

CAAM 453 · NUMERICAL ANALYSIS I

Problem Set 4 · Solutions

Posted Thursday 8 October 2009. Due Monday, 19 October 2009. [Minor typos corrected 16 October 2009]
CAAM 453 students should complete 4 problems (including problem 5).
CAAM 553 students should complete 5 problems (including problem 5).
Students are welcome to attempt more problems if they wish.

1. [25 points]

Determine *by hand calculation*, the singular value decompositions of the matrices

$$(a) \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix} \quad (b) \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \quad (c) \begin{bmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad (d) \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \quad (e) \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

and, for each matrix, write down the optimal (2-norm) rank-1 approximation.

[Trefethen and Bau, problem 4.1]

Solution.

(a) For this matrix \mathbf{A} , we form

$$\mathbf{A}^* \mathbf{A} = \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^*.$$

From this decomposition of $\mathbf{A}^* \mathbf{A}$ we set

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and

$$\sigma_1 = \sqrt{9} = 3, \quad \sigma_2 = \sqrt{4} = 2.$$

Finally, we find \mathbf{u}_1 and \mathbf{u}_2 as

$$\mathbf{u}_1 = \frac{1}{\sigma_1} \mathbf{A} \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{u}_2 = \frac{1}{\sigma_2} \mathbf{A} \mathbf{v}_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

We assemble these components into the SVD

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^*.$$

The optimal rank-1 approximation is

$$\mathbf{X}_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^* = \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix}.$$

(b) The trick behind this problem is swapping the order the diagonal entries:

$$\mathbf{A}^* \mathbf{A} = \begin{bmatrix} 4 & 0 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^*.$$

From this we find

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$$\sigma_1 = \sqrt{9} = 3, \quad \sigma_2 = \sqrt{4} = 2,$$

and

$$\mathbf{u}_1 = \frac{1}{\sigma_1} \mathbf{A} \mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{u}_2 = \frac{1}{\sigma_2} \mathbf{A} \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

We assemble these components into the SVD

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^*.$$

Note that \mathbf{A} is a Hermitian positive definite matrix: for such matrices, the SVD is also an eigenvalue decomposition, i.e., the left and right singular vectors is identical. (This is not true for Hermitian indefinite matrices, as seen in question (a).)

The optimal rank-1 approximation is

$$\mathbf{X}_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^* = \begin{bmatrix} 0 & 0 \\ 0 & 3 \end{bmatrix}.$$

(c) The matrix \mathbf{A} is rectangular and rank-deficient. We have

$$\mathbf{A}^* \mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^*,$$

from which we find

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$$\sigma_1 = \sqrt{4} = 2, \quad \sigma_2 = \sqrt{0} = 0,$$

and

$$\mathbf{u}_1 = \frac{1}{\sigma_1} \mathbf{A} \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

For \mathbf{u}_2 and \mathbf{u}_3 , we simply need to find two unit vectors that are orthogonal to \mathbf{u}_1 . For example,

$$\mathbf{u}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{u}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

We assemble these components into the SVD

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^*.$$

Since \mathbf{A} is rank-1, the optimal rank-1 approximation is $\mathbf{X}_1 = \mathbf{A}$.

(d) Again we have a rank-1 matrix, but now the singular vectors are a bit more interesting. We compute

$$\mathbf{A}^* \mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2^{-1/2} & 2^{-1/2} \\ 2^{-1/2} & -2^{-1/2} \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2^{-1/2} & 2^{-1/2} \\ 2^{-1/2} & -2^{-1/2} \end{bmatrix} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^*,$$

and we determine

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

$$\sigma_1 = \sqrt{2}, \quad \sigma_2 = \sqrt{0} = 0,$$

and

$$\mathbf{u}_1 = \frac{1}{\sigma_1} \mathbf{A} \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Since \mathbf{A} is rank-deficient, we find \mathbf{u}_2 as a unit vector orthogonal to \mathbf{u}_1 :

$$\mathbf{u}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

giving the SVD

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2^{1/2} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2^{-1/2} & 2^{-1/2} \\ 2^{-1/2} & -2^{-1/2} \end{bmatrix} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^*.$$

Since \mathbf{A} is rank-1, the optimal rank-1 approximation is $\mathbf{X}_1 = \mathbf{A}$.

- (e) Note that this matrix is positive-semidefinite. (It is a multiple of an orthogonal projector.) We have

$$\mathbf{A}^* \mathbf{A} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 2^{-1/2} & 2^{-1/2} \\ 2^{-1/2} & -2^{-1/2} \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2^{-1/2} & 2^{-1/2} \\ 2^{-1/2} & -2^{-1/2} \end{bmatrix} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^*,$$

and we determine

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

$$\sigma_1 = \sqrt{4} = 2, \quad \sigma_2 = \sqrt{0} = 0,$$

and

$$\mathbf{u}_1 = \frac{1}{\sigma_1} \mathbf{A} \mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Since \mathbf{A} is rank-deficient, we find \mathbf{u}_2 as a unit vector orthogonal to \mathbf{u}_1 :

$$\mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

giving the SVD

$$\mathbf{A} = \begin{bmatrix} 2^{-1/2} & 2^{-1/2} \\ 2^{-1/2} & -2^{-1/2} \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2^{-1/2} & 2^{-1/2} \\ 2^{-1/2} & -2^{-1/2} \end{bmatrix} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^*.$$

Since \mathbf{A} is rank-1, the optimal rank-1 approximation is $\mathbf{X}_1 = \mathbf{A}$.

2. [25 points]

- (a) Suppose $\mathbf{A} \in \mathbb{C}^{m \times n}$, $m \geq n$, has full rank. The exact solution to the least squares problem

$$\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{A} \mathbf{x} - \mathbf{b}\|_2$$

is given by $\mathbf{x} = \mathbf{A}^+ \mathbf{b}$, where the *pseudoinverse* \mathbf{A}^+ is defined, for full rank \mathbf{A} , by

$$\mathbf{A}^+ = (\mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^*.$$

Show that the matrix $\mathbf{\Pi} = \mathbf{A} \mathbf{A}^+$ is an *orthogonal projector* onto $\text{Ran}(\mathbf{A})$.

- (b) If $\mathbf{A} \in \mathbb{C}^{m \times n}$, $m \geq n$, has full-rank, develop an expression for \mathbf{A}^+ in terms of the singular value decomposition of \mathbf{A} . (You may find it most convenient to work with the dyadic form of the SVD, $\mathbf{A} = \sum_{j=1}^n \sigma_j \mathbf{u}_j \mathbf{v}_j^*$.)
- (c) From (b), deduce a formula for \mathbf{A}^+ that is suitable when $\text{rank}(\mathbf{A}) = r < n$ for $\mathbf{A} \in \mathbb{C}^{m \times n}$, $m \geq n$. Your formula should satisfy two properties: $\mathbf{x} = \mathbf{A}^+ \mathbf{b}$ should be a vector that minimizes $\|\mathbf{b} - \mathbf{A} \mathbf{x}\|_2$, and $\mathbf{A} \mathbf{A}^+$ should be an orthogonal projector onto $\text{Ran}(\mathbf{A})$. Confirm that both hold.

Solution.

- (a) To check that $\mathbf{\Pi} = \mathbf{A}\mathbf{A}^+$ is an orthogonal projection, we must verify that $\mathbf{\Pi}^2 = \mathbf{\Pi}$ and $\mathbf{\Pi} = \mathbf{\Pi}^*$. These conditions can be checked with direct computations:

$$\begin{aligned}\mathbf{\Pi}^2 &= \mathbf{A}\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-2}\mathbf{A}^* = \mathbf{\Pi}; \\ \mathbf{\Pi}^* &= (\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*)^* = (\mathbf{A}^*)^*((\mathbf{A}^*\mathbf{A})^{-1})^*\mathbf{A}^* = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{\Pi},\end{aligned}$$

where we have used the fact that Hermitian matrices have Hermitian inverses.

To prove that $\mathbf{\Pi}$ projects onto $\text{Ran}(\mathbf{A})$, we must show that $\text{Ran}(\mathbf{\Pi}) \subseteq \text{Ran}(\mathbf{A})$ and $\text{Ran}(\mathbf{A}) \subseteq \text{Ran}(\mathbf{\Pi})$.

The first containment is simple, as for any $\mathbf{x} \in \mathbb{C}^m$, $\mathbf{\Pi}\mathbf{x} = \mathbf{A}(\mathbf{A}^*\mathbf{x}) = \mathbf{A}\mathbf{w}$ for some $\mathbf{w} \in \mathbb{C}^n$. Hence $\mathbf{\Pi}\mathbf{x} \in \text{Ran}(\mathbf{A})$, and so $\text{Ran}(\mathbf{\Pi}) \subseteq \text{Ran}(\mathbf{A})$.

For the reverse containment, we want to show that for every $\mathbf{y} \in \text{Ran}(\mathbf{A})$ we must have $\mathbf{y} \in \text{Ran}(\mathbf{\Pi})$. That is, we seek some $\mathbf{x} \in \mathbb{C}^m$ such that $\mathbf{\Pi}\mathbf{x} = \mathbf{y}$. But note that since $\mathbf{y} \in \text{Ran}(\mathbf{A})$, we can write $\mathbf{y} = \mathbf{A}\mathbf{w}$ for some $\mathbf{w} \in \mathbb{C}^n$. Now

$$\mathbf{\Pi}\mathbf{y} = \mathbf{A}\mathbf{A}^+\mathbf{A}\mathbf{w} = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A}\mathbf{w} = \mathbf{A}\mathbf{w} = \mathbf{y},$$

and so $\mathbf{\Pi}\mathbf{y} = \mathbf{y}$. This implies $\mathbf{y} \in \text{Ran}(\mathbf{\Pi})$, and hence $\text{Ran}(\mathbf{A}) \subseteq \text{Ran}(\mathbf{\Pi})$. (We have relied on the idea that a projector leaves fixed any vector in its range.)

From these two containments, we see that $\text{Ran}(\mathbf{\Pi}) = \text{Ran}(\mathbf{A})$, so $\mathbf{\Pi}$ projects onto $\text{Ran}(\mathbf{A})$.

- (b) Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be a full SVD of \mathbf{A} , and write the reduced (skinny) SVD as

$$\mathbf{A} = (\mathbf{U}_1 \ \mathbf{U}_2) \begin{pmatrix} \widehat{\mathbf{\Sigma}} \\ \mathbf{0} \end{pmatrix} \mathbf{V}^*.$$

Note that $\mathbf{A}^*\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \mathbf{V}\mathbf{\Sigma}^*\mathbf{\Sigma}\mathbf{V}^*$. Furthermore, $\mathbf{\Sigma}^*\mathbf{\Sigma} = \widehat{\mathbf{\Sigma}}^*\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^2 = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$; since \mathbf{A} is full rank, all these singular values are nonzero and hence $\mathbf{\Sigma}^*\mathbf{\Sigma}$ is invertible. We have

$$(\mathbf{A}^*\mathbf{A})^{-1} = \mathbf{V}\widehat{\mathbf{\Sigma}}^{-2}\mathbf{V}^*,$$

and since $\widehat{\mathbf{\Sigma}}^{-2}\mathbf{\Sigma}^* = (\widehat{\mathbf{\Sigma}} \ \mathbf{0})$,

$$\begin{aligned}\mathbf{A}^+ &= (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* \\ &= \mathbf{V}\widehat{\mathbf{\Sigma}}^{-2}\mathbf{V}^*\mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^* \\ &= \mathbf{V}(\widehat{\mathbf{\Sigma}}^{-1} \ \mathbf{0}) \begin{pmatrix} \mathbf{U}_1^* \\ \mathbf{U}_2^* \end{pmatrix} \\ &= \mathbf{V}\widehat{\mathbf{\Sigma}}^{-1}\mathbf{U}_1^*.\end{aligned}$$

Alternatively, if you write the SVD in the dyadic form

$$\mathbf{A} = \sum_{j=1}^n \sigma_j \mathbf{u}_j \mathbf{v}_j^*,$$

then you can see that

$$\mathbf{A}^+ = \sum_{j=1}^n \frac{1}{\sigma_j} \mathbf{v}_j \mathbf{u}_j^*.$$

Also note that $\mathbf{A}\mathbf{A}^+$ takes the form

$$\mathbf{\Pi} = \mathbf{A}\mathbf{A}^+ = \mathbf{U}_1\mathbf{U}_1^*.$$

Since the columns of \mathbf{U}_1 are orthonormal, $\mathbf{U}_1\mathbf{U}_1^*$ is an orthogonal projector onto $\text{Ran}(\mathbf{U}_1)$. (Recall that $\text{Ran}(\mathbf{U}_1) = \text{Ran}(\mathbf{A})$.)

- (c) For the rank-deficient case, one can follow the lead of the full-rank case. From the dyadic decomposition, one could define

$$\mathbf{A}^+ = \sum_{j=1}^r \frac{1}{\sigma_j} \mathbf{v}_j \mathbf{u}_j^*.$$

Alternatively, one can partition the full SVD in the form

$$\mathbf{A} = (\mathbf{U}_1 \quad \mathbf{U}_2) \begin{pmatrix} \widehat{\boldsymbol{\Sigma}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}_1^* \\ \mathbf{V}_2^* \end{pmatrix},$$

where $\mathbf{U}_1 \in \mathbb{C}^{m \times r}$, $\mathbf{U}_2 \in \mathbb{C}^{m \times (m-r)}$, $\boldsymbol{\Sigma} \in \mathbb{C}^{r \times r}$, $\mathbf{V}_1 \in \mathbb{C}^{n \times r}$, $\mathbf{V}_2 \in \mathbb{C}^{n \times (n-r)}$. Note that $\widehat{\boldsymbol{\Sigma}}$ is a square, invertible matrix, since $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$. With this partitioning, one can define

$$\mathbf{A}^+ = \mathbf{V}_1 \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{U}_1^*.$$

Note that $\mathbf{A}\mathbf{A}^+$ takes the form

$$\mathbf{A}\mathbf{A}^+ = (\mathbf{U}_1 \quad \mathbf{U}_2) \begin{pmatrix} \widehat{\boldsymbol{\Sigma}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}_1^* \\ \mathbf{V}_2^* \end{pmatrix} \mathbf{V}_1 \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{U}_1^* = \mathbf{U}_1 \mathbf{U}_1^*.$$

Recall that $\text{Ran}(\mathbf{U}_1) = \text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ forms a basis for $\text{Ran}(\mathbf{A})$, and, as in part (b), $\mathbf{U}_1 \mathbf{U}_1^*$ forms an orthogonal projector onto $\text{Ran}(\mathbf{U}_1) = \text{Ran}(\mathbf{A})$.

3. [25 points]

Prove the following three pseudoinverse identities for arbitrary $\mathbf{A} \in \mathbb{C}^{m \times n}$.

(a) $\mathbf{A}^+ = \lim_{t \rightarrow 0} (\mathbf{A}^* \mathbf{A} + t\mathbf{I})^{-1} \mathbf{A}^*.$

(b) $\mathbf{A}^+ = \int_0^\infty e^{-\mathbf{A}^* \mathbf{A} t} \mathbf{A}^* dt.$

- (c) Let Γ be a closed contour in the complex plane that encloses all nonzero eigenvalues of $\mathbf{A}^* \mathbf{A}$ but does not enclose the origin. Then

$$\mathbf{A}^+ = \frac{1}{2\pi i} \int_\Gamma \frac{1}{z} (z\mathbf{I} - \mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* dz.$$

[Stewart; Campbell & Meyer]

Solution.

- (a) Begin with the full SVD $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$. Let r denote the rank of \mathbf{A} , so that $\sigma_1, \dots, \sigma_r > 0$. We have $\mathbf{A}^* \mathbf{A} = \mathbf{V}\boldsymbol{\Sigma}^* \mathbf{U}^* \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^* = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^*$, where $\boldsymbol{\Lambda} := \boldsymbol{\Sigma}^* \boldsymbol{\Sigma} \in \mathbb{C}^{n \times n}$ is a diagonal matrix with nonnegative entries,

$$\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n).$$

Here $\lambda_j = \sigma_j^2$ if $1 \leq j \leq r$, and $\lambda_j = 0$ if $r < j \leq n$. Hence we can write

$$\mathbf{A}^* \mathbf{A} + t\mathbf{I} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^* + t\mathbf{V}\mathbf{V}^* = \mathbf{V}(\boldsymbol{\Lambda} + t\mathbf{I})\mathbf{V}^*$$

with $\boldsymbol{\Lambda} + t\mathbf{I}$ invertible for all $t > 0$. Thus,

$$(\mathbf{A}^* \mathbf{A} + t\mathbf{I})^{-1} = \mathbf{V}(\boldsymbol{\Lambda} + t\mathbf{I})^{-1} \mathbf{V}^*.$$

It is perhaps cleanest to write the product on the right in its dyadic form,

$$(\mathbf{A}^* \mathbf{A} + t\mathbf{I})^{-1} = \mathbf{V}(\boldsymbol{\Lambda} + t\mathbf{I})^{-1} \mathbf{V}^* = \sum_{j=1}^n (\lambda_j + t)^{-1} \mathbf{v}_j \mathbf{v}_j^*.$$

In this form, can form $(\mathbf{A}^* \mathbf{A} + t\mathbf{I})^{-1} \mathbf{A}^*$ as

$$\begin{aligned} (\mathbf{A}^* \mathbf{A} + t\mathbf{I})^{-1} \mathbf{A}^* &= \left(\sum_{j=1}^n (\lambda_j + t)^{-1} \mathbf{v}_j \mathbf{v}_j^* \right) \left(\sum_{k=1}^r \sigma_k \mathbf{v}_k \mathbf{v}_k^* \right) \\ &= \sum_{j=1}^r \frac{\sigma_j}{\lambda_j + t} \mathbf{v}_j \mathbf{v}_j^* = \sum_{j=1}^r \frac{\sigma_j}{\sigma_j^2 + t} \mathbf{v}_j \mathbf{v}_j^*. \end{aligned}$$

Now it is easy to take the limit as $t \rightarrow 0$, since $\sigma_j > 0$ for $j \leq r$:

$$\lim_{t \rightarrow 0} (\mathbf{A}^* \mathbf{A} + t\mathbf{I})^{-1} \mathbf{A}^* = \lim_{t \rightarrow 0} \sum_{j=1}^r \frac{\sigma_j}{\sigma_j^2 + t} \mathbf{v}_j \mathbf{v}_j^* = \sum_{j=1}^r \frac{1}{\sigma_j} \mathbf{v}_j \mathbf{v}_j^*,$$

which agrees with the formula for \mathbf{A}^+ developed in the last problem.

- (b) As in part (a), we take the SVD $\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^*$ and form $\mathbf{A}^* \mathbf{A} = \mathbf{V} \mathbf{\Sigma}^* \mathbf{U}^* \mathbf{U} \mathbf{\Sigma} \mathbf{V}^* = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^*$, where $\mathbf{\Lambda} := \mathbf{\Sigma}^* \mathbf{\Sigma} \in \mathbb{C}^{n \times n}$ is a diagonal matrix with nonnegative entries,

$$\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n).$$

Here $\lambda_j = \sigma_j^2$ if $1 \leq j \leq r$, and $\lambda_j = 0$ if $r < j \leq n$, where r is the rank of \mathbf{A} .

From the definition of the matrix exponential,

$$e^{\mathbf{Z}} := \mathbf{I} + \mathbf{Z} + \frac{1}{2} \mathbf{Z}^2 + \frac{1}{3!} \mathbf{Z}^3 + \dots,$$

we can deduce that

$$\begin{aligned} e^{-\mathbf{A}^* \mathbf{A} t} &= \mathbf{I} + \mathbf{V}(-\mathbf{\Lambda} t) \mathbf{V}^* + \frac{1}{2} \mathbf{V}(-\mathbf{\Lambda} t)^2 \mathbf{V}^* + \frac{1}{3!} \mathbf{V}(-\mathbf{\Lambda} t)^3 \mathbf{V}^* + \dots \\ &= \mathbf{V}(\mathbf{I} - t\mathbf{\Lambda} + \frac{1}{2} t^2 \mathbf{\Lambda}^2 - \frac{1}{3!} t^3 \mathbf{\Lambda}^3 + \dots) \mathbf{V}^* \\ &= \mathbf{V} e^{-t\mathbf{\Lambda}} \mathbf{V}^*, \end{aligned}$$

where

$$e^{-t\mathbf{\Lambda}} = \begin{bmatrix} e^{-t\lambda_1} & & \\ & \ddots & \\ & & e^{-t\lambda_n} \end{bmatrix}.$$

Similar to (a), write the expression for $e^{t\mathbf{A}^* \mathbf{A}}$ in dyadic form and multiply it against \mathbf{A}^* :

$$e^{-\mathbf{A}^* \mathbf{A} t} \mathbf{A}^* = \left(\sum_{j=1}^n e^{-t\lambda_j} \mathbf{v}_j \mathbf{v}_j^* \right) \left(\sum_{k=1}^r \sigma_k \mathbf{v}_k \mathbf{v}_k^* \right) = \sum_{j=1}^r \sigma_j e^{-t\lambda_j} \mathbf{v}_j \mathbf{v}_j^* = \sum_{j=1}^r \sigma_j e^{-t\sigma_j^2} \mathbf{v}_j \mathbf{v}_j^*.$$

Now integrate this sum,

$$\int_0^\infty \sum_{j=1}^r \sigma_j e^{-t\sigma_j^2} \mathbf{v}_j \mathbf{v}_j^* dt = \sum_{j=1}^r \left(\int_0^\infty \sigma_j e^{-t\sigma_j^2} dt \right) \mathbf{v}_j \mathbf{v}_j^* = \sum_{j=1}^r \left[-\frac{e^{-t\sigma_j^2}}{\sigma_j} \right]_{t=0}^{t=\infty} \mathbf{v}_j \mathbf{v}_j^* = \sum_{j=1}^r \frac{1}{\sigma_j} \mathbf{v}_j \mathbf{v}_j^*,$$

which again agrees with the formula for \mathbf{A}^+ derived in the last problem.

- (c) Using notation and techniques from part (a) and (b), we find that

$$(z\mathbf{I} - \mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* = \left(\sum_{j=1}^n \frac{1}{z - \lambda_j} \mathbf{v}_j \mathbf{v}_j^* \right) \left(\sum_{k=1}^r \sigma_k \mathbf{v}_k \mathbf{v}_k^* \right) = \sum_{j=1}^r \frac{\sigma_j}{z - \lambda_j} \mathbf{v}_j \mathbf{v}_j^* = \sum_{j=1}^r \frac{\sigma_j}{z - \sigma_j^2} \mathbf{v}_j \mathbf{v}_j^*,$$

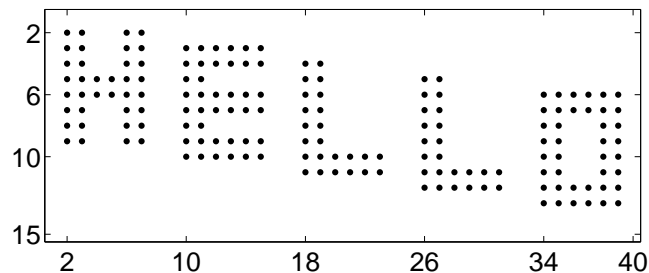
provided $z \neq \sigma_j^2$ for $j = 1, \dots, r$. Hence the integral takes the form

$$\frac{1}{2\pi i} \int_{\Gamma} \frac{1}{z} (z\mathbf{I} - \mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* dz = \sum_{j=1}^r \left(\frac{1}{2\pi i} \int_{\Gamma} \frac{\sigma_j/z}{z - \sigma_j^2} dz \right) \mathbf{v}_j \mathbf{v}_j^* = \sum_{j=1}^r \frac{\sigma_j}{\sigma_j^2} \mathbf{v}_j \mathbf{v}_j^* = \sum_{j=1}^r \frac{1}{\sigma_j} \mathbf{v}_j \mathbf{v}_j^*,$$

where we have evaluated the integral using the Cauchy integral formula. Once again, this formula agrees with the form for \mathbf{A}^+ that we developed in the last problem.

4. [25 points] Recall the example of image compression using the singular value decomposition that we saw in class. Now it is your turn to compress an image, this time a simple bitmap.

- (a) Write a MATLAB routine to construct the 15×40 matrix \mathbf{A} that is zero everywhere except for ones in the positions marked in the figure below. The upper-left point of the ‘H’ is in the (2,2) entry, and the bottom-right point of the ‘O’ is in the (13,39) entry.



- (b) What are the singular values of \mathbf{A} ? (Use MATLAB’s `svd` and `format long` to print 14 digits after the decimal point.) By counting the number of independent rows and columns of \mathbf{A} , determine the exact rank of \mathbf{A} . Does this agree with your output from `svd`?
- (c) For each k from 1 to $\text{rank}(\mathbf{A})$, compute the rank- k matrix \mathbf{A}_k that best approximates \mathbf{A} in the 2-norm. Visualize this matrix using the commands:

```
imagesc(Ak)
colormap(flipud(gray))
```

[Adapted from Trefethen and Bau, problem 9.3]

Solution. *Graders: The determination of rank in part (b) does not need to be particularly rigorous: students may just observe that there are 10 independent columns. In part (c), students need not print out all of the low-rank approximations; they should show at least a few to show that their code is correct.*

```
(a) A = zeros(15,40);
    A(2:9,2:3) = 1; % H
    A(5:6,4:5) = 1;
    A(2:9,6:7) = 1;

    A(3:10,10:11) = 1; % E
    A(3:4,12:15) = 1;
    A(6:7,12:15) = 1;
```

```

A(9:10,12:15) = 1;

A(4:11,18:19) = 1; % L
A(10:11,20:23) = 1;

A(5:12,26:27) = 1; % L
A(11:12,28:31) = 1;

A(6:13,34:35) = 1; % 0
A(6:7,36:37) = 1;
A(12:13,36:37) = 1;
A(6:13,38:39) = 1;

```

- (b) The easiest way to determine the rank of \mathbf{A} is to count the number of independent columns. Each letter contributes two independent columns. (Notice that the second, fifth, and sixth columns of 'H' are identical to the first column; the third and fourth are also identical, etc.) Thus, the exact rank of \mathbf{A} is ten.

MATLAB's `svd` command produces the following values:

```

10.38326912715624
 5.13177886450726
 3.18775603322296
 2.93826394575594
 2.06251516861703
 1.83760005749223
 1.22698575147038
 0.98791787255892
 0.73798536940384
 0.63288499887766
 0.00000000000000
 0.00000000000000
 0
 0
 0

```

There are ten non-zero singular values that are clearly non-zero. MATLAB also produces three exactly zero singular values. This is a consequence of the fact that \mathbf{A} has three identically zero rows. There are two other singular that look like zero when printed out using `format long`. Actually, they are

```

0.36277592792674e-15
0.18620124418176e-15

```

that is, they are on the order of roundoff error.

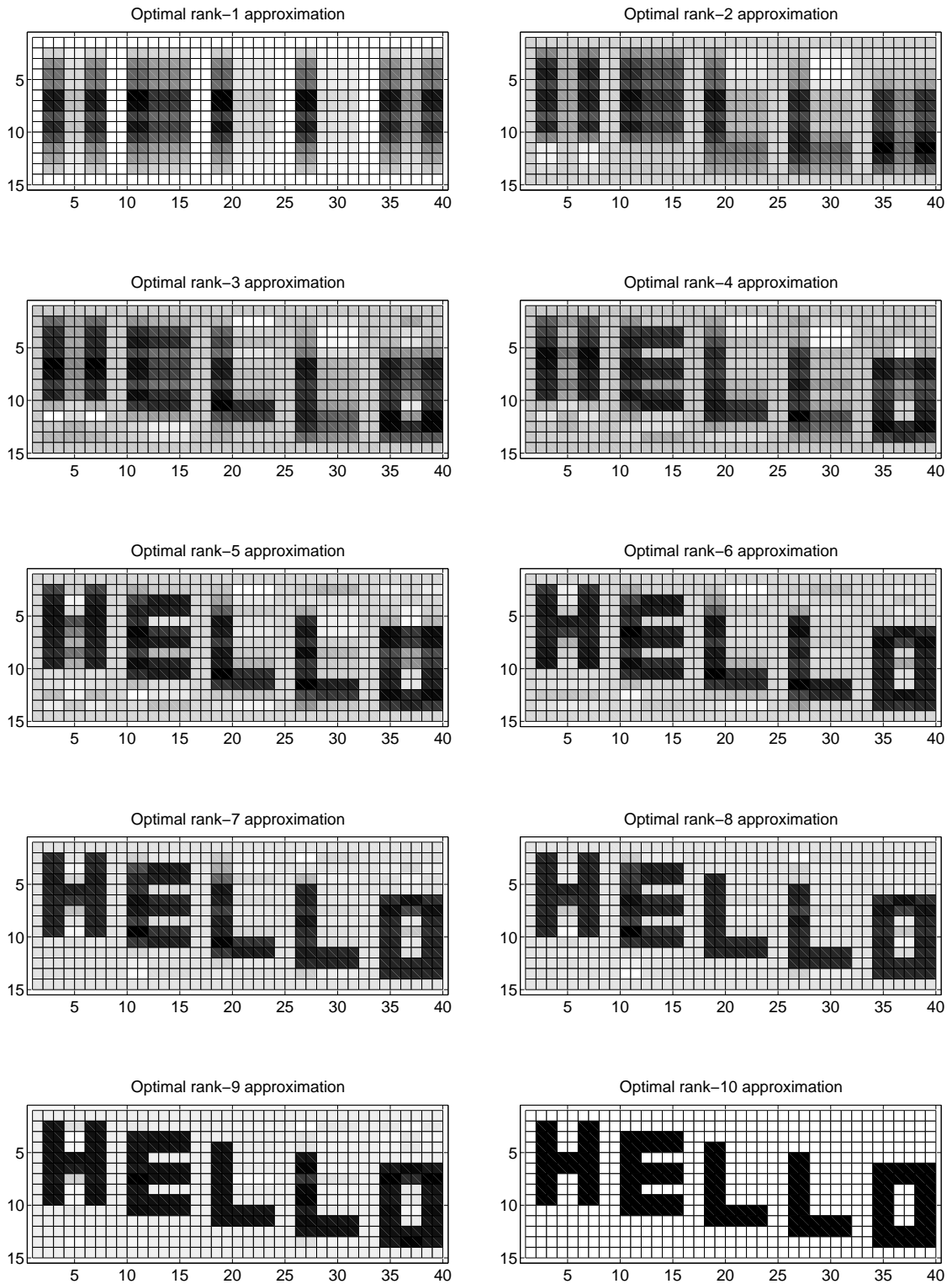
- (c) MATLAB code for computing (and printing postscript files) for each of the rank- k approximations for $k = 1, \dots, 10$ are shown below.

```

% compute and show low rank-approximations of A
% rank of A is 10
[U,S,V] = svd(A);
figure(1), clf
for k=1:10
    B = U(:,1:k)*S(1:k,1:k)*V(:,1:k)'; % construct the rank-k approximation
    pcolor(B)
    set(gca,'YDir','reverse')
    axis equal, axis([.5 40.5 .5 15.5])
    colormap(flipud(gray))
    set(gca,'fontsize',17)
    title(sprintf('Optimal rank-%d approximation',k),'fontsize',17)
    eval(sprintf('print -depsc2 hello%d.eps',k))
end

```

The plots this code produces are shown below.



5. [25 points] Let $\mathbf{A} \in \mathbb{C}^{m \times n}$ be a full rank matrix, with $m > n$. In general, $\mathbf{A}\mathbf{x} \neq \mathbf{b}$ for all $\mathbf{x} \in \mathbb{C}^n$.

The least squares problem amounts to finding the optimal approximation to $\mathbf{b} \in \mathbb{C}^m$ from $\text{Ran}(\mathbf{A})$:

$$\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 = \min_{\widehat{\mathbf{b}} \in \text{Ran}(\mathbf{A})} \|\mathbf{b} - \widehat{\mathbf{b}}\|_2.$$

In other words, the standard least squares problem seeks the smallest perturbation $\delta\mathbf{b}$ such that there exists some \mathbf{x} for which $\mathbf{A}\mathbf{x} = \mathbf{b} + \delta\mathbf{b}$. Implicitly, we are thus assuming that the matrix \mathbf{A} is exact, but the data \mathbf{b} has some errors.

An alternative approach, called *total least squares*, allows for errors in both \mathbf{A} and \mathbf{b} . Now we look for the smallest $\delta\mathbf{A}$ and $\delta\mathbf{b}$ such that there exists some \mathbf{x} for which $(\mathbf{A} + \delta\mathbf{A})\mathbf{x} = \mathbf{b} + \delta\mathbf{b}$, i.e.,

$$[\mathbf{A} + \delta\mathbf{A} \quad \mathbf{b} + \delta\mathbf{b}] \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \mathbf{0}. \quad (*)$$

This equation implies that the matrix $[\mathbf{A} + \delta\mathbf{A} \quad \mathbf{b} + \delta\mathbf{b}] \in \mathbb{C}^{m \times (n+1)}$ has rank less than $n + 1$. (Recall that $m > n$.)

- Use the singular value decomposition of the matrix $[\mathbf{A} \quad \mathbf{b}]$ to describe how to compute the matrix $[\delta\mathbf{A} \quad \delta\mathbf{b}]$ that makes $[\mathbf{A} + \delta\mathbf{A} \quad \mathbf{b} + \delta\mathbf{b}]$ rank-deficient and minimizes $\|[\delta\mathbf{A} \quad \delta\mathbf{b}]\|_2$.
- Use the optimal $[\delta\mathbf{A} \quad \delta\mathbf{b}]$ in (a) to write a simple formula for the solution \mathbf{x} in (*) in terms of appropriate singular values and/or vectors of $[\mathbf{A} \quad \mathbf{b}]$.
- Explain why $\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ cannot be smaller than the smallest singular value of $[\mathbf{A} \quad \mathbf{b}]$.
- Compute (in MATLAB) the solution \mathbf{x} produced by (i) standard least squares and (ii) total least squares for

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix}.$$

For (i), also report $\delta\mathbf{b} = \mathbf{A}\mathbf{x} - \mathbf{b}$ and $\|\delta\mathbf{b}\|$; for (ii), report $\delta\mathbf{A}$, $\delta\mathbf{b}$, and $\|[\delta\mathbf{A} \quad \delta\mathbf{b}]\|$ as in part (b).

Solution.

- Write the SVD of $[\mathbf{A} \quad \mathbf{b}]$ in the dyadic form

$$[\mathbf{A} \quad \mathbf{b}] = \sum_{j=1}^{n+1} \sigma_j \mathbf{u}_j \mathbf{v}_j^*.$$

The distance of $[\mathbf{A} \quad \mathbf{b}]$ from the set of rank-deficient matrices (i.e., matrices of rank n or less) is given by σ_{n+1} , and the matrix

$$\sum_{j=1}^n \sigma_j \mathbf{u}_j \mathbf{v}_j^*$$

is the best approximation to $[\mathbf{A} \quad \mathbf{b}]$ from the set of matrices of degree n or less. We thus define the perturbation $[\delta\mathbf{A} \quad \delta\mathbf{b}]$ to be the difference between $[\mathbf{A} \quad \mathbf{b}]$ and the best rank- n approximation:

$$[\delta\mathbf{A} \quad \delta\mathbf{b}] = -\sigma_{n+1} \mathbf{u}_{n+1} \mathbf{v}_{n+1}^*.$$

- Notice that

$$[\mathbf{A} + \delta\mathbf{A} \quad \mathbf{b} + \delta\mathbf{b}] \mathbf{v}_{n+1} = \left(\sum_{j=1}^n \sigma_j \mathbf{u}_j \mathbf{v}_j^* \right) \mathbf{v}_{n+1} = \mathbf{0},$$

and hence $\mathbf{v}_{n+1} \in \text{Ker}([\mathbf{A} + \delta\mathbf{A} \ \mathbf{b} + \delta\mathbf{b}])$. For this problem it is fine to assume that the last component of \mathbf{v}_{n+1} , which we will denote via the MATLAB notation $\mathbf{v}_{n+1}(n+1)$, is nonzero. In this case, \mathbf{v}_{n+1} can be viewed as a scaled version of the vector in the kernel in (*). Thus, we obtain \mathbf{x} via a scaling of the first n components of \mathbf{v}_{n+1} :

$$\mathbf{x} = -\frac{\mathbf{v}_{n+1}(1:n)}{\mathbf{v}_{n+1}(n+1)}.$$

Though not necessary for this problem, it is interesting to note that we can have $\mathbf{v}_{n+1}(n+1) = 0$, even when \mathbf{A} is full rank. For example, consider the $n = 1$ case

$$\mathbf{A} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 2 \end{bmatrix},$$

for which the singular value decomposition is

$$[\mathbf{A} \ \mathbf{b}] = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} = 2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} [0 \ 1] + 1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} [1 \ 0].$$

That is, $\mathbf{v}_2 = \mathbf{v}_{n+1} = [10]^T$, so $\mathbf{v}_{n+1}(n+1) = 0$. How should one adapt the Total Least Squares approach for cases like this one?

- (c) The usual least squares error, $\min_{\mathbf{x} \in \mathbb{C}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$, is a measure of the smallest perturbation $\delta\mathbf{b}$ that puts $\mathbf{b} + \delta\mathbf{b} \in \text{Ran}(\mathbf{A})$, whereas the smallest singular value of $[\mathbf{A} \ \mathbf{b}]$ measures the smallest perturbation $[\delta\mathbf{A} \ \delta\mathbf{b}]$ that makes $\mathbf{b} + \delta\mathbf{b} \in \text{Ran}(\mathbf{A} + \delta\mathbf{A})$. One option for this latter perturbation would be to take $\delta\mathbf{A} = \mathbf{0}$ and let $\delta\mathbf{b}$ be the same vector as in the standard least squares problem. However, it is possible that nonzero choices for $\delta\mathbf{A}$ give smaller overall values of $[\delta\mathbf{A} \ \delta\mathbf{b}]$. By optimizing over both $\delta\mathbf{A}$ and $\delta\mathbf{b}$, we are minimizing over a larger set, and thus must obtain a perturbation that is no larger than when minimizing only over $\delta\mathbf{b}$.
- (d) The code to perform these elementary computations is given below.

```
A = [1 2; 2 1; 1 2]
b = [1;0;-2];

x1 = A\b; % solve standard least squares problem
db1 = A*x1-b; % compute perturbation

[m,n] = size(A);
[U,S,V] = svd([A b]); % compute SVD of [A b]
dAdb = -U(:,m)*S(n+1,n+1)*V(:,n+1)'; % construct [dA db]
dA2 = dAdb(1:m,1:n); % extract dA
db2 = dAdb(1:m,n+1); % extract db
x2 = (A+dA2)\(b+db2); % construct solution, x

[x1 x2] % compare standard LS and TLS answers
[db1] % standard LS perturbation
[dA2 db2] % TLS perturbation
[norm(db1) norm([dA2 db2])] % compare standard LS and TLS errors
```

The solutions produced by the two methods are quite different. To four digits of accuracy,

$$\mathbf{x}_{\text{LS}} = \begin{bmatrix} 0.1667 \\ -0.3333 \end{bmatrix}, \quad \mathbf{x}_{\text{TLS}} = \begin{bmatrix} 5.8686 \\ -4.8058 \end{bmatrix}.$$

The standard least squares perturbation [Note: graders, the definition of $\delta\mathbf{b}_{\text{LS}}$ had a mistake in the statement of the problem; please do not take off points for sign errors!] is given by

$$\delta\mathbf{b}_{\text{LS}} = \begin{bmatrix} -1.5000 \\ 0.0000 \\ 1.5000 \end{bmatrix}.$$

On the other hand, the total least squares perturbation is

$$\delta \mathbf{A}_{\text{TLS}} = \begin{bmatrix} 0.4755 & -0.3894 \\ -0.6949 & 0.5691 \\ 0.1747 & -0.1431 \end{bmatrix}, \quad \delta \mathbf{b}_{\text{TLS}} = \begin{bmatrix} -0.0810 \\ 0.1184 \\ -0.0298 \end{bmatrix}.$$

The error is smaller for TLS, as expected:

$$\|\delta \mathbf{b}_{\text{LS}}\|_2 = 2.1213, \quad \|[\delta \mathbf{A}_{\text{TLS}} \ \delta \mathbf{b}_{\text{TLS}}]\|_2 = 1.1211.$$

6. [25 points]

In a celebrated 2003 article in the *Proceedings of the National Academy of Sciences*, Lawrence Sirovich analyzed US Supreme Court voting patterns using the singular value decomposition. Thanks to your skills with the SVD, and data kindly supplied by Prof. Keith Poole at the University of Houston, you can do the same analysis for yourself.

The data is found in `supreme_court.mat`, available from the course web site. With this file in your working directory, the command `load supreme_court` will create an 493-by-9 matrix \mathbf{A} . Each row of this matrix corresponds to a decision made by the between the period that Justice Stephen Breyer joined the court in 1994, and approximately 2002. (The period from Breyer's appointment to Renquist's death in 2005 was one of the longest intervals over which the same nine justices served the court.) Each of the nine columns of \mathbf{A} corresponds to one of the justices. The (j, k) entry of \mathbf{A} corresponds to the opinion of justice k on case j : agreement with the majority is denoted by 1, dissent by 0.

Here is how four famous recent Supreme Court cases would be encoded. The *Bush v. Gore* decision ended the Florida vote recount after the 2000 presidential election; the *Lawrence v. Texas* decision overturned the Texas sodomy law in 2003; the 2004 *Hamdi v. Rumsfeld* decision stated that Yaser Hamdi could challenge his 'enemy combatant' status in court; the *Kelo v. New London* decision, handed down in summer 2005, allows municipalities to seize the land of private citizens for the municipality's economic benefit. These results would contribute the following rows:

	Rehnquist	Stevens	O'Connor	Scalia	Kennedy	Souter	Thomas	Ginsberg	Breyer
<i>Bush v. Gore</i>	1	0	1	1	1	0	1	0	0
<i>Lawrence v. Texas</i>	0	1	1	0	1	1	0	1	1
<i>Hamdi v. Rumsfeld</i>	1	1	1	1	1	1	0	1	1
<i>Kelo v. New London</i>	0	1	0	0	1	1	0	1	1

- Compute the singular values of \mathbf{A} .
- From your answer to part (a), explain why \mathbf{A} might be well approximated by a rank-2 matrix. Report the value of $\|\mathbf{A} - \mathbf{A}_2\|_2$, where \mathbf{A}_2 is that optimal rank-2 approximation to \mathbf{A} .
- The leading right singular vectors \mathbf{v}_1 and \mathbf{v}_2 indicate properties of the most common rows of \mathbf{A} (though the entries will be positive and negative entries and the vectors will be scaled to have norm one).
 What are the first two right singular vectors, \mathbf{v}_1 and \mathbf{v}_2 ?
 What voting patterns do these two vectors correspond to? That is, for each vector, list the justices most likely to side with the majority, and those that dissent.

From \mathbf{v}_2 , can you deduce the two most frequent ‘swing votes’ — that is, those least tied to the other members of the \mathbf{v}_2 majority?

[Here is an example of how you would interpret the *last* singular vector,

$$\mathbf{v}_9 \approx (0.02, -0.02, -0.03, -0.70, -0.04, 0.02, 0.70, 0.10, -0.03)^T.$$

There are two large entries, -0.70 corresponding to Scalia, and 0.70 corresponding to Thomas. As the *signs* of these entries are opposite, we would interpret this vector as representing decisions in which Scalia and Thomas *disagreed*. (Such cases are quite rare, which explains why this is the least significant singular vector!) The other components, be they positive or negative, are small, so we would regard those justices as ‘swing voters’.]

N.B. For those unfamiliar with the US Supreme Court: the Renquist Court was said to consist of a conservative wing (Rehnquist, Scalia, Thomas), a liberal wing (Stevens, Souter, Ginsberg, Breyer), and two swing voters (O’Connor and Kennedy) who usually aligned with the conservatives but broke from them on key votes, such as *Lawrence v. Texas* above. The last part of question (c) is essentially asking you to assess how well the singular values confirm this conventional wisdom.

Solution.

(a) The singular values of \mathbf{A} are:

57.06248385159180
14.94162242875403
7.26614095897825
6.68143317176955
6.30008628524612
5.85495617812596
5.06666224940201
4.99819620724594
3.94434677657177

(b) We can approximate \mathbf{A} by a rank-2 matrix because the leading two singular values are significantly larger than the remaining seven singular values. The third singular value tells us the error in the optimal rank-2 approximation:

$$\min_{\text{rank}(\mathbf{X})=2} \|\mathbf{A} - \mathbf{X}\|_2 = \sigma_3 = 7.26614 \dots$$

(c) The leading two right singular vectors are

-0.34423037277379 0.29695542977526
-0.28023233037069 -0.50393605916673
-0.35576457664256 0.12452929660229
-0.32904801839704 0.39532941045656
-0.36250487994562 0.09989431828919
-0.34238142399287 -0.25551189974737
-0.32518095848651 0.42397238536922
-0.33068345397801 -0.32738727651611
-0.32311798014588 -0.35195569947798

The first right singular vector corresponds to a *unanimous decision*: all the entries are roughly the same size and have the same sign.

The second right singular vector corresponds to a 5-4 split (Renquist, O’Connor, Scalia, Kennedy, Thomas in the majority). Note that the third and fifth entries have the smallest magnitude; this identifies the swing voters: O’Connor and Kennedy.

Supplemental Problems

S1. Recall that the Frobenius norm is defined as

$$\|\mathbf{A}\|_F = \left(\sum_{j=1}^m \sum_{k=1}^n |a_{jk}|^2 \right)^{1/2}.$$

- (a) Prove that if $\mathbf{Q}_1 \in \mathbb{C}^{m \times m}$ and $\mathbf{Q}_2 \in \mathbb{C}^{n \times n}$ are unitary, then $\|\mathbf{Q}_1 \mathbf{A} \mathbf{Q}_2\|_F = \|\mathbf{A}\|_F$.
 (b) Determine a formula for $\|\mathbf{A}\|_F$ in terms of the singular values of \mathbf{A} , and use this to conclude that

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

- (c) What is the nearest rank-1 matrix to

$$\mathbf{A} = \begin{pmatrix} 1 & M \\ 0 & 1 \end{pmatrix}$$

in the Frobenius norm, where $M \in \mathbb{R}$?

[Golub and Van Loan]

S2. Suppose that $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$.

- (a) Suppose $r < n$, and let \mathcal{Z} be any subset of r distinct integers from $1, \dots, m$.
 Prove that there always exists some nonzero $\mathbf{z} \in \mathbb{R}^n$ such that $\mathbf{q} := \mathbf{A}\mathbf{z}$ is zero in each entry in \mathcal{Z} , i.e., $q_j = 0$ for all $j \in \mathcal{Z}$.

The rest of this problem walks through a proof (adapted from Powell's *Approximation Theory and Methods*) of the following remarkable fact: there exists some vector $\mathbf{w} \in \mathbb{R}^n$ for which

$$\|\mathbf{b} - \mathbf{A}\mathbf{w}\|_1 = \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_1$$

and the residual $\mathbf{b} - \mathbf{A}\mathbf{w}$ has at least n zero entries, i.e., an optimal 1-norm approximation exactly satisfies at least n of the equations in $\mathbf{A}\mathbf{x} \approx \mathbf{b}$.

Suppose that $\|\mathbf{b} - \mathbf{A}\mathbf{w}\|_1 = \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_1$, and set $\mathbf{y} := \mathbf{A}\mathbf{w}$. Let \mathcal{Z} denote the subset of the indices $1, \dots, m$ that correspond to zero entries of the vector $\mathbf{b} - \mathbf{A}\mathbf{w}$. We wish to show that \mathcal{Z} must contain at least n distinct indices.

If \mathcal{Z} contains $r < n$ distinct entries, then, from part (a), we can construct some vector $\mathbf{q} = \mathbf{A}\mathbf{z}$ that is zero in every entry in \mathcal{Z} , i.e., $q_j = 0$ for all $j \in \mathcal{Z}$.

Define the function

$$\gamma(\theta) := \|\mathbf{b} - \mathbf{A}(\mathbf{w} + \theta\mathbf{z})\|_1 = \sum_{j=1}^n |b_j - (y_j + \theta q_j)|.$$

- (b) Explain why γ is a continuous and piecewise linear function of θ , why γ must have a minimum at $\theta = 0$, and why $\gamma(\theta) \rightarrow \infty$ as $|\theta| \rightarrow \infty$.
 (c) Explain why γ is constant in a neighborhood about $\theta = 0$. (Hint: use the zero structure in \mathbf{q} .)
 (d) Since $\gamma(\theta) \rightarrow \infty$ as $|\theta| \rightarrow \infty$, $\gamma(\theta)$ cannot be constant for all θ . Explain why, as θ increases from $\theta = 0$, the value of $\gamma(\theta)$ must remain constant until θ reaches some distinguished value $\hat{\theta}$, where $b_j - (y_j + \hat{\theta}q_j) = 0$ for some value of $j \in \{1, \dots, m\}$ for which $q_j \neq 0$, i.e., $j \notin \mathcal{Z}$.
 (e) Explain why this implies that $\mathbf{w} + \hat{\theta}\mathbf{z}$ must be as good an approximation as \mathbf{w} itself was, but that $\mathbf{b} - \mathbf{A}(\mathbf{w} + \hat{\theta}\mathbf{z})$ has (at least) $r + 1$ zeros.

Repeatedly applying this argument, we see that there must exist some best approximation, say $\hat{\mathbf{w}}$ for which $\mathbf{b} - \mathbf{A}\hat{\mathbf{w}}$ has at least n zero entries.