MAPL 600 / CMSC 760 Fall 2007
Take-Home Exam 3
Partial Solution

NOTATION: In all of these problems, assume that $\mathbf{A}$ and $\mathbf{M}$ are symmetric and positive definite, with $\mathbf{A} = \mathbf{M} - \mathbf{N}$.

1a. (10) Show that if the rank of the $n \times n$ matrix $\mathbf{R}$ is $k < n$, then the matrix $\mathbf{I} + \mathbf{R}$ has $n - k$ eigenvalues equal to 1.

Answer:
Since $\mathbf{R}$ has rank k, its range is a k-dimensional subspace. Let $\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n$ be an orthonormal basis for the complement of this space.

Then
$$(\mathbf{I} + \mathbf{R})\mathbf{v}_j = \mathbf{v}_j, \quad j > k,$$
so $\mathbf{I} + \mathbf{R}$ has $(n - k)$ eigenvalues equal to 1.

1b. (10) Suppose the rank of $\mathbf{N}$ is $k < n$. How many iterations will cg take to solve the linear system $\mathbf{Ax} = \mathbf{b}$ using $\mathbf{M}$ as a preconditioner? Justify your answer.

Answer:
$\mathbf{M}^{-1}\mathbf{A} = \mathbf{M}^{-1}(\mathbf{M} - \mathbf{N}) = \mathbf{I} - \mathbf{M}^{-1}\mathbf{N}$. The matrix $\mathbf{M}^{-1}\mathbf{N}$ has rank k (since $\mathbf{N}$ does), so the preconditioned matrix has $(n-k)$ eigenvalues equal to 1. Therefore, it has at most $k+1$ distinct eigenvalues, so pcg will converge in at most $k+1$ iterations; i.e., there is a polynomial of degree $k+1$ that is zero at each of the eigenvalues, so the error will be zero.

2a. (10) Prove that
$$(\mathbf{A} - \mathbf{Z}\mathbf{V}^T)^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{Z}(\mathbf{I} - \mathbf{V}^T\mathbf{A}^{-1}\mathbf{Z})^{-1}\mathbf{V}^T\mathbf{A}^{-1}$$
by verifying that the product of this matrix with $\mathbf{A} - \mathbf{Z}\mathbf{V}^T$ is the identity matrix $\mathbf{I}$. Dimensions: $\mathbf{A}$ is $n \times n$, and $\mathbf{Z}$ and $\mathbf{V}$ are $n \times k$ with $k < n$. (In particular, $\mathbf{Z}$ and $\mathbf{V}$ are not square so they have no inverse.) This is called the Sherman-Morrison-Woodbury (SMW) Formula.

Answer: We compute
$$\begin{aligned}
& (\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{Z}(\mathbf{I} - \mathbf{V}^T\mathbf{A}^{-1}\mathbf{Z})^{-1}\mathbf{V}^T\mathbf{A}^{-1})(\mathbf{A} - \mathbf{Z}\mathbf{V}^T) \\
= & \ \mathbf{I} - \mathbf{A}^{-1}\mathbf{Z}\mathbf{V}^T + \mathbf{A}^{-1}\mathbf{Z}(\mathbf{I} - \mathbf{V}^T\mathbf{A}^{-1}\mathbf{Z})^{-1}\mathbf{V}^T - \mathbf{A}^{-1}\mathbf{Z}(\mathbf{I} - \mathbf{V}^T\mathbf{A}^{-1}\mathbf{Z})^{-1}\mathbf{V}^T\mathbf{A}^{-1}\mathbf{Z}\mathbf{V}^T \\
= & \ \mathbf{I} - \mathbf{A}^{-1}\mathbf{Z}\mathbf{V}^T + \mathbf{A}^{-1}\mathbf{Z}(\mathbf{I} - \mathbf{V}^T\mathbf{A}^{-1}\mathbf{Z})^{-1}(\mathbf{I} - \mathbf{V}^T\mathbf{A}^{-1}\mathbf{Z})\mathbf{V}^T \\
= & \ \mathbf{I} - \mathbf{A}^{-1}\mathbf{Z}\mathbf{V}^T + \mathbf{A}^{-1}\mathbf{Z}^T\mathbf{V}^T \\
= & \ \mathbf{I}.
\end{aligned}$$

Therefore, the inverse of $(\mathbf{A}-\mathbf{Z}\mathbf{V}^T)$ is $(\mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{Z}(\mathbf{I}-\mathbf{V}^T\mathbf{A}^{-1}\mathbf{Z})^{-1}\mathbf{V}^T\mathbf{A}^{-1})$.

2b. (10) Write an algorithm that uses the SMW formula and a factorization $\mathbf{M}=\mathbf{L}\mathbf{U}$ to solve the linear system $\mathbf{A}\mathbf{x}=\mathbf{b}$ where $\mathbf{N}=\mathbf{E}\mathbf{E}^T$ and $\mathbf{E}$ is $n\times k$. (Do not factor any other matrices of size $n\times n$, do not use the inverse of any matrices, and make the algorithm as efficient as you can.)

Answer: Using Matlab notation:

- Let $\mathbf{y}=\mathbf{U}\backslash(\mathbf{L}\backslash\mathbf{b})$.
- Let $\mathbf{E}=\mathbf{U}\backslash(\mathbf{L}\backslash\mathbf{Z})$.
- Let $\mathbf{t}=(\mathbf{I}-\mathbf{V}^T\mathbf{E})\backslash(\mathbf{V}^T\mathbf{y})$.
- Then $\mathbf{x}=\mathbf{y}+\mathbf{E}\mathbf{t}$.

Note the use of parentheses to keep the work $O(n^2)$ instead of $O(n^3)$.

3. (10) Use the Gerschgorin theorem to show that eigenvalues of $\mathbf{M}^{-1}\mathbf{A}$ are close to those of $\mathbf{I}$ if $\|\mathbf{N}\|<\epsilon$ and ϵ is small enough. (You may use any matrix norm that you find convenient.) Under this assumption, what can you say about the convergence of cg using $\mathbf{M}$ as a preconditioner for $\mathbf{A}$? How small is "small enough"?

Answer:
Lemma: The eigenvalues of $\mathbf{A}$ are bounded by the maximum rowsum of $|\mathbf{A}|$.
Proof: This is a minor variant of Gershgorin's theorem. Consider the kth eigenvalue, write $\mathbf{A}\mathbf{x}=\lambda_k\mathbf{x}$, (i.e., $\mathbf{x}$ is the corresponding eigenvector) and let the ith component of $\mathbf{x}_k$ be the one with maximum magnitude. Then

$$\lambda_k x_i=\sum_{j=1}^{n} a_{ij}x_j,$$

so

$$|\lambda_k|=|\sum_{j=1}^{n} a_{ij}x_j/x_i|\le|\sum_{j=1}^{n} a_{ij}|\le\sum_{j=1}^{n}|a_{ij}|.$$

This concludes the proof of the lemma.

Now suppose that

$$\|\mathbf{M}^{-1}\|_\infty\|\mathbf{N}\|_\infty\le\delta.$$

Then

$$\mathbf{M}^{-1}\mathbf{A}-\mathbf{I}=-\mathbf{M}^{-1}\mathbf{N},$$

so, by the lemma, the eigenvalues of $\mathbf{M}^{-1}\mathbf{A}-\mathbf{I}$ are bounded above by the rowsums of $|\mathbf{M}^{-1}\mathbf{N}|$, which are bounded by $\|\mathbf{M}^{-1}\mathbf{N}\|_\infty$, which is bounded by

δ. Similarly, the eigenvalues are bounded below by $-\|\mathbf{M}^{-1}\mathbf{N}\|_\infty$. Therefore, the condition number of $\mathbf{M}^{-1}\mathbf{A}$ is at most

$$\kappa = \frac{1+\delta}{1-\delta}.$$

Therefore the error in pcg is reduced by at least

$$\frac{1-\sqrt{\kappa^{-1}}}{1+\sqrt{\kappa^{-1}}},$$

so this will be fast if δ is small.

4. Let $\mathbf{T}$ be a *Toeplitz matrix*. Such a matrix is specified by a set of parameters $\boldsymbol{\beta}$ so that $t_{ij} = \beta_{i-j}$, $i, j = 1, \ldots, n$.

A *circulant matrix* is defined by specifying the first row of the matrix. Then any other row is equal to the preceding row, right circularly shifted by 1. For example,

$$\mathbf{C} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{bmatrix}$$

is a circulant matrix. (Note that every circulant is a Toeplitz matrix, but the converse is not true.) One nice property of a circulant matrix $\mathbf{C}$ is that its eigenvectors are the columns of the Fourier transform matrix $\mathbf{F}$ defined by

$$f_{k\ell} = \frac{1}{\sqrt{n}} e^{2\pi i k\ell},$$

$k, \ell = 1, \ldots, n$, and its eigenvalues are computed by taking $\mathbf{F}$ times the first column of $\mathbf{C}$. Using the discrete Fourier transform algorithm, multiplication by $\mathbf{F}$ can be performed in $O(n \log_2 n)$ operations if n is a power of two. (If n has a lot of small prime factors, then the multiplication is also fast.)

4a. (10) Given a Toeplitz matrix $\mathbf{T}$, construct a circulant matrix $\mathbf{C}$ of larger size that has $\mathbf{T}$ as its upper left block. Describe an algorithm for forming $\mathbf{Tp}$ for an arbitrary vector $\mathbf{p}$ in the time it takes to form $\mathbf{C}$ times a vector, and show how to make this time at most $O(N \log_2 N)$, where N is the smallest power of 2 that is larger than $2n$.

Answer:
As an example, if $\mathbf{T}$ is tridiagonal, then we can embed it in a circulant of dimension $(n+1) \times (n+1)$ by adding a row and column:

$$\mathbf{C} = \begin{bmatrix} \beta_0 & \beta_{-1} & & & \beta_1 \\ \beta_1 & \beta_0 & \beta_{-1} & & \\ & & & & \\ & \ddots & \ddots & \ddots & \\ & & & & \\ & & \beta_1 & \beta_0 & \beta_{-1} \\ \beta_{-1} & & & \beta_1 & \beta_0 \end{bmatrix}$$

We are allowed to add more than one row and column, so we can make the dimension a power of 2, if that is convenient.

If $\mathbf{T}$ is pentadiagonal, then we add two rows and columns, and if $\mathbf{T}$ is dense, we add $n-1$ rows and columns, so that the matrix is square and circulant with the first row equal to

$$\beta_0 \beta_{-1} \ldots \beta_{-n+1} \beta_{n-1}, \ldots, \beta_1.$$

Again, we can add extra rows and columns to make the dimension a power of 2, by putting some zeros between β_{-n+1} and β_{n-1}.

If we multiply $\mathbf{Cz}$, where the first n components of $\mathbf{z}$ equal $\mathbf{p}$ and the others are zero, then the first n components of the product are equal to $\mathbf{Tp}$.

We multiply by a circulant matrix by taking the DFT of the vector, multiplying component-by-component by the DFT of the first column of the matrix, and then taking the inverse DFT of the result. The work is thus twice the cost of a DFT plus one multiplication per component.

4b. (10) Suppose $\mathbf{A}$ is a symmetric positive definite tridiagonal Toeplitz matrix. Find a circulant preconditioner so that cg converges in at most 3 iterations

Answer:
If we use the matrix $\mathbf{C}$ written above, but make its dimension n, then

$$\mathbf{C} = \mathbf{T} + \beta_1 \mathbf{e}_1 \mathbf{e}_n^T + \beta_{-1} \mathbf{e}_n \mathbf{e}_1^T,$$

where $\mathbf{e}_j$ is the jth column of the identity matrix. Therefore, $\mathbf{C}$ differs from $\mathbf{T}$ by a matrix of rank 2. Apply the Sherman-Morrison-Woodbury formula, with $\mathbf{A} = \mathbf{T}$, $\mathbf{Z} = [\mathbf{e}_1, \mathbf{e}_n]$, $\mathbf{V} = [\beta_1 \mathbf{e}_n, \beta_{-1} \mathbf{e}_1]$. Multiplying through by $\mathbf{A}$, we see that

$$\mathbf{C}^{-1}\mathbf{T} = \mathbf{I} + \mathbf{E}$$

where $\mathbf{E}$ has rank 2. By problem 1a, we see that $\mathbf{C}^{-1}\mathbf{T}$ has at most 3 distinct eigenvalues, so pcg will converge in at most 3 iterations.