



Figure 1: Distribution of the singular values  $\sigma_i$  in the image of Fig. 2, showing that they decrease faster than exponentially with  $i$ .

## 18.335 Problem Set 3 Solutions

### Problem 1: SVD and low-rank approximations (10+10+10 pts)

- (a) It is sufficient to show that the reduced SVD  $A\hat{U} = \hat{U}\hat{\Sigma}$  is real, since the remaining columns of  $U$  and  $V$  are formed as a basis for the orthogonal complement of the columns of  $\hat{U}$  and  $\hat{V}$ , and if the latter are real then their complement is obviously also real. Furthermore, it is sufficient to show that  $\hat{U}$  can be chosen real, since  $A^*u_i/\sigma_i = v_i$  for each column  $u_i$  of  $\hat{U}$  and  $v_i$  of  $\hat{U}$ , and  $A^*$  is real. The columns  $u_i$  are eigenvectors of  $A^*A = B$ , which is a real-symmetric matrix, i.e.  $Bu_i = \sigma_i^2 u_i$ . Suppose that the  $u_i$  are *not* real. Then the real and imaginary parts of  $u_i$  are themselves eigenvectors with eigenvalue  $\sigma_i^2$  (proof: take the real and imaginary parts of  $Bu_i = \sigma_i^2 u_i$ , since  $B$  and  $\sigma_i^2$  are real). Hence, taking either the real or imaginary parts of the complex  $u_i$  (whichever is nonzero) and normalizing them to unit length, we obtain a new purely real  $\hat{U}$ . Q.E.D.<sup>1</sup>
- (b) We just need to show that, for any  $A \in \mathbb{C}^{m \times n}$  with  $\text{rank} < n$  and for any  $\varepsilon > 0$ , we can find a full-rank matrix  $B$  with  $\|A - B\|_2 < \varepsilon$ . Form the SVD  $A = U\Sigma V^*$  with singular values  $\sigma_1, \dots, \sigma_r$  where  $r < n$  is the rank of  $A$ . Let  $B = U\tilde{\Sigma}V^*$  where  $\tilde{\Sigma}$  is the same as  $\Sigma$  except that it has  $n - r$  additional nonzero singular values  $\sigma_{k>r} = \varepsilon/2$ . From equation 5.4 in the book,  $\|B - A\|_2 = \sigma_{r+1} = \varepsilon/2 < \varepsilon$ , noting that  $A = B_r$  in the notation of the book.
- (c) Take any grayscale photograph (either one of your own, or off the web). Scale it down to be no more than  $1500 \times 1500$  (but not necessarily square), and read it into Matlab as a matrix  $A$  with the `imread` command (type “`doc imread`” for instructions).
- This is plotted on a semilog scale in Fig 1, showing that the singular values  $\sigma_i$  decrease *faster* than exponentially with  $i$ .
  - Figure 2 shows an image of a handsome fellow, both at full resolution (200 singular values), and using only 16 and 8 singular values. Even with just 8 singular values (4% of the data), the image

<sup>1</sup>There is a slight wrinkle if there are repeated eigenvalues, e.g.  $\sigma_1 = \sigma_2$ , because the real or imaginary parts of  $u_1$  and  $u_2$  might not be orthogonal. However, taken together, the real and imaginary parts of any multiple eigenvalues must span the same space, and hence we can find a real orthonormal basis with Gram-Schmidt or whatever.



Figure 2: Left: full resolution image (albeit JPEG-compressed). Middle: 16% of the singular values. Right: 4% of the singular values.

is still at least somewhat recognizable. If the image were larger (this one is only  $282 \times 200$ ) then it would probably compress even more.

### Problem 2: QR and orthogonal bases (10+10+(5+5+5) pts)

- (a) If  $A = QR$ , then  $B = RQ = Q^*AQ = Q^{-1}AQ$  is a similarity transformation, and hence has the same eigenvalues as shown in the book. Numerically (and as explained in class and in lecture 28), doing this repeatedly for a Hermitian  $A$  (the unshifted QR algorithm) converges to a diagonal matrix  $\Lambda$  of the eigenvalues in descending order. To get the eigenvectors, we observe that if the  $Q$  matrices from each step are  $Q_1, Q_2$ , and so on, then we are computing  $\cdots Q_2^*Q_1^*AQ_1Q_2 \cdots = \Lambda$ , or  $A = Q\Lambda Q^*$  where  $Q = Q_1Q_2 \cdots$ . By comparison to the formula for diagonalizing  $A$ , the columns of  $Q$  are the eigenvectors.
- (b) The easiest way to approach this problem is probably to look at the explicit construction of  $\hat{R}$  via the Gram-Schmidt algorithms. By inspection,  $r_{ij} = q_i^*v_j$  is zero if  $i$  is even and  $j$  is odd or vice-versa. Because of this,  $\hat{R}$  will have a checkerboard pattern of nonzero values:

$$\hat{R} = \begin{pmatrix} \times & & & & & & \\ & \times & & & & & \\ & & \times & & & & \\ & & & \times & & & \\ & & & & \times & & \\ & & & & & \times & \\ & & & & & & \times \end{pmatrix}.$$

- (c) Trefethen, problem 10.4:

- (i) e.g. consider  $\theta = \pi/2$  ( $c = 0, s = 1$ ):  $Je_1 = -e_2$  and  $Je_2 = e_1$ , while  $Fe_1 = e_2$  and  $Fe_2 = e_1$ .  $J$  rotates clockwise in the plane by  $\theta$ .  $F$  is easier to interpret if we write it as  $J$  multiplied on the right by  $[-1, 0; 0, 1]$ : i.e.,  $F$  corresponds to a mirror reflection through the  $y$  ( $e_2$ ) axis followed by clockwise rotation by  $\theta$ . More subtly,  $F$  corresponds to reflection through a mirror

plane corresponding to the  $y$  axis rotated clockwise by  $\theta/2$ . That is, let  $c_2 = \cos(\theta/2)$  and  $s_2 = \sin(\theta/2)$ , in which case (recalling the identities  $c_2^2 - s_2^2 = c$ ,  $2s_2c_2 = s$ ):

$$\begin{pmatrix} c_2 & s_2 \\ -s_2 & c_2 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} c_2 & -s_2 \\ s_2 & c_2 \end{pmatrix} = \begin{pmatrix} -c_2 & s_2 \\ s_2 & c_2 \end{pmatrix} \begin{pmatrix} c_2 & -s_2 \\ s_2 & c_2 \end{pmatrix} = \begin{pmatrix} -c & s \\ s & c \end{pmatrix} = F,$$

which shows that  $F$  is reflection through the  $y$  axis rotated by  $\theta/2$ .

- (ii) The key thing is to focus on how we perform elimination under a single column of  $A$ , which we then repeat for each column. For Householder, this is done by a single Householder rotation. Here, since we are using  $2 \times 2$  rotations, we have to eliminate under a column one number at a time: given 2-component vector  $x = \begin{pmatrix} a \\ b \end{pmatrix}$  into  $Jx = \begin{pmatrix} \|x\|_2 \\ 0 \end{pmatrix}$ , where  $J$  is clockwise rotation by  $\theta = \tan^{-1}(b/a)$  [or, on a computer,  $\text{atan2}(b, a)$ ]. Then we just do this working “bottom-up” from the column: rotate the bottom two rows to introduce one zero, then the next two rows to introduce a second zero, etc.
- (iii) The flops to compute the  $J$  matrix itself are asymptotically irrelevant, because once  $J$  is computed it is applied to many columns (all columns from the current one to the right). To multiply  $J$  by a single 2-component vector requires 4 multiplications and 2 additions, or 6 flops. That is, 6 flops per row per column of the matrix. In contrast, Householder requires each column  $x$  to be rotated via  $x = x - 2v(v^*x)$ . If  $x$  has  $m$  components,  $v^*x$  requires  $m$  multiplications and  $m - 1$  additions, multiplication by  $2v$  requires  $m$  more multiplications, and then subtraction from  $x$  requires  $m$  more additions, for  $4m - 1$  flops overall. That is, asymptotically 4 flops per row per column. The 6 flops of Givens is 50% more than the 4 of Householder.

### Problem 3: Schur fine (10 + 15 points)

- (a) First, let us show that  $T$  is normal: substituting  $A = QTQ^*$  into  $AA^* = A^*A$  yields  $QTQ^*QT^*Q^* = QT^*Q^*QTQ^*$  and hence (cancelling the  $Q$ s)  $TT^* = T^*T$ .

The (1,1) entry of  $T^*T$  is the squared  $L_2$  norm ( $\|\cdot\|_2^2$ ) of the first column of  $T$ , i.e.  $|t_{1,1}|^2$  since  $T$  is upper triangular, and the (1,1) entry of  $TT^*$  is the squared  $L_2$  norm of the first row of  $T$ , i.e.  $\sum_i |t_{1,i}|^2$ . For these to be equal, we must obviously have  $t_{1,i} = 0$  for  $i > 1$ , i.e. that the first row is diagonal.

We proceed by induction. Suppose that the first  $j - 1$  rows of  $T$  are diagonal, and we want to prove this of row  $j$ . The  $(j, j)$  entry of  $T^*T$  is the squared norm of the  $j$ -th column, i.e.  $\sum_{i \leq j} |t_{i,j}|^2$ , but this is just  $|t_{j,j}|^2$  since  $t_{i,j} = 0$  for  $i < j$  by induction. The  $(j, j)$  entry of  $TT^*$  is the squared norm of the  $j$ -th row, i.e.  $\sum_{i \geq j} |t_{j,i}|^2$ . For this to equal  $|t_{j,j}|^2$ , we must have  $t_{j,i} = 0$  for  $i > j$ , and hence the  $j$ -th row is diagonal. Q.E.D.

- (b) The eigenvalues are the roots of  $\det(T - \lambda I) = \prod_i (t_{i,i} - \lambda) = 0$ —since  $T$  is upper-triangular, the roots are obviously therefore  $\lambda = t_{i,i}$  for  $i = 1, \dots, m$ . To get the eigenvector for a given  $\lambda = t_{i,i}$ , it suffices to compute the eigenvector  $x$  of  $T$ , since the corresponding eigenvector of  $A$  is  $Qx$ .

$x$  satisfies

$$0 = (T - t_{i,i}I)x = \begin{pmatrix} T_1 & u & B \\ & 0 & v^* \\ & & T_2 \end{pmatrix} \begin{pmatrix} x_1 \\ \alpha \\ x_2 \end{pmatrix},$$

where we have broken up  $T - t_{i,i}I$  into the first  $i - 1$  rows ( $T_1 u B$ ), the  $i$ -th row (which has a zero on the diagonal), and the last  $m - i$  rows  $T_2$ ; similarly, we have broken up  $x$  into the first  $i - 1$  rows  $x_1$ , the  $i$ -th row  $\alpha$ , and the last  $m - i$  rows  $x_2$ . Here,  $T_1 \in \mathbb{C}^{(i-1) \times (i-1)}$  and  $T_2 \in \mathbb{C}^{(m-i) \times (m-i)}$  are upper-triangular,

and are non-singular because by assumption there are no repeated eigenvalues and hence no other  $t_{j,j}$  equals  $t_{i,i}$ .  $u \in \mathbb{C}^{i-1}$ ,  $v \in \mathbb{C}^{m-i}$ , and  $B \in \mathbb{C}^{(i-1) \times (m-i)}$  come from the upper triangle of  $T$  and can be anything. Taking the last  $m-i$  rows of the above equation, we have  $T_2 x_2 = 0$ , and hence  $x_2 = 0$  since  $T_2$  is invertible. Furthermore, we can scale  $x$  arbitrarily, so we set  $\alpha = 1$ . The first  $i-1$  rows then give us the equation  $T_1 x_1 + u = 0$ , which leads to an upper-triangular system  $T_1 x_1 = -u$  that we can solve for  $x_1$ .

Now, let us count the number of operations. For the  $i$ -th eigenvalue  $t_{i,i}$ , to solve for  $x_1$  requires  $\sim (i-1)^2 \sim i^2$  flops to do backsubstitution on an  $(i-1) \times (i-1)$  system  $T_1 x_1 = -u$ . Then to compute the eigenvector  $Qx$  of  $A$  (exploiting the  $m-i$  zeros in  $x$ ) requires  $\sim 2mi$  flops. Adding these up for  $i = 1 \dots m$ , we obtain  $\sum_{i=1}^m i^2 \sim m^3/3$ , and  $2m \sum_{i=0}^{m-1} i \sim m^3$ , and hence the overall cost is  $\sim \frac{4}{3}m^3$  flops ( $K = 4/3$ ).