Linear algebra

We shall turn our attention to solving linear systems of equations $Ax = b$

$$A \in \mathbb{R}^{m \times n}$$
$$x \in \mathbb{R}^{n}$$
$$b \in \mathbb{R}^{m}.$$

We already saw examples of methods that required the solution of a linear system as part of the overall algorithm, e.g. the Vandermonde system for interpolation, which was a square system ($m = n$).

Another category of methods that leads to rectangular systems with $m > n$ is least square methods. They answer questions of the form:

→ What is the best n-order polynomial we can use to approximate (not interpolate) $(m+1)$ data points (where $m > n$).

→ More generally, find the solution that most closely satisfies m equations, in the presence of n ($n < m$) unknowns.

All these algorithms need to be conscious about <u>error</u>, and there are at least 3 sources for it.

→ Some algorithms are "imperfect" in the sense that they require several iterations to generate a good quality approximation. Thus, intermediate results are subject to <u>error</u>

→ Sometimes, it is not possible to find an "ideal" solution, e.g. because we have more equations than unknowns. In this case, not all equations will be satisfied exactly, and we need a notion of the "error" incurred in not satisfying certain equations fully.

→ Inputs to an algorithm are often corrupted by noise, roundoff error, etc. For example, instead of solving an "intended" system $Ax = b$ we may be solving $A^* x = b^*$ where the entries in A^* & b^* have been subject to noise and inaccuracy. It is important to know how those translate to errors in determining x.

(Vector and Matrix) Norms :

3/10/11

Norms are valuable tools in arguing about the extent and magnitude of error. We will introduce some concepts that we will use broadly later on.

Definition A vector norm is a function from $\mathbb{R}^n$ to $\mathbb{R}$, with a certain number of properties. If $\underline{x} \in \mathbb{R}^n$, we symbolize its norm by $\|\underline{x}\|$. The defining properties of a norm are:

(i) $\|\underline{x}\| \geq 0$ for all $x \in \mathbb{R}^n$

also $\|\underline{x}\| = 0$ iff $\underline{x} = \underline{0}$

(ii) $\|\alpha \underline{x}\| = |\alpha| \cdot \|\underline{x}\|$ for all $\alpha \in \mathbb{R}$, $\underline{x} \in \mathbb{R}^n$

(iii) $\|\underline{x}+\underline{y}\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$

Note that the properties above do not determine a unique for of a "norm" function, in fact many different valid norms exist. Typically, we will use subscripts ($\|\cdot\|_a, \|\cdot\|_b$) to denote different types of norms.

Vector norms – Why are they needed?

When dealing e.g with the solution of a nonlinear equation $f(x) = 0$, the error $e = x_{approx} - x_{exact}$ is a single number, thus the absolute value $|e|$ gives us a good idea of the "extent" of error.

When solving a system of linear equations $A\underline{x} = \underline{b}$, the exact solution $\underline{x}_{exact}$ as well as any approximation $\underline{x}_{approx}$ are vectors, and the error:

$$\underline{e} = \underline{x}_{approx} - \underline{x}_{exact}$$

is a vector, too. It is not as straightforward to assess the "magnitude" of such a vector-valued error.

e.g. Consider $\underline{e}_1, \underline{e}_2 \in \mathbb{R}^{1000}$, and

$$\underline{e}_1 = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \\ \vdots \\ \vdots \\ 0.1 \end{pmatrix} \qquad \underline{e}_2 = \begin{pmatrix} 100 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Which one is worse? e_1 has a modest amount of

error, distributed over all components. In $\underline{e}_2$, all but one component are exact, but one of them has a huge discrepancy.

Exactly how we quantify and assess the extent of error is application-dependent. Vector norms are alternative ways to measure this magnitude, and different norms would be appropriate for different tasks.

Def. A vector norm is a function from $\mathbb{R}^n \to \mathbb{R}^+$, which maps a vector $\underline{v}$ to the real number $\|\underline{v}\|$. This symbol must satisfy the properties:

(i) $\|\underline{x}\| \geq 0$ for all $\underline{x} \in \mathbb{R}^n$. Also $\|\underline{x}\| = 0$ iff $\underline{x} = 0$

(ii) $\|\alpha \underline{x}\| = |\alpha| \cdot \|\underline{x}\| \quad \forall \underline{x} \in \mathbb{R}^n, \alpha \in \mathbb{R}$

(iii) $\|x+y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathbb{R}^n$ (triangle inequality).

Some norms which can be proven to satisfy these properties, are: (Here $\underline{x} = (x_1, x_2, \ldots, x_n)$)

1. The L_1-norm (or 1-norm)

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

2. The L_2-norm (or 2-norm, or Euclidean norm)

$$\|\underline{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

3. The infinity norm (or max-norm)

$$\|\underline{x}\|_\infty = \max_{1 \le i \le n} |x_i|$$

4. (Less common) L_p-norm

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

It is relatively easy to show that these satisfy the defining properties of a norm. e.g. for $\|\ \|_1$:

- $\|\underline{x}\|_1 = \sum_{i=1}^{n} |x_i| \geq 0$

- if $\underline{x} = \underline{0}$, then $\|x\|_1 = 0$

  if $\|x\|_1 = 0 \Rightarrow \sum_{i=1}^{n} |x_i| = 0 \Rightarrow |x_i| = 0 \ \forall i \Rightarrow \underline{x} = 0$

- $\|\alpha \underline{x}\| = \sum_{i=1}^{n} |\alpha x_i| = |\alpha| \sum_{i=1}^{n} |x_i| = |\alpha| \, \|\underline{x}\|_1$

- $\|\underline{x}+\underline{y}\| = \sum_{i=1}^{n} |x_i + y_i| \leq \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| = \|x\|_1 + \|y\|_1$

Similar proofs can be given for $\|\cdot\|_\infty$ (just as easy), $\|\cdot\|_2$ (a bit more difficult) and $\|\cdot\|_p$ (rather complicated).

We can actually define norms for (square) matrices, as well. A matrix norm is a function

$\|\cdot\| : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}$ which satisfies:

(i) $\|M\| \geq 0 \quad \forall M \in \mathbb{R}^{n\times n}$. $\|M\| = 0$ iff $M = 0$

(ii) $\|\alpha M\| = |\alpha| \, \|M\|$

(iii) $\|M+N\| \leq \|M\| + \|N\|$

(iv) $\|M \cdot N\| \leq \|M\| \cdot \|N\|$.

(Property (iv) is the one that has slightly different flavor than vector norms).

Although <u>more types of matrix norms exist</u>, one common category is that of matrix norms <u>induced</u> by vector norms.

<u>Def</u>. If $\|\cdot\|_*$ is a valid vector norm, its <u>induced</u> matrix norm is defined as

$$\|M\|_* = \max_{x\in\mathbb{R}^n \, x\neq 0} \left\{ \frac{\|Mx\|_*}{\|x\|_*} \right\}$$

or equivalently:

$$\|M\|_* = \max_{\substack{x \in \mathbb{R}^n \\ \|x\|_* = 1}} \{\|Mx\|_*\}$$

Note, again, that not all valid matrix norms are induced by vector norms. One notable example is the very commonly used Frobenius norm:

$$\|M\|_F = \sqrt{\sum_{i,j=1}^{n} M_{ij}^2}$$

We can easily show though that induced norms satisfy properties (i) through (iv). (i)-(iii) are rather trivial, eg:

$$\|M+N\| = \max_{\underline{x} \neq 0} \frac{\|(M+N)\underline{x}\|}{\|\underline{x}\|} \leq \max_{\underline{x} \neq 0} \frac{\|M\underline{x}\| + \|N\underline{x}\|}{\|\underline{x}\|}$$

$$= \max_{\underline{x} \neq 0} \frac{\|Mx\|}{\|x\|} + \max_{\underline{x} \neq 0} \frac{\|Nx\|}{\|x\|} = \|M\| + \|N\|.$$

Property (iv) is slightly trickier to show.

First, a lemma:

Lemma If $\|\cdot\|$ is a matrix norm induced by a vector norm $\|\cdot\|$, then:

$$\|Ax\| \leq \|A\| \cdot \|x\|$$

Proof: Since $\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$, we have that for an arbitrary $y \in \mathbb{R}^n$: $(y \neq 0)$

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \geq \frac{\|Ay\|}{\|y\|} \Rightarrow$$

$\Rightarrow \|Ay\| \leq \|A\| \|y\|$. This holds for $y \neq 0$, but we can see it is also true for $y = 0$

Prop. (iv)

$$\|MN\| = \max_{\|x\|=1} \|MNx\| \leq \max_{\|x\|=1} \|M\| \|Nx\| =$$

$$= \|M\| \cdot \max_{\|x\|=1} \|Nx\| = \|M\| \cdot \|N\| \Rightarrow$$

$$\|MN\| \leq \|M\| \cdot \|N\|.$$

Although the definition of an induced norm allowed us to prove certain properties, it does not necessarily provide a convenient formula for evaluating the matrix norm.

Fortunately, such formulas do exist for the L_1 and L_∞ induced matrix norms. Given here without proof:

$$\|A\|_1 = \max_j \sum_{i=1}^{n} |A_{ij}| \qquad \text{(max. absolute column sum)}$$

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |A_{ij}| \qquad \text{(max. absolute row sum)}.$$

($\|\cdot\|_2$ is much more complicated!)

Where do these vector/matrix norms come handy?

Useful properties of matrix & vector norms

3/24/11 L

We previously saw that

$$\|Ax\| \leq \|A\| \cdot \|x\| \qquad (1)$$

for any matrix A, and any vector x (of dimensions $m \times m$ and $m \times 1$, respectively).

Note that, when writing an expression such as (1), the matrix norm $\|A\|$ is understood to be the inferred norm from the vector norm used in $\|Ax\|$ and $\|x\|$. Thus

$$\|Ax\|_1 \leq \|A\|_1 \cdot \|x\|_1$$

and

$$\|Ax\|_\infty \leq \|A\|_\infty \cdot \|x\|_\infty$$

are both valid, but we cannot mix and match, e.g.:

$$\|Ax\|_\infty \leq \|A\|_2 \cdot \|x\|_1$$ → NOT CORRECT

When solving a linear system $A\underline{x} = \underline{b}$, computer algorithms are only providing an approximation ($\underline{x}_{app}$) to the exact solution ($\underline{x}_{ex}$). This is due to factors such as finite precision, roundoff errors or even imperfect solution algorithms. In either case, we have an <u>error</u> (error vector, in fact) defined as

$$\underline{e} = \underline{x}_{ap} - \underline{x}_{ex}$$

Naturally, we would like to have an understanding of the <u>magnitude</u> of this error (e.g. some appropriate norm $\|\underline{e}\|$). The problem is that we do not know the exact, pristine solution $\underline{x}_{ex}$!

One remedy is offered via the <u>residual vector</u> defined as:

$$\underline{r} = \underline{b} - A\underline{x}_{app}$$

The vector $\underline{r}$ is something we can compute practically since it involves only known quantities ($\underline{b}, A, \underline{x}_{app}$).

Furthermore, we have:

$$\underline{r} = \underline{b} - A\,\underline{x}_{app}$$

$$= A\,\underline{x}_{ex} - A\,\underline{x}_{app}$$

$$= -A\left(\underline{x}_{app} - \underline{x}_{ex}\right)$$

$$= -A\,\underline{e} \quad \Longrightarrow \quad \underline{r} = -A\,\underline{e}$$

$$\underline{e} = -A^{-1}\,\underline{r}$$

The last equation links the error with the residual.

Furthermore, we can write

$$\|\underline{e}\| = \|A^{-1} r\| \leq \|A^{-1}\|\,\|r\|$$

This equation provides a <u>bound</u> for the error, as a function of $\|A^{-1}\|$ and the norm of the computable vector $\underline{r}$! Note that:

→ We can obtain this estimate <u>without</u> knowing the exact solution, <u>but</u>

→ We need $\|A^{-1}\|$ and generally, computing A^{-1} is just as difficult (if not more) than finding $\underline{x}_{ex}$. <u>However</u> there are special cases where an estimate of $\|A^{-1}\|$ can be obtained.

A different source of error :

Sometimes, the right-hand-side (b) of $Ax = b$ has errors that make it deviate from its intended value. For example in the Vandermonde matrix method for polynomial interpolation, $\underline{b}$ contains the samples $(y_1 = f(x_1), y_2, \ldots, y_n)$ where $y_i = f(x_i)$. An error in a measuring device supposed to sample $f(x)$ could lead to erroneous readings y_i^* instead of y_i. In general, measuring inaccuracies can lead to the right-hand-side vector $\underline{b}$ being mis-represented as $\underline{b}^*$ $(\neq b)$.

In this case, instead of the intended solution $x = A^{-1}b$ we in fact compute $x^* = A^{-1}b^*$.

How important is the error $e = x^* - x$ that is caused by this misrepresentation of b?

Let us introduce some notation:

Let $\delta \underline{b} := \underline{b}^* - \underline{b}$

$\delta \underline{x} := \underline{x}^* - \underline{x}$

$$A\underline{x} = \underline{b}$$
$$A\underline{x}^* = \underline{b}^*$$

$$A(\underline{x}^* - \underline{x}) = \underline{b}^* - \underline{b}$$

$$A\,\delta\underline{x} = \delta\underline{b}$$

$$\delta x = A^{-1}\,\delta b$$

Taking norms:

$$\|\delta x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\|\,\|\delta b\| \qquad (1)$$

Thus the error in the computed solution δx is proportional to the error in b.

An even more relevant question is: How does the relative error $\frac{\|\delta x\|}{\|x\|} = \frac{\|x^* - x\|}{\|x\|}$ compare to the relative error in $\underline{b}$ $\frac{\|\delta b\|}{\|b\|}$? This may be more useful to know, since $\|\delta b\|$ may be impossible to compute (if we don't know the real $\underline{b}$!).

For this, we write

$$Ax = b \Rightarrow \|b\| = \|Ax\| \leq \|A\| \cdot \|x\|$$

$$\Rightarrow \frac{1}{\|x\|} \leq \|A\| \cdot \frac{1}{\|b\|} \quad (2)$$

Multiplying (1) & (2) we get:

$$\frac{\|\delta x\|}{\|x\|} \leq \underbrace{\|A\| \cdot \|A^{-1}\|}_{\kappa(A)} \cdot \frac{\|\delta\|}{\|b\|}$$

Thus the relative error in $\underline{x}$ is bounded by a multiple of the relative error in $\underline{b}$! The multiplicative constant $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ is called the condition number of A, and is an important measure of the sensitivity of a linear system $Ax=b$ to being solved on a computer, in the presence of inaccurate values.

e.g. If the relative error $\frac{\|\delta b\|}{\|b\|}$ is 0.0001%, but $\kappa(A) = 100.000$ (could happen!), Then we could have up to a 10% error in the computed $\underline{x}$!

Simply, almost <u>any</u> b will have <u>some</u>, small relative error, due to the fact it is represented on a computer up to machine precision! The relative error will be at least as much as the machine epsilon due to roundoff!

$$\frac{\|\delta b\|_\infty}{\|b\|} \geq \varepsilon \cong 10^{-7} \quad \text{(for } \underline{\text{floats}}\text{)}.$$

But, how bad can the condition number get?

<u>Very</u> bad, at times. For example:

Hilbert matrices $H_n \in \mathbb{R}^{n\times n}$ $(H_n)_{ij} = \dfrac{1}{i+j-1}$

$$H_5 = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 \\ 1/3 & 1/4 & 1/5 & 1/6 & 1/7 \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 \end{bmatrix}$$

$$\kappa_\infty(H_5) = \|H_5\|_\infty \cdot \|H_5^{-1}\|_\infty \cong 10^6 !$$

Thus, any attempt at solving $H_5 x = b$ would be subject to a relative error up to 10% <u>just due</u> to roundoff errors in <u>b</u>!

Another case : near-singular matrices

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 6+\varepsilon \end{bmatrix}$$

As $\varepsilon \to 0$ A becomes singular (non-invertible)

In this case, $\kappa(A) \to \infty$.

What is the best case for $\kappa(A)$?

Lemma For any vector-induced matrix norm, we have $\|I\| = 1$.

Proof From definition:

$$\|I\| = \max_{x \neq 0} \frac{\|Ix\|}{\|x\|} = \max_{x \neq 0} \frac{\|x\|}{\|x\|}$$

Using property (iv) of matrix norms, we get:

$$I = A \cdot A^{-1} \Rightarrow 1 = \|I\| = \|A \cdot A^{-1}\| \leq \|A\| \cdot \|A^{-1}\|$$

Thus $\boxed{\kappa(A) \geq 1}$

The "best" conditioned matrices are of the form $A = c \cdot I$!

And have $\kappa(A) = 1$.