UNIVERSITY OF CALIFORNIA, SANTA CRUZ
BOARD OF STUDIES IN COMPUTER ENGINEERING

CMPE 240: INTRODUCTION TO LINEAR DYNAMICAL SYSTEMS

PROBLEM SET 9

1. *Iterative solution of linear equations.* In many applications we need to solve a set of linear equations $Ax = b$, where A is nonsingular (square) and x is very large (*e.g.*, $x \in \mathbf{R}^{100000}$). We assume that Az can be computed at reasonable cost, for any z, but the standard methods for computing $x = A^{-1}b$ (*e.g.*, LU decomposition) are not feasible.

   A common approach is to use an *iterative* method, which computes a sequence $x(1), x(2)$ $, \ldots$ that *converges* to the solution $x = A^{-1}b$. These methods rely on another matrix $\hat{A}$, which is supposed to be 'close' to A. More importantly, $\hat{A}$ has the property that $\hat{A}^{-1}z$ is easily or cheaply computed for any given z.

   As a simple example, the matrix $\hat{A}$ might be the diagonal part of the matrix A (which, presumably, has relatively small off-diagonal elements). Obviously computing $\hat{A}^{-1}z$ is fast; it's just scaling the entries of z. There are many, many other examples.

   A simple iterative method, sometimes called *relaxation*, is to set $\hat{x}(0)$ equal to some approximation of x (*e.g.*, $\hat{x}(0) = \hat{A}^{-1}b$) and repeat, for $t = 0, 1, \ldots$

   $$r(t) = A\hat{x}(t) - b; \qquad \hat{x}(t+1) = \hat{x}(t) - \hat{A}^{-1}r(t);$$

   (The hat reminds us that $\hat{x}(t)$ is an approximation, after t iterations, of the true solution $x = A^{-1}b$.) This iteration uses only 'cheap' calculations: multiplication by A and $\hat{A}^{-1}$. Note that $r(t)$ is the residual after the tth iteration.

   (a) Let $\beta = \|\hat{A}^{-1}(A - \hat{A})\|$ (which is a measure of how close $\hat{A}$ and A are). Show that if we choose $\hat{x}(0) = \hat{A}^{-1}b$, then $\|\hat{x}(t) - x\| \le \beta^{t+1}\|x\|$. Thus if $\beta < 1$, the iterative method works, *i.e.*, for any b we have $\hat{x}(t) \to x$ as $t \to \infty$. (And if $\beta < 0.8$, say, then convergence is pretty fast.)

   (b) Find the exact conditions on A and $\hat{A}$ such that the method works for any starting approximation $\hat{x}(0)$ and any b. Your condition can involve norms, singular values, condition number, and eigenvalues of A and $\hat{A}$, or some combination, etc. Your condition should be as explicit as possible; for example, it should not include any limits.

   Try to avoid the following two errors:

- Your condition guarantees convergence but is too restrictive. (For example: $\beta = \|\hat{A}^{-1}(A - \hat{A})\| < 0.8$)
- Your condition doesn't guarantee convergence.

2. A matrix can have all entries large and yet have small gain in some directions, that is, it can have a small $\sigma_{\min}$. For example,
$$A = \begin{bmatrix} 10^6 & 10^6 \\ 10^6 & 10^6 \end{bmatrix}$$
has "large" entries while $\|A[1 \ \ -1]^T\| = 0$.

   Can a matrix have all entries small and yet have a large gain in some direction, that is, a large $\sigma_{\max}$? Suppose, for example, that $|A_{ij}| \leq \epsilon$ for $1 \leq i, j \leq n$. What can you say about $\sigma_{\max}(A)$?

3. *Condition number.* Show that $\kappa(A) = 1$ if and only if A is a multiple of an orthogonal matrix. Thus the best conditioned matrices are precisely (scaled) orthogonal matrices.

4. *Tightness of the condition number sensitivity bound.* Suppose A is invertible, $Ax = y$, and $A(x+\delta x) = y+\delta y$. In the lecture notes we showed that $\|\delta x\|/\|x\| \leq \kappa(A)\|\delta y\|/\|y\|$. Show that this bound is not conservative, *i.e.*, there are x, y, δx, and δy such that equality holds.

   Conclusion: the bound on relative error can be taken on, if the data x is in a particularly unlucky direction and the data error δx is in (another) unlucky direction.

5. *Detecting linear relations.* Suppose we have N measurements $y_1, \ldots, y_N$ of a vector signal $x_1, \ldots, x_N \in \mathbf{R}^n$:
$$y_i = x_i + d_i, \ i = 1, \ldots, N.$$
Here d_i is some small measurement or sensor noise. We hypothesize that there is a linear relation among the components of the vector signal x, *i.e.*, there is a nonzero vector q such that $q^T x_i = 0$, $i = 1, \ldots, N$. The geometric interpretation is that all of the vectors x_i lie in the hyperplane $q^T x = 0$. We will assume that $\|q\| = 1$, which does not affect the linear relation.

   Even if the x_i's do lie in a hyperplane $q^T x = 0$, our measurements y_i will not; we will have $q^T y_i = q^T d_i$. These numbers are small, assuming the measurement noise is small. So the problem of determing whether or not there is a linear relation among the components of the vectors x_i comes down to finding out whether or not there is a unit-norm vector q such that $q^T y_i$, $i = 1, \ldots, N$, are all small.

   We can view this problem geometrically as well. Assuming that the x_i's all lie in the hyperplane $q^T x = 0$, and the d_i's are small, the y_i's will all lie close to the hyperplane. Thus a scatter plot of the y_i's will reveal a sort of flat cloud, concentrated near the hyperplane $q^T x = 0$. Indeed, for any z and $\|q\| = 1$, $|q^T z|$ is the distance from the vector z to the hyperplane $q^T x = 0$. So we seek a vector q, $\|q\| = 1$, such that all the measurements $y_1, \ldots, y_N$ lie close to the hyperplane $q^T x = 0$ (that is, $q^T y_i$ are all small).

How can we determine if there is such a vector, and what is its value? We define the following normalized measure:

$$\rho = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(q^T y_i)^2} \Bigg/ \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|y_i\|^2}.$$

This measure is simply the ratio between the *root mean square distance* of the vectors to the hyperplane $q^T x = 0$ and the *root mean square length* of the vectors. If ρ is small, it means that the measurements lie close to the hyperplane $q^T x = 0$. Obviously, ρ depends on q.

Here is the problem: explain how to find the minimum value of ρ over all unit-norm vectors q, and the unit-norm vector q that achieves this minimum, given the data set $y_1, \ldots, y_N$.

6. *Frobenius norm of a matrix.* The Frobenius norm of a matrix $A \in \mathbf{R}^{n\times n}$ is defined as $\|A\|_{\mathrm{F}} = \sqrt{\mathbf{Tr}\, A^T A}$. (Recall $\mathbf{Tr}$ is the trace of a matrix, *i.e.*, the sum of the diagonal entries.)

   (a) Show that

   $$\|A\|_{\mathrm{F}} = \left(\sum_{i,j}|A_{ij}|^2\right)^{1/2}.$$

   Thus the Frobenius norm is simply the Euclidean norm of the matrix when it is considered as an element of $\mathbf{R}^{n^2}$. Note also that it is much easier to compute the Frobenius norm of a matrix than the (spectral) norm (*i.e.*, maximum singular value).

   (b) Show that if U and V are orthogonal, then $\|UA\|_{\mathrm{F}} = \|AV\|_{\mathrm{F}} = \|A\|_{\mathrm{F}}$. Thus the Frobenius norm is not changed by a pre- or post- orthogonal transformation.

   (c) Show that $\|A\|_{\mathrm{F}} = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$, where $\sigma_1, \ldots, \sigma_r$ are the singular values of A. Then show that $\sigma_{\max}(A) \le \|A\|_{\mathrm{F}} \le \sqrt{r}\sigma_{\max}(A)$. In particular, $\|Ax\| \le \|A\|_{\mathrm{F}}\|x\|$ for all x.

7. *Minimum energy required to steer the state to zero.* Consider a controllable discrete-time system $x(t+1) = Ax(t) + Bu(t)$, $x(0) = x_0$. Let $E(x_0)$ denote the minimum energy required to drive the state to zero, *i.e.*

$$E(x_0) = \min\left\{ \sum_{\tau=0}^{t-1}\|u(\tau)\|^2 \mid x(t) = 0 \right\}.$$

An engineer argues as follows:

> This problem is like the minimum energy reachability problem, but 'turned backwards in time' since here we steer the state from a given state to zero, and in the reachability problem we steer the state from zero to a given state. The system $z(t+1) = A^{-1}z(t) - A^{-1}Bv(t)$ is the same as the given one, except

time is running backwards. Therefore $E(x_0)$ is the same as the minimum energy required for z to reach x_0 (a formula for which can be found in the lecture notes).

Either justify or refute the engineer's statement. You can assume that A is invertible.

8. *Sensor selection and observer design.* Consider the system $\dot{x} = Ax$, $y = Cx$, with

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

(This problem concerns observer design so we've simplified things by not even including an input.)

We consider observers that (exactly and instantaneously) reconstruct the state from the output and its derivatives. Such observers have the form

$$x(t) = F_0 y(t) + F_1 \frac{dy}{dt}(t) + \cdots + F_k \frac{d^k y}{dt^k}(t),$$

where $F_0, \ldots, F_k$ are matrices that specify the observer. (Of course we require this formula to hold for any trajectory of the system and any t, *i.e.*, the observer has to work!)

Consider an observer defined by $F_0, \ldots, F_k$. We say the *degree* of the observer is the largest j such that $F_j \neq 0$. The degree gives the highest derivative of y used to reconstruct the state.

If the ith columns of $F_0, \ldots, F_k$ are all zero, then the observer doesn't use the ith sensor signal $y_i(t)$ to reconstruct the state. We say the observer *uses* or *requires* the sensor i if at least one of the ith columns of $F_0, \ldots, F_k$ is nonzero.

(a) What is the minimum number of sensors required for such an observer? List all combinations (*i.e.*, sets) of sensors, of this minimum number, for which there is an observer using only these sensors.

(b) What is the minimum degree observer? List all combinations of sensors for which an observer of this minimum degree can be found.