1. Put an equidistant mesh \( a = x_0 < x_1 < \cdots < x_N = b \) on the interval \( \Omega = [a, b] \) and let \( \phi_j \) be the continuous piecewise-linear nodal basis functions satisfying \( \phi_j(x_i) = \delta_{ij} \), with boundary conditions \( Au(a) - Bu(b) = 0 \). Suppose \( N = 2^n \) for some \( n \geq 0 \) and let \( \theta_{kj} \) be the hierarchical basis functions for the same solution space for \( 0 \leq k \leq n, 0 \leq j \leq 2^k \) or so. Express each \( \theta_{kj} \) as (a) scaled and shifted versions of a single function \( \theta \) and (b) linear combinations of the neighboring finer-level functions \( \theta_{k+1,p} \) for \( |p - 2j| \leq 1 \). Explain what the coefficients in the linear combination mean (hint: smoothing). Carry out the dual computations for the discontinuous piecewise-constant basis \( \tau_j \). As a result, write the matrix \( L_h \) in the original basis in factored form and explain the significance of the factors.

Define the function

\[
\theta(x) = \max\{0, 1 - |x|\}.
\]

Then for \( x \in \Omega \),

\[
(1) \quad \theta_{kj}(x) = \theta\left(\frac{2^k}{b-a} \left(x - \left(a + j \frac{b-a}{2^k}\right)\right)\right)
\]

for \( 0 \leq k \leq n \) and \( 0 \leq j \leq 2^k \), and

\[
(2) \quad \theta_{kj}(x) = \frac{1}{2}\theta_{k+1,2j-1}(x) + \theta_{k+1,2j}(x) + \frac{1}{2}\theta_{k+1,2j+1}(x)
\]

for \( 0 < k < n \) and \( 0 < j < 2^k \). The hierarchical basis of continuous piecewise-linear nodal basis functions (ignoring the boundary conditions) is given by the set

\[
\{\theta_{00}, \theta_{01}\} \cup \{\theta_{kj} | 0 < k \leq n, 0 < j < 2^k, j \text{ odd}\}.
\]

In this framework, the standard basis functions \( \phi_j \) (such that \( \phi_j(x_i) = \delta_{ij} \)) are exactly the \( \theta_n \) functions for \( 0, n \leq 2^n = N \).

Note that for \( j \) even, \( \theta_{kj} \) is not part of the hierarchical basis. The relation (2) represents \( \theta_{kj} \) as a kind of weighted average of the hat functions on the next coarsest level, and can be viewed as a smoothing process. The stencil \( \frac{1}{2}[1 \ 2 \ 1] \) is not exactly a weighted average, and this is due to each level having a different element size (differing by a factor of two).

In order to construct a corresponding dual basis of discontinuous piecewise-constant functions, we define the function

\[
\psi(x) = \begin{cases} 
1 & \text{if } x \in [-1, 0] \\
-1 & \text{if } x \in (0, 1] \\
0 & \text{else}
\end{cases}
\]

Then for \( x \in \Omega \), let

\[
\psi_{kj}(x) = \psi\left(\frac{2^k}{b-a} \left(x - \left(a + j \frac{b-a}{2^k}\right)\right)\right)
\]

1
for $0 < k \leq n$ and $0 < j < 2^k$, and so

$$\psi_{kj}(x) = \psi_{k+1,2j-1}(x) + 2\psi_{k+1,2j}(x) + \psi_{k+1,2j+1}(x)$$

for $0 < k < n$ and $0 < j < 2^k$. Noting $\psi_{01}(x) = 1$ for $x \in \Omega$, the hierarchical basis of discontinuous piecewise-constant functions is given by the set

$$\{\psi_{01}\} \cup \{\psi_{kj} \mid 0 < k \leq n, 0 < j < 2^k, j \text{ odd}\}.$$

Here, the basis functions are mutually orthogonal and all, except $\psi_{01}$, integrate to 0. Given the standard basis functions

$$\tau_j(x) = \begin{cases} 
1 & \text{if } x \in [x_j, x_{j+1}] \\
0 & \text{else}
\end{cases}$$

for $0 \leq j < N$, the hierarchical basis functions can be written as linear combinations of $\tau_j$ such that all the coefficients are 0, 1, or $-1$ (like a difference operator).

Here are the bases for the solution and dual spaces (standard on left, hierarchical on right) in the case $N = 4$:
For $N = 8$, the change of basis matrices are

\[
P \Phi = \begin{bmatrix}
1 & 7 & 3 & 5 & 1 & 3 & 1 & 1 & 0 \\
0 & 8 & 4 & 1 & 7 & 3 & 1 & 1 & 1 \\
1 & 0 & 8 & 4 & 1 & 7 & 3 & 1 & 1 \\
0 & 1 & 0 & 8 & 4 & 1 & 7 & 3 & 1 \\
0 & 0 & 1 & 0 & 8 & 4 & 1 & 7 & 3 \\
0 & 0 & 0 & 1 & 0 & 8 & 4 & 1 & 7 \\
0 & 0 & 0 & 0 & 1 & 0 & 8 & 4 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 8 & 1
\end{bmatrix}
\begin{bmatrix}
\phi_0 \\
\phi_1 \\
\phi_2 \\
\phi_3 \\
\phi_4 \\
\phi_5 \\
\phi_6 \\
\phi_7 \\
\phi_8
\end{bmatrix} = \begin{bmatrix}
\theta_{00} \\
\theta_{01} \\
\theta_{11} \\
\theta_{21} \\
\theta_{23} \\
\theta_{31} \\
\theta_{33} \\
\theta_{35} \\
\theta_{37}
\end{bmatrix} = \Theta
\]

and

\[
Q' \tau = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\
1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & -1 \\
1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1
\end{bmatrix}
\begin{bmatrix}
\tau_0 \\
\tau_1 \\
\tau_2 \\
\tau_3 \\
\tau_4 \\
\tau_5 \\
\tau_6 \\
\tau_7
\end{bmatrix} = \begin{bmatrix}
\psi_{01} \\
\psi_{11} \\
\psi_{21} \\
\psi_{23} \\
\psi_{31} \\
\psi_{33} \\
\psi_{35} \\
\psi_{37}
\end{bmatrix} = \Psi.
\]

Then if $(L_h)_{ij} = a(\tau_i, \phi_j)$, $0 \leq i \leq N$ and $0 \leq j < N$ and $(L_h)_{N0} = A$, $(L_h)_{NN} = -B$, we modify $Q'$ slightly, defining

\[
Q = \begin{bmatrix}
Q' & 0
\end{bmatrix}
\]

so that $Q$ has dimension $(N + 1) \times (N + 1)$ and premultiplying $L_h$ by $Q$ will not affect the boundary conditions. Thus, we can rewrite the discrete problem $L_h u_h = F_h$ as

\[
\hat{L}_h \hat{u}_h = (QL_hP)(P^{-1}u_h) = QF_h = \hat{F}_h
\]

and solve a better conditioned system (although not as sparse) for $\hat{u}_h$ (multiplying by $P$ to obtain $u_h$).

2. (3.14) Let $\Omega = (0,1)^2$. Consider the problem $-\Delta u + u = 1$ in $\Omega$ and $u|_{\partial \Omega} = 0$. Approximate its solution with $P_1$, $H^1$-conforming elements.

(i) Let $\{\lambda_0, \lambda_1, \lambda_2\}$ be the barycentric coordinates in the triangle $K_h$ shown in the figure. Compute the entries of the elementary stiffness matrix $A_{ij} = \int_{K_h} \nabla \lambda_i \cdot \nabla \lambda_j + \int_{K_h} \lambda_i \lambda_j$, and the right-hand side vector $\int_{K_h} \lambda_i$.

Let $v_0 = (x_1^0, x_2^0)$ be vertex 1 (at the right angle) of triangle $K_h$. Then defining $v_1 = (x_1^0 + h, x_2^0)$ and $v_2 = (x_1^0, x_2^0 + h)$, we have the barycentric coordinates $(\lambda_0, \lambda_1, \lambda_2)$ given by

\[
\lambda_0(x_1, x_2) = 1 - \frac{1}{h}(x_1 - x_1^0) - \frac{1}{h}(x_2 - x_2^0)
\]

3
\[ \lambda_1(x_1, x_2) = \frac{1}{h}(x_1 - x_2^2) \]
\[ \lambda_2(x_1, x_2) = \frac{1}{h}(x_2 - x_1^2) \]

for \((x_1, x_2) \in K_h\) so that \(v_0 = (1, 0, 0), v_1 = (0, 1, 0),\) and \(v_2 = (0, 0, 1).\) Then
\[
\nabla \lambda_0 = \left( -\frac{1}{h}, -\frac{1}{h} \right), \quad \nabla \lambda_1 = \left( \frac{1}{h}, 0 \right), \quad \text{and} \quad \nabla \lambda_2 = \left( 0, \frac{1}{h} \right).
\]

Elements of the 3 \times 3 elementary stiffness matrix are given by
\[
A_{ij} = \int_{K_h} \nabla \lambda_i \cdot \nabla \lambda_j + \int_{K_h} \lambda_i \lambda_j.
\]

Thus, to compute \(A_{00},\) we compute the integral of \(\nabla \lambda_0 \cdot \nabla \lambda_0,\) a constant, and the integral of \(\lambda_0^2,\) a quadratic polynomial. We can compute the second integral exactly using numerical quadrature of three points (see Table 8.2 in the text).

Thus,
\[
A_{00} = \int_{K_h} \frac{2}{h^2} + \frac{\hbar^2}{6} \left( \frac{1}{4} + \frac{1}{4} + 0 \right) = 1 + \frac{\hbar^2}{12}.
\]

The other entries are computed similarly, yielding
\[
A = \begin{bmatrix}
1 + \frac{\hbar^2}{12} & -\frac{1}{2} + \frac{\hbar^2}{24} & -\frac{1}{2} + \frac{\hbar^2}{24} \\
-\frac{1}{2} + \frac{\hbar^2}{24} & \frac{1}{2} + \frac{\hbar^2}{12} & \frac{\hbar^2}{24} \\
-\frac{1}{2} + \frac{\hbar^2}{24} & \frac{\hbar^2}{24} & \frac{1}{2} + \frac{\hbar^2}{12}
\end{bmatrix}.
\]

The corresponding values of the right hand side are given by \(\int_{K_h} \lambda_j \cdot 1,\) the integral of a linear polynomial which can be computed exactly using one-point quadrature (again see Table 8.2). Since the evaluation point is \((\frac{1}{3}; \frac{1}{3}; \frac{1}{3})\), each of the three integrals is given by
\[
\int_{K_h} \lambda_j = \frac{\hbar^2}{2} \left( \frac{1}{3} \right) = \frac{\hbar^2}{6}.
\]

(ii) Consider the two meshes shown in the figure. Assemble the stiffness matrix and the right-hand side in both cases and compute the solution. For a fine mesh composed of 800 elements, \(u_h(1/2, 1/2) \approx 0.0702.\) Comment.

Since there is only one internal node in each of the meshes, the stiffness matrices are \(1 \times 1.\) The solution \(u_h\) is a multiple of the sole basis function \(\phi,\) i.e. \(u_h = u_0 \phi,\) and must satisfy the equation \(a(u_h, \phi) = (1, \phi),\) or
\[
u_0 \int_{\Omega} \nabla \phi \cdot \nabla \phi + \phi^2 = \int_{\Omega} \phi.
\]
Each mesh consists of 8 elements, so we can evaluate these integrals by partitioning them over the elements, yielding

$$u_0 \sum_{i=1}^{8} \int_{K_i} \nabla \phi|_{K_i} \cdot \nabla \phi|_{K_i} + \phi^2|_{K_i} = \sum_{i=1}^{8} \int_{K_i} \phi|_{K_i}.$$  

In both meshes, each element is an isosceles right triangle with leg length 1/2 (like the triangle $K_h$ in part (i) with $h = 1/2$). In order to use the previous results, we number the nodes of each element counterclockwise such that the right angle node is node 0. In this way, the sole interior node is numbered alternately 1 and 2 in the first mesh and always 0 in the second mesh. The global shape function $\phi$, when restricted to element $K_i$, is exactly $\lambda_j$ where the interior node is numbered $j$ in element $K_i$. We note that in the second mesh, the four elements whose nodes all lie on the boundary contribute nothing to the global shape function (they’re fixed at 0). Since the elements are all the same size and we saw from part (i) that $\lambda_1$ and $\lambda_2$ yield the same results, the equation for the first mesh becomes

$$8u_0 \int_{K_{1/2}} \nabla \lambda_0 \cdot \nabla \lambda_1 + \lambda_1^2 = 8 \int_{K_{1/2}} \lambda_1$$

and the equation for the second mesh becomes

$$4u_0 \int_{K_{1/2}} \nabla \lambda_0 \cdot \nabla \lambda_0 + \lambda_0^2 = 4 \int_{K_{1/2}} \lambda_0.$$  

We computed these integrals for general $h$ in part (i), and since $A_{00} = 1 + h^2/12$, $A_{11} = 1/2 + h^2/12$, and $\int_{K_h} \lambda_j = h^2/6$ for each $j$, we have

$$u_0 \left( 1 + \frac{(1/2)^2}{12} \right) = \frac{(1/2)^2}{6}$$

for the first mesh and

$$u_0 \left( \frac{1}{2} + \frac{(1/2)^2}{12} \right) = \frac{(1/2)^2}{6}$$

for the second mesh. Solving these equations, we find $u_0 = 2/25 = .08$ for the first mesh and $u_0 = 2/49 \approx .04$ for the second mesh. Since $u_0 = u_h(1/2, 1/2)$, we see that the first mesh yields a solution much closer to the finer mesh solution at the central point. Although the second mesh contains 8 elements, since 4 of the elements consist of all boundary nodes, those elements are forced to zero, effectively shrinking the domain. The lesson here, I think, is to never create an element all of whose nodes are constrained by the boundary. The second mesh could work if the elements had more than three degrees of freedom (or at least degrees of freedom which did not lie on the legs of the right triangles).
3.17

(i) Let everything be as in the question. We assume that the global shape functions $\phi_i$ are such that they are unity on a particular triangle vertex and zero at all other vertices, i.e. the shape functions correspond to Lagrange interpolation.

To show that $A_{ij} = \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j \leq 0$ if $i \neq j$, it suffices to show that $\nabla \phi_i \cdot \nabla \phi_j \leq 0$. This is obviously true if the support of $\phi_i$ and $\phi_j$ are disjoint, so assume that they are not. Then $\phi_i$ and $\phi_j$ are both nonzero only on some common triangle $T$. Since $\nabla \phi_i$ is proportional to the outwards pointing normal of the triangle face opposite the vertex that $\phi_i$ is unity on, and similarly for $\phi_j$, it follows by simple geometry that $\nabla \phi_i \cdot \nabla \phi_j = -|\nabla \phi_i||\nabla \phi_j| \cos \theta$, where $\theta$ is the angle between the mentioned faces. Because all such angles are acute on the given mesh, we have that $\nabla \phi_i \cdot \nabla \phi_j \leq 0$.

Fix $i$ and consider $\sum_j A_{ij} = \int_{\Omega} \nabla \phi_i \cdot \nabla (\sum_j \phi_j)$. The function $\sum_j \phi_j$ is unity on most of the domain $\Omega$, except on triangles touching $\partial \Omega$. Hence $\nabla (\sum_j \phi_j)$ is zero except on triangles touching $\partial \Omega$. If the support of $\phi_i$ does not touch $\partial \Omega$ then this implies that $\sum_j A_{ij} = 0$. Otherwise, if $\phi_i$ is nonzero near the boundary, for each triangle $T$ such that $\phi_i|_T \neq 0$, consider $\int_T \nabla \phi_i \cdot \nabla (\sum_j \phi_j)$. By a similar argument used in the previous paragraph, it is straightforward to find that $\nabla \phi_i \cdot \nabla (\sum_j \phi_j) \geq 0$ on $T$. This is essentially because $\sum_j \phi_j$ is equal to 1 minus a shape function situated on a boundary vertex. Summing over all such $T$ gives that $\sum_j A_{ij} \geq 0$ for $i$ such that $\text{spt } \phi_i \cap \partial \Omega \neq \emptyset$. In summary, $\sum_j A_{ij} \geq 0$ for all $i$.

Thus $A$ is an M-matrix.
(ii) Since \( u_h \) is the finite element solution with right-hand side \( f \), we have \( Au_h = f \).
Since \( A \) is an \( M \)-matrix, we may rewrite it as a difference of two positive matrices
\( A = D - M \) where \( D \) is diagonal and \( M \) has zeros along the diagonal. Thus, we
have \( (D - M)u_h = f \) and since \( f \leq 0 \) by supposition,

\[
(Du_h)_i \leq (Mu_h)_i \quad \text{for} \quad 1 \leq i \leq N.
\]

If we let \( u_i = (u_h)_i \), we can rewrite this equation as

\[
u_i \leq \frac{1}{d_{ii}} \sum_{i \neq j} m_{ij} u_j
\]

which is true for each \( i \). Let \( u_k \) be the max value in the vector \( u_h \), then

\[
u_k \leq \frac{1}{d_{kk}} \sum_{k \neq j} m_{kj} u_j \leq \frac{1}{d_{kk}} \sum_{k \neq j} m_{kj} u_k
\]
or

\[
\left(1 - \frac{1}{d_{kk}} \sum_{k \neq j} m_{kj}\right) u_k \leq 0.
\]

Since \( A \) is an \( M \)-matrix, the diagonal entries are greater in absolute value than
the sum of the off-diagonal entries, so \( 0 \leq \frac{1}{d_{ii}} \sum m_{ij} \leq 1 \) for all \( i \). Thus, the
first term in the product above is non-negative, and we have \( u_k \leq 0 \) as long as
\( \frac{1}{d_{ii}} \sum m_{ij} \neq 1 \). Thus, assuming the \( k \)th row sum of \( A \) is strictly positive, since \( u_k \)
is the maximum value of \( u_h \), \( u_h \leq 0 \). Now consider the case that the row sum
does vanish. This implies \((Au_h)_k = f_k\), or

\[
\sum_{j \neq k} m_{kj} u_k - \sum_{j \neq k} m_{kj} u_j = f_k \leq 0
\]

which implies

\[
\sum_{j \neq k} m_{kj} (u_k - u_j) \leq 0
\]

where each \( m_{kj} \) is non-negative. Since we also have \( u_k \geq u_j \) for all \( j \), \( u_h \) must
be a constant function, and with Dirichlet boundary conditions, it must be the
trivial function, and therefore \( u_h \leq 0 \) in this case as well.