Nummerical Programming II

Personal summary by Gilberto Lem

Summer 2018

Contents

1 ODE's: Initial Value Problems **2**
1.1 General Theory . . . 2
1.2 Runge Kutta and Multistep Methods . . . 4
1.3 Stiffness of IVP . . . 5

2 ODE's: Boundary Value Problems **6**
2.1 Solution Schemes . . . 6
2.2 Practical Issues . . . 6
2.3 BVPs as optimization problems . . . 7
2.4 BVPs as optimal control problems . . . 8

3 PDE's: Poisson Equation **9**
3.1 General Aspects of PDE's . . . 9
3.2 Poisson Equation: Fast Poisson Solvers . . . 9
3.3 Poisson Equation: Montecarlo Method "Random Walk on Spheres" . . . 11

4 PDE's: Finite Element **12**
4.1 Finite Element Spaces . . . 13
4.2 Assemblying the linear system . . . 14
4.3 Solving the Linear System. Iterative Solvers. . . . 15
4.4 Error Estimates . . . 17

1 ODE's: Initial Value Problems

1.1 General Theory

An Ordinary Differential Equation is presented in the following form:

$$y'(t) = f(t, y(t))$$

Where f is a given function, t is the independent (scalar) variable, and y is the dependent variable we want to find.

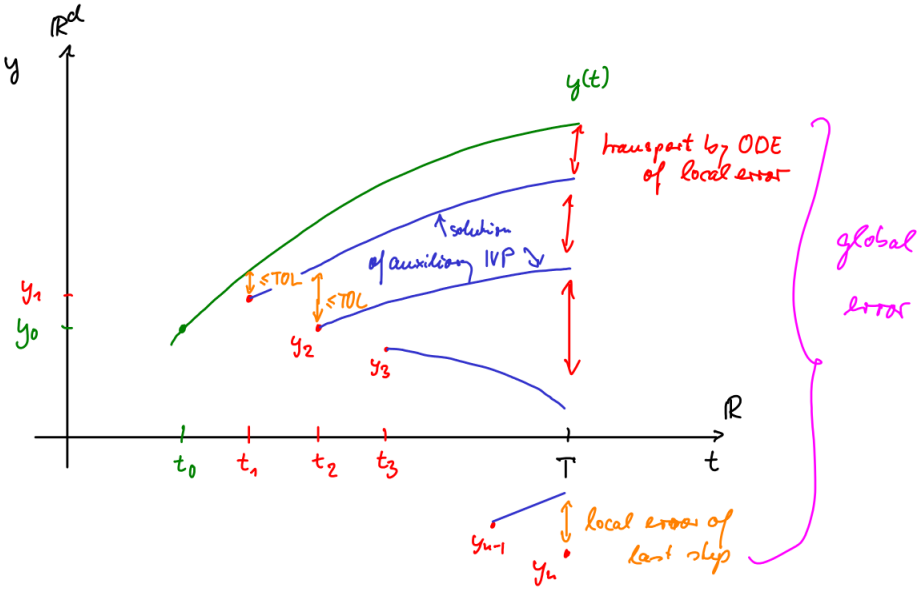
Solution lines for ODE's never cross, they can only touch, in which case uniqueness is lost.

An Initial Value Problem (IVP) means that we are provided with the value of the dependent variable at a specific value of the independent value: $y(t_0) = y_0$. IVP's have a general theory about existence and uniqueness:

- If f is continuous, then it **exists** a solution
- If both f and $\frac{\partial f}{\partial y}$ are continuous, then it exists a **unique** solution

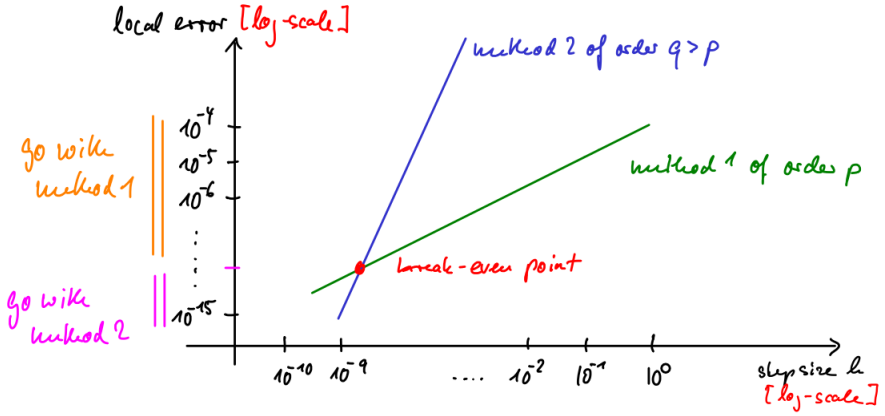
For already-made numeric implementations for solving IVP's, as the ones in MATLAB, it is important to consider the uniqueness criteria, for the machine to be able to obtain a valid solution, as well as the sensitivity to inputs of the ODE (take into account numerical error of initial values).

Local Error vs Global Error Before going into Error Estimation we need to define two types of errors: Local and Global. **Local Error** is the difference between the real solution of the ODE starting in the last grid-point and the new point y_k given by the numerical method. **Global Error** is the accumulation of those local errors, resulting in the difference between the last y_k given by the numerical method and the analytical solution of the ODE. Both kinds of error are illustrated in the following figure:



Order and Error of the Method The order p of the numerical method tells you how much the error changes when you modify the stepsize. One can express the error in terms of the order. The local error is $\mathcal{O}(h_k^{p+1})$, where h_k is the local stepsize, while the global error is $\mathcal{O}(h_{max}^p)$, where h_{max} is the maximum stepsize in the domain. It is important to note that a bigger order does not necessarily

imply a lower error. It depends on the stepsize. There comes a point where there is a break-even point, as we can see in the following figure:



In the IVP packages supplied for example by MATLAB, the machine controls just the **local** error to match the user supplied tolerance TOL. As in real life one does not have the analytical solution for the ODE to evaluate the real local error, one can use a better numerical method as the "real" solution. The standard today is consider the difference between the methods as the local error and still use as a solution the one provided by the "better" numerical method, even though the error we calculated would correspond to the error of the "worse" numerical method. This results in a conservative error estimate, and is called **local extrapolation mode**.

Choosing Tolerances In practice, for the error estimation, one uses the sum of two tolerances: the Relative Tolerance, which takes into account the value of the solution itself, and the Absolute Tolerance, for not being too strict when the value of the solution (and thus the RelTOL) is very small. To choose the values of the different tolerances one can use the following rule of thumb:

- **RelTOL:** adjusted to the measurement errors in the physical system
- **AbsTOL:** adjusted to the problem and physical units

MATLAB ODE methods In MATLAB, IVP's can be solved with different implemented methods, here is a summary with some of their characteristics:

Name	**Order** (stepsize control)	**Single/Multi step**	**Stiff/Nonstiff**
ode45	5(4)	s	nonstiff
ode23	3(2)	s	nonstiff
ode113	1-13	m	nonstiff
ode23s	3(2)	s	stiff
ode15s	1-5	m	stiff

1.2 Runge Kutta and Multistep Methods

Runge Kutta and Multistep Methods are used to solve IVP's. The idea behind them is to do a local estimation of the next value for the solution with the following exact expression:

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(\tau, y(\tau))\mathrm{d}\tau$$

Runge Kutta methods approximate the integral by a quadrature formula, while Multistep Methods replace f with a polynomial obtained interpolating its values in the last m steps.

1.2.1 Runge Kutta Methods

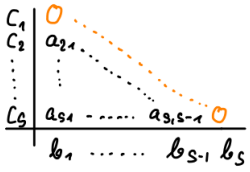
As said, Runge Kutta methods use a quadrature formula for approximating the integral. In a code-friendly form for evaluating the expression, one uses already-taken-care-of tables of coefficients c_i, a_{ij}, and b_j to compute:

$$y_{n+1} = y_n + h_n \sum_{j=1}^{n} b_j k_j$$

with

$$k_i = f(t_n + c_i h_n, \quad y_n + h_n \sum_{j=1}^{i-1} a_{ij} k_j)$$

The tables are commonly written in a Butcher-scheme:



As we have said, the estimation of the local error in all methods is done using two numerical methods of different order, and thus, in Runge Kutta methods we have to compute two different quadratures for each step. A nice thing of this approach is that between the two different methods only the b_j coefficients change, and thus we get to recycle the f-evaluations, which could be expensive.

Another trick used for making the computation efficient is is the **FSAL** (first-same-as-last). Here one uses the last f-evaluation of the last step as the first evaluation of the current step, to save one evaluation.

Another nice trick of Runge Kutta methods is that we can use them to generate dense outputs of y, i.e. not only get y_k, y_{k+1},etc. but get also any value between those points too. This is acheived by changing the coefficients b_j for polynomials $b_j(\theta)$, where $\theta \in [0, 1]$ and $\theta = 0$ corresponds to y_k and $\theta = 1$ corresponds to y_{k+1}. We can construct the polynomials to match the same error as the corresponding to the order of our method.

1.2.2 Multistep Methods

On the other side, instead of a quadrature use an interpolation for f taking the information of the last m points. The error of the interpolation is $\mathcal{O}(h^m)$, and the method is order m. Note thas this needs the computing of only one f-evaluation per step, and thus is useful for very expensive f.

A useful characteristic of multistep methods is that you can make very easily a numerical method variable order, however it has disadvantages, as that rapid stepsize changes degrade the quality of the interpolated polynomial, and thus step sizes have to be changed in a very conservative way.

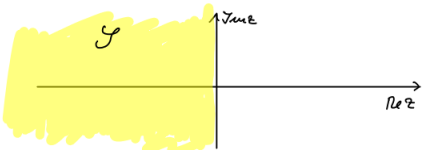
1.3 Stiffness of IVP

A problem is stiff if the solution being sought varies slowly, but there are nearby solutions that vary rapidly, so the numerical method must take small steps to obtain satisfactory results. It depends on the differential equation, the initial conditions, and the numerical method. What can be done about stiff problems? You don't want to change the differential equation or the initial conditions, so you have to change the numerical method. Methods intended to solve stiff problems efficiently do more work per step, but can take much bigger steps. Stiff methods are implicit.

Stiffness imply that one needs certain restrictions on the stepsize for the numerical method to be stable. Specifically, we need the real part of the eigenvalues of the linearized operator $D_y f$ around the point of interest to be negative. This defines some regions of stability for each numerical method. In general, the restrictions have the following form:

$$h|\lambda| \leq c_\alpha$$

A method is called A-stable if the stability region is all space such that $\mathrm{Re} z < 0$:

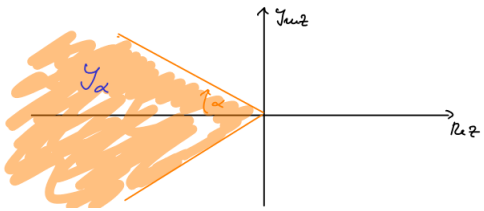


For example, for the `ode45` method in MATLAB, $c_\alpha = 3.31$.

Stiff solvers use implicit methods to overcome the stability problem. This methods imply solving nonlinear systems of equations, which we should solve using Newton iterations (iterate by fixed-point iteration would be a disguised explicit method). The cost for each step of this kind of methods is:

- $\#f$-evaluations = s
- Build the $D_y f$ (Jacobian) matrix. (Automatic differentiation)
- # steps in Newton iteration

Multistep methods add an additional restriction for the stability region. Depending on the number of steps taken, there is an angle α that restricts the region. Here the method is called A(α)-stable.



2 ODE's: Boundary Value Problems

Initial Value Problems have one or more conditions in one point of the independent variable domain. Boundary Value Problems have conditions in more than one point of the independent variable domain.

Unlike in IVP's, there is no general theory for BVP's regarding existence and uniqueness. There can be infinitely many solutions to a problem or none.

Many times, the problem is formulated with an additional parameter, which is part of the ODE and is going to change with the solution.

2.1 Solution Schemes

Shooting Method One solution technique to solving BVP's is the **shooting method**. It solves an IVP for an initial parameter in the ODE, setting with that a nonlinear system of equations and using Newton's method tries to satisfy all the boundary conditions. This method can lead to numerical instability.

Multiple Shooting Another solution scheme is **multiple shooting**. This adds multiple shooting points and solves a more stable but bigger nonlinear system of equations.

Global Solvers for BVP's This is the most general scheme. They take an implicit method (for example the trapezoidal rule) and with that method applied to every point on the grid, plus the boundary conditions, define a $(N+1)d$-dimensional nonlinear system of equations, where N is the number of grid points and d the dimension of the problem. The nonlinear system is solved by Newton iteration. This methods are available in MATLAB with the names bvp4c and bvp5c, where the number 4 or 5 refer to the order of the numerical scheme used (for example the trapezoidal rule would be order 2).

2.2 Practical Issues

2.2.1 Initial Guesses

Newton's method is iterative, and thus we need a useful first guess for the solutions and for the parameters needed. These guesses are made with some preliminary mathematical analysis (as taking approximations), or prior experience from the problem.

Take for example the following ODE:

$$\epsilon y' = \sin^2(x) - \lambda \frac{\sin^4(x)}{y}$$

where λ is the parameter, and ϵ is a small number. As ϵ is small, we could approximate $\epsilon y' = 0$ and with that get an initial guess for $\lambda = 1$ and $y(x) = \sin^2(x)$. We could generalize this to $\epsilon y' = \mathcal{O}(\epsilon)$ to see the order of magnitude of the approximation. This is called **asymptotic analysis.**

Another technique is **Homotopy (parameter continuation).** Imagine we have a problem that has a fixed parameter (for example a fixed Reynolds number in a fluid problem), and that value makes the ODE difficult to solve numerically. One way to deal with this is to solve the BVP for a different and easier value of the parameter, and take that solution as an initial guess for solving the problem with a parameter a little closer to the real value, and iterate.

2.2.2 Parameters not part of the ODE

Imagine the case where the parameter is not part of the ODE, but it is the length of the interval of the independent variable e.g. parameter T in $t \in [0, T]$. In this case we could make a change of variable, using an **auxiliary independent variable** $t = Ts$, $s \in [0, 1]$, and modify the BVP accordingly.

2.3 BVPs as optimization problems

Suppose we want to optimize a function R. In the optimal point:

$$\nabla R = 0$$

Now, if we want to have an additional restriction on the solution, we could express the restriction as $G = 0$, where G is a function. To get a solution for the minimization problem that satisfies the restriction condition, we need the following expression to be true:

$$\nabla R = \lambda \nabla G$$

where λ is a scalar called **Lagrange multiplier**. This means that the gradient of R has the same direction as the gradient of G. For compact notation, we can define an additional function:

$$I = R - \lambda G,$$

and state the problem as simply

$$\nabla I = 0$$

Now imagine that the problem can be stated as an integral equation. Then our optimization problem could be expressed as:

$$I = \int_0^1 F\left(x, y, y'\right) dx = \min!$$

The solution to that problem is provided by **Calculus of Variations** techniques, and is given by the y that solves the **Euler-Lagrange equation**:

$$\frac{\partial F}{\partial y} = \frac{\mathrm{d}}{\mathrm{d}x} \frac{\partial F}{\partial y'}$$

This transforms the problem into a differential problem. If we additionally had some restrictions for y in different parts of the domain of x, as $y(a) = y_a$ and $y(b) = y_b$, then this optimization problem is equivalent to solving a BVP.

If our problem needs to satisfy some restrictions, then our function F should be of the form $R(x, y, y') - \lambda G(x, y, y')$.

One common trick for solving this kind of problems, when the restriction is given in an integral form, is to define an additional differential equation with boundary problems, which represents the restriction.

2.4 BVPs as optimal control problems

Suppose we have a physical phenomenon described by an ODE:

$$x' = f\left(x(t), u(t)\right),$$

with its respective boundary conditions $x(0) = x_0$, $x(1) = x_1$. Here the differential part of the model is for $x(t)$, however, the problem also considers an additional function $u(t)$, which is called **control**. This could be for example a car moving, where the control is how much the accelerator is being pushed.

Suppose we have a quantity of interest from the problem which we want to optimize, for example, minimize the pollution of the car. And that that quantity can be expressed as a function of $x(t)$ and $u(t)$. Then we can formulate the optimization problem:

$$\mathcal{J} = \int_0^1 \varphi\left(x(t), u(t)\right) dt = \min!$$

We want to find the function $u^*(t)$ which optimizes $\mathcal{J}$.

This problem could be equivallently expressed as an optimization problem for $\mathcal{J}$,with a restriction given by $x' = f$. So we can define a Lagrangian:

$$\mathcal{L} = \varphi(x, u) + \lambda\left(f(x, u) - x'\right)$$

Applying Euler-Lagrange equations with respect to x, λ and u (considering all of them as functions) we get:

- w.r.t. x: We call it the **adjoint equation**
- w.r.t. λ : The original ODE of the problem
- w.r.t. u: An ODE on u, which solves for u^*

For easying notation, we can define the **Hamilton function** as:

$$H(x, \lambda, u) = \varphi(x, u) + \lambda f(x, u)$$

Note that this is very similar to the Lagrangian, just removing the x'.

Then the three equations are:

- Adjoint equation: $\lambda' = -\frac{\partial H}{\partial x}$
- Original ODE: $x' = f$
- Equation for u^*: $\frac{\partial H}{\partial u} = 0 \Rightarrow u^* = g(x, \lambda)$

In practice we need to solve this problem with the computer. We have a system of ODE's (adjoint and original) and we have the computation of the optimal u for each of the numerical method's solutions for x and λ. As this is a boundary value problem, we need an initial solution. This is generally done proposing a clever guess for $u(t)$,and with that guess solve for an initial solution for x with the original ODE and an initial solution for λ with the adjoint equation.

3 PDE's: Poisson Equation

PDE's depend on more than one independent variable on a domain Ω with boundary $\delta\Omega$. This is often stated as one independent vector $x \in \mathbb{R}^n$, and time is often separated: $x = (x', t)$. There is no general theory for solving PDE's, even in the linear case i.e. when there is a spacial dimension and a time dimension.

3.1 General Aspects of PDE's

3.1.1 Types of PDE's

There are several types of PDE's, and each one has a different kind of set conditions which lead to proper problems.

- **Elliptic PDE's**, such as Laplace Equation or Poisson Equation, need stationary **boundary conditions**, such as Dirichlet or Neumann conditions on $\delta\Omega$.
- **Hiperbolic PDE's** need a time variable, for example the wave equation $u_{tt} = c^2\Delta u$, and need only an **initial condition** $u(x, t = 0) = u_0(x)$, $x \in \Omega$.
- Parabolic PDE's model infinite speed of information transport, for example the heat equation $u_t = k\Delta u$. They need **initial and boundary conditions**, i.e. they need $u(x, t = 0) = u_0(x)$ on Ω, and $u(x, t) = \varphi_t(x)$ on $\delta\Omega$ for $t \geq 0$.

3.1.2 Well-Posedness

The PDE is well posed if and only if there **exists** a **unique** solution for every input, and the solution **depends continuously** on the input i.e. noise in data does not blow up in the solution.

3.1.3 Principles of deducing PDE's

One common way to deduce PDE's is the **variational principle**, which uses the Euler-Lagrange equations for reformulating an integral optimization problem into a differential problem. It may be useful to use Taylor expansions or other simplifications to be able to formulate a simpler problem.

One important tool used in PDE theory is multidimensional integration by parts:

$$\int_\Omega \frac{\partial u}{\partial x_j} v \mathrm{d}x = \int_{\delta\Omega} uv \cdot n_j d\sigma - \int_\Omega u \frac{\partial v}{\partial x_j} \mathrm{d}x$$

When u is a scalar field, this yelds the **divergence theorem.**

3.2 Poisson Equation: Fast Poisson Solvers

One of the most common PDE's is the Poisson equation. It has the following form:

$$-\Delta u = f, \qquad \text{on } \Omega \subset \mathbb{R}^d$$

$$u|_{\delta\Omega} = 0$$

We will consider the case of the hypercube $\Omega = [-1, 1]^d$

3.2.1 1D problem

In the 1D problem, one can discretize and get the operator Δ as a matrix T_n with stencil $[-1, 2, -1]$, a tridiagonal symmetric positive definite matrix.

$$T_n u_n = h^2 f_n,$$

where h is the discretization step. This needs $\mathcal{O}(n)$ flops for solving.

$\mathcal{O}(n)$ is pretty cheap, however in the problem with more dimension, this is nice complexity not achieved. As an alternative we can diagonalize T_n in the following form:

$$S_n T_n S_n = D_n,$$

where $D_n$is a diagonal matrix with the eigenvalues of T_n, and S_n is a unitary matrix with the eigenvectors of T_n, called the **discrete sine transform (DST)**. This transformation is useful because the DST can be easily obtained from to the Discrete Fourier Transform, and we can obtain the DFT with the Fast Fourier Transform algorithm in $\mathcal{O}(n \log n)$ operations (note that $n \log n$ is bigger than n, so for 1D this is not very useful).

3.2.2 2D problem

Now we have

$$-\Delta u = (-\frac{\partial^2}{\partial x_1^2} - \frac{\partial^2}{\partial x_2^2})u = f$$

Discretizing, u becomes a matrix, where x_1 varies in the first dimension and x_2 varies in the second dimension. We can apply the T_n matrix to the row indices by multiplying u by the left, and apply it to the columns by multiplying it by the right. The linear system is now:

$$T_n u_n + u_n T_n = h^2 f_n,$$

which has $N = n^2$ unknowns and equations to be solved, and as the system is not tridiagonal anymore, it needs $\mathcal{O}(N^3)$ flops to be solved. But we can transform the problem with the DST. Multiplying the equation from both sides with S_n :

$$D_n \tilde{u}_n + \tilde{u}_n D_n = h^2 \tilde{f}_n$$

As D_n is diagonal, we can solve easily for the element j, k:

$$(\tilde{u}_n)_{jk} = h^2 \frac{\tilde{(f_n)}_{jk}}{\lambda_j + \lambda_k}$$

As an algorithm:

1. Compute $\tilde{f}_n = S_n f_n S_n$ $[\mathcal{O}(N \log N)]$
2. Compute $\tilde{u}_n = h^2 \tilde{f} / D_n$ (elementwise division) $[\mathcal{O}(N)]$
3. Backtransform $u_n = S_n \tilde{u}_n S_n$ $[\mathcal{O}(N \log N)]$

So in total this has $\mathcal{O}(N \log N)$, which is useful even for orders of 10^{6-9}

3.2.3 Hypercube

The same algorithm is used for $d > 2$, where we have $N = n^d$. Here the DST is successily applied to each coordinate, and instead of matrices we have d-dimensional tensors.

In general, the error is given by the error of the central difference approximation made: $\mathcal{O}(h^2)$. As the number of points in each dimension is $N^{1/d}$, we have $h = N^{-1/d}$, and thus our error is $\mathcal{O}(N^{-2/d})$. Note that this is bad, because when $d \to \infty$, $N^{-2/d} \to 1$, and thus our algorithm has error of $\mathcal{O}(1)$. This is called **curse of dimensionality**. If we want to acheieve an absolute error of $TOL = N^{-2/d} \Rightarrow N = TOL^{-d/2}$, then we need a complexity of $\mathcal{O}(N \log N) = \mathcal{O}(d{\cdot}TOL^{-d/2} \log TOL)$.

3.3 Poisson Equation: Montecarlo Method "Random Walk on Spheres"

We take as a model the homogeneous Poisson equation (Laplace equation):

$$-\Delta u = 0, \qquad \text{on } \Omega \subset \mathbb{R}^d$$

$$u|_{\delta\Omega} = w$$

Functions that satisfy this equation are called harmonic, and satisfy the **mean-value property**. This property says that the value of $u(x_0)$ is equal to the mean value of $u(x)$ on a sphere with center on x_0. This added with brownian motion magic lead to the following algorithm, called Random Walk on Spheres (WoS):

1. Get the radius of the largest sphere that fits into Ω around x_j
2. Sample uniformily from that sphere to get x_{j+1}
3. If $x_{j+1} \in \delta\Omega$ (with a tolerance h), stop and return $w(x_{j+1})$ as an estimate of $u(x)$

3.3.1 Error

Monte-Carlo methods have a statistical error of $\mathcal{O}(1/\sqrt{N})$, where N is the number of samples taken, and the error of introducing a tolerance in the boundary is $\mathcal{O}(h)$, thus the total error is $\mathcal{O}(1/\sqrt{N}) + \mathcal{O}(h)$. To make both error contributions equally large, we define $h \coloneqq 1/\sqrt{N}$, and finally our error is $\mathcal{O}(1/\sqrt{N})$.

3.3.2 Complexity

The expected number of steps to teach the boundary scales as $\mathcal{O}(d\,|\log h|) = \mathcal{O}(d\,|\log N^{-1/2}|) = \mathcal{O}(d \log N)$. Additionally, the number of flops per step is $\mathcal{O}(d)$. So, our total complexity is $N{\cdot}\mathcal{O}(d \log N){\cdot}\mathcal{O}(d) = \mathcal{O}(d^2 N \log N)$

If we want to acheive an absolute error $TOL = 1/\sqrt{N}$, then we have a complexity of $\mathcal{O}(d^2 TOL^{-2} \log TOL)$. Note that for WoS we do not have d in the exponent, whereas in the best grid-based method we have a complexity of $\mathcal{O}(TOL^{-d/2})$. This sets a break even point. For example for $TOL = 10^{2-3}$ the breakeven is in $d \approx 5 - 6$, and thus for that tolerance, if $d \geq 7$, we would go with WoS.

3.3.3 Summary

WoS is nice because it is mildly dependent on d, embarassingly parallel, simple to program and uses only pointwise evaluations. However it has a limited accuracy.

4 PDE's: Finite Element

Weak Formulation

Finite Element Method works works for differential equations with elliptic operators in divergence form:

$$Pu(x) = -\nabla \cdot (\mathbb{A}(x)\nabla u(x)) + a_0(x)u(x)$$

Where $\mathbb{A}(x)$ is called the diffusion matrix, and it is symmetric and uniformly positive definite. One can define a BVP with this kind of operator:

$$Pu = f \ \text{ on } \Omega, \qquad u|_{\Gamma_D} = 0, \ \frac{\partial u}{\partial n_{\mathbb{A}}} = g|_{\Gamma_N}$$

where f is the load function, Γ_D is a Dirichlet part of the boundary $\delta\Omega$ and Γ_N is the Neumann part of the boundary. $n_{\mathbb{A}}$ is the normal direction in the boundary rotated by $\mathbb{A}$ (if $\mathbb{A} = I$, then $n_{\mathbb{A}} = n$).

The idea of the weak formulation is to go from having two derivatives (divergence and gradient) to having just one:

$$Pu = f \implies \int_\Omega Puv\mathrm{d}x = \int_\Omega fv\mathrm{d}x$$

The week formulation and the original PDE are equivalent if the PDE has a solution. Here v is a test function. We introduce v such that $v \in V$, where V is the space of functions which have a value of 0 in the Dirichlet boundary (to satisfy the PDE Dirichlet boundary). Applying integration by parts to the first part of the new approach we can throw one derivative from P to v. Applying the divergence theorem and the boundary conditions, we get the following form:

$$\int_\Omega \langle \mathbb{A}\nabla u, \nabla v\rangle \, dx + \int_\Omega a_0 uv dx = \int_\Omega fv dx + \int_{\partial\Omega} gv d\sigma$$

We notice that the result of the Left Hand Side of the equation is a number, and it is linear w.r.t. both u and v. This is called a **bilinear form.** The RHS is a **linear form** with respect to v. Naming the bilinear form a and the linear form λ, we can reexpress as:

$$a(u, v) = \lambda(v)$$

There is another way to get this last equation, an alternative to the weak formulation that yields the same result. This alternative parts from the definition and the optimization of a quadratic functional dependent on u and $v \in V$. In this approach, the functional is minimized when our PDE is satisfied. However, according to Dirichlet's principle, not every quadratic functional bounded from below has a minimum, which would mean that this approach does not find a solution for our PDE. For finding a solution, one would have to find the correct function space V to make the functional to have a minimum. We should have a solution if V is taken large enough, but also we do not want it to be too large because we want u to be unique and computable. **Sobolev spaces**[1] satisfy this conditions. For our particular problem, $H^1(\Omega)$ is enough.

The equation with the linear forms, together with the restrictions on V, define the **weak formulation:**

$$a(u, v) = \lambda(v), \qquad v \in V \coloneqq \left\{u \in H^1(\Omega) : \ v|_{\Gamma_D} = 0\right\}$$

[1] Intuitively, a Sobolev space is a space of functions with sufficiently many **(weak)** derivatives for some application domain, such as partial differential equations, and equipped with a norm that measures both the size and regularity of a function. The norm of a function w.r.t. a Sobolev Space $H^m(\Omega)$ is the sum of all the L2 norms of its derivatives of order 0 to m.

Ritz-Galerkin Method

The idea of the Ritz-Galerkin method is to take a finite dimensional function subspace $V_h \subset V$, where h aludes to a grid size. Being $\{\varphi_i\}$ the basis of V_h, then we could express u_h and v_h (which are discrete representations of u and v) as a linear combination of these basis functions. Plugging these into our equation would result in the formation of a linear system of equations $A_h u_h = f_h$.

This process is called **assemblying.** A_h is called the **stiffness matrix**, f_h is called the **load vector**.

The stiffness matrix, as results from a bilinear form, is symmetric and positive definite.

To do List for FEM

For any problem that we would like to solve with FEM, defined in Ω, we would have the following steps to follow:

1. Choose the suitable geometric spaces to describe our domain
2. Choose the suitable function spaces $V_h \subset H^1(\Omega)$ to represent our solution. These are general piecewise polynomials. The fact of being piecewise allows geometric flexibility and numerical stability, and the fact of being polynomials allow different approximation powers by interpolation.
3. Assembly (set up the linear system $A_h u_h = b_h$). This requires a choice of basis. As we want A_h to be sparse for storage and computing reasons, we would like to choose a basis that is as local as possible.
4. Solve the linear system
5. Estimate the error of our solution (a posteriori)
6. Adaptively refine V_h until reaching a desired error criteria

The following sections elaborate on these steps.

4.1 Finite Element Spaces

For describing our domain, we must divide it in cells. For example in 2D we could choose only triangles, only hexagons, or any combination of several 2D figures. We call this discretization a **mesh**. Here are some definitions and concepts used in the geometry aspects of FEM:

- The set of cells at a certain level h is notated as J_h, and a specific cell of J_h as T.
- h represents the maximum diameter of a cell $T \in J_h$.
- The union of all the cells is an approximation of our domain, notated by $\bar{\Omega}$, and the intersection between different T_j is null or a lower dimensional space representing a boundary.
- X_h is the space of piecewise polynomials on J_h. A specific polynomial is $P_T = \{v_h|_T : v_h \in X_h\}$.
- A mesh can be characterized with the number of nodes, the number of edges and the number of edges. These variables are notated by n_0, n_1 and n_2 respectively.
- One could express restriction on the admissibility of a mesh in terms of the mesh variables. For example, if the piecewise polynomials chosen are of the Lagrange family of order m, then J_h is an admissible mesh if $\dim X_h = n_0 + (m-1)n_1 + \frac{1}{2}(m-1)(m-2)$. If they are chosen Cubic Hermite, which means having information of the value in each node and the derivative in each node, and an extra center of gravity, we would need $\dim X_h = 9n_0 + n_2$.

- In every 2D domain, the following geometric relation is satisfied: $n_0 - n_1 + n_2 = 1 - n_{holes} =: \chi$. n_{holes} if the number of physical holes inside the domain, independently of their size. χ is called the **Euler characteristic**, and it is a property of the domain, independently of its subdivision.

4.2 Assemblying the linear system

Assemblying means setting up the linear system $A_h u_h = b_h$. This requires a choice of basis. As we want A_h to be sparse for storage and computing reasons, we would like to choose a basis that is as local as possible. One example of a choice for a basis is the **nodal basis**. It is defined:

$$\varphi_j(x_k) = \delta_{jk}$$

This basis is commonly utilized because the coefficients of u_h with respect to this basis are just the function values at the grid points, so u_h is then just the piecewise linear interpolation of the values in the nodes.

However, independently of the chosen basis, the process of assemblying the system is the same. These are the two main ideas behind the process:

- Express the integral over the whole domain as a sum of integrals in each element $T \in J_h$.
- Transform T to a reference cell $\hat{T}$, to reach a simple, efficient framework and to be able to precompute a lot. (In FENICS the precomputing is done in the JIT-compile step). One can define this transformation to the reference cell with a map given by a matrix notated as B_T, which would be the Jacobian of the transformation.

4.2.1 Load Vector

In the weak formulation, we had the load function in the term $\int_\Omega f v dx$. After applying the Ritz-Galerkin method (expressing v as a linear combination of the basis of the function space V) we now want to find $(b_h)_j = \int_\Omega f \varphi_j dx$

We can express f too as a linear combination of our basis functions $f \approx \sum_k \alpha_k \varphi_k$ and then compute all the possible combinations of $\langle \varphi_j, \varphi_k \rangle$. For computational reasons, instead of iterating over j, i.e. going element by element of the resulting vector, and then check the cells in which the integral is not zero, we iterate once over all cells and then update accordingly. As an algorithm:

- $b_h = 0$
- for all $T \in J_h$:
  - for k, j being nodes of T :
    - update $(b_h)_j \to (b_h)_j + |\det B_T| \, M \alpha_k$

Where $M = \int_{\hat{T}} \hat{\varphi}_j \hat{\varphi}_k \mathrm{d}\hat{x}$ has the information about the integrals for different configurations, which is invariant and can be precomputed.

4.2.2 Stiffness Matrix

One can iterate over the elements of the matrix, to have

$$\int_T \langle \nabla \varphi_k, \nabla \varphi_j \rangle \, \mathrm{d}x = |\det B_T| \int_{\hat{T}} \left\langle \left(B_T^T\right)^{-1} \nabla \hat{\varphi}_k, \left(B_T^T\right)^{-1} \nabla \hat{\varphi}_j \right\rangle \mathrm{d}\hat{x}$$

4.2.3 Matrix-Free Assembly in FEM

If the linear system $A_h u_h = b_h$ gotten from a FEM analysis of a PDE is solved by an iterative method, that employs just matrix-vector multiplications of the form $A_h r_h$, then a black-box for setting up the operation $r_h \to A_h r_h$ would suffice to solve it. This means that we do not need to assembly the matrix A_h, saving storage and possibly computing time. This way of computing also allows reuse of the code of assemblying the load vector. This is useeful when the number of iterations k is not too large:

$$k \lesssim \frac{1}{2} c_A \frac{\text{nnz}(A)}{\#DOF}$$

where $\#DOF$ is the number of nodes, and $\text{nnz}(A)$ is the number of nonzero values of A_h. c_A is a constant typically between 1 and 2.

4.3 Solving the Linear System. Iterative Solvers.

We part from the linear system of equations we got from the FEM approach, for a PDE with boundary conditions valid on $\mathbb{R}^d$:

$$A_h u_h = b_h,$$

where A_h is symmetric, positive definite and sparse. The dimension of the system is $N = \mathcal{O}(h^{-d})$. One way to solve this system would be to get the Cholesky decomposition of A_h, however, as the dimension of the system increases this starts being too computationally expensive.

The idea of an iterative method consists on having a solver that gets better with time, and can be interrupted. This kind of solvers can be modeled and we can get the number of iterations necessary to reach the discretization error (our better approximation of the solution, the error we would get by solving the system directly). The following table includes relevant information for some iterative solvers:

Method	**Complexity (2D)**	**Complexity (3D)**	**Iters. to reach discr. error**
Gauss-Seidel method	$\mathcal{O}(N^2)$	$\mathcal{O}(N^{5/3})$	$\mathcal{O}(h^{-2})$
Krylov-space method: CG	$\mathcal{O}(N^{3/2})$	$\mathcal{O}(N^{4/3})$	$\mathcal{O}(h^{-1})$
Multigrid/Multilevel methods	$\mathcal{O}(N)$	$\mathcal{O}(N)$	$\mathcal{O}(1)$

4.3.1 Preconditioning

If we take our equation and multiply it by a matrix B_h we get

$$B_h A_h u_h = B_h b_h$$

The number of iterations needed to reach a certain RelTOL is dependent on the condition number of the matrix ($\mathcal{O}(\sqrt{\kappa(A_h)})$). Then if we reduced the condition number we could reduce the number of iteration. We would like to find a matrix B_h such that $\kappa(B_h A_h) \ll \kappa(A_h)$. The optimal thing would be to choose A_h^{-1} to get a condition number of 1, but that is unrealistic.

4.3.2 Classical Iterative Methods

Another idea is to decompose the matrix. One option is $A = M - N$, which would yield:

$$Mx = Nx + b$$

We can establish with this equation a fixed point iteration and get

$$x^{k+1} = M^{-1} N x^k + M^{-1} b$$

If we define M as the diagonal of the matrix A (call it D), and $-N$ to be the rest, then we would easily get $M^{-1} = D^{-1}$. This is called the **Jacobi Method.**

If we instead choose M to be the diagonal **and** the left triangular part of A (call them D and L respectively), then we would get $M^{-1} = (D-L)^{-1}$. This may sound like a difficult matrix to invert, but this configuration means that $M^{-1}r$ is simply a forward substitution, which would be $O(N)$ because of the sparsity. This is called **Gauss-Seidel Method.** A variant of that would be its **symmetric version** which consists on using $B \coloneqq (D-R)^{-1}D(D-L)^{-1}$.

Incomplete Cholesky Decomposition As we have commented, Cholesky decomposition is too expensive for large N, however we could define an alternative Cholesky decomposition where

$$A_h \text{ " } = \text{ " } \tilde{L}_h \tilde{L}_h^T,$$

where that equality is valid only in the non-zero elements of A_h. This provides a preconditioner $B = \tilde{L}\tilde{L}^T$,called **ICC preconditioner**. This takes $\mathcal{O}(N)$ operations.

4.3.3 Conjugate Gradient method (CG)

Conjugate Gradient method is an iterative algorithm to find the numerical solution of systems of linear equations whose matrix is symmetric and positive-definite. We can apply Galerkin's idea of taking the inner product of the whole equation with a vector y to get:

$$Ax = b$$

$$\langle Ax, y\rangle = \langle b, y\rangle$$

$$a(x_m, y_m) = \lambda(y_m), \quad \forall y_m \in V_m$$

The idea in CG is to define $V_m = \text{span}\left\{b, Ab, AAb, ..., A^{m-1}b\right\}$, where m would be the number of iteration. This provides that one only needs A and b to generate V_m.

The CG algorithm is the following:

- Define initial x_0, $r_0 = b - Ax_0$, $p_1 = r_0$
- for $m = 1, 2, 3, ...$
  - $\alpha_m = \frac{\langle r_{m-1}, r_{m-1}\rangle}{\langle p_{m-1}, Ap_m\rangle}$
  - $x_m = x_{m-1} + \alpha_m p_m$ -> Update x
  - $r_m = r_{m-1} - \alpha_m A p_m$ -> Update residual
  - $\beta_m = \frac{\langle r_m, r_m\rangle}{\langle r_{m-1}, r_{m-1}\rangle}$ -> How much did the residual changed in the last iteration
  - $p_{m+1} = r_m + \beta_m r_m$ -> Modified residual that includes information about the change of the residual with respect to the last iteration

Complexity The complexity of this method is the number of iterations times the complexity per iteration.

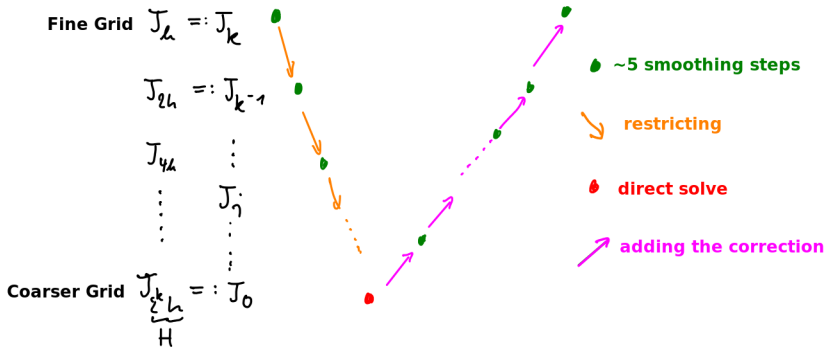
Each iteration needs $\mathcal{O}(N)$ flops for the inner product and scalings, plus the flops needed for matrix-vector-multiplication (which is $\mathcal{O}(N)$ in FE/FD methods, because of sparsity). So every iteration needs in total $\mathcal{O}(N)$.

The number of iterations needed with this method to reach a Relative Residual of TOL is $\approx \frac{1}{2}\sqrt{\kappa(A_h)}\log\frac{2}{TOL}$. As the condition number of this kind of matrix is $\mathcal{O}(h^{-2})$, then the number of iterations is $\mathcal{O}(h^{-1})$.

In total, the algorithm needs $\mathcal{O}(h^{-1}N) = \mathcal{O}(N^{1+1/d})$

4.3.4 Multigrid Methods

The general idea of Multigrid Methods is to use a fixed number of smoothing steps (making the residual smooth) on a fine grid ($\mathcal{O}(1)$ operations), then switch to a coarser grid to make the residual rough again and repeat. The process is the following:



This is called a V-cycle, and can be used as a preconditioner for CG. The complexity of the V-cycle is $\mathcal{O}(N_k) = \mathcal{O}(2^{dk}N_0)$, where N_0 is the number of points in the coarser grid.

The restriction and correction steps are very simple if we arrange for the FE spaces to have nested functions. That is why Multigrid Methods imply choosing a different basis of functions than the one we usually utilize in FEM. We use a **hierarchical basis** instead of a nodal basis:

At the left of the picture we have the nodal basis and at the right the hierarchical basis. In the hierarchical basis each color represents a different level.

For the case of the Poisson equation, applying the bilinear form in the hierarchical basis results in a diagonal A_h matrix. This means that if we take $D_h = A_h^{-1}$ as a preconditioner, we would get $B_h A_h = I$, which implies a condition number of 1. In general for elliptic PDEs in 1D, the matrix is no longer diagonal, but still the condition number that results of its preconditioning is independent of h (it is $\mathcal{O}(1)$).

There could be the case where we do do not have access to a history of grid refinement. In this case we could apply an alternative method called **Algebraic Multi-Grid Method (AMG)**, which builds a refinement structure from the sparsity pattern of A_h on a purely algebraic level.

There could be also the case where there is refinement but you cannot or do not want to keep it in storage. A well suited method for this case is the **Cascadic Multigrid.** This consists on starting with a coarse grid and solving the system directly. Then take that solution and interpolate it to a finer mesh, to take as initial guess in an iterative method (a preconditioned CG); then repeat the process until reaching the finest desired grid.

4.4 Error Estimates

From interpolation theory, we are supposed to know that to interpolate a function u using V_h, we get the *maximum* error

$$\|u - w_h\|_\infty \leq ch^{k+1}\|u^{(k+1)}\|_\infty,$$

where $w_h \in V_h$ is the interpolation of u. This would lead to think, as in FEM we are making an interpolation, that the error would be:

$$\|u - u_h\|_{L^2} \leq ch^{k+1}\|u^{(k+1)}\|_{H^{k+1}}$$

In short, one would expect error of $\mathcal{O}(h^{k+1})$.

If we have $u \in V$ and $u_h \in V_h$, and we compute the weak form for both:

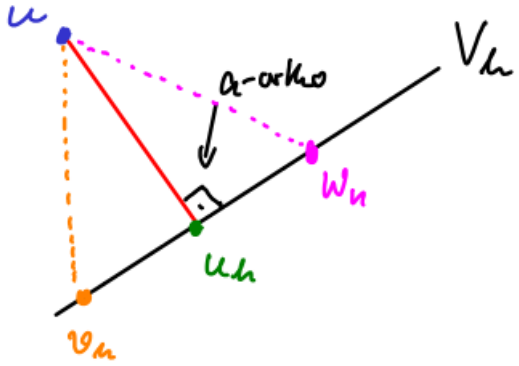
$$a(u, v_h) = \lambda(v_h)$$

$$a(u_h, v_h) = \lambda(v_h)$$

If we substract them we get that $a(u - u_h, v_h) = 0$, which means that the error is orthogonal to V_h with respect to a. [2]This property of the error is called the **Galerkin orthogonality of the FE solution.** Note that this means that u_h is the projection of u on V_h, and thus:

$$\|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a$$

One can see this geometrically with the following figure:



Note that in the figure we are considering the same approach for the interpolation w_h,[3] so now we have two restrictions: $\|u - u_h\|_a \leq \|u - v_h\|_a$, and $\|u - u_h\|_a \leq \|u - w_h\|_a$. From the last expression, we would get $\|u - u_h\|_a = \mathcal{O}(h^{k+1})$, but as $\|u - u_h\|_a$ measures also the derivative, we loose one power of h, so $\|u - u_h\|_a = \mathcal{O}(h^k)$.

$\|\cdot\|_a$ and $\|\cdot\|_{H^1}$ are equivalent up to constants, thus we have $\|u - u_h\|_{H^1} = \mathcal{O}(h^k)$. For other norms as L^2 and L^∞, the approach to get the error is completely different.

[2]In applications, $\|\cdot\|_a$ often has a physical meaning.

[3]For dimensions ≥ 3, one would have to use alternative approximation schemes as Clement-interpolation.