Mathematics 372 – Numerical Linear Algebra
Final Problem Set
April 30, 2007, *due*: 5:00pm on May 9, 2007

General Information and Ground Rules

This is an open-book, open notes, open Swords 335 assignment. You must work *independently* on these problems, and you may not consult with other faculty members or students. You may consult other books, sources in the library, or information from the Internet, but if you find something relevant, you should *footnote* it for proper acknowledgment of sources. When you have finished your solutions, please write out a short statement that you have abided by these ground rules, sign it, and hand it in with your work.

I. *Some additional properties of matrix norms.* Recall that from Theorem 4.2.1 in Watkins, using the SVD, we know $\|A\|_2 = \sigma_1$, the largest singular value of A. This gives additional properties and estimates for the matrix 2-norm.

A) (10) On the midterm problem set, recall that we showed

$$\|A\|_F \le \sqrt{n} \cdot \|A\|_2$$

for all $n \times n$ matrices. Show the following more general and sharper form of this inequality: For all $A \in M_{n\times m}(\mathbf{R})$,

$$\|A\|_F \le \sqrt{\text{rank}(A)} \cdot \|A\|_2.$$

B) (10) Show that for all matrices $A \in M_{n\times m}(\mathbf{R})$,

$$\|A\|_2 \le \sqrt{\|A\|_1 \|A\|_\infty}$$

(this is sometimes useful for estimating $\|A\|_2$ without computing it exactly). (Hints: How are singular values of A related to eigenvalues of A^tA? What happens if you apply the 1-norm to an equation $A^tAz = \lambda z$?)

C) (10) Show using MATLAB that the inequalities in parts A and B are satisfied for the matrix $B - I_{60}$, where B is the 60×60 `bucky` matrix.

II. *More on condition numbers.* Let $n > m$. Recall that if $A \in M_{n\times m}(\mathbf{R})$, the condition number $\kappa_2(A) = \sigma_1/\sigma_m$ measures the susceptibility of the least-squares solution of $Ax = b$ to round-off errors.

A) (10) Show that if an additional column $y \in \mathbf{R}^n$ is appended to A, to yield

$$\overline{A} = \begin{pmatrix} A & y \end{pmatrix} \in M_{n\times(m+1)}(\mathbf{R}),$$

then

$$\sigma_1\left(\overline{A}\right) \geq \sigma_1\left(A\right) \quad \text{and} \quad \sigma_{m+1}\left(\overline{A}\right) \leq \sigma_m\left(A\right).$$

What does this say about $\kappa_2(\overline{A})$ vs. $\kappa_2(A)$?

B) (10) Show that if an additional row, w^t for $w \in \mathbf{R}^m$, is appended to A to yield

$$\overline{A} = \begin{pmatrix} A \\ w^t \end{pmatrix} \in M_{(n+1)\times m}(\mathbf{R}),$$

then

$$\sigma_1\left(\overline{A}\right) \leq \sqrt{\sigma_1(A)^2 + \|w\|_2^2} \quad \text{and} \quad \sigma_m\left(\overline{A}\right) \geq \sigma_m\left(A\right).$$

What does this say about $\kappa_2(\overline{A})$ vs. $\kappa_2(A)$?

III. *An image-processing application of the SVD.* A large, detailed, image stored as a matrix of gray-scale pixel values can take a large amount of storage space. If the information contained in such an image can be *compressed* without losing too much image quality, that is a good thing for transmission and storage. One possible method for image compression is based on the the theoretical result on SVD's in part A below. The later parts will show you how this works in practice.

A) (10) Recall from Theorem 4.1.12 (Exercise 4.1.13) in Watkins that if $A = U\Sigma V^t$ is an SVD of A, then if A has rank r,

$$(*) \qquad A = \sum_{j=1}^{r} \sigma_j u_j v_j^t,$$

where u_j and v_j are the columns of the U and V matrices respectively, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ as usual. If we keep only the largest k singular values for some $k < r$, then the resulting matrix

$$(**) \qquad A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^t,$$

can be thought of as a rank k *approximation to* A (this makes especially good sense if the omitted singular values $\sigma_{k+1}, \ldots, \sigma_r$ are "small" compared to the others). Prove that:

(1) $\|A - A_k\|_2 = \sigma_{k+1}$, and

(2) if A' is any other matrix of rank $\leq k$, then

$$\|A - A'\|_2 \geq \|A - A_k\|_2.$$

In words, A_k is the closest matrix to A among all matrices of rank k or less. Hint: For part (2), show that

$$\mathcal{N}(B) \cap \operatorname{Span}\{v_1, \ldots, v_{k+1}\} \neq \{0\}$$

and see what $A - A'$ does to a unit vector in the intersection.

B) (5) One measure of the size of an image A is the total number of real numbers needed to write the vectors u_j, v_j and the σ_j in (*) or (**). If A is 200×320 and has rank 200, what is the

$$\text{compression ratio} = \frac{\text{size using } (**)}{\text{size using } (*)}$$

achieved if we replace the original expression (*) for A with (**), using $k = 5, 10, 20, 25$?

C) (10) Using MATLAB, test out this compression scheme using the image file `clown.mat`:

```
load clown.mat;
colormap('gray');
```

This will store the image file as a 200×320 full matrix called X. You can display the full image using

```
image(X)
```

Now, compute the SVD of X calling the factors U, S, V. To compute the "compressed" matrices A_k for various k, you can use commands like this:

```
U(:,1:k)*S(1:k,1:k)*V(:,1:k)'
```

(you supply the values of k). Note: the MATLAB syntax $A(:, a : b)$ means: form the submatrix of A taking all rows and columns a through b.) For $k = 5, 10, 20, 25$, compute A_k, display the resulting images, and comment on how well they represent the full image ($k = 200$). As a more precise measure of image quality, also compute $\|A - A_k\|_2$ for each of these k values.

IV. *More on iterative methods.* Recall that the Jacobi and Gauss-Seidel iterative methods for square systems $Ax = b$ can be derived by splitting the coefficient matrix A as a sum. The general idea would be to write $A = M + N$ for some square matrices M, N with M invertible.

A) (5) Show that however this is done, the resulting iteration can be written as a correction based on the *residual* $r^{(k)} = b - Ax^{(k)}$:

$$x^{(k+1)} = x^{(k)} + M^{-1} r^{(k)}.$$

B) (5) For the remainder of this problem, assume that *all eigenvalues of A are real and non-negative.* The method obtained with $M = \frac{1}{\omega}I$ for some $\omega > 0$ and $N = A - M$ is called *Richardson's method.* Richardson's method iteration in the fixed point form is

$$x^{(k+1)} = (I - \omega A)x^{(k)} + \omega b.$$

Show that Richardson iteration converges only for $\omega < \frac{2}{\lambda_{max}}$, where λ_{max} is the largest eigenvalue of A.

C) (10) Show that the omega that minimizes the spectral radius of the "Richardson G-matrix" $I - \omega A$ is

$$\omega_{opt} = \frac{2}{\lambda_{max} + \lambda_{min}},$$

where "opt" stands for "optimal – explain why this would be the best value of ω to use.

D) (5) Refer to the system from Example 7.2.3 in the text. Using MATLAB, determine ω_{opt} for Richardson on this system, and determine the number of Richardson iterations needed to yield a solution that is accurate to 8 decimal places using $\omega = .17$, ω_{opt}, and $\omega = .1$.

Have an enjoyable and productive summer, and congratulations to our graduating seniors!