Mathematical Methods

Course Overview

Carles Batlle Arnau
(carles.batlle@upc.edu)

Departament de Matemàtica Aplicada 4
and
Institut d'Organització i Control de Sistemes Industrials

Universitat Politècnica de Catalunya

Course goals

- To present tools from advanced linear algebra that are used in a variety of control problems (over- and underconstrained systems, QR and SVD matrix decompositions).
- To present basic ideas of partial differential equations: modeling origins, classification, analytical and numerical tools.

Outline

1. Linear algebra review.
2. QR and least squares estimation.
3. Least squares applications.
4. SVD factorization and applications.
5. Partial differential equations.
6. First order PDE. The method of characteristics.
7. Second-order PDE in two variables. Separation of variables for the heat and wave equations.
8. Elliptic equations. Separation of variables for the Laplace equation.
9. Variational methods.
10. Numerical methods.

References

Course slides will be posted on the intranet before each session. They are based on the following references:

MW A. Megretski and J. Wyatt, Linear Algebra and Functional Analysis for Signals and Systems, Lecture notes for MIT graduate course 6.972. Available at `http://web.mit.edu/6.241/www/b6972.pdf`

BL S. Boyd and S. Lall, Introduction to Linear Dynamical Systems, Stanford Course EE263. Available at `http://www.stanford.edu/class/ee263/`

PR Y. Pinchover and J. Rubinstein, An introduction to partial differential equations, Cambridge University Press, Cambridge, UK (2005).

Course grading

- Grading is entirely based on homework.
- At the end of each session several short exercises (typically 3 or 4) will be proposed. They must be turned in either electronically or by hand before the next session. Electronic submissions must be in the form of a PDF file, either produced by any appropriate software (LaTeX, Word) or scanned from your handwriting, and made preferably through the intranet, although e-mail attachments will also be accepted.
- Late submissions will be penalized.
- You can discuss in group, but I expect that you will independently write up the actual solutions that you turn in.
- To compute the final average, you can discard the lowest grade.

Mathematical Methods – Lecture 1

Linear Algebra Review

Carles Batlle Arnau

Departament de Matemàtica Aplicada 4
and
Institut d'Organització i Control de Sistemes Industrials

Universitat Politècnica de Catalunya

Lecture goals

- To review the basic definitions and results of elementary linear algebra.
- To introduce normed space vectors and inner products, and to present a first version of the orthogonality principle.
- To introduce the abbreviated notation for matrix operations (Einstein's sum convention).
- To introduce the underconstrained and overconstrained problems.

Outline

- Vector spaces. Examples.
- Subspaces.
- Linear independence and basis.
- Normed vector spaces.
- Inner product. Cauchy-Schwartz inequality.
- The projection theorem.
- Matrices. Operations and notation. Right and left nullvectors.
- Determinants. Basic properties.
- Linear maps. Range and nullspace.
- Matrix associated to a map. Change of basis.
- Systems of equations. Overconstrained and underconstrained systems.
- Eigenvectors and eigenvalues.

References

MW Parts of chapters 1, 2 and 3.

DDV1 M. Dahleh, M.A. Dahleh and G. Verghese, Lectures on Dynamic Systems and Control, Chapter 1, MIT Course 6.241 (2003). (available in Atenea)

Strang G. Strang, *Algebra lineal y sus aplicaciones*, Addison-Wesley Iberoamericana, 1986.

Vector spaces

A vector space over a field K is a set E endowed with an internal operation $+$ such that $(E, +)$ is a commutative group, and with and external operation (multiplication by an element of K, or scalar), such that the following *compatibility conditions* hold:

V1 $1\ x = x$,

V2 $(c_1 c_2)x = c_1(c_2 x)$,

V3 $c(x + y) = cx + cy$,

V4 $(c_1 + c_2)x = c_1 x + c_2 x$.

Here 1 is the unity of the field K, $c, c_1, c_2 \in K$, $x, y \in E$, and notice that we are using the same notation for different operations (those of the field, of the $(E, +)$ group and the mixed ones).
Elements of E are called vectors.

Examples

- $\mathbb{R}^n$ is a vector space over $\mathbb{R}$.
- $\mathbb{C}^n$ is a vector space both over $\mathbb{R}$ and $\mathbb{C}$.
- The set of real continuous functions on the real line is a vector space over $\mathbb{R}$.
- The set of $m \times n$ matrices with real coefficients is a vector space over $\mathbb{R}$.
- The set of solutions of $y'' + 3y' = 0$ is a vector space over $\mathbb{R}$.
- The set of solutions of $y'' + 3 \sin x \; y' = 0$ is a vector space over $\mathbb{R}$.
- The set of solutions of $y'' + 3y' = 3$ is not a vector space.
- The set of solutions of $y' + y^2 = 0$ is not a vector space.

Examples

- The set of points $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ satisfying $x_1^2 + x_2^2 + x_3^2 = 1$ is not a vector space.
- Consider the set $X = [0, 1)$ and let $\{h\}$ denote difference between h and the largest integer not greater than h, so that $\{h\} \in X$ for any $h \in \mathbb{R}$. On X we define the operations

$$x_1 \oplus x_2 = \{x_1 + x_2\}, \quad c * x_1 = \{cx_1\},$$

for any $x_1, x_2 \in X$, $c \in \mathbb{R}$. This does not provide a vector space structure for X. For instance, **V3** is not satisfied. Just take $c = x_1 = x_2 = 0.5$. Then

$$0.5 * (0.5 \oplus 0.5) = 0.5 * \{1\} = 0.5 * 0 = \{0.5 \cdot 0\} = \{0\} = 0$$

but

$$\begin{aligned} (0.5 * 0.5) \oplus (0.5 * 0.5) &= \{0.5 \cdot 0.5\} \oplus \{0.5 \cdot 0.5\} = \{0.25\} \oplus \{0.25\} = 0.25 \oplus 0.25 \\ &= \{0.25 + 0.25\} = \{0.5\} = 0.5 \end{aligned}$$

Subspaces

A subset S of a vector space E over K is a **linear subspace** if

$$c_1 x + c_2 y \in S \quad \text{for any } c_1, c_2 \in K \text{ and any } x, y \in S.$$

Examples

- The set of solutions to $y''' = 0$ such that $y(0) = 0$ is a subspace.
- The set of all linear combinations of a given set of vectors forms a subspace, called the subspace *generated* by these vectors, or also their *span*.
- The intersection of two subspaces is again a subspace, but their union is not.
- The *direct sum* of two subspaces, formed by the vectors that can be written as the sum of two vectors drawn from each subspace, is again a subspace.

Linear independence. Basis

- A set (finite or infinite) of vectors $\{v_i\}_{i\in I}$ is called *linearly independent* if for any finite linear combination set to zero

$$\sum_{k\in J\subset I} c_k v_k = 0, \quad \text{with } J \text{ a finite subset of indices, } c_k \in K,$$

  there exists only the trivial solution $c_k = 0$ for all k. Otherwise the set is called *linearly dependent*.
- A basis of E is a linearly independent set such that its span is E. Similarly, one defines a basis of a subspace.
- Given a vector space (or subspace) all its basis have the same number of elements, called the *dimension* of the vector space (or subspace).
- If a space has a set of n independent vectors for any n, then the space is called *infinite dimensional*.

Normed spaces

Given a vector space E over $K = \mathbb{R}, \mathbb{C}$, a *norm* is a map

$$||\cdot|| : E \longrightarrow \mathbb{R}^+ \cup \{0\}$$

satisfying

N1 $||x|| = 0$ iff $x = 0$.

N2 $||cx|| = |c|\ ||x||$, for any $c \in K$ and any $x \in E$.

N3 (triangle inequality) $||x + y|| \leq ||x|| + ||y||$, for any $x, y \in E$.

Here $|c|$ denotes either the absolute value (if $K = \mathbb{R}$) or the modulus (if $K = \mathbb{C}$) of c.

A vector space endowed with a norm is a normed space.

Examples

- $\mathbb{R}^n$ with the usual Euclidean norm $||x|| = \sqrt{x'x}$, with x' denoting the transpose of x, is a normed space.
- A complex matrix Q is called Hermitian if $Q^\dagger = Q$, where $Q^\dagger$ is the transpose and complex conjugate of Q; if Q is real this condition boils down to Q being symmetric. A matrix is positive definite if $x^\dagger Q x > 0$ (this implies that $x^\dagger Q x$ must be real) for $x \neq 0$.
- $\mathbb{C}^n$ with $||x|| = \sqrt{x^\dagger Q x}$ is a normed space for Q Hermitian and positive definite.
- $\mathbb{R}^n$ is a normed space with either

$$||x||_1 = \sum_{i=1}^{n} |x_i| \quad \text{or} \quad ||x||_\infty = \max_i |x_i|.$$

Normed functional spaces

Let us turn now to functional vector spaces. One can consider

- 1-norm
$$||u||_1 = \int_{-\infty}^{+\infty} |u(t)| \, \mathrm{d}t.$$
- 2-norm
$$||u||_2 = \left(\int_{-\infty}^{+\infty} u^2(t) \, \mathrm{d}t \right)^{1/2}.$$
- ∞-norm
$$||u||_\infty = \sup_{t \in \mathbb{R}} |u(t)|.$$

These define norms in $PC(\mathbb{R}) \bigcap L_1(\mathbb{R})$, $PC(\mathbb{R}) \bigcap L_2(\mathbb{R})$ and $PC(\mathbb{R}) \bigcap L_\infty(\mathbb{R})$, respectively.

Examples

$PC(\mathbb{R})$ denotes the set of piecewise continuous functions in $\mathbb{R}$, $L_1(\mathbb{R})$ is the set of absolutely integrable functions in $\mathbb{R}$, $L_2(\mathbb{R})$ is the set of square integrable functions in $\mathbb{R}$, and $L_\infty(\mathbb{R})$ is the set of bounded functions in $\mathbb{R}$.

These restrictions must be imposed for the norm of a function to be a real (finite!) number. In all cases, $\mathbb{R}$ can be replaced by appropriate subsets.

- For $u(t) = \theta(t)$ (step function) we have $u \notin L_1(\mathbb{R})$, $u \notin L_2(\mathbb{R})$, but $u \in L_\infty(\mathbb{R})$ and $||u||_\infty = 1$.
- For $u(t) = (1 - e^{-t})\theta(t)$ we have $||u||_\infty = 1$.
- For $u(t) = \frac{1}{\sqrt{t}}\theta(1-t)\theta(t)$, $||u||_1 = 2$ but $u \notin L_2(\mathbb{R})$, $u \notin L_\infty(\mathbb{R})$ (actually, this is not a $PC(\mathbb{R})$ function).

Inner product

A vector space can be provided with further structure, the *inner product*, yielding an inner product space, or Euclidean space. For $K = \mathbb{R}$ or $K = \mathbb{C}$, an inner product is a map

$$\langle \cdot, \cdot \rangle : E \times E \longrightarrow K$$

satisfying, for any $x, y, z \in E$ and for any $a, b \in K$,

E1 $\langle x, x \rangle > 0$ for $x \neq 0$.

E2 $\langle x, y \rangle = (\langle y, x \rangle)^*$, where * denotes the complex conjugate.

E3 $\langle x, ay + bz \rangle = a\langle x, y \rangle + b\langle x, z \rangle$, and hence $\langle ay + bz, x \rangle = a^* \langle y, x \rangle + b^* \langle z, x \rangle$.

Given an Euclidean space, one obtains a normed space by means of the associated norm

$$||x|| = \sqrt{\langle x, x \rangle}.$$

Examples

- In $\mathbb{C}^n$, $\langle x, y\rangle = x^\dagger Q y$ defines an inner product if Q is Hermitian and positive definite.
- For continuous (or just integrable) real functions in $[0, 1]$

$$\langle u, v\rangle = \int_0^1 u(t)v(t)\ \mathrm{d}t$$

  defines an inner product.
- For complex functions of a real variable in $[a, b]$

$$\langle u, v\rangle = \int_a^b u^*(t)v(t)\ \mathrm{d}t$$

  is also an inner product. This is the inner product of quantum mechanics.

The Cauchy-Schwartz inequality

Given a inner product space with its associated (or induced) norm, one has the

Cauchy-Schwartz inequality

$$|\langle x, y\rangle| \leq ||x||\ ||y||,$$

with equality holding only if $x = \alpha y$ for some scalar α.

- Two vectors x, y are said to be orthogonal if $\langle x, y\rangle = 0$.
- Two sets X and Y are called orthogonal if every vector of X is orthogonal to every vector of Y.
- The orthogonal complement of X is the set of vectors orthogonal to X, and is denoted by $X^{\perp}$.
- The orthogonal complement of any set is a subspace.

The projection theorem

Let M be a subspace in an Euclidean space E, and let y be a given element in E. Consider the problem of minimizing the distance of M to y, that is

$$\min_{m \in M} ||y - m||,$$

where the norm is the one induced by the inner product.

Projection theorem

The optimal solution $\hat{m}$ to the above minimizing problem satisfies

$$(y - \hat{m}) \perp M.$$

This result has an obvious geometric interpretation in low dimension spaces.

Operations and notation

- We denote the elements of an $m \times n$ matrix A by A_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$.
- The product of two matrices $A_{m \times n}$ and $B_{n \times p}$ is given by

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, p.$$

- Einstein's summation convention gets rid of the summation sign and abbreviates the above to

$$(AB)_{ij} = A_{ik} B_{kj},$$

  i.e. it is understood that repeated indices are summed over the appropriate range.
- In particular, the elements of the vector resulting from the action of a matrix A on a vector v are given by $(Av)_i = A_{ij} v_j$.

Operations and notation (cont'd)

- The trace of a square matrix $A_{n\times n}$ is defined as

$$\text{Tr } A = \sum_{i=1}^{n} A_{ii} \quad \text{or, in Einstein's notation,} \quad \text{Tr } A = A_{ii}.$$

- Notice that

$$\text{Tr } (AB) = (AB)_{ii} = A_{ij}B_{ji} = B_{ji}A_{ij} = (BA)_{jj} = \text{Tr } (BA).$$

- Other examples of notation:
  - $v'Au = v_i A_{ij} u_j$.
  - $(A')_{ij} = A_{ji}$.
  - $(A'B)_{ij} = (A')_{ik}B_{kj} = A_{ki}B_{kj}$.

Operations and notation (cont'd)

- The exponential of a square matrix $A_{n\times n}$ is again a $n \times n$ matrix e^A defined as

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k.$$

- Notice that, in general

$$e^A e^B \neq e^{A+B} \neq e^B e^A,$$

the equality being true only if the commutator of A and B,

$$[A, B] = AB - BA,$$

vanishes, $[A, B] = 0$, which means that the matrices commute.

- In general one has the famous Baker-Campbell-Hausdorff formula

$$e^A e^B = e^{A+B+\frac{1}{2}[A,B]+\frac{1}{12}[A,[A,B]]-\frac{1}{12}[B,[A,B]]+\cdots}$$

Right and left nullvectors

- A right nullvector, or simply nullvector, of a matrix A is a vector u satisfying
$$Au = 0.$$
- A left nullvector, of a matrix A is a vector v satisfying
$$v'A = 0.$$
- The matrix A need not be square.
- $Au = 0$ indicates that the column vectors of A are dependent, with the coefficients of u providing the linear combination.
- Similarly, $v'A = 0$ means that the row vectors of A are dependent.

Determinants

The determinant of a square matrix $A_{n\times n}$ is the real number given by

$$\det A = \sum_{\sigma\in S_n} \epsilon(\sigma) A_{1\sigma(1)} A_{2\sigma(2)} \cdots A_{n\sigma(n)}$$

where the sum is over the permutations of the symmetric group S_n (which has $n!$ elements), and $\epsilon(\sigma) = \pm 1$ is the parity of the permutation.

For instance

$$\begin{aligned} \det \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} &= \sum_{\sigma\in S_2} \epsilon(\sigma) A_{1\sigma(1)} A_{2\sigma(2)} \\ &= (+1)A_{11}A_{22} + (-1)A_{12}A_{21} = A_{11}A_{22} - A_{12}A_{21}. \end{aligned}$$

Determinants (cont'd)

- The column vectors or the row vectors of A are independent iff $\det A \neq 0$.
- The value of the determinant does not change if to any column (row) we add a linear combination of the remaining columns (rows).
- $\det(AB) = \det A \det B = \det(BA)$.
- $\det A' = \det A$.
- A matrix A has an inverse A^{-1} such that $AA^{-1} = A^{-1}A = \mathbb{I}$ iff $\det A \neq 0$.
- If A^{-1} exists, then $\det A^{-1} = (\det A)^{-1}$.
- $\det e^A = e^{\mathrm{Tr}\ A}$, or

$$\log \det e^A = \mathrm{Tr}\ A.$$

Linear maps. Range and nullspace

Given vector spaces E and F over the same field K, a map $f : E \longrightarrow F$ is called linear if

$$f(ax + by) = af(x) + bf(y), \quad \forall\ x, y \in E,\ \forall\ a, b \in K.$$

- The range, or image, of f, denoted by Im f, is the subspace of F spanned by the images of all the elements of E.
- The nullspace, or *kernel*, of f, denoted by Ker f, is the subspace of E spanned by all the elements $x \in E$ such that $f(x) = 0$.
- A fundamental result in linear algebra is that

$$\dim \text{ Im } f + \dim \text{ Ker } f = \dim\ E.$$

Matrix associated to a map

Given basis $\{u_i\}_{i=1,\dots,m}$ of E and $\{v_i\}_{i=1,\dots,n}$ of F, a linear map $f : E \longrightarrow F$ can be completely specified by giving the images of the vectors of the basis of E:

$$f(u_i) = \sum_{j=1}^{n} a_{ij} v_j, \quad i = 1, \dots, m.$$

The image of any $x = \sum_{i=1}^{m} x_i u_i$ of E can then be computed as

$$f(x) = f(\sum_{i=1}^{m} x_i u_i) = \sum_{i=1}^{m} x_i f(u_i) = \sum_{i=1}^{m} x_i \sum_{j=1}^{n} a_{ij} v_j = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} a_{ij} x_i \right) v_j.$$

This means that the components of the image of x are given by

$$y_j = \sum_{i=1}^{m} a_{ij} x_i \overset{\text{Einstein's notation}}{=} A_{ji} x_i, \quad \text{where } A_{ji} = a_{ij}.$$

Matrix associated to a map (cont'd)

- Hence, in the given basis,

$$y = Ax.$$

- The matrix A is the matrix associated to the map in the given basis.
- The matrix A changes if there is a change of basis either in E or F, or in both.
- More specifically, if in some basis $y = Ax$ and we perform a change of basis both in E and F so that the new components are given by $\tilde{x} = Mx$ and $\tilde{y} = Ny$, then the matrix of the linear map in those new basis is

$$\tilde{A} = NAM^{-1}.$$

Matrix associated to a map (cont'd)

- The column vectors of A are the images of the basis vectors of E, expressed in the given basis of F.
- The number of independent column vectors of A is the dimension of Im f, and is called the rank of A.
- Imf is denoted as $\mathcal{R}(A)$, and is the subspace spanned by the column vectors of A.
- Ker$f$ is denoted as $\mathcal{N}(A)$, and is the subspace spanned by the (right)nullvectors of A.

Basic results about linear systems

Consider a linear system with m equations and n unknowns:

$$A_{m\times n}x_{n\times 1} = y_{m\times 1}.$$

- $Ax = y$ has at least a solution (is compatible) iff $y \in \mathcal{R}(A)$. Hence $Ax = y$ has a solution iff rank $([A|y])$ = rank (A).
- If x is a solution and $\mathcal{N}(A) \neq \{0\}$, then $x + x_0$, where $x_0 \in \mathcal{N}(A)$, is also a solution. Hence, a compatible system has a unique solution iff $\mathcal{N}(A) = \{0\}$.

In particular,

- if $m = n$ and $\det A \neq 0$, there exists a unique solution $x = A^{-1}y$.
- if $m = n$ and $y = 0$, the only solution is the trivial one, $x = 0$, *unless* $\det A = 0$.

Overconstrained and underconstrained problems

In system and control theory, two common situations arise:

- If $m > n$, *i.e.* there are more equations than unknowns, the system may be *overconstrained*. In fact, in many cases y will not lie in the range of A, and hence the system will be inconsistent. This is the situation encountered in estimation or identification problems, where x is a parameter vector of low dimension compared to the measurements y available. One then looks for an x that comes closest to achieving $Ax = y$, according to some error criterion.
- If $m < n$, *i.e.* there are fewer equations than unknowns, the system is *underconstrained*. In this case $\mathcal{N}(A)$ is guaranteed to be nontrivial (why?) and, if the system has a solution, then it has infinitely many. This is the situation that occurs in many control problems, where the control objectives do not uniquely determine the control. One then typically searches among the available solutions the ones that are optimal according to some performance criteria.

Eigenvectors and eigenvalues

Consider a vector space E and a linear map from E to E, *i.e.* a linear **endomorphism**

$$f : E \longrightarrow E.$$

We say that $x \in E$, $x \neq 0$, is an **eigenvector** of f if there exists $\lambda \in \mathbb{R}$, called the associated **eigenvalue**, such that

$$f(x) = \lambda x.$$

In particular, λ may be zero, and in this case the eigenvector belongs to $\mathrm{Ker} f$.

Eigenvectors and eigenvalues (cont'd)

If A is the matrix associated to f for a given basis of E, we have

$$Ax = \lambda x \quad \text{or} \quad (A - \lambda \mathbb{I})x = 0.$$

from this it follows that x is a (right)nullvector of $A - \lambda \mathbb{I}$. From the results about solutions of linear systems, it follows that the necessary and sufficient condition for $x \neq 0$ to exist is that

$$\det(A - \lambda \mathbb{I}) = 0.$$

This is called the **characteristic equation** of the linear map, and if $\dim E = n$, it is a polynomial of degree n in λ; furthermore, it is independent of the basis used for E.

Exercises

1. Prove that

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

defines a norm in $\mathbb{R}^n$.

2. Consider $M_2(\mathbb{R})$, the set of 2×2 matrices with real coefficients. With the standard matrix operations, this is a vector space over $\mathbb{R}$, with dimension 4. Let $S \subset M_2(\mathbb{R})$ denote the subset of symmetric matrices.
   - Prove that S is a subspace, and find out its dimension.
   - Write down a basis for S.
   - Repeat with A the subset of skew-symmetric matrices.
   - Generalize all the above results to $M_n(\mathbb{R})$.
   - Prove that any square matrix can be uniquely written as the sum of a symmetric matrix and a skew-symmetric one.

3. Prove that
   - the kernel and the image of a linear map are subspaces.
   - the characteristic equation of an endomorphism does not depend on the basis used for the vector space.

Mathematical Methods – Lecture 2

QR decomposition

Carles Batlle Arnau

Departament de Matemàtica Aplicada 4
and
Institut d'Organització i Control de Sistemes Industrials

Universitat Politècnica de Catalunya

Lecture goals

- To define orthonormal sets of vectors and basis, and the associated orthogonal transformations and their properties.
- To present the Gram-Schmidt procedure and the QR and full QR factorizations.

Outline

- Orthonormal sets of vectors. Geometric properties.
- Orthogonal basis and transformations.
- The Gram-Schmidt procedure.
- The QR decomposition.
- General Gram-Schmidt procedure.
- Full QR factorization.
- Linear algebra applications of QR.

References

BL4 S. Boyd and S. Lall, lecture 4 of Introduction to Linear Dynamical Systems, Stanford Course EE263. Available at `http://www.stanford.edu/class/ee263/`

Orthonormal sets (I)

- Let E be an Euclidean space, *i.e.* a vector space with an inner product $\langle \cdot, \cdot \rangle$ and associated Euclidean norm $||x|| = \langle x, x \rangle^{1/2}$.
- A set of vectors $\{u_1, u_2, \ldots, u_k\} \subset E$ is
  - *normalized* if $||u_i|| = 1$, $i = 1, 2, \ldots, k$.
  - *orthogonal* if $u_i \perp u_j$, that is, $\langle u_i, u_j \rangle = 0 \ \forall \ i \neq j = 1, \ldots, n$.
  - *orthonormal* if both, that is $\langle u_i, u_j \rangle = \delta_{i,j} \ \forall \ i, j = 1, \ldots, n$.
- If E is finite dimensional, say $\dim E = n$, $\{u_1, u_2, \ldots, u_k\}$, $k \leq n$, is an orthonormal set and U is the $n \times k$ matrix whose ith column is made of the components of u_i, then

$$U^T U = \mathbb{I}_{k \times k},$$

but notice that $UU^T \neq \mathbb{I}_{n \times n}$ if $k < n$.

Orthonormal sets (II)

- Orthonormal vectors are independent:

$$\sum_{i=1}^{k} \alpha_i u_i = 0 \;\Rightarrow \sum_{i=1}^{k} \alpha_i \langle u_j, u_i \rangle = 0 \;\Rightarrow \sum_{i=1}^{k} \alpha_i \delta_{i,j} = 0$$

  and hence $\alpha_j = 0,\ \forall j = 1, \ldots k.$

  In fact this is also true for orthogonal vectors, provided that none of them is zero.

- Hence, an orthonormal set is a basis for its span, *i.e* for the range of the matrix U

$$\text{span}(u_1, u_2, \ldots, u_k) = \mathcal{R}(U).$$

Geometric properties

- Let $E = \mathbb{R}^n$, so that the inner product is just $\langle x, y \rangle = x^T y$.
- Let the columns of $U = [u_1 \; u_2 \; \cdots \; u_k]$ be orthonormal.
- Let $w = Uz$. The action of U does not change norms:

$$\begin{aligned} ||w||^2 &= ||Uz||^2 = \langle Uz, Uz \rangle = (Uz)^T (Uz) = z^T U^T U z \\ &= z^T z = \langle z, z \rangle = ||z||^2. \end{aligned}$$

- It also preserves inner products. If $w = Uz$ and $\tilde{w} = U\tilde{z}$,

$$\langle \tilde{w}, w \rangle = \langle U\tilde{z}, Uz \rangle = (U\tilde{z})^T (Uz) = \tilde{z}^T U^T U z = \tilde{z}^T z = \langle \tilde{z}, z \rangle.$$

- Hence, U preserves angles:

$$\cos \angle(\tilde{w}, w) = \frac{\langle \tilde{w}, w \rangle}{||\tilde{w}||||w||} = \frac{\langle \tilde{z}, z \rangle}{||\tilde{z}||||z||} = \cos \angle(\tilde{z}, z).$$

- The transformation given by U is called *orthogonal* (not orthonormal!). It preserves distances and angles.

Orthonormal basis (I)

- Let $\{u_1, \ldots, u_n\}$ be an orthonormal basis for E. Then the $n \times n$ matrix $U = [u_1 \ \cdots \ u_n]$ is called orthogonal and satisfies *both*

$$U^T U = \mathbb{I}_{n\times n} \quad \textit{and} \quad UU^T = \mathbb{I}_{n\times n}.$$

  This means that both the vector columns and the row columns of U are orthonormal.

- We can write $x = UU^T x$ or, in components

$$x_i = \sum_{j=1}^{n}\sum_{k=1}^{n} U_{ij}U^T_{jk}x_k = \sum_{j=1}^{n}\sum_{k=1}^{n} U_{ij}U_{kj}x_k.$$

Orthonormal basis (II)

- Since U_{ij} is the ith component of u_j, that is $U_{ij} = (u_j)_i$, we get

$$x_i = \sum_{j=1}^{n} \sum_{k=1}^{n} (u_j)_i (u_j)_k x_k = \sum_{j=1}^{n} (u_j)_i u_j^T x = \sum_{j=1}^{n} \left(u_j^T x \right) (u_j)_i$$

or, in pure matrix notation,

$$x = \sum_{j=1}^{n} \left(u_j^T x \right) u_j ,$$

which expresses x in the basis $\{u_j\}$, with components

$$a_i = u_i^T x = \sum_{j=1}^{n} (u_i)_j x_j = \sum_{j=1}^{n} U_{ji} x_j = \sum_{j=1}^{n} U_{ij}^T x_j = (U^T x)_i .$$

- Matricially,

$$a = U^T x$$

which is called the *resolution* of x in the orthonormal basis.

- Then, from $x = UU^T x$,

$$x = Ua,$$

which is the *reconstruction* of x in the given orthonormal basis.

Orthogonal transformations - geometric interpretation

- The action of U on a vector $w = Uz$, can be interpreted either as a change of basis of the same object (passive interpretation) or as a transformation into a new object in the same basis (active interpretation).
- An example is provided by rotations in the plane. If $x \in \mathbb{R}^2$ and $y = U_\theta x$ with

$$U_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

then y is the vector x rotated counterclockwise by an angle θ. Indeed, if x has components $x_1 = r\cos\theta_1$, $x_2 = r\sin\theta_1$ then $y_1 = r\cos(\theta_1 + \theta)$ and $y_2 = r\sin(\theta_1 + \theta)$. It is easy to see that $U_\theta^T U_\theta = \mathbb{I}_{2\times 2}$.

- Another example is provided by reflections about the X axis, given by

$$R_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

giving $y_1 = x_1$, $y_2 = -x_2$. Again, $R_0^T R_0 = \mathbb{I}_{2\times 2}$.

- It is geometrically clear that any of these transformations preserves lengths and angles.

Gram-Schmidt procedure (I)

- This is a method to compute an orthonormal set from a given set of vectors.
- Given independent vectors $a_1, \ldots, a_k \in \mathbb{R}^n$, one wants to find k independent orthonormal vectors $q_1, \ldots, q_k$ spanning the same subspaces:

$$\text{span}(a_1, \ldots, a_r) = \text{span}(q_1, \ldots, q_r) \quad \forall r \leq k,$$

  and, in particular, for $r = k$.
- The general idea is to orthogonalize each vector with respect the previous ones, and then normalize.

Gram-Schmidt procedure (II)

- step 1a (initialize): $\tilde{q}_1 = a_1$.
- step 1b (normalize): $q_1 = \tilde{q}_1 / ||\tilde{q}_1||$.
- step 2a (remove q_1 component from a_2): $\tilde{q}_2 = a_2 - (q_1^T a_2) q_1$.
- step 2b (normalize): $q_2 = \tilde{q}_2 / ||\tilde{q}_2||$.
- step 3a (remove q_1, q_2): $\tilde{q}_3 = a_3 - (q_1^T a_3) q_1 - (q_2^T a_3) q_2$.
- step 3b (normalize): $q_3 = \tilde{q}_3 / ||\tilde{q}_3||$.
- $\vdots$
- step ka (remove $\{q_j\}_{j=1 \ldots k-1}$): $\tilde{q}_k = a_k - \sum_{j=1}^{k-1} (q_j^T a_k) q_j$.
- step kb (normalize): $q_k = \tilde{q}_k / ||\tilde{q}_k||$.

Gram-Schmidt procedure (III)

- It is easy to see that the above procedure yields an orthonormal set $\{q_1, \ldots, q_k\}$ (see exercise).
- Since the q are orthonormal, they are independent, and, being linear combinations of the a, they span the same subspace.
- In a more algorithmic form

$r = 0$

for $i = 1, \ldots, k$

$\{$

$\quad \tilde{q} = a_i - \sum_{j=1}^{r} (q_j^T a_i) q_j;$

$\quad r = r + 1;$

$\quad q_r = \tilde{q}/||\tilde{q}||;$

$\}$

Inverse Gram-Schmidt procedure

- One can invert the Gram-Schmidt (G-S) procedure to express each a_i in terms of the q_i. Notice that, since q_i is normalized and in the direction of $\tilde{q}_i$, $\tilde{q}_i = ||\tilde{q}_i||q_i$.
- From the "a" steps in G-S one obtains
  - $a_1 = \tilde{q}_1 = ||\tilde{q}_1||q_1$.
  - $a_2 = \tilde{q}_2 + (q_1^T a_2)q_1 = (q_1^T a_2)q_1 + ||\tilde{q}_2||q_2$.
  - $a_3 = \tilde{q}_3 + (q_1^T a_3)q_1 + (q_2^T a_3)q_2 = (q_1^T a_3)q_1 + (q_2^T a_3)q_2 + ||\tilde{q}_3||q_3$.
  - $\vdots$
  - $a_k = \tilde{q}_k + \sum_{j=1}^{k-1}(q_j^T a_k)q_j = \sum_{j=1}^{k-1}(q_j^T a_k)q_j + ||\tilde{q}_k||q_k$.
- One can express this as

$$\begin{aligned} a_i &= (q_1^T a_i)q_1 + (q_2^T a_i)q_2 + \cdots + (q_{i-1}^T a_i)q_{i-1} + ||\tilde{q}_i||q_i \\ &= r_{1i}q_1 + r_{2i}q_2 + \ldots + r_{i-1i}q_{i-1} + r_{ii}q_i. \end{aligned}$$

Notice that the r_{ij} come directly from the $G-S$ procedure, and that $r_{ii} = ||\tilde{q}_i|| > 0$.

QR factorization (I)

- The above expression of the a_i in terms of the q_i can be given the matrix form $A = QR$:

$$\underbrace{(a_1 \ a_2 \ \cdots \ a_k)}_{A_{n\times k}} = \underbrace{(q_1 \ q_2 \ \cdots \ q_k)}_{Q_{n\times k}} \underbrace{\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ 0 & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & r_{kk} \end{pmatrix}}_{R_{k\times k}}$$

- This is called the QR decomposition, or factorization, of A.
- Notice that $Q^TQ = I_k$, and that R is upper triangular and invertible, since $\det R = r_{11}r_{22}\cdots r_{kk} > 0$.
- The columns of Q provide an orthonormal basis for $\mathcal{R}(A)$.

Generalized Gram-Schmidt procedure (I)

- In basic G-S, $a_1, a_2, \ldots, a_k$ are assumed to be independent.
- If they are not, then one has that, for some j, a_j is linearly dependent on $a_1, \ldots, a_{j-1}$, and this implies in turn that a_j belongs to the subspace spanned by $q_1, \ldots, q_{j-1}$. Hence, when removing the $q_1, \ldots, q_{j-1}$ components from a_j in the ja step of G-S, one gets $\tilde{q}_j = 0$.
- A modified G-S procedure must then be used, where, if $\tilde{q}_j = 0$, one must skip to the next a_{j+1} and continue:

  $r = 0$

  for $i = 1, \ldots, k$

  $\{$

  $\quad \tilde{q} = a_i - \sum_{j=1}^{r} (q_j^T a_i) q_j;$

  $\quad$ if $\tilde{q} \neq 0 \quad \{r = r + 1;\ q_r = \tilde{q}/||\tilde{q}||;\}$

  $\}$

Generalized Gram-Schmidt procedure (II)

- On exit, the above procedure yields $q_1, \ldots, q_r$, with $r \leq k$, which are an orthonormal basis for $\mathcal{R}(A)$, and hence $r = \text{rank}(A)$. The r vectors q form an $n \times r$ matrix Q_r satisfying $Q_r^T Q_r = \mathbb{I}_{r \times r}$.
- Each a_i is a linear combination of the previously generated q_j, with coefficients given by the elements of the $r \times k$ matrix R_r.
- In matrix notation one has

$$A = Q_r R_r.$$

  The matrix R_r is in *upper staircase form*, *i.e.* upper triangular but with some 0s on the diagonal; the column index of the diagonal zeros indicate which as are dependent on the previous ones.
- The r rows of R_r are independent, and hence R_r^T has full column rank, a fact that will be used later.

QR factorization (II)

- Consider again an $n \times k$ matrix A whose column vectors may or may not be independent. As above, we write $A = Q_r R_r$ and recast it as

$$\underbrace{A}_{n\times k} = \begin{bmatrix} \underbrace{Q_r}_{n\times r} & \underbrace{Q_c}_{n\times(n-r)} \end{bmatrix} \begin{bmatrix} \underbrace{R_r}_{r\times k} \\ \underbrace{0}_{(n-r)\times k} \end{bmatrix},$$

  where the matrix Q_c is chosen so that $Q = [Q_r \; Q_c]$ is orthogonal.

- To find Q_c, one must choose any matrix A_c such that $[A \; A_c]$ is full rank. For instance one may overkill and set $A_c = \mathbb{I}_{n\times n}$.

- The general G-S is applied then to $[A \; A_c]$, and Q_r and R_r can then be read from the result.

QR factorization (III)

- $Q = [Q_r \; Q_c]$ gives a (non unique, since it depends on A_c) orthonormal basis for $\mathbb{R}^n$, in such a way that

$$A = QR \quad \text{with} \quad R = \begin{bmatrix} R_r \\ 0 \end{bmatrix}.$$

  This is called a full QR factorization of A.

- $\mathcal{R}(Q_r)$ and $\mathcal{R}(Q_c)$ are called complementary subspaces since
  1. they are orthogonal: each vector in the first subspace is orthogonal to each vector in the second one,
  2. their sum is $\mathbb{R}^n$: each vector in $\mathbb{R}^n$ can be uniquely written as the sum of a vector in $\mathcal{R}(Q_r)$ and a vector in $\mathcal{R}(Q_c)$.

QR factorization (IV)

- In `Matlab`, the full QR factorization is implemented as `[Q,R]=qr(A)` (several options are available; see `Matlab` help). Notice however that, depending on the `Matlab` version, there might be an overall minus sign for both Q and R.
- Example: A with independent column vectors but not forming a base of the corresponding space.

$$A = \begin{pmatrix} 1 & 2 \\ -1 & 1 \\ 0 & -3 \end{pmatrix} \quad Q = \begin{pmatrix} 0.7071 & 0.4082 & 0.5774 \\ -0.7071 & 0.4082 & 0.5774 \\ 0 & -0.8165 & 0.5774 \end{pmatrix} \quad R = \begin{pmatrix} 1.4142 & 0.7071 \\ 0 & 3.6742 \\ 0 & 0 \end{pmatrix}$$

from which

$$Q_r = \begin{pmatrix} 0.7071 & 0.4082 \\ -0.7071 & 0.4082 \\ 0 & -0.8165 \end{pmatrix}.$$

The row of zeros in R reflects the fact that the columns of A form a 2-dimensional subspace spanned by the two first columns of the orthogonal matrix Q.

QR factorization (V)

- Example: A with dependent column vectors.

For

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 3 \\ -1 & -2 & -3 \\ 4 & 4 & 8 \end{pmatrix}$$

one gets

$$Q = \begin{pmatrix} 0.2132 & -0.5415 & -0.4866 & 0.6515 \\ 0.4264 & -0.4874 & -0.2394 & -0.7234 \\ -0.2132 & -0.6498 & 0.7260 & 0.0719 \\ 0.8528 & 0.2166 & 0.4228 & 0.2168 \end{pmatrix} \quad R = \begin{pmatrix} 4.6904 & 4.2640 & 8.9544 \\ 0 & 1.6787 & 1.6787 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

from which

$$Q_r = \begin{pmatrix} 0.2132 & -0.5415 \\ 0.4264 & -0.4874 \\ -0.2132 & -0.6498 \\ 0.8528 & 0.2166 \end{pmatrix}.$$

Some applications of QR factorization

- Our main application of QR will be in the least-squares problem, but many results in linear algebra can be obtained as well.
- First of all, $\mathcal{R}(Q_r) = \text{rang}(A)$. Consider now

$$A^T = \begin{bmatrix} R_r^T & 0 \end{bmatrix} \begin{bmatrix} Q_r^T \\ Q_c^T \end{bmatrix}.$$

  This implies that $A^T z = 0$ iff $R_r^T Q_r^T z = 0$, and since R_r^T is full-rank, this is iff $Q_r^T z = 0$, that is iff $z \in \mathcal{R}(Q_c)$. Hence $\mathcal{R}(Q_c) = \mathcal{N}(A^T)$.
- From these two properties and the complementarity of $\mathcal{R}(Q_r)$ and $\mathcal{R}(Q_c)$ we conclude that

$$\mathcal{R}(A) \text{ and } \mathcal{N}(A^T) \text{ are complementary spaces.}$$

- This is called the orthogonal decomposition of $\mathbb{R}^n$ induced by $A \in \mathbb{R}^{n \times k}$. This has applications in many fields, for instance in formulating Kirchhoff laws for circuit theory.

Exercises

1. Applying the G-S algorithm, find an orthonormal basis for the subspace of $\mathbb{R}^3$ spanned by

$$v_1 = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \quad v_2 = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}.$$

2. Find an orthonormal basis for the space of polynomials of degree ≤ 2, with respect to the scalar product

$$\langle u, v \rangle = \int_0^1 u(x)v(x)\mathrm{d}x.$$

Hint: start with the basis $\{1, x, x^2\}$ and apply the G-S procedure.

3. Using `Matlab`, compute QR decompositions for the matrices

$$\begin{aligned} A &= \begin{pmatrix} -1 & 2 & 1 \\ 0 & 4 & 1 \end{pmatrix}, \\ B &= \begin{pmatrix} -1 & 2 \\ 1 & 0 \\ 4 & 1 \end{pmatrix}, \\ C &= (c_{kl})_{k,l=1,\ldots,10}, \text{ with } c_{kl} = k + l. \end{aligned}$$

Check the results.

Mathematical Methods – Lecture 3

Least squares estimation

Carles Batlle Arnau

Departament de Matemàtica Aplicada 4
and
Institut d'Organització i Control de Sistemes Industrials

Universitat Politècnica de Catalunya

Lecture goals

- To compute the least-squares solution to overdetermined systems.
- To compute the least-norm solution to undetermined systems.

Outline

- Overdetermined linear equations and the least-squares approximate solution.
- Orthogonality theorem revisited.
- Least-squares via QR factorization.
- Multi-objective least-squares.
- Underdetermined linear equations and the least norm solution.
- Least norm solution via QR.

References

BL5678 S. Boyd and S. Lall, lectures 5, 6, 7 and 8 of Introduction to Linear Dynamical Systems, Stanford Course EE263. Available at `http://www.stanford.edu/class/ee263/`

Overdetermined linear systems (I)

- Consider $y = Ax$ where $A \in \mathbb{R}^{m \times n}$ is *skinny*, that is $m > n$. This is an *overdetermined* set of linear equations, since there are more equations than unknowns.
- For most y, those not belonging to $\mathcal{R}(A)$, there is no solution.
- When there is no solution, one can try to find an approximate solution:
  - define the residual or error $r(x) = Ax - y$.
  - minimize $||r(x)||$ over all $x \in \mathbb{R}^n$ and find

$$x_{ls} = \arg \min_{x \in \mathbb{R}^n} ||Ax - y||.$$

- x_{ls} is called the least-squares solution to the overdetermined system. If $y \in \mathcal{R}(A)$, then $r(x_{ls}) = 0$ and x_{ls} is an exact solution.

Overdetermined linear systems (II)

- As an example, suppose we make m some measurements y_i, $i = 1, \ldots, m$ of an unknown function $f(t)$ at points t_i. We want to find the polynomial of degree $n-1$, with n free parameters, $g(t) = \sum_{i=0}^{n-1} \alpha_i t^i$, $n < m$, which best describes y_i.
- We write $y_i - g(t_i) = r_i$ and the goal is to minimize $\sum_{i=1}^{n-1} r_i^2$.
- This can be given the $Ax = y$ form as follows:

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}}_{y} = \underbrace{\begin{pmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^{n-1} \\ \vdots & & & \vdots & \\ 1 & t_m & t_m^2 & \ldots & t_m^{n-1} \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_{n-1} \end{pmatrix}}_{x}.$$

Least-squares approximate solution (I)

- Assume A is full rank and skinny (this means it is full *column* rank). If it is not full column rank, one can always redefine the x so that dependent columns are eliminated.
- To find x_{ls}, let us minimize the norm of the residual squared

$$||r||^2 = x^T A^T A x - 2y^T A x + y^T y.$$

- Setting the gradient (column vector) to zero

$$\partial_x ||r||^2 = 2A^T A x - 2A^T y = 0,$$

one gets the *normal equations*

$$A^T A x = A^T y.$$

- If A is full column rank then $A^T A$ is invertible, and the minimizing vector is

$$x_{ls} = (A^T A)^{-1} A^T y.$$

Least-squares approximate solution (II)

- If A is square, one can expand the inverse of the product and obtain

$$x_{ls} = (A^T A)^{-1} A^T y = A^{-1} y.$$

  Obviously, if A is square, since we assume that it is full column rank, it is also invertible, and hence the above result. In this case one also has that $y \in \mathcal{R}(A)$.

- The *pseudo-inverse* of A is defined as

$$A^\dagger = (A^T A)^{-1} A^T.$$

- The pseudo-inverse $A^\dagger$ is a left inverse of the skinny, full (column) rank A:

$$A^\dagger A = (A^T A)^{-1} A^T A = \mathbb{I}.$$

Least-squares approximate solution (III)

- The projection operator on $\mathcal{R}(A)$, denoted by $\mathcal{P}_{\mathcal{R}(A)}$, is given by

$$\mathcal{P}_{\mathcal{R}(A)}(y) = A(A^T A)^{-1} A^T y, \quad \forall\, y \in \mathbb{R}^n.$$

- Indeed, it maps any vector into $\mathcal{R}(A)$, since the result is the image by A of $(A^T A)^{-1} A^T y$. Furthermore, it is a *projection* operator, since it is idempotent:

$$\left(\mathcal{P}_{\mathcal{R}(A)}\right)^2 = A(A^T A)^{-1} A^T A(A^T A)^{-1} A^T = A(A^T A)^{-1} A^T = \mathcal{P}_{\mathcal{R}(A)},$$

  i.e. applying it twice is the same that applying it once, as must be the case for a projection.

- We know already the projection theorem, which states that the optimal residual is orthogonal to the approximating subspace. We are going to show that the residual associated to x_{ls} is indeed optimal (the gradient calculation done above is only a necessary condition).

Least-squares approximate solution (IV)

- The optimal residual is $r = Ax_{ls} - y$, and the approximating subspace is $\mathcal{R}(A)$, *i.e* the set of vectors of the form Az for any z. Then, using that the transpose of the symmetric matrix $(A^TA)^{-1}$ is the same matrix,

$$\begin{aligned} r^T(Az) &= (A((A^TA)^{-1}A^Ty) - y)^TAz = y^T(A(A^TA)^{-1}A^T - \mathbb{I})Az \\ &= y^T(A(A^TA)^{-1}A^TA - A)z = y^T(A - A)z = 0. \end{aligned}$$

  Hence $Ax_{ls} - y \perp \mathcal{R}(A)$.

- In particular, $Ax_{ls} - y \perp A(x - x_{ls})$ for any x. Then

$$\begin{aligned} ||Ax - y||^2 &= ||(Ax_{ls} - y) + A(x - x_{ls})||^2 \\ &= ||(Ax_{ls} - y)||^2 + ||A(x - x_{ls})||^2 + 2\underbrace{\langle Ax_{ls} - y, A(x - x_{ls})\rangle}_{=0} \\ &= ||(Ax_{ls} - y)||^2 + ||A(x - x_{ls})||^2 \geq ||(Ax_{ls} - y)||^2. \end{aligned}$$

- Hence the residual for any x is not less than the residual for x_{ls}

$$||Ax - y|| \geq ||Ax_{ls} - y|| \quad \forall x,$$

  and equality is attained only at $x = x_{ls}$.

Least-squares via QR (I)

- We can obtain expressions for both the approximate least-squares solution and the optimal error in terms of the QR factorization of A. This is not only numerically advantageous but also yields further inside into the basic result.
- Let us perform a full QR factorization of the skinny ($m > n$), full column rank $A \in \mathbb{R}^{m \times n}$:

$$\underbrace{A}_{m \times n} = \begin{bmatrix} \underbrace{Q_1}_{m \times n} & \underbrace{Q_2}_{m \times (m-n)} \end{bmatrix} \begin{bmatrix} \underbrace{R_1}_{n \times n} \\ \underbrace{0}_{(m-n) \times n} \end{bmatrix},$$

with $[Q_1 \ Q_2] \in \mathbb{R}^{m \times m}$ orthogonal, and $R_1 \in \mathbb{R}^{n \times n}$ upper triangular and invertible.

Least-squares via QR (II)

- Using this one has

$$||Ax - y||^2 = \left|\left| \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - y \right|\right|^2 .$$

- Since an orthogonal transformation does not change the norm this is

$$\begin{aligned} ||Ax - y||^2 &= \left|\left| \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}^T \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}^T y \right|\right|^2 \\ &= \left|\left| \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} y \right|\right|^2 = \left|\left| \begin{bmatrix} R_1 x - Q_1^T y \\ -Q_2^T y \end{bmatrix} \right|\right|^2 \\ &= ||R_1 x - Q_1^T y||^2 + ||Q_2^T y||^2. \qquad (*) \end{aligned}$$

- The second contribution in the last expression does not depend on our selection of x, and thus cannot be reduced; however we can minimize the first contribution by selecting

$$x = x_{ls} = R_1^{-1} Q_1^T y.$$

This is the least-squares approximate solution to the overdetermined system of equations in terms of the QR factorization.

Least-squares via QR (III)

- As a bonus we get also the expression of the optimal residual

$$\begin{aligned} Ax_{ls} - y &= \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} R_1^{-1} Q_1^T y - y \\ &= Q_1 Q_1^T y - y = -(\mathbb{I} - Q_1 Q_1^T) y. \end{aligned}$$

- But from the orthogonality of $[Q_1 \; Q_2]$ one gets immediately $Q_1 Q_1^T + Q_2 Q_2^T = \mathbb{I}$, and hence

$$Ax_{ls} - y = -Q_2 Q_2^T y,$$

with norm $|| - Q_2 Q_2^T y|| = ||Q_2^T y||$, as can also be seen directly from $(*)$ in the previous slide.

Multi-objective least-squares

- In many applications, one has two (or more) objectives of the type

$$J_1 = ||Ax - y||^2 \text{ small} \quad \text{and} \quad J_2 = ||Fx - g||^2 \text{ small.}$$

- No matter the number of equations in $Ax - y = 0$, $Fx - g = 0$, the two objectives are generally competing, and no exact solution exists.
- We can apply the same procedure we used for overdetermined systems; this can be justified if some matrices are invertible.
- In the plane (J_1, J_2) a point can either correspond to values which can be achieved for some $x \in \mathbb{R}^n$ or to values such that either J_1, J_2 or both are not achieved. This splits the positive (J_1, J_2) quadrant into two regions, separated by a boundary called the *optimal trade-off curve*; the corresponding values of x are called *Pareto optimal*.
- If $J_1 = 0$ (resp. $J_2 = 0$) can be achieved, then $J_1 = 0$ (resp. $J_2 = 0$) is an asymptote of the optimal trade-off curve.

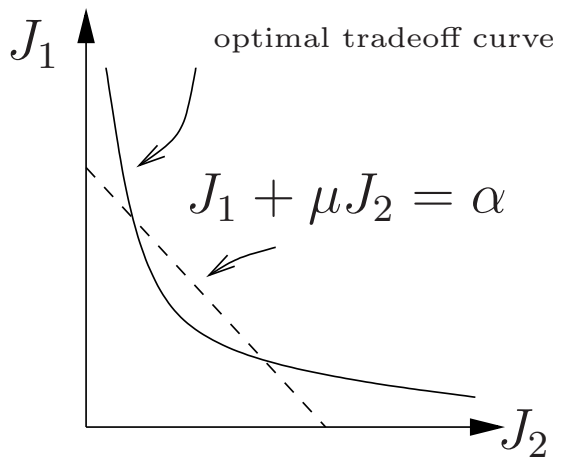
Weighted-sum objective

- In order to find Pareto optimal points, one can minimize a weighted-sum objective

$$J_1 + \mu J_2 = ||Ax - y||^2 + \mu ||Fx - g||^2,$$

  where the parameter $\mu \geq 0$ gives the relative weight of J_1 and J_2.

- Points with constant weighted sum, $J_1 + \mu J_2 = \alpha$, correspond to a segment with slope $-\mu$ on the first quadrant. By varying μ from 0 to $+\infty$ one can sweep out the entire optimal tradeoff curve.



Minimizing the weighted-sum objective

- The weighted-sum objective can be expressed as an ordinary least-squares objective:

$$\begin{aligned} ||Ax-y||^2+\mu||Fx-g||^2 &= \left|\left|\begin{bmatrix} A \\ \sqrt{\mu}F \end{bmatrix} x - \begin{bmatrix} y \\ \sqrt{\mu}g \end{bmatrix}\right|\right|^2 \\ &= ||\tilde{A}x-\tilde{y}||^2, \end{aligned}$$

  with an obvious notation.

- Assuming that $\tilde{A}$ is full rank, the solution is given by

$$\begin{aligned} x &= (\tilde{A}^T\tilde{A})^{-1}\tilde{A}^T\tilde{y} \\ &= (A^TA+\mu F^TF)^{-1}(A^Ty+\mu F^Tg). \end{aligned}$$

- The corresponding value of $J_1+\mu J_2$ yields the value of α such that the line $J_1+\mu J_2$ touches the optimal tradeoff curve at a single point, for the given value of μ.

Regularized least-squares (I)

- For $F = \mathbb{I}$, $g = 0$, one has the special objectives

$$J_1 = ||Ax - y||^2, \quad J_2 = ||x||^2.$$

- The corresponding weighted-sum objective is called *regularized least-squares*, with solution

$$x = (A^T A + \mu \mathbb{I})^{-1} A^T y,$$

  also known as *Tychonov regularization*.

- For $\mu > 0$, this works for any A, with no shape or rank restriction.

Regularized least-squares (II)

- As an example, consider a unit mass at rest subject to piecewise constant forces x_i for $i-1 < t < i$, $i = 1, 2, \ldots, 10$.
- Using repeatedly the formulae for uniformly accelerated movement

$$y(i) = y(i-1) + v(i-1) \cdot 1 + \frac{1}{2} x_i \cdot 1^2 = y(i-1) + v(i-1) + \frac{1}{2} x_i, \quad v(i) = \sum_{k=1}^{i} x_k,$$

one gets

$$y(10) = \sum_{i=1}^{10} \frac{21 - 2i}{2} x_i = Ax,$$

with $x \in \mathbb{R}^{10}$ and $A \in \mathbb{R}^{1 \times 10}$ with elements $(21-2i)/2$, $i = 1, \ldots, 10$.

- The solution to the regularized least squares with desired final position $y(10) = y_d$ is then

$$x = (A^T A + \mu \mathbb{I})^{-1} A^T y_d.$$

- The following table displays the values obtained for $y_d = 5$ and several values of μ, and illustrates the competing minimizing goals:

μ	$y(10)$	$\|\|x\|\|$
10^{-6}	5.0000	0.2742
1	4.9850	0.2734
100	3.8439	0.2108
10^{6}	0.0017	$9.1143 \cdot 10^{-5}$

Underdetermined linear systems

- Consider an underdetermined linear system $y = Ax$, where $A \in \mathbb{R}^{m \times n}$ and $m < n$, that is A is *fat*.
- Since there are more variables than equations, one has that $\mathcal{N}(A) \neq \{0\}$ and, given a solution x_p (if it exists), any

$$x = x_p + z \quad \text{with} \quad z \in \mathcal{N}(A),\ z \neq 0$$

will be a different solution.

- We assume that A has the maximum column rank possible, *i.e.* m, so that there is always a solution for each y and, furthermore,

$$\dim \mathcal{N}(A) = n - \dim \mathcal{R}(A) = n - m,$$

meaning that there are $n - m$ degrees of freedom to get solutions from a given one.

Least-norm solution (I)

- Since A is also full row rank (m), one has that AA^T is invertible and a solution to $Ax = y$ is given by

$$x_{ln} = A^T(AA^T)^{-1}y.$$

- Assume that there is another solution x, $Ax = y$, so that $A(x - x_{ln}) = y - y = 0$. Then

$$(x-x_{ln})^T x_{ln} = (x-x_{ln})^T A^T(AA^T)^{-1}y = (A(x-x_{ln}))^T(AA^T)^{-1}y = 0,$$

and we conclude that $(x - x_{ln}) \perp x_{ln}$.

- Then

$$||x||^2 = ||x_{ln} + x - x_{ln}||^2 = ||x_{ln}||^2 + ||x - x_{ln}||^2 \geq ||x_{ln}||^2,$$

so that x_{ln} is the least-norm solution.

Least-norm solution (II)

- The pseudo-inverse of the full rank, fat A is defined as

$$A^{\ddagger} = A^T (A A^T)^{-1}.$$

- $A^{\ddagger}$ is a *right-inverse* of A.
- $\mathbb{I} - A^{\ddagger} A$ gives the projection onto $\mathcal{N}(A)$.
- Remember that, for a full rank, skinny matrix A,
  - $A^{\dagger} = (A^T A)^{-1} A^T$ is called the pseudo-inverse of A.
  - $A^{\dagger}$ is a *left-inverse* of A.
  - $A A^{\dagger}$ gives the projection onto $\mathcal{R}(A)$.

Least-norm solution via QR

- The least-norm solution can be computed effectively from the QR factorization of A^T (which is different from that of A!).
- Write $A^T = QR$ with $Q \in \mathbb{R}^{n\times m}$, $Q^TQ = \mathbb{I}_{m\times m}$, and $R \in \mathbb{R}^{m\times m}$ upper triangular and nonsigular.
- Then

$$\begin{aligned} x_{ln} &= A^T(AA^T)^{-1}y = QR(R^TQ^TQR)^{-1}y = QR(R^TR)^{-1}y \\ &= QRR^{-1}R^{-T}y \\ &= QR^{-T}y, \end{aligned}$$

  where the fact that R is square and nonsingular allows to compute $(R^TR)^{-1} = R^{-1}R^{-T}$.

- Furthermore,

$$||x_{ln}|| = ||R^{-T}y||.$$

Exercises

1. Consider the overdetermined linear system $Ax = y$.

$$y = \begin{pmatrix} 1 \\ 2 \\ 7 \\ 1 \end{pmatrix}, \qquad A = \begin{pmatrix} 1 & 0 & 1 \\ 2 & -1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

   Find the least-squares approximate solution 1) by hand, without QR, 2) using `Matlab` and its QR factorization (beware of the sign conventions!).

2. Solve the least-squares weighted sum problem for the unit mass particle developed in the lecture, with the same data but with the additional approximate goal of making the final velocity zero.

3. With the same data as the previous problem, solve the underdetermined problem of getting the final position $y_d = 5$ and the final position $v_d = 0$, and compute the minimum of the norm of the force. Compare the results of the two problems.

Mathematical Methods – Lecture 4

SVD factorization and applications

Carles Batlle Arnau

Departament de Matemàtica Aplicada 4
and
Institut d'Organització i Control de Sistemes Industrials

Universitat Politècnica de Catalunya

Lecture goals

- To motivate the problem of matrix perturbation.
- To define norms for matrices.
- To introduce the Singular Valued Decomposition (SVD).
- To show how the SVD can be used to compute some matrix norms.
- To discuss how much a matrix can be perturbed (in the additive and multiplicative senses) before becoming singular, and obtaining the results in terms of singular values.
- To retake the matrix inversion conditioning problem and solve it in terms of singular values.

Outline

- Motivation: inversion of an ill-conditioned matrix.
- Induced matrix norms.
- The Frobenius norm.
- The SVD.
- Using the SVD to compute matrix norms.
- Additive perturbation.
- Multiplicative perturbation.
- Conditioning of matrix inversion.

References

DDV45 M. Dahleh, M.A. Dahleh and G. Verghese, Lectures on Dynamic Systems and Control, chapters 4 and 5, MIT Course 6.241.

BL1516 S. Boyd and S. Lall, lectures 15 and 16 of Introduction to Linear Dynamical Systems, Stanford Course EE263. Available at `http://www.stanford.edu/class/ee263/`

Perturbing matrices

- In this lecture, we will obtain results relating the norm of a matrix to how much some of its characteristics, for instance invertibility, changes under small variations of the matrix elements.
- A special kind of decomposition of a matrix, the **Singular Value Decomposition** (SVD), will be instrumental to obtain the results.
- We begin with a motivating example. Let

$$A = \begin{pmatrix} 100 & 100 \\ 100.2 & 100 \end{pmatrix}.$$

- A quick calculation shows that

$$A^{-1} = \begin{pmatrix} -5 & 5 \\ 5.01 & -5 \end{pmatrix}.$$

- Now we perturb A slightly...

$$A + \Delta A = \begin{pmatrix} 100 & 100 \\ 100.1 & 100 \end{pmatrix},$$

Perturbing matrices (cont'd)

- ... and we get

$$(A + \Delta A)^{-1} = \begin{pmatrix} -10 & 10 \\ 10.01 & -10 \end{pmatrix}.$$

- A 0.1% change in one entry of A has brought a 100% change in the inverse!
- The same happens if we want to solve $Ax = b$ and perturb A.
- This situation is much worse than what happens with scalars. If $a \in \mathbb{R}$ (or $\mathbb{C}$) depends on a parameter λ, one has

$$\frac{1}{a^{-1}} \frac{\mathrm{d}a^{-1}}{\mathrm{d}\lambda} = -\frac{1}{a} \frac{\mathrm{d}a}{\mathrm{d}\lambda}$$

  so the fractional change in a^{-1} is of the same order of magnitude than that of a.
- The above example shows a purely matrix phenomenon, related to the fact that the two columns of A are nearly dependant,and A is thus nearly singular.
- We need to quantify this sensitivity.

Matrix norms

- An $m \times n$ matrix A can be viewed as an operator between vector spaces of dimension n and m, respectively.
- If those vector spaces are provided with a norm, we can define an **induced norm** for matrices.
- For instance, the **induced 2-norm** is

$$||A||_2 \overset{(*)}{\equiv} \sup_{x \neq 0} \frac{||Ax||_2}{||x||_2} \overset{(**)}{=} \max_{||x||_2=1} ||Ax||_2 .$$

- The definition $(*)$ shows that the induced norm measures how much lengths of vectors are amplified by the matrix.
- The existence of $\max_{||x||_2=1} ||Ax||_2$ follows from the fact that the norm is a continuous function of x, and $||x||_2 = 1$ is a compact set.
- To prove equality $(**)$, notice that, if $x \neq 0$, $x/||x||_2$ has norm 1, and that

$$\frac{||Ax||_2}{||x||_2} = \left|\left| A \frac{x}{||x||_2} \right|\right|_2$$

so, in fact, the supremum is computed on unitary vectors.

Matrix norms (cont'd)

- We can obtain an induced matrix norm for any p-norm (usually $p = 1, 2, \infty$) on the vector spaces:

$$||A||_p \equiv \sup_{x \neq 0} \frac{||Ax||_p}{||x||_p} = \max_{||x||_p = 1} ||Ax||_p.$$

- It is easy to see that these norms are indeed norms, *i.e.* they satisfy

  (N1) $||A||_p \geq 0$ and $||A||_p = 0$ if and only if $A = 0$.
  (N2) $||\alpha A||_p = |\alpha| ||A||_p$, for all $\alpha \in \mathbb{C}$.
  (N3) $||A + B||_p \leq ||A||_p + ||B||_p$.

- Induced norms have two very important additional properties.
- The first one follows from the properties of the supremum:

$$||Ax||_p \leq ||A||_p ||x||_p.$$

- The second one is called **submultiplicative property**. For any $m \times n$ A and $n \times r$ B,

$$||AB||_p \leq ||A||_p ||B||_p.$$

Matrix norms (cont'd)

- Induced norms $||\cdot||_1$ and $||\cdot||_\infty$ can be computed quite easily:

$$||A||_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|.$$

$$||A||_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|.$$

- There are matrix norms that are not induced by a norm in the subjacent vector spaces. One of them is the **Frobenius norm**:

$$||A||_F \equiv \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}}.$$

- Notice that this is just the vector norm when the elements of A are arranged as an $m \cdot n$ vector. The Frobenius norm has also the submultiplicative property, even if it is not an induced norm.
- The Frobenius norm can be computed as (for A complex, $'$ means transpose *plus* complex conjugation, also known as **hermitian transpose**)

$$||A||_F = (\text{trace}(A'A))^{\frac{1}{2}}.$$

Some matrix definitions

The following facts will be used:

- A complex matrix $U \in \mathbb{C}^{n \times n}$ is **unitary** if $U'U = UU' = \mathbb{I}$.
- A real matrix $O \in \mathbb{R}^{n \times n}$ is **orthogonal** if $O'O = OO' = \mathbb{I}$.
- If U is unitary, $||Ux||_2 = ||x||_2$.
- If $S = S'$ (we say S is **hermitian**) then it can be diagonalized by an unitary matrix, *i.e.* $U'SU =$ diagonal.
- For any matrix, both $A'A$ and AA' are hermitian; they can be diagonalized by unitary matrices.
- For any matrix A, the eigenvalues of $A'A$ and AA' are always real and non-negative.

Some matrix definitions (cont'd)

- Let us prove the last statement. Let x be an eigenvector of $A'A$ with eigenvalue $\lambda \in \mathbb{C}$: $A'Ax = \lambda x$. Then $\langle x, A'Ax \rangle = x'A'Ax$ can be computed **as**

$$x'\lambda x = \lambda ||x||_2^2,$$

or as

$$(A'Ax)'x = (\lambda x)'x = \lambda^* ||x||_2^2,$$

from which $\lambda^* = \lambda$, *i.e.* $\lambda \in \mathbb{R}$.

- Now assume $\lambda < 0$. One has

$$||Ax||_2^2 = \langle Ax, Ax \rangle = \langle x, A'Ax \rangle = \langle x, \lambda x \rangle = \lambda \langle x, x \rangle = \lambda ||x||_2^2,$$

which is a contradiction, since λ is the only negative term.

The SVD

Singular Value Decomposition, or SVD. Any matrix $A \in \mathbb{C}^{m\times n}$ can be written as

$$A = \overset{m\times m}{U} \quad \overset{m\times n}{\Sigma} \quad \overset{n\times n}{V'},$$

where U, V are unitary: $U'U = \mathbb{I} = V'V$, and

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},$$

with $\Sigma_1 = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, and

$$\sigma_i = \sqrt{i\text{th nonzero eigenvalue of } A'A}.$$

The SVD (cont'd)

- The σ_i are termed the **singular values** of A, and are arranged in descending magnitude:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

  This ordering makes sense, since $A'A$ is hermitian and hence all its nonzero eigenvalues are positive.

- In `Matlab`, SVD are obtained using `[u,s,v]=svd(A)`.
- If U and V are expressed in terms of their columns:

$$\begin{aligned} U &= [u_1|u_2|\ldots|u_m], \\ V &= [v_1|v_2|\ldots|v_n], \end{aligned}$$

  then

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i'.$$

The SVD (cont'd)

- The u_i are termed the **left singular vectors** of A, while the v_i are the **right singular vectors**.
- Each term of the form uv', that is column vector times row vector, is called a *dyad*, and a matrix of the form $A = uv'$ is a *dyadic* matrix. The above result shows that any matrix can be expressed as a linear combination of dyads.
- Using the dyadic form of the SVD one gets

$$Ax = \sum_{i=1}^{r} \sigma_i u_i \underbrace{v_i' x}_{\text{projection}} = \sum_{i=1}^{r} w_i u_i,$$

  which is a weighted sum of the u_i, with weights w_i equal to the product of the singular value σ_i and the projection of x on v_i.
- The last observation shows that $\mathcal{R}(A) = \text{span}\{u_1, u_2, \ldots, u_r\}$.

The SVD (cont'd)

- Since the columns of U are independent (U' provides a change of basis in C^m which diagonalizes AA'), so are the first r columns, and hence
$$\dim \mathcal{R}(A) = r.$$
- The null space (or kernel) of A is given by span$\{v_{r+1}, v_{r+2}, \ldots, v_n\}$:
$$U\Sigma V'x = 0 \overset{U'U=\mathbb{I}}{\iff} \Sigma V'x = 0 \iff \begin{bmatrix} \sigma_1 v_1'x \\ \vdots \\ \sigma_r v_r'x \end{bmatrix} = 0 \iff v_i'x = 0,$$
from which $x \in$ span$\{v_{r+1}, v_{r+2}, \ldots, v_n\}$, and, since the columns of V are independent, dim ker$(A) = n - r$.
- The above facts can be used to solve $Ax = b$ very efficiently, no matter the dimensions and ranks.

SVD and matrix norms

- The SVD can be used to compute the induced 2-norm of a matrix:

$$||A||_2 = \sigma_1 \equiv \sigma_{max}(A).$$

  The maximum amplification is given by the maximum singular value.

- Given $A \in \mathbb{C}^{m \times n}$, suppose it has full column rank. Then $\min_{||x||_2=1} ||Ax||_2 = \sigma_n \equiv \sigma_{min}(A)$, which shows that the minimum amplification on the unit sphere is equal to the minimum singular value. If $\text{rank}(A) < n$ then there is an $x \neq 0$ such that $Ax = 0$.

- The Frobenius norm can also be given in terms of singular values:

$$||A||_F = \left(\sum_{i=1}^{r} \sigma_i^2 \right)^{\frac{1}{2}}.$$

Additive perturbation

- Suppose $A \in \mathbb{C}^{m \times n}$ has full column rank. Then

$$\min_{\Delta \in \mathbb{C}^{m \times n}} \{||\Delta||_2 \mid \text{rank}(A + \Delta) < n\} = \sigma_{min}(A) > 0.$$

- To show this, suppose $A + \Delta$ has rank $< n$. Then there exists $x \neq 0$, $||x||_2 = 1$ such that $(A + \Delta)x = 0$, from which

$$||\Delta x||_2 = ||Ax||_2 \geq \sigma_{min}(A).$$

- Since $||\Delta||_2 = ||\Delta||_2||x||_2 \geq ||\Delta x||_2$, we arrive at $||\Delta||_2 \geq \sigma_{min}$.
- To complete the proof, we must show that this bound is actually attained. Thus, we must construct Δ such that $A + \Delta$ has rank $< n$ and $||\Delta||_2 = \sigma_{min}$.
- Take $\Delta = -\sigma_{min}(A)u_n v_n'$. From the expression of a matrix in terms of left and right singular vectors and the computation of 2-norms using SVD, it follows that $||\Delta||_2 = || - \Delta||_2 = \sigma_{min}$.
- Moreover,

$$(A + \Delta)v_n = \left(\sum_{i=1}^{n} \sigma_i u_i v_i' - \sigma_{min} u_n v_n' \right) v_n = \sigma_{min} u_n - \sigma_{min} u_n v_n' v_n = 0,$$

which completes the proof.

Low rank approximations

- The additive perturbation problem can be generalized to the problem of finding a matrix of rank $\leq k$ such that the error has minimum 2-norm.
- Given $A \in \mathbb{C}^{m \times n}$ with full column rank and $k \leq n$, find

$$e_k = \min_{B \in \mathbb{C}^{m \times n}} \{||A - B||_2 \mid \text{rank}(B) \leq k\}.$$

- The solution is given by

$$e_k = \sigma_{k+1} \quad \text{and the optimal } B \text{ is} \quad B_k = \sum_{i=1}^{k} \sigma_i u_i v_i'.$$

- For $k = n - 1$ we recover the additive perturbation problem, with $\Delta = B - A$.
- If $\sigma_{k+1} \ll \sigma_k$, it is said that A has *numerical rank* k, because it can be approximated with small error by a matrix of rank k.

Low rank approximations (cont'd)

- Suppose $A \in \mathbb{R}^{10000 \times 10000}$ is a dense matrix, so that computing $y = Ax$, with $x \in \mathbb{R}^{10000}$ requires 10^8 real multiplications.
- If A has singular values $\sigma_1 = 100$, $\sigma_2 = 35$, $\sigma_3 = 10$, $\sigma_4 = 2$ and $\sigma_5 = 0.001$, then the optimal rank 4 aproximant is

$$B_4 = \sum_{i=1}^{4} \sigma_i u_i v_i'.$$

- The approximate value of $y = Ax$ is then

$$y_4 = B_4 x = 100(v_1'x)u_1 + 35(v_2'x)u_2 + 10(v_3'x)u_3 + 2(v_4'x)u_4,$$

which requires only $\sim 4 \times 10^4$ multiplications, at the cost of an error given by

$$||y - y_4||_2 = ||(A - B_4)x||_2 \leq ||A - B_4||_2 ||x||_2 \leq 0.001||x||_2.$$

Multiplicative perturbation

- Given $A \in \mathbb{C}^{m \times n}$, without any rank assumption, then

$$\min_{\Delta \in \mathbb{C}^{n \times m}} \{||\Delta||_2 \mid \mathbb{I} - A\Delta \text{ is singular}\} = \frac{1}{\sigma_{max}(A)}.$$

- As can be seen in the proof, the bound is achieved with

$$\Delta = \frac{1}{\sigma_{max}} v_1 u_1'.$$

- This result is known as the (algebraic) **small gain theorem** because it guarantees that

$$\mathbb{I} - A\Delta$$

is nonsingular provided that

$$||\Delta||_2 < \frac{1}{\sigma_{max}} = \frac{1}{||A||_2},$$

which is more commonly written as

$$||\Delta||_2 ||A||_2 < 1.$$

Conditioning of matrix inversion

- Just for completeness, we can now turn back to our motivating problem, that of the sensitivity of a matrix under inversion.
- Assume A is invertible. Taking differentials in $A^{-1}A = \mathbb{I}$, one immediately gets

$$\mathrm{d}(A^{-1}) = -A^{-1}\mathrm{d}AA^{-1}.$$

- Taking norms and using twice the submultiplicative property yields

$$||\mathrm{d}(A^{-1})|| \leq ||A^{-1}||^2||\mathrm{d}A||,$$

or, equivalently,

$$\frac{||\mathrm{d}(A^{-1})||}{||A^{-1}||} \leq ||A||||A^{-1}||\frac{||\mathrm{d}A||}{||A||}.$$

- The factor $||A||||A^{-1}||$ is called the **condition number** of A, and is denoted by $K(A)$.
- An index is attached if one wants to specify the norm. For instance, from the SVD it follows immediately that

$$K_2(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}.$$

Conditioning of matrix inversion (cont'd)

- The importance of

$$\frac{||\mathrm{d}(A^{-1})||}{||A^{-1}||} \leq ||A||||A^{-1}|| \frac{||\mathrm{d}A||}{||A||}$$

  is that the bound can be saturated, so a perturbation can be found that makes the situation as bad as possible: the relative change in the inverse of A can be $K(A)$ times as large as the relative change of A.
- For instance, for 2-norms, the bound is saturated by perturbing along $u_n v_n'$:

$$\mathrm{d}A = -\mathrm{d}\sigma u_n v_n'.$$

- Hence, a large condition number corresponds to a matrix whose inverse is very sensitive to small variations; such a matrix is said to be **ill conditioned** or **poorly conditioned**.
- A high condition number also indicates that the matrix is close to losing rank, in the sense that a perturbation Δ of small 2-norm ($\sigma_{min}(A)$) relative to the norm of A ($\sigma_{max}(A)$) exists, such that $A + \Delta$ has lower rank than A.
- All this definitions and considerations can be extended to non-square matrices A.
- Notice that, for an scalar, the condition number is always equal to 1; hence the good behavior in that case.

Assignments for lecture 4

- Given the SVD of a matrix A, $A = U\Sigma V'$, show that the change of basis given by U' diagonalizes AA' and that the one associated to V' diagonalizes $A'A$, and compute the corresponding diagonal forms.
- Show that

$$||A||_F \geq ||A||_2,$$

  and that, for a dyad, both norms are actually equal.
- Using `Matlab`, generate some big matrices, get their SVD and use appropriate low rank approximations to compute matrix-vector products. Check the theoretical bound on the error.

Mathematical Methods – Lecture 5

Partial differential equations

Carles Batlle Arnau

Departament de Matemàtica Aplicada 4
and
Institut d'Organització i Control de Sistemes Industrials

Universitat Politècnica de Catalunya

Lecture goals

- To present some general definitions and notations about PDE, and the concept of well-posedness.
- To present some PDE coming from physics and engineering.

Outline

- ODE review.
- PDE basic concepts. Well-posedness.
- Classification. Strong and weak solutions.
- Modeling.
- Initial and boundary conditions.

References

PR Pinchover, Y., & J. Rubinstein, *An introduction to partial differential equations*, Cambridge University Press, Cambridge, UK (2005), chapter 1.

GL Guenther, R.B., & J.W. Lee, *Partial differential equations of mathematical physics and integral equations*, Prentice-Hall, Englewood Cliffs, NJ, USA (1988), chapter 1.

ODE review (I)

Let $x \in \mathbb{R}^n$ and consider a system of n first-order ordinary differential equations (ODE)

$$\dot{x} = f(x, t)$$

or, in components

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, \ldots, x_n, t), \\ &\vdots \\ \dot{x}_n &= f_n(x_1, \ldots, x_n, t), \end{aligned}$$

where the dot denotes the derivative with respect to a t. A solution to the above system is a (vector)function $x(t) = (x_1(t), \ldots, x_n(t))$ such that, together with $\dot{x}(t) = (\dot{x}_1(t), \ldots, \dot{x}_n(t))$ satisfies the above system identically.

ODE review (II)

- Under suitable conditions on f, a unique solution $x(t)$ exists such that, given t_0 and $(x_{10}, \ldots, x_{n0})$, it satisfies

$$x_1(t_0) = x_{10}, \ldots, x_n(t_0) = x_{n0}.$$

  In general, solutions are only guaranteed to exist in an open set $(t_0 - \epsilon, t_0 + \epsilon)$ around the initial time.

- If f does not depend on t the system of ODE is called *autonomous*. Otherwise it is *nonautonomous*.

- If f is a linear function of $x_1, \ldots, x_n$ (and with arbitrary dependence on t), the system is called *linear*.

- The *general solution* to the above system os ODE contains n arbitrary constants C_i, $i = 1, \ldots, n$, allowing to satisfy the *initial condition* $(x_{10}, \ldots, x_{n0})$ at t_0.

ODE review (III)

- For an autonomous system, a solution $x(t)$ is a curve in $\mathbb{R}^n$ parameterized by t, and such that at each point the velocity tangent vector is given by $f(x(t))$.
- Nonautonomous systems can be promoted to autonomous ones in $\mathbb{R}^{n+1}$ by considering t as a dependent variable, introducing a new independent variable τ, and adding the new ODE $t' = 1$, where $'$ denotes derivation with respect to τ:

$$\begin{aligned} x' &= f(x,t), \\ t' &= 1. \end{aligned}$$

- ODE of higher order

$$F(x^{(n)}, x^{(n-1)}, \ldots, \dot{x}, x, t) = 0$$

can always be converted, nonuniquely, to a system of n first order ODE. The converse is not necessarily true.

ODE review (IV)

- Generically, ODE systems cannot be solved exactly and numerical methods starting with given initial conditions must be used to obtain approximations to the solutions.
- Both uniqueness and existence results, as well as numerical methods, are usually formulated for systems of first order ODE.
- In the context of ordinary differential equations, a *control system* is a system of first-order ODE of the form

$$\dot{x} = f(x, t, u)$$

where u, an unspecified function, is called the *input*, together with and *output function*

$$y = g(x, u, t).$$

In *open loop*, u is thought of as an external function of t, $u = u(t)$, while in *closed loop* u is driven by the state, $u = \beta(x)$.

Review of LTI systems (I)

- A linear time-invariant system of ODE is

$$\begin{aligned} \dot{x} &= Ax + Bu, \quad x \in \mathbb{R}^n, \ u \in \mathbb{R}^m \\ y &= Cx + Du, \quad y \in \mathbb{R}^p, \end{aligned}$$

  where A, B, C, D are constant matrices of appropriate sizes.

- State-transition matrix: $\Phi_A(t) = e^{At}$.

- Solution in the time domain:

$$y(t) = Ce^{At} \int_0^t e^{-A\tau} Bu(\tau) \, \mathrm{d}\tau + Ce^{At}x(0) + Du(t).$$

Review of LTI systems (II)

- Laplace-transformed system

$$\begin{aligned} sX(s) - x(0) &= AX(s) + BU(s), \quad x \in \mathbb{R}^n, \ u \in \mathbb{R} \\ Y(s) &= CX(s) + DU(s), \quad y \in \mathbb{R}. \end{aligned}$$

- Solution

$$Y(s) = H(s)U(s) + C(s\mathbb{I} - A)^{-1}x(0).$$

- Transfer function $H(s) = C(s\mathbb{I} - A)^{-1}B + D$.
- $H(s)$ is a matrix of proper rational functions of s, and it satisfies

$$\lim_{s \to \infty} H(s) = 0 \Leftrightarrow D = 0.$$

Definitions and notation (I)

- The general form of a partial differential equation (PDE) for a function $u(x_1, x_2, \ldots, x_n)$ is

$$F(x_1, x_2, \ldots, x_n, u, u_1, u_2, \ldots, u_n, u_{11}, u_{12}, \ldots) = 0$$

  where

$$u_{i_1 i_2 \ldots i_r} = \frac{\partial^r u}{\partial x_{i_1} \partial x_{i_2} \ldots x_{i_r}}$$

  is an rth order partial derivative of u. It is assumed that the number of partial derivatives appearing in F is finite.

- The order of the maximum partial derivative appearing in F is the order of the PDE.
- We will deal mainly with single PDE, and only occasionally consider systems of PDE.

Definitions and notation (II)

- In some cases, one of the independent variables represents time and is denoted by t, and the other variables are then, usually, associated to an n-dimensional spatial domain (so that one has $n+1$ independent variables). For instance, in an obvious notation,

  $$u_t = u_{xx} + u_{yy}$$

  is a second order PDE for a function $u(t, x, y)$ depending on time and 2 spatial variables.
- When time is involved, the PDE is accompanied by a set of *initial conditions*, specifying u and some of its time-derivatives at a given instant and for all the points in the spatial domain.
- Regardless of whether initial conditions are specified or not, *boundary conditions* can also be given, specifying the values of the solution and of some of its spatial partial derivatives at some spatial boundary for all values of time.

Well-posedness

- The fundamental theoretical question is whether the problem consisting of the PDE and its associated conditions (initial and/or boundary) is well-posed.
- A problem (and this is not restricted to PDE) is *well-posed* if it satisfies all of the following criteria (Hadamard, 1865-1963):
  - Existence. The problem has a solution.
  - Uniqueness. The problem has only one solution.
  - Stability. A small change in the equation or in the associated conditions gives rise to a small change in the solution.
- If any of the above criteria fails, the problem is called *ill-posed*.
- Most PDE coming from physics are well-posed, but some bad modeling decision can yield an ill-posed PDE.

Classification

- As said before, the order of a PDE is the order of the highest derivative appearing in it. Hence $u_t + u_{yy} + u_{xxxy} = 0$ is a fourth-order PDE.
- A PDE is called *linear* if F is a linear function of u and of all its derivatives. Hence

$$x^5 u_x + \sin x \; u_{tt} = 0 \qquad \text{is a linear second-order PDE.}$$
$$u_x u_y + u_{xxx} = 0 \qquad \text{is a nonlinear third-order PDE.}$$

- A PDE is called *quasilinear* if the highest-order derivatives appear linearly:

$$u_{xx} + u_{yy} = u^3 \qquad \text{is a quasilinear second-order PDE.}$$
$$u_x u_y + u_{xxx} = 0 \qquad \text{is a quasilinear third-order PDE.}$$
$$u_x u_{yy}^2 + u = 0 \qquad \text{is a nonlinear non quasilinear second-order PDE.}$$
$$u_x u_y + u = 0 \qquad \text{is a nonlinear non quasilinear first-order PDE.}$$

Strong and weak solutions

- A function has to be k times differentiable in order to be a solution of a kth order PDE. We define $C^k(D)$ as the set of all functions that have continuous derivatives up to order k in the domain D. In particular $C^0(D)$, or $C(D)$, denotes the set of continuous functions on D.
- A function in the set $C^k(D)$ that satisfies a kth order PDE is called a *strong* or *classical* solution of the PDE. Solutions of the PDE that do not satisfy this condition are called *weak solutions*.
- For a strong solution of a PDE to be a strong solution of the full problem (PDE+associated conditions) it must satisfy the associated conditions in a smooth way.

The heat equation (I)

- Let D be a fixed spatial domain and ∂D its boundary. Assume that the material in D is homogeneous and that the mass density and the heat capacity are constant in time. We scale them to 1 and identify hence internal energy density with temperature $u(x, y, z, t)$.
- The change in the energy stored in D between t and $t + \Delta t$ is

$$\int_D \left(u(x, y, z, t + \Delta t) - u(x, y, z, t)\right) \mathrm{d}V = \int_t^{t+\Delta t} \int_D q(x, y, z, t, u) \mathrm{d}V \ \mathrm{d}t - \int_t^{t+\Delta t} \int_{\partial D} \vec{B}(x, y, z, t) \cdot \mathrm{d}\vec{S} \ \mathrm{d}t,$$

  where q is the rate of heat production in D, and $\vec{B}$ is the heat flux through the boundary.

- The heat production is determined by external sources, although in some cases, such as in an air conditioner controlled by a thermostat, it may depend on the temperature itself. Hence we assume $q = q(x, y, z, t, u)$ but no dependence on the derivatives of u is considered.

The heat equation (II)

- The functional dependence of $\vec{B}$ on u can be determined from the experimental observation that heat flows from hotter to colder places. Mathematically this can be implemented by

$$\vec{B}(x,y,z,t) = -k(x,y,z)\vec{\nabla}u(x,y,z,t).$$

  This assumption is called Fourier's law of heat conduction, and k is the heat conduction coefficient, which should be a constant for an homogeneous material.

- The assumptions on the functional dependence of q and $\vec{B}$ on u are called *constitutive laws*.

- Using Fourier's law into the energy balance equation, approximating the time integrals using the mean value theorem and letting $\Delta t \to 0$

$$\int_D u_t \, \mathrm{d}V = \int_D q(x,y,z,t,u) \, \mathrm{d}V + \int_{\partial D} k(x,y,z)\vec{\nabla}u \cdot \mathrm{d}\vec{S}.$$

The heat equation (III)

- Use of the divergence (or Gauss) theorem allows one to convert the boundary integral into an integral of the divergence over the spatial domain, so finally

$$\int_D \left(u_t - q - \vec{\nabla} \cdot (k \vec{\nabla} u) \right) \mathrm{d}V = 0.$$

- It is easy to show that, if the integrand in the above expression is assumed to be a continuous function, and the equality to zero valid for any domain, then the integrand must be identically zero:

$$u_t = q + \vec{\nabla} \cdot (k \vec{\nabla} u).$$

This is a second-order PDE, and is linear if the dependence of q on u is linear.

The heat equation (IV)

- In the special case of no heat production and a constant k, one gets the classical heat equation

$$u_t = k\triangle u = k(u_{xx} + u_{yy} + u_{zz}).$$

- We have assumed that the function u and some of its derivatives are continuous functions. Since we do not know the solutions yet, this is a rather bold step, leading to classical or strong solutions.
- The integral energy balance obtained after using Fourier's law and Gauss theorem

$$\int_D \left(u_t - q - \vec{\nabla} \cdot (k\vec{\nabla}u)\right) \mathrm{d}V = 0.$$

provides a formulation more general than the one associated to the final PDE, one able to deal with non continuous solutions (or with non continuous derivatives), and hence with weak solutions.

An ill-posed problem (I)

- Consider

$$u_t = -u_{xx}$$

  which looks like a one-dimensional heat equation with $k = 1$ but with the wrong sign.

- Any problem associated to this PDE is ill-defined, because it violates the third requirement for well-posedness, namely stability with respect to small perturbations of the data.
- Indeed, consider an initial condition such that $u(x, 0) = u_0$ for all x. This corresponds to an uniform initial distribution, and it is easy to see that the unique solution is $u(x, t) = u_0$ for all x and all $t > 0$.
- Now consider a small perturbation of the initial data, in the form of a kink around $x = a$, *i.e.* $u(x, 0) > u_0$ if $x \in (a - \epsilon, a + \epsilon)$ and $u(x, 0) = u_0$ otherwise, in such a way that $u(x, 0)$ is nevertheless sufficiently smooth.

An ill-posed problem (II)

- From $u_t = -u_{xx}$, it is easy to see that those regions for which $u(x,0)$ is convex, *i.e.* $u_{xx}(x,0) < 0$, will evolve initially with increasing u, and thus making u_{xx} still more negative, and the other way around, reinforcing the effect.
- Hence, after some time the solution will differ considerably from the solution $u(x,t) = u_0$ corresponding to the uniform initial distribution, and eventually will run away.
- This does not depend on the size of the initial kink, and will spread in fact outside of the initially perturbed region, depending on the smoothness of $u(x,0)$ at $a \pm \epsilon$.
- Physically, this would describe heat flowing from colder to hotter places, something that is clearly unstable.
- Notice that one may think of the minus sign as a "correct" heat equation but integrated backwards in time.

Hydrodynamics

- The motion of a viscous fluid with constant density ρ and with no external forces can be described by a system of PDE for the velocity field $\vec{u}(x, y, z, t)$ and the pressure $p(x, y, z, t)$:

$$\begin{aligned} \vec{u}_t + (\vec{u} \cdot \vec{\nabla})\vec{u} &= \frac{\mu}{\rho}\triangle\vec{u} - \frac{1}{\rho}\vec{\nabla}p, \\ \vec{\nabla} \cdot \vec{u} &= 0, \end{aligned}$$

  called (specifically the first equation) the Navier-Stokes equation.

- This is a quasilinear system of second-order PDE, and is of foremost importance for many engineering applications involving fluid dynamics. No general closed-form solutions are known except for special cases, and numerical methods requiring enormous computational efforts are required to obtain approximate solutions.

- Despite its importance, the well-posedness of the Navier-Stokes PDE has not yet been established. This is essentially one of the Millennium Problems of the Clay Mathematics Institute, with a prize of $ 1 million.

The convection equation

- Many problems in chemistry, biology and geology involve the spread of some substrate being convected by a given velocity field $\vec{u}(x, y, z, t)$ associated, for instance, with the movement of a fluid.
- If the concentration of the substrate is denoted by $C(x, y, z, t)$, the PDE describing this is given by the convection equation

$$C_t + \vec{\nabla} \cdot (C\vec{u}) = 0.$$

- If $C = \rho$, the mass density of the fluid, one gets the mass transport equation $\rho_t + \vec{\nabla} \cdot (\rho\vec{u}) = 0$.
- Integrating the convection equation over a fixed spatial domain D and using Gauss theorem, one gets

$$\int_D C_t \, \mathrm{d}V = -\int_{\partial D} C\vec{u} \cdot \mathrm{d}\vec{S} \quad \text{or} \quad \frac{\mathrm{d}}{\mathrm{d}t}\int_D C \, \mathrm{d}V = -\int_{\partial D} C\vec{u} \cdot \mathrm{d}\vec{S}$$

expressing that the variation of the substance in the volume is due only to the flow through the boundary. If creation of the substance in the volume is allowed, an additional term must be included.

- In fact, as was the case for the heat equation, the integral formulation is more fundamental than the PDE one.

The diffusion equation

- Besides being convected by a fluid, a substrate can also vary its concentration by *diffusion*, which can be explained microscopically by the thermal movement of the molecules and has the macroscopic consequence that more molecules travel from higher concentration regions to lower concentration ones than the other way around.
- Fick's law of diffusion states that the flow of the substrate is then

$$\vec{q} = -D\vec{\nabla}C$$

  where $D > 0$ is the coefficient of diffusion.
- Assuming D to be constant, the diffusion equation is obtained:

$$C_t = D\triangle C,$$

  which is the same as the heat equation (with k constant).
- If both convection and diffusion are relevant, one gets

$$C_t = D\triangle C - \vec{\nabla}\cdot(C\vec{u}).$$

Vibrations of a string

- Consider a uniform one-dimensional string undergoing transversal motion whose amplitude is denoted by $u(x,t)$, where $x \in [0, L]$, and such an external force with density $f(x,t)$ acts on it.
- Under the assumption that the internal elastic forces of the string act only in the tangential direction one gets

$$u_{tt} - \frac{c^2}{\sqrt{1+u_x^2}} u_{xx} = \frac{f(x,t)}{\rho},$$

  where ρ is the (constant) density, and c is a constant that can be computed from ρ and the elasticity coefficient of the material.

- The above is a quasilinear second order PDE. Under the assumption of small slope movement, *i.e.* $|u_x| \ll 1$, one gets

$$u_{tt} - c^2 u_{xx} = \frac{1}{\rho} f(x,t).$$

- If no external forces are applied, the classical one-dimensional wave equation $u_{tt} - c^2 u_{xx} = 0$ is obtained.

Geometrical optics

- Consider the wave equation in space for a quantity $v(\vec{x}, t)$, and with non uniform wave speed $c(\vec{x})$:

$$v_{tt} - c^2(\vec{x})\triangle v = 0.$$

- Looking for solutions that are oscillatory in time, $v(\vec{x}, t) = e^{i\omega t}\psi(\vec{x})$, one gets for ψ

$$\triangle\psi + k^2 n^2(\vec{x})\psi = 0$$

where $k = \omega/c_0$ is the *wavenumber*, $n(\vec{x}) = c_0/c(\vec{x})$ is the *refraction index*, and c_0 is an average wave velocity in the medium.

- If solutions for ψ of the form

$$\psi(\vec{x}) = A(\vec{x}, k)e^{ikS(\vec{x})}$$

are seek, in the small wavelength limit $2\pi k^{-1} \to 0$ one gets for S, assuming that A is a bounded function of k,

$$|\vec{\nabla} S| = n(\vec{x}),$$

which is the *eikonal equation*, describing the geometrical optics limit of full electromagnetism, and postulated first by Hamilton in 1827.

Random motion

- Consider a particle in a two-dimensional region which in time δt travels a distance δr randomly and with equal probability in any direction, and such it "dies" upon reaching the boundary ∂D.
- What is the life expectancy $u(x, y)$ of a particle that starts life in (x, y)? Obviously, $u(x, y) = 0$ on ∂D.
- In the limit $\delta r \to 0$, $\delta t \to 0$ but $(\delta r)^2/(2\delta t) = k > 0$, one gets the two-dimensional *Poisson equation*

$$\triangle u = -\frac{1}{k}.$$

- This model has many applications, for instance in modeling stock prices. If a broker buys a stock at price x and decides to sell it if the price reaches a lower bound x_1 or an upper bound x_2, and the stock price is assumed to vary randomly, as a consequence of the cumulative effects of many variables, then the equation governing the time $u(x)$ that the broker holds the stock bought at price x is given by a one-dimensional version of the above PDE:

$$ku''(x) = -1, \quad u(x_1) = u(x_2) = 0.$$

The Laplace equation

- Many of the models presented so far include the Laplace operator

$$\triangle u = u_{xx} + u_{yy} + u_{zz}$$

  or any of its lower dimensional versions.

- Probably, the "most important" PDE is the Laplace equation

$$\triangle u = 0.$$

- Solutions to the Laplace equation are called *harmonic functions*. The equation was introduced first by Laplace in 1780 in relation to gravity, but its application reaches nearly any field of physics and engineering.

- For instance, the electrostatic potential $V(x, y)$ is a two-dimensional domain D without electrical charges satisfies

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0,$$

  together with boundary conditions specifying $V(x, y)$ or its normal derivative on ∂D.

The Schrödinger equation

- One of the fundamental equations of quantum mechanics was proposed by Schrödinger in 1926. For the wavefunction $\psi(x,t)$ of a particle of mass m moving in a one-dimensional potential $V(x)$, it reads

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\psi_{xx} + V\psi,$$

  where $\hbar$ is the rationalized Planck's constant $\hbar = (2\pi)^{-1}h$.

- Notice that due to the presence of i, the wavefunction solution to this PDE is complex. Its modulus squared $|\psi(x,t)|^2$ represents the probability density of founding the particle at x at time t.

- If $V = 0$ the above equation resembles a heat or diffusion equation. However the properties of the solutions are quite different due to the Schrödinger equation being a complex one.

Other models and summary

- There are many other PDE that appear in science and technology. To name just a few: Maxwell equations of electromagnetism, reaction-diffusion equations in chemical engineering, the biharmonic equation in elasticity, the Korteweg-de Vries equation for solitons, the Ginzburg-Landau equations of superconductivity, Einstein's equations of gravity, the telegrapher's equations for transmission lines, and many others.
- Some linear second-order differential operators have appeared several times. Disregarding constants and nonhomogeneous terms, they give rise to the three basic second-order PDE of mathematical physics:
  - the elliptic PDE $\triangle u = 0$, *i.e.* the Laplace equation.
  - the parabolic PDE $u_t = \triangle u$, *i.e.* the heat equation.
  - the hyperbolic PDE $u_{tt} = \triangle u$, *i.e.* the wave equation.
- The above equations and some related to them can be solved exactly on simple geometries. However, the general rule is that solution of PDE requires the use of numerical methods.

Initial conditions (I)

- Consider the convection equation in one spatial dimension for $C(x,t)$ and with velocity $u(x,t)$ given:

$$C_t + C_x u + C u_x = 0.$$

- It is natural to formulate a problem in which one gives the concentration at time t_0 and wants to find the concentration at later times. Hence one specifies $C(x,t_0) = C_0(x)$. This is an initial condition for a PDE of first order with respect to time.
- Another PDE for which it makes sense to impose initial conditions is the heat equation. In this case one gives the initial temperature profile $u(\vec{x},t_0) = u_0(\vec{x})$.

Initial conditions (II)

- For the string equation, which is of second-order with respect to time derivatives, one expects to have to specify both the initial displacement profile $u(x,0) = u_0(x)$ *and* the initial velocity profile $u_t(x,0) = v_0(x)$, as it is fit for an equation coming just from the application of Newton's second law.
- It can be shown that for the convection and the string equations, the PDE, together with the stated initial conditions, leads to well-posed problems, but this is not the case for the heat equation.

Boundary conditions (I)

- Boundary conditions are restrictions on the behavior of the solution and its spatial derivatives at the boundary of the spatial domain under consideration.
- Consider again the heat equation on a bounded spatial domain Ω

$$u_t = k\triangle u, \quad (x, y, z) \in \Omega,\ t > 0.$$

- It turns out that, in order to obtain a unique solution, in addition to the initial condition previously considered, one has to provide information on u at $\partial\Omega$ for all time.
- Excluding rare exceptions, there are three kinds of boundary conditions found in applications (for the heat equation as well as others).

Boundary conditions (II)

- The first kind, when the values of the temperature u at the boundary are provided

$$u(x, y, z, t) = f(x, y, z, t), \quad (x, y, z) \in \partial\Omega, \ t > 0,$$

  is called a *Dirichlet condition*. If the domain Ω is submerged in an isothermal bath, one would set $f = T_0$.

- Alternatively, one can supply the normal derivative of the temperature on the boundary, a quantity which is proportional to the heat flow through the boundary:

$$\partial_n u(x, y, z, t) = f(x, y, z, t), \quad (x, y, z) \in \partial\Omega, \ t > 0.$$

  This is called a *Neumann condition*. For an insulating boundary one would set $f = 0$.

Boundary conditions (III)

- A third situation involves a mixture of values of u and its normal derivative,

$$\alpha(x,y,z)u(x,y,z,t) + \beta(x,y,z)\partial_n u(x,y,z,t) = f(x,y,z,t)$$

  for $(x,y,z) \in \partial\Omega,\ t > 0$. This is a boundary condition of the third kind, sometimes also known as a *Robin condition*.

- Other situations are also possible. For instance, one can supply Dirichlet conditions on part of the boundary and Neumann conditions on the rest of it (*mixed boundary condition*), or one can specify conditions involving arbitrary, non tangent, derivatives on the boundary (*oblique boundary condition*), or even *nonlocal boundary conditions* relating values at the boundary with integrals of the the solution over all the domain.

Boundary conditions (IV)

- For the string PDE, although the initial conditions are enough to get a well-posed problem, it turns out that the addition of boundary conditions does not destroy the well-posedness, but just restricts the kind of solutions that are obtained to those corresponding to definite modes of vibration. This is of paramount importance, for instance, for the mathematical theory of music instruments or for the theory of electromagnetic waves in resonant cavities or in waveguides.
- Whatever the case, one should always keep in mind that one of the central issues is to know whether the PDE, together with the initial and boundary conditions, yields a well-posed problem. Elucidating this must usually be done on a case by case basis, and involves advanced tools from functional analysis.

Assignments for lecture 5

1. Let $p : \mathbb{R} \to \mathbb{R}$ be a differentiable function. Prove that the equation

$$u_t = p(u)u_x, \quad t > 0,$$

has solutions of the form $u(x,t) = f(x + p(u)t)$, with f an arbitrary differentiable function. In particular, find a solution for each of the following PDE: $u_t = ku_x$, $u_t = uu_x$, and $u_t = u\sin(u)\ u_x$.

2. Consider the equation $u_{xx} + 2u_{xy} + u_{yy} = 0$. Write it in the coordinates $s = x$, $t = x - y$ and find its general solution.

3. Solve the stock broker equation presented in the text (this is a trivial ODE), and compute the average time for which the broker holds the stock. Interpret the result in terms of x_1, x_2 and k.

Mathematical Methods – Lecture 6

First order PDE. The method of characteristics.

Carles Batlle Arnau

Departament de Matemàtica Aplicada 4
and
Institut d'Organització i Control de Sistemes Industrials

Universitat Politècnica de Catalunya

Lecture goals

- To present the method of characteristics for first-order quasilinear PDE, and the associated existence and uniqueness theorem.

Outline

- First-order quasilinear PDE.
- The method of characteristics.
- The existence and uniqueness theorem.
- Other methods and general nonlinear first-order PDE.

References

PR Pinchover, Y., & J. Rubinstein, *An introduction to partial differential equations*, Cambridge University Press, Cambridge, UK (2005), chapter 2.

GL Guenther, R.B., & J.W. Lee, *Partial differential equations of mathematical physics and integral equations*, Prentice-Hall, Englewood Cliffs, NJ, USA (1988), chapter 2.

First-order quasilinear PDE (I)

- A general first-order PDE in n variables is of the form

$$F(x_1, x_2, \ldots, x_n, u, u_1, u_2, \ldots, u_n) = 0.$$

- First-order PDE have many applications in physics and engineering, but nevertheless they appear less frequently than second-order ones.
- We will consider only the case of 2 independent variables

$$F(x, y, u, u_x, u_y) = 0.$$

- This is both for the sake of simplicity and because the method of characteristics can be best understood in this case, although it can be generalized to higher dimensions. A solution to this PDE can be visualized as a surface $z = u(x, y)$ in $\mathbb{R}^3$.

First-order quasilinear PDE (II)

- A quasilinear first-order PDE in two independent variables is of the form
$$a(x,y,u)u_x + b(x,y,u)u_y = c(x,y,u).$$
- The special case of linear first-order PDE is given by
$$a(x,y)u_x + b(x,y)u_y = c_0(x,y)u + c_1(x,y).$$
- To warm up, consider the linear, constant coefficient equation
$$u_x = c_0 u + c_1(x,y).$$
- The natural condition for a first-order PDE is a curve lying on the solution surface $u(x,y)$.
- Take, for instance, $u(0,y) = y$. This is a curve given on the section $x = 0$ of the surface.

First order quasilinear PDE (III)

- Since u_y does not appear in the PDE, we have in fact an ODE, with y parameterizing the set of initial conditions. The solution is

$$u(x,y) = e^{c_0 x}\left(\int_0^x e^{-c_0\xi} c_1(\xi,y)\mathrm{d}\xi + y\right).$$

- Notice that the PDE specifies the evolution along the x axis. This constraints in fact the selection of the curve of initial conditions. Consider for instance $c_1(x,y) = 0$ and change the initial curve to $u(x,0) = 2x$. The solution solution to the PDE is now $u(x,y) = e^{c_0 x}T(y)$, where $T(y)$ is determined from the initial conditions. But this yields $2x = u(x,0) = e^{c_0 x}T(0)$, which is clearly impossible.

- Also, with the same example with $c_1 = 0$, consider $u(x,0) = 2e^{c_0 x}$. Then $2e^{c_0 x} = e^{c_0 x}T(0)$, so that $2 = T(0)$. This means that we have a solution for any function $T(y)$ satisfying $T(0) = 2$.

- We conclude that the initial condition must be checked to guarantee existence and/or uniqueness. Notice also that for ODE, problems with existence and uniqueness are related to lack of smoothness of the functions appearing in the ODE, which is not the case for our example. Hence, this reflects genuine PDE phenomena.

The method of characteristics (I)

- The method of characteristics was developed by Hamilton in the middle of the nineteenth century when trying to solve the eikonal equation.
- The quasilinear equation $a(x,y,u)u_x + b(x,y,u)u_y = c(x,y,u)$ can be written as an scalar product:

$$(a(x,y,u), b(x,y,u), c(x,y,u)) \cdot (u_x, u_y, -1) = 0.$$

- Given a surface $z = u(x,y)$, at each point a set of two independent tangent vectors is given by $\vec{e}_x(x,y) = (1, 0, u_x(x,y))$ and $\vec{e}_y(0, 1, u_y(x,y)$. Hence, a normal vector at each point is given by

$$\vec{n}(x,y) = (u_x(x,y), u_y(x,y), -1).$$

- We can thus interpret the PDE as imposing that the vector $(a(x,y,u), b(x,y,u), c(x,y,u))$ must be in the tangent plane to the surface at each point $(x, y, z = u(x,y))$ of the surface.

The method of characteristics (II)

- Consider now a curve $(x(t), y(t), u(t)) \in \mathbb{R}^3$ and impose that it lays on $u(x, y)$. This means that the tangent vector to the curve must belong to the tangent plane at each point. This can be guaranteed if we impose

$$\begin{aligned} \dot{x} &= a(x, y, u), \\ \dot{y} &= b(x, y, u), \\ \dot{u} &= c(x, y, u). \end{aligned}$$

- This is a system of ODE, called the *characteristic equations* of the PDE, and the solutions are called *characteristic curves*.
- In order to determine an specific curve, we need an initial condition. For each initial condition we will get a solution curve, and we can parameterize the set of initial conditions by a parameter s such that

$$x(0, s) = x_0(s), \quad y(0, s) = y_0(s), \quad u(0, s) = u_0(s).$$

The method of characteristics (III)

- Notice that $(x_0(s), y_0(s), u_0(s))$ is a curve in $\mathbb{R}^3$, called the initial curve, parameterized by s.
- The set of solution curves is hence given by $x = x(t, s)$, $y = y(t, s)$, $z = u(t, s)$, and, under suitable conditions, this parameterizes a surface in $\mathbb{R}^3$, with parameters (t, s).
- The projection of a characteristic curve on the plane xy is called a *characteristic*.
- The problem consisting of the characteristic equations together with the initial curve is called the *Cauchy problem* for the quasilinear PDE.

The method of characteristics (IV)

- As a first example, consider

$$u_x + u_y = 2$$

with the initial condition $u(x, 0) = x^2$.

- The characteristic equations are $\dot{x} = 1$, $\dot{y} = 1$, $\dot{u} = 2$, and the initial condition can be parameterized as $x(0, s) = s$, $y(0, s) = 0$ and $u(0, s) = s^2$.
- The general solution to the characteristic equations is

$$x(t, s) = t + f_1(s), \quad y(t, s) = t + f_2(s), \quad u(t, s) = 2t + f_3(s),$$

and, upon imposing the initial conditions one gets the following parametric representation for the solution surface

$$x(t, s) = t + s, \quad y(t, s) = t, \quad u(t, s) = 2t + s^2.$$

The method of characteristics (V)

- In order to get an explicit solution $u(x, y)$, one needs to invert $x(t, s)$ and $y(t, s)$ and express (t, s) in terms of (x, y).
- This is easy in our case and one gets $t = y$, $s = x - y$, so that the explicit representation of the surface solution is

$$u(x, y) = 2y + (x - y)^2.$$

  Notice that for $y = 0$ one indeed gets $u(x, 0) = x^2$.
- From the simplicity of this example one must not conclude that each initial value problem for a quasilinear first-order PDE has a unique solution. We have in fact seen that this is not the case, even in the linear case.

The method of characteristics (VI)

There are several dragons lying hidden in the above procedure:

1. Even if the PDE is linear the characteristic equations are generally nonlinear. Hence, existence of solutions can only be guaranteed locally, and solutions to PDE can develop singularities in finite time even if the PDE is linear and perfectly smooth.
2. $x = x(t, s)$, $y = y(t, s)$ may be difficult to invert, or it can even do not define t and s and functions of x and y at each point.
3. A characteristic curve may intersect the initial curve more than once. There is then a potential conflict between the initial value at a point and the value at that point at a later time from the solution computed from another initial point.

The method of characteristics (VII)

- Let us explore in more detail the problem 2 enumerated above.
- For (t, s) to be a function of (x, y), even if not explicitly computable, it is necessary, from the inverse function theorem, that the Jacobian is different from zero:

$$J(t,s) = \begin{vmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial s} \\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial s} \end{vmatrix} \neq 0.$$

- An explicit computation using the characteristic equations and the initial curve shows, denoting derivation with respect to s by $'$, that

$$J(t,s) = \begin{vmatrix} a(x(t,s),y(t,s),u(t,s)) & x'(t,s) \\ b(x(t,s),y(t,s),u(t,s)) & y'(t,s) \end{vmatrix}.$$

- Hence, for the inversion to be possible, at least locally, it is necessary that the 2-vectors $(a(x(t,s),y(t,s),u(t,s)), b(x(t,s),y(t,s),u(t,s))$ and $(x'(t,s), y'(t,s))$ are linearly independent.

The method of characteristics (VIII)

- Consider t given and let s vary, so that we obtain a curve parameterized by s.
- The geometrical meaning of $J = 0$ at a given point of the surface corresponding to (t, s) is that the projection of the tangent vector to the curve $(x(t, s), y(t, s), u(t, s))$ on the plane xy, that is $(x'(t, s), y'(t, s))$, is on the same line that the projection on the same plane of the vector tangent to the characteristic curve, *i.e.* $(a(x(t, s), y(t, s), u(t, s)), b(x(t, s), y(t, s), u(t, s))$.
- For $t = 0$, this has the interpretation of representing a conflict between the information given by the initial curve and the information propagated by the characteristic curves.
- The condition $J \neq 0$ is called the *transversality condition*.

The method of characteristics (IX)

- As an (extremely trivial) example, consider $u_x = 1$ subject to the initial condition $u(0, y) = g(y)$.
- The system of characteristic equations is $\dot{x} = 1$, $\dot{y} = 0$, $\dot{u} = 1$, with general solution $x(t, s) = t + f_1(s)$, $y(t, s) = f_2(s)$, $u(t, s) = t + f_3(s)$.
- The initial curve is $x(0, s) = 0$, $y(0, s) = s$, $u(0, s) = g(s)$, and this yields the parameterized solution surface $x(t, s) = t$, $y(t, s) = s$, $u(t, s) = t + g(s)$. with explicit form $u(x, y) = x + g(y)$.
- On the other hand, if we choose the initial curve $u(x, 0) = h(x)$, *i.e.* $x(0, s) = s$, $y(0, s) = 0$, $u(0, s) = h(s)$, one gets $x(t, s) = t + s$, $y(t, s) = 0$, $u(t, s) = t + h(s)$, and now $(x(t, s), y(t, s))$ cannot be inverted. This could have been foreseen because

$$J = \begin{vmatrix} a & b \\ x' & y' \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 1 & 0 \end{vmatrix} = 0.$$

The method of characteristics (X)

As a slightly less trivial example, consider

$$3u_x + 5u_y = u,$$

with an initial curve $(s, 0, f(s))$ where f is arbitrary. We have to solve $\dot{x} = 3$, $\dot{y} = 5$, $\dot{u} = u$.
The solution satisfying the initial conditions is

$$\begin{aligned} x(t,s) &= 3t + s, \\ y(t,s) &= 5t, \\ u(t,s) &= f(s)e^t. \end{aligned}$$

From the first two equations we get $t = y/5$ and $s = x - 3y/5$, and the corresponding solution surface is

$$u(x,y) = f(x - \frac{3y}{5})e^{\frac{y}{5}}.$$

The existence and uniqueness theorem (I)

- Consider a quasilinear equation with initial conditions $\Gamma(s) = (x_0(s), y_0(s), u_0(s))$. Let us formulate the transversality condition specifically for the initial curve.
- We say that the equation and the initial curve satisfy the *transversality condition* at a point s on Γ if the characteristic emanating from the projection of $\Gamma(s)$ intersects the projection of Γ nontangentially, *i.e.*

$$J|_{t=0} = \dot{x}(0,s)y_0'(s) - \dot{y}(0,s)x_0'(s) = \begin{vmatrix} a & b \\ x_0' & y_0' \end{vmatrix} \neq 0.$$

The existence and uniqueness theorem (II)

Existence and uniqueness theorem for first-order quasilinear PDE

- Assume that the functions a, b, c are smooth functions of their variables in a neighborhood of the initial curve Γ.
- Assume further that the transversality condition holds at each point s in the interval $(s_0 - \delta, s_0 + \delta)$ on the initial curve.

Then the Cauchy problem has a unique solution in the neighborhood $(t, s) = (-\epsilon, \epsilon) \times (s_0 - \delta, s_0 + \delta)$ of the initial curve. Furthermore, if the transversality condition fails in all the points of an open interval around s_0, then the Cauchy problem has either no solution at all or it has infinitely many solutions. If the transversality condition does not hold in an isolated point s_0, the situation must be analyzed on a case by case basis.

Lagrange's method and general nonlinear first order PDE

- Besides the method of characteristics, there is another method, developed by Lagrange before Hamilton, which can also be applied for quasilinear first-order PDE. Its value is more historical than practical, except for the case of certain canonical equations (see section 2.3 of [PR] for more details).
- The method of characteristics can be generalized to deal with general nonlinear first order PDE.
- In the quasilinear case, the family of planes envelopes a unique line through (x_0, y_0, u_0) which determines the characteristic system of equations. In the nonlinear case, the family of planes envelopes a cone, called the Monge cone in honor of Gaspard Monge (1746-1818).
- See Section 2.9 of [PR] for further details.

Assignments for lecture 6

1. Solve $-yu_x + xu_y = u$ with $u(x, 0) = \psi(x)$.
2. (Exercise 2.10 in [PR]) A river is defined by the domain

$$D = \{(x, y) \mid |y| < 1, -\infty < x < +\infty\}.$$

A factory spills a contaminant into the river. The contaminant is further spread and convected by the flow in the river. The velocity field of the fluid in the river is only in the x direction. The concentration of the contaminant at a point (x, y) in the river and at time τ is denoted by $u(x, y, \tau)$. Conservation of matter and momentum implies that u satisfies the first-order PDE

$$u_\tau - (y^2 - 1)u_x = 0.$$

The initial condition is $u(x, y, 0) = e^y e^{-x^2}$.

1. Find the concentration u for all (x, y, τ).
2. A fish lives near the point $(x, y) = (2, 0)$ at the river. The fish can tolerate contaminant concentration levels up to 0.5. If the concentration exceeds this level, the fish will die at once. Will the fish survive? If yes, explain why. If no, find the time in which the fish will die.

Hint: Notice that y appears in the PDE just as a parameter.

Mathematical Methods – Lecture 7

Second-order PDE in two variables. Separation of variables for the heat and wave equations.

Carles Batlle Arnau

Departament de Matemàtica Aplicada 4
and
Institut d'Organització i Control de Sistemes Industrials

Universitat Politècnica de Catalunya

Lecture goals

- To present the classification of second-order linear PDE in two variables.
- To solve the $1+1$ heat equation using separation of variables and Fourier series.
- To solve the $1+1$ wave equation using separation of variables and Fourier series.
- To present the energy method as a tool for proving uniqueness.

Outline

- Classification of second-order linear PDE. Elliptic, parabolic and hyperbolic points.
- Canonical forms.
- Separation of variables for the heat equation in $1+1$ with Dirichlet conditions. Fourier expansion of the initial data.
- Separation of variables for the wave equation in $1+1$ with Neumann conditions. Fourier expansion of the initial data.
- The energy method.

References

PR Pinchover, Y., & J. Rubinstein, *An introduction to partial differential equations*, Cambridge University Press, Cambridge, UK (2005), chapters 3 and 5.

ABBP Antonijuan, J., C. Batlle, S. Boza, & J. d'Arc Prat, *Matemàtiques de la Telecomunicació*, Aula Politècnica 68, Edicions UPC (2001), capítol 7 (Sèries de Fourier).

Principal part of a second-order linear PDE

- Consider a general second-order linear PDE in two variables, $L[u] = g(x, y)$, where

$$\begin{aligned} L[u] &= a(x,y)u_{xx} + 2b(x,y)u_{xy} + c(x,y)u_{yy} \\ &+ d(x,y)u_x + e(x,y)u_y + f(x,y)u. \end{aligned}$$

  It is assumed that the coefficients a, b and c do not vanish at the same time anywhere.

- The *principal part* of L consists of the higher order terms:

$$L_0[u] = a(x,y)u_{xx} + 2b(x,y)u_{xy} + c(x,y)u_{yy}.$$

- The discriminant (of the principal part) of the operator is defined as

$$\delta(L)(x,y) = b^2(x,y) - a(x,y)c(x,y).$$

Type of a second-order linear PDE at a point (I)

- The type of a second order linear PDE at a point (x, y) is said to be
  - hyperbolic if $\delta(L)(x, y) > 0$.
  - parabolic if $\delta(L)(x, y) = 0$.
  - elliptic if $\delta(L)(x, y) < 0$.
- Given a non-empty open connected set $\Omega \in \mathbb{R}^2$, the PDE is said to be hyperbolic (resp. parabolic, elliptic) if its type is hyperbolic (resp.parabolic, elliptic) at all the points in Ω.
- The transformation $(\xi, \eta) = (\xi(x, y), \eta(x, y))$ is called a *change of coordinates* in Ω if its Jacobian

$$J(x, y) = \xi_x \eta_y - \xi_y \eta_x$$

does not vanish at any point in Ω.

Type of a second-order linear PDE at a point (II)

Theorem. The type of a linear second order PDE is invariant under changes of coordinates.

- To prove it, one has to use the chain rule to compute $w_{\xi\xi}$, $w_{\xi\eta}$, $w_{\eta\eta}$, w_ξ, w_η in terms of u_{xx}, u_{xy}, u_{yy}, u_x, u_y, where $w(\xi,\eta) = u(x(\xi,\eta), y(\xi,\eta))$.
- If A, B, C are the coefficients of the principal part in the new coordinates, one gets

$$\begin{pmatrix} A & B \\ B & C \end{pmatrix} = \begin{pmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{pmatrix}^T,$$

and the proof follows at once since then $B^2 - AC = J^2(b^2 - ac)$.

Type of a second-order linear PDE at a point (III)

- The types of the three fundamental equations of mathematical physics are
  - for the wave equation, $u_{tt} - c^2(x)u_{xx} = 0$, $a = 1$, $c = -c^2(x)$, $b = 0$, and $\delta = c^2(x)$, *i.e.* hyperbolic everywhere.
  - for the heat equation, $u_t - k(x)u_{xx} = 0$, $a = 0$, $c = -k(x)$, $b = 0$, and $\delta = 0$, *i.e.* parabolic everywhere.
  - for the Laplace equation, $u_{xx} + u_{yy} = 0$, $a = 1$, $c = 1$, $b = 0$, and $\delta = -1$, *i.e.* elliptic everywhere.
- The importance of these three equations stems both from their intrinsic importance in the description of diverse physical phenomena and the fact that any second-order linear PDE can, under a change of coordinates, be given a principal part which coincides with the one of the corresponding fundamental equation of the same type.

Canonical forms (I)

- The *canonical form of a hyperbolic equation* is

$$w_{\xi\eta} + l_1[w] = G(\xi, \eta),$$

where l_1 is a first-order linear differential operator.

- The *canonical form of a parabolic equation* is

$$w_{\xi\xi} + l_1[w] = G(\xi, \eta),$$

where l_1 is a first-order linear differential operator.

- The *canonical form of an elliptic equation* is

$$w_{\xi\xi} + w_{\eta\eta} + l_1[w] = G(\xi, \eta),$$

where l_1 is a first-order linear differential operator.

- Note that the principal part of the canonical form of the hyperbolic equation *is not* the wave operator.

Canonical forms (II)

- **Theorem**. Suppose that $L[u] = g(x, y)$ is of hyperbolic (resp. parabolic) type in the domain Ω. Then there exists in Ω a change of coordinates such that in the new coordinates the PDE has the canonical hyperbolic (resp. parabolic) form.
- **Theorem**. Suppose that $L[u] = g(x, y)$ is of elliptic type in the domain Ω and that the coefficients a, b, c are real analytic in Ω. Then there exists in Ω a change of coordinates such that in the new coordinates the PDE has the canonical elliptic form.
- The proofs of these theorems, as well as the actual construction of the changes of coordinates, involves solving pairs of first-order linear PDE for $\xi(x, y)$ and $\eta(x, y)$. This can be solved by the method of characteristics, and the corresponding *characteristics* are called the *characteristics of the second-order PDE*. In the elliptic case, these characteristics are in fact defined in the complex plane and do not exist as real ones (this is why we need a separate theorem for this case). See chapter 3 of [PR] for details and examples.

Canonical forms (III)

- As an example, consider the Euler-Tricomi equation, which appears in the study of transonic flows:

$$u_{xx} + xu_{yy} = 0.$$

  This PDE is hyperbolic for $x < 0$, and this is the case we will consider here.

- We will construct a change of coordinates $\xi = \xi(x, y)$, $\eta = \eta(x, y)$ such that the principal part of the PDE for $w(\xi, \eta)$ is $w_{\xi\eta}$. To this end, we have to impose that the other pieces of the principal part are zero.

- One gets that the $w_{\xi\xi}$ coefficient is $A = \xi_x^2 + x\xi_y^2$, and that of $w_{\eta\eta}$ is $C = \eta_x^2 + x\eta_y^2$. Hence we impose

$$0 = \xi_x^2 + x\xi_y^2 = (\xi_x + \sqrt{-x}\xi_y)(\xi_x - \sqrt{-x}\xi_y),$$

  and the same equation for η.

Canonical forms (IV)

- The previous PDE for ξ is nonlinear but decomposes as a product of two linear first-order PDE which can be solved by the method of characteristics.
- For the first factor

$$\xi_x + \sqrt{-x}\xi_y = 0$$

the characteristic system is $\dot{x} = 1$, $\dot{y} = \sqrt{-x}$, $\dot{\xi} = 0$. Hence ξ is constant on the characteristics. The equations for (x, y) can be rewritten as an ODE for the characteristic $y(x)$ as

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \sqrt{-x}, \quad \text{with solution} \quad \frac{3}{2}y + (-x)^{\frac{3}{2}} = \text{constant}.$$

Canonical forms (V)

- Since ξ is constant on these curves, it turns out that ξ is an arbitrary function of $\frac{3}{2}y + (-x)^{\frac{3}{2}}$, and we choose the simplest one:

$$\xi(x, y) = \frac{3}{2}y + (-x)^{\frac{3}{2}}.$$

- Notice that this means that $\xi(x, y)$ is not, in fact, a solution of the other linear factor. However η obeys the same quadratic PDE and we can choose it to satisfy this second linear factor, so as to obtain an independent quantity, as required for a change of coordinates.
- It is then immediate that η can be chosen as

$$\eta(x, y) = \frac{3}{2}y - (-x)^{\frac{3}{2}}.$$

Canonical forms (VI)

- The inverse change is given by

$$x(\xi,\eta) = -\left(\frac{\xi-\eta}{2}\right)^{2/3}, \quad y(\xi,\eta) = \frac{1}{2}(\xi+\eta).$$

- A simple computation shows then that the PDE becomes

$$0 = u_{xx} + xu_{yy} = -9\left(\frac{\xi-\eta}{2}\right)^{2/3}\left[w_{\xi\eta} - \frac{1}{6}\frac{w_\xi - w_\eta}{\xi-\eta}\right].$$

- Since the terms in $[\ ]$ must be zero, one gets indeed that the principal part in the new coordinates is that of the canonical form.
- Had we selected the arbitrary function of $\frac{3}{2}y \pm (-x)^{\frac{3}{2}}$ more carefully, the extra factors in the equation could have been disposed of.

Heat equation with homogeneous boundary conditions

- Consider the following heat conduction problem in a finite interval:

$$\begin{aligned} u_t - k u_{xx} &= 0, && 0 < x < L,\ t > 0, \\ u(0,t) = u(L,t) &= 0, && t \geq 0, \\ u(x,0) &= f(x), && 0 \leq x \leq L, \end{aligned}$$

  where f is a given initial condition (satisfying $f(0) = f(L) = 0$), and k is a positive constant.

- This problem represents a (one-dimensional) rod of length L with ends kept at zero temperature and whose initial temperature profile is known.

Separation of variables (I)

- We start by looking for solutions that satisfy the boundary conditions and have the special, separated variables form

$$u(x,t) = X(x)T(t).$$

  We exclude the trivial solution $u(x,t) = 0$, which could not satisfy the initial condition in any way.

- Differentiation and substitution into the PDE yields

$$X(x)T'(t) = kX''(x)T(t).$$

- Next we move all the dependence on t to the left:

$$\frac{1}{k}\frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)}.$$

Separation of variables (II)

- The left-hand side depends only on t, but the equation says that it is equal to the right-hand side, a function of x. Hence both sides must be constant, and we write

$$\frac{1}{k}\frac{T'(t)}{T(t)} = -\lambda, \quad \frac{X''(x)}{X(x)} = -\lambda,$$

  where λ, the *separation constant*, is a constant to be determined (by the boundary conditions, as it turns out), and the minus sign is arbitrary but convenient.

- We get hence the ODE system

$$X'' = -\lambda X,\ 0 < x < L, \quad T' = -\lambda k T,\ t > 0,$$

  coupled by the separation constant λ.

Separation of variables (III)

- Since $u(0,t) = X(0)T(t)$ and $u(L,t) = X(L)T(t)$ must be zero, any nontrivial solution implies $X(0) = 0 = X(L)$. Hence, we are left with the following two-point boundary problem for $X(x)$:

$$X'' + \lambda X = 0,\ 0 < x < L, \quad X(0) = X(L) = 0.$$

- A nontrivial solution of this problem is called an *eigenfunction* of the problem with *eigenvalue* λ. Notice that this is not a Cauchy problem, and it is not clear that there exists a solution for arbitrary values of λ. For instance, it can be shown that there is no solution for λ non real. Hence, we will consider only $\lambda \in \mathbb{R}$.
- For $\lambda < 0$ the solutions are exponentials, while for $\lambda = 0$ they are first-order polynomials; in either case, it is impossible to satisfy the boundary conditions except for the trivial solution $X(x) = 0$.

Separation of variables (IV)

- We are left then with the case $\lambda > 0$, for which

$$X(x) = \alpha \cos(\sqrt{\lambda} x) + \beta \sin(\sqrt{\lambda} x),$$

  with α, β arbitrary.

- Imposing $X(0) = 0$ sets $\alpha = 0$, and then $X(L) = 0$ boils down to $\sin(\sqrt{\lambda} L) = 0$, or $\sqrt{\lambda} L = n\pi$, with $n \in \mathbb{Z}$.

- Since negative values of n give rise to the same eigenfunctions, we will consider only the eigenvalues

$$\lambda = \left(\frac{n\pi}{L}\right)^2, \ n = 1, 2, \ldots, \text{ with eigenfunctions } X(x) = \sin \frac{n\pi x}{L}.$$

Separation of variables (V)

- Summing up, the set of all solutions for the spatial part of the heat equation with homogeneous Dirichlet boundary conditions is spanned by

$$X_n(x) = \sin \frac{n\pi x}{L}, \ \lambda_n = \left(\frac{n\pi}{L}\right)^2, \ n = 1, 2, 3, \ldots$$

- Let us now return to the time-dependent part. The general solution to the ODE $T' = -k\lambda T$ is $T(t) = Be^{-k\lambda t}$, with B arbitrary. Notice that, from the physical point of view, it is deduced again that λ must be positive, since the solutions should decay in time.
- Inserting the allowed values of λ one obtains the set

$$T_n(t) = B_n e^{-k\left(\frac{n\pi}{L}\right)^2 t}, \ n = 1, 2, 3, \ldots$$

Separation of variables (VI)

- Putting everything together, we have the sequence of separated solutions

$$u_n(x,t) = B_n \sin \frac{n\pi x}{L} e^{-k\left(\frac{n\pi}{L}\right)^2 t}, \ n = 1, 2, 3, \ldots$$

- It is obvious that a finite sum of solutions of this type cannot satisfy an arbitrary initial condition, except if it is itself a finite combination of $\sin \frac{n\pi x}{L}$. However, the theory of Fourier series can be invoked to write down a formal solution

$$u(x,t) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi x}{L} e^{-k\left(\frac{n\pi}{L}\right)^2 t}.$$

- This is a formal solution in the sense that it involves a series, and the question of the smoothness of the result, at least to order 2 in x and order 1 in t, must be studied. We will, however, proceed with this solution for the time being.

Separation of variables (VII)

- Imposing the initial condition one gets

$$f(x) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi x}{L}.$$

- Under suitable conditions, the theory of Fourier series establishes that, given a periodic function $F(x)$ with period T, it can be represented by the series

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos \frac{2\pi n x}{T} + \sum_{n=1}^{\infty} b_n \sin \frac{2\pi n x}{T},$$

where the coefficients are computed as (the integrals can be in fact computed over any interval of length T)

$$\begin{aligned} a_0 &= \frac{2}{T} \int_0^T F(x)\, \mathrm{d}x, \\ a_n &= \frac{2}{T} \int_0^T F(x) \cos \frac{2\pi n x}{T} \mathrm{d}x, \ n = 1, 2, 3, \ldots \\ b_n &= \frac{2}{T} \int_0^T F(x) \sin \frac{2\pi n x}{T} \mathrm{d}x, \ n = 1, 2, 3, \ldots \end{aligned}$$

Separation of variables (VIII)

- In order to use the Fourier results to compute the B_n from $f(x) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi x}{L}$, we rewrite the later as

$$f(x) = \sum_{n=1}^{\infty} B_n \sin \frac{2n\pi x}{2L}$$

  which indicates that we have the sine, or odd, part of the Fourier series of a function of period $2L$. However, we face the problem of $f(x)$ being defined only on $[0, L]$, and also that the cosine, or even, part is missing.

- Both problems can be solved at once introducing the odd-extension f_O of f from $[0, L]$ to $[-L, L]$:

$$f_O(x) = \begin{cases} f(x) & x \in [0, L], \\ -f(-x) & x \in [-L, 0). \end{cases}$$

Separation of variables (IX)

- The function f_O coincides with f on the interval of interest $[0, L]$. Furthermore, for an odd-symmetric function as f_O, all the a_n coefficients are zero, and this is consistent with the expansion for $f(x)$. Finally, for $x \in [0, L]$,

$$f(x) = f_O(x) = \sum_{n=1}^{\infty} b_n \sin \frac{2n\pi x}{2L}$$

  so we get that $B_n = b_n$ for $n = 1, 2, \ldots$

- The B_n are thus computed as

$$B_n = b_n = \frac{2}{2L} \int_{-L}^{L} f_O(x) \sin \frac{2\pi n x}{2L} \mathrm{d}x = \frac{1}{L} \int_{-L}^{L} f_O(x) \sin \frac{\pi n x}{L} \mathrm{d}x.$$

Separation of variables (X)

- Using the odd-symmetry of f_O and the sine function, the integral over $[-L, L]$ equals twice the integral over $[0, L]$, where furthermore f_O coincides with f, and we get

$$B_n = \frac{2}{L} \int_0^L f(x) \sin \frac{\pi n x}{L} \mathrm{d}x,$$

  which is the final expression for the coefficients of the series expansion. This way, given an initial temperature profile, the solution is completely specified, at least formally.

- As an example, consider $L = \pi$, $k = 1$ and the tent-like, nonsmooth profile

$$f(x) = \begin{cases} x & 0 \leq x \leq \pi/2, \\ \pi - x & \pi/2 \leq x \leq \pi. \end{cases}$$

Separation of variables (XI)

- The series solution is

$$u(x,t) = \sum_{n=1}^{\infty} B_n \sin nx \; e^{-n^2 t}.$$

- One gets

$$\begin{aligned} B_n &= \frac{2}{\pi}\int_0^{\pi} f(x)\sin nx \; \mathrm{d}x = \frac{2}{\pi}\int_0^{\pi/2} x \sin nx \; \mathrm{d}x + \frac{2}{\pi}\int_{\pi/2}^{\pi} (\pi - x)\sin nx \; \mathrm{d}x \\ &= \frac{4}{\pi n^2}\sin\frac{n\pi}{2}. \end{aligned}$$

- But

$$\sin\frac{n\pi}{2} = \begin{cases} 0 & \text{if } n = 2m, \\ (-1)^{m+1} & \text{if } n = 2m-1. \end{cases}$$

- Therefore, we obtain the formal solution

$$u(x,t) = \frac{4}{\pi}\sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{(2m-1)^2}\sin[(2m-1)x] \; e^{-(2m-1)^2 t}.$$

Separation of variables (XII)

- It can be shown that this solution is in fact a classical or strong solution, *i.e.* it can be differentiated once with respect to time and twice with respect to space.
- In fact, $u(x,t)$ is a $C^\infty((0,L)\times(0,\infty))$ function. The lack of smoothness of the initial data disappears immediately: heat conduction, as any other diffusion phenomena, irons out any initial irregularities. This effect is known to hold also in more general parabolic problems, in contrast with the hyperbolic case, where singularities propagate along characteristics and persist in time.
- Notice that

$$\lim_{t\to+\infty} u(x,t) = 0,$$

something that is physically immediate from the boundary conditions.

A string with clamped but free ends

Consider the initial and boundary value PDE problem

$$\begin{aligned} u_{tt} - c^2 u_{xx} &= 0, && 0 < x < L,\ t > 0, \\ u_x(0,t) = u_x(L,t) &= 0, && t \geq 0, \\ u(x,0) &= f(x), && 0 \leq x \leq L, \\ u_t(x,0) &= g(x), && 0 \leq x \leq L, \end{aligned}$$

corresponding to a string of length L with no transverse force at the free ends, and with initial position and velocity profiles $f(x)$ and $g(x)$, satisfying not only $f'(0) = f'(L) = 0$, but also $g'(0) = g'(L) = 0$, since the boundary conditions are imposed for all $t \geq 0$ and hence their time derivative must be also zero.

Separation of variables (XIII)

- We will apply again the method of separation of variables, *i.e.* we will try to satisfy the boundary conditions by means of nontrivial solutions of the form

$$u(x,t) = X(x)T(t).$$

- The same arguments used for the heat equation lead immediately to the pair of ODE $X'' = -\lambda X$, for $0 < x < L$ and $T'' = -\lambda c^2 T$ for $t > 0$.
- Taking into account the boundary conditions, it turns out that the function X must be a solution of the eigenvalue problem

$$X'' + \lambda X = 0,\ 0 < x < L, \quad X'(0) = 0,\ X'(L) = 0.$$

Separation of variables (XIV)

- Again, non real values of λ, as well as negative ones, can be discarded. This time, however, $\lambda = 0$ is a valid eigenvalue, with constant eigenfunction $X_0(x) = 1$ (or any other constant). This is a consequence of the Neumann conditions considered; this zero-mode would have not been obtained with Dirichlet conditions, just like as in the heat equation.
- For $\lambda > 0$ the general solution of the ODE is

$$X(x) = \alpha \cos(\sqrt{\lambda}x) + \beta \sin(\sqrt{\lambda}x).$$

- Imposing the boundary conditions selects again $\lambda = \left(\frac{n\pi}{L}\right)^2$, $n = 1, 2, 3, \ldots$ but now with eigenfunctions

$$X(x) = \cos \frac{n\pi x}{L}.$$

Separation of variables (XV)

- We can write together the $\lambda = 0$ and $\lambda > 0$ cases as

$$X_n(x) = \cos\frac{n\pi x}{L}, \quad \lambda_n = \left(\frac{n\pi}{L}\right)^2, n = 0, 1, 2, 3, \ldots$$

- The corresponding solutions for the time-dependent part are

$$\begin{aligned} T_0(t) &= \frac{A_0 + B_0 t}{2}, \\ T_n(t) &= A_n \cos\frac{c\pi n t}{L} + B_n \sin\frac{c\pi n t}{L}, \ n = 1, 2, 3, \ldots, \end{aligned}$$

where the $1/2$ factor in T_0 has been chosen for convenience.

Separation of variables (XVI)

- Finally, the formal series solution satisfying the boundary conditions is

$$u(x,t) = \frac{A_0 + B_0 t}{2} + \sum_{n=1}^{\infty} \left(A_n \cos \frac{c\pi n t}{L} + B_n \sin \frac{c\pi n t}{L} \right) \cos \frac{n\pi x}{L}.$$

- Since

$$\begin{aligned} \cos \frac{c\pi n t}{L} \cos \frac{n\pi x}{L} &= \frac{1}{2} \left(\cos \frac{n\pi}{L}(ct+x) + \cos \frac{n\pi}{L}(ct-x) \right), \\ \sin \frac{c\pi n t}{L} \cos \frac{n\pi x}{L} &= \frac{1}{2} \left(\sin \frac{n\pi}{L}(ct+x) + \sin \frac{n\pi}{L}(ct-x) \right), \end{aligned}$$

and

$$\frac{A_0 + B_0 t}{2} = \left[\frac{A_0}{4} + \frac{B_0}{2c}(ct+x) \right] + \left[\frac{A_0}{4} + \frac{B_0}{2c}(ct-x) \right],$$

we see that $u(x,t)$ is, formally, of the form $F(ct+x) + G(ct-x)$, *i.e.* the sum of functions of the characteristics $ct \pm x$ of the wave equation $u_{tt} - c^2 u_{xx} = 0$, which is its general solution, made of forward and backward moving arbitrary profiles.

Separation of variables (XVII)

- It remains to find the coefficients A_0, B_0, A_n, B_n from the initial conditions.
- At $t = 0$ we have, first,

$$f(x) = u(x, 0) = \frac{A_0}{2} + \sum_{n=1}^{\infty} A_n \cos \frac{n\pi x}{L}.$$

- In order to use the Fourier series results, we have to extend again $f(x)$ to $[-L, L]$, but this time with even symmetry, in order to get zero coefficients for the sine terms, not present in $f(x)$. Repeating the same steps carried out for the heat equation, the final result is

$$\begin{aligned} A_0 &= \frac{2}{L} \int_0^L f(x) \mathrm{d}x, \\ A_n &= \frac{2}{L} \int_0^L \cos \frac{n\pi x}{L} f(x)\, \mathrm{d}x, \ n = 1, 2, 3, \ldots \end{aligned}$$

Separation of variables (XVIII)

- Deriving $u(x,t)$ with respect to time and setting $t=0$ one gets

$$g(x) = u_t(x,0) = \frac{B_0}{2} + \sum_{n=1}^{\infty} B_n \frac{c\pi n}{L} \cos \frac{n\pi x}{L}.$$

- Hence, the B_0 and B_n can be computed from the Fourier series of the even symmetry extended $g_E(x)$. This way one gets, taking into account the extra factor $\frac{c\pi n}{L}$,

$$\begin{aligned} B_0 &= \frac{2}{L}\int_0^L g(x)\mathrm{d}x, \\ B_n &= \frac{2}{c\pi n}\int_0^L \cos\frac{n\pi x}{L} g(x)\ \mathrm{d}x,\ n=1,2,3,\ldots \end{aligned}$$

- This solves the problem formally for any initial position and velocity profiles. Notice that for the wave equation one does not have the decreasing exponentials in time, which cause any nonsmoothness of the initial profile to disappear for $t>0$: hyperbolic evolution preserves the singularities of the initial data, and the question of whether the obtained solution is a classical one is more delicate.

The energy method (I)

- The energy method has many applications on the theory of PDE, but we will use it here in the framework of proving uniqueness of the solution to an initial and boundary value problem for a PDE.
- It is inspired by the physical principle of energy conservation, although it is applied to functions which actually may not be the energy of the system, or even applied for nonphysical systems.
- The basic idea of the method is as follows. For certain homogeneous problems it is possible to define a function ("the energy") that is nonnegative and nonincreasing (as a function of t). If, in addition, the energy is zero at $t = 0$ it will remain zero for all $t > 0$. If the only zero of the energy corresponds to a zero solution, it follows that the solution is zero for all $t \geq 0$. If this is applied to the difference of two solutions, one gets that the two solutions are actually the same.
- Instead of giving a general formulation, we will illustrate the method with an example.

The energy method (II)

- Consider the Neumann problem for the vibrating string

$$u_{tt} - c^2 u_{xx} = F(x,t), \quad 0 < x < L,\ t > 0,$$

with boundary conditions $u_x(0,t) = a(t)$, $u_x(L,t) = b(t)$, for $t \geq 0$, and initial ones $u(x,0) = f(x)$, $u_t(x,0) = g(x)$, for $0 \leq x \leq L$.

- Let u_1, u_2 be two solutions of the problem. Then the function $w = u_1 - u_2$ is a solution of the homogeneous problem

$$w_{tt} - c^2 w_{xx} = 0, \quad 0 < x < L,\ t > 0,$$

with boundary conditions $w_x(0,t) = 0$, $w_x(L,t) = 0$, for $t \geq 0$, and initial ones $w(x,0) = 0$, $w_t(x,0) = 0$, for $0 \leq x \leq L$.

The energy method (III)

- Define the (total) energy of the solution w at time t as

$$E[w](t) = \frac{1}{2}\int_0^L \left(w_t^2(x,t) + c^2 w_x^2(x,t)\right) \mathrm{d}x.$$

- Using the PDE, one has

$$\frac{\mathrm{d}}{\mathrm{d}t}E = \int_0^L \left(w_t w_{tt} + c^2 w_x w_{xt}\right) \mathrm{d}x = c^2 \int_0^L \left(w_t w_{xx} + w_x w_{xt}\right) \mathrm{d}x.$$

- The term inside the integral is a total derivative in x and so

$$\frac{\mathrm{d}}{\mathrm{d}t}E = c^2 \int_0^L \frac{\partial}{\partial x}\left(w_x w_t\right) \mathrm{d}x = c^2 \left(w_x w_t\right)\Big|_{x=0}^{x=L}.$$

The energy method (IV)

- The boundary conditions $w_x(0,t) = 0$, $w_x(L,t) = 0$ imply that $\dot{E}(t) = 0$ and hence $E(t) = E(0)$. But

$$E(0) = \frac{1}{2}\int_0^L \left(w_t^2(x,0) + c^2 w_x^2(x,0)\right) \mathrm{d}x,$$

  and, due to the initial conditions $w(x,0) = 0$ (which implies $w_x(x,0) = 0$) and $w_t(x,0) = 0$, one gets $E[w](t) = 0$ for all $t \geq 0$.

- Now the integrand $e(x,t) = w_t^2 + c^2 w_x^2$ is nonnegative and, since its integral over $[0, L]$ is zero, it follows that $w_t^2(x,t) + c^2 w_x^2(x,t) = 0$, which in turn implies $w_t(x,t) = 0$ and $w_x(x,t) = 0$. Hence $w(x,t)$ is constant for all $x \in [0, L]$ and $t \geq 0$.

- Finally, from the initial condition $w(x,0) = 0$ one gets $w(x,t) = 0$ and $u_1(x,t) = u_2(x,t)$, showing that the two solutions are actually the same and completing the uniqueness proof.

Assignments for lecture 7

1. (Exercise 5.3 in [PR]) Using the separation of variables method find a (formal) solution of a vibrating string with fixed ends:

$$\begin{aligned} u_{tt} - c^2 u_{xx} &= 0, \quad 0 < x < L, 0 < t, \\ u(0,t) = u(L,t) &= 0, \quad t \geq 0, \\ u(x,0) &= f(x), \quad 0 \leq x \leq L, \\ u_t(x,0) &= g(x), \quad 0 \leq x \leq L. \end{aligned}$$

Prove that the solution can be represented as a superposition of a forward and a backward wave.

Mathematical Methods – Lecture 8

Elliptic equations. Separation of variables for the Laplace equation

Carles Batlle Arnau

Departament de Matemàtica Aplicada 4
and
Institut d'Organització i Control de Sistemes Industrials

Universitat Politècnica de Catalunya

Lecture goals

- To present some basic properties of elliptic PDE and why initial value problems are not well-defined for them.
- To present the maximum principle and Green's identities, and what they imply for the solution of elliptic PDE.
- To discuss separation of variables for elliptic problems, and how to apply it to rectangular and circular domains.

Outline

- Basic properties of elliptic problems.
- The maximum principle and its application to the Dirichlet problem.
- Green's identities and their application to the Neumann problem.
- Separation of variables for elliptic problems. Rectangular and circular domains.

References

PR Pinchover, Y., & J. Rubinstein, *An introduction to partial differential equations*, Cambridge University Press, Cambridge, UK (2005), chapter 7.

Elliptic problems (I)

- We will consider the Laplace equation in two variables in a planar domain (non-empty, open connected set)

$$\Delta u \equiv u_{xx} + u_{yy} = 0,\ (x, y) \in \Omega \subset \mathbb{R}^2,$$

whose solutions are called *harmonic functions* in Ω, as well as the Poisson equation

$$\Delta u = F(x, y),\ (x, y) \in \Omega \subset \mathbb{R}^2.$$

- The problem defined by the Poisson equation and Dirichlet boundary conditions $u(x, y) = g(x, y)$ on $\partial\Omega$ is called, specifically, the Dirichlet problem.
- The problem defined by the Poisson equation and Neumann boundary conditions $\partial_n u(x, y) = g(x, y)$ on $\partial\Omega$ is called, specifically, the Neumann problem.

Elliptic problems (II)

- One can also consider mixed boundary problems, called specifically Robin problems in the context of elliptic PDE.
- The question of existence of solutions to each of these problems is not easy, specially when non smooth boundaries are considered. In this lecture we will consider only classical solutions, *i.e.* $u \in C^2(\Omega)$.
- **Lemma**. A necessary condition for the existence of a solution to the Neumann problem is

$$\int_{\partial\Omega} g(x(s), y(s))\ \mathrm{d}s = \int_{\Omega} F(x, y)\ \mathrm{d}x\mathrm{d}y,$$

where $(x(s), y(s))$ is a parametrization of the closed curve $\partial\Omega$.

Elliptic problems (III)

- In order to prove this result, consider Gauss's (or divergence) theorem in two dimensions

$$\int_{\Omega} \vec{\nabla} \cdot \vec{\psi}(x,y) \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} \vec{\psi}(x(s),y(s)) \cdot \vec{n}(s)\ \mathrm{d}s,$$

  where $(x(s), y(s))$ is a parametrization of the boundary $\partial\Omega$ and $\vec{n}$ is an unitary outwards normal. The theorem holds for any $\vec{\psi} \in C^1(\Omega) \cap C(\bar{\Omega})$and any bounded piecewise smooth domain Ω.

- Let $\vec{\Psi} = \vec{\nabla} u$. Using that $\vec{\nabla} \cdot \vec{\nabla} u = \Delta u$ and that $\vec{\nabla} u \cdot \vec{n} = \partial_n u$, one gets

$$\int_{\Omega} \Delta u(x,y) \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} \partial_n u\ \mathrm{d}s$$

  and the result follows from $\Delta u = F$ on Ω and $\partial_n u = g$ on $\partial\Omega$.

Elliptic problems (IV)

- Notice that, as a special case of the above property, for harmonic functions, *i.e.* solutions to $\Delta u = 0$, one has, for any boundary condition

$$\int_\Gamma \partial_n u \, \mathrm{d}s = 0,$$

where Γ is $\partial\Omega$ or any closed curve contained in Ω.

- Initial value problems are not well defined for elliptic PDE, *i.e.* one gets into problems if one considers y to be a time-like coordinate and specifies, for instance, $u(x,0)$ and $u_y(x,0)$.
- As an example, consider $u_{xx} + u_{yy} = 0$ on the upper half plane $-\infty < x < +\infty$, $y > 0$, with initial conditions, depending on an arbitrary integer $n > 0$:

$$u^n(x,0) = 0, \quad u^n_y(x,0) = \frac{\sin nx}{n}, \quad -\infty < x < \infty.$$

Elliptic problems (V)

- It is easy to check that

$$u^n(x,y) = \frac{1}{n^2}\sin nx \sinh ny$$

  is a harmonic function on the upper half-plane.

- Choosing n very large, the initial condition is arbitrarily close to $u^0(x,0) = 0$, $u^0_y(x,0) = 0$, with trivial solution $u(x,y) = 0$.

- On the other hand, for any $y > 0$, the solution grows exponentially fast as $n \to \infty$. Hence, the Cauchy problem for the Laplace equation is not stable with respect to the stated initial conditions and is not well posed.

Elliptic problems (VI)

- A *harmonic polynomial of degree* n is a harmonic function $P_n(x, y)$ of the form

$$P_n(x, y) = \sum_{0 \leq i+j \leq n} a_{i,j} x^i y^j.$$

- A *homogeneous harmonic polynomial of degree* n is a harmonic function $P_n(x, y)$ of the form

$$P_n(x, y) = \sum_{i+j=n} a_{i,j} x^i y^j.$$

If we consider the set of harmonic homogeneous polynomials of degree equal to n as a vector space over $\mathbb{R}$, this subspace, called V_n, is of dimension 2 for any n (this holds only for two variables). For instance,

$$V_1 = \langle x, y \rangle, \; V_2 = \langle xy, x^2 - y^2 \rangle, \; V_3 = \langle x^3 - 3xy^2, y^3 - 3yx^2 \rangle.$$

Elliptic problems (VII)

- The most important solution of the Laplace equation is the one symmetric around the origin. To compute it in two dimensions, we express the Laplacian in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$, with $w(r,\theta) = u(x(r,\theta), y(r,\theta))$. One gets

$$\Delta w = w_{rr} + \frac{1}{r} w_r + \frac{1}{r^2} w_{\theta\theta} = 0.$$

- The radial symmetric solution $w(r)$ satisfies thus

$$w'' + \frac{1}{r} w' = 0,$$

which is a linear second-order ODE, called the Euler equation, with two fundamental solutions, namely $w(r) = 1$ and

$$w(r) = -\frac{1}{2\pi} \log r.$$

This is called *the fundamental solution of the Laplace equation*.

Weak form of the maximum principle

- THE WEAK MAXIMUM PRINCIPLE. Let Ω be a bounded domain, $\bar{\Omega} = \Omega \cup \partial\Omega$, and let $u(x, y) \in C^2(\Omega) \cap C(\bar{\Omega})$ be an harmonic function on Ω. Then the maximum of u in $\bar{\Omega}$ is achieved on the boundary $\partial\Omega$.
- Comments:
  - Since $\min_A u = -\max_A(-u)$, and since, if u is harmonic in Ω so is $-u$, the minimum is also attained on the boundary.
  - The result does not exclude that the maximum is attained also at an interior point.
  - he result can be extended to a large class of elliptic problems.
  - The proof is based on the elementary result that on a local maximum of v, $\Delta v \leq 0$ (see Section 7.3 of [PR]).

Applications of the maximum principle (I)

Theorem. The Dirichlet problem in a bounded domain Ω

$$\begin{aligned} \Delta u &= f(x,y), && (x,y) \in \Omega, \\ u(x,y) &= g(x,y), && (x,y) \in \partial\Omega, \end{aligned}$$

has at most one solution in $C^2(\Omega) \cap C(\bar{\Omega})$.

PROOF. Assume that u_1 and u_2 are solutions. Then $v = u_1 - u_2$ is a harmonic function satisfying $v = 0$ on $\partial\Omega$. By the weak maximum (and minimum) principle, one has $0 \leq v(x,y) \leq 0$ on Ω and hence $v = 0$ on Ω.

The maximum principle can also be used to prove results for the heat equation (see 7.6 of [PR]).

Applications of the maximum principle (II)

Theorem. Let Ω be a bounded domain and let u_1, u_2 be functions in $C^2(\Omega) \cap C(\bar{\Omega})$ that solve $\Delta u = f$ with Dirichlet conditions g_1 and g_2, respectively. Let $M_g = \max_{\partial\Omega} |g_1(x,y) - g_2(x,y)|$. Then

$$\max_{(x,y)\in\Omega} |u_1(x,y) - u_2(x,y)| \leq M_g.$$

PROOF. Let $v = u_1 - u_2$, which is harmonic in Ω satisfying $v = g_1 - g_2$ in $\partial\Omega$. From the weak form of the maximum (and minimum) principle

$$\min_{\partial\Omega}(g_1 - g_2) \leq v(x,y) \leq \max_{\partial\Omega}(g_1 - g_2), \ \forall\ (x,y) \in \Omega,$$

and the theorem follows immediately from $|B| \leq b \Leftrightarrow -b \leq B \leq b$ and $\min_A u = -\max_A(-u)$.

Green's identities (I)

- Consider again Gauss's theorem in two dimensions

$$\int_{\Omega} \vec{\nabla} \cdot \vec{\psi}(x,y) \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} \vec{\psi}(x(s),y(s)) \cdot \vec{n}(s)\ \mathrm{d}s.$$

- Selecting $\vec{\psi} = \vec{\nabla} u$ leads to

$$\int_{\Omega} \Delta u\ \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} \partial_n u\ \mathrm{d}s,$$

which is *Green's first identity*, as was already used to prove a necessary condition for Neumann's problem.

Green's identities (II)

- Selecting instead $\vec{\psi} = v\vec{\nabla}u - u\vec{\nabla}v$ obtains *Green's second identity*

$$\int_\Omega (v\Delta u - u\Delta v)\ \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} (v\,\partial_n u - u\,\partial_n v)\ \mathrm{d}s.$$

  (Notice that there is a cancelation of the $\vec{\nabla}u \cdot \vec{\nabla}v$ terms in the left-hand side).

- Finally, setting $\vec{\psi} = v\vec{\nabla}u$, the above mentioned cancelation disappears and we get *Green's third identity*

$$\int_\Omega \vec{\nabla}u \cdot \vec{\nabla}v\ \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} v\,\partial_n u\ \mathrm{d}s - \int_\Omega v\Delta u\ \mathrm{d}x\mathrm{d}y,$$

  which will be used to prove uniqueness results for the Poisson equation.

Uniqueness theorem for Poisson's equation (I)

- **Theorem**. Let Ω be a smooth domain and consider the problems associated to the Poisson equation.
  1. The Dirichlet problem has at most one solution.
  2. For the Robin problem $u(x,y) + \alpha(x,y)\partial_n u(x,y) = g(x,y)$, $(x,y) \in \partial\Omega$, if $\alpha \geq 0$ then there is at most one solution.
  3. If u solves the Neumann problem, then any other solution is of the form $v = u + c$ with $c \in \mathbb{R}$.
- Comments:
  - Statement 1 is a special case of 2.
  - That adding a constant does not destroy a solution of the Neumann problem is obvious. The nontrivial part is that all the solutions are obtained from a given one, if it exists, by adding constants.

Uniqueness theorem for Poisson's equation (II)

PROOF OF THE THEOREM. Suppose u_1 and u_2 are two solutions of the Robin problem. Then $v = u_1 - u_2$ is harmonic in Ω and satisfies the boundary condition $v + \alpha \partial_n v = 0$. Setting $u = v$ in the third Green's identity yields

$$\int_\Omega \left|\vec{\nabla} v\right|^2 \, \mathrm{d}x\mathrm{d}y = -\int_{\partial\Omega} \alpha (\partial_n v)^2 \, \mathrm{d}s.$$

Since the left-hand side is nonnegative and the right-hand one is non positive, both must vanish. Hence $\vec{\nabla} v = 0$ in Ω and $0 = \alpha \partial_n v = -v$ on $\partial\Omega$. Therefore v is constant in Ω and vanishes in $\partial\Omega$. Thus $v(x, y) = 0$. To prove 3, use again the third identity but now with $v = u_1 - u_2$ which is a solution of the homogeneous Neumann problem. This implies that

$$\int_\Omega \left|\vec{\nabla} v\right|^2 \, \mathrm{d}x\mathrm{d}y = 0,$$

and hence v is a constant. Since we do not have a constraint on the value of v at $\partial\Omega$, the constant is free.

Classical solutions for elliptic problems

- Since we will be solving elliptic PDE by separation of variables and Fourier series, the question of the strong validity of the obtained series arises. For the Dirichlet problem, this is answered by the following general result
- **Theorem**. Consider the Dirichlet problem $\Delta u = 0$ for $(x, y) \in \Omega$, $u(x, y) = g(x, y)$ for $(x, y) \in \partial\Omega$ in a bounded domain Ω. Let

$$u(x, y) = \sum_{n=1}^{\infty} u_n(x, y)$$

be a formal solution of the problem, with each u_n a harmonic function in Ω and continuous in $\bar{\Omega}$. If the series converges uniformly on $\partial\Omega$ to g, then it converges uniformly on $\bar{\Omega}$ and is a classical solution of the problem.

Separation of variables on rectangular domains (I)

- Consider the Dirichlet problem for the Laplace equation on the rectangular domain $\Omega = (a, b) \times (c, d)$, with the boundary conditions

$$u(a, y) = f(y),\ u(b, y) = g(y),\ u(x, c) = h(x),\ u(x, d) = k(x).$$

- In order to be able to apply the separation of variables technique, we split first u as $u(x, y) = u_1(x, y) + u_2(x, y)$, with

$$\begin{aligned} &u_1(a, y) = f(y),\ u_1(b, y) = g(y),\ u_1(x, c) = 0,\ u_1(x, d) = 0, \\ &u_2(a, y) = 0,\ u_2(b, y) = 0,\ u_2(x, c) = h(x),\ u_2(x, d) = k(x), \end{aligned}$$

  so that the sum satisfies the original boundary conditions.

- This has the advantage that each problem can be separately solved by separation of variables.

Separation of variables on rectangular domains (I)

- Indeed, consider for instance the problem for u_1: $\Delta u_1 = 0$ on Ω and $u_1(a,y) = f(y),\ u_1(b,y) = g(y),\ u_1(x,c) = 0,\ u_1(x,d) = 0$, and search for solutions of the form $u_1(x,y) = X_1(x)Y_1(y)$.
- One gets, using the standard separation of variables reasoning,

$$\begin{aligned} X_1''(x) - \lambda X_1(x) &= 0, \qquad a < x < b, \\ Y_1''(x) + \lambda Y_1(x) &= 0, \qquad c < y < d. \end{aligned}$$

- The homogeneous boundary condition at $y = c, d$ implies $Y(c) = Y(d) = 0$, and a sequence of eigenvalues λ_n and $Y_n(y)$ can be constructed. The obtained values of λ_n can then be used to solve the equation for the associated $X_n(x)$, without imposing any boundary condition. Each X_n will thus contain two arbitrary constants, A_n and B_n.

Separation of variables on rectangular domains (II)

- The homogeneity of the boundary conditions then implies that

$$u_1(x,y) = \sum_n X_n(x)Y_n(y)$$

satisfies also the boundary at $y = c, d$.

- Finally, we impose the remaining boundary conditions at $x = a, b$:

$$\begin{aligned} f(y) &= u_1(a,y) = \sum_n X_n(a)Y_n(y), \\ g(y) &= u_1(b,y) = \sum_n X_n(b)Y_n(y). \end{aligned}$$

Each of them gives rise to a Fourier series allows the computation of the A_n and B_n in X_n.

Separation of variables on circular domains (I)

- Consider a circular domain of radius a given in polar coordinates $B_a = \{(r, \theta),\ 0 < r < a,\ 0 \leq \theta \leq 2\pi\}$, and the corresponding Laplace equation for $w(r, \theta)$

$$w_{rr} + \frac{1}{r} w_r + \frac{1}{r^2} w_{\theta\theta} = 0,$$

  with Dirichlet conditions $w(a, \theta) = h(\theta)$, with $h(0) = h(2\pi)$.

- Separation of variables $w(r, \theta) = R(r)\Theta(\theta)$ leads immediately to

$$\begin{aligned} r^2 R''(r) + r R'(r) - \lambda R(r) = 0, &\qquad 0 < r < a, \\ \Theta''(\theta) + \lambda \Theta(\theta) = 0, &\qquad 0 \leq \theta \leq 2\pi. \end{aligned}$$

- In order that the solution is of class C^2, we need to impose that $\Theta(0) = \Theta(2\pi)$ and $\Theta'(0) = \Theta'(2\pi)$; no independent condition on Θ'' is necessary, since $\Theta''(0) = \Theta''(2\pi)$ automatically from the differential equation for Θ.

Separation of variables on circular domains (II)

- The solution to the angular equation satisfying all the periodicity conditions is

$$\Theta_n(\theta) = A_n \cos n\theta + B_n \sin n\theta, \quad \lambda_n = n^2, n = 0, 1, 2, \ldots.$$

- The corresponding equation for the radial part is

$$r^2 R_n'' + r R_n' - n^2 R_n = 0,$$

with general solution $R_0(r) = C_0 + D_0 \log r$ for $n = 0$ and

$$R_n(r) = C_n r^n + D_n r^{-n}, \ n = 1, 2, \ldots$$

- Since only smooth solutions are considered, the r^{-n} terms must be suppressed since $r = 0$ belongs to B_a (the situation changes if the equation is solved in a ring, where both terms must be kept, or in the exterior of the circle, where it is the r^n terms that must be disregarded). Hence, $D_n = 0$, $n = 1, 2, \ldots$, and also $D_0 = 0$.

Separation of variables on circular domains (III)

- Combining the remaining terms, we form the series

$$w(r,\theta) = \frac{\alpha_0}{2} + \sum_{n=1}^{\infty} r^n \left(\alpha_n \cos n\theta + \beta_n \sin n\theta\right),$$

  where the coefficients have been redefined appropriately. Each term in the series is harmonic.

- Finally, the boundary condition is imposed as

$$h(\theta) = w(a,\theta) = \frac{\alpha_0}{2} + \sum_{n=1}^{\infty} a^n \left(\alpha_n \cos n\theta + \beta_n \sin n\theta\right).$$

- This corresponds to the Fourier series of a function of period 2π, and the coefficients can be computed straightforwardly after dividing by a^n.

Separation of variables on circular domains (IV)

- Indeed, one gets

$$\begin{aligned}
\alpha_0 &= \frac{1}{\pi}\int_0^{2\pi} h(\phi)\mathrm{d}\phi, \\
\alpha_n &= \frac{1}{\pi a^n}\int_0^{2\pi} h(\phi)\cos n\phi \ \mathrm{d}\phi, \ n=1,2,\ldots \\
\beta_n &= \frac{1}{\pi a^n}\int_0^{2\pi} h(\phi)\sin n\phi \ \mathrm{d}\phi, \ n=1,2,\ldots
\end{aligned}$$

- Due to the a^{-n} coefficients, the series converges uniformly on any disk of radius $b<a$. This can be used to show that the series solution is a classical one, even for piecewise smooth boundary conditions $h(\theta)$. However, the convergence deteriorates rapidly when approaching $r=a$. For an example with $h(\theta)=V_0,-V_0,V_0,-V_0$ on the four quadrants, see the solved problem in the Campus.

Assignments for lecture 8

1. (Exercise 7.2 in [PR]) Prove uniqueness for the Dirichlet and Neumann problems for the reduced Helmholtz equation

$$\Delta u - ku = 0$$

in a bounded planar domain Ω, where k is a positive constant. Hint: use Green's third identity.

2. **Potential in a coaxial conductor**. Consider a long conducting cylinder of interior radius b with a coaxial wire of radius r_0. The central wire is kept at potential V_0, while the cylinder is divided into equal quarters, with alternate segments being held at potentials $+V$ and $-V$. Solve Laplace's equation for the potential in the interior of the cylinder. Hint: the r^{-n} and $\log r$ solutions for the radial equation cannot be disregarded for this problem, because $r = 0$ is outside the region of interest.

Mathematical Methods – Lecture 9

Numerical methods

Carles Batlle Arnau

Departament de Matemàtica Aplicada 4
and
Institut d'Organització i Control de Sistemes Industrials

Universitat Politècnica de Catalunya

Lecture goals

- To present the idea of *numerical scheme* and an overview of problems and methods.
- To introduce some finite differences schemes for the heat equation, and discuss the stability, consistency and convergence of some of them.
- To overview some techniques for the solution of large linear algebraic systems.

Outline

- Introduction: numerical schemes.
- Finite differences.
- The heat equation: explicit and implicit schemes, stability, consistency and convergence.
- Numerical solution of large linear algebraic equations.

References

PR Pinchover, Y., & J. Rubinstein, *An introduction to partial differential equations*, Cambridge University Press, Cambridge, UK (2005), chapter 11.

- Further discussions, especially for finite differences, can be found in Morton, K.W., & D.F. Mayers, *Numerical solution of partial differential equations*, Cambridge University Press, Cambdrige, UK (1994).
- A good introductory reference, including C++ code, for numerical methods in general, is Press, W.H, Teukolsky, S.A., Vetterling, W.T., and B.P. Flannery, *Numerical Recipes. The art of scientific computing*, Third Edition, Cambridge University Press, New York, NY (2007).

Numerical schemes (I)

- PDE with nonconstant coefficients, in complicated domains or nonlinear cannot, in general, be solved analytically, using the methods we have seen (characteristics, separation of variables) or others.
- Alternatives:
  - qualitative analysis: normally not acceptable from the point of view of engineering.
  - numerical solutions: facilitated by the advances in digital computation.
- A numerical solution is only approximate; however, this is not a serious drawback as long as the approximation can be improved with more effort. In fact, even exact analytical solutions must be evaluated numerically with some error.
- Different kinds of PDE require different numerical methods.

Numerical schemes (II)

- A numerical method replaces the PDE, formulated by a single equation with a single real unknown function, by a discrete (finite) set of algebraic equations in finitely many unknowns, typically corresponding to the values of the function at selected points in space and/or time.
- The discrete problem obtained from a PDE is called a *numerical scheme*.
- In the linear case one obtains a large linear algebraic system, generally with some structure (diagonal banded, for instance), for which special techniques can be used in order to increase the efficiency (computation time, memory storage, accuracy).
- The two most popular numerical methods are based on finite differences (FDM) or finite elements (FEM). Both methods can be used for most PDE.

Finite differences (I)

- Consider a function $u(x, y)$ defined on $D = [0, a] \times [0, b]$, and a $N \times M$ discrete grid, net or mesh on D:

$$(x_i, y_j) = (i\Delta x, j\Delta y), \quad 0 \leq i \leq N-1,\ 0 \leq j \leq M-1,$$

  where $\Delta x = a/(N-1)$, $\Delta y = b/(M-1)$.

- We define further the discretized values of u on the mesh

$$U_{i,j} = u(x_i, y_j).$$

- Assuming the target function u is smooth enough, we can use a Taylor expansion to compute

$$\begin{aligned} u(x_{i+1}, y_j) &= u(x_i, y_j) + \partial_x u(x_i, y_j)\Delta x \\ &+ \frac{1}{2}\partial_x^2 u(x_i, y_j)(\Delta x)^2 + \frac{1}{6}\partial_x^3 u(x_i, y_j)(\Delta x)^3 + \cdots \end{aligned}$$

Finite differences (II)

- From this it follows that

$$\partial_x u(x_i, y_j) = \frac{U_{i+1,j} - U_{i,j}}{\Delta x} + O(\Delta x).$$

- Disregarding the higher order terms, we obtain the *forward difference formula* for u_x:

$$\partial_x u(x_i, y_j) = \frac{U_{i+1,j} - U_{i,j}}{\Delta x}.$$

- Similarly, using a Taylor expansion for $u(x_{i-1}, y_j)$

$$\begin{aligned} u(x_{i-1}, y_j) &= u(x_i, y_j) + \partial_x u(x_i, y_j)(-\Delta x) \\ &+ \frac{1}{2}\partial_x^2 u(x_i, y_j)(-\Delta x)^2 + \frac{1}{6}\partial_x^3 u(x_i, y_j)(-\Delta x)^3 + \cdots \end{aligned}$$

we obtain the *backward difference formula* for u_x:

$$\partial_x u(x_i, y_j) = \frac{U_{i,j} - U_{i-1,j}}{\Delta x}.$$

Finite differences (III)

- The error induced by these approximations is called a *truncation error*. In order to minimize it, Δx should be very small. Since $\Delta x = O(1/N)$, this means that N should be very large. However this is expensive, both in terms of computation time and memory requirements.
- If we write the Taylor expansion for $u(x_{i-1}, y_j)$ around $u(x_i, y_j)$ and subtract it from the one for $u(x_{i+1}, y_j)$, we are able to cancel the terms in $(\Delta x)^2$ and obtain

$$\partial_x u(x_i, y_j) = \frac{U_{i+1,j} - U_{i-1,j}}{2\Delta x} + O((\Delta x)^2).$$

- The approximation

$$\partial_x u(x_i, y_j) = \frac{U_{i+1,j} - U_{i-1,j}}{2\Delta x}$$

is called a *central finite difference*, or second-order approximation for u_x. Since the error is $O((\Delta x)^2) = O(1/N^2)$, a smaller error can be obtained than in the case of forward or backward differences with the same N.

Finite differences (IV)

- Similarly, the central finite difference approximation for u_y is given by

$$\partial_y u(x_i, y_j) = \frac{U_{i,j+1} - U_{i,j-1}}{2\Delta y}.$$

- Central differences can be obtained also for higher order derivatives. If we add, instead of subtract, the series for $u(x_{i+1}, y_j)$ and $u(x_{i-1}, y_j)$ around $u(x_i, y_j)$ we obtain

$$\partial_x^2 u(x_i, y_j) = \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{(\Delta x)^2} + O((\Delta x)^2),$$

and, similarly,

$$\partial_y^2 u(x_i, y_j) = \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{(\Delta y)^2} + O((\Delta y)^2).$$

- Central finite differences can also be obtained for mixed or higher order derivatives, and finite differences of higher order, involving more mesh points, can be computed as well, but we will not use them here.

Discretizing the heat equation (I)

- One may think that building a numerical scheme for a PDE is just a matter of choosing finite differences for each partial derivative appearing in the PDE, choosing the mesh and straightforwardly solving the resulting algebraic system. However, serious problems do appear when one tries to carry out this process, and each kind of PDE requires specific techniques in order to obtain meaningful results.
- Consider the Dirichlet problem for the heat equation:

$$\begin{aligned} u_t &= ku_{xx}, \quad 0 < x < \pi,\ t > 0, \\ u(0,t) &= u(\pi,t) = 0,\ t \geq 0, \quad u(x,0) = f(x),\ 0 \leq x \leq \pi, \end{aligned}$$

where $f(0) = f(\pi) = 0$.

Discretizing the heat equation (II)

- Construct a grid $\Delta x = \pi/(N-1)$, fix $\Delta t > 0$ and let $x_i = i\Delta x$, $t_n = n\Delta t$, $u_{i,n} = u(x_i, t_n)$. A forward first order finite difference for u_t and a second order central one for u_{xx} yields

$$\frac{U_{i,n+1} - U_{i,n}}{\Delta t} = k\frac{U_{i+1,n} - 2U_{i,n} + U_{i-1,n}}{(\Delta x)^2},\ 1 \leq i \leq N-2,\ n \geq 0.$$

- Notice that i varies between 1 and $N-2$, otherwise we would get expressions involving $U_{-1,n}$ or $U_{N,n}$, which are not defined. This is not a problem, because the boundary conditions imply

$$U_{0,n} = U_{N-1,n} = 0 \quad \forall\ n \geq 0.$$

- One gets a numerical scheme given by the above relations and the $N-2$ difference equations ($i = 1, \ldots, N-2$)

$$U_{i,n+1} = U_{i,n} + \alpha\left(U_{i+1,n} - 2U_{i,n} + U_{i-1,n}\right),\ \alpha = k\frac{\Delta t}{(\Delta x)^2},\ n > 0,$$

with initial conditions $U_{i,0} = f(x_i)$.

Stability of a numerical scheme (I)

- We have derived a simple algorithm for a numerical solution of the heat equation. However, it turns out that, unless Δt is chosen so that
$$\Delta t < \frac{1}{2k}(\Delta x)^2$$
the scheme is unstable: any small perturbation of the initial conditions will grow very fast in time. Since the representation of real numbers in the computer is always finite, any initial condition will develop an increasing *round-off error* and the numerical solution will be meaningless after some time.
- In order to define precisely what stability means in this context, let us denote by V the vector of unknowns, and F the vector that contains the known parameters of the problem (boundary or initial conditions). Then any numerical scheme can be written as $T(V) = F$, or $AV = F$ for an appropriate matrix A in the linear case.

Stability of a numerical scheme (II)

- Let $T(V) = F$ be a numerical scheme, and let V^i, $i = 1, 2$, be two solutions corresponding to different boundary or initial conditions F^i. We say that the scheme is stable if for each $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that $|F^1 - F^2|_F < \delta$ implies $|V^1 - V^2|_V < \epsilon$. In other words, a small change in the problem's data implies a small change in the solution. Here $|\cdot|_{F,V}$ denote appropriate norms for data and solutions, respectively.
- Let us study the stability of the numerical scheme obtained for the heat equation. We will consider as perturbations of the initial data, obeying the boundary conditions, $p_k(x) = \sin kx$, where $k \in \mathbb{N}$. To simplify the computations, we will work with complex exponentials and consider instead $p_k(x) = e^{jkx}$.
- Since any perturbation of the initial condition can be expanded in (sinus) Fourier series, stability for all k implies stability for any perturbation. Conversely, if the scheme is unstable for any value of k, it will be unstable for a general perturbation.

Stability of a numerical scheme (III)

- Remember from our discussion of the separation of variables method for the heat equation than the individual solutions were (sinusoidal functions vanishing at the boundary) times (functions of time). Hence, we can seek a solution to our numerical scheme for initial data $p_k(x)$, representing the difference between a given solution and a perturbed one, in the form

$$U_{i,n} = A_n e^{jki\Delta x}.$$

- Substitution in the numerical scheme yields, after canceling a $e^{jk\Delta x}$ term,

$$A_{n+1} = A_n + \alpha A_n (e^{jk\Delta x} - 2 + e^{-jk\Delta x}) = A_n \left(1 - 4\alpha \sin^2 \frac{k\Delta x}{2}\right).$$

- If we want the A_n not to grow unbounded, we must demand that

$$\left|1 - 4\alpha \sin^2 \frac{k\Delta x}{2}\right| \leq 1.$$

Stability of a numerical scheme (IV)

- Since $1 - 4\alpha \sin^2 \frac{k\Delta x}{2} \leq 1$, it follows that the necessary and sufficient condition for stability is

$$1 - 4\alpha \sin^2 \frac{k\Delta x}{2} \geq -1, \quad \text{or} \quad 2\alpha \sin^2 \frac{k\Delta x}{2} \leq 1.$$

- This must hold for any integer k. Since, for general Δx, the sinus can be made arbitrarily close to 1, the only solution is that $2\alpha \leq 1$, which leads to the bound $\Delta t < \frac{1}{2k}(\Delta x)^2$.
- This is quite restrictive, since Δx is chosen already small and hence the time step is forced to be very small. A large number of time steps will thus be necessary to compute the solution for interesting times and the round-off error will accumulate.
- Before presenting more favorable numerical schemes, we will discuss two further theoretical aspects: *consistency* and *convergence* of an scheme.

Consistency of a numerical scheme

- A numerical scheme is said to be *consistent* if the solution of the PDE satisfies the scheme in the limit where the grid tends to zero.
- To examine the consistency of our numerical scheme for the heat equation, define, for any function $v(x,t)$,

$$R[v] = \frac{v(x_i, t_{n+1}) - v(x_i, t_n)}{\Delta t} - k\frac{v(x_{i+1}, t_n) - 2v(x_i, t_n) + v(x_{i-1}, t_n)}{(\Delta x)^2}.$$

- Let $u(x,t)$ be a solution to the PDE problem. Using the finite difference approximation to the several derivatives, one has

$$\begin{aligned} R[u] &= u_t(x_i,t_n) + \frac{1}{2}u_{tt}(x_i,t_n)\Delta t + O((\Delta t)^2) \\ &- ku_{xx}(x_i,t_n) - k\frac{1}{12}(\Delta x)^2 u_{xxxx}(x_i,t_n) + O((\Delta x)^4) \\ &\overset{u_t = ku_{xx}}{=} \frac{1}{2}u_{tt}(x_i,t_n)\Delta t - k\frac{1}{12}(\Delta x)^2 u_{xxxx}(x_i,t_n) + O((\Delta t)^2, (\Delta x)^4). \end{aligned}$$

- Hence $R[u] \to 0$ as $\Delta t, \Delta x \to 0$ and the scheme is consistent.

Convergence of a numerical scheme

- We say that a numerical scheme for the heat equation is *convergent* if the solution to the discrete numerical problem converges in the limit $\Delta x \to 0$, $\Delta t \to 0$ to the solution of the original PDE.
- **Theorem**. Any consistent and stable numerical scheme for our heat equation problem is convergent.
- The importance of the above theorem is that stability and consistency are much more easier to check than convergency.
- Hence, our numerical scheme for the heat equation is convergent, since it is consistent and it is also stable, as long as $\Delta t < \frac{1}{2k}(\Delta x)^2$.
- Similar theorems can be stated for other PDE problems.

Other numerical schemes for the heat equation (I)

- The numerical scheme derived for the heat equation has the problem of requiring very small time steps if high accuracy is required. One may think that the problem lies in the first-order time difference.
- To examine this, let us replace the forward finite difference in time by a central one, so that we get

$$\frac{U_{i,n+1} - U_{i,n-1}}{2\Delta t} = k\frac{U_{i+1,n} - 2U_{i,n} + U_{i-1,n}}{(\Delta x)^2},\ 1 \leq i \leq N-2,\ n \geq 0.$$

- The (minor) obstacle that $U_{i,-1}$ does not exists can be solved by using our first scheme for the first step and then continuing with the new one.

Other numerical schemes for the heat equation (II)

- Surprisingly, the stability of the new scheme gets worse: it is unstable for any choice of Δt and Δx!
- Indeed, proceeding as before, we obtain the following second order recurrence for the amplitude of an arbitrary harmonic perturbation

$$A_{n+1} = A_{n-1} - 8\alpha \sin^2 \frac{k\Delta x}{2} A_n.$$

- The characteristic polynomial for this recurrence is

$$r^2 + 8\alpha \sin^2 \frac{k\Delta x}{2} r - 1 = 0$$

which has real solutions, one of them with absolute value always greater than 1. Hence the scheme is always unstable and the problem is not (only) with the forward difference in time.

Other numerical schemes for the heat equation (III)

- The problem of finding a stable and efficient scheme for the heat equation (and others) was an intense area of research in the middle of the XX century. One of the most popular schemes that was proposed is the *Crank-Nicolson* scheme, defined by

$$\frac{U_{i,n+1} - U_{i,n}}{\Delta t} = k \left(\frac{U_{i+1,n} - 2U_{i,n} + U_{i-1,n}}{2(\Delta x)^2} + \frac{U_{i+1,n+1} - 2U_{i,n+1} + U_{i-1,n+1}}{2(\Delta x)^2} \right),$$

  for $1 \leq i \leq N-2$, $n \geq 0$.

- Notice that u_{xx} has been approximated by a combination of the present and *future* central second order differences. This may sound strange, but it can be shown that the resulting scheme is consistent, and is stable for any value of Δx and Δt. Hence it is also convergent, which is all that matters.
- The Crank-Nicolson cannot be written as an explicit expression for computing each $U_{i,n+1}$ in terms of the $\{U_{i,m}\}_{\forall i, m \leq n}$. Such schemes are called *implicit*, while the former ones were *explicit*. This means that at each step in time a system for the $U_{i,n+1}$ must be solved, greatly increasing the numerical burden of the method. However, the advantages of being able to choose the mesh arbitrarily more than compensate for this.
- As a rule of thumb, implicit numerical schemes are more stable than explicit ones.
- The discussion presented here can be extended to many other PDE. See 11.4 and 11.5 of [PR] for the Laplace and wave equations.

Large linear systems

- From the study of the Crank-Nicolson scheme, and also from that of other numerical schemes associated to Laplace or wave equations, it is clear that at each step one has to solve large linear systems. This is the domain of *numerical linear algebra*, probably the most important area of numerical analysis.
- Large linear systems **must not** be solved by inverting the system matrix; the round off error accumulated by the process of computing a large inverse is unacceptable. An alternative is Gaussian elimination, which is well behaved with respect to error propagation but has a high complexity: a direct solution by the Gauss elimination method for a system with K unknowns requires $O(K^3)$ multiplications.
- The methods used for large linear systems are mostly iterative, although there are some direct methods based on special decompositions of the system matrix (LU and our well-known QR and SVD). We will present three iterative methods, which exploit the special form of the the systems coming from PDE: the *Jacobi method*, the *Gauss-Seidel* method and the *successive over relaxation* (SOR) method.
- We will illustrate everything with the Crank-Nicolson scheme for the heat equation.

Iterative methods (I)

- We rewrite the Crank-Nicholson scheme as

$$U_{i,n+1} = \frac{\alpha}{2}(U_{i+1,n+1} - 2U_{i,n+1} + U_{i-1,n+1}) + r_{i,n} \qquad (1)$$

where

$$r_{i,n} = \frac{\alpha}{2}(U_{i+1,n} - 2U_{i,n} + U_{i-1,n}) + U_{i,n}.$$

- The values of $r_{i,n}$ are known at the nth step for all i, and the unknowns are the $U_{i,n+1}$ for $i = 1, \ldots, N-2$. We consider a given n and solve (1) iteratively.
- The solution at the pth iteration will be denoted by $V^p_{i,n+1}$, with the process starting with some $V^0_{i,n+1}$ which, for instance, can be chosen as $U_{i,n}$.
- In the Jacobi method we compute $V^{p+1}_{i,n+1}$ using (1) and the values of $V^p_{i-1,n+1}$ and $V^p_{i+1,n+1}$, which are known from the previous iteration.

Iterative methods (II)

- One gets

$$V_{i,n+1}^{p+1} = \frac{\alpha}{2}(V_{i+1,n+1}^{p} - 2V_{i,n+1}^{p+1} + V_{i+1,n+1}^{p}) + r_{i,n}$$

from which the Jacobi formula is obtained

$$V_{i,n+1}^{p+1} = \frac{\alpha}{2\alpha+2}(V_{i-1,n+1}^{p} + V_{i+1,n+1}^{p}) + \frac{1}{\alpha+1} r_{i,n}. \qquad (2)$$

- Inspecting (2), one sees that, when computing $V_{i,n+1}^{p+1}$, the values of $V_{i-1,n+1}^{p}$ are used, while in fact the updated values $V_{i-1,n+1}^{p+1}$ are already known. It is not surprising, then, that better results are obtained with the Gauss-Seidel formula

$$V_{i,n+1}^{p+1} = \frac{\alpha}{2\alpha+2}(V_{i-1,n+1}^{p+1} + V_{i+1,n+1}^{p}) + \frac{1}{\alpha+1} r_{i,n}. \qquad (3)$$

Iterative methods (III)

- We shall see that the Gauss-Seidel method yields a rate of convergence twice as fast as the Jacobi method. Furthermore, the Gauss-Seidel method is also more efficient memory wise: it is possible to update the vector $V_{i,n}$ in place, without keeping old values along the new ones.
- The Gauss-Seidel method can be improved further by the SOR method. Rewrite the Gauss-Seidel formula as

$$V_{i,n+1}^{p+1} = V_{i,n+1}^{p} + \left[\frac{\alpha}{2\alpha+2}(V_{i-1,n+1}^{p+1} + V_{i+1,n+1}^{p}) + \frac{1}{\alpha+1} r_{i,n} - V_{i,n+1}^{p} \right].$$

- The term in square brackets is the change from $V_{i,n+1}^{p}$ to $V_{i,n+1}^{p+1}$. The SOR method multiplies this by a *relaxation parameter* ω:

$$V_{i,n+1}^{p+1} = V_{i,n+1}^{p} + \omega \left[\frac{\alpha}{2\alpha+2}(V_{i-1,n+1}^{p+1} + V_{i+1,n+1}^{p}) + \frac{1}{\alpha+1} r_{i,n} - V_{i,n+1}^{p} \right].$$

- The Gauss-Seidel method is recovered for $\omega = 1$, but a clever choice of ω in $(1, 2)$ yields a scheme which converges much faster than Gauss-Seidel's.

Convergence of iterative methods (I)

- All the numerical schemes that we have presented can be written as

$$AV = b.$$

- For instance, for the Crank-Nicholson scheme with $N = 7$ at time step $n + 1$ one has

$$V_i = U_{i,n+1}, \; b_i = r_{i,n}, \; i = 1, 2, 3, 4, 5$$

and

$$A = \begin{pmatrix} 1+\alpha & -\alpha/2 & 0 & 0 & 0 \\ -\alpha/2 & 1+\alpha & -\alpha/2 & 0 & 0 \\ 0 & -\alpha/2 & 1+\alpha & -\alpha/2 & 0 \\ 0 & 0 & -\alpha/2 & 1+\alpha & -\alpha/2 \\ 0 & 0 & 0 & -\alpha/2 & 1+\alpha \end{pmatrix}.$$

Convergence of iterative methods (II)

- Let us decompose A as

$$A = L + D + U,$$

where L, D and U are matrices whose nonzero entries are below, on and above the diagonal, respectively.

- For the Jacobi method we have

$$A_{ii} V_i^{p+1} = -\sum_{j \neq i} A_{ij} V_j^p + b_i,$$

and hence, in vector notation, the $(p+1)$ iterate at time step $n+1$ is

$$V^{p+1} = -D^{-1}(L+U)V^p + D^{-1}b.$$

Convergence of iterative methods (III)

- Similarly, for the Gauss-Seidel formula,

$$\sum_{j \leq i} A_{ij} V_j^{p+1} = -\sum_{j>i} A_{ij} V_j^p + b_i,$$

from which

$$V^{p+1} = -(D+L)^{-1} U V^p + (D+L)^{-1} b.$$

- Finally, the SOR method can be cast in this formulation as

$$V^{p+1} = -(D+\omega L)^{-1} \left\{ [(1-\omega)D - \omega U] V^p + \omega b \right\}.$$

- All the iterative methods can be given the general form

$$V^{p+1} = MV^p + Qb$$

with appropriate M and Q.

Convergence of iterative methods (IV)

- Obviously, the solution $V = U_n$ we are trying to obtain iteratively is a fixed point of the iteration:

$$V = MV + Qb.$$

- Let $\Xi^{p+1} = V^{p+1} - V$ be the difference between the $(p+1)$th iteration and the exact solution. One gets

$$\Xi^{p+1} = (MV^p + Qb) - (MV + Qb) = M(V^p - V) = M\Xi^p.$$

- In order to prove convergence, we need to check that the sequence $\{\Xi^p\}$ solution of the above recurrence goes to zero.
- For simplicity, we assume that M is diagonalizable, and we denote its eignevalues by λ_i and the corresponding eigenvectors by ω_i.

Convergence of iterative methods (V)

- We expand Ξ^0 in terms of the eigenvectors of M as $\Xi^0 = \sum_i \beta_i \omega_i$.
- Then $\Xi^1 = M\Xi^0 = \sum_i \beta_i M\omega_i = \sum_i \beta_i \lambda_i \omega_i$ and

$$\Xi^p = \sum_i \beta_i \lambda_i^p \omega_i. \tag{4}$$

- We define the *spectral radius* of the matrix M as

$$\lambda(M) = \max_i |\lambda_i|.$$

- It is evident from (4) that the iterative method converges if and only if $\lambda(M) < 1$. If there is any eigenvector with modulus equal to or greater than 1, picking an initial guess such that the difference with the fixed point has a component along that eigenvector will make the difference constant or exponentially increasing.

Convergence of iterative methods (VI)

- It is also obvious that the smaller the spectral radius the greater the rate of convergence of the iterative method.
- For instance, it can be shown that the spectral radius of the Jacobi, Gauss-Seidel and SOR methods for the Crank-Nicolson scheme are given by

$$\begin{aligned}
\lambda_{\text{Jacobi}} &= \frac{\alpha}{1+\alpha}\cos\Delta x \overset{\Delta x \text{ small}}{\approx} \frac{\alpha}{1+\alpha}\left(1-\frac{(\Delta x)^2}{2}\right) < 1, \\
\lambda_{\text{Gauss-Seidel}} &\approx \frac{\alpha}{1+\alpha}\left(1-(\Delta x)^2\right) < 1, \\
\lambda_{\text{SOR}} &\approx \frac{\alpha}{1+\alpha}(1-2\Delta x) < 1.
\end{aligned}$$

- Hence, the three iterative methods are convergent, although the rate of convergence decreases for all of them as $\Delta x \to 0$. Furthermore, for a given value of Δx,

$$\lambda_{\text{SOR}} < \lambda_{\text{Gauss-Seidel}} < \lambda_{\text{Jacobi}},$$

  meaning that SQR converges more rapidly than Gauss-Seidel's, and the later more rapidly than Jacobi's.

- There exist several sophisticated methods, such as the *multi-grid* method, that accelerate the convergence rate well beyond the methods we have presented.

Assignments for lecture 9

1. Obtain the central finite difference approximation for $\partial_x^2 u(x_i, y_i)$, with the order of the error.
2. (Exercise 11.5 in [PR]) Consider the heat equation on $[0, \pi]$

$$u_t = u_{xx}, \quad 0 < x < \pi, \ t > 0,$$

with boundary and initial conditions

$$u(0,t) = u(\pi,t) = 0, \quad u(x,0) = x(\pi - x).$$

- Solve the problem numerically (in spatial grids of 25, 61, and 101 points) using the Crank-Nicolson scheme. Compute the solution at the point $(x,t) = (\pi/4, 2)$ for each one of the grids.
- Solve the same problem analytically using 2, 7, and 20 Fourier terms. Construct a table to compare the analytic solution at the point $(x,t) = (\pi/4, 2)$ with the numerical ones.

Mathematical Methods – Lecture 10

Variational methods

Carles Batlle Arnau

Departament de Matemàtica Aplicada 4
and
Institut d'Organització i Control de Sistemes Industrials

Universitat Politècnica de Catalunya

Lecture goals

- To present the basic ideas of the calculus of variations, and some applications.
- To introduce some ideas about Hilbert spaces.
- To present the Ritz method.
- To present the Galerkin method and the weak formulation of a PDE.
- To present the finite elements method as a special case of Galerkin's method.

Outline

- Calculus of variations. Examples: classical mechanics, minimal surfaces and reconstruction of a function from its gradient.
- The second variation.
- Hilbert spaces and weak formulation.
- The Ritz method.
- Weak solutions and the Galerkin method.
- Finite elements: an example.

References

PR Pinchover, Y., & J. Rubinstein, *An introduction to partial differential equations*, Cambridge University Press, Cambridge, UK (2005), chapter 10.

KB Kwon, Y.W., & H. Bang, *The finite element method using Matlab*, CRC Press, Boca Raton (1997).

BF To fully appreciate some of the concepts related to this lecture a course in advanced real analysis is needed. An introductory exposition can be found, for instance, in Batlle, C., & E. Fossas, *Apunts d'Anàlisi Real*, Facultat de Matemàtiques i Estadística, UPC, imprès per `Ahlens, S.L.`, D.L.: B-8830-2002 (2002).

Calculus of variations (I)

- Let Γ be a simple closed curve in $\mathbb{R}^3$. A surface whose boundary is Γ is said to be *spanned* by Γ.
- Let S be spanned by Γ and assume that S is the graph of a function $z = u(x, y)$, with $(x, y) \in \Omega$ such that $\partial\Omega$ is the projection of Γ on the xy plane.
- The normal vector to S associated with this parametrization is given by $\vec{n}(x, y) = (u_x(x, y), u_y(x, y), -1)$.
- The area of S is given by

$$E(u) = \int_\Omega |\vec{n}(x, y)| \mathrm{d}x\mathrm{d}y = \int_\Omega \sqrt{1 + u_x^2(x, y) + u_y^2(x, y)}\ \mathrm{d}x\mathrm{d}y.$$

- E is a function of the function u; this kind of object is called a *functional*. Such a dependence is denoted usually by $E[u]$.

Calculus of variations (II)

- The surface S is called *locally minimal* if its area is not greater than the area of any other surface spanned by Γ that is close to S in an appropriate sense.
- To be more precise, S parameterized by u is locally minimal if $E[u] \leq E[v]$, for any v close to u that is *admissible*, *i.e* a $C^1(\Omega)$ function that parameterizes a surface spanned also by Γ.
- Notice that the functional, being a surface integral, does not depend on the specific admissible parametrization, but only on the geometric surface. It may change only if the surface is changed.
- In order to obtain a condition for u to parameterize a minimal surface, we write $v = u + \epsilon\psi$, with $\epsilon \in \mathbb{R}$ and ψ in

$$\mathcal{A} = \{\psi \in C^1(\Omega) \cap C(\bar{\Omega}),\ \psi(x, y) = 0 \text{ for } (x, y) \in \partial\Omega\}.$$

Calculus of variations (III)

- The minimality condition is then $E[u] \leq E[u+\epsilon\psi]$, for small $|\epsilon|$ and for all $\psi \in \mathcal{A}$.
- Considering $E[u+\epsilon\psi]$ as a real function of ϵ with u and ψ fixed, the necessary condition of minimum for smooth functions is

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} E[u+\epsilon\psi]\bigg|_{\epsilon=0} = 0.$$

- The expression of the left-hand side is called *the first variation of* E *at* u, and it is denoted by $\delta E[u,\psi]$ or $\delta E[u]$.
- As an special but important case, let us assume that the minimizer function u has small derivatives, so that the square root can be approximated by $\sqrt{1+x} \approx 1 + \frac{1}{2}x$. Then

$$E[u] = \text{area of } \Omega + \frac{1}{2}\int_\Omega \left(u_x^2 + u_y^2\right) \mathrm{d}x\mathrm{d}y.$$

Calculus of variations (IV)

- Since the area of Ω does not depend on ϵ, we can replace the problem of minimizing E with the problem of minimizing

$$G[u] = \frac{1}{2}\int_\Omega \left(u_x^2 + u_y^2\right) \mathrm{d}x\mathrm{d}y = \frac{1}{2}\int_\Omega \left|\vec{\nabla} u\right|^2 \mathrm{d}x\mathrm{d}y.$$

  This is called the *Dirichlet functional* or *Dirichlet integral* associated to Ω.

- It is easy to check that

$$G[u+\epsilon\psi] = G[u] + \epsilon\int_\Omega \vec{\nabla} u\cdot\vec{\nabla}\psi \ \mathrm{d}x\mathrm{d}y + \epsilon^2 G[\psi].$$

- Hence

$$\delta G[u] = \int_\Omega \vec{\nabla} u\cdot\vec{\nabla}\psi \ \mathrm{d}x\mathrm{d}y.$$

Calculus of variations (V)

- We conclude then that a necessary condition for u to be a local minimizer is that

$$\int_\Omega \vec{\nabla} u \cdot \vec{\nabla} \psi \; \mathrm{d}x\mathrm{d}y = 0 \quad \forall \; \psi \in \mathcal{A}.$$

- Using Green's third identity and taking into account the condition on ψ at the boundary $\partial\Omega$ we get

$$\int_\Omega \vec{\nabla} u \cdot \vec{\nabla} \psi \; \mathrm{d}x\mathrm{d}y = \int_{\partial\Omega} \psi \partial_n u \mathrm{d}s - \int_\Omega \psi \Delta u \mathrm{d}x\mathrm{d}y = -\int_\Omega \psi \Delta u \mathrm{d}x\mathrm{d}y.$$

- The minimum condition is then

$$\int_\Omega \Delta u \; \psi \; \mathrm{d}x\mathrm{d}y = 0 \quad \forall \; \psi \in \mathcal{A}.$$

Calculus of variations (VI)

- Assuming now that $u \in C^2(\Omega)$, we obtain

$$\Delta u = 0 \quad \text{in } \Omega.$$

- Furthermore, by construction the values of u at the boundary are fixed, since S is spanned by Γ. Hence

$$u(x, y) = g(x, y),$$

  where g is the graph of u over $\partial\Omega$, which can be computed from Γ.
- We have thus proved that the minimizer of the Dirichlet functional is the solution of the Dirichlet problem for the Laplace equation.
- The PDE that is obtained by equating the first variation of a functional to zero is called the *Euler-Lagrange equation*.

Dirichlet problem for the Laplace equation = Euler-Lagrange equation for the Dirichlet functional.

Calculus of variations (VII)

- Let us return to the general case of computing the first variation of an arbitrary functional $K[u]$ of the form

$$K[u] = \int_\Omega F(x_1, \ldots, x_n, u, u_1, \ldots, u_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n,$$

  where we are considering an arbitrary number of independent variables but restricting ourselves to a dependency at most in the first derivatives of u (although this restriction can be removed and a similar method can be applied), and where F is a smooth function.

- One has, since $(u + \epsilon\psi)_i = u_i + \epsilon\psi_i$,

$$K[u + \epsilon\psi] = K[u] + \int_\Omega \left(\frac{\partial F}{\partial u} \epsilon\psi + \sum_{i=1}^{n} \frac{\partial F}{\partial u_i} \epsilon\psi_i \right) \mathrm{d}x\mathrm{d}y + O(\epsilon^2).$$

Calculus of variations (VIII)

- Hence

$$\delta K[u] = \int_\Omega \left(\frac{\partial F}{\partial u}\psi + \sum_{i=1}^{n} \frac{\partial F}{\partial u_i}\psi_i \right) \mathrm{d}x\mathrm{d}y.$$

- Integrating the terms in the sum by parts, using Gauss theorem, and taking into account the condition for ψ on $\partial\Omega$, one gets

$$\delta K[u] = \int_\Omega \left(\frac{\partial F}{\partial u} - \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\frac{\partial F}{\partial u_i}\right) \right) \psi \mathrm{d}x\mathrm{d}y.$$

- Since this must be zero for arbitrary $\psi \in \mathcal{A}$, assuming that the integrand is a continuous function, which amounts to u being in $C^2(\Omega)$, one gets the Euler-Lagrange equations for K:

$$\frac{\partial F}{\partial u} - \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\frac{\partial F}{\partial u_i}\right) = 0 \quad x \in \Omega.$$

Classical mechanics (I)

- As a special case, consider $n = 1$, $x_1 = t$, $u = q$ and $\Omega = (t_a, t_b)$, and let the functional be

$$J[q] = \int_{t_a}^{t_b} L(q, \dot{q}) \mathrm{d}t,$$

  where $L(q, \dot{q}) = T(\dot{q}) - V(q)$ is the Lagrangian of the mechanical system with generalized coordinate q, $T(\dot{q})$ is the kinetic energy and $V(q)$ is the potential energy.

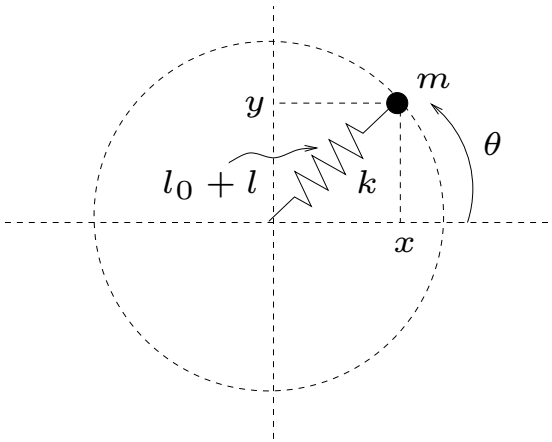
- Then the Euler-Lagrange equations of J are

$$0 = \frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}}\right) = -\frac{\partial V}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T}{\partial \dot{q}}\right).$$

- The functional J is called the *action* of the mechanical system, and the Euler-lagrange equations for the action are just Newton's equations. $\delta J[q] = 0$ is called *Hamilton's principle*, and it states that a mechanical system evolves between two instants of time in such a way that the action is locally minimal. The principle can be generalized to arbitrary dynamical systems.
- If the functional depends on several functions one gets an Euler-Lagrange equation for each of them.

Classical mechanics (II)

- Consider for instance the following 2-dimensional system



$$\begin{aligned} V &= \frac{1}{2}kl^2, \\ T &= \frac{1}{2}m(\dot{x}^2 + \dot{y}^2). \end{aligned}$$

where l is the elongation of the spring with respect to its relaxed length l_0, $l = \sqrt{x^2 + y^2} - l_0$.

- Changing to polar coordinates $x = (l_0 + l)\cos\theta$, $y = (l_0 + l)\sin\theta$, so that $l = r - l_0$, one gets

$$L(r, \dot{r}, \dot{\theta}) = \frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2) - \frac{1}{2}k(r - l_0)^2.$$

- The Euler-Lagrange equations associated to θ and r are

$$\begin{aligned} \frac{\partial L}{\partial \theta} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\theta}} &= -\frac{\mathrm{d}}{\mathrm{d}t}(mr^2\dot{\theta}) = -mr(2\dot{r}\dot{\theta} + r\ddot{\theta}), \\ \frac{\partial L}{\partial r} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{r}} &= mr\dot{\theta}^2 - k(r - l_0) - m\frac{\mathrm{d}}{\mathrm{d}t}\dot{r} = mr\dot{\theta}^2 - k(r - l_0) - m\ddot{r}. \end{aligned}$$

Classical mechanics (III)

- The final equations of motion are thus

$$\begin{aligned} m\ddot{r} &= \underbrace{mr\dot{\theta}^2}_{\text{centrifugal force}} \underbrace{-k(r-l_0)}_{\text{elastic force}}, \\ \ddot{\theta} &= \underbrace{-\frac{2}{r}\dot{r}\dot{\theta}}_{\textit{skater effect}}, \end{aligned}$$

  where the skater effect term is due to the conservation of angular momentum and the variation of the moment of inertia with r.

- Conservation of angular momentum is a consequence of L not depending on θ.

- Notice that the kinetic energy in polar coordinates has developed a dependence in the configuration variable r, as is common in robotic systems. This brings about both the centrifugal term and the angular acceleration due to the change in r.

Minimal surfaces (I)

- For the minimal surface problem without the small derivatives assumption, the integrand in the functional is

$$F(u_x, u_y) = \sqrt{1 + u_x^2 + u_y^2}.$$

- Since $\partial_u F = 0$, the Euler-lagrange equations are

$$\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) =$$

$$\frac{\partial}{\partial x}\left(\frac{u_x}{\sqrt{1 + u_x^2 + u_y^2}}\right) + \frac{\partial}{\partial y}\left(\frac{u_y}{\sqrt{1 + u_x^2 + u_y^2}}\right) = 0.$$

- Notice that this equation is of elliptic type, although the principal part is not in canonical form.

Minimal surfaces (II)

- As a (solved) example, consider the surface spanned by two circumferences of radii d and $2d$ at heights $z_1 = d$ arccosh 1 and $z_2 = d$ arccosh 2, respectively.
- It can be checked that

$$u(x,y) = d \text{ arccosh} \frac{\sqrt{x^2+y^2}}{d},$$

called a *catenoid*, is a solution of the minimal surface PDE, and indeed satisfies

$$u(x,y)|_{x^2+y^2=d^2} = d \text{ arccosh } 1, \quad u(x,y)|_{x^2+y^2=4d^2} = d \text{ arccosh } 2.$$

Reconstruction from the gradient (I)

- Many applications in image analysis require a surface $u(x, y)$ to be computed from measurements of its gradient, which are only approximated.
- Denote by $\vec{f}(x, y) = (f_1(x, y), f_2(x, y))$ the measured vector that approximates the gradient of u. Since the measure is not exact, one has that $\partial_y f_1 \neq \partial_x f_2$ and u cannot be (locally) computed by simple integration. Instead, a least squares estimation may be defined by

$$\min_u K[u] = \int_\Omega \left| \vec{\nabla} u - \vec{f} \right|^2 \mathrm{d}x\mathrm{d}y,$$

  where Ω is the region over which $\vec{f}$ is measured.
- The integrand in this functional is

$$F(u_x, u_y) = \left| \vec{\nabla} u - \vec{f} \right|^2 = \left| \vec{\nabla} u \right|^2 - 2\vec{\nabla} u \cdot \vec{f} + \left| \vec{f} \right|^2.$$

Reconstruction from the gradient (II)

- The first variation of $K[u]$ is

$$\delta K[u] = 2 \int_\Omega \left(\vec{\nabla} u - \vec{f} \right) \cdot \vec{\nabla} \psi \, \mathrm{d}x\mathrm{d}y.$$

  Notice that for the time being we do not have any boundary condition on ψ.

- Integrating by parts and using Gauss theorem one gets

$$\frac{1}{2}\delta K[u] = \int_\Omega (-\Delta u + \vec{\nabla} \cdot \vec{f})\psi \, \mathrm{d}x\mathrm{d}y + \int_{\partial\Omega} (\partial_n u - \vec{f} \cdot \vec{n})\psi \, \mathrm{d}s.$$

- Since the first variation must vanish for arbitrary ψ, we can first consider those ψ that are zero on $\partial\Omega$. This cancels the last term and then, using the standard assumption of continuity of the integrand,

$$\Delta u = \vec{\nabla} \cdot \vec{f}, \quad (x, y) \in \Omega.$$

Reconstruction from the gradient (III)

- Then the first variation reduces to

$$\int_{\partial\Omega} (\partial_n u - \vec{f} \cdot \vec{n})\psi \, \mathrm{d}s.$$

- Since this must be zero also for those ψ that are nonzero on $\partial\Omega$, we obtain

$$\partial_n u = \vec{f} \cdot \vec{n}, \quad (x, y) \in \partial\Omega.$$

- The problem of computing u from approximate measures of its gradient reduces thus to solving a Neumann problem for the Poisson equation.
- We have seen that Dirichlet and Neumann conditions arise naturally when considering minimization problems. Because of this, they are called *natural* boundary conditions.

Second variation

- From elementary calculus it is known that equating the first derivative to zero only yields a necessary condition for a smooth function to have a minimum: it could be a maximum or a point of inflexion (a saddle point for functions of several variables).
- Consider a functional $Q[u]$ such that its first variation is zero at u. The *second variation* at u is defined as

$$\delta^2 Q[u,\psi] = \delta^2 Q[u] = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}\epsilon^2} Q[u+\epsilon\psi]\Big|_{\epsilon=0},$$

  and a sufficient condition for u to be a local minimizer of Q is that it is strictly positive.
- For instance, for the Dirichlet functional

$$G[u+\epsilon\psi] = G[u] + \epsilon \int_\Omega \vec{\nabla} u \cdot \vec{\nabla}\psi \,\mathrm{d}x\mathrm{d}y + \epsilon^2 G[\psi],$$

  one has that $\delta^2 G[u] = G[\psi] > 0$ for any nontrivial ψ. Hence the harmonic function that we had identified as a candidate for minimizer is indeed a local minimizer.
- Functional Q such that $\delta^2 Q[u,\psi] > 0$ for all appropriate u and ψ are called *strictly convex*, and they have unique local minimizers.

Hilbert spaces (I)

- Let V be a vector space, finite or infinite dimensional, with an inner product $\langle\ ,\ \rangle$. A norm can be defined then by $||f|| = \langle f, f\rangle^{\frac{1}{2}}$.
- We say that a sequence of vectors (v_n) converges *strongly*, or simply converges, to v if $\lim_{n\to\infty} ||v_n - v|| = 0$.
- A sequence of vectors (v_n) is a *Cauchy sequence* if for each $\epsilon > 0$ there exits $\nu(\epsilon) \in \mathbb{N}$ such that $||f_n - f_m|| < \epsilon$ if $n, m > \nu(\epsilon)$. Every strongly convergent sequence is a Cauchy sequence but the converse is not true.
- For instance, consider $V = \mathbb{Q}$, the set of rational numbers, which is a 1-dimensional vector space over $\mathbb{Q}$, and let $||q|| = |q| = \sqrt{q \cdot q}$. The sequence of rational numbers defined by the recurrence $x_{n+1} = x_n/2 + 1/x_n$, $x_0 = 1$ is a Cauchy one, but it is not convergent: its limit is $\sqrt{2}$, which is not in $\mathbb{Q}$. It is said that $\mathbb{Q}$ is not *complete*.

Hilbert spaces (II)

- As a second example, consider the space $V = C([-\pi, \pi])$ of continuous functions in $[-\pi, \pi]$, with inner product

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(x) g(x) \mathrm{d}x,$$

and let

$$f_n(x) = \frac{4}{\pi} \sum_{k=0}^{n} \frac{1}{2k+1} \sin(2k+1)x.$$

- This is a Cauchy sequence of functions in V, since each f_n is a finite sum of continuous functions and, assuming $m \leq n$,

$$f_n(x) - f_m(x) = \frac{4}{\pi} \sum_{k=m+1}^{n} \frac{1}{2k+1} \sin(2k+1)x$$

with norm

$$||f_n - f_m||^2 = \frac{16}{\pi} \sum_{k=m+1}^{n} \frac{1}{(2k+1)^2}$$

which can be made arbitrarily small by taking n, m large enough, since the numerical series $\sum_{k=0}^{\infty} 1/(2k+1)^2$ is convergent.

- However, this sequence in not convergent in V, since its limit, in the above norm, is in fact the discontinuous function $f(x) = \text{sign } x$. Hence V is not complete.

Hilbert spaces (III)

- An inner product space in which any Cauchy sequence converges is said to be complete. Complete inner product spaces are also called Hilbert spaces, in honor of David Hilbert (1862-1943). If an inner product space is not complete, it can be completed by adding in, in an appropriate sense, the limits of all its Cauchy sequences.
- Examples of Hilbert spaces include $\mathbb{R}^n$ with the standard inner product, or the space of integrable functions on $[a, b]$ with the inner product

$$\langle f, g \rangle = \int_a^b f(x)g(x)\mathrm{d}x.$$

  This space is called $L_2([a, b])$.
- Although any complete inner product space is a Hilbert space, sometimes the word is reserved to the infinite dimensional case, such as the L_2 spaces discussed above, or the space of infinite sequences $l_2 = \{(q_n),\ q_n \in \mathbb{R},\ \sum_{k=0}^{\infty} q_n^2 < \infty\}$.

Hilbert spaces (IV)

- A subset W of a Hilbert space H is said to be *dense* if for any $f \in H$ and for every $\epsilon > 0$ there exists f_ϵ in in W such that $||f - f_\epsilon|| < \epsilon$, *i.e.* any element in H can be approximated with arbitrary precision by an element in W.
- A set B of functions in a Hilbert space H is said to be a *basis* of H if its vectors are linearly independent, that is, any (finite!) linear combination is zero iff all the coefficients are zero, and the linear closure of B, that is, the set of all the (finite!) linear combinations of elements of B, is dense in H.
- A very important Hilbert space is the one obtained from $C^1(\Omega)$ equipped with the inner product

$$\langle f, g \rangle = \int_\Omega (fg + \vec{\nabla} f \cdot \vec{\nabla} g) \mathrm{d}\vec{x}.$$

- The Hilbert space obtained by completion of this set is called a Sobolev space in honour of Sergei Sobolev (1908-1989), and is denoted by $H_1(\Omega)$. Other Hilbert spaces can be obtained by considering functions satisfying special boundary conditions at $\partial\Omega$.

The Ritz method

- Consider the problem of minimizing a functional $G[u]$ in some Hilbert space H. Select a basis $B = \{\phi_i\}$ in H, preferably orthonormal, that is, such that

$$\langle \phi_i, \phi_j \rangle = \delta_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \end{cases} .$$

The advantage of orthonormality is that it simplifies the computation of scalar products of linear combinations of ϕ_i.

- The Ritz (Walter Ritz, 1878-1909) method is based on expressing the unknown minimizer in the basis B

$$u = \sum_{k=1}^{\infty} \alpha_k \phi_k ,$$

and, since the series is expected to converge, to truncate with enough terms so that the error is acceptable

$$u \approx \sum_{k=1}^{N} \alpha_k \phi_k .$$

This converts the problem of minimizing a functional into the problem of minimizing a function of N variables, $\alpha_1, \ldots, \alpha_N$, obtained by substituting the truncated u in $G[u]$ and evaluating the integral using the orthonormality of the ϕ_i, if that is the case.

Weak solutions (I)

- We will apply the ideas developed in this lecture to a more specific problem, and obtain the Galerkin (Boris Galerkin, 1871-1945) or Ritz-Galerkin method, which ultimately leads to some of the more popular numerical methods for solving PDE, such as the finite element method.
- Consider the minimization problem

$$\min_u Y[u] = \int_\Omega \left(\frac{1}{2} |\vec{\nabla} u|^2 + \frac{1}{2} u^2 + f u \right) \mathrm{d}\vec{x},$$

where Ω is a bounded domain in $\mathbb{R}^n$ and f is a given continuous function satisfying, without loss of generality, $|f| \leq 1$ in Ω.

- The first variation is

$$\delta Y[u] = \int_\Omega \left(\vec{\nabla} u \cdot \vec{\nabla} \psi + u\psi + f\psi \right) \mathrm{d}\vec{x}.$$

Weak solutions (II)

- We seek a minimizer in the Sobolev space $H_1(\Omega)$, the Hilbert space with the inner product involving both the functions and their gradients, and hence ψ must belong also to $H_1(\Omega)$. Therefore the condition on the minimizer u is

$$\int_\Omega \left(\vec{\nabla} u \cdot \vec{\nabla} \psi + u\psi + f\psi \right) \mathrm{d}\vec{x} = 0 \quad \forall\ \psi \in H_1(\Omega). \qquad (*)$$

- If we assume that the minimizer is in the class $C^2(\Omega) \cap C^1(\bar{\Omega})$ and Ω has a smooth boundary, the standard Green plus continuity of the integrand argument leads to the Euler-Lagrange equation for this functional

$$-\Delta u + u = f,\ \vec{x} \in \Omega, \quad \partial_n u = 0,\ \vec{x} \in \partial\Omega. \qquad (**)$$

- The integral version is more general since it holds under the weaker assumption u is only once continuously differentiable. Hence we call $(*)$ the *weak formulation* of $(**)$.

The Galerkin method (I)

- **Theorem**. The weak formulation has unique solution u^*, which is actually a (local) minimizer of $Y[u]$, *i.e* it has a positive second variation.
- The proof of the theorem is not constructive: it defines u^* as the limit of an unknown sequence, which is guaranteed to exist due to the compactness properties in Hilbert spaces.
- The Galerkin method provides a practical algorithm to generate a sequence which converges to u^*. The idea is to construct a chain of subspaces $H^{(1)}, H^{(2)}, \ldots, H^{(k)}, \ldots$ with the properties

$$H^{(k)} \subset H^{(k+1)}, \quad \dim H^{(k)} = k, \quad \bigcup_{k=1}^{\infty} H^{(k)} = H_1(\Omega).$$

- This means that it exists a basis $\{\phi_k\}$ of $H_1(\Omega)$ such that $\phi_1, \phi_2, \ldots, \phi_k \in H^{(k)}$.

The Galerkin method (II)

- In each subspace $H^{(k)}$ we select a basis $\phi_1^k, \phi_2^k, \ldots, \phi_k^k$, where the superindex is added because the basis, although spanning also the same previous subspace, may be changed when going from $H^{(j)}$ to $H^{(j+1)}$.
- We denote by v^k the minimizer of $Y[u]$ in $H^{(k)}$. Remembering the definition of the Sobolev inner product,

$$\langle f, g \rangle_{H_1(\Omega)} = \int_\Omega (fg + \vec{\nabla} f \cdot \vec{\nabla} g) \mathrm{d}\vec{x}.$$

and since ψ is now any element of $H^{(k)}$, we can write the weak formulation of the minimizing problem as

$$\langle v^k, \phi_i^k \rangle_{H_1(\Omega)} = -\int_\Omega f \phi_i^k \ \mathrm{d}\vec{x}, \quad i = 1, 2, \ldots, k.$$

The Galerkin method (III)

- Expanding v^k in terms of the basis of $H^{(k)}$

$$v^k = \sum_{j=1}^{k} \alpha_j^k \phi_j^k$$

and substituting into the $H_1(\Omega)$ inner product one gets

$$\sum_{j=1}^{k} K_{ij}^k \alpha_j^k = d_i, \quad i = 1, 2, \ldots, k, \quad \text{GS}$$

$$\text{where} \quad K_{ij}^k = \langle \phi_i^k, \phi_j^k \rangle_{H_1(\Omega)} \quad \text{and} \quad d_i = -\int_\Omega f \phi_i^k \, \mathrm{d}\vec{x}.$$

- The Galerkin system (GS) is an algebraic system of k equations in k unknowns α_i^k. It can be shown that it has a unique solution for each k and that the v^k that are obtained converge strongly to u^*.

The Galerkin method (IV)

- Notice that the difference between the Ritz and Galerkin methods is that the expansion in the selected basis is performed at the level of the functional for the former and on the first variation for the later. Galerkin method yields directly a system of equations for the unknown coefficents, while in the Ritz method one gets a functional which depends on those coefficients and has then to be minimized with respect to them.
- It can be shown that, if in the Ritz method the same basis as in the Galerkin method is used, both yield the same system of algebraic equations. This is why the methods are sometimes fused and labeled as the Ritz-Galerkin method (or even the Rayleigh-Ritz-Galerkin method, in honor of Lord Rayleigh, 1842-1919, Nobel prize of Physics 1904).
- The Galerkin method, however, is more general than the Ritz one, since it can be applied to any weak formulation, whether derived from a a variational problem or not. Indeed, given any PDE of the form $L[u] = f$, where L is a linear or nonlinear differential operator, we can write

$$\langle L[u] - f, \psi \rangle = 0 \quad \forall\, \psi \in H,$$

  where H is a suitable Hilbert space. Using integral identities, some of the derivatives on u can sometimes be cast on ψ, obtaining thus a formulation that requires less regularity of the solution.
- The big issue is how to choose the spaces $H^{(k)}$, that is the basis $\{\phi_k\}$ of the Hilbert space. In some problems good approximations can be obtained using basis whose elements are well known functions, such as trigonometric functions, Bessel functions, Hermite polynomials or others, but generally one has to use numerically constructed basis. A very important class of such basis constitute the foundation for a numerical method called *finite elements*, which will be presented in the next lecture.

Finite elements (I)

- The *finite elements method* (FEM) is a numerical method for PDE that starts from a point of view completely different from that of the finite differences methods that we saw in the previous lecture.
- Indeed, while finite differences methods aim to discretize the PDE, yielding difference equations, FEM are an instance of the Galerkin method, and thus start from an integral point of view.
- To illustrate the essentials of FEM we will consider a canonical elliptic problem in two dimensions, namely the Poisson equation with Dirichlet homogeneous conditions:

$$-\Delta u = f,\ \vec{x} \in \Omega, \quad u = 0,\ \vec{x} \in \partial\Omega.$$

Finite elements (II)

- Remember that the weak formulation of a PDE is obtained by multiplying the equation by a test function in a suitable Hilbert space and then integrating by parts. In our case we get, using the boundary condition for u,

$$\int_\Omega \vec{\nabla} u \cdot \vec{\nabla} \psi \, \mathrm{d}\vec{x} = \int_\Omega f \psi \, \mathrm{d}\vec{x}.$$

- In view of this, the relevant Hilbert space is the one obtained from the completion of the $C^1(\Omega)$ functions that vanish on $\partial\Omega$. We denote this space by $\dot{H}_1$.

- Notice that we are using an scalar product that differs from the one defined previously for the Sobolev space, because the scalar product involves only the gradients of the functions and not the functions themselves. This does not destroy the nondegeneracy of the product, because any constant function must be the zero function due to the boundary condition.

Finite elements (III)

- Proceeding with the Galerkin method, we have to choose u in an increasing sequence of Hilbert spaces $\{H^{(k)}\}$ with basis $\phi_1, \phi_2, \ldots, \phi_k$, and write

$$u = \sum_{i=1}^{k} \alpha_i \phi_j,$$

  and then use the basis elements as test functions.

- This leads to a system $K\alpha = d$ with

$$K_{ij} = \int_\Omega \vec{\nabla}\phi_i \cdot \vec{\nabla}\phi_j \; \mathrm{d}\vec{x}, \quad d_i = \int_\Omega f\phi_i \; \mathrm{d}\vec{x}.$$

- The FEM was invented by Courant in 1943, and was originally extensively developed by mechanical engineers, although it has found applications in all areas of engineering. Due to its mechanical origins, the matrix K is usually called the *stiffness matrix*, and f is the force vector. The mathematical justification in terms of the Galerkin method came later.

Finite elements (IV)

- The special feature of the FEM lies in the choice of the family ϕ_i. The idea is to localize the test functions ϕ_i to facilitate the computation of the stiffness matrix. There are many variants of how to do this, but a very popular one is to use triangles to decompose the domain Ω into smaller regions.
- Assume that for a given (in general approximate) partition of Ω we have the triangles $\{T_j\}$ and the vertices $\{V_i\}$. We will not dwell into details, but a clever numbering is important in practice.
- Each test function is constructed to be linear in each triangle and continuous at the vertices (this, together with linearity, implies continuity at the edges, too). The shape taken by a test function in the triangles is called an *element*.

Finite elements (V)

- We associate a privileged test function to each vertex, according to

$$\phi_i(V_j) = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases}$$

- Since ϕ_i is linear on each triangle, the three conditions on the three vertices of each triangle determine ϕ_i uniquely. Furthermore, if the vertex V_i does not belong to the triangle T_j then ϕ_i is identically zero on T_j. The test functions look as tents of height unity over a given vertex, going linearly to zero at all the adjacent vertices, and remaining zero further away.

- Due to the localization of the test functions, the stiffness matrix is sparse, and, if a clever numbering is employed, the nonzero elements will be banded around the diagonal, greatly simplifying storage and computations.

- Another important consequence of this choice of test functions is that the numerical approximation of the solution at the vertices is

$$u(V_i) = \sum_j \alpha_j \phi_j(V_i) = \alpha_i .$$