Matrix norms, condition number etc.

Recall the following facts from linear algebra:

For $x \in \mathbb{R}^n$ we have several norms $\|x\|$

$\|x\|_1 = \sum_{i=1}^{n} |x_i|$

$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$, unit square

$\|x\|_2 = \left( \sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$, unit ball

Matrix norms could be introduced as operator norms since $A \in \mathbb{R}^{n \times n}$ is an operator $\mathbb{R}^n \to \mathbb{R}^n$.

$\|A\| = \sup \left\{ \frac{\|Ax\|}{\|x\|} : x \in \mathbb{R}^n, \|x\|=1 \right\} = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$

This is called matrix norm subordinate to the vector norm $\|x\|_1$.

Remark: Frobenius norm $\|A\|_F = \left( \sum_{i,j} a_{ij}^2 \right)^{1/2}$ is not an operator norm, i.e. it is not subordinate to a vector norm.
One can deduce various properties of the matrix norm

\[ \|AX\| = \|A\| \|X\| \]

\[ \|A + B\| \leq \|A\| + \|B\| \]

\[ \|AB\| \leq \|A\| \|B\| \]

**Condition Number:** Motivation and importance

Consider the system \( Ax = b \)

\[ \varepsilon(A) = \|A\| \|A^{-1}\| \]

Assume \( A \) invertible and the systems \( Ax = b \) and \( \tilde{A}x = \tilde{b} \)

where \( b - \tilde{b} \) is "small" (in some norm). Say \( \|b - \tilde{b}\| \leq \varepsilon = 10 \)

What can be said about

\[ \frac{\|x - \tilde{x}\|}{\|x\|} \]

E.g., if we compute \( \tilde{b} \) with 6 decimal digits how accurate \( x \) will be assuming exact calculations in the forward process?

Or what is the worst error \( \|x - \tilde{x}\| \)?

\[ \|x - \tilde{x}\| = \|A^{-1}b - A^{-1}\tilde{b}\| \leq \|A^{-1}\| \|b - \tilde{b}\| \]

\[ \leq \|A^{-1}\| \|b\| \|b - \tilde{b}\| \leq \|A^{-1}\| \|A\| \|x\| \|b - \tilde{b}\| \]

\[ \Rightarrow \frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A^{-1}\| \|A\| \|b - \tilde{b}\| \|b\| \|x\| \]

\[ \leq \varepsilon(A) \frac{\|b - \tilde{b}\|}{\|b\|} \]

\[ = \varepsilon(A) \frac{\|b - \tilde{b}\|}{\|b\|} \]
Theorem: Let $e = x - \hat{x}$ and $r = b - Ax$ - residual error.

Then

$$\frac{1}{\kappa(A)} \frac{\|e\|}{\|e\|} \leq \frac{\|e\|}{\|r\|} \leq \kappa(A) \frac{\|r\|}{\|e\|}$$

Proof. The second inequality has been already proven. Now the lower one

$$\|r\| \leq \|A\| \|x\| \|Ax - b\| \leq \|A\| \|x - \hat{x}\| \|A\| \|\hat{x}\|$$

$$\leq \|A\| \|e\| \|A\| \|A\| \|\hat{x}\|$$

$$\Rightarrow \frac{1}{\kappa(A)} \frac{\|r\|}{\|e\|} \leq \frac{\|e\|}{\|r\|} \leq \kappa(A) \frac{\|r\|}{\|e\|}$$

This is not very reliable estimate, but it says that the error of the r.h.s. could be magnified by the condition number compared with the error of the solution.

Def. We say that the system is ill-conditioned if

the condition number is very large.

Hilbert matrix

$$a_{ij} = \frac{1}{i+j-1} \quad ij = 1, \ldots, n$$

$$\kappa(A) \approx e^{cn}$$
The idea of iterative improvement (refinement)

\[ A \approx LU \quad (LU)^{-1} \] is easy to compute with

\[ LU x = \mathbf{6} \quad x^{(0)} \] which differs from \( x \)

Then compute the residual \( r^{(0)} = Ax^{(0)} + \mathbf{6} \). Maybe using this residual we can improve the solution \( x^{(0)} \) by correcting it

\[
e^{(0)} = (LU)^{-1} r^{(0)} \quad \Rightarrow \quad x^{(1)} = x^{(0)} + e^{(0)}
\]

Iterative improvement

Now we formalize this into the following; let

\[ x^{(0)} = L B \] where \( B \) is easy to compute with

Obviously if \( B = A^{-1} \) we are done

But \( A^{-1} \) is very costly. We take \( B \) to be some cheaply computed

Obviously, \( x^{(0)} \) is not a solution. Therefore we can find

\[
r^{(0)} = \mathbf{6} - Ax^{(0)} \quad e^{(0)} = Br^{(0)} \quad x^{(1)} = x^{(0)} + e^{(0)}
\]

Now we make this in many steps

\[
\begin{align*}
  r^{(k)} &= \mathbf{6} - Ax^{(k)} \\
  x^{(k+1)} &= x^{(k)} + Br^{(k)} \\
  k &= 0, 1, \ldots
\end{align*}
\]

With the hope that this in a few steps will (produce) converge to \( x \)
\[ x^{(k+1)} = x^{(k)} + B(6 - Ax^{(k)}) = (I - BA)x^{(k)} + B6 \]

Assume for a while that \( x^{(0)} = y \)

\[ y = y - BAy + B6 \quad BAy = B6 \]

\( B \) non-singular \( \Rightarrow Ay = 6 \) \( y = x \)

If the process converges, then it converges to the desired solution of the system \( Ax = 6 \)

\[ x = x - BAx + B6 \]

\[ e^{(k+1)} = (I - BA)e^{(k)} \quad \text{error equation} \]

\[ \| e^{(k)} \| \to 0 \]

\[ \| e^{(k+1)} \| \leq \| I - BA \| \| e^{(k)} \| \leq \cdots \leq \delta^{k+1} \| e^{(0)} \| \]

\( \delta \) if \( \delta < 1 \) then the process converges

\[ e^{(k)} = (I - BA)^k e^{(0)} \]

Obviously \( \| I - BA \| = \delta < 1 \) is sufficient for the convergence of the iterative method (1).
**Theorem:** Assume that $\|I-BA\| < 1$ for some matrix norm subordinate to a vector norm. Then

$$x^{(0)} = B b, \quad x^{(k+1)} = (I-BA)x^{(k)} + Bb, \quad k=0,1,\ldots$$

converges to the solution of $Ax=b$.

**Note:** We did not assume $Ax=b$ has a solution. In fact $\|I-BA\| < 1$ ensures us that $A$ is a nonsingular matrix. Indeed, if $A$ were singular, then $Ac=0 \iff c \neq 0$.

Then

$$1 > \|I-BA\| = \sup_{\|x\| \leq 1, x \neq 0} \frac{\|(I-BA)x\|}{\|x\|} \geq \frac{\|(I-BA)c\|}{\|c\|} = \frac{\|c\|}{\|c\|} = 1 \Rightarrow 1 > 1$$

Contradiction!

**Proof:**

$$x^{(k+1)} - x = x^{(k)} - x + B(\varepsilon - A x^{(k)})$$

$$e^{(k+1)} = e^{(k)} - BAe^{(k)} = (I-BA)e^{(k)}$$

$$\|e^{(k+1)}\| \leq \delta^{(k+1)} \|e^{(0)}\| \quad e^{(k)} \to 0 \quad k \to \infty$$

One more thing

$$x^{(k+1)} - x = (I-BA)(x^{(k)} - x + x^{(k+1)} - x^{(k+1)})$$

$$\|x^{(k+1)} - x\| \leq \delta \|x^{(k)} - x^{(k+1)}\| + \delta \|x^{(k+1)} - x\|$$

$$(\delta - \varepsilon) \|x^{(k+1)} - x\| \leq \delta \|x^{(k)} - x^{(k+1)}\| \quad \varepsilon \|x^{(k+1)} - x\| \leq \frac{\delta}{1-\varepsilon} \|x^{(k)} - x^{(k+1)}\|$$
Denoting by \( Q = B^{-1} \sim A \) we aim at splitting of \( A \)

\[
x^{(k+1)} = x^{(k)} + B(B - Ax^{(k)})
\]

\[
Q(x^{(k+1)} - x^{(k)}) = B - Ax^{(k)} = r^{(k)}
\]

\[
Q x^{(k+1)} = (Q - A)x^{(k)} + b
\]

we can look at \( Q \) as some piece of \( A \)

this is the reason \( Q \) is called splitting of \( A \)

Basic concept of standard iterative methods is to find a nonsingular splitting of \( A \) such that

1. \( x^{(k)} \) is easy to compute \( \equiv Q x^{(k)} \) - easy to solve

2. \( x^{(k)} \) converges rapidly to \( x = Q^{-1} b \) - good approximation

we know that \( \| I - Q^{-1} A \| = \delta < 1 \) is sufficient to ensure smaller \( \delta \) provides faster convergence
More facts from linear algebra

1. \( \det(A) \cdot \det(AB) = \det(A) \cdot \det(B) \)

2. Eigenvalues and eigenvectors of \( A \in \mathbb{R}^{n \times n} \)
\[
\det(A - \lambda I) = (-1)^n + (a_{11} + \ldots + a_{nn}) \lambda^{n-1} + \ldots + \det(A) = 0
\]
characteristic polynomial of \( A \)

counting their multiplicities \( \lambda_1, \ldots, \lambda_m \) in the complex plane

Example: \( \det(I - \lambda I) = (1-\lambda)^n \) \( \lambda_1 = \lambda_2 = \ldots = \lambda_n = 1 \)

Take \( j \), \( A\psi_j = \lambda_j \psi_j \), \( \psi_j \): eigenvector

One of the most difficult questions in LA is to answer how many linearly independent eigenvectors we can get?

\[
D = \begin{bmatrix}
d_{11} & 0 & \cdots & 0 \\
0 & d_{22} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & d_{nn}
\end{bmatrix}
\]
\[
\det(D - \lambda I) = \prod_{i=1}^n (d_{ii} - \lambda)
\]
\[
\lambda_i = d_{ii} \quad i = 1 \ldots n
\]
eigenvectors will be \( e_i \)
\((\lambda_i, e_i)\) eigenspair
There is a class of matrices that this question is fully understood and solved, the class of symmetric matrices.

If $A$ is a symmetric matrix in $\mathbb{R}^{n \times n}$, then:

1. It has real eigenvalues.
2. It has full set of eigenvectors, i.e., $(\psi_1, \ldots, \psi_n)$ span the space $\mathbb{R}^n$ (or form a basis for $\mathbb{R}^n$).

We can orthonormalize them so that $(\psi_i, \psi_j) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$

$x = \sum_{j=1}^{n} g_j \psi_j \quad g_j = (x, \psi_j)$ "Fourier" representation

$\|x\|_2^2 = \sum_{j=1}^{n} g_j^2 \quad \ell_2$-norm

**Def:** The matrix $W$ is unitary (complex elements)

\[ W^* W = I, \quad W^T = W^* \]

**Example**

\[
W = \begin{bmatrix}
\frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}} \\
\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{bmatrix} \quad W^* = \begin{bmatrix}
\frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\
-\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{bmatrix}
\]

$W W^* = \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}$