Mathematics 372 – Numerical Linear Algebra
Solutions for Final Problem Set
May 9, 2007

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

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

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

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

Solution: Say $\text{rank}(A) = r$. Let $A = U\Sigma V^t$ be an svd for A. Then by the result of Exercise 4.2.3,

$$\|A\|_F^2 = \|\Sigma\|_F^2 = \sigma_1^2 + \cdots + \sigma_r^2 \leq r\sigma_1^2 = \text{rank}(A)\sigma_1^2 = \text{rank}(A)\|A\|_2^2.$$

Taking square roots completes the proof.

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

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

(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 $A^tAz = \lambda z$?)

Solution: The eigenvalues of A^tA are the squares of the singular values of A by Exercise 5.2.17. As in the Hint, let z be an eigenvector of A^tA with eigenvalue σ_1^2, and assume z has been normalized so $\|z\|_1 = 1$. Then $\|A^tAz\|_1 = \|\sigma_1^2 z\|_1 = \sigma_1^2$. Hence

$$\sigma_1^2 \leq \max_{\|x\|_1=1} \|A^tAx\|_1 = \|A^tA\|_1 \leq \|A^t\|_1\|A\|_1 = \|A\|_\infty\|A\|_1,$$

(since the matrix 1-norm is the "maximum column sum" and the matrix ∞-norm is the "maximum row sum.") Taking square roots gives

$$\|A\|_2 = \sigma_1 \leq \sqrt{\|A\|_1\|A\|_\infty}$$

which is what we wanted to show.

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.

Solution: For part A, using MATLAB, we find the matrix $B - I_{60}$ has rank 51, so

$$\|B - I_{60}\|_F = 15.4919 \leq 25.8379 = \sqrt{51} \cdot \|B - I_{60}\|_2.$$

For part B,

$$\|B - I_{60}\|_2 = 3.6180 \leq 4.0 = \sqrt{\|B\|_1 \|B\|_\infty}.$$

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

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

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. $\kappa_2(A)$?

Solution: Let v_1 be a singular vector of A corresponding to the largest singular value (that is, column 1 of the V matrix in an svd). We know $Av_1 = \sigma_1 u_1$ and v_1, u_1 are unit vectors. Now consider the vector $\overline{v} = \begin{pmatrix} v_1 \\ 0 \end{pmatrix} \in \mathbf{R}^{m+1}$. Since v_1 is a unit vector in $\mathbf{R}^m$, this is also a unit vector. We have

$$\|\overline{A}\overline{v}\|_2 = \|(A\ y)\begin{pmatrix} v_1 \\ 0 \end{pmatrix}\|_2 = \|Av_1\|_2 = \sigma_1(A).$$

Hence

$$\sigma_1(\overline{A}) = \|\overline{A}\|_2 = \max_{\|\overline{x}\|_2=1} \|A\overline{x}\|_2 \geq \sigma_1(A),$$

where the maximum here is over all unit vectors in $\mathbf{R}^{m+1}$.
Similarly, if v_m is a singular vector of A corresponding to the smallest singular value (that is, column m of the V matrix in an svd). We know $Av_m = \sigma_m u_m$ and v_m, u_m are unit vectors. Now consider the vector $\overline{v} = \begin{pmatrix} v_m \\ 0 \end{pmatrix} \in \mathbf{R}^{m+1}$. Since v_m is a unit vector in $\mathbf{R}^m$, this is also a unit vector. We have

$$\|\overline{A}\overline{v}\|_2 = \|(A\ y)\begin{pmatrix} v_m \\ 0 \end{pmatrix}\|_2 = \|Av_m\|_2 = \sigma_m(A).$$

Hence

$$\sigma_{m+1}(\overline{A}) = \min_{\|\overline{x}\|_2=1} \|A\overline{x}\|_2 \leq \sigma_m(A),$$

since the minimum here is over all unit vectors in $\mathbf{R}^{m+1}$. (See Exercise 4.2.5 in the text – that is stated for square matrices, but the same is true in general.)
Finally, combining the two inequalities shown here,

$$\kappa_2(\overline{A}) = \frac{\sigma_1(\overline{A})}{\sigma_{m+1}(\overline{A})} \geq \frac{\sigma_1(A)}{\sigma_m(A)} = \kappa_2(A).$$

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. $\kappa_2(A)$?

Solution: We have, taking maximum over vectors $x \in \mathbf{R}^m$,

$$\begin{aligned}\sigma_1(\overline{A}) &= \max_{\|x\|_2=1} \|\overline{A}x\|_2 \\ &= \max_{\|x\|_2=1} \| \begin{pmatrix} A \\ w^t \end{pmatrix} x\|_2 \\ &= \max_{\|x\|_2=1} \sqrt{\|Ax\|_2^2 + (w^t x)^2}\end{aligned}$$

Now, by Cauchy-Schwarz, since x is a unit vector,

$$(w^t x)^2 = \langle w, x\rangle^2 \leq \|w\|_2^2 \|x\|_2^2 = \|w\|_2^2.$$

Moreover, $\max \|Ax\|_2^2 = \|A\|_2^2 = \sigma_1(A)^2$. Hence,

$$\sigma_1(\overline{A}) \leq \sqrt{\sigma_1(A)^2 + \|w\|_2^2},$$

as claimed.
Using the same formulas as above, for all unit vectors $x \in \mathbf{R}^m$,

$$\|\overline{A}x\|_2 = \sqrt{\|Ax\|_2^2 + \langle w, x\rangle^2} \geq \|Ax\|_2.$$

Hence the minimum over all x of $\|\overline{A}x\|_2$ must also be greater than or equal to the minimum over all x of $\|Ax\|_2$. This shows $\sigma_m(\overline{A}) \geq \sigma_m(A)$.

Unlike the situation in part A, we cannot say anything about the relative sizes of $\kappa_2(\overline{A})$ and $\kappa_2(A)$ here. Either one could be larger.

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

Solution: From (*) and (**) above,

$$A - A_k = \sum_{j=k+1}^{r} \sigma_j u_j v_j^t.$$

It follows that the singular values of $A - A_k$ are $\sigma_{k+1}, \ldots, \sigma_r$ and the rest zero. Since $\sigma_{k+1} \geq \sigma_j$ for $j > k+1$, this says $\|A - A_k\|_2 = \sigma_{k+1}$.

(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

$$\mathcal{N}(A') \cap \text{Span}\{v_1, \ldots, v_{k+1}\} \neq \{0\}$$

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

Solution: If $\text{rank}(A') \le k$, then by the fundamental equation

$$\dim \text{rank}(A') + \dim \mathcal{N}(A') = m,$$

we see $\dim \mathcal{N}(A') \ge m - k$. Now $V = \mathcal{N}(A')$ and $W = \text{Span}\{v_1, \ldots, v_{k+1}\}$ are two subspaces of $\mathbf{R}^m$ with $\dim(V) + \dim(W) \ge (m-k) + (k+1) = m+1$. It follows by a general linear algebra fact in this situation that $V \cap W \ne \{0\}$ (that is, there is some nonzero vector in the intersection of the two subspaces). Take any such $z \in V \cap W$. We can normalize z to obtain a unit vector (divide z by $\|z\|_2$). Then since $z \in W$, we have

$$z = c_1 v_1 + \cdots + c_{k+1} v_{k+1}$$

and at least one $c_i \ne 0$. But now, since $z \in V = \mathcal{N}(A')$ as well,

$$(A - A')z = Az - A'z = Az - 0 = Az.$$

Then, since $v_1, \ldots, v_{k+1}$ and $u_1, \ldots, u_{k+1}$ are orthogonal unit vectors, by the Pythagorean Theorem:

$$\begin{aligned}
\|(A - A')z\|_2 = \|Az\|_2 &= \|A(c_1 v_1 + \cdots + c_{k+1} v_{k+1})\|_2 \\
&= \|c_1 \sigma_1 u_1 + \cdots + c_{k+1} \sigma_{k+1} u_{k+1})\|_2 \\
&= \sqrt{(c_1 \sigma_1)^2 + \cdots + (c_{k+1} \sigma_{k+1})^2} \\
&\ge \sigma_{k+1} \sqrt{c_1^2 + \cdots + c_{k+1}^2} \\
&= \sigma_{k+1} \|z\|_2 \\
&= \sigma_{k+1}.
\end{aligned}$$

This implies $\|A - A'\|_2 \ge \sigma_{k+1}$, which, when combined with part (1), is what we wanted to show.

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 $k = 5, 10, 20, 25$?

Solution: With $k = 5$:

$$\frac{5 + 5 \times 200 + 5 \times 320}{200 + 200 \times 200 + 200 \times 320} = .025$$

With $k = 10$:

$$\frac{10 + 10 \times 200 + 10 \times 320}{200 + 200 \times 200 + 200 \times 320} = .05$$

With $k = 20$:

$$\frac{20 + 20 \times 200 + 20 \times 320}{200 + 200 \times 200 + 200 \times 320} = .1$$

With $k = 25$:

$$\frac{25 + 25 \times 200 + 25 \times 320}{200 + 200 \times 200 + 200 \times 320} = .125$$

Comment: A fairer comparison might actually be to take the numerator here over the size of the matrix form of A. For instance with $k = 25$:

$$\frac{25 + 25 \times 200 + 25 \times 320}{200 \times 320} = .2035$$

By that measure, the "compressed" form with $k = 25$ is about 1/5 the size of the original image in the matrix format.

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.

Solution: You should have seen that the image quality (judging by eye) steadily improved with $k = 5, 10, 20, 25$, until the image with $k = 25$ was hardly distinguishable from the original. The precise measures of image quality are (all $\times 10^3$):

$$\|A - A_5\|_2 = \sigma_6 = 1.0699$$
$$\|A - A_{10}\|_2 = \sigma_{11} = .6250$$
$$\|A - A_{20}\|_2 = \sigma_{21} = .3288$$
$$\|A - A_{25}\|_2 = \sigma_{26} = .2610$$

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* $r^{(k)} = b - Ax^{(k)}$:

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

Solution: The rearrangement to fixed-point form based on the splitting $A = M + N$ is found by starting from $Ax = (M + N)x = b$:

$$x = -M^{-1}Nx + M^{-1}b$$

This leads to the iteration formula

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

Now, if we rewrite N as $A - M$, this becomes

$$x^{(k+1)} = -M^{-1}(A - M)x^{(k)} + M^{-1}b = x^{(k)} + M^{-1}(b - Ax^{(k)}) = x^{(k)} + M^{-1}r^{(k)},$$

which is what we wanted to show.

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.

Solution: This is similar to one problem from Problem Set/Lab 11. The Richardson iteration converges if and only if the spectral radius of $I - \omega A$ is < 1. The spectrum of this matrix is the set

$$\sigma(I - \omega A) = \{1 - \omega\lambda : \lambda \in \sigma(A)\}.$$

Since all $\lambda \geq 0$ and $\omega > 0$,

$$\sigma(I - \omega A) \subset [1 - \omega\lambda_{max}, 1 - \omega\lambda_{min}] \subset (-\infty, 1) \subset \mathbf{R}$$

for all ω. The spectral radius is < 1 only if

$$1 - \omega\lambda_{max} > -1 \iff \omega < \frac{2}{\lambda_{max}}.$$

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.

Solution: As a function of ω, the spectral radius of $I - \omega A$ is given by the maximum of $|1 - \omega\lambda_{max}|$ and $|1 - \omega\lambda_{min}|$. Which one is larger depends on whether $1 - \omega\lambda_{max}$ is closer to -1 or $1 - \omega\lambda_{min}$ is closer to 1.

$$\begin{aligned} \rho(I - \omega A) &= \max\{|1 - \omega\lambda_{max}|, |1 - \omega\lambda_{min}|\} \\ &= \begin{cases} |1 - \omega\lambda_{max}| & \text{if } \omega\lambda_{min} < 2 - \omega\lambda_{max} \\ |1 - \omega\lambda_{min}| & \text{if } \omega\lambda_{min} > 2 - \omega\lambda_{max} \end{cases} \end{aligned}$$

Plotting the two parts of this function separately on the interval $(0, 2/\lambda_{max})$, we see that both start at 1 with $\omega = 0$. The first is a vee-shaped graph that goes down to zero at $\omega = 1/\lambda_{max}$ and comes back up to 1 at $\omega = 2/\lambda_{max}$. The second is (part of) a second vee-shaped curve sloping down *more gradually* from 1 at $\omega = 0$, possibly hitting the ω axis, and coming back up to $|1 - 2\lambda_{min}/\lambda_{max}|$ at $\omega = 2/\lambda_{max}$. Note: this could happen before you hit the "vee," depending on the exact values of λ_{max} and λ_{min}. The smallest max occurs when the two lines intersect, which is when

$$\omega\lambda_{min} = 2 - \omega\lambda_{max} \iff \omega = \frac{2}{\lambda_{max} + \lambda_{min}}.$$

This value is optimal in the sense that the spectral radius is minimized. The smaller the largest eigenvalue is (in absolute value), the more rapid the convergence will be.

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 $\omega = .1$.

Solution: We have $\lambda_{max} = 11.2561$, and $\lambda_{min} = 3.3182$, so

$$\omega_{opt} = \frac{2}{11.2561 + 3.3182} = .1372.$$

The exact solution is $(4, 3, 2, 1)^t$. Iterations needed for 8 decimal place accuracy:

$$\begin{aligned} \omega = .17 &\quad \text{about } 224 \\ \omega = \omega_{opt} = .1372 &\quad \text{about } 34 \\ \omega = .1 &\quad \text{about } 48. \end{aligned}$$