Computational Techniques: 233 – 2

Peter Harrison

Department of Computing, Imperial College London

March 7, 2011

Contents of part 2 of Course 233

The following topics will be covered:

- Conditioning
- Convergence and fixed point problems
- Iterative solution of linear equations
- Laplace transforms
- Functions of several variables
- Introduction to continuous optimisation

Note that Sparse Matrix Techniques will not be covered: they are discussed in the "Background Notes" however.

A polynomial equation

Consider the polynomial equation

$$(x-1)^4 = 0$$

Four coincident roots, $x = 1$

A polynomial equation

Consider the polynomial equation

$$(x-1)^4 = 0$$

Four coincident roots, $x = 1$

Now try $(x-1)^4 = 10^{-8}$

A polynomial equation

Consider the polynomial equation

$$(x-1)^4 = 0$$

Four coincident roots, $x = 1$

Now try $(x-1)^4 = 10^{-8}$

One exact root is now 1.01

A polynomial equation

Consider the polynomial equation

$$(x-1)^4 = 0$$

Four coincident roots, $x = 1$

Now try $(x-1)^4 = 10^{-8}$

One exact root is now 1.01

The change of 10^{-8} in a 'parameter' has caused a change of 10^{-2} in the solution: a ratio of 1000000 !

Linear equations

Now consider the equations

$$x + y = 1$$
$$x + \alpha y = 0$$

Linear equations

Now consider the equations

$$x + y = 1$$
$$x + \alpha y = 0$$

The solution is $x = 1 - 1/(1-\alpha), \;\; y = 1/(1-\alpha)$

Linear equations

Now consider the equations

$$x + y = 1$$
$$x + \alpha y = 0$$

The solution is $x = 1 - 1/(1-\alpha), \;\; y = 1/(1-\alpha)$

A small change in a coefficient (α) leads to a large change in the solution when $\alpha \sim 1$

Now consider the equations

$$x + y = 1$$
$$x + \alpha y = 0$$

The solution is $x = 1 - 1/(1-\alpha), \;\; y = 1/(1-\alpha)$

A small change in a coefficient (α) leads to a large change in the solution when $\alpha \sim 1$

If α increases from 0.999 to 0.9999, the solution changes from $(-999, 1000)$ to $(-9999, 10000)$: an increase by a factor of 10 from a change of about 0.1% in the parameter α.

Mathematical conditioning

The *condition number*, or just *condition*, of a problem P is the maximum size-ratio:

$$\kappa(P) = \max_{d_1, d_2} \frac{\|s(d_1) - s(d_2)\|}{\|d_1 - d_2\|}$$

where d_1, d_2 are alternate inputs to P and $s(d_1), s(d_2)$ are the corresponding solutions (outputs)

Mathematical conditioning

The *condition number*, or just *condition*, of a problem P is the maximum size-ratio:

$$\kappa(P) = \max_{d_1, d_2} \frac{\|s(d_1) - s(d_2)\|}{\|d_1 - d_2\|}$$

where d_1, d_2 are alternate inputs to P and $s(d_1), s(d_2)$ are the corresponding solutions (outputs)

In the above examples, the ratios were big: 10^6 and 10^4 ... and these are not worst cases!

Mathematical conditioning

The *condition number*, or just *condition*, of a problem P is the maximum size-ratio:

$$\kappa(P) = \max_{d_1, d_2} \frac{\|s(d_1) - s(d_2)\|}{\|d_1 - d_2\|}$$

where d_1, d_2 are alternate inputs to P and $s(d_1), s(d_2)$ are the corresponding solutions (outputs)

In the above examples, the ratios were big: 10^6 and 10^4 ... and these are not worst cases!

A relatively small number (near to 1) implies that even in the worst case, the solution will not be too sensitive to small changes in the input ... will not "blow up".

Key point is that we're looking at the *worst case*.

Matrix norm

Define, for each vector norm $(1, 2, \infty)$, the subordinate matrix norm:

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|u\|=1} \|Au\|$$

Matrix norm

Define, for each vector norm $(1, 2, \infty)$, the subordinate matrix norm:

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|u\|=1} \|Au\|$$

A lower bound on the norm is:

$$\|A\| \geq \frac{\|Ax\|}{\|x\|}$$

for any specific non-zero vector x.

Subordinate norms

Proposition: The subordinate matrix norms are those defined previously for the corresponding vector norms.

Subordinate norms

Proposition: The subordinate matrix norms are those defined previously for the corresponding vector norms.

We'll prove for 1-norms.

Subordinate norms

Proposition: The subordinate matrix norms are those defined previously for the corresponding vector norms.

We'll prove for 1-norms.

$$\begin{aligned}
\|Ax\|_1 &= \|(\ldots, \sum_j a_{ij}x_j, \ldots)\|_1 \\
&\leq \sum_{i,j} |a_{ij}x_j| \\
&= \sum_{i,j} |a_{ij}||x_j| \\
&= \sum_j \|a_j\|_1 |x_j| \\
&\leq \max_j \|a_j\|_1 \|x\|_1 \\
&= \|x\|_1 \|A\|_1
\end{aligned}$$

with equality for some vector x.

Perturbation

Consider the linear equations $Ax = b$

Perturbation

Consider the linear equations $Ax = b$

Let the perturbed equation, with b changed to $b + \delta b$, have solution $x + \delta x_b$, so

$$A(x + \delta x_b) = b + \delta b$$

Perturbation

Consider the linear equations $Ax = b$

Let the perturbed equation, with b changed to $b + \delta b$, have solution $x + \delta x_b$, so

$$A(x + \delta x_b) = b + \delta b$$

Hence $A\delta x_b = \delta b$ and so $\delta x_b = A^{-1}\delta b$

Perturbation

Consider the linear equations $Ax = b$

Let the perturbed equation, with b changed to $b + \delta b$, have solution $x + \delta x_b$, so

$$A(x + \delta x_b) = b + \delta b$$

Hence $A\delta x_b = \delta b$ and so $\delta x_b = A^{-1}\delta b$

Therefore $\|\delta x_b\| \leq \|A^{-1}\|\|\delta b\|$ and equality can be attained.

Perturbation

Consider the linear equations $Ax = b$

Let the perturbed equation, with b changed to $b + \delta b$, have solution $x + \delta x_b$, so

$$A(x + \delta x_b) = b + \delta b$$

Hence $A\delta x_b = \delta b$ and so $\delta x_b = A^{-1}\delta b$

Therefore $\|\delta x_b\| \leq \|A^{-1}\|\|\delta b\|$ and equality can be attained.

Relative perturbation is

$$\frac{\|\delta x_b\|}{\|x\|} \leq \|A^{-1}\|\|A\|\frac{\|\delta b\|}{\|b\|}$$

Condition number

Similarly we can perturb the elements of the matrix A.

Condition number

Similarly we can perturb the elements of the matrix A.

$$(A + \delta A)(x + \delta x_A) = b$$

Condition number

Similarly we can perturb the elements of the matrix A.

$$(A + \delta A)(x + \delta x_A) = b$$

Ignoring second-order quantities and recalling $Ax = b$ we get (not quite the same as the notes)

Condition number

Similarly we can perturb the elements of the matrix A.

$$(A + \delta A)(x + \delta x_A) = b$$

Ignoring second-order quantities and recalling $Ax = b$ we get (not quite the same as the notes)

$$\frac{\|\delta x_A\|}{\|x\|} \leq \|A^{-1}\|\|A\|\frac{\|\delta A\|}{\|A\|}$$

Condition number

Similarly we can perturb the elements of the matrix A.

$$(A + \delta A)(x + \delta x_A) = b$$

Ignoring second-order quantities and recalling $Ax = b$ we get (not quite the same as the notes)

$$\frac{\|\delta x_A\|}{\|x\|} \leq \|A^{-1}\|\|A\|\frac{\|\delta A\|}{\|A\|}$$

Therefore define:

$$\mathrm{cond}(A) = \|A^{-1}\|\|A\|$$

Condition number

Similarly we can perturb the elements of the matrix A.

$$(A + \delta A)(x + \delta x_A) = b$$

Ignoring second-order quantities and recalling $Ax = b$ we get (not quite the same as the notes)

$$\frac{\|\delta x_A\|}{\|x\|} \leq \|A^{-1}\|\|A\|\frac{\|\delta A\|}{\|A\|}$$

Therefore define:

$$\mathrm{cond}(A) = \|A^{-1}\|\|A\|$$

Then

$$\mathrm{cond}(A) \geq \max\left\{\frac{\|\delta x_b\|/\|x\|}{\|\delta b\|/\|b\|}, \frac{\|\delta x_A\|/\|x\|}{\|\delta A\|/\|A\|}\right\}$$

Does a big norm mean a matrix is ill-conditioned?

Consider

$$A = \begin{bmatrix} 10^8 & 0 \\ 0 & 10^8 \end{bmatrix}$$

Does a big norm mean a matrix is ill-conditioned?

Consider

$$A = \begin{bmatrix} 10^8 & 0 \\ 0 & 10^8 \end{bmatrix}$$

has norm 10^8, but *not* ill-conditioned (why?).
Gives same (big) result $\|Ax\|/\|x\|$ for all vectors x

Does a big norm mean a matrix is ill-conditioned?

Consider

$$A = \begin{bmatrix} 10^8 & 0 \\ 0 & 10^8 \end{bmatrix}$$

has norm 10^8, but *not* ill-conditioned (why?).
Gives same (big) result $\|Ax\|/\|x\|$ for all vectors x

In contrast consider:

$$A = \begin{bmatrix} 10^4 & 0 \\ 0 & 10^{-4} \end{bmatrix} \qquad x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad x' = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Does a big norm mean a matrix is ill-conditioned?

Consider

$$A = \begin{bmatrix} 10^8 & 0 \\ 0 & 10^8 \end{bmatrix}$$

has norm 10^8, but *not* ill-conditioned (why?).
Gives same (big) result $\|Ax\|/\|x\|$ for all vectors x

In contrast consider:

$$A = \begin{bmatrix} 10^4 & 0 \\ 0 & 10^{-4} \end{bmatrix} \qquad x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad x' = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Then $\|Ax\|/\|x\| = 10^4$ but $\|Ax'\|/\|x'\| = 10^{-4}$

Does a big norm mean a matrix is ill-conditioned?

Consider

$$A = \begin{bmatrix} 10^8 & 0 \\ 0 & 10^8 \end{bmatrix}$$

has norm 10^8, but *not* ill-conditioned (why?).
Gives same (big) result $\|Ax\|/\|x\|$ for all vectors x

In contrast consider:

$$A = \begin{bmatrix} 10^4 & 0 \\ 0 & 10^{-4} \end{bmatrix} \qquad x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad x' = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Then $\|Ax\|/\|x\| = 10^4$ but $\|Ax'\|/\|x'\| = 10^{-4}$

$\|A\| = \|A^{-1}\| = 10^4$ so Condition of A is 10^8

Application to Least Squares

- In least squares problems there are m equations in n variables, where $m > n$
- Gives a non-square ($m \times n$) matrix A, for which the condition *can* be calculated (see below)

Application to Least Squares

- In least squares problems there are m equations in n variables, where $m > n$
- Gives a non-square ($m \times n$) matrix A, for which the condition *can* be calculated (see below)
- Condition of $A^T A$ is the square of the condition of A, *for* ℓ_2*-norm*
- So unfortunately 4 norms get multiplied together, often giving a very big condition number

Application to Least Squares

- In least squares problems there are m equations in n variables, where $m > n$
- Gives a non-square ($m \times n$) matrix A, for which the condition *can* be calculated (see below)
- Condition of $A^T A$ is the square of the condition of A, *for ℓ_2-norm*
- So unfortunately 4 norms get multiplied together, often giving a very big condition number
- For the normal equation, we calculate the condition of $A^T A$ directly:

$$\mathrm{cond}(A^T A) = \|A^{-1}(A^T)^{-1}\| \|A^T A\|$$

Example

Consider:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 10^{-4} \end{bmatrix}$$

Normal equation matrix is

$$A^T A = \begin{bmatrix} 2 & 2 \\ 2 & 2 + 10^{-8} \end{bmatrix}$$

with inverse

$$(A^T A)^{-1} = 0.5 \times 10^8 \times \begin{bmatrix} 2 + 10^{-8} & -2 \\ -2 & 2 \end{bmatrix}$$

Example

Consider:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 10^{-4} \end{bmatrix}$$

Normal equation matrix is

$$A^T A = \begin{bmatrix} 2 & 2 \\ 2 & 2 + 10^{-8} \end{bmatrix}$$

with inverse

$$(A^T A)^{-1} = 0.5 \times 10^8 \times \begin{bmatrix} 2 + 10^{-8} & -2 \\ -2 & 2 \end{bmatrix}$$

So

$$\mathrm{cond}(A^T A) = (4 + 10^{-8})(4 + 10^{-8}) \times 0.5 \times 10^8 \simeq 8 \times 10^8$$

Why are the notes different?

- Notes say $\mathrm{cond}(A^T A) = 7.8 \times 10^8$
- They also say "condition of $A^T A$ is the square of the condition of A" but this is *not true* for ℓ_1- and ℓ_∞-norms (try it!)
- So a big difference is using the ℓ_2-norm.

Why are the notes different?

- Notes say $\text{cond}(A^T A) = 7.8 \times 10^8$
- They also say "condition of $A^T A$ is the square of the condition of A" but this is *not true* for ℓ_1- and ℓ_∞-norms (try it!)
- So a big difference is using the ℓ_2-norm.
- Singular value decomposition gives $\sigma_1(A) = \sigma_1(A^T) = 2.0$ and $\sigma_1(A^{-1}) = 14142.1$
- Thus, using these ℓ_2-norms,

$$\text{cond}_2(A) = 28284.3$$

- This agrees with the notes' 2.8×10^4 to 2 s.f. ... see page 78

Why are the notes different?

- Notes say $\mathrm{cond}(A^T A) = 7.8 \times 10^8$
- They also say "condition of $A^T A$ is the square of the condition of A" but this is *not true* for ℓ_1- and ℓ_∞-norms (try it!)
- So a big difference is using the ℓ_2-norm.
- Singular value decomposition gives $\sigma_1(A) = \sigma_1(A^T) = 2.0$ and $\sigma_1(A^{-1}) = 14142.1$
- Thus, using these ℓ_2-norms,

$$\mathrm{cond}_2(A) = 28284.3$$

- This agrees with the notes' 2.8×10^4 to 2 s.f. ... see page 78
- Squaring 2.8×10^4 does indeed give 7.8×10^8 to 2 s.f.
- However, squaring $\mathrm{cond}_2(A)$ gives 8×10^8 !!

Definition: A *metric space* is a non-empty set S of points (or objects) together with a function $d : S \times S \to \mathbb{R}$ (the metric of the space) satisfying:

1. $d(x, x) = 0$;
2. $d(x, y) > 0$ if $x \neq y$;
3. $d(x, y) = d(y, x)$
4. $d(x, y) \leq d(x, z) + d(z, y)$

for all points $x, y, z \in S$.

Metric Spaces

Definition: A *metric space* is a non-empty set S of points (or objects) together with a function $d : S \times S \rightarrow \mathbb{R}$ (the metric of the space) satisfying:

1. $d(x, x) = 0$;
2. $d(x, y) > 0$ if $x \neq y$;
3. $d(x, y) = d(y, x)$
4. $d(x, y) \leq d(x, z) + d(z, y)$

for all points $x, y, z \in S$.

- Obviously true for distances in real spaces such as $\mathbb{R}, \mathbb{R}^2, \mathbb{R}^3$;
- Actually all we need to define notions of limits and convergence.

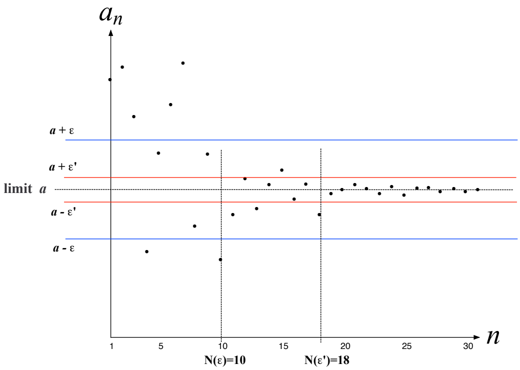
Convergence - revision

Definition: A sequence $a_1, a_2, \ldots$ converges to a limit $\ell \in \mathbb{R}$, written $a_n \to \ell$ as $n \to \infty$ or $\lim_{n\to\infty} a_n = \ell$, iff

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ such that } \forall n > N, |a_n - \ell| < \epsilon$$

- equivalently, $\ell - \epsilon < a_n < \ell + \epsilon$
- 'tramlines' ϵ away from the limit value ℓ

Illustration of convergence



Need a bigger N as ϵ decreases

Cauchy test in $\mathbb{R}$

Theorem (Cauchy): The sequence $a_1, a_2, \ldots$ is convergent if and only if $\forall \epsilon > 0, \exists N$ such that $|a_n - a_m| < \epsilon$ for all $n, m > N$.

Cauchy test in $\mathbb{R}$

Theorem (Cauchy): The sequence $a_1, a_2, \ldots$ is convergent if and only if $\forall \epsilon > 0, \exists N$ such that $|a_n - a_m| < \epsilon$ for all $n, m > N$.

- Useful because you don't need to know what the limit is (when it exists), e.g. when a_n is defined by a recurrence relation or a recursive function
- Also a test for divergence

Example

$$a_n = \sum_{i=1}^{n} \frac{1}{i(i+1)}$$

Example

$$a_n = \sum_{i=1}^{n} \frac{1}{i(i+1)}$$

$$a_n - a_m = \frac{1}{(m+1)(m+2)} + \ldots + \frac{1}{n(n+1)}$$

Example

$$a_n = \sum_{i=1}^{n} \frac{1}{i(i+1)}$$

$$\begin{aligned}
a_n - a_m &= \frac{1}{(m+1)(m+2)} + \ldots + \frac{1}{n(n+1)} \\
&= \left(\frac{1}{m+1} - \frac{1}{m+2}\right) + \left(\frac{1}{m+2} - \frac{1}{m+3}\right) + \\
&\quad \ldots + \left(\frac{1}{n} - \frac{1}{n+1}\right) \\
&= \frac{1}{m+1} - \frac{1}{n+1} \to 0 \quad \text{as } n > m \to \infty
\end{aligned}$$

Cauchy sequence

End of revision! More generally, in metric spaces ...

End of revision! More generally, in metric spaces ...

Definition: A sequence $x_1, x_2, \ldots$ in a metric space S converges to a limit $\ell \in S$ iff $\forall \epsilon > 0, \exists N \in \mathbb{N}$ such that $\forall n > N, d(x_n, \ell) < \epsilon$.

Cauchy sequence

End of revision! More generally, in metric spaces ...

Definition: A sequence $x_1, x_2, \ldots$ in a metric space S converges to a limit $\ell \in S$ iff $\forall \epsilon > 0, \exists N \in \mathbb{N}$ such that $\forall n > N, d(x_n, \ell) < \epsilon$.

Definition: A sequence $\{x_n\}$ in a metric space S is called a *Cauchy sequence* if for all $\epsilon > 0$ there exists an integer N such that for all $n, m \geq N, \quad d(x_n, x_m) < \epsilon$.

Cauchy sequence

End of revision! More generally, in metric spaces ...

Definition: A sequence $x_1, x_2, \ldots$ in a metric space S converges to a limit $\ell \in S$ iff $\forall \epsilon > 0, \exists N \in \mathbb{N}$ such that $\forall n > N, d(x_n, \ell) < \epsilon$.

Definition: A sequence $\{x_n\}$ in a metric space S is called a *Cauchy sequence* if for all $\epsilon > 0$ there exists an integer N such that for all $n, m \geq N, \quad d(x_n, x_m) < \epsilon$.

Definition: A metric space S in which every Cauchy sequence has a limit in S is called *complete*.

Cauchy sequence

End of revision! More generally, in metric spaces ...

Definition: A sequence $x_1, x_2, \ldots$ in a metric space S converges to a limit $\ell \in S$ iff $\forall \epsilon > 0, \exists N \in \mathbb{N}$ such that $\forall n > N, d(x_n, \ell) < \epsilon$.

Definition: A sequence $\{x_n\}$ in a metric space S is called a *Cauchy sequence* if for all $\epsilon > 0$ there exists an integer N such that for all $n, m \geq N, \quad d(x_n, x_m) < \epsilon$.

Definition: A metric space S in which every Cauchy sequence has a limit in S is called *complete*.

Theorem: Suppose a sequence $\{x_n\}$ converges in a metric space S. Then $\{x_n\}$ is Cauchy.

Application and examples

- These theorems mean that in any complete metric space a sequence is convergent if and only if it is Cauchy, which is relatively easy to test.
- It can be shown that $\mathbb{R}^k$ is complete for all $k \geq 1$.

Application and examples

- These theorems mean that in any complete metric space a sequence is convergent if and only if it is Cauchy, which is relatively easy to test.
- It can be shown that $\mathbb{R}^k$ is complete for all $k \geq 1$.
  Proof hint:
  - A cauchy sequence is bounded and so has a convergent sub-sequence (Fundamental axiom / Bolzano-Weierstrass theorem).
  - It therefore has the same limit as the subsequence

Application and examples

- These theorems mean that in any complete metric space a sequence is convergent if and only if it is Cauchy, which is relatively easy to test.
- It can be shown that $\mathbb{R}^k$ is complete for all $k \geq 1$.
  Proof hint:
  - A cauchy sequence is bounded and so has a convergent sub-sequence (Fundamental axiom / Bolzano-Weierstrass theorem).
  - It therefore has the same limit as the subsequence
- This is very useful computationally, for example in numerical iterations.
- Examples in $\mathbb{R}$ given in the background notes.
- Used to prove convergence of certain fixed point iterations later.

Uniqueness of limits

Theorem: A convergent sequence $\{x_n\}$ in $\mathbb{R}^k$ has a unique limit

Uniqueness of limits

Theorem: A convergent sequence $\{x_n\}$ in $\mathbb{R}^k$ has a unique limit

Proof: Suppose $x_n \to \ell_1$ and $x_n \to \ell_2$ as $n \to \infty$. Then

$$d(\ell_1, \ell_2) \leq d(\ell_1, x_n) + d(x_n, \ell_2)$$

for all n. But given any $\epsilon > 0$, for sufficiently large n, both terms on the RHS are less than $\epsilon/2$ so

$$d(\ell_1, \ell_2) < \epsilon$$

Hence $d(\ell_1, \ell_2) = 0$ since ϵ was arbitrary and so $\ell_1 = \ell_2$ by the definition of a metric.

Fixed point equations

Definition: Let $f : S \rightarrow S$ be a function from a metric space S to itself. A point p is S is called a *fixed point* (sometimes fixpoint) of f if $f(p) = p$. The function f is a *contraction* of S if there exists a real number α, $0 < \alpha < 1$, called a *contraction constant*, such that

$$d(f(x), f(y)) \leq \alpha d(x, y) \qquad \forall x, y \in S$$

Fixed point equations

Definition: Let $f : S \to S$ be a function from a metric space S to itself. A point p is S is called a *fixed point* (sometimes fixpoint) of f if $f(p) = p$. The function f is a *contraction* of S if there exists a real number α, $0 < \alpha < 1$, called a *contraction constant*, such that

$$d(f(x), f(y)) \le \alpha d(x, y) \qquad \forall x, y \in S$$

Theorem (Fixed point theorem): A continuous contraction f of a complete metric space S has a unique fixed point.

Proof of fixed point theorem

For any point $p \in S$, define the sequence $\{p_n\}$ by

$$p_0 = x, \quad p_{n+1} = f(p_n), \quad n = 0, 1, 2, \ldots$$

Proof of fixed point theorem

For any point $p \in S$, define the sequence $\{p_n\}$ by

$$p_0 = x, \quad p_{n+1} = f(p_n), \quad n = 0, 1, 2, \ldots$$

Then

$$d(p_{n+1}, p_n) = d(f(p_n), f(p_{n-1})) \leq \alpha d(p_n, p_{n-1}) \leq \ldots \leq c\alpha^n$$

where $c = d(p_1, p_0)$.

Proof of fixed point theorem

For any point $p \in S$, define the sequence $\{p_n\}$ by

$$p_0 = x, \quad p_{n+1} = f(p_n), \quad n = 0, 1, 2, \ldots$$

Then

$$d(p_{n+1}, p_n) = d(f(p_n), f(p_{n-1})) \leq \alpha d(p_n, p_{n-1}) \leq \ldots \leq c\alpha^n$$

where $c = d(p_1, p_0)$.

Hence, for $m > n$, by the triangle inequality,

$$d(p_m, p_n) \leq \sum_{k=n}^{m-1} d(p_{k+1}, p_k) \leq c \sum_{k=n}^{m-1} \alpha^k = c\frac{\alpha^n - \alpha^m}{1-\alpha} < \frac{c\alpha^n}{1-\alpha}$$

Since $\alpha < 1$, $d(p_m, p_n) \to 0$ as $m, n \to \infty$, and so $\{p_n\}$ is a Cauchy sequence.

Since S is complete, $p_n \to p$ for some point $p \in S$.

Since S is complete, $p_n \longrightarrow p$ for some point $p \in S$.

By continuity of f,

$$f(p) = f\left(\lim_{n\to\infty} p_n\right) = \lim_{n\to\infty} f(p_n) = \lim_{n\to\infty} p_{n+1} = p$$

Since S is complete, $p_n \to p$ for some point $p \in S$.

By continuity of f,

$$f(p) = f\left(\lim_{n\to\infty} p_n\right) = \lim_{n\to\infty} f(p_n) = \lim_{n\to\infty} p_{n+1} = p$$

Finally, if p and p' are both fixed points,

$$d(p, p') = d(f(p), f(p')) \leq \alpha d(p, p')$$

where $0 < \alpha < 1$, so $d(p, p') = 0$ and the fixed points are the same and hence unique.

Iterative solutions of linear equations

A splitting of a square matrix A is defined by a nonsingular matrix $M = A - N$.

Suppose we are solving the equation $Ax = b$. Then we may write $Mx = b - Nx$ so that $x = M^{-1}b - M^{-1}Nx$.

If $G = -M^{-1}N$ and $c = M^{-1}b$, we need to solve

$$x = Gx + c$$

and we can define the iteration

$$x^{(k+1)} = Gx^{(k)} + c$$

together with some starting value $x^{(0)}$

Convergence of the iteration

Theorem : For any matrix norm, if $\|G\| < 1$, then $x^{(k+1)} = Gx^{(k)} + c$ converges for any starting point $x^{(0)}$.

Proof : Given that x is the correct solution, let $y^{(k)} = x^{(k)} - x$ for $k = 0, 1, 2, \ldots$. Then $y^{(k+1)} = G(y^{(k)})$ and so

$$\|y^{(k+1)}\| \leq \|G\| \|y^{(k)}\| \leq \ldots \leq \|G\|^{k+1} \|y^{(0)}\| \to 0$$

as $k \to \infty$ since $\|G\| < 1$

Convergence conditions

Proposition : A sufficient condition for convergence is any of:

1. $\lim_{k\to\infty} G^k = 0$
2. $\lim_{k\to\infty} G^k\vec{x} = \vec{0} \quad \forall \vec{x} \in \mathbb{R}^m$
3. $\rho(G) < 1$

where $\rho(G) = \max_i |\lambda_i|$ is the largest of the absolute values of the eigenvalues of G, called the *spectral radius* of G.

Convergence conditions

Proposition : A sufficient condition for convergence is any of:

1. $\lim_{k\to\infty} G^k = 0$
2. $\lim_{k\to\infty} G^k\vec{x} = \vec{0} \quad \forall \vec{x} \in \mathbb{R}^m$
3. $\rho(G) < 1$

where $\rho(G) = \max_i |\lambda_i|$ is the largest of the absolute values of the eigenvalues of G, called the *spectral radius* of G.

Proof : Looking at the proof of the Theorem,

$$y^{(k)} = G(y^{(k-1)}) = \ldots = G^k y^{(0)}$$

Thus either of conditions 1 or 2 implies $y^{(k)} \to 0$ and so $x^{(k)} \to x$ as $k \to \infty$. For condition 3, diagonalising G (more generally, when there are multiple eigenvlues, putting G into Jordan Normal Form) we have

$$G^k = V^{-1}D^kV \to 0 \quad \text{as } k \to \infty$$

where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ and $|\lambda_i| < 1 \;\; \forall i$.

Choosing G for efficient solution

Given G, the rate of convergence is $r(G) = -\log_{10} \rho(G)$.

Choosing G for efficient solution

Given G, the rate of convergence is $r(G) = -\log_{10} \rho(G)$.

Why? ... look at previous proof.

Choosing G for efficient solution

Given G, the rate of convergence is $r(G) = -\log_{10} \rho(G)$.

Why? ... look at previous proof.

So we want to choose $G = -M^{-1}N$ so that both

1. $-M^{-1}Nx$ and $c = M^{-1}b$ are easy to compute
2. $\rho(-M^{-1}N)$ is small.

Choosing G for efficient solution

Given G, the rate of convergence is $r(G) = -\log_{10} \rho(G)$.

Why? ... look at previous proof.

So we want to choose $G = -M^{-1}N$ so that both

1. $-M^{-1}Nx$ and $c = M^{-1}b$ are easy to compute
2. $\rho(-M^{-1}N)$ is small.

For example:

- $M = I$ is good for 1.
- $M = A$ is good for 2. but probably not for 1.

Common splitting of A

Assuming A has no zeros on the diagonal (can often relabel the variables x_i to permute the rows and columns, if necessary, to avoid this),

$$A = D - \tilde{L} - \tilde{U} = D(I - L - U)$$

where D is the diagonal of A, $-\tilde{L}, \tilde{U}$ are the strict lower and upper triangular parts of A, $L = D^{-1}\tilde{L}$ and $U = D^{-1}\tilde{U}$

Common splitting of A

Assuming A has no zeros on the diagonal (can often relabel the variables x_i to permute the rows and columns, if necessary, to avoid this),

$$A = D - \tilde{L} - \tilde{U} = D(I - L - U)$$

where D is the diagonal of A, $-\tilde{L}, \tilde{U}$ are the strict lower and upper triangular parts of A, $L = D^{-1}\tilde{L}$ and $U = D^{-1}\tilde{U}$

Example :

$$A = \begin{bmatrix} 2 & -4 & 2 \\ -3 & 1 & -5 \\ 6 & -2 & 2 \end{bmatrix}$$

Jacobi method

Using this splitting gives $N = \tilde{L} + \tilde{U}$ and the fixed point equation

$$x = D^{-1}b - D^{-1}Nx$$

with $G = -D^{-1}N = I - D^{-1}A = L + U, \;\; c = D^{-1}b.$

Jacobi method

Using this splitting gives $N = \tilde{L} + \tilde{U}$ and the fixed point equation

$$x = D^{-1}b - D^{-1}Nx$$

with $G = -D^{-1}N = I - D^{-1}A = L + U, \;\; c = D^{-1}b.$

The iteration $x^{(k+1)} = Gx^{(k)} + c$ now yields

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right)$$

Jacobi method

Using this splitting gives $N = \tilde{L} + \tilde{U}$ and the fixed point equation

$$x = D^{-1}b - D^{-1}Nx$$

with $G = -D^{-1}N = I - D^{-1}A = L + U, \;\; c = D^{-1}b.$

The iteration $x^{(k+1)} = Gx^{(k)} + c$ now yields

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right)$$

- Update of x_i requires access to only row i of A

Jacobi method

Using this splitting gives $N = \tilde{L} + \tilde{U}$ and the fixed point equation

$$x = D^{-1}b - D^{-1}Nx$$

with $G = -D^{-1}N = I - D^{-1}A = L + U, \;\; c = D^{-1}b.$

The iteration $x^{(k+1)} = Gx^{(k)} + c$ now yields

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right)$$

- Update of x_i requires access to only row i of A
- Good for parallel computation (see Course 429)

Gauss-Seidel method

This is intended to give faster convergence by using more up-to-date data:

Gauss-Seidel method

This is intended to give faster convergence by using more up-to-date data:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{m} a_{ij} x_j^{(k)} \right)$$

Gauss-Seidel method

This is intended to give faster convergence by using more up-to-date data:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{m} a_{ij} x_j^{(k)} \right)$$

In matrix form, $(D - \tilde{L})x^{(k+1)} = \tilde{U}x^{(k)} + b$, i.e.

$$x^{(k+1)} = (I - L)^{-1} U x^{(k)} + (I - L)^{-1} D^{-1} b$$

Gauss-Seidel method

This is intended to give faster convergence by using more up-to-date data:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{m} a_{ij} x_j^{(k)} \right)$$

In matrix form, $(D - \tilde{L})x^{(k+1)} = \tilde{U}x^{(k)} + b$, i.e.

$$x^{(k+1)} = (I - L)^{-1} U x^{(k)} + (I - L)^{-1} D^{-1} b$$

So now $G = (I - L)^{-1} U$ as opposed to $L + U$ for Jacobi.

Successive over-relaxation, SOR(ω)

A generalisation of Gauss-Seidel by "smoothing":

Successive over-relaxation, SOR(ω)

A generalisation of Gauss-Seidel by "smoothing":

$$x_i^{(k+1)} := (1-\omega)x_i^{(k)} + \omega x_i^{(k+1)}$$

where $x_i^{(k+1)}$ on the RHS is computed by Gauss-Seidel.
ω is the relaxation parameter, $0 < \omega < 2$ (see later).

Successive over-relaxation, SOR(ω)

A generalisation of Gauss-Seidel by "smoothing":

$$x_i^{(k+1)} := (1-\omega)x_i^{(k)} + \omega x_i^{(k+1)}$$

where $x_i^{(k+1)}$ on the RHS is computed by Gauss-Seidel.
ω is the relaxation parameter, $0 < \omega < 2$ (see later). Thus

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{m} a_{ij}x_j^{(k)}\right)$$

Successive over-relaxation, SOR(ω)

A generalisation of Gauss-Seidel by "smoothing":

$$x_i^{(k+1)} := (1-\omega)x_i^{(k)} + \omega x_i^{(k+1)}$$

where $x_i^{(k+1)}$ on the RHS is computed by Gauss-Seidel.
ω is the relaxation parameter, $0 < \omega < 2$ (see later). Thus

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{m} a_{ij}x_j^{(k)}\right)$$

In matrix form, $(D-\omega\tilde{L})x^{(k+1)} = ((1-\omega)D + \omega\tilde{U})x^{(k)} + \omega b$, i.e.

$$x^{(k+1)} = (I-\omega L)^{-1}((1-\omega)I + \omega U)x^{(k)} + \omega(I-\omega L)^{-1}D^{-1}b$$

Successive over-relaxation, SOR(ω)

A generalisation of Gauss-Seidel by "smoothing":

$$x_i^{(k+1)} := (1-\omega)x_i^{(k)} + \omega x_i^{(k+1)}$$

where $x_i^{(k+1)}$ on the RHS is computed by Gauss-Seidel.
ω is the relaxation parameter, $0 < \omega < 2$ (see later). Thus

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{m} a_{ij}x_j^{(k)}\right)$$

In matrix form, $(D-\omega\tilde{L})x^{(k+1)} = ((1-\omega)D + \omega\tilde{U})x^{(k)} + \omega b$, i.e.

$$x^{(k+1)} = (I-\omega L)^{-1}((1-\omega)I + \omega U)x^{(k)} + \omega(I-\omega L)^{-1}D^{-1}b$$

Now $G = (I-\omega L)^{-1}((1-\omega)I + \omega U)$, so $\omega = 1$ gives Gauss-Seidel.

Convergence conditions

Definitions: A square matrix A is *strictly row diagonally dominant* if $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$.

Convergence conditions

Definitions: A square matrix A is *strictly row diagonally dominant* if $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$.

Theorem: A *sufficient* condition for both Jacobi and Gauss-Seidel to converge is that A is strictly row diagonally dominant. G-S is faster.

Sketch of the proof

1. Using the subordinate ℓ_∞-norm, we show $\|G\| < 1$ where $G = -D^{-1}N$.

Sketch of the proof

1. Using the subordinate ℓ_∞-norm, we show $\|G\| < 1$ where $G = -D^{-1}N$.
2. Strict diagonal dominance implies all absolute row sums of G are less than one.

Sketch of the proof

1. Using the subordinate ℓ_∞-norm, we show $\|G\| < 1$ where $G = -D^{-1}N$.

2. Strict diagonal dominance implies all absolute row sums of G are less than one.

3. Hence the *maximum* absolute row sum is less than one, and so therefore is the ℓ_∞-norm.

Sketch of the proof

1. Using the subordinate ℓ_∞-norm, we show $\|G\| < 1$ where $G = -D^{-1}N$.

2. Strict diagonal dominance implies all absolute row sums of G are less than one.

3. Hence the *maximum* absolute row sum is less than one, and so therefore is the ℓ_∞-norm.

4. Any norm less than one leads to convergence, by the theorem above.

Other results

1. If A is *weakly* row diagonally dominant *and* irreducible, both Jacobi and G-S still converge, and G-S is faster.

Other results

1. If A is *weakly* row diagonally dominant *and* irreducible, both Jacobi and G-S still converge, and G-S is faster.

   A square matrix A is *irreducible* if, by symmetric permutation of rows and columns (i.e. variable renaming in a system of linear equations) it cannot take the form:

$$\begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix}$$

   where the block matrices A_{11} and A_{22} are square.

Other results

1. If A is *weakly* row diagonally dominant *and* irreducible, both Jacobi and G-S still converge, and G-S is faster.

   A square matrix A is *irreducible* if, by symmetric permutation of rows and columns (i.e. variable renaming in a system of linear equations) it cannot take the form:

   $$\begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix}$$

   where the block matrices A_{11} and A_{22} are square.

2. A necessary condition for SOR(ω) to converge is that $0 < \omega < 2$. This condition is also sufficient if A is positive definite (see later for definition).