Lecture Note #1 **MTH653: Advanced Numerical Analysis**
Spring 2025 Instructor: Will Pazner

Advection Equation and Finite Volume Method

1 Model Problems

We consider as model problems the **advection equation** and **advection–reaction equation**. In both cases, there is a prescribed *advection velocity* $\boldsymbol{\beta} : \Omega \to \mathbb{R}^d$.

Time-dependent advection equation

We look for solution $u : \Omega \times [0, T] \to \mathbb{R}$ satisfying

$$(*) \begin{cases} \dfrac{\partial u}{\partial t} + \nabla \cdot (\boldsymbol{\beta} u) = 0 & \text{in } \Omega, \\ u(\boldsymbol{x}, 0) = u_0 & \text{in } \Omega, \\ u = g & \text{on } \Gamma_{\text{in}}. \end{cases}$$

Here, u_0 is the prescribed *initial condition*, and g is called the *inflow boundary condition.* $\Gamma_{\text{in}} \subseteq \partial\Omega$ is the *inflow* part of the boundary, defined by

$$\Gamma_{\text{in}} := \{\boldsymbol{x} \in \partial\Omega : \boldsymbol{\beta} \cdot \boldsymbol{n} < 0\},$$

where $\boldsymbol{n}$ is the outward facing normal.

In certain cases, for the purposes of analysis, we consider the *steady* (non-time-dependent) version of the problem.

Steady advection–reaction equation

We look for $u : \Omega \to \mathbb{R}$ satisfying

$$(**) \begin{cases} \nabla \cdot (\boldsymbol{\beta} u) + cu = f & \text{in } \Omega, \\ u = g & \text{on } \Gamma_{\text{in}}, \end{cases}$$

where Γ_{in} is the inflow part of the boundary as before. We assume that the *reaction coefficient* $c \in L^\infty(\Omega)$ and

$$c + \frac{1}{2}\nabla \cdot \boldsymbol{\beta} \geq \gamma_0 > 0.$$

Note that in contrast to most of the equations we have been considering so far (Poisson, elasticity, and Stokes, which are **elliptic** and heat equation which is **parabolic**), the advection equation is **hyperbolic**.

We will often switch between these two. A discretization for one of these problems can easily be modified to give a discretization for the other. To illustrate this, if we use the *method of lines* to discretize the time variable in $(*)$ using (for example) backward Euler, then we obtain

$$\frac{u_{n+1} - u_n}{\Delta t} + \nabla \cdot (\boldsymbol{\beta} u_{n+1}) = 0.$$

Rearranging,

$$\Delta t \nabla \cdot (\boldsymbol{\beta} u_{n+1}) + u_{n+1} = u_n, \tag{1}$$

which has exactly the same form as $(**)$. Generally, a discretization suitable for $(**)$ will be suitable for (1), and vice versa. **Note:** in this illustration, we used backward Euler for the time integration, but for the time-dependent equation $(*)$, **explicit** methods are usually more common.

2 Attempt: H^1-Conforming Discretization

To begin, we will derive a variational formulation and associated H^1-conforming finite element method for $(**)$. Then, we will see a numerical example that illustrates why this method has some undesirable features. (This numerical example will later be backed up by error analysis, but for now we consider only the numerics). This will motivate the development of other types of discretizations for this problem.

There are two ways of enforcing the inflow boundary condition: weakly or strongly. It is more common to use weak enforcement of the boundary conditions, but we explain both methods here.

Strongly enforced boundary conditions. For simplicity, we consider homogeneous boundary conditions, $g = 0$. Consider the function space $V \subseteq H^1(\Omega)$ defined by

$$V = \{v \in H^1(\Omega) : u = 0 \text{ on } \Gamma_{\text{in}}\}.$$

Following the standard methodology, we multiply $(**)$ by a test function $v \in V$ and integrate over Ω.

$$\begin{aligned} & \nabla \cdot (\boldsymbol{\beta} u) + cu = f \\ \implies \quad & \nabla \cdot (\boldsymbol{\beta} u) v + cuv = fv \\ \implies \quad & \int_\Omega \nabla \cdot (\boldsymbol{\beta} u) v \, dx + \int_\Omega cuv \, dx = \int_\Omega fv \, dx \end{aligned}$$

So, the variational problem is: find $u \in V$ such that, for all $v \in V$,

$$\int_\Omega \nabla \cdot (\boldsymbol{\beta} u) v \, dx + \int_\Omega cuv \, dx = \int_\Omega fv \, dx.$$

Weakly enforced boundary conditions. Instead of working in the space V (which has the boundary conditions "built-in"), we choose u and v to be drawn from the whole space $H^1(\Omega)$. Then, multiplying by $v \in H^1(\Omega)$ and integrating the fist term by parts,

$$\begin{aligned}
& \nabla \cdot (\boldsymbol{\beta} u) + cu = f \\
\implies \quad & \nabla \cdot (\boldsymbol{\beta} u) v + cuv = fv \\
\implies \quad & \int_\Omega \nabla \cdot (\boldsymbol{\beta} u) v \, dx + \int_\Omega cuv \, dx = \int_\Omega fv \, dx \\
\implies \quad & \int_{\partial\Omega} uv \boldsymbol{\beta} \cdot \boldsymbol{n} \, ds - \int_\Omega u \boldsymbol{\beta} \cdot \nabla v \, dx + \int_\Omega cuv \, dx = \int_\Omega fv \, dx
\end{aligned}$$

In the first term, we can replace u on Γ_{in} by the inflow condition g, and integrate by parts a second time to obtain

$$\int_\Omega \nabla \cdot (\boldsymbol{\beta} u) v \, dx - \int_{\Gamma_{\text{in}}} uv \boldsymbol{\beta} \cdot \boldsymbol{n} \, ds + \int_\Omega cuv \, dx = - \int_{\Gamma_{\text{in}}} gv \boldsymbol{\beta} \cdot \boldsymbol{n} + \int_\Omega fv \, dx.$$

2.1 Example: Spurious Oscillations

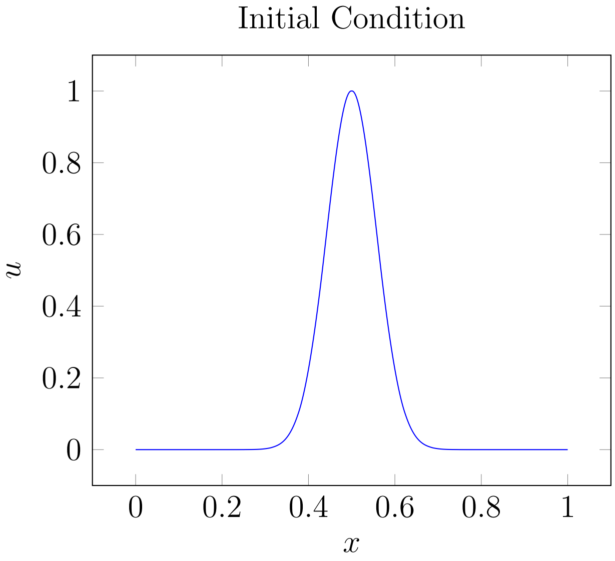
We consider a simple 1D time-dependent example with $\boldsymbol{\beta} \equiv 1$. Then, the exact solution is given by

$$u(x,t) = u_0(x - t).$$

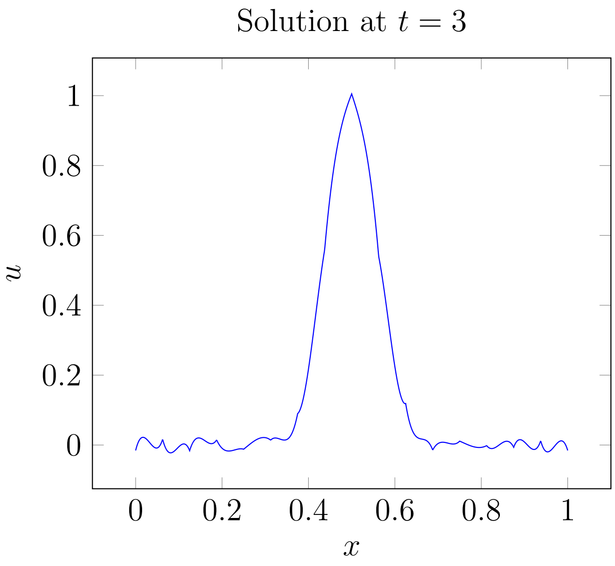
Let $\Omega = [0, 1]$ with periodic boundary conditions, and choose initial condition

$$u_0(x) = \exp(-150(x - 1/2)^2).$$

The initial condition is shown in the following plot.



Using the H^1-conforming discretization above, we integrate in time using the fourth-order Runge–Kutta methods (RK4) until a final time $T = 3$. Note that at time $t = 1$, the solution has been advected one full period, and $u = u_0$. So when $t = 3$, the solution has performed three periods. The numerical solution is shown below.



We see that while the overall shape of the solution looks roughly correct, there are also visible "spurious oscillations". These are the wiggles in the numerical solution that should not be present (since the exact solution should be the same as the initial condition). Our goal is to come up with a discretization that avoids (as much as possible) these spurious oscillations. The eventual discretization we will come up with will be the **discontinuous Galerkin** method. In order to motivate the derivation of this method, we consider first the **finite volume method**.

3 Finite Volume Method

Consider the domain Ω subdivided into "cells" $\mathcal{T} = \{\kappa_i\}$. These could be triangles like in the finite element method, or they could be more general shapes (we will assume they are polygons here). We will use the notation u_i to denote the **cell average** of u on κ_i,

$$u_i := \frac{1}{|\kappa_i|} \int_{\kappa_i} u \, dx.$$

Let h_i denote the measure of cell i, $h_i := |\kappa_i|$. The finite volume method represents the numerical solution as a vector of cell averages $\boldsymbol{u} = (u_1, \ldots, u_N)$. (Compare this to the piecewise linear finite element method, which represents the numerical solution as a vector of vertex values).

Consider the time-dependent equation

$$\frac{\partial u}{\partial t} + \nabla \cdot (\boldsymbol{\beta} u) = 0.$$

Integrating over cell κ_i, we obtain

$$\int_{\kappa_i} \frac{\partial u}{\partial t} \, dx + \int_{\kappa_i} \nabla \cdot (\boldsymbol{\beta} u) \, dx = 0.$$

Switching the order of the derivative and integral,

$$\frac{\partial}{\partial t}\int_{\kappa_i} u\,dx + \int_{\kappa_i} \nabla\cdot(\boldsymbol{\beta} u)\,dx = 0.$$

Then,

$$h_i \frac{\partial u_i}{\partial t} + \int_{\kappa_i} \nabla\cdot(\boldsymbol{\beta} u)\,dx = 0.$$

At this point, we integrate the second term by parts (apply the divergence theorem) to obtain

$$h_i \frac{\partial u_i}{\partial t} - \int_{\partial\kappa_i} u\boldsymbol{\beta}\cdot\boldsymbol{n}\,ds = 0. \tag{2}$$

It is at this point that we make the key approximation in the finite volume method. Suppose we do not know the full solution u, but only its cell-average values $u_1, \ldots, u_N$. Then, we need to approximate the integrals of the form $\int_{\partial\kappa_i} u\boldsymbol{\beta}\cdot\boldsymbol{n}\,ds$, since given cell averages, we do not know the value of u on $\partial\kappa_i$.

The key approximation in the finite volume method is to replace u on cell boundaries with its **upwind value**. For hyperbolic problems (like the advection equation), the solution evolves along **characteristic curves**. In this case, the solution is **constant** along the curves traced out by the velocity $\boldsymbol{\beta}$. This means that the information is propagating in the direction of $\boldsymbol{\beta}$. We replace u on cell boundaries with the cell-average value of u from the cell in the opposite direction of $\boldsymbol{\beta}$.

Let e denote an edge in the mesh, so that e is given by the intersection of two cells, $e = \kappa^+ \cap \kappa^-$. The cell averages are u^+ and u^-, respectively, and the outward facing normals are $\boldsymbol{n}^+$ and $\boldsymbol{n}^-$ (and so $\boldsymbol{n}^- = -\boldsymbol{n}^+$). On e, u is approximated by the constant upwind value

$$u_e = \begin{cases} u^+ & \text{if } \boldsymbol{\beta}\cdot\boldsymbol{n}^+ > 0, \\ u^- & \text{otherwise.} \end{cases}$$

Then, (2) becomes

$$h_i \frac{\partial u_i}{\partial t} - \sum_{e\in\partial\kappa_i} |e| u_e \boldsymbol{\beta}\cdot\boldsymbol{n} = 0,$$

which gives a coupled system of ODEs for the cell averages u_i. These ODEs can be integrated using whichever time integration method you prefer.

3.1 1D Example

To make the method more concrete, consider $\Omega = [a,b]$, and $\kappa_i = [x_i, x_{i+1}]$, $h_i = x_{i+1} - x_i$. Consider $\beta \equiv 1$. The equation is then $\partial u/\partial_t = -\partial u/\partial x$. Then, since information is propagating from left to right, the *upwind value* at x_i is always the cell average from the cell to the left, u_{i-1}. The discretization becomes

$$h_i \frac{\partial u_i}{\partial t} + u_{i-1} - u_i = 0.$$

Rearranging,

$$\frac{\partial u_i}{\partial t} = -\frac{u_i - u_{i-1}}{h_i},$$

which can be seen to be a simple backward-difference approximation for the term $\partial u/\partial x$. A simple Taylor series argument shows that this method is first-order accurate.

That this method uses the **backward-difference** approximation for the spatial derivative is very important. If we were to replace the approximation either the forward-difference approximation or centered difference, the method would be **unstable**.

Similarly, if we had $\beta \equiv -1$, then we should use the forward difference quotient, and not the backward difference quotient. This is because it is important to respect the directionality of the propagation of information along characteristics in hyperbolic problems by always using the **upwind value**.

4 H^1-Conforming Method: Analysis

Define the bilinear form $a(\cdot\,,\cdot) : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ by

$$a(u, v) := \int_\Omega \nabla \cdot (\boldsymbol{\beta} u) v \, dx - \int_{\Gamma_{\text{in}}} uv \boldsymbol{\beta} \cdot \boldsymbol{n} \, ds + \int_\Omega cuv \, dx$$

and the linear form $F : H^1(\Omega) \to \mathbb{R}$ by

$$F(v) := -\int_{\Gamma_{\text{in}}} gv\boldsymbol{\beta} \cdot \boldsymbol{n} + \int_\Omega fv \, dx.$$

The (**non-symmetric**) variational problem is: find $u \in H^1(\Omega)$ such that

$$a(u, v) = F(v) \tag{3}$$

for all $v \in H^1(\Omega)$. It is clear (based on the derivation of the variational problem) that the exact solution u of $(**)$ satisfies (3).

Continuity and coercivity of $a(\cdot\,,\cdot)$ (and boundedness of F) are sufficient to ensure that the variational problem is well-posed, and then it will be possible to apply the usual finite element analysis.

Lemma 1. *For $u \in H^1(\Omega)$,*

$$a(u, u) \geq \gamma_0 \|u\|^2_{L^2(\Omega)} + \frac{1}{2} \int_{\partial\Omega} |\boldsymbol{\beta} \cdot \boldsymbol{n}| u^2 \, ds.$$

Proof. By the product rule,

$$\begin{aligned}
\nabla \cdot (\boldsymbol{\beta} u) u &= (\nabla \cdot \boldsymbol{\beta}) u^2 + (\boldsymbol{\beta} \cdot \nabla u) u \\
&= (\nabla \cdot \boldsymbol{\beta}) u^2 + \frac{1}{2} \boldsymbol{\beta} \cdot \nabla u^2 \\
&= (\nabla \cdot \boldsymbol{\beta}) u^2 + \frac{1}{2} \nabla \cdot (\boldsymbol{\beta} u^2) - \frac{1}{2} (\nabla \cdot \boldsymbol{\beta}) u^2 \\
&= \frac{1}{2} (\nabla \cdot \boldsymbol{\beta}) u^2 + \frac{1}{2} \nabla \cdot (\boldsymbol{\beta} u^2)
\end{aligned}$$

Therefore,

$$\begin{aligned}
a(u,u) &= \int_\Omega \nabla\cdot(\boldsymbol{\beta} u)u\,dx - \int_{\Gamma_{\text{in}}} u^2 \boldsymbol{\beta}\cdot\boldsymbol{n}\,ds + \int_\Omega cu^2\,dx \\
&= \int_\Omega (c + \tfrac{1}{2}\nabla\cdot\boldsymbol{\beta})u^2\,dx + \frac{1}{2}\int_\Omega \nabla\cdot(\boldsymbol{\beta}u^2)\,dx - \int_{\Gamma_{\text{in}}} u^2\boldsymbol{\beta}\cdot\boldsymbol{n}\,ds \\
&\geq \int_\Omega \gamma_0 u^2\,dx + \frac{1}{2}\int_\Omega \nabla\cdot(\boldsymbol{\beta}u^2)\,dx - \int_{\Gamma_{\text{in}}} u^2\boldsymbol{\beta}\cdot\boldsymbol{n}\,ds \\
&= \int_\Omega \gamma_0 u^2\,dx + \frac{1}{2}\int_{\Gamma_{\text{out}}} u^2\boldsymbol{\beta}\cdot\boldsymbol{n}\,ds - \int_{\Gamma_{\text{in}}} u^2\boldsymbol{\beta}\cdot\boldsymbol{n}\,ds \\
&= \int_\Omega \gamma_0 u^2\,dx + \frac{1}{2}\int_{\partial\Omega} |\boldsymbol{\beta}\cdot\boldsymbol{n}|u^2\,ds. \qquad \square
\end{aligned}$$

This implies that $a(\cdot\,,\cdot)$ is coercive with respect to

$$|||u|||^2 := \|u\|^2_{L^2(\Omega)} + \int_{\partial\Omega} |\boldsymbol{\beta}\cdot\boldsymbol{n}|u^2\,ds.$$

However, this is **not** a norm on $H^1(\Omega)$, and so this result does not show coercivity on $H^1(\Omega)$.

Let V_h be the standard piecewise linear finite element space, and let u_h satisfy

$$a(u_h, v_h) = F(v_h)$$

for all test functions $v_h \in V_h$. Note that since the exact solution $u \in H^1(\Omega)$ satisfies $a(u,v) = F(v)$ for all $v \in H^1(\Omega)$, then we have the **Galerkin orthogonality** property

$$a(u - u_h, v_h) = 0$$

for all $v_h \in V_h$. We can prove the following result concerning accuracy of the H^1-conforming method.

Theorem 1. *Suppose the exact solution u to (**) is in $H^2(\Omega)$. Then,*

$$\|u-u_h\|^2_{L^2(\Omega)} + \int_{\partial\Omega} |\boldsymbol{\beta}\cdot\boldsymbol{n}||u-u_h|^2\,ds \lesssim h^2|u|^2_{H^2(\Omega)}.$$

Proof. We begin with some notation. Denote the error $e = u - u_h$. Let $I_h : H^2(\Omega) \to V_h$ denote the nodal interpolation operator, so that $I_h u$ is the unique element of V_h such that $I_h u(\boldsymbol{x}_i) = u(\boldsymbol{x}_i)$ for all vertices $\boldsymbol{x}_i$ of the mesh. Then,

$$e = u - u_h = u - I_h u + I_h u - u_h = (u - I_h u) - (u_h - I_h u).$$

It then suffices to bound $|||u - I_h u|||$ and $|||u_h - I_h u|||$. The bound on $|||u - I_h u|||$ follows from standard approximation properties of the nodal interpolant, and so we consider $|||u_h - I_h|||$. We compute

$$\begin{aligned}
|||u_h - I_h|||^2 &\leq a(u_h - I_h u, u_h - I_h u) && \text{(coercivity)} \\
&= a(u_h - u + u - I_h u, u_h - I_h u) \\
&= a(u - I_h u, u_h - I_h u) && \text{(Galerkin orthogonality).}
\end{aligned}$$

We introduce notation,

$$\eta := u - I_h u, \qquad \xi := u_h - I_h u,$$

and so the last line can be written

$$\begin{aligned} a(u - I_h u, u_h - I_h u) &= a(\eta, \xi) \\ &= \int_\Omega (\nabla \cdot (\boldsymbol{\beta}) + c)\, \eta\xi \, dx - \int_{\Gamma_{\text{in}}} \eta\xi \boldsymbol{\beta} \cdot \boldsymbol{n} \, ds \end{aligned}$$

□

We make note of two points regarding the previous result. First, note that the L^2 error is suboptimal by one order (i.e. compared with the error of the normal interpolant). Second, the result only holds if the exact solution is in $H^2(\Omega)$. In fact, if the exact solution $u \in H^1(\Omega)$ but $u \notin H^2(\Omega)$, then the method is not convergent in H^1.