Inner products

An **inner product** $(\cdot,\cdot)$ is a mapping from (S,S), where S is a (vector, matrix, function, etc.) space to **R**, with the following properties: Let $x,y,z \in S$
1) $(x,x) \geq 0$ and $(x,x) = 0$ iff $x = 0$
2) $(\alpha x, y) = \alpha(x,y)$ for all scalars α
3) $(x+z,y) = (x,y) + (z,y)$ (triangle equality)
4) $(x,y) = (y,x)$.

A commonly used inner product for vectors is the dot product, $(x,y) = x^T \cdot y = \sum_{i=1}^n x_i \cdot y_i$.

Vector norms and matrix norms

A **norm** $||\cdot||$ is a mapping from a space S to $\mathbf{R}^+ \cup \{0\}$, with the following properties:
Let $x,y \in S$
1) $||x|| \geq 0$ and $||x|| = 0$ iff $x = 0$
2) $||\alpha x|| = |\alpha| ||x||$ for all scalars α
3) $||x+y|| \leq ||x|| + ||y||$ (triangle inequality).

Norms provide a means for quantifying multiple-component quantities by a single number. Thus, we can quantify the magnitude of a single vector, matrix, or function, as well as the difference (**distance**) between two vectors, or two functions, etc.

Let $x = (x_1, x_2, \ldots, x_n)^T$.
The following are vector norms:
- **max (infinity) norm**: $||x||_\infty \equiv \max_{i=1}^n \{|x_i|\}$
- **Euclidean norm (length):** $||x||_2 \equiv \sqrt{(x,x)} \equiv (\sum_{i=1}^n x_i^2)^{1/2}$
- **one-norm:** $||x||_1 \equiv \sum_{i=1}^n |x_i|$.

The following can be proved:
- **Schwarz inequality:** $|(x,y)| \leq ||x||_2 ||y||_2$

A common matrix norm is:
p-**norm** (or **induced** norm), $p = 1, 2, \infty$:
$||A||_p \equiv \max_{x \neq 0} \{\frac{||Ax||_p}{||x||_p}\}$.

For the p-norms three more properties are valid:
4) $||Ax|| \leq ||A|| ||x||$
5) $||I|| = 1$
6) $||AB|| \leq ||A|| ||B||$.

The following can be proved:
- $\max_{x \neq 0} \{\frac{||Ax||_p}{||x||_p}\} = \max_{||x||_p = 1} \{||Ax||_p\}$
- $||A||_\infty = \max_{i=1}^n \{\sum_{j=1}^n |a_{ij}|\}$(**row norm**)
- $||A||_1 = \max_{j=1}^n \{\sum_{i=1}^n |a_{ij}|\}$ (**column norm**)
- $||A||_2 = \sqrt{\rho(A^T A)}$

The vector $x = (-1, 1, -2)^T$ has norms:
$||x||_\infty = \max\{|-1|, |1|, |-2|\} = 2$
$||x||_2 = \sqrt{(-1)^2 + (1)^2 + (-2)^2} = \sqrt{6}$
$||x||_1 = |-1| + |1| + |-2| = 4$

The matrix

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 3 & -1 \\ 5 & -1 & 1 \end{bmatrix}$$

has norms:
$||A||_\infty = \max\{4, 4, 7\} = 7$
$||A||_1 = \max\{6, 6, 3\} = 6$

Accuracy of solutions of linear systems

Consider the system $Ax = b$

$$\begin{bmatrix} 0.780 & 0.563 \\ 0.457 & 0.330 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.217 \\ 0.127 \end{bmatrix}$$

Perform GE in 3 decimal digits

$$\begin{bmatrix} 0.780 & 0.563 \\ 0 & 0.0000820 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.217 \\ -0.000162 \end{bmatrix}$$

By BS

$x_2 = \frac{-0.162 \times 10^{-3}}{0.820 \times 10^{-4}} = -1.98$,

$x_1 = \frac{0.217 - 0.563 \times x_2}{.780} = 1.71$

The **residual**, defined as $r = \tilde{b} - \tilde{A}\tilde{x}$, where $\tilde{b}$ and $\tilde{A}$ are the computer representation of b and A, respectively, and $\tilde{x}$ is the computed solution, is

$$\begin{bmatrix} -0.00206 \\ -0.00107 \end{bmatrix} \approx 10^{-3} \sim E$$

But the exact solution is $x = (1, -1)^T$.
GE guarantees small residuals but small residuals do not guarantee small errors.

Condition number of matrix

Original problem: $Ax = b$
Representation on computer : $\tilde{A}\tilde{x} = \tilde{b}$
Final problem solved : $\tilde{A}\tilde{\tilde{x}} = \tilde{b} - r$

The representation on the computer is due to round-off errors, while the final problem solved is due to operations' errors.

The **condition number** of a non-singular matrix A is given by $\kappa_a(A) = ||A||_a ||A^{-1}||_a$ and is a measure of the relative sensitivity of the solution x of $Ax = b$ to relative changes in A and b.

The following can be proved:

$$\frac{||x - \tilde{x}||_a}{||x||_a} \leq \kappa_a(A) \left(\frac{||b - \tilde{b}||_a}{||b||_a} + \frac{||A - \tilde{A}||_a}{||A||_a} + \frac{||r||_a}{||b||_a} \right)$$

If A and b are exactly representable then

$$\frac{||x - \tilde{x}||_a}{||x||_a} \leq \kappa_a(A) \frac{||r||_a}{||b||_a},$$

where $r = b - A\tilde{x}$.

Iterative refinement

Let A and b be exactly representable, and let $Ax = b$ be the system to be solved.

Due to operations' errors, $\tilde{x}$ is computed, such that $A\tilde{x} = b - r$. Thus, $r = b - A\tilde{x} = Ax - A\tilde{x} = A(x - \tilde{x}) = Ae$, where $e \equiv x - \tilde{x}$ is the error in the computed soln. Therefore, e is the soln to $Ae = r$, and $x = \tilde{x} + e$.

However, to compute e, we need to have r computed accurately, and $Ae = r$ solved accurately. We compute r in some reasonable high accuracy, then solve $Ae = r$, then compute $\tilde{\tilde{x}} = \tilde{x} + e$. Usually, $\tilde{\tilde{x}}$ is a better approximation to x than $\tilde{x}$.

Iterative refinement algorithm

choose a tolerance (tol) and max. no. of iterations (maxit)
obtain LU factorization of A by GE
solve $Ly = b$
solve $Ux = y$

for $k = 1$ to maxit
    compute $r = b - Ax$ in **double precision**
    solve $Ly = r$
    solve $Ue = y$
    compute $x = x + e$
    if $||e|| <$ tol exit loop
endfor

It is important to note that all computations should be performed in low precision, except for the computation of the residual which is done in high precision.

Overdetermined systems

An **overdetermined linear system** $Ax = b$ has m equations and n unknowns, where $m > n$ and the m equations are inconsistent. Such a system has no solution.

Examples:
$2x_1 + x_2 = 4, 3x_1 - x_2 = 1, x_1 + 2x_2 = 3$
is a 3×2 overdetermined system.
$x_1 = 2, 2x_1 = 4$ is not overdetermined.

Overdetermined systems arise frequently in physical modelling. Certain unknowns parameters must satisfy some physical laws. In order to determine the unknowns we make experimental measurements, often more than necessary in theory.

When such a system arises, there is no x that satisfies all equations of the system. Therefore, we seek to find an x that minimizes the residual $r = b - Ax$ in the Euclidean norm. This x is called the **least squares solution** to the overdetermined system.

Normal equations

Let $Ax = b$, with $A \in \mathbf{R}^{m\times n}$ and $m > n$, be an overdetermined system. We are looking for a vector x that minimizes $||r||_2$ or equivalently
$||r||_2^2 = r_1^2 + r_2^2 + \ldots + r_m^2$. Set

$$\frac{\partial}{\partial x_j}(r_1^2 + \ldots + r_m^2) = 0, j = 1, \ldots, n \qquad (1)$$

we have

$$\frac{\partial}{\partial x_j}(r_1^2 + \ldots + r_m^2) = 2r_1\frac{\partial r_1}{\partial x_j} + \ldots + 2r_m\frac{\partial r_m}{\partial x_j} \qquad (2)$$

And $r_i = b_i - a_{i1}x_1 - \ldots - a_{in}x_n$.
Thus $\frac{\partial r_i}{\partial x_j} = -a_{ij}$, and (2) becomes

$$\frac{\partial}{\partial x_j}(r_1^2 + \ldots + r_m^2) = -2r_1a_{1j} - \ldots - 2r_ma_{mj} \qquad (3)$$

From (1) and (3) we have:

$$\begin{array}{c} a_{11}r_1 + a_{21}r_2 + \ldots + a_{m1}r_m = 0 \\ a_{12}r_1 + a_{22}r_2 + \ldots + a_{m2}r_m = 0 \\ \ldots \\ a_{1n}r_1 + a_{2n}r_2 + \ldots + a_{mn}r_m = 0 \end{array} \qquad (4)$$

Equations (4) can be equivalently be written in matrix form:
$A^Tr = 0 \Rightarrow A^T(b - Ax) = 0 \Rightarrow A^TAx = A^Tb$
This is a $n \times n$ linear system. The corresponding equations are called **normal equations** for the overdetermined system $Ax = b$.

Method of normal equations for solving overdetermined systems $Ax = b$:
Compute A^TA
Compute A^Tb
Solve $(A^TA)x = A^Tb$

Notes:
- The x computed above does not satisfy $Ax = b$.
- A is $m \times n$, A^T is $n \times m$, thus A^TA is $n \times n$.
- $(A^TA)_{ij}$ is dot product of columns i and j of A.
- A^TA is symmetric.
- The matrix A^TA is often ill-conditioned, especially for large problems.

Operation counts:
Compute $A^TA : \frac{n(n+1)}{2}m$ flops.
Compute $A^Tb : mn$ flops.
Solve $(A^TA)x = A^Tb : \frac{n^3}{6} + n^2$ flops (Sym GE and F/BS).
Approximately $\frac{n^2m}{2} + \frac{n^3}{6}$ flops total.

Underdetermined systems

An **underdetermined linear system** $Ax = b$ is a linear system with m equations and n unknowns, where $m < n$. An underdetermined linear system has more than one solutions.

Example: $2x_1 + x_2 = 4$ is a 1×2 under. lin. sys.

Such systems arise frequently in optimization theory and in economic modelling. Certain unknown parameters must satisfy some constraints, but the number of constraints may not be enough to uniquely determine the unknowns.

When such systems arise, we often seek the minimum(in Euclidean norm) solution x that satisfies the equations. This x is called the **least squares solution** to the underdetermined system. It can be shown that $x = A^T t$, for some unknown vector t. Then $Ax = b \Rightarrow AA^T t = b$. The corresponding equations are called **normal equations**.

Method of normal equations for solving underdetermined systems $Ax = b$:
Compute AA^T
Solve $(AA^T)t = b$
Compute $x = A^T t$

Notes:
- The x computed above satisfies $Ax = b$.
- A is $m \times n$, A^T is $n \times m$, thus AA^T is $m \times m$.
- $(AA^T)_{ij}$ is dot product of rows i and j of A.
- AA^T is symmetric.
- The matrix AA^T is often ill-conditioned, especially for large problems.

Operation counts:
Compute AA^T : $\frac{m(m+1)}{2}n$ flops.
Compute $A^T t : mn$ flops.
Solve $(AA^T)t = b$: $\frac{m^3}{6} + m^2$ flops (Sym GE and F/BS).
Approximately $\frac{nm^2}{2} + \frac{m^3}{6}$ flops total.