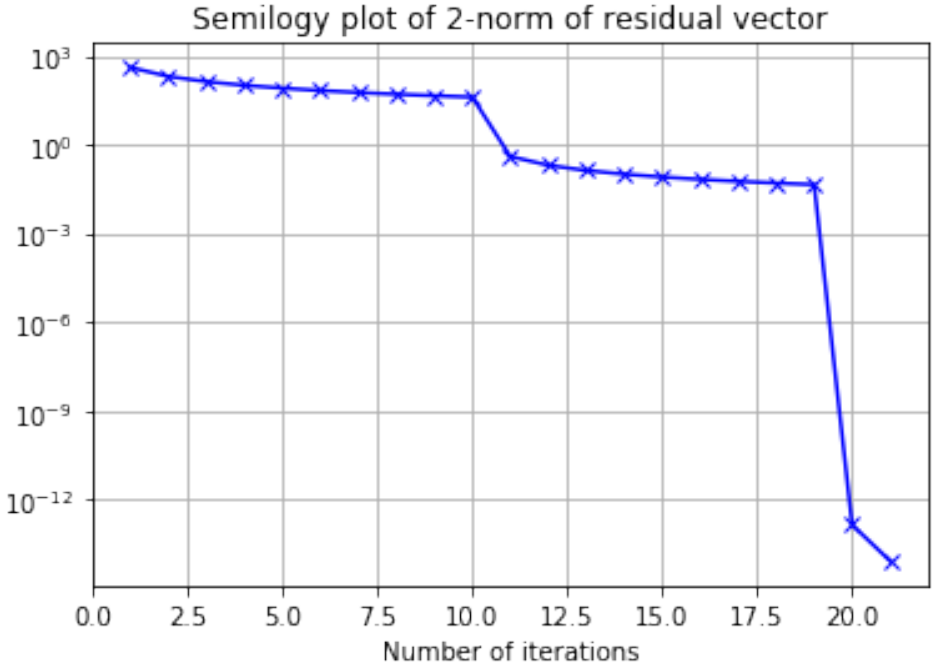


Let $\epsilon = 10^{-3}$.

We produce semilogy plots of the max-norm of the error and the 2-norm of the residual as a function of iteration k.

Again, we are able to converge in 20 iterations with $m = 19$. We see a similar behavior with the max norm as it does not decrease much for 18 iterations but then drops rapidly. The 2 norm of the residual vector actually has a more interesting shape. It decreases steeply at iteration 10, then decreases less and then has a sudden drop at iteration 18. We observe that right piece of the piecewise solution does better as it achieves plotting accuracy first. Larger values of κ seem to do better.


Semilogy plot of max error
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Number of iterations


Semilogy plot of 2-norm of residual vector
10^{3}
10^{0}
10^{-3}
10^{-6}
10^{-9}
10^{-12}
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Number of iterations

Problem 4. Consider the same problem as in Problem 3 but now on the interval $0 \le x \le 4$ with Dirichlet boundary conditions and with $m = 3$ internal grid points (so $h = 1$ for convenience). Now put the jump in κ at the midpoint $x = 2$:

$$\kappa(x) = \begin{cases} \epsilon & \text{if } x < 2, \\ 1 & \text{if } x > 2. \end{cases}$$

(a) Write out the 3×3 matrix A explicitly in this case.

(b) Write out the matrix M that would be used as the "diagonal preconditioner" in this case. Also compute $B = M^{-1}A$ and observe that it is not symmetric.

(c) In this case we can choose C to be $\text{diag}(\sqrt{M_{ii}})$. Write out the matrix $\tilde{A} = C^{-T}AC^{-1}$ in this 3×3 case and observe that it is symmetric.

(d) For the case $\epsilon = 10^{-4}$ compute the eigenvalues and 2-norm condition number of A and B (recall that those of $\tilde{A}$ agree with those of B, but B is easier to work with). You can use the `eig` function in Numpy or Matlab, or do it by hand.

(e) Note that as $\epsilon \to 0$ the matrix A approaches a singluar matrix and the condition number blows up. What does the condition number of B approach as $\epsilon \to 0$? (You should be able to compute this analytically by looking at the limiting matrix.)

Solution:

a) Let $m = 3$. Then, $h = \frac{4}{m+1} = 1$.

The gridpoints are $x_0 = 0, x_1 = 1, x_2 = 2, x_3 = 3$ and $x_4 = 4$. Let

$$\kappa(x) = \begin{cases} \epsilon & \text{if } x < 2, \\ 1 & \text{if } x > 2. \end{cases}$$

.

We will write out the 3×3 matrix A explicitly in this case.

$$A = \tfrac{1}{h^2} = 1 \begin{bmatrix} -(\kappa_{\frac{1}{2}} + \kappa_{\frac{3}{2}}) & \kappa_{\frac{3}{2}} & 0 \\ \kappa_{\frac{3}{2}} & -(\kappa_{\frac{3}{2}} + \kappa_{\frac{5}{2}}) & \kappa_{\frac{5}{2}} \\ 0 & \kappa_{\frac{5}{2}} & -(\kappa_{\frac{5}{2}} + \kappa_{\frac{7}{2}}) \end{bmatrix}$$

$$= \begin{bmatrix} -(\epsilon + \epsilon) & \epsilon & 0 \\ \epsilon & -(\epsilon + 1) & 1 \\ 0 & 1 & -2 \end{bmatrix}$$

$$= \begin{bmatrix} -2\epsilon & \epsilon & 0 \\ \epsilon & -(\epsilon + 1) & 1 \\ 0 & 1 & -2 \end{bmatrix}$$

b)

We will write out the matrix M that could be used as the "diagonal preconditioner".

So, let $M = \text{diag}(-A_{ii})$ since A is symmetric negative definite. This is because any symmetric negative definite matrix will have negative diagonal entries.

$$M = \begin{bmatrix} 2\epsilon & 0 & 0 \\ 0 & (\epsilon+1) & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Since, M is a diagonal matrix, M^{-1} is just an diagonal matrix with the reciprocals of the diagonal entries of M.

Then,

$$M^{-1} = \begin{bmatrix} \frac{1}{2\epsilon} & 0 & 0 \\ 0 & \frac{1}{(\epsilon+1)} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$$

Now, let $B = M^{-1}A$.

Then,

$$B = \begin{bmatrix} \frac{1}{2\epsilon} & 0 & 0 \\ 0 & \frac{1}{(\epsilon+1)} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} -2\epsilon & \epsilon & 0 \\ \epsilon & -(\epsilon+1) & 1 \\ 0 & 1 & -2 \end{bmatrix} = \begin{bmatrix} -1 & \frac{1}{2} & 0 \\ \frac{\epsilon}{(\epsilon+1)} & -1 & \frac{1}{(\epsilon+1)} \\ 0 & \frac{1}{2} & -1 \end{bmatrix}$$

We can already see that the matrix is not symmetric in general. It will be symmetric if $\epsilon = 1$ but then our $\kappa(x)$ would not be piecewise constant.

c)

Let $C = \text{diag}(\sqrt{M_{ii}})$.

Then,

$$C = \begin{bmatrix} \sqrt{2\epsilon} & 0 & 0 \\ 0 & \sqrt{(\epsilon+1)} & 0 \\ 0 & 0 & \sqrt{2} \end{bmatrix}$$

C is a diagonal matrix and so C^{-1} is just a diagonal matrix with the diagonal entries being the reciprocal of the diagonal entries of C.

$$C^{-1} = \begin{bmatrix} \frac{1}{\sqrt{2\epsilon}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{(\epsilon+1)}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Notice: $(C^{-1})^T = C^{-1}$.

Let $\tilde{A} = C^{-T}AC^{-1}$. Then,

$$\tilde{A} = \begin{bmatrix} \frac{1}{\sqrt{2\epsilon}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{(\epsilon+1)}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} -2\epsilon & \epsilon & 0 \\ \epsilon & -(\epsilon+1) & 1 \\ 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2\epsilon}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{(\epsilon+1)}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{2\epsilon}{\sqrt{2\epsilon}} & \frac{\epsilon}{\sqrt{2\epsilon}} & 0 \\ \frac{\epsilon}{\sqrt{(\epsilon+1)}} & -\frac{\epsilon+1}{\sqrt{(\epsilon+1)}} & \frac{1}{\sqrt{(\epsilon+1)}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{2}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2\epsilon}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{(\epsilon+1)}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$= \begin{bmatrix} -1 & \frac{\epsilon}{\sqrt{2\epsilon}\sqrt{(\epsilon+1)}} & 0 \\ \frac{\epsilon}{\sqrt{2\epsilon}\sqrt{(\epsilon+1)}} & -1 & \frac{1}{\sqrt{2}\sqrt{(\epsilon+1)}} \\ 0 & \frac{1}{\sqrt{2}\sqrt{(\epsilon+1)}} & -1 \end{bmatrix}$$

Thus, $\tilde{A}$ is symmetric.

d)

Let $\epsilon = 10^{-4}$. We will compute the eigenvalues of A and B and the 2-norm condition number of A and $\tilde{A}$.

Eigenvalues of A:

$$\begin{gathered} -2.618 \\ -0.3820 \\ -1.99998 \times 10^{-4} \end{gathered}$$

Eigenvalues of B:

$$\begin{gathered} \text{-1} \\ \text{-1.7071} \\ \text{-0.2929} \end{gathered}$$

We can notice that the eigenvalues of B are the same eigenvalues as $\tilde{A}$.

Condiition number of A: 1.3092×10^4

Condition number of $\tilde{A}$: 5.8284

d)

We will compute the condition number of $\tilde{A}$ as $\epsilon \to 0$.

Notice:

$$\lim_{\epsilon \to 0} \frac{\epsilon}{\epsilon + 1} = 0$$

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon + 1} = 1$$

So the limiting matrix of B becomes:

$$\begin{bmatrix} -1 & \frac{1}{2} & 0 \\ 0 & -1 & 1 \\ 0 & \frac{1}{2} & -1 \end{bmatrix}$$

We seek to find the eigenvalues of this matrix B. These will be the same as the eigenvalues of $\tilde{A}$. We need $\det(B - \lambda I) = 0$.

Therefore,

$$\det(\begin{bmatrix} -1-\lambda & \frac{1}{2} & 0 \\ 0 & -1-\lambda & 1 \\ 0 & \frac{1}{2} & -1-\lambda \end{bmatrix}) = 0$$

$$\implies (-1-\lambda)((-1-\lambda)^2 - \frac{1}{2}) - \frac{1}{2}(0) + 0 = 0$$

$$\implies (-1-\lambda)(-1-\lambda)^2 - \frac{1}{2}) = 0$$

So, $\lambda = -1$ or $(-1-\lambda)^2 - \frac{1}{2} = 0$.

$$\lambda^2 + 2\lambda + \frac{1}{2} = 0 \implies \lambda = \frac{-2 \pm \sqrt{2}}{2}$$

So, $\lambda = -1 + \frac{\sqrt{2}}{2}, -1 - \frac{\sqrt{2}}{2}$.
Therefore, the eigenvalues are:

$$-1 + \frac{\sqrt{2}}{2} - 1 - \frac{\sqrt{2}}{2} - 1$$

Since, these are going to be the same eigenvalues of $\tilde{A}$ and given that $\tilde{A}$ is symmetric:

$$\kappa_2(\tilde{A}) = \frac{\max\limits_i |\lambda_i|}{\min\limits_i |\lambda_i|} = \frac{1 + \frac{\sqrt{2}}{2}}{1 - \frac{\sqrt{2}}{2}} \approx 5.8284$$