


Matematicko-fyzikální
fakulta
Univerzita Karlova

Numerical analysis of nonlinear PDE

Alexei Gazca

February 25, 2026

Preliminary version

Contents

Contents ii

1 Introduction 1
1.1 Convergence of numerical schemes . . . 1
1.2 Nonlinear PDE . . . 3

2 Things you should know... 4
2.1 ...about functional analysis . . . 4
2.2 ...about Lebesgue spaces . . . 7
2.3 ...about Sobolev spaces . . . 8
2.4 ...about Bochner spaces . . . 11
2.5 ...about convex analysis . . . 13
2.6 ...about finite elements . . . 15

Compactness methods 18

3 Linear Elliptic Problems 19
3.1 The Galerkin scheme . . . 20
3.2 Problems arising from a potential . . . 22
Possible convergence issues . . . 24
Γ-convergence . . . 25
3.3 Exercises . . . 28

4 Linear Parabolic Problems 31
4.1 Fully discrete approximation . . . 32
4.2 Exercises . . . 38

Introduction 1

1.1 Convergence of numerical schemes

Suppose we are trying to solve a problem of the form

$$\text{Find } u \in X \text{ such that:} \qquad A(u) = f, \tag{1.1}$$

where $A\colon X \to X^*$ is a given nonlinear differential operator, $f \in X^*$ is a given forcing term, and u is the solution, which belongs to some Banach space X. Now consider a numerical scheme consisting of:

1. A finite dimensional space X_k that approximates X in some sense.
2. Appropriate approximations $A_k \in X_k^*$ and $f_k \in X_k^*$ of A and f, respectively.
3. An approximate solution $u_k \in X_k$ solving $A_k(u_k) = f_k$.

Ideally, the goal is to develop a numerical scheme in such a way that not only the existence of the u_k's is guaranteed, but also stability and boundedness properties. In the end this should eventually lead to *convergence*:

$$u_k \to u \qquad \text{as} \qquad k \to \infty. \tag{1.2}$$

For linear operators $A \in \mathscr{L}(X; X^*)$ this is a consequence of the celebrated *Lax equivalence principle*, [1] which guarantees convergence assuming *consistency* and *stability*. There are various ways of defining notions of stability, but one possibility is the following: suppose a representative $u_k^\star \in X_k$ of the exact solution $u \in X$ is available (for example some interpolation or projection of u onto X_k); in particular, in view of condition (1), we expect that $u_k^\star \to u$ as $k \to \infty$. The consistency error can then be defined as [2]

1: Insert picture of Lax

2: For finite difference schemes this is sometimes called the truncation error.

$$\left\| A_k u_k^\star - f_k \right\|_{X_k^*}, \tag{1.3}$$

and a consistent method is one for which the consistency error vanishes for $k \to \infty$; this means that in a sense the discrete problem $A_k u_k = f_k$ approaches the original problem $Au = f$ and hints at what "appropriate" should mean in condition (2).

We now turn to a notion of stability: suppose the data is slightly perturbed $f_k \mapsto \widetilde{f}_k := f_k + \epsilon_k$ and call the resulting discrete solution $\widetilde{u}_k$. A stable scheme should result in a perturbed solution $\widetilde{u}_k$ that is close to u_k, assuming that ϵ_k is small. The relative error can be quantified in terms of the operator norm $\left\| A_k^{-1} \right\|_{\mathscr{L}(X_k^*; X_k)}$:

$$\frac{\|\widetilde{u}_k - u_k\|_{X_k}}{\left\| f_k - \widetilde{f}_k \right\|_{X_k^*}} = \frac{\left\| A_k^{-1} \varepsilon_k \right\|_{X_k}}{\|\varepsilon_k\|_{X_k^*}} \le \left\| A_k^{-1} \right\|_{\mathscr{L}(X_k^*; X_k)}. \tag{1.4}$$

A stable scheme is then one for which $\left\| A_k^{-1} \right\|_{\mathscr{L}(X_k^*; X_k)}$ is under control (e.g. stays bounded uniformly in k). Note that if we assume that the data approximation $f \mapsto f_k$ is bounded, this implies in particular bounds for the numerical approximations in terms of the data:

$$\|u_k\|_{X_k} \le \left\| A_k^{-1} \right\|_{\mathscr{L}(X_k^*; X_k)} \left\| f_k \right\|_{X_k^*} \le c \left\| f \right\|_{X^*}. \tag{1.5}$$

A simple computation reveals the relationship between these quantities:

$$\left\|u_k^\star - u_k\right\|_{X_k} \leq \left\|A_k^{-1}\right\|_{\mathscr{L}(X_k^*;X_k)} \left\|A_k u_k^\star - f_k\right\|_{X_k^*}, \tag{1.6}$$

which leads directly to the claimed Lax equivalence principle.

Example 1.1.1 Suppose X is a Hilbert space and $A \in \mathscr{L}(X;X^*)$ is elliptic, meaning that

$$\alpha \|v\|_X \leq |\langle Av, v\rangle_{X^*;X}|, \tag{1.7}$$

for some $\alpha > 0$. Then the Lax–Milgram Theorem implies that A is an isomorphism. Moreover, the condition (1.7) implies that $\left\|A^{-1}\right\|_{\mathscr{L}(\mathfrak{X}^*;\mathfrak{X})} \leq \alpha^{-1}$.
Now consider a conforming approximation $X_k \subset X$. If we denote the inclusion map as $i_k \colon X_k \to X$, we can simply set $A_k = i_k^* A i_k$ and $f_k = i_k^* f$, where i_k^* is the (Hilbert) adjoint of i_k. In other words, the approximate solution $u_k \in X_k$ is defined through

$$\langle Au_k, v_k\rangle_{X^*;X} = \langle f, v_k\rangle_{X^*;X} \qquad \forall v_k \in X_k \tag{1.8}$$

Since the approximation is conforming, we have $\|A_k\|_{\mathscr{L}(X_k;X_k^*)} \leq \|A\|_{\mathscr{L}(X;X^*)}$ and $\left\|A_k^{-1}\right\|_{\mathscr{L}(\mathfrak{X}_k^*;\mathfrak{X}_k)} \leq \alpha^{-1}$, so the method is uniformly stable in k. Now, defining $u_k^* := \mathcal{I}_k u \in X_k$ as some interpolant of the exact solution, the estimate (1.6) becomes

$$\|u_k - \mathcal{I}_k u\|_X \leq \frac{\|A\|_{\mathscr{L}(X;X^*)}}{\alpha} \|u - \mathcal{I}_k u\|_X. \tag{1.9}$$

Convergence of the approximations follows then from the approximation properties of $\mathcal{I}_k$; namely $u_k \to u$ as $k \to \infty$ as long as $\mathcal{I}_k u \to u$. Assuming some regularity of the exact solution, this often leads to a convergence rate

$$\|u - u_k\|_X = O(N_k^{-\gamma}), \tag{1.10}$$

where $N_k = \dim(X_k)$ and $\gamma > 0$ determines the rate of convergence.

When the operator A is nonlinear, things become much more challenging. For example, consider the problem associated to the p-Laplacian

$$\text{Find } u \in W^{1,p}(\Omega) \text{ s.t. } \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v = \int_\Omega fv \qquad \forall v \in W_0^{1,p}(\Omega), \tag{1.11}$$

where $f \in L^{p'}(\Omega)$ is given. This problem is well-posed in $W^{1,p}(\Omega)$, but if one tries to measure the error of some numerical approximation u_k in the norm of this space[3]

$$\|\nabla u - \nabla u_k\|_{L^p(\Omega)},$$

then in general only sub-optimal rates of convergence will be obtained. In this case it is necessary to develop notions of error/distance that are specifically tailored to the structure of the nonlinear problem. [4] More drastically, for nonlinear problems the uniqueness of solutions might be lost, so one has to be more careful when dealing with expressions like (1.10). [5]

For complicated nonlinear PDE, simply proving convergence is challenging enough; reaching back to our discussion of the Lax equivalence

3: Which seems like a perfectly reasonable thing to try!

4: This will be the so-called 'quasi-norm' or 'natural distance'.

5: To *which* solution u do you converge to as $k \to \infty$?

principle, in a nonlinear setting bounds like (1.5) can be rather weak, and we will have one more requirement in addition to stability and consistency, namely *compactness*. This means that, possibly up to a subsequence, $u_k \to u$ in *some sense*; in most cases this will be weak convergence in some Sobolev space, and in addition we will usually have

$$A_k(u_k) \rightharpoonup \overline{A} \quad \text{as } k \to \infty, \tag{1.12}$$

where $\overline{A} \in X^*$ satisfies $\overline{A} = f$. The question is now whether $\overline{A} = A(u)$. Since the convergence $u_k \rightharpoonup u$ can be rather weak, this is in general not easy to prove and requires us to use the structure of the nonlinear operator. In this course we will for example be able to prove this when the operator A is *monotone*.

1.2 Nonlinear PDE

[Insert some nice motivation here...]

Things you should know... 2

Disclaimer: you are not meant to go through this chapter at once. It is rather a collection of concepts and results that will be useful later on. [1]

1: It would still be a good idea to skim through it, to get used to the notation.

2.1 ...about functional analysis

A vector space V equipped with a norm $\|\cdot\|_V$ is called a *Banach space* it is complete: every Cauchy sequence in V converges in V [2]. Two norms $\|\cdot\|_1, \|\cdot\|_2$ on V are said to be equivalent if [3]

$$\|v\|_1 \lesssim \|v\|_2 \lesssim \|v\|_1, \quad \text{for all } v \in V. \tag{2.1}$$

2: We will call this strong (or norm) convergence

3: The notation $a \lesssim b$ means that $a \leq c \cdot b$ with a generic constant $c > 0$.

A useful fact in numerical analysis is that if V is finite-dimensional, all norms are equivalent. In this case it is important to analyse the dependence of the constants in (2.1) on problem parameters like mesh size or polynomial degree.

A linear map between two vector spaces $A\colon V \to W$ is said to be bounded (or continuous), if

$$\|A\|_{\mathscr{L}(V;W)} \coloneqq \sup_{v \in V} \frac{\|A(v)\|_W}{\|v\|_V} < +\infty. \tag{2.2}$$

The vector space of all such maps is denoted $\mathscr{L}(V;W)$, which is a Banach space whenever W is. The (topological) dual space of a vector space V is then defined as $V^* \coloneqq \mathscr{L}(V;\mathbb{R})$, and we usually write the action of $v^* \in V^*$ on a vector $v \in V$ with a duality bracket: $\langle v^*, v\rangle_V \coloneqq v^*(v)$.

If a space V is equipped with an inner product $(\cdot,\cdot)_V$ [4] is complete with the norm induced by it ($\|\cdot\|_V \coloneqq \sqrt{(\cdot,\cdot)_V}$), we call V a *Hilbert space*. A very imporant result is the *Riesz representation theorem*, which states that if $v^* \in V^*$, with V a Hilbert space, there is a unique element $v \in V$, such that

$$\langle v^*, w\rangle_V = (v, w)_V \qquad \text{for all } w \in V, \tag{2.3}$$

and furthermore $\|v^*\|_{V^*} = \|v\|_V$. The mapping $J_V\colon V^* \to V$ that delivers the representative just described, is called the *Riesz map*.

4: Note that we use brackets for duality pairings and parantheses for inner products. The subindices will be ommitted if the space is clear from context.

We say that a sequence $\{v_k\}_{k\in\mathbb{N}}$ of elements belonging to a Banach space X converges weakly to $v \in X$ in X (written $v_k \rightharpoonup v$), if

$$\langle v^*, v_k\rangle \xrightarrow[k\to\infty]{} \langle v^*, v\rangle \quad \text{for all } v^* \in X^*. \tag{2.4}$$

In this case one has[5]

$$\|v\|_X \leq \liminf_{k\to\infty} \|v_k\|_X. \tag{2.5}$$

5: I.e. the norm is *weakly lower semicontinuous.*

We say that a sequence $\{v_k^*\}_{k\in\mathbb{N}}$ of elements belonging to X^* converges weakly* to $v^* \in X^*$ in X^* (written $v_k \overset{*}{\rightharpoonup} v$), if

$$\langle v_k^*, v\rangle \xrightarrow[k\to\infty]{} \langle v^*, v\rangle \quad \text{for all } v \in X. \tag{2.6}$$

The norm on X^* is also weakly* lower semicontinuous:

$$\|v^*\|_{X^*} \leq \liminf_{k\to\infty} \|v_k^*\|_{X^*}. \tag{2.7}$$

A subset K of a normed space V is said to be *(sequentially) compact* if it is closed and every sequence $\{v_k\}_{k\in\mathbb{N}} \subset K$ contains a subsequence that converges in K. A subset $K \subset V$ is called *relatively compact* (or precompact) if its closure is compact. We also say that a mapping $f : V \to W$ between normed spaces is *compact*, if it is continuous and maps bounded sets to precompact sets.

A space V is called separable if there is a countable subset that is dense [6] The *Banach–Alaoglu theorem* states that bounded sets in the dual V^* of a normed separable space V are (sequentially) precompact in the weak* topology. As a consequence, on separable reflexive spaces, bounded sets are sequentially precompact with respect to the weak topology.

6: This is crucial in numerical analysis, where we try to construct approximations with sequences of discrete spaces.

If $X \subset W$ are two vector spaces such that $\|v\|_W \lesssim \|v\|_X$ for all $v \in X$, we say that X is *continuously embedded into* W, written $X \hookrightarrow W$ If the embedding (inclusion map) is compact, we say that X is *compactly embedded into* V, written $X \overset{c}{\hookrightarrow} W$. This implies that if a sequence is bounded in X, then it is precompact in W.

Also very important when proving existence of solutions to diverse problems are fixed point theorems. *Banach's fixed point theorem* states that if X is a Banach space [7] and $f : X \to X$ is a contraction, meaning that for some $\alpha \in (0, 1)$:

$$\|f(v) - f(w)\|_X \leq \alpha \|v - w\|_X, \quad \text{for all } v, w \in X, \tag{2.8}$$

7: This is true also in metric spaces, but for us this is enough.

then there exists a unique fixed point $\overline{v} \in X$, i.e. $f(\overline{v}) = \overline{v}$. Moreover, the sequence $\{v_k\}_{k\in\mathbb{N}}$ defined recursively via $v_k = f(v_{k-1})$ (with $v_0 \in X$ arbitrary), converges linearly to $\overline{v}$.

Another important result is *Brouwer's fixed point theorem*, which states that in a finite dimensional space V, a continuous map $f : \overline{B_1(0)} \to \overline{B_0(1)}$ on the closed unit ball $\overline{B_0(1)}$ always has at least one fixed point. However, it is actually a corollary of Brouwer's fixed point theorem that will prove very useful when proving existence of solutions to discretised problems.

Corollary 2.1.1 (Zeros via Brouwer's fixed point theorem) *Suppose V is a finite dimensional Hilbert space and let $f : V \to V$ be a continuous mapping. Assume that there is $R > 0$ such that*

$$(f(v), v)_V \geq 0 \qquad \textit{for all } v \in V \quad \textit{with } \|v\|_V = R. \tag{2.9}$$

Then there is $\overline{v} \in V$ such that $f(\overline{v}) = 0$ and $\|\overline{v}\|_V \leq R$.

Of a similar flavour, we quote a result that is useful whenever the problem can be interpreted as a "perturbation" of another problem that is easier to solve.

Proposition 2.1.2 (Zeros via topological degree) *Suppose that V is a finite dimensional normed space. Let $f : V \times [0, 1] \to V$ be a continuous function and $R > 0$ be such that:*

1. *$f(\cdot, 0)$ is an affine function and the equation $f(v, 0) = 0$ has a solution $v \in V$ such that $\|v\|_V < R$.*

2. *For any* $(v, \gamma) \in V \times [0,1]$, *the condition* $f(v,\gamma) = 0$ *implies* $\|v\|_V \neq R$.

Then there exists $\overline{v} \in V$ *such that* $f(\overline{V}, 1) = 0$ *and* $\|v\|_V < R$.

Essential in nonlinear analysis is also the concept of a derivative. If X and Y are Banach spaces, and $U \subset X$ is an open set, we say that a function $f : U \subset X \to Y$ is *Fréchet differentiable* (or simply differentiable) at a point $a \in U$ if there is an element $f'(a) \in \mathscr{L}(X;Y)$ such that

$$\lim_{h\to 0} \frac{\left\| f(a+h) - f(h) - f'(a)h \right\|_Y}{\|h\|_X} = 0. \tag{2.10}$$

One difficulty with this definition is that we need to somehow come up with a candidate for $f'(a)$; it does not tell us how to compute it. In this direction the concept of a *Gâteaux derivative* is useful. The *Gâteaux derivative* (or directional derivative) of a function $f : U \subset X \to Y$ at a point $a \in U$ in the direction $h \in V$ is defined as:

$$D_G f(a;h) \coloneqq \lim_{t\to 0} \frac{f(a+th) - f(a)}{t}. \tag{2.11}$$

We say that f is Gâteaux differentiable at $a \in U$ if the directional derivative exists for all directions $h \in V$, and furthermore $D_G f(a;\cdot) \in \mathscr{L}(X;Y)$ [8] A Fréchet differentiable function is always Gâteaux differentiable, and the derivatives coincide, but the converse is not true in general. However, it is known that a function $f : U \subset X \to Y$ is continuously differentiable [9] if and only if $v \in U \mapsto D_G f(v;\cdot) \in \mathscr{L}(X;Y)$ is continuous, in which case also both derivatives coincide.

Just like in calculus, if $f : U \subset X \to \mathbb{R}$, with $U \subset X$ open, is differentiable at a local extremum $a \in U$ [10], then necessarily

$$f'(a) = 0 \tag{2.12}$$

Turning to measure theory, $(\Omega, \mathscr{A})$ is a *measurable space* if Ω is a set, and $\mathscr{A}$ is a σ-algebra: a collection of subsets of Ω that includes $\emptyset, \Omega$, is closed under complements and countable unions. A set function $\mu : \mathscr{A} \to [0, +\infty]$ is called a *measure* if $\mu(\emptyset) = 0$ and it is countably additive: if $\{E_j\}_{j\in\mathbb{N}} \subset \mathscr{A}$ are mutually disjoint, then $\mu(\bigcup_{j\in\mathbb{N}} E_j) = \sum_{j\in\mathbb{N}} \mu(E_j)$. We say that μ is a *finite measure* if $\mu(\Omega) < +\infty$.

In a measurable space $(\Omega, \mathscr{A})$ we say that $\lambda : \mathscr{A} \to [-\infty, \infty]$ is a *signed measure* if $\lambda(\emptyset) = 0$, it takes at most one of the two values $+\infty$ and $-\infty$, and $\lambda(\bigcup_{j\in\mathbb{N}} E_j) = \sum_{j\in\mathbb{N}} \lambda(E_j)$ [11] For a signed measure λ we define the set function $\|\lambda\| : \mathscr{A} \to [0, +\infty]$ as

$$\|\lambda\|(E) \coloneqq \sup \left\{ \sum_{j\in\mathbb{N}} |\lambda(E_j)| \mid \{E_j\}_{j\in\mathbb{N}} \text{ is a partition of } E \right\}, \tag{2.13}$$

and call $\|\lambda\|_{\mathscr{M}(\Omega)} \coloneqq \|\lambda\|(\Omega)$ the *total variation* of λ.

If Ω is a topological space, the smallest σ-algebra containing all the open sets of Ω is called the *Borel σ-algebra* of Ω, usually denoted $\mathscr{B}(\Omega)$. If Ω is a locally compact metric space that is also σ-compact (i.e. it is the countable union of compact sets) [12], we say that a Borel measure μ is a *Radon measure* if it is finite on compact sets of Ω. We then call λ a *signed Radon measure* if $\|\lambda\|$ is a Radon measure, and the space of all *finite signed Radon measures*

8: The directional derivative (2.11) may fail to be linear or continuous, so this is a necessary requirement.

9: I.e. $f' : U \to \mathscr{L}(X;Y)$ is continuous.

10: Meaning that it it achieves the maximum or minimum value around a neighbourhood of a.

11: I.e. the partial sums $\sum_{j=1}^{k} \lambda(E_j)$ converge as $k \to \infty$ to the left-hand-side in $[-\infty, \infty]$.

12: For us this will usually be (a subset of) $\mathbb{R}^d$.

is denoted $\mathcal{M}(\Omega)$; this is a Banach space with the total variation norm $\|\cdot\|_{\mathcal{M}(\Omega)}$. A very useful result, is that with this one can characterise the dual space of continuous functions. The space of continuous functions with compact support is defined as

$$C_c(\Omega) := \{v \in C(\Omega) \mid \exists \text{ compact } K \subset \Omega \text{ such that } v|_{\Omega\setminus K} \equiv 0\}. \quad (2.14)$$

The closure of $C_c(\Omega)$ with respect to the supremum norm $\|v\|_{C_0} := \sup_{x\in\Omega} |v(x)|$ is denoted as $C_0(\Omega)$.

Theorem 2.1.3 (Riesz–Alexandroff representation theorem) *Let Ω be a locally compact Hausdorff space, which is σ-compact, and $\ell : C_c(\Omega) \to \mathbb{R}$ a linear functional.*

1. *If ℓ is positive, i.e. $\ell(v) \geq 0$ for all $v \in C_0(\Omega)$ with $v \geq 0$, then there is a unique Radon measure $\mu : \mathcal{B}(\Omega) \to [0, +\infty]$, such that*

$$\ell(v) = \int_\Omega v \, \mathrm{d}\mu \qquad \text{for all } v \in C_c(\Omega), \quad (2.15)$$

   and one has $\mu(\Omega) = \|\ell\|_{C_0^*}$.
2. *If ℓ is bounded, then there is a unique finite signed Radon measure $\lambda \in \mathcal{M}(\Omega)$, such that*

$$\ell(v) = \int_\Omega v \, \mathrm{d}\lambda \qquad \text{for all } v \in C_c(\Omega), \quad (2.16)$$

   and one has $\|\lambda\|_{\mathcal{M}(\Omega)} = \|\ell\|_{C_0^*}$.

Note that since $C_c(\Omega) \hookrightarrow C_0(\Omega)$ densely, one can identify $C_c(\Omega)^* \cong C_0(\Omega)^* \cong \mathcal{M}(\Omega)$. Given this duality result, we can also equip $\mathcal{M}(\Omega)$ with a weak* topology, and by the Banach–Alaoglu theorem, we know that if $\{\mu_j\}_{j\in\mathbb{N}}$ is a bounded sequence in $\mathcal{M}(\Omega)$, there is a subsequence $\{\lambda_{j_k}\}_{k\in\mathbb{N}}$ such that $\lambda_{j_k} \overset{*}{\rightharpoonup} \lambda$ weak* as $k \to \infty$ to some $\lambda \in \mathcal{M}(\Omega)$; i.e.

$$\int_\Omega v \, \mathrm{d}\lambda_{j_k} \xrightarrow[k\to\infty]{} \int_\Omega v \, \mathrm{d}\lambda. \qquad \text{for all } v \in C_0(\Omega). \quad (2.17)$$

2.2 ...about Lebesgue spaces

For an open set $\Omega \subset \mathbb{R}^d$, the *Lebesgue spaces*[13] are defined for $p \in [1, \infty]$ as $L^p(\Omega) := \{v : \Omega \to \mathbb{R} \text{ (Lebesgue) measurable } \mid \|f\|_{L^p(\Omega)} < \infty\}$, where[14]:

$$\|v\|_{L^p(\Omega)} := \left(\int_\Omega |v(x)|^p \, \mathrm{d}x\right)^{1/p}, \text{ for } p \in [1, \infty) \quad (2.18)$$

$$\|v\|_{L^\infty(\Omega)} := \operatorname*{esssup}_{x\in\Omega} |v(x)| := \inf\{M \in \mathbb{R} \mid |v(x)| \leq M \text{ for a.e. } x \in \Omega\} \quad (2.19)$$

These are all Banach spaces, which are separable for $p \in [1, \infty)$ and reflexive for $p \in (1, \infty)$. The space of smooth functions with compact support $C_c^\infty(\Omega)$ is also dense in $L^p(\Omega)$ for all $p \in [1, \infty)$. For the space of $\mathbb{R}^n$-valued functions whose components belong to $L^p(\Omega)$ we will write $L^p(\Omega)^n$. The space $L^2(\Omega)$ is also a Hilbert space with the inner product [15]:

13: We will mostly consider Ω equipped with the classical Lebesgue measure but ocassionally with other Radon measures.

14: Here $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$, but we will also use $|A|$ to denote the (d-dimensional) Lebesgue measure of a Lebesgue measurable set $A \subset \mathbb{R}^d$.

15: We will ommit the volume measure when it is clear from context. And of course, the surface measure will be employed for $(d-1)$-dimensional sets and so on.

$$(v, w)_{L^2(\Omega)} \coloneqq \left(\int_\Omega v \cdot w \right)^{1/2} \tag{2.20}$$

To simplify the notation we will often just write $(v, w)_\Omega$ instead of $(v, w)_{L^2(\Omega)}$. The dual space $(L^p(\Omega))^*$ can be identified with $L^{p'}(\Omega)$, where $p' \in [1, \infty]$ is the Hölder conjugate of $p \in [1, \infty]$, defined through $\frac{1}{p} + \frac{1}{p'} = 1$. One also has Hölder's inequality:

$$\int_\Omega |v \cdot w| \leq \|v\|_{L^p(\Omega)} \|w\|_{L^{p'}(\Omega)}, \tag{2.21}$$

which guarantees that $v \cdot w \in L^1(\Omega)$, whenever $v \in L^p(\Omega)$ and $w \in L^{p'}(\Omega)$. This allows for *interpolation* between L^p-spaces as well: if $p, p_1, p_2 \in [1, \infty]$ and $\theta \in [0, 1]$, then

$$\frac{1}{p} = \frac{\theta}{p_1} + \frac{1-\theta}{p_2} \qquad \Longrightarrow \qquad \|v\|_{L^p(\Omega)} \leq \|v\|^{\theta}_{L^{p_1}(\Omega)} \|v\|^{1-\theta}_{L^{p_2}(\Omega)}. \tag{2.22}$$

The L^p-norms are also weakly lower semicontinuous: if $v_k \rightharpoonup v$ weakly in $L^p(\Omega)$ ($\overset{*}{\rightharpoonup}$ if $p = \infty$), then

$$\|v\|_{L^p(\Omega)} \leq \liminf_{k\to\infty} \|v\|_{L^p(\Omega)} \leq \sup_{k\in\mathbb{N}} \|v\|_{L^p(\Omega)} < \infty. \tag{2.23}$$

The spaces $L^p(\Omega)$ are also uniformly convex for $p \in (1, \infty)$, which in particular implies that if $v_n \rightharpoonup v$ weakly in $L^p(\Omega)$, and the norms converge $\|v_k\|_{L^p(\Omega)} \to \|v\|_{L^p(\Omega)}$, then $v_k \to v$ *strongly* in $L^p(\Omega)$.

Useful when carrying out compactness arguments is the *Dunford–Pettis theorem* will also come in handy, which states that, for a bounded set $\mathcal{F} \subset L^1(\Omega)$, the following statements are equivalent [16]:

16: We assume here that Ω has finite measure.

1. $\mathcal{F}$ is relatively weakly compact in $L^1(\Omega)$.
2. The set $\mathcal{F}$ is *equi-integrable*: for all $\varepsilon > 0$ there is $\delta > 0$ such that if $|E| \leq \delta$, then

$$\int_E |v(x)| \, \mathrm{d}x \leq \varepsilon \qquad \text{for all } v \in \mathcal{F}. \tag{2.24}$$

And when passing to a limit in nonlinear terms, *Vitali's convergence theorem*[17] will be very useful: it states that for $p \in [1, \infty)$, and $v_k, v \colon \Omega \to \mathbb{R}$ measurable functions ($k \in \mathbb{N}$), v_k converges strongly to v in $L^p(\Omega)$, if

17: This result is more powerful than Lebesgue's dominated convergence theorem.

1. $v_k \to v$ pointwise a.e. in Ω.
2. $\{|v_k|^p\}_{k\in\mathbb{N}}$ is equi-integrable.

Conversely, if $v_k \to v$ strongly in $L^p(\Omega)$, (2.) is satisfied and (1.) holds for a subsequence [18] . As a consequence, if $v_k \rightharpoonup v$ weakly in $L^1(\Omega)$, then it converges strongly if and only if it converges in measure. Another useful fact is that for $p \in (1, \infty)$, if $\{v_k\}_{k\in\mathbb{N}}$ is bounded in $L^p(\Omega)$, and v_k converges pointwise (or in measure) to a function v, then $v \in L^p(\Omega)$, and $v_n \rightharpoonup v$ weakly in $L^p(\Omega)$.

18: You can use convergence in measure in (i) instead, and then no subsequence is needed for the converse.

2.3 ...about Sobolev spaces

We say that a locally integrable function $u \in L^1_{\mathrm{loc}}(\Omega)$ [19] has a *weak partial*

19: I.e. $v \in L^1(\omega)$ for all open $\omega \subset \Omega$ that are relatively compact and $\overline{\omega} \subset \Omega$.

derivative in the direction $i \in \{1, \dots, d\}$ if there is $v_i \in L^1_{\text{loc}}(\Omega)$, such that

$$\int_\Omega u \partial_i \varphi = -\int_\Omega v\varphi \qquad \text{for all } \varphi \in C_c^\infty(\Omega). \tag{2.25}$$

We say that u is *weakly differentiable* if all the weak partial derivatives exist for $i \in \{1, \dots, d\}$, and we define the *weak derivative* (or weak gradient) as $\nabla u \coloneqq (\partial_i u)_{i=1}^d$. For $p \in [1, \infty]$ we define the *Sobolev space* $W^{1,p}(\Omega) \coloneqq \{u \in L^p(\Omega) \mid \nabla u \in L^p(\Omega)^d\}$. These are Banach spaces with the norm:

$$\|v\|_{W^{1,p}(\Omega)} \coloneqq \left(\|v\|^p_{L^p(\Omega)} + \|\nabla v\|^p_{L^p(\Omega)}\right)^{1/p}, \text{ for } p \in [1, \infty) \tag{2.26}$$

$$\|v\|_{W^{1,\infty}(\Omega)} \coloneqq \max\{\|v\|_{L^\infty(\Omega)}, \|\nabla v\|_{L^\infty(\Omega)}\}. \tag{2.27}$$

For $p = 2$ we write $H^1(\Omega) \coloneqq W^{1,2}(\Omega)$, which is a Hilbert space with the inner product:

$$(v, w)_{H^1(\Omega)} \coloneqq (v, w)_{L^2(\Omega)} + (\nabla v, \nabla w)_{L^2(\Omega)}. \tag{2.28}$$

Higher order derivatives and Sobolev spaces are defined analogously; e.g. $W^{k,p}(\Omega)$ is the set of functions in $L^p(\Omega)$ with all weak derivatives up to order $k \in \mathbb{N}$ belonging to $L^p(\Omega)$. Similarly to the Lebesgue spaces, $W^{k,p}(\Omega)$ is separable for $p \in [1, \infty)$ and $k \in \mathbb{N}$, and reflexive (hence also uniformly convex) for $p \in (1, \infty)$. The set $C^\infty(\Omega) \cap W^{k,p}(\Omega)$ is also dense in $W^{k,p}(\Omega)$ for $k \in \mathbb{N}$ and $p \in [1, \infty)$.

To have many of the useful results related to Sobolev spaces available, we need to assume some regularity on the boundary of Ω. For our purposes, assuming that Ω is an open bounded set with Lipschitz boundary will suffice [20]. In this case, defining the Sobolev exponent as

$$p^* \coloneqq \begin{cases} \frac{dp}{d-p}, & \text{if } p < d, \\ \text{an arbitrary large real}, & \text{if } p = d, \\ \infty, & \text{if } p > d. \end{cases} \tag{2.29}$$

20: So $\partial\Omega$ looks locally like the graph of a Lipschitz function.

then the Sobolev embedding $W^{1,p}(\Omega) \hookrightarrow L^{p^*}(\Omega)$ holds, and if $p > d$ one has in fact $W^{1,p}(\Omega) \hookrightarrow C^{0,\alpha}(\overline{\Omega})$ for $\alpha \coloneqq 1 - \frac{d}{p}$. For higher order we similarly have that $W^{k,p}(\Omega)$ embeds continuously into $L^{\frac{dp}{d-kp}}(\Omega)$ whenever $kp < d$, into arbitrary $L^q(\Omega)$ with $q \in [1, \infty)$, and into $C^{0,\alpha}(\overline{\Omega})$ for $\alpha = 1 - \frac{d}{kp}$. In addition, also the embedding $W^{d,1}(\Omega) \hookrightarrow C(\overline{\Omega})$ holds. The *Rellich–Kondrachov theorem* then states that the following embeddings for $k \in \mathbb{N}$ and $p \in [1, \infty]$ are compact:

$$W^{k,p}(\Omega) \overset{c}{\hookrightarrow} \begin{cases} L^q(\Omega) \text{ for all } q \in [1, \frac{pd}{d-kp}), & \text{if } kp \le d, \\ C(\overline{\Omega}) & \text{if } kp > d, \\ W^{m,p}(\Omega) & \text{if } m < k. \end{cases} \tag{2.30}$$

Regarding the restriction to the boundary, for $p \in [1, \infty)$ there is a liner bounded operator (called the *trace operator*) $\gamma \colon W^{1,p}(\Omega) \to L^p(\partial\Omega)$, such that $\gamma(v) = v|_{\partial\Omega}$ for all $\gamma \in C^1(\overline{\Omega})$. This operator is actually continuous as a mapping $\gamma \colon W^{1,p}(\Omega) \to L^{p^\sharp}(\partial\Omega)$, where [21]

$$p^\sharp \coloneqq \begin{cases} \frac{dp-p}{d-p}, & \text{if } p < d, \\ \text{an arbitrary large real}, & \text{if } p = d, \\ \infty, & \text{if } p > d, \end{cases} \tag{2.31}$$

21: The operator is surjective onto the space $W^{\frac{1}{p'},p}(\partial\Omega)$.

and compact as a mapping $\gamma\colon W^{1,p}(\Omega) \to L^q(\partial\Omega)$ for all $q < p^\sharp$. To simplify the notation, we will mostly write $u|_{\partial\Omega}$ (or simply u) instead of $\gamma(u)$. With this, it is possible to write down an integration by parts formula for Sobolev spaces: if $u \in W^{1,p}(\Omega)$ and $v \in W^{1,q}(\Omega)$, with $p, q \in [1,\infty)$, are such that

$$\begin{cases} \frac{1}{p} + \frac{1}{q} \leq 1 + \frac{1}{d}, & \text{if } p, q \in [1, d), \\ q > 1 & \text{if } p \geq d, \\ p > 1 & \text{if } q \geq d, \end{cases} \tag{2.32}$$

then the functions $uv\mathbf{n}_j$[22] lie in $L^1(\partial\Omega)$ for all $j \in \{1, \dots, d\}$, and

$$(u, \partial_j v)_\Omega + (\partial_j u, v)_\Omega = (u, v\mathbf{n}_j)_{\partial\Omega} \tag{2.33}$$

Note that this will be satisfied if $q = p'$ with $p \in (1,\infty)$.

22: Here $\mathbf{n}$ denotes the unit outer normal vector along $\partial\Omega$.

We define the Sobolev space with zero boundary conditions as $W^{1,p}_0(\Omega) \coloneqq \ker(\gamma) = \{v \in W^{1,p}(\Omega) \mid v|_{\partial\Omega} = 0\}$ [23]. The dual of this space will be denoted $W^{-1,p'}(\Omega) \coloneqq (W^{1,p}_0(\Omega)^*)$. We will also make use of the following spaces for $p \in (1,\infty)$:

$$W^p(\mathrm{div};\Omega) \coloneqq \left\{ \mathbf{r} \in L^p(\Omega)^d \mid \mathrm{div}\,\mathbf{r} \coloneqq \sum_{j=1}^d \partial_j \mathbf{r}_j \in L^p(\Omega) \right\}, \tag{2.34}$$

23: On Lipschitz domains this is equivalent to say that $W^{1,p}_0(\Omega)$ is the closure of $C^\infty_c(\Omega)$ with respect to $\|\cdot\|_{W^{1,p}(\Omega)}$

and write $H(\mathrm{div};\Omega) \coloneqq W^2(\mathrm{div};\Omega)$. Combining the integration by parts formula (2.33) with the surjectivity of the trace operator γ onto $W^{\frac{1}{p'},p}(\partial\Omega)$, we can define by duality a *normal trace operator* $\gamma_{\mathbf{n}}\colon W^p(\mathrm{div};\Omega) \to W^{-\frac{1}{p},p}(\partial\Omega)$:

$$\langle \gamma_{\mathbf{n}}(\mathbf{r}), v\rangle_\Omega = (\mathbf{r}, \nabla v)_\Omega + (\mathrm{div}\,\mathbf{r}, v)_\Omega \qquad \forall\, v \in W^{1,p'}(\Omega). \tag{2.35}$$

Where for ease of notation we wrote $\langle \gamma_{\mathbf{n}}(\mathbf{r}), v\rangle_{\partial\Omega} \coloneqq \langle \gamma_{\mathbf{n}}(\mathbf{r}), v\rangle_{W^{\frac{1}{p},p'}(\partial\Omega)}$; we will also usually write $\mathbf{r}\cdot\mathbf{n}$ instead of $\gamma_{\mathbf{n}}(\mathbf{r})$.

Also very useful is the *Poincaré–Steklov* inequality for $p \in [1,\infty]$:

$$\|v - v_\Omega\|_{L^p(\Omega)} \leq C_{\mathrm{ps}}\ell_D \|\nabla v\|_{L^p(\Omega)} \qquad \forall\, v \in W^{1,p}(\Omega), \tag{2.36}$$

where $\ell_D = \mathrm{diam}(\Omega)$ is the diameter of Ω, $v_\Omega \coloneqq \frac{1}{|\Omega|}\int_\Omega v$ is the mean of v on Ω, and $C_{\mathrm{ps}} > 0$ is a constant that depends only on p [24]. More generally, such inequalities will be satisfied as long as one gets rid of the constant functions (i.e. the kernel of ∇); if $p \in [1,\infty)$ and f is a bounded linear functional on $W^{1,p}(\Omega)$ whose restriction to constant functions does not vanish, then one has:

$$\|v\|_{L^p(\Omega)} \leq \widehat{C}_{\mathrm{ps}} \left[\ell_D \|\nabla v\|_{L^p(\Omega)} + |f(v)|\right] \qquad \forall\, v \in W^{1,p}(\Omega), \tag{2.37}$$

where again $\widehat{C}_{\mathrm{ps}} > 0$ depends on p. In particular, for instance by taking $f(v) = \int_{\partial\Omega} v$ and restricting to functions with zero traces, we obtain:

$$\|v\|_{L^p(\Omega)} \leq \widehat{C}_{\mathrm{ps}}\ell_D \|\nabla v\|_{L^p(\Omega)} \qquad \forall\, v \in W^{1,p}_0(\Omega). \tag{2.38}$$

This inequality implies that the right-hand-side is an equivalent norm to the usual $W^{1,p}_0(\Omega)$-norm, and we will write $\|v\|_{W^{1,p}_0(\Omega)} \coloneqq \|\nabla v\|_{L^p(\Omega)}$ for all $v \in W^{1,p}_0(\Omega)$.

24: One has $C_{\mathrm{ps}} = \pi^{-1}$ for $p = 2$

2.4 ...about Bochner spaces

When working with time-dependent problems, we will interpret space-time functions $v(t,x)$ as trajectories on a function space; i.e. for a time t, we interpret $v(t)$ as an element of a space of functions of x, usually a Sobolev space. We therefore need a theory for integrating Banach-valued functions.

Suppose $I \subset \mathbb{R}$ is a bounded interval[25] and V is a Banach space. A *simple function* $v\colon I \to V$ is a function that takes only a finite number of values $\{v_k\}_{k=1}^N$, and is such that the sets $A_k := v^{-1}(v_k) \subset I$ are Lebesgue measurable. For such functions one can define their *Bochner integral* as $\int_I v(t)\,\mathrm{d}t := \sum_{k=1}^N v_k|A_k|$ [26].

25: We will usually consider something like $I = [0,T]$ with $T > 0$.

26: Note that the Bochner integral is an element of V.

A function $v\colon I \to V$ is called *strongly measurable* (or Bochner measurable) if there is a sequence of simple functions $\{v_k\}_{k\in\mathbb{N}}$ such that $\|v_k(t) - v(t)\|_V \to 0$ for a.e. $t \in I$. A theorem from Pettis guarantees that if V is separable, a function v is strongly measurable if and only if it is *weakly measurable*, meaning that $t \in I \mapsto \langle v^*, v(t)\rangle_V \in \mathbb{R}$ is Lebesgue measurable for all $v^* \in V^*$. If v is strongly measurable, then the norm $t \in I \mapsto \|v(t)\|_V \in \mathbb{R}$ is Lebesgue measurable. Also, semidiscrete functions of the type $v = \sum_{j=1}^N \varphi_j(t)v_j$, with $\varphi \in L^1(I)$ and $\{v_j\}_{j=1}^N$ belonging to a finite-dimensional subspace of V, are strongly measurable.

We say that a strongly measurable function $v\colon I \to V$ is *Bochner integrable*, if for a sequence of simple functions $\{v_k\}_{k\in\mathbb{N}}$ with $v_k(\cdot) \to v(\cdot)$ a.e. in I, one has $\int_I \|v_k(t) - v(t)\|_V\,\mathrm{d}t \to 0$. In this case one defines $\int_I v(t)\,\mathrm{d}t$ as $\lim_{k\to\infty}\int_I v_k(t)\,\mathrm{d}t$ [27] An important fact is that v is Bochner integrable if and only if $\|v(\cdot)\|_V$ is Lebesgue integrable, and in this case $\left\|\int_I v(t)\,\mathrm{d}t\right\|_V \le \int_I \|v(t)\|_V\,\mathrm{d}t$. Also, for linear forms $v^* \in V^*$ one has $\langle v^*, \int_I v(t)\,\mathrm{d}t\rangle_V = \int_I \langle v^*, v(t)\rangle_V\,\mathrm{d}t$.

27: This exists and is independent of the approximating sequence v_k.

We define the *Bochner spaces* for $p \in [1,\infty]$ as $L^p(I;V) := \{v\colon I \to V \text{strongly measurable} \mid \|v\|_{L^p(I;V)} < \infty\}$, where

$$\|v\|_{L^p(I;V)} := \left(\int_I \|v(t)\|_V^p\,\mathrm{d}t\right)^{1/p}, \text{ for } p \in [1,\infty) \tag{2.39}$$

$$\|v\|_{L^\infty(I;V)} := \operatorname*{esssup}_{t\in I} \|v(t)\|_V \tag{2.40}$$

These are Banach spaces, and simple functions are dense in $L^p(I;V)$ for $p \in [1,\infty)$. Moreover, if V^* is separable, or if V is reflexive, for $p \in [1,\infty)$ we can identify $(L^p(I;V))^* \cong L^{p'}(I;V^*)$, with duality pairing:

$$\langle v^*, v\rangle_{L^p(I;V)} := \int_I \langle v^*(t), v(t)\rangle_V\,\mathrm{d}t. \tag{2.41}$$

In particular, if V is reflexive and $p \in (1,\infty)$, then $L^p(I;V)$ is also reflexive.

We say that a function $u \in L^1_{\mathrm{loc}}(I;V)$ has a *weak time derivative* if there is $v \in L^1_{\mathrm{loc}}(I;V)$ such that

$$\int_I u(t)\varphi'(t)\,\mathrm{d}t = -\int_I v(t)\varphi(t)\,\mathrm{d}t \qquad \forall\,\varphi \in C_c^\infty(I). \tag{2.42}$$

We usually write $\partial_t u := v$. In this case one also has that $t \in I \mapsto \langle v^*, u(t)\rangle_V \in \mathbb{R}$ is weakly differentiable in $\mathbb{R}$ for all $v^* \in V^*$, and $\partial_t\langle v^*, u(t)\rangle_V = \langle v^*, \partial_t u(t)\rangle_V$.

Suppose V_1, V_2 are Banach spaces such that $V_1 \hookrightarrow V_2$ continuously. We define the following space for $p, q \in [1, \infty]$:

$$W^{1,p,q}(I; V_1, V_2) := \{v \in L^p(I; V_1) \mid \partial_t u \in L^q(I; V_2)\}. \tag{2.43}$$

This is a Banach space with the norm $\|v\|_{W^{1,p,q}(I;V_1,V_2)} := \|v\|_{L^p(I;V_1)} + \|\partial_t v\|_{L^q(I;V_2)}$. The space $C^1(\bar{I}; V_1)$ is dense in $W^{1,p,q}$. We will also ocassionally use the space in which the distributional time derivative (2.42) is not represented by a locally integrable function, but by a finite Radon measure: $\partial_t u \in \mathcal{M}(I; V_2)$ [28]. This space will be denoted $W^{1,p,\mathcal{M}}(I; V_1, V_2)$. A very important fact is that $W^{1,p,q}(I; V_1, V_2) \hookrightarrow C(\bar{I}; V_2)$ continuously.

Suppose V is a Banach that is continuously and densely embedded in a Hilbert space H. In many parabolic problems, one makes use of a so-called *Gelfand triple*: $V \hookrightarrow H \cong H^* \hookrightarrow V^*$, by working on the space $W^{1,p,p'}(I; V, V^*)$. The space H is called a *pivot space*, and with this identification one can interpret the duality pairing in V as a continuous extension of the inner product from H:

$$(u, v)_H = \langle u, v \rangle_V \qquad \forall\, u \in H,\, v \in V. \tag{2.44}$$

Now, suppose that $I = [0, T]$ for some $T > 0$. Then one has that $W^{1,p,p'}(I; V, V^*) \hookrightarrow C(I; H)$ continuously [29], and the following integration by parts formula holds for $u, v \in W^{1,p,p'}(I; V, V^*)$ and $0 \le t_1 \le t_2 \le T$:

$$(u(t_2), v(t_2))_H - (u(t_1), v(t_1))_H = \int_{t_1}^{t_2} \langle \partial_t u(t), v(t) \rangle_V + \int_{t_1}^{t_2} \langle \partial_t v(t), u(t) \rangle_V. \tag{2.45}$$

In particular, by taking $v = u$, it follows that the function $t \in I \mapsto \frac{1}{2}\|u(t)\|_H^2$ is absolutely continuous, and therefore differentiable (in the classical sense) a.e. in I, and

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_H^2 = \langle \partial_t u(t), u(t) \rangle_V \qquad \text{for a.e. } t \in I. \tag{2.46}$$

For compactness arguments, it is essential to have a result that allows us to turn compactness in space (e.g. from the Rellich–Kondrachov theorem) into space-time compactness. In the analysis of PDE, this is often obtained via the *Aubin–Lions lemma*, which states for instance, that if V_1, V_2, V_3 are Banach spaces, with V_1 reflexive and separable, $V_1 \overset{c}{\hookrightarrow} V_2$ compactly, and $V_2 \hookrightarrow V_3$ continuously, then for $p \in (1, \infty)$ and $q \in [1, \infty]$, the embedding $W^{1,p,q}(I; V_1, V_3) \overset{c}{\hookrightarrow} L^p(I; V_2)$ is compact. One problem with this classical result in numerical analysis, is that it requires control over the time derivative in some Bochner space, which in many cases is not attainable at the discrete level[30]. In certain cases a generalisation to measures is therefore useful: if V_3 has a predual space, then the embedding $W^{1,p,\mathcal{M}}(I; V_1, V_3) \overset{c}{\hookrightarrow} L^p(I; V_2)$ is also compact. That said, the following result will prove very handy when looking for compactness, since it does not invoke a time derivative explicitly [31].

Theorem 2.4.1 (Simon's convergence Lemma) *Let V and H be Banach spaces such that V is compactly embedded into H. Suppose that $\mathcal{F} \subset L^p(I; H)$, with $p \in [1, \infty)$, satisfies:*

1. *$\mathcal{F}$ is bounded in $L^1_{\mathrm{loc}}(I; V)$;*

28: Elements ν of $\mathcal{M}(I; V)$ are countably additive maps $\nu \colon I \to V$, such that their total variation (defined analogously to (2.13)) is finite. We similarly have $(C_0(I; V))^* \cong \mathcal{M}(I; V^*)$, and since I will usually be a compact interval, $C_0(I; V) = C_c(I; V) = C(I; V)$.

29: In particular, there is a continuous trace operator $u \in W^{1,p,p'}(I; V, V^*) \mapsto u(s) \in H$ for all $s \in I$.

30: E.g. if we use piecewise-constant (in time) approximations, which cannot have a weak derivative, like the implicit Euler method.

31: It can be interpreted as a generalisation of the Riesz–Kolmogorov theorem to Bochner spaces.

2. $\int_0^{T-\varepsilon} \|v(s+\varepsilon,\cdot) - v(s,\cdot)\|_H^p \, \mathrm{d}s \to 0$. *as* $\varepsilon \to 0$, *uniformly for* $v \in \mathcal{F}$.

Then $\mathcal{F}$ *is relatively compact in* $L^p(I;H)$.

While this result is very useful, there are cases where the finite dimensional spaces $\{V_k\}_{k\in\mathbb{N}}$ where the discrete solutions lie, are not subspaces of V[32]. Suppose now that we have a sequence $\{m_k\}_{k\in\mathbb{N}}$ with $m_k \to \infty$ and define uniform time steps $\tau_k := \frac{T}{m_k}$; this induces a uniform partition of the interval $I = [0,T]$ for each k: $[0,T] = \cup_{j=1}^{m_k} I_j$, with $I_j := (t_{j-1}, t_j]$ and $t_j = j/m_k$. The following theorem is a generalisation of the Aubin–Lions lemma to the situation where the approximation is piecewise-constant on each subinterval I_j.

32: Here we say that the approximation is non-conforming in V.

Theorem 2.4.2 (Discrete Aubin–Lions lemma) *Suppose* $\{V_k\}_{k\in\mathbb{N}}$ *is a family of finite-dimensional subspaces of a Hilbert space* H. *Define for each* $k \in \mathbb{N}$ *the norm* $\|\cdot\|_{Y_k}$ *on* V_k *via duality:*

$$\|v_k\|_{Y_k} := \sup_{w_k \in V_k} \frac{(v_k, w_k)_H}{\|w_k\|_{V_k}} \qquad \forall\, v_k \in V_k. \tag{2.47}$$

Suppose further that any sequence $\{v_k\}_{k\in\mathbb{N}} \subset H$ *with* $v_k \in V_k$ *for all* $k \in \mathbb{N}$, *satisfying* $\|v_k\|_{V_k} \le c$ *with* $c > 0$ *independent of* k, *is necessarily precompact in* H. *Then, if* $\{z_k\}_{k\in\mathbb{N}}$ *is a sequence such that:*

1. z_k *is piecewise-constant in time with values in* V_k: $z_k(t)|_{I_j} = z_k^j \in V_k$.
2. *There is* $c > 0$ *and* $q \in [1,\infty)$ *such that*

$$\tau_k \sum_{j=1}^{m_k} \left\| z_k^j \right\|_{V_k}^q + \tau_k \sum_{j=2}^{m_k} \left\| \frac{z_k^j - z_k^{j-1}}{\tau_k} \right\|^q \le c \qquad \forall\, k \in \mathbb{N}. \tag{2.48}$$

Then $\{z_k\}_{k\in\mathbb{N}}$ *is precompact in* $L^q(I;H)$.

2.5 ...about convex analysis

In a vector space V, a set $K \subset V$ is called *convex* if $\lambda v + (1-\lambda)w \in K$, whenever $v, w \in V$ and $\lambda \in (0,1)$. Convex sets are closed if and only if they are weakly closed. A function $f : K \to \overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$[33] defined on a convex set $K \subset V$ is called *convex* if $f(\lambda v + (1-\lambda)w) \le \lambda v + (1-\lambda)w$, for all $v, w \in V$ and $\lambda \in (0,1)$, and we say that it is *strictly convex* if the inequality is strict for $v \neq w$, whenever $f(v)$ and $f(w)$ are finite [34]. We say a function $f : K \to \overline{R}$ is *(strictly) concave*, if $-f$ is (strictly) convex. The *(effective) domain* of a convex function $f : K \to \overline{\mathbb{R}}$ is defined as:

$$\mathrm{dom}(f) := \{v \in K \mid f(v) < +\infty\}, \tag{2.49}$$

which is itself a convex set. A convex function is called *proper* if $\mathrm{dom}(f) \neq \emptyset$ and $f(v) > -\infty$ for all v. The epigraph of a function $f : V \times \overline{R}$ is defined as $\mathrm{epi}(f) := \{(v,\alpha) \in V \times \mathbb{R} \mid \alpha \ge f(x)\}$ [35]. The function f is convex/lower semicontinuous if and only if $\mathrm{epi}(f)$ is convex/closed.

If f is a convex function, then any local minimum is a global minimum, and the set of minima $\mathrm{argmin}(f)$ is convex. If f is additionally strictly convex and proper, then $\mathrm{argmin}(f)$ contains at most one element (i.e.

33: The extended real line works mostly as you would expect; by convention we consider $\infty - \infty = \infty$.

34: One could just work with the restriction of f to the points where it is finite, but in convex analysis it is often convenient to consider functions defined on the whole set and allow them to take infinite values.

35: Note that this is a subset of $V \times \mathbb{R}$ and not $V \times \overline{\mathbb{R}}$.

minimisers are unique). In the setting where V is a Banach space and $f\colon U \subset V \to \overline{R}$, where U is open, we have a generalisation of the necessary condition for a local extremum (2.12) to the case where the extremum a belongs to a convex subset $K \subset U$. In this case, for instance if a is a local minimum relative to the set K [36], then necessarily

$$f'(a)(v-a) \geq 0 \quad \text{for all } v \in K. \tag{2.50}$$

36: Meaning that for a neighbourhood $\widetilde{U} \subset U$ of a, $f(v) \geq f(a)$ for all $v \in \widetilde{U} \cap K$.

In addition, for differentiable functions f defined on an open convex set $U \subset V$, the following conditions are necessary and sufficient for the convexity of f on U:

1. $\langle f'(v) - f'(w), v - w\rangle_V \geq 0$ for all $v, w \in U$; i.e. the derivative $f'\colon U \to V^*$ is *monotone*.
2. $f(w) \geq f(v) + \langle f'(v), w - v\rangle_V$ for all $v, w \in U$.
3. For twice-differentiable f, $f''(v)$ is positive semi-definite [37] for all $v \in U$: $f''(v)[w_1, w_2] \geq 0$ for all $w_1, w_2 \in U$.

37: Here we identified $\mathscr{L}(V; V^*)$ with the space of bounded bilinear forms on V.

For convex functions one can define a useful generalisation of the notion of a derivative: the *subdifferential* of a proper convex function $f\colon V \to \mathbb{R} \cup \{+\infty\}$ at a point $u \in \overline{\mathrm{dom}(f)}$ [38] is defined as

$$\partial f(u) \coloneqq \{v^\star \in V^* \mid \langle v^*, v - u\rangle_V + f(u) \leq f(v) \text{ for all } v \in V\}. \tag{2.51}$$

38: Or we simply set $\partial f(u) = \emptyset$ if $u \notin \overline{\mathrm{dom}(f)}$.

For set-valued maps such as this one, we will write $u \rightrightarrows \partial f(u)$; this map is closed and convex valued, meaning that the set $\partial f(u) \subset V^*$ is closed and convex for all $u \in \overline{\mathrm{dom}(f)}$. Elements v^* of $\partial f(u)$ are often called *subgradients of f at u*. If $\partial f(u) \neq \emptyset$, then necessarily f is lower semicontinuous at u [39]; this will always be the case for $u \in \mathrm{int}(\mathrm{dom}(f))$. Whenever f is Gâteaux-differentiable at u, then in fact $\partial f(u) = \{D_G(u;\cdot)\}$.

39: The subdifferential $\partial f(u)$ is also non-empty whenver f is continuous and finite at u.

The indicator function χ_K of a set $K \subset V$ is defined as $\chi_K(v) = 0$ if $v \in K$ and set to ∞ otherwise. This function is convex/proper/lower semicontinuous, whenever K is convex/non-empty/closed.

In applications one is often interested in integral functionals of the form

$$\int_\Omega f(x, u(x), \nabla u(x))\,\mathrm{d}x, \tag{2.52}$$

where u belongs to some Sobolev space $W^{1,p}(\Omega)$ [40]. A crucial question is when are these functionals lower semicontinuous; this is a topic with a vast history and numerous results with different levels of generality exist. We do not try to state here the result with optimal hypotheses.

40: Recall that we are working with a bounded open set $\Omega \subset \mathbb{R}^d$.

Many cases of interest will be covered by so-called convex normal integrands. We say that a function $f\colon \Omega \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a *normal integrand* if it is $\mathscr{L} \otimes \mathscr{B}$-measurable [41], and the function $(s, z) \mapsto f(x, s, z)$ is lower semicontinuous on $\mathbb{R}^m \times \mathbb{R}^n$ for a.e. $x \in \Omega$. Some examples of normal integrands include:

- **Carathéodory integrands:** $f(\cdot, s, z)$ is measurable for all $(s, z) \in \mathbb{R}^m \times \mathbb{R}^n$ and $f(x, \cdot, \cdot)$ is continuous for a.e. $x \in \Omega$.
- **Autonomous integrands:** $f(x, s, z) \equiv g(s, z)$ with $g\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ lower semicontinuous.
- **Integrands with constraints:** the indicator $\chi_K\colon \Omega \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ of a mapping $K\colon \Omega \rightrightarrows \mathbb{R}^m \times \mathbb{R}^n$ is a normal integrand whenever K is measurable and closed valued; this is defined as[42]:

41: $\mathscr{L}$ and $\mathscr{B}$ denote the Lebesgue and Borel σ-algebras of Ω and $\mathbb{R}^m \times \mathbb{R}^n$, respectively.

42: Here $\xi = (s, z) \in \mathbb{R}^m \times \mathbb{R}^n$.

$$\chi_K(x,\xi) := \chi_{K(x)}(\xi) = \begin{cases} 0 & \text{if } \xi \in K(x), \\ +\infty & \text{if } \xi \notin K(x), \end{cases} \tag{2.53}$$

Moreover, if f_0 is a normal integrand, the function $f = f_0 + \chi_K$, with:

$$f(x,\xi) := \begin{cases} f_0(x,\xi) & \text{if } \xi \in K(x), \\ +\infty & \text{if } \xi \notin K(x), \end{cases} \tag{2.54}$$

is also a normal integrand with $\mathrm{dom}(x,f) = \mathrm{dom}(x,f_0) \cap K(x)$.

Theorem 2.5.1 (Lower semicontinuity of convex functionals) *Let $f : \Omega \times \mathbb{R}^m \times \mathbb{R}^n \to [0,+\infty]$ be a normal integrand* [43]*, such that $z \mapsto f(x,s,z)$ is convex for all $s \in \mathbb{R}^m$ and a.e. $x \in \Omega$. Then the functional*

$$(u,v) \mapsto \int_\Omega f(x,u(x),v(x))\,\mathrm{d}x, \tag{2.55}$$

is sequentially lower semicontinuous with respect to strong convergence in $L^1(\Omega)^m$ and weak convergence in $L^1(\Omega)^n$. In particular, if $u_k \to u$ strongly in $L^1(\Omega)^m$ and $v_k \rightharpoonup v$ weakly in $L^1(\Omega)^n$, then:

$$\int_\Omega f(x,u(x),v(x)) \le \liminf_{k\to\infty} \int_\Omega f(x,u_k(x),v_k(x)). \tag{2.56}$$

43: Note the positivity assumption; this can be relaxed.

One of the most important applications of this result is the existence of minimisers of convex functions $J : V \to \mathbb{R} \cup \{+\infty\}$ defined on a Banach space V; for the above result, this will typically be some Sobolev space. This is achieved with the *direct method of the calculus of variations*. The idea is to start with a *minimising sequence* $\{v_k\}_{k\in\mathbb{N}} \subset V$; i.e. a sequence such that $J(v_k) \to \inf_{v\in V} J(v)$, as $k \to \infty$ [44] To carry on, the first important ingredient is the *(sequential) τ-inf-compactness* of J: we assume that the sub-level sets

44: This always exists, thanks to the definition of the infimum.

$$\mathrm{lev}_{\le\alpha}(J) := \{v \in V \mid J(v) \le \gamma\}, \qquad \alpha \in \mathbb{R}, \tag{2.57}$$

are (sequentially) relatively compact in V for a given topology τ. This will be the case for instance if $\mathrm{dom}\, J$ is bounded or if J is *coercive* ($J(v) \to +\infty$ for $\|v\| \to \infty$), and V is reflexive and one considers the weak topology[45]. This guarantees that there is a τ-convergent subsequence $v_{k_j} \to \overline{v} \in V$, and then one can prove that $\overline{v}$ is the desired minimiser by assuming that J is (sequentially) τ-lower semicontinuous (or apply a result like Theorem 2.5.1):

45: Or the norm topology on a finite-dimensional space, useful for discretised problems!

$$J(\overline{v}) \le \lim_{j\to\infty} J(v_{k_j}) = \inf_{v\in V} J(v) \le J(\overline{v}). \tag{2.58}$$

2.6 ...about finite elements

Many of the results we will present here are valid for many types of (Galerkin) discretisations, but one of the main motivations is the finite element method, and ocassionally we will work with specific finite element spaces. We therefore only need some relatively basic assumptions and do not go into further details about assembly, efficiency of the implementation, etc.

For a bounded polyhedral Lipschitz domain $\Omega \subset \mathbb{R}^d$, a triangulation $\mathcal{T}$ is a partition of Ω into closed d-simplices: $\overline{\Omega} = \cup_{K\in\mathcal{T}} K$. The diameter of an element $K \in \mathcal{T}$ will be denoted $h_K := \mathrm{diam}(K)$, and the (maximum)

mesh size of $\mathcal{T}$ is denoted $h := \max_{K\in\mathcal{T}} h_K$. We also define the (local) mesh-size function $h_{\mathcal{T}} : \overline{\Omega} \to \mathbb{R}$ as the piecewise-constant function such that $h_{\mathcal{T}}|_K := h_K$ for all $K \in \mathcal{T}$. We will exclusively deal with *conforming triangulations*, meaning that the intersection of two elements $K \cap \widehat{K}$ from $\mathcal{T}$ can only be empty, a vertex, or a ℓ-dimensional facet, with $\ell \in \{1, \dots, d-1\}$. The sets of $(d-1)$-dimensional facets $\mathcal{F}$, interior facets $\mathcal{F}^i$, and boundary facets $\mathcal{F}^\partial$ of $\mathcal{T}$, are defined as

$$\begin{aligned}
\mathcal{F}^i &:= \{K \cap \widehat{K} \mid K, \widehat{K} \in \mathcal{T},\ \dim_{\mathcal{H}}(K \cap \widehat{K}) = d-1\},\\
\mathcal{F}^\partial &:= \{K \cap \partial\Omega \mid K \in \mathcal{T},\ \dim_{\mathcal{H}}(K \cap \partial\Omega) = d-1\},\\
\mathcal{F} &:= \mathcal{F}^i \cup \mathcal{F}^\partial,
\end{aligned}$$

where $\dim_{\mathcal{H}}(E)$ denotes the Hausdorff dimension of a set $E \subset \mathbb{R}^d$. The diameter of a facet $F \in \mathcal{F}$ will be denoted $h_F := \operatorname{diam}(F)$.

For purposes of discretisation we will in fact consider a family of triangulations $\{\mathcal{T}_k\}_{k\in\mathbb{N}}$, where the mesh sizes h_k vanish as $k \to \infty$. Moreover, we shall always assume that the family of triangulations $\{\mathcal{T}_k\}_{k\in\mathbb{N}}$ is *shape-regular*: if $\rho_K > 0$ denotes the supremum of the diameters of inscribed balls in $K \in \mathcal{T}_k$, we assume that there is a constant $\omega_{\mathcal{T}} > 0$ (independent of k), such that $\max_{K\in\mathcal{T}_k} \frac{h_k}{\rho_K} \le \omega_{\mathcal{T}}$. This implies that locally all mesh sizes are equivalent:

$$|K|^{\frac{1}{d}} \simeq h_K \simeq h_F. \tag{2.59}$$

For the sake of consistency with other works, we will often use h as an index and write $\mathcal{T}_h$ instead of $\mathcal{T}_k$ and write $h \to 0$ as shorthand for $k \to \infty$.

We know that all functions such that $v|_K \in W^{1,p}(K)$ for all $K \in \mathcal{T}_h$, with $p \in [1, \infty]$, have a trace $v|_F$ in $L^p(F)$ for all facets $F \in \mathcal{F}_h$. The following *multiplicative trace inequality* quantifies precisely this:

$$\|v\|_{L^p(F)} \le c\, \|v\|_{L^p(K)}^{\frac{1}{p'}} \left[h_K^{-\frac{1}{p}} \|v\|_{L^p(K)}^{\frac{1}{p}} + \|\nabla v\|_{L^p(K)}^{\frac{1}{p}} \right], \tag{2.60}$$

where $c > 0$ depends on p, d, and the shape-regularity constant $\omega_{\mathcal{T}_h}$.

For use in the description of discrete spaces, we define the space of *broken polynomials of degree* $k \in \mathbb{N}_0$ [46]:

$$\mathbb{P}^k(\mathcal{T}_h) := \{v_h \in L^\infty(\Omega) \mid v_h|_K \in \mathbb{P}^k(K) \text{ for all } K \in \mathcal{T}_h\}. \tag{2.61}$$

For the subset of globally continuous polynomials of degree at most k we write $\mathbb{P}^k_{\mathrm{c}}(\mathcal{T}_h) := \mathbb{P}^k(\mathcal{T}_h) \cap C(\overline{\Omega})$ [47].

Let us now take a generic discrete space on the mesh $\mathcal{T}_h$:

$$V_h := \{v_h \in L^1_{\mathrm{loc}}(\Omega) \mid v_h|_K \in \mathcal{P}_K\}, \tag{2.62}$$

where $\mathcal{P}_K$ is a finite-dimensional space such that $\mathbb{P}^{r_0}(K) \subset \mathcal{P}_K \subset \mathbb{P}^{r_1}(K)$, for some $r_0 \le r_1 \in \mathbb{N}_0$; the largest r_0 with that property is usually referred to as the *degree* of the finite element space [48]. These spaces have usually a locally defined basis: if ϕ_h is a basis function of V_h and $\phi_h|_K \neq 0$ on some element $K \in \mathcal{T}_h$, then $\operatorname{supp}(\phi_h) \subset \omega(K)$, where $\omega(K) := \operatorname{int}(\bigcup_{\widehat{K}\cap K\neq\emptyset} \widehat{K})$ denotes the *neighbouring patch of* K [49].

Since $\mathbb{P}^{r_1}(K)$ is a finite-dimensional space, as a consequence of the norm-equivalence in finite-dimensional spaces one obtains the following *inverse*

46: Here $\mathbb{P}^k(E)$ denotes polynomials of degree at most k on a set E.

47: In other words, $\mathbb{P}^k(\mathcal{T}_h)$ and $\mathbb{P}^k_{\mathrm{c}}(\mathcal{T}_h)$ are none other than the $\mathrm{DG}(k)$ and $\mathrm{CG}(k)$ finite element spaces.

48: Usually $\mathcal{P}_K$ will arise from the affine transformation of a space defined on a reference simplex $\widehat{K}$.

49: This way the resulting matrices are sparse.

inequality for $p, q \in [1, \infty]$, $\ell \in \mathbb{N}_0$, and $j \in \{0, \dots, \ell\}$:

$$\left\|\nabla^\ell v_h\right\|_{L^p(K)} \leq c h_K^{j-\ell+d(\frac{1}{p}-\frac{1}{q})} \left\|\nabla^j v_h\right\|_{L^q(K)} \qquad \forall\, v_h \in V_h. \tag{2.63}$$

where the constant $c > 0$ depends only on r_1, d, and the shape of K (i.e. does not depend on its size h_K). Here $\nabla^\ell v$ denotes all the derivatives of v of degree $\ell \in \mathbb{N}_0$. We highlight the following useful particular cases:

$$\|\nabla v_h\|_{L^p(K)} \leq c h_K^{-1} \|v_h\|_{L^p(K)}, \tag{2.64a}$$

$$\|v_h\|_{L^p(K)} \leq c h_K^{d(\frac{1}{p}-\frac{1}{q})} \|v_h\|_{L^q(K)}. \tag{2.64b}$$

In addition, the *inverse facet-to-element inequality* also holds:

$$\|v_h\|_{L^p(F)} \leq c h_K^{-\frac{1}{p}+d(\frac{1}{p})-\frac{1}{q}} \|v_h\|_{L^q(K)}, \tag{2.65}$$

holds for all $F \in \mathcal{F}_h$ and $K \in \mathcal{T}$.

Finally, we recall that there is a linear projection operator $\mathcal{I}_{V_h} : W^{1,1}(\Omega) \to V_h$, such that $\mathcal{I}_{V_h} v_h = v_h$ for all $v_h \in \mathbb{P}^{r_0}(\mathcal{T}_h)$, and the following *local approximation property* holds [50]

$$\left\|\nabla^\ell (v - \mathcal{I}_{V_h} v)\right\|_{L^p(K)} \leq c h_K^{m-\ell} \|\nabla^m v\|_{L^p(\omega(K))} \qquad \forall\, v \in W^{1,1}(\Omega), \tag{2.66}$$

for any $p \in [1, \infty)$, $m \in \{1, \dots, r_0 + 1\}$, $\ell \in \{0, 1\}$, where $c > 0$ does not depend on h_K. In addition, the interpolation operator $\mathcal{I}_{V_h}$ can be constructed in such a way that it preserves zero boundary conditions: $\mathcal{I}_{V_h} v \in V_h \cap W_0^{1,1}(\Omega)$ if $v \in W_0^{1,1}(\Omega)$ [51]. Combining this with the density of smooth functions in Sobolev spaces, we obtain the *global approximation property*:

$$\left\|v - \mathcal{I}_{V_h} v\right\|_{W^{1,p}(\Omega)} \xrightarrow[h \to 0]{} 0, \tag{2.67}$$

for all $v \in W^{1,p}(\Omega)$ with $p \in [1, \infty)$.

50: Of course, this is only meaningful when $v \in W^{m,p}(\omega(K))$, so the right-hand-side is finite.

51: For instance, the Scott–Zhang interpolation operator satisfies this property.

Compactness methods

3 Linear Elliptic Problems

In this section we will do a quick recap of the finite element approximation of linear PDE; we will focus first on elliptic problems and proceed to parabolic problems in the next chapter. Although most of the material is probably known to you, we will take a slightly different point of view, focusing for instance more on compactness arguments. This approach will sometimes result in slightly suboptimal statements and proofs[1], but will have the advantage that it can be generalised to the nonlinear setting.

1: For example, for such problems one could prove error estimates directly.

Consider the following homogeneous Dirichlet boundary value problem posed on a bounded Lipschitz polyhedral domain $\Omega \subset \mathbb{R}^d$:

$$\begin{aligned} -\operatorname{div}(\underline{a}\nabla u) + \underline{b}\cdot\nabla u + \underline{c} = f \quad & \text{in } \Omega, \\ u = 0 \quad & \text{on } \partial\Omega, \end{aligned} \tag{3.1}$$

where $f \in H^{-1}(\Omega)$, $\underline{a} \in L^\infty(\Omega)^{d\times d}$, $\underline{b} \in L^\infty(\Omega)^d$, and $\underline{c} \in L^\infty(\Omega)$; to make sure that the problem is elliptic, we assume here that there is a positive constant $c > 0$ such that:

$$z^\top \underline{a}(\cdot) z \geq c|z|^2 \qquad \forall\, z \in \mathbb{R}^d, \text{ a.e. in } \Omega. \tag{3.2}$$

The weak formulation of the problem (3.1) consists in finding a function u belonging to the Hilbert space $H^1_0(\Omega)$ such that

$$\int_\Omega \left[\underline{a}\nabla u\cdot\nabla v + \underline{b}\cdot\nabla u v + \underline{c} u v\right] = \langle f, v\rangle_{H^1_0(\Omega)} \qquad \forall\, v \in H^1_0(\Omega). \tag{3.3}$$

By introducing a linear operator $A\colon H^1_0(\Omega) \to H^{-1}(\Omega)$ defined as:

$$\langle Av, w\rangle_{H^1_0(\Omega)} \coloneqq \int_\Omega \left[\underline{a}\nabla v\cdot\nabla w + \underline{b}\cdot\nabla v w + \underline{c} v w\right], \tag{3.4}$$

the weak formulation (3.3) can be re-written as

$$Au = f. \tag{3.5}$$

Under certain assumptions on the coefficients, it can be shown that the operator A is bounded and coercive, meaning that there exist two constants $M, \alpha > 0$ such that:

$$\langle Av, v\rangle_{H^1_0(\Omega)} \geq \alpha \|v\|^2_{H^1_0(\Omega)} \qquad \forall\, v \in H^1_0(\Omega), \tag{3.6a}$$

$$|\langle Av, w\rangle_{H^1_0(\Omega)}| \leq M \|v\|_{H^1_0(\Omega)} \|w\|_{H^1_0(\Omega)} \qquad \forall\, v, w \in H^1_0(\Omega). \tag{3.6b}$$

In this case, the Lax–Milgram theorem implies that a unique solution $u \in H^1_0(\Omega)$ exists. However, we will follow a different path, imagining that we do not know in advance that the solution exists. Instead, we will produce numerical approximations u_n and then show that they converge to a function $u \in H^1_0(\Omega)$ that satisfies the weak formulation (3.5).[2]

2: This can be interpreted as a *constructive* proof of existence of the weak solution.

3.1 The Galerkin scheme

Consider now a sequence of finite dimensional subspaces $V_n \subset H_0^1(\Omega)$, $n \in \mathbb{N}$, that approach $H_0^1(\Omega)$,[3] meaning that

$$\forall v \in H_0^1(\Omega): \quad \lim_{n\to\infty} \inf_{v_n \in V_n} \|v - v_n\|_{H_0^1(\Omega)} = 0. \tag{3.7}$$

3: Why does such a sequence exist?

Example 3.1.1 In the Finite Element setting described in Section 2.6, the condition (3.7) is a consequence of the existence of an interpolation operator $\mathcal{I}_n : H_0^1(\Omega) \to V_n$ with good approximation properties, such as (2.66). In this case one has

$$\|\mathcal{I}_n v - v\|_{H_0^1(\Omega)} \xrightarrow[n\to\infty]{} 0, \tag{3.8}$$

with a convergence rate that depends on the regularity of v.

The Galerkin scheme then consists in finding $u_n \in V_n$ such that:

$$A_n u_n = f_n, \tag{3.9}$$

where $A_n : V_n \to V_n^*$ and $f_n \in V_n^*$ are simply defined by restricting to V_n:[4]

$$\langle A_n v_n, w_n\rangle_{V_n} := \langle A v_n, w_n\rangle_{H_0^1(\Omega)} \qquad \forall v_n, w_n \in V_n \tag{3.10a}$$

$$\langle f_n, w_n\rangle_{V_n} := \langle f, w_n\rangle_{H_0^1(\Omega)} \qquad \forall w_n \in V_n. \tag{3.10b}$$

4: This is possible since we are considering a *conforming* approximation: $V_n \subset H_0^1(\Omega)$.

The coercivity property (3.6a) implies that $\ker(A_n) = \{0\}$, which guarantees the existence of a unique solution $u_n \in V_n$.[5]

5: Proving existence is often much easier at the discrete level!

Now, taking $v_n = u_n$ as a test function in the discrete formulation (3.9) and recalling the coercivity and boundedness properties (3.6), we see that

$$\alpha \|u_n\|^2_{H_0^1(\Omega)} \le \langle A u_n, u_n\rangle_{H_0^1(\Omega)} = \langle f, u_n\rangle_{H_0^1(\Omega)} \le \|f\|_{H^{-1}(\Omega)} \|u_n\|_{H_0^1(\Omega)}, \tag{3.11}$$

which in turn implies:

$$\|u_n\|_{H_0^1(\Omega)} \le \frac{\|f\|_{H^{-1}(\Omega)}}{\alpha} \quad \forall n \in \mathbb{N}. \tag{3.12}$$

Since $H^1(\Omega)$ is reflexive, this means that the sequence is relatively weakly compact:[6]

$$u_n \rightharpoonup u \quad \text{weakly in } H^1(\Omega), \tag{3.13}$$

6: To simplify the notation, we will not usually introduce additional indices for the subsequences.

where $u \in H_0^1(\Omega)$ is going to be the solution of the original problem, and as the following proposition shows, no subsequence is necessary and the convergence is in fact strong.

Theorem 3.1.1 *The discrete formulation (3.9) has a unique solution $u_n \in V_n$ for all $n \in \mathbb{N}$, and we have as $n \to \infty$ that*

$$u_n \to u \quad \textit{strongly in } H^1(\Omega), \tag{3.14}$$

where $u \in H_0^1(\Omega)$ solves the weak formulation (3.3).

Proof. Following the above discussion, what remains is to check that u is

indeed a solution of the original problem, and that the convergence is strong.

Now take an arbitrary $v \in H^1_0(\Omega)$; the approximation property (3.7) of the spaces V_n guarantees the existence of a sequence $v_n \in V_n$ such that $v_n \to v$ strongly in $H^1(\Omega)$. Combining this with the weak convergence of u_n (3.13) is enough to pass to the limit; for instance, for the first term in the weak formulation it is clear: [7]

$$\int_\Omega \underline{a}\nabla u_n \cdot \nabla v_n \to \int_\Omega \underline{a}\nabla u \cdot \nabla v. \tag{3.15}$$

7: Remember that the product of a weakly and a strongly convergent sequence converges to the product of the limits.

One can also argue more abstractly by considering the adjoint operator $A^\top : H^1_0(\Omega) \to H^{-1}(\Omega)$ defined by

$$\langle A^\top w_1, w_2\rangle_{H^1_0(\Omega)} \coloneqq \langle A w_2, w_1\rangle_{H^1_0(\Omega)} \quad \forall\, w_1, w_2 \in H^1_0(\Omega). \tag{3.16}$$

It is clear that $A^\top v_n \to A^\top v$ as $n \to \infty$, [8] and therefore

8: Why is $A^\top$ continuous?

$$\langle Au_n, v_n\rangle_{H^1_0(\Omega)} = \langle A^\top v_n, u_n\rangle_{H^1_0(\Omega)} \xrightarrow[n\to\infty]{} \langle A^\top v, u\rangle_{H^1_0(\Omega)} = \langle Au, v\rangle_{H^1_0(\Omega)}. \tag{3.17}$$

In addition, one clearly has $\ell(v_n) \to \ell(v)$, so the limiting function u is a solution of (3.3). The coercivity property (3.6a) guarantees that u is the only solution.

Regarding strong convergence, note first that by the Rellich-Kondrachov theorem we also have (up to a subsequence) that

$$u_n \to u \quad \text{strongly in } L^2(\Omega), \tag{3.18}$$

as $n \to \infty$. In particular, this means that

$$\langle f, u_n\rangle_{H^1_0(\Omega)} - \int_\Omega \left[\underline{b}\cdot\nabla u_n u_n + \underline{c}|u_n|^2\right] \xrightarrow[n\to\infty]{} \langle f, u\rangle_{H^1_0(\Omega)} - \int_\Omega \left[\underline{b}\cdot\nabla u u + \underline{c}|u|^2\right], \tag{3.19}$$

and so by testing respectively the weak formulation (3.3) with $v = u$, and the discrete formulation (3.9) with $v_n = u_n$, we see that

$$\begin{aligned}
\lim_{n\to\infty} \int_\Omega \underline{a}\nabla u_n \cdot \nabla u_n &= \lim_{n\to\infty} \left[\langle f, u_n\rangle_{H^1_0(\Omega)} - \int_\Omega \left[\underline{b}\cdot\nabla u_n u_n + \underline{c}|u_n|^2\right]\right] \\
&= \langle f, u\rangle_{H^1_0(\Omega)} - \int_\Omega \left[\underline{b}\cdot\nabla u u + \underline{c}|u|^2\right] \\
&= \int_\Omega \underline{a}\nabla u \cdot \nabla u.
\end{aligned}$$

The fact that strong convergence follows from this is left as an exercise.

To finish the proof, observe that the uniform boundedness (3.12) implies that any subsequence of $\{u_n\}_{n\in\mathbb{N}}$ will in turn have a subsequence that converges to some function in $\tilde{u} \in H^1_0(\Omega)$. Following the same argument above, this limiting function will satisfy the weak formulation, so by uniqueness we must have $\tilde{u} = u$. The Urysohn subsequence principle [9] then implies that the whole sequence $\{u_n\}_{n\in\mathbb{N}}$ converges to u, and no subsequences are needed.

9: The Urysohn subsequence principle: if every subsequence of a given sequence $\{x_n\}_{n\in\mathbb{N}}$ has itself a subsequence that converges to the same limit x, then the whole sequence converges to x.

□

Comment 3.1.1

In a basic finite element course one proves convergence via Céa's lemma, which asserts that

$$\|u_n - u\|_{H_0^1(\Omega)} \lesssim \inf_{v_n \in V_n} \|v_n - u\|_{H_0^1(\Omega)}, \tag{3.20}$$

which together with the approximation property (3.7) yields convergence. Then, for example with an interpolation operator as in Example 3.1.1 one can obtain convergence rates, assuming additional regularity of u.
The proof of Theorem 3.1.1 is instead more involved, and delivers only convergence without rates, but has the advantage that it can be more easily generalised to more complicated problems. Some general principles that you can keep in mind:

- Existence of solutions at the discrete level is often easier, since one has access to facts such as the equivalence of norms in finite-dimensional spaces.
- Uniform estimates like (3.12) and compactness properties of the space guarantee the existence of a (weak) limit.
- Some sort of (weak) continuity property is needed to pass to the limit and prove that the limit is a solution of the original problem. (Note that coercivity was enough to guarantee existence of discrete approximations, but boundedness -continuity- was needed to pass to the limit.) This is often one of the hardest steps.
- Strong monotonicity or uniform convexity properties like (3.6a) can turn weak convergence into strong convergence.
- Uniqueness of solutions means one can avoid the need for subsequences.

3.2 Problems arising from a potential

Let us now examine the case where $\underline{b}, \underline{c} = 0$ in the elliptic problem (3.1), and in addition $\underline{a} = \underline{a}^\top$. The fact that $\underline{a}$ is symmetric and (uniformly) positive definite (recall (3.2)), implies that the following defines an inner product on $H_0^1(\Omega)$, equivalent to the usual one:

$$(v, w)_a := \int_\Omega \underline{a} \nabla v \cdot \nabla w \qquad \text{for } v, w \in H_0^1(\Omega). \tag{3.21}$$

Define now the following energy functional $I\colon H_0^1(\Omega) \to \mathbb{R}$:

$$I(v) := \frac{1}{2} \int_\Omega \underline{a} \nabla v \cdot \nabla v - \langle f, v \rangle_{H_0^1(\Omega)}, \tag{3.22}$$

where $f \in H^{-1}(\Omega)$ is given. A simple computation[10] reveals that this function is Fréchet-differentiable, and one has:

$$\langle I'(v), w \rangle_{H_0^1(\Omega)} = (v, w)_a - \langle f, w \rangle_{H_0^1(\Omega)} \qquad \text{for } v, w \in H_0^1(\Omega). \tag{3.23}$$

10: Just plug this into the definition (2.10). The norm on a Hilbert space is differentiable, and its derivative is twice the inner product.

Thus, if $u \in H^1_0(\Omega)$ is a critical point of the energy functional I, one must have for all $v \in H^1_0(\Omega)$ (recall (2.12)):

$$\begin{aligned} 0 &= (u,v)_a - \langle f, v\rangle_{H^1_0(\Omega)} \\ &= \int_\Omega \underline{a}\nabla u \cdot \nabla v - \langle f, v\rangle_{H^1_0(\Omega)}, \end{aligned} \tag{3.24}$$

which is none other than the weak formulation (3.3). Moreover, since the inner product is bilinear, another simple computation shows that I is twice Fréchet-differentiable, with its second derivative is given by [11]:

$$I''(v)[w_1, w_2] = (w_1, w_2)_a \qquad \text{for } v, w_1, w_2 \in H^1_0(\Omega). \tag{3.25}$$

11: Recall that we can interpret the second derivative at a point as a bilinear form.

In particular, the second derivative is strongly positive definite, which means that if u is a critical point, then it is necessarily a minimum [12](and there is at most one). In other words, we just proved the equivalence:

$$u \text{ is a minimiser of } I \iff u \text{ solves the weak formulation (3.24).}$$

12: *Warning:* in finite-dimensional spaces it is enough to have positivity $I''(v)[w,w] > 0$ for all w to guarantee that a critical point v is a minimum, but in infinite dimensions this is not enough; this would hold e.g. if the stronger condition $I''(v)[w,w] \gtrsim \|w\|^2$ is satisfied.

And in fact, one could have *defined* a weak solution as the minimiser of I. This observation is very useful because, among other things, allows us to employ methods and techiniques from optimisation to study the problem; in particular one could focus solely on the minimisation principle when proving existence and developing discretisation schemes, and not work with the PDE at all.

Since many applications are framed from the beginning as optimisation problems, it is worth studying them in their own right. Similarly to the previous section, we present some ideas that are useful for proving convergence of numerical approximations, which can be generalised to the nonlinear setting. As before, take a sequence of conforming finite-dimensional spaces $V_n \subset H^1_0(\Omega)$, $n \in \mathbb{N}$, that satisfy the approximation property (3.7).

At the discrete level the problem consists in finding $u_n \in V_n$ such that

$$I(u_n) = \min_{v_n \in V_n} I(v). \tag{3.26}$$

As mentioned in Section 2.5, one of the main methods for proving existence of solutions to variational problems is the *direct method of the calculus of variations*, which we now apply to the discrete problem (3.26). Take then a minimising sequence $\{w_n^j\}_{j\in\mathbb{N}} \subset V_n$; i.e. a sequence such that[13]

$$I(w_n^j) \xrightarrow[j\to\infty]{} \inf_{v_n \in V_n} I(v_n). \tag{3.27}$$

13: This exists by definition of the infimum.

Note that this infimum is not $-\infty$, since the energy functional is bounded from below:

$$\begin{aligned} I(v_n) &\geq \alpha \|v_n\|^2_{H^1_0(\Omega)} - \|f\|_{H^{-1}(\Omega)} \|v_n\|_{H^1_0(\Omega)} \\ &\geq \tfrac{\alpha}{2} \|v_n\|^2_{H^1_0(\Omega)} - \tfrac{1}{2\alpha} \|f\|_{H^{-1}(\Omega)} \\ &\geq -\tfrac{1}{2\alpha} \|f\|_{H^{-1}(\Omega)} > -\infty, \end{aligned}$$

where we used the ellipticity condition (3.6a), the boundedness of f, and Young's inequality. In fact, the same computation also shows that I is *coercive* on V_n: $I(v_n) \to +\infty$ whenever $\|v_n\|_{H^1_0(\Omega)} \to +\infty$. This coercivity property means that necessarily the minimising sequence is uniformly

bounded

$$\left\| w_n^j \right\|_{H_0^1(\Omega)} \leq c \qquad \text{for all } j, n \in \mathbb{N}. \tag{3.28}$$

Since V_n is finite-dimensional, the Heine–Borel theorem implies that (up to a subsequence) $w_n^j \to u_n$ as $j \to \infty$, for some $u_n \in V_n$; alternatively, as in (3.13) one could argue that the sequence is relatively weakly compact in $H_0^1(\Omega)$, and note the fact that on finite-dimensional spaces weak and strong convergence are equivalent.

Now, noting that the norm $\|\cdot\|_a$ induced by the inner product $(\cdot,\cdot)_a$ is continuous in $H_0^1(\Omega)$, and f is continuous as well, we immediately see that

$$I(u_n) = \lim_{j\to\infty} I(w_n^j) = \inf_{v_n \in V_n} I(v_n). \tag{3.29}$$

This proves that u_n is the solution of the discrete problem (3.26), which is unique thanks to the strong convexity of I. Moreover, also by continuity of the norm, we see from the uniform bound (3.28) that

$$\|u_n\|_{H_0^1(\Omega)} \leq c \qquad \text{for all} n \in \mathbb{N}. \tag{3.30}$$

Since the bound is uniform in n, this sequence is relatively weakly precompact in $H_0^1(\Omega)$: $u_n \rightharpoonup u$ to some $u \in H_0^1(\Omega)$ (up to subsequences). It only remains to prove that u is the minimiser on the whole space $H_0^1(\Omega)$. Here this could be proved directly, but we will now introduce a couple of concepts that are more generally applicable to the convergence analysis of variational problems.

Possible convergence issues

Above we constructed a sequence of discrete solutions $u_n \in V_n$ that even have a limit $u \in H_0^1(\Omega)$. For general problems this is however not sufficient to guarantee convergence to the original problem. Consider for instance the following one-dimensional example due to Manià: the goal is to minimise the energy functional

$$I(v) = \int_0^1 (x - v(x)^3)^2 |v'(x)|^6 \, \mathrm{d}x, \tag{3.31}$$

among the class of functions

$$\mathcal{A} = \{v \in W^{1,1}(0,1) \mid v(0) = 0,\ v(1) = 1\}. \tag{3.32}$$

The exact solution to this problem is $u(x) = x^{1/3}$, but if we consider for instance discrete spaces V_n consisting of piecewise linear functions[14], one actually has that

$$0 = \min_{v\in\mathcal{A}} I(v) < \min_{v\in\mathcal{A}\cap W^{1,\infty}(0,1)} I(v) \leq \min_{v_n\in V_n} I(v_n). \tag{3.33}$$

This means in particular that convergence is impossible, since

$$\min_{v_n\in V_n} I(v_n) \not\xrightarrow[n\to\infty]{} \inf_{v\in\mathcal{A}} I(v). \tag{3.34}$$

This is an example of a phenomenon called a *Lavrentiev gap*, where the minima over different spaces are not the same, and it matters to which approximation class the discrete approximations belong [15]. One can also construct examples where this phenomenon occurs, in which the

14: Which in particular are Lipschitz and so belong to $W^{1,\infty}(0,1)$.

15: One possible remedy is to try *non-conforming approximations*.

energy is coercive, and even some arising from applications in nonlinear elasticity.

We also note that even if the convergence of the energies (3.34) does hold, and the discrete minimisers converge, it is still possible that the limit is not a solution to the original problem. To see this consider the energy

$$I(v) := \int_0^1 [|v'|^2 - 1]^2 + v^4, \tag{3.35}$$

defined on the space $W^{1,4}(0,1)$, and define V_n as the space of continuous piecewise linear functions on a uniform mesh of $(0,1)$ of width $1/n$. In this case one has that $\inf_{v\in W^{1,4}(0,1)} I(v) = 0$, and [16]

16: Try with sawtooth functions with slope ± 1.

$$\left| \min_{v_n \in V_n} I(v_n) - \inf_{v\in W^{1,4}(0,1)} I(v) \right| \lesssim n^{-4}. \tag{3.36}$$

Morever, the sequence of discrete minimisers u_n converges weakly to zero in $W^{1,4}(0,1)$, but $u = 0$ cannot be the solution, since $\inf_{v\in W^{1,4}(0,1)} I(v) = 0 < 1 = I(0)$, a contradiction. The problem in this case is that the functional I is *not* weakly lower semicontinuous [17].

17: Convergence is actually attained for the *relaxation of* I: the greatest lower semicontinuous functional that minorises I:

$$\int_0^1 [|v'|^2 - 1]_+^2 + v^4.$$

Γ-convergence

The concept of Γ-convergence, introduced by De Giorgi, provides a useful tool for studying the convergence of variational problems. Out of a sequence of variational problems associated to the minimisation of an energy functional I_n, with $n \in \mathbb{N}$, this theory provides a suitable notion of a limiting functional I as $n \to \infty$, that is meant to capture the essential properties of the sequence $\{I_n\}_{n\in\mathbb{N}}$. In the calculus of variations this is applied for instance to analyse singularly-perturbed problems (e.g. the model for an elastic plate with thickness $\varepsilon \to 0$), or homogenisation (to derive a macroscopic model arising from a microscopic model with features of length $\varepsilon \to 0$), among other things. For us, the main application will be to obtain limits of discrete variational problems.

In a Banach space X, we say that the energy functional $I\colon X \to \mathbb{R}\cup\{+\infty\}$ is the *(sequential)* Γ*-limit* (with respect to a topology τ) of the sequence of functionals $I_n\colon X \to \mathbb{R}\cup\{+\infty\}$ as $n \to \infty$, if the following two conditions are satisfied:

1. For any sequence $\{u_n\}_{n\in\mathbb{N}} \subset X$ with $u_n \to_\tau u$ in X, the *lim-inf-inequality* holds:
$$I(u) \le \liminf_{n\to\infty} I_n(u_n). \tag{3.37}$$
2. For any $u \in X$ there is a *recovery sequence* $\{u_n\}_{n\in\mathbb{N}} \subset X$, such that $u_n \to_\tau u$ in X, and[18]
$$I(u) = \lim_{n\to\infty} I_n(u_n). \tag{3.38}$$

18: Using (1.) you can see that it is enough to require that $I(u) \ge \limsup_{n\to\infty} I_n(u_n)$.

If it exists, the Γ-limit is unique, and we write $I = \Gamma\text{-}\lim_{n\to\infty} I_n$. This limit is also necessarily τ-lower semicontinuous. Furthermore, as the following proposition shows, (almost) minimisers of I_n necessarily converge to a minimiser of I.

Proposition 3.2.1 (Γ-convergence) *Let $I_n\colon X \to \mathbb{R}\cup\{+\infty\}$ be a sequence of energy functionals on a Banach space X that* Γ*-converges to a functional*

$I\colon X \to \mathbb{R}$ with respect to a topology τ. Then the following holds:

1. *Suppose $u_n \in X$ is such that $I_n(u_n) \leq \inf_{v_n \in X} I_n(v_n) + \varepsilon_n$, where $\{\varepsilon_n\}_{n\in\mathbb{N}} \subset \mathbb{R}_{\geq 0}$ is a null sequence. Assume that $\{u_n\}_{n\in\mathbb{N}}$ is τ-precompact. Then any accumulation point $\overline{u} \in X$ of this sequence is a minimiser of I, and $\inf_{v_n\in X} I_n(v_n) \xrightarrow[n\to\infty]{} I(\overline{u})$.*
2. *If $J\colon X \to \mathbb{R}$ is τ-continuous, then $\{I_n + J\}_{n\in\mathbb{N}}$ Γ-converges to $I + J$.*

Proof.
[Proof of (1.)] First note that since $\{u_n\}_{n\in\mathbb{N}}$ are almost minimisers, one has

$$\liminf_{n\to\infty} I_n(u_n) = \liminf_{n\to\infty} \inf_X I_n. \tag{3.39}$$

Now let $\overline{u} \in X$ be an accumulation point of $\{u_n\}_{n\in\mathbb{N}}$; in particular we know that there is a subsequence $\{u_{n_j}\}_{j\in\mathbb{N}}$ such that $u_{n_j} \xrightarrow[j\to\infty]{} \overline{u}$. Define then a sequence $\{\widehat{u}_n\}_{n\in\mathbb{N}}$ as

$$\widehat{u}_n := \begin{cases} u_{n_j} & \text{if } n = n_j \text{ for some } j \in \mathbb{N}, \\ \widetilde{u} & \text{otherwise.} \end{cases} \tag{3.40}$$

In particular one has $\widehat{u}_n \to \overline{u}$ as $n \to \infty$. Recalling the lim-inf-inequality (3.37), we see that

$$I(u) \leq \liminf_{n\to\infty} I_n(\widehat{u}_n) \leq \liminf_{j\to\infty} I_{n_j}(u_{n_j}) \leq \liminf_{j\to\infty} \inf_X I_{n_j}. \tag{3.41}$$

On the other hand, take an arbitrary $v \in X$ and denote a recovery sequence by $\{v_n\}_{n\in\mathbb{N}}$; in particular one has:

$$I(v) \geq \limsup_{n\to\infty} I_n(v_n) \geq \limsup_{j\to\infty} I_{n_j}(v_{n_j}). \tag{3.42}$$

Combining these inequalities we arrive at

$$I(\overline{u}) \leq \liminf_{j\to\infty} \inf_X I_{n_j} \leq \limsup_{j\to\infty} \inf_X I_{n_j} \leq \limsup_{j\to\infty} I_{n_j}(v_{n_j}) \leq I(v). \tag{3.43}$$

Since v was arbitrary, this proves that $I(\overline{u}) = \min_X I$; i.e. $\overline{u}$ is a minimiser. In addition, taking $v = \overline{u}$ above, we also see that $\inf_X I_{n_j} \to \min_X I$ as $j \to \infty$, but since all subsequences of $\inf_X I_n$ admit a subsequence converging to $\min_X I$, in the end one has $\inf_X I_n \to \min_X I$ as $n \to \infty$.

[Proof of (ii)] Since $J(u_n) \to J(u)$ whenever $u_n \to_\tau u$, the Γ-convergence of $I_n + J$ is an immediate consequence of that of I_n. □

Proposition 3.2.1 shows that if a precompact sequence of (near) minimisers is at hand, then essentially the problem is solved. In practice this is often a consequence of a uniform coercivity property of the functionals I_n, and compactness properties of the space; combining this with the previous section we can obtain a convergence result for the discrete variational problems (3.26). Since these results make use of functionals defined on a single function space X, we simply define the discrete functionals $I_n\colon X \to \mathbb{R} \cup \{+\infty\}$ as

$$I_n(v) := \begin{cases} I(v) & \text{if } v \in V_n, \\ +\infty & \text{if } v \notin V_n. \end{cases} \tag{3.44}$$

This definition forces the minimiser u_n of I_n to belong to the space V_n. Alltogether we have the following convergence result.

Theorem 3.2.2 *The sequence of functionals I_n given by (3.44), with the functional I defined in (3.22) and a sequence of finite-dimensional spaces $V_n \subset H_0^1(\Omega)$ that satisfy the approximation property (3.7), Γ-converges to I as $n \to \infty$ with respect to weak convergence in $H_0^1(\Omega)$. In particular, the sequence of minimisers $u_n \in V_n$ of I_n converges strongly in $H_0^1(\Omega)$ to the minimiser u of I in $H_0^1(\Omega)$, and*

$$I_n(u_n) \xrightarrow[n\to\infty]{} I(u). \tag{3.45}$$

Proof. To prove the claim regarding the Γ-convergence, take an arbitrary weakly convergent sequence $v_n \rightharpoonup v$ in $H_0^1(\Omega)$. To prove the lim-inf-inequality, simply notice that I is the sum of a (squared) norm on $H_0^1(\Omega)$[19] and a weakly continuous (linear) term. The inequality follows directly:

19: Recall that norms are weakly lower semicontinuous.

$$\liminf_{n\to\infty} I_n(v_n) \geq \liminf_{n\to\infty} I(v_n) \geq I(v). \tag{3.46}$$

As for the recovery sequence, the approximation property (3.7) guarantees that for arbitrary $v \in H_0^1(\Omega)$ there is a sequence $v_n \in V_n$ that converges strongly in $H_0^1(\Omega)$ to v. Since the norm in a Hilbert space is continuous, one clearly has $I_n(v_n) = I(v_n) \to I(v)$ as $n \to \infty$.

In Section 3.2 we proved already that the functional I_n has a unique minimiser $u_n \in V_n$ for all $n \in \mathbb{N}$, and that up to a subsequence $u_n \rightharpoonup u$ weakly in $H_0^1(\Omega)$ to some $u \in H_0^1(\Omega)$. Proposition 3.2.1 (1.) implies that $I_n(u_n) \to I(u) = \min_{H_0^1(\Omega)} I$. Moreover, since I is strongly convex, the minimiser is unique and hence the whole sequence converges.

Finally, one can prove that the convergence of the energies (3.45) implies that u_n actually converges *strongly* to u in $H_0^1(\Omega)$; this is left as an exercise. □

Comment 3.2.1

The Γ-convergence result in Theorem 3.2.2 was straightforward, since for every $n \in \mathbb{N}$ we had essentially the same functional I. As you can imagine, this would require a bit more work in cases where the functional I_n does change with n; for instance, if one considers discrete approximations of the coefficients, or penalisation (see Exercise 3.3). Similarly, the construction of a recovery sequence could be more involved if the approximations are non-conforming: $V_n \not\subset H_0^1(\Omega)$.

Comment 3.2.2

The fact that the convergence of the energies allows us to upgrade weak to strong convergence is a reflection of the *strong convexity* of I; compare this with the strong monotonicity of the differential operator (see Exercise 3.2). As in the convergence analysis of a basic finite element course (see (3.20)), one can obtain convergence rates assuming higher regularity of the exact solution u by working with an appropriate recovery sequence.

3.3 Exercises

Exercise 3.1 (Strongly monotone operators) Let V be a Hilbert space and let $A\colon V \to V^*$ be strongly monotone and Lipschitz continuous; i.e. there are two constants $m, M > 0$ such that for any $v_1, v_2 \in V$ one has:

$$\langle A(v_1) - A(v_2), v_1 - v_2\rangle_V \geq m \|v_1 - v_2\|_V^2 , \tag{3.47a}$$
$$\|A(v_1) - A(v_2)\|_{V^*} \leq M \|v_1 - v_2\| . \tag{3.47b}$$

Show that there is a constant $\theta > 0$ such that the sequence $\{u_k\}_{k\in\mathbb{N}} \subset V$ generated by:

$$u_k = u_{k-1} - \theta J_V(A(u_{k-1}) - f), \tag{3.48}$$

where $f \in V^*$ is given, $u_0 \in V$ chosen arbitrarily, and $J_V\colon V^* \to V$ is the Riesz map, converges to a limit $u \in V$ that satisfies $A(u) = f$ [20]

20: Hint: Banach's fixed point theorem.

Comment 3.3.1

The method (3.48) is known as a Zarantonello iteration, and it is used as a linearisation method to compute the solution to nonlinear problems, often with more sophisticated methods for choosing the parameter θ.
One word of caution: the presence of the Riesz map J_V is essential. After discretisation, where one writes e.g. $u_k^n = \sum_{j=1}^n U_{k;j}^n \varphi_j^n$ in terms of a basis $\{\varphi_j^n\}_{j=1}^n$ of a discrete space V_n, one might be tempted to implement the discretisation of the iterative scheme above as

$$U_k = U_{k-1} - \theta(A_n(U_{k-1}) - f_n), \tag{3.49}$$

where $A_n\colon \mathbb{R}^n \to \mathbb{R}^n$ and $f_n \in \mathbb{R}^n$ are restrictions of A and f to the finite dimensional space of coefficients.
The problem is that $A(u) - f$ is a dual object (its natural space is V^*), whereas u is a primal object (it lies in V), and this distinction should be kept also at the finite dimensional level; otherwise the iteration counts can deteriorate badly with n^*. In this case the remedy is found by inspecting the action of J_V on the basis:

$$\langle J_V^{-1}\varphi_j^n, \varphi_i^n\rangle_V = (\varphi_j^n, \varphi_i^n)_V =: M_n \tag{3.50}$$

This means that the matrix that represents J_V at the discrete level is precisely the inverse of the mass matrix M_n^{-1}, and so the iteration above should be written as

$$M_n U_k^n = M_n U_{k-1}^n - \theta(A_n(U_{k-1}^n) - f_n). \tag{3.51}$$

Note that J_V depends on the choice of inner product on V, so one could interpret it as a preconditioner.

*We shall refer to this as the primal-dual cardinal sin.

Exercise 3.2 (Strong convergence) In the proof of Theorem 3.1.1 we saw that the weakly converging Galerkin approximations $u_n \rightharpoonup u$ satisfy:

$$\lim_{n\to\infty} \langle A(u_n), u_n\rangle_{H_0^1(\Omega)} = \langle A(u), u\rangle_{H_0^1(\Omega)}, \tag{3.52}$$

where $A\colon H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined for all $v, w \in H_0^1(\Omega)$ as $\langle A(v), w\rangle_{H_0^1(\Omega)} := \int_\Omega \underline{a}\nabla v \cdot \nabla w$.

(a) Prove the claim we made: the condition (3.52) implies that actually

$u_n \to u$ *strongly* in $H^1_0(\Omega)$.

(b) Prove that the strong convergence still holds for strongly monotone and Lipschitz A, with the weaker condition

$$\limsup_{n\to\infty} \langle A(u_n), u_n\rangle_{H^1_0(\Omega)} \le \langle \overline{A}, u\rangle_{H^1_0(\Omega)}, \tag{3.53}$$

where $\overline{A}$ is the weak* limit of $A(u_n)$ in $H^{-1}(\Omega)$ [21]; i.e. one has not yet proved that $\overline{A} = A(u)$. Conclude that $\overline{A} = A(u)$.

21: Why does this limit exist?

Exercise 3.3 (Nitsche's method) Take a family of shape-regular triangulations $\{\mathcal{T}_h\}_{h>0}$ and associated finite element spaces $V_h \subset H^1(\Omega)$, as described in Section 2.6. Consider the Laplace equation:

$$(\nabla u, \nabla v)_\Omega = (f, v)_\Omega \qquad \forall\, \in H^1_0(\Omega), \tag{3.54}$$

with $f \in L^2(\Omega)$. In the *Nitsche method* for the weak imposition of boundary values, we look for a discrete function $u_h \in V_h$ that solves [22]:

$$\begin{aligned}(\nabla u_h, \nabla v_h)_\Omega - (\nabla u_h \cdot \mathbf{n}, v_h)_{\partial\Omega} - \theta(\nabla v_h \cdot \mathbf{n}, u_h)_{\partial\Omega} \\ + \alpha(h_{\mathcal{F}}^{-1} u_h, v_h)_{\partial\Omega} = (f, v_h)_\Omega,\end{aligned} \tag{3.55}$$

22: Note that u_h does *not* belong to the target space $H^1_0(\Omega)$.

for all $v_h \in V_h$; here $\theta \in \{-1, 0, 1\}$, and $\alpha > 0$ is a parameter to be chosen. Also, $h_{\mathcal{F}} := \partial\Omega \to (0,\infty)$ is the piecewise constant function with $h_{\mathcal{F}}|_F = h_F$ for a boundary facet $F \in \mathcal{F}_h^\partial$.

(a) Prove that for large enough $\alpha > 0$, the discrete solution $u_h \in V_h$ exists, and it is uniformly bounded in the following h-dependent norm [23]:

$$\|v_h\|_h := \left(\|\nabla v_h\|^2_\Omega + \left\| h_{\mathcal{F}}^{-\frac{1}{2}} v_h \right\|^2_{\partial\Omega} \right)^{1/2}. \tag{3.56}$$

23: Hint: use the inverse trace inequality (2.65)

(b) Prove that u_h converges weakly in $H^1(\Omega)$ to a limit $u \in H^1_0(\Omega)$, and this limit is the weak solution of the Laplace equation.

(c) For the skew-symmetric variant of the Nitsche method ($\theta = -1$), prove that the convergence of u_h is in fact strong.

(d) Prove that the symmetric variant ($\theta = 1$) can also be interpreted as the optimality condition $J_h'(u_h) = 0$ in V_h^*, for the minimisation of an appropriately defined energy $J_h : V_h \to \mathbb{R}$.

(e) **[Bonus]** Prove that the discrete variational problems from (d) Γ-converge to the usual Dirichlet principle for the Laplace problem.

Comment 3.3.2

With Nitsche's method it is straightforward to modify the formulation to approximate the problem with a non-zero Dirichlet boundary condition $u|_{\partial\Omega} = u_b$, with $u_b \in H^1(\Omega)$ given. One simply needs to make the following changes in the bilinear form:

$$\begin{aligned}\theta(\nabla v_h \cdot \mathbf{n}, u_h)_{\partial\Omega} &\mapsto \theta(\nabla v_h \cdot \mathbf{n}, u_h - u_b)_{\partial\Omega}, \\ \alpha(h_{\mathcal{F}}^{-1} u_h, v_h)_{\partial\Omega} &\mapsto \alpha(h_{\mathcal{F}}^{-1} u_h - u_b, v_h)_{\partial\Omega}.\end{aligned}$$

Similarly, if the Dirichlet condition is only desired on a portion of the boundary $\Gamma_D \subset \partial\Omega$, then the surface integrals should be on Γ_D instead of $\partial\Omega$.
By considering more general piecewise polynomials $u_h \in \mathbb{P}^k(\mathcal{T}_h)$ (note that in this case $u_h \notin H^1(\Omega)$), and using the idea above to

weakly enforce the continuity of u_h across facets, one obtains the *Discontinuous Galerkin method*.

Exercise 3.4 (Strong convergence for strongly convex problems) Take the sequence of functionals I_n on $H_0^1(\Omega)$ from Theorem 3.2.2 and their Γ-limit I.

(a) Prove the claim we made about strong convergence: if $u_n \rightharpoonup u$ weakly in $H_0^1(\Omega)$ and $I_n(u_n) \to I(u)$, then $u_n \to u$ strongly in $H_0^1(\Omega)$.

(b) Suppose that V_n are finite element spaces of degree at least 1 with maximal mesh size h_n, and that $u \in H^2(\Omega) \cap H_0^1(\Omega)$. Prove that we then obtain the convergence rate

$$\|u_n - u\|_{H_0^1(\Omega)} \leq c h_n^2 \tag{3.57}$$

(c) **[Bonus]** Could you prove similar results for the Nitsche formulation from Exercise 3.3 (e)?

4 Linear Parabolic Problems

We now turn to the analysis of a linear parabolic problem, where, similarly to the previous section, arguments that carry over to nonlinear problems will be presented.

Let $T > 0$ be a given (finite) final time and write $I \coloneqq (0, T)$. Consider now the following unsteady problem posed on the space-time domain $Q \coloneqq I \times \Omega$:

$$\begin{aligned} \partial_t u - \operatorname{div}(\underline{a} \nabla u) + \underline{b} \cdot \nabla u + \underline{c} &= f && \text{in } I \times \Omega, \\ u &= 0 && \text{in } I \times \partial\Omega, \\ u(0, \cdot) &= u_0(\cdot) && \text{in } \Omega. \end{aligned} \tag{4.1}$$

where $f \in C(\bar{I}; H^{-1}(\Omega))$, $\underline{a} \in L^\infty(Q)^{d \times d}$, $\underline{b} \in L^\infty(Q)^d$, $\underline{c} \in L^\infty(Q)$, and we are given an initial datum $u_0 \in L^2(\Omega)$. To avoid technicalities, we also assume that the coefficients $\underline{a}, \underline{b}, \underline{c}$ are continuous with respect to the time variable. [1]

1: As usual, we interpret a space-time function $v(t, x)$ as a family (parametrised by t) of x-dependent functions.

We will make use of the Gelfand triple $H_0^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$, and define the family of linear operators $A(t) \colon H_0^1(\Omega) \to H^{-1}(\Omega)$, with $t \in I$, through:

$$\langle A(t) v, w \rangle_{H_0^1(\Omega)} \coloneqq \int_\Omega \left[\underline{a}(t) \nabla v \cdot \nabla w + \underline{b}(t) \cdot \nabla v w + \underline{c}(t) v w \right]. \tag{4.2}$$

The weak formulation then consists in finding a function [2] $u \in L^2(I; H_0^1(\Omega))$ with $\partial_t u \in L^2(I; H^{-1}(\Omega))$, such that for almost every $t \in I$ one has:

2: In other words, u belongs to $W^{1,2,2}(I; H_0^1(\Omega), H^{-1}(\Omega))$

$$\partial_t u(t) - A(t) u = f(t) \qquad \text{in } H^{-1}(\Omega), \tag{4.3a}$$

$$u(0) = u_0 \qquad \text{in } L^2(\Omega). \tag{4.3b}$$

Note that it makes sense to interpret the initial condition in this sense, since $L^2(I; H_0^1(\Omega)) \cap H^1(I; H^{-1}(\Omega)) \hookrightarrow C(\bar{I}; L^2(\Omega))$. In addition, similarly to the elliptic case, the coefficients are taken such that the following coercivity and boundedness properties hold for a.e. $t \in I$:

$$\langle A(t) v, v \rangle_{H_0^1(\Omega)} \geq \alpha \|v\|_{H_0^1(\Omega)}^2 \qquad \forall v \in H_0^1(\Omega), \tag{4.4a}$$

$$|\langle A(t) v, w \rangle_{H_0^1(\Omega)}| \leq M \|v\|_{H_0^1(\Omega)} \|w\|_{H_0^1(\Omega)} \qquad \forall v, w \in H_0^1(\Omega). \tag{4.4b}$$

Note that under these assumptions, for $v \in L^2(I; H_0^1(\Omega))$, the mapping $A(v) \colon I \to H^{-1}(\Omega)$ is strongly measurable, and $\|A(v)\|_{L^2(I;H^{-1}(\Omega))} \leq M \|v\|_{L^2(I;H_0^1(\Omega))}$. Hence, the weak formulation (4.3a) can also be written using time integrals:

$$\int_I \langle \partial_t u(t), \phi(t) \rangle_{H_0^1(\Omega)} \, \mathrm{d}t + \int_I \langle A(t) u(t), \varphi(t) \rangle_{H_0^1(\Omega)} \, \mathrm{d}t = \int_I \langle f(t), \varphi(t) \rangle_{H_0^1(\Omega)} \, \mathrm{d}t, \tag{4.5}$$

for all $\varphi \in L^2(I; H_0^1(\Omega))$, plus the initial condition (4.3b); or more succintly:

$$\partial_t u - A u = f \qquad \text{in } L^2(I; H^{-1}(\Omega)), \tag{4.6a}$$

$$u(0) = u_0 \qquad \text{in } L^2(\Omega). \tag{4.6b}$$

From this formulation it is straightforward to see that there is at most one solution: suppose there is another solution $\widetilde{u}$; substracting the corresponding PDEs and and testing with $\varphi(t) = \chi_{[0,s]}(t)(u(t) - \widetilde{u}(t))$, with $s \in (0, T]$ arbitrary, we get:

$$\tfrac{1}{2}\|u(s) - \widetilde{u}(s)\|_\Omega^2 + \int_0^s \langle A(t)(u-\widetilde{u})(t), (u-\widetilde{u})(t)\rangle_{H_0^1(\Omega)}\,\mathrm{d}t = \tfrac{1}{2}\|u(0) - \widetilde{u}(0)\|_\Omega^2\,, \tag{4.7}$$

where we used the integration by parts formula (2.45). Since $u(0) = u_0 = \widetilde{u}(0)$, from the ellipticity of A it follows that $u = \widetilde{u}$.

4.1 Fully discrete approximation

Now, just as in the steady case, we will prove existence of a weak solution by proving that a numerical scheme is convergent. Sometimes it can be useful to analyse a semi-discrete problem, and there are essentially two options:

- The method of lines: a space discretisation is introduced, but the time derivative stays continuous. One can then prove existence for instance using results from ODEs.
- Rothe's method: a time discretisation is introduced, but the space variable remains continuous. One can then reduce things to the elliptic case.

We will however analyse directly a fully discrete approximation, which is what people usually implement, and prove convergence to a weak solution. For the space discretisation we consider the same setting as for the steady problem; in particular we have a sequence of finite-dimensional spaces $V_n \subset H_0^1(\Omega)$ that satisfy the approximation property (3.7). Regarding the time approximation, let us consider for simplicity uniform partitions of $\overline{I}$: for a sequence $m_k \to \infty$ as $k \to \infty$ (representing the number of subintervals), define a *time step* $\tau = \tau_k := \frac{T}{m_k}$, the *time nodes* $t_j := j\tau_k$, and the subintervals $I_j := (t_{j-1}, t_j]$ [3] for $j \in \{0, \dots, m_k\}$. The time nodes t_j and subintervals I_j should also show the index k, but we ommit it to keep the notation simple.

3: Note also that I_j contains its right endpoint t_j. We also set $I_0 := \{t_0\} = \{0\}$.

As commonly done when carrying out numerical analysis for minimal regularity weak solutions, we focus mainly on the implicit Euler scheme, where the time derivative is approximated with a backwards difference quotient [4]:

$$\partial_t v(t_j) \rightsquigarrow \frac{v(t_j) - v(t_{j-1})}{\tau} \tag{4.8}$$

4: Implicit schemes are more appropriate for parabolic problems

As a trial space, we will then employ the space of piecewise constant functions with respect to the time partition $\mathcal{J}_k := \{I_j\}_{j=1}^{m_k}$, with values in V_n:

$$\mathbb{P}^0(\mathcal{J}_k; V_n) := \{v_{k,n} : [0,T] \to V_n \mid v_{k,n}|_{I_j} \in \mathbb{P}^0(I_j; V_n) \text{ for all } j = 1, \dots, m_k\}. \tag{4.9}$$

The discrete formulation then consists in finding a function $u_{k,n} \in \mathbb{P}^0(\mathcal{J}_k; V_n)$ by carrying out the following [5]:

- For $j = 0$ set $u_{k,n}(0) := \Pi_h u_0$, where $\Pi_h : H_0^1(\Omega) \to V_n$ denotes the L^2-orthogonal projection.

5: Since $u_{k,n}$ is piecewise constant in time, it is uniquely determined by its values at $\{t_j\}_{j=0}^{m_k}$

- For $j \in \{1, \dots, m_k\}$, given the value $u_{k,n}(t_{j-1}) \in V_n$, compute the next value $u_{k,n}(t_j) \in V_n$ by solving:

$$(u_{k,n}(t_j) - u_{k,n}(t_{j-1}), v_n)_\Omega + \tau \langle A(t_j) u_{k,n}(t_j), v_n \rangle_{H_0^1(\Omega)} = \tau \langle f(t_j), v_n \rangle_{H_0^1(\Omega)}, \tag{4.10}$$

for all $v_n \in V_n$.

Comment 4.1.1

One can also write down the discrete formulation above as a space-time variational problem, similarly to (4.5). First, denote the zeroth order Lagrange interpolant of $A(t)$ associated to the time nodes $\{t_j\}_{j=1}^{m_k}$ as $A_k(t) := \sum_{j=1}^{m_k} A(t_j)\chi_{I_j}(t)$ (similarly with f). Then the discrete formulation can be written as:

$$\sum_{j=1}^{m_k} (u_{k,n}(t_j) - u_{k,n}(t_{j-1}), v_{k,n}(t_j))_\Omega + \int_I \langle A_k(t) u_{k,n}(t), v_{k,n}(t) \rangle_{H_0^1(\Omega)} \, \mathrm{d}t = \int_I \langle f_k(t), v_{k,n}(t) \rangle \, \mathrm{d}t, \tag{4.11}$$

for all $v_{k,n} \in \mathbb{P}_0(\mathcal{I}_k; V_n)$. If one applies a lowest-order quadrature to the time integrals using the nodes $\{t_j\}_{j=1}^{m_k}$, it is not necessary to introduce the approximations A_k and f_k above.

One advantage of this formulation is that it is straightforward to generalise to higher-order. One simply needs to employ a higher-order (right-sided) Gauß–Radau quadrature*, modify appropriately the time-derivative, and use $\mathbb{P}_\ell(\mathcal{I}_k; V_n)$, with $\ell \in \mathbb{N}$, as a trial and test space. This is known as the DG(ℓ) time discretisation of order ℓ ("Discontinuous Galerkin in time"), and is equivalent to the Radau IIA Implicit Runge–Kutta method, which enjoys nice stability properties that make it suitable for parabolic problems.

*Sometimes the method is defined without the Gauß–Radau quadrature, but this way one might lose the equivalence to a Runge–Kutta method in general.

Just as in the steady case, the ellipticity condition (4.4a) guarantees the exisence of a discrete solution $u_{k,n}$. Moreover, by taking $v_n = u_{k,n}(t_j)$ as a test function in (4.10) we also find that

$$(u_{k,n}(t_j) - u_{k,n}(t_{j-1}), u_{k,n}(t_j))_\Omega + \tau \langle A(t_j) u_{k,n}(t_j), u_{k,n}(t_j) \rangle_{H_0^1(\Omega)} \leq \tau \left\| f(t_j) \right\|_{H^{-1}(\Omega)} \left\| u_{k,n}(t_j) \right\|_{H_0^1(\Omega)}$$

Using the elementary identity $2a(a-b) = a^2 - b^2 + (a-b)^2$ for $a, b \in \mathbb{R}$, we can re-write the first term as

$$(u_{k,n}(t_j) - u_{k,n}(t_{j-1}), u_{k,n}(t_j))_\Omega = \tfrac{1}{2} \left\| u_{k,n}(t_j) \right\|_\Omega^2 - \tfrac{1}{2} \left\| u_{k,n}(t_{j-1}) \right\|_\Omega^2 + \tfrac{1}{2} \left\| u_{k,n}(t_j) - u_{k,n}(t_{j-1}) \right\|_\Omega^2.$$

As for the right-hand-side, we can apply Young's inequality and the ellipticity condition (4.4a):

$$\begin{aligned} \tau \left\| f(t_j) \right\|_{H^{-1}(\Omega)} \left\| u_{k,n}(t_j) \right\|_{H_0^1(\Omega)} &\leq \frac{\tau}{\sqrt{\alpha}} \left\| f(t_j) \right\|_{H^{-1}(\Omega)} \langle A(t_j) u_{k,n}(t_j), u_{k,n}(t_j) \rangle_{H_0^1(\Omega)}^{1/2} \\ &\leq \frac{\tau}{2\alpha} \left\| f(t_j) \right\|_{H^{-1}(\Omega)}^2 + \frac{\tau}{2} \langle A(t_j) u_{k,n}(t_j), u_{k,n}(t_j) \rangle_{H_0^1(\Omega)} \end{aligned}$$

Hence, if we take an arbitrary $\ell \in \{1, \dots, m_k\}$ and sum over all $j \in$

$\{1, \dots, \ell\}$, we obtain the bound:

$$\frac{1}{2}\|u_{k,n}(t_\ell)\|_\Omega^2 - \frac{1}{2}\|u_{k,n}(t_0)\|_\Omega^2 + \frac{1}{2}\sum_{j=1}^{\ell}\left\|u_{k,n}(t_j) - u_{k,n}(t_{j-1})\right\|_\Omega^2 + \frac{\tau}{2}\sum_{j=1}^{\ell}\langle A(t_j)u_{k,n}(t_j), u_{k,n}(t_j)\rangle_{H_0^1(\Omega)} \le \frac{\tau\ell}{2\alpha}\left\|f\right\|_{C(\bar{I};H^{-1}(\Omega))}^2. \tag{4.12}$$

Noting that $\|u_{k,n}(t_0)\|_\Omega \le \|u_0\|_\Omega$ [6], by taking the maximum over $\ell \in \{1, \dots, m_k\}$, we obtain the stability bound:

$$\|u_{k,n}\|_{L^\infty(I;L^2(\Omega))}^2 + \sum_{j=1}^{m_k}\left\|u_{k,n}(t_j) - u_{k,n}(t_{j-1})\right\|_\Omega^2 + \tau\sum_{j=1}^{\ell}\langle A(t_j)u_{k,n}(t_j), u_{k,n}(t_j)\rangle_{H_0^1(\Omega)} \le \frac{T}{\alpha}\left\|f\right\|_{C(\bar{I};H^{-1}(\Omega))}^2 + \|u_0\|_\Omega^2. \tag{4.13}$$

Now, observe that the function $u_{k,n} \in \mathbb{P}_0(\mathcal{I}_k; V_n)$, which means that it posseses a time derivative that can be written in terms of Dirac distributions; in particular we can write [7]:

$$\sum_{j=1}^{m_k}\left\|u_{k,n}(t_j) - u_{k,n}(t_{j-1})\right\|_\Omega = \left\|\sum_{j=1}^{m_k}(u_{k,n}(t_j) - u_{k,n}(t_{j-1}))\delta_{t_{j-1}}\right\|_{\mathcal{M}(I;L^2(\Omega))} = \|\partial_t u_{k,n}\|_{\mathcal{M}(I;L^2(\Omega))}. \tag{4.14}$$

Finally, recalling the useful consequence of Jensen's inequality $(\sum_{j=1}^{m_k} a_k)^2 \le k\sum_{j=1}^{m_k} a_k^2$ and the ellipticity assumption (4.4a), we arrive at:

$$\|u_{k,n}\|_{L^\infty(I;L^2(\Omega))}^2 + \alpha\|u_{k,n}\|_{L^2(I;H_0^1(\Omega))}^2 + \tfrac{\tau}{T}\|\partial_t u_{k,n}\|_{\mathcal{M}(I;L^2(\Omega))}^2 \le \frac{T}{\alpha}\left\|f\right\|_{C(\bar{I};H^{-1}(\Omega))}^2 + \|u_0\|_\Omega^2. \tag{4.15}$$

This is a uniform estimate with respect to $(k, n) \in \mathbb{N}^2$, and in particular it means that, as $(k, n) \to \infty$:

$$u_{k,n} \overset{*}{\rightharpoonup} u \qquad \text{weakly* in } L^\infty(I; L^2(\Omega)), \tag{4.16a}$$

$$u_{k,n} \rightharpoonup u \qquad \text{weakly in } L^2(I; H_0^1(\Omega)). \tag{4.16b}$$

Similarly to Theorem 3.1.1, we prove now that u is in fact the (unique) weak solution to the problem, and that the convergence is actually strong.

Theorem 4.1.1 *The unsteady discrete formulation (4.10) has a unique solution $u_{k,n} \in \mathbb{P}_0(\mathcal{I}_k; V_n)$ for all $(k, n) \in \mathbb{N}^2$, and we have as $(k, n) \to \infty$*

$$u_{k,n} \to u \qquad \text{strongly in } L^2(I; H^1(\Omega)), \tag{4.17}$$

where $u \in L^2(I; H_0^1(\Omega)) \cap H^1(I; H^{-1}(\Omega))$ solves the weak formulation (4.3).

Proof. The existence of a weak limit $u \in L^2(I; H_0^1(\Omega)) \cap L^\infty(I; L^2(\Omega))$ has already been established, so it remains to prove that this is the weak solution of the system.

Similarly to the steady case, strong convergence in $L^2(Q)$ would come very handy. We will obtain this with the help of Simon's lemma (Theorem

6: Almost by definition of the L^2-projection.

7: This term often gets neglected in the analysis, but it can be useful.

2.4.1). For this, note that $H^1_0(\Omega) \overset{c}{\hookrightarrow} L^2(\Omega)$, and that $\{u_{k,n}\}_{k,n}$ is bounded uniformly in $L^2(Q)$, so to apply the lemma we only need to estimate the time differences. Noting that the integrand is piecewise constant and only nonzero on $(t_j - \varepsilon, t_j]$ for $j \in \{1, \dots, m_k - 1\}$ [8], we have

8: Take $0 < \varepsilon < \tau$ without loss of generality.

$$\int_0^{T-\varepsilon} \|u_{k,n}(s+\varepsilon) - u_{k,n}(s)\|^2_\Omega \, \mathrm{d}s = \varepsilon \sum_{j=2}^{m_k} \left\|u_{k,n}(t_j) - u_{k,n}(t_{j-1})\right\|^2_\Omega , \tag{4.18}$$

and therefore tends to zero uniformly in k, n as $\varepsilon \to 0$, thanks to the a priori estimate (4.13). Hence, as $(k, n) \to \infty$:

$$u_{k,n} \to u \qquad \text{strongly in } L^2(Q). \tag{4.19}$$

For the limit passage, we will make use of the space-time formulation (4.11) To handle the time derivative, we will apply the summation by parts formula:

$$\sum_{j=1}^{m_k} (u_{k,n}(t_j) - u_{k,n}(t_{j-1}), v_{k,n}(t_j))_\Omega = (u_{k,n}(T), v_{k,n}(T))_\Omega - (u_{k,n}(0), v_{k,n}(0))_\Omega$$
$$- \sum_{j=1}^{k} (u_{k,n}(t_{j-1}), v_{k,n}(t_j) - v_{k,n}(t_{j-1}))_\Omega$$

Let us now denote the zeroth and first order Lagrange interpolants on I associated to the time nodes $\{t_j\}_{j=1}^{m_k}$ as $\mathcal{I}^0_k$ and $\mathcal{I}^1_k$, respectively. We will test the discrete formulation with $v_{k,n} = \mathcal{I}^0_k(\varphi) v_n \in \mathbb{P}_0(\mathcal{I}_k; V_n)$, where $\varphi \in C^\infty(\bar{I})$ is arbitrary, and $\{v_n\}_{n\in\mathbb{N}} \subset V_n$ converges in $H^1(\Omega)$ to a given arbitrary $v \in H^1_0(\Omega)$ [9] Noting that $\partial_t(\mathcal{I}^1_k(\varphi)) \in \mathbb{P}_0(\mathcal{I}_k; V_n)$ corresponds to backward different quotients, we get from above:

9: As before, this exists thanks to Assumption 3.7.

$$\sum_{j=1}^{m_k} (u_{k,n}(t_j) - u_{k,n}(t_{j-1}), v_{k,n}(t_j))_\Omega = -\int_I (u_{k,n}(t), v_n)_\Omega \partial_t(\mathcal{I}^1_k(\varphi)(t)) \, \mathrm{d}t$$
$$+ \sum_{j=1}^{m_k} \tau (u_{k,n}(t_{j-1}) - u_{k,n}(t_j), v_n)_\Omega \partial_t(\mathcal{I}^1_k(\varphi)(t_j)) + (u_{k,n}(T), v_n)_\Omega \varphi(T)$$
$$- (u_{k,n}(0), v_n)_\Omega \varphi(0)$$

where we used that $\mathcal{I}^0_k(\varphi)(0) = \varphi(0)$ and $\mathcal{I}^0_k(\varphi)(T) = \varphi(T)$. Observing that $\partial_t(\mathcal{I}^1_k(\varphi)) \to \partial_t \varphi$ strongly in $L^2(I)$ as $k \to \infty$, [10] as well as $u_{k,n}(0) = \Pi_n u_0 \to u_0$ strongly in $L^2(\Omega)$ as $n \to \infty$, and $u_{k,n}(T) \rightharpoonup u_T$ weakly in $L^2(\Omega)$, for some $u_T \in L^2(\Omega)$ [11], the convergence (4.19) and stability estimate (4.13) yield:

10: Why?

11: Thanks to the uniform bound (4.13), up to a subsequence as usual.

$$\lim_{(k,n)\to\infty} \sum_{j=1}^{m_k} (u_{k,n}(t_j) - u_{k,n}(t_{j-1}), v_{k,n}(t_j))_\Omega = -\int_Q u v \partial_t \varphi + \int_\Omega u_T v \varphi(T)$$
$$- \int_\Omega u_0 v \varphi(0). \tag{4.20}$$

Passing to the limit in the remaining terms is left as an exercise. All in all, we obtained the following weak formulation for arbitrary $v \in H^1_0(\Omega)$

and $\varphi \in C^\infty(\bar{I})$:

$$-(u, v\partial_t\varphi)_{L^2(Q)} + \int_I \langle A(t)u(t), v\rangle_{H_0^1(\Omega)}\varphi(t)\,\mathrm{d}t = \int_I \langle f(t), v\rangle_{H_0^1(\Omega)}\varphi(t)\,\mathrm{d}t - (u_T, v)_\Omega\varphi(T) + (u_0, v)_\Omega\varphi(0). \tag{4.21}$$

Combining this equation with the boundedness property (4.4b), and the fact that functions of the type $\sum_{\ell=1}^L \phi_k(t)v_k(x)$ (with $\phi_\ell \in C_c^\infty(I)$, $v_\ell \in H_0^1(\Omega)$) are dense[12] in $L^2(I; H_0^1(\Omega))$, we see that actually $\partial_t u \in L^2(I; H^{-1}(\Omega))$, and u satisfies the space-time weak formulation (4.6a).

12: Since they contain in particular simple functions!

To verify the initial condition (4.6b), note that combining (4.21) with the integration by parts formula (2.45), we get for any $v \in C_c^\infty(\Omega)$ and $\varphi \in C^\infty(\bar{I})$:

$$\begin{aligned}\int_I \langle \partial_t u, v\rangle_{H_0^1(\Omega)}\varphi\,\mathrm{d}t &= \int_I \langle \partial_t(u\varphi), v\rangle_{H_0^1(\Omega)}\,\mathrm{d}t - (u, v\partial_t\varphi)_{L^2(Q)} \\ &= (u(T)\varphi(T) - u(0)\varphi(0), v)_\Omega - (u_T\varphi(T) - u_0\varphi(0), v)_\Omega \\ &\quad + \int_I \langle f(t) - A(t)u(t), v\rangle_{H_0^1(\Omega)}\varphi(t)\,\mathrm{d}t.\end{aligned}$$

Since the time-integrated weak formulation (4.6a) tells us that $\partial_t u = f - Au$ in $L^2(I; H^{-1}(\Omega))$, this implies that

$$(u(T)\varphi(T) - u(0)\varphi(0), v)_\Omega = (u_T\varphi(T) - u_0\varphi(0), v)_\Omega \tag{4.22}$$

Choosing in particular φ such that $\varphi(0) = 1$ and $\varphi(T) = 0$ yields that $u(0) = u_0$, thus proving (4.6b); in an analogous manner we also get that $u_T = u(T)$. All in all, we proved that u solves (4.6), and since the weak solution is unique, no subsequences are needed.

In regards to upgrading to strong convergence, from the discrete balance of energy leading to (4.12) (cf. (4.11)), we see that [13]:

13: As in (4.11), $\underline{a}_k, \underline{b}_k, \underline{c}_k$ are piecewise constant interpolants of a, b, c associated to $\{t_j\}_{j=1}^{m_k}$

$$\limsup_{(k,n)\to\infty} \int_Q \underline{a}_k \nabla u_{k,n}\cdot\nabla u_{k,n} \le \limsup_{(k,n)\to\infty} \Big[\tfrac{1}{2}\|u_{k,n}(0)\|_\Omega^2 - \tfrac{1}{2}\|u_{k,n}(T)\|_\Omega^2 \int_I \langle f_k(t), u_{k,n}(t)\rangle_{H_0^1(\Omega)}\,\mathrm{d}t - \int_Q [\underline{b}_k\cdot\nabla u_{k,n}u_{k,n} + \underline{c}_k|u_{k,n}|^2]\Big]$$

Since the interpolants $f_k, \underline{b}_k, \underline{c}_k$ converge uniformly to their respective limits [14], using the fact that $u_{k,n}$ converges to u, weakly in $L^2(I; H_0^1(\Omega))$ and strongly in $L^2(Q)$, as well as the strong (resp. weak) convergence of $u_{k,n}(0)$ to u_0 (resp. $u_{k,n}(T)$ to $u(T)$) in $L^2(\Omega)$, and recalling the weak-lower semicontinuity of the L^2-norm, we obtain

14: Why?

$$\limsup_{(k,n)\to\infty} \int_Q \underline{a}_k \nabla u_{k,n}\cdot\nabla u_{k,n} \le \tfrac{1}{2}\|u_0\|_\Omega^2 - \tfrac{1}{2}\|u(T)\|_\Omega^2 + \int_I \langle f(t), u(t)\rangle_{H_0^1(\Omega)}\,\mathrm{d}t - \int_Q [\underline{b}\cdot\nabla u u + \underline{c}|u|^2] = \int_Q \underline{a}\nabla u\cdot\nabla u$$

As in the steady case, strong convergence is an easy consequence of this and is left as an exercise.

□

In PDE analysis you might be used to employing the Aubin–Lions Lemma to extract strongly convergent subsequences in $L^2(Q)$; this re-

quires however uniform bounds on the time derivative $\partial_t u_{k,n}$ in spaces like $L^2(I;H^{-1}(\Omega))$ or $\mathcal{M}(I;H^{-1}(\Omega))$, which does not immediately fit the structure of the time discretisations. However, such an argument is possible by assuming additionally that the $L^2(\Omega)$-orthogonal projection is $H^1(\Omega)$-stable[15]:

15: Uniformly in n!

$$\|\Pi_n v\|_{H^1_0(\Omega)} \leq c\,\|v\|_{H^1_0(\Omega)} \quad \forall v \in H^1_0(\Omega). \tag{4.23}$$

This way we could take arbitrary $v \in H^1_0(\Omega)$, $\varphi \in C^\infty_c(I)$, and from (4.10) we see that [16]

16: Recall that $\partial_t u_{k,n}$ can be interpreted as a sum of Dirac deltas (in time).

$$\begin{aligned}
\langle \partial_t u_{k,n}, \varphi v\rangle_{C_0(I;H^1_0(\Omega))} &= \sum_{j=1}^{k} (u_{k,n}(t_j) - u_{k,n}(t_{j-1}), \varphi(t_{j-1})v)_\Omega \\
&= \sum_{j=1}^{k} (u_{k,n}(t_j) - u_{k,n}(t_{j-1}), \Pi_n v)_\Omega \varphi(t_{j-1}) \\
&= \sum_{j=1}^{m_k} \tau \langle f(t_j) - A(t_j)u_{k,n}(t_j), \Pi_n v\rangle_{H^1_0(\Omega)} \varphi(t_{j-1}) \\
&\leq c\,\|\varphi\|_{C_0(I)}\,\|\Pi_n v\|_{H^1_0(\Omega)} \\
&\leq c\,\|\varphi\|_{C_0(I)}\,\|v\|_{H^1_0(\Omega)},
\end{aligned}$$

which implies that that $\|\partial_t u_{k,n}\| \leq c$, from which one can apply a generalised form of Aubin–Lions. The more classical version can be applied by using the piecewise linear linterpolant $\widetilde{u}_{k,n} \in \mathbb{P}_1(\mathcal{J}_k; V_n)$ associated to the nodes and values $\{(t_j, u_{k,n}(t_j)\}_{j=1}^{m_k}$ and noting that $\partial_t \widetilde{u}_{k,n}|_{I_j} = \frac{u_{k,n}(t_j) - u_{k,n}(t_{j-1})}{\tau}$. A similar argument as above yields then uniform bounds for $\partial_t \widetilde{u}_{k,n}$ in $L^2(I;H^{-1}(\Omega))$, and one can easily prove that $\widetilde{u}_{k,n}$ and $u_{k,n}$ both converge to the same limit[17]. One advantage of this approach is that these bounds lead, respectively, to the convergences:

17: Try!

$$\partial_t u_{k,n} \overset{*}{\rightharpoonup} u \qquad \text{weakly* in } \mathcal{M}(I;H^{-1}(\Omega)), \tag{4.24a}$$

$$\partial_t u_{k,n} \rightharpoonup u \qquad \text{weakly in } L^2(I;H^{-1}(\Omega)). \tag{4.24b}$$

From this it is then straightforward to pass to the limit in the time derivative term without the need to go through (4.20). In addition, the identification of the initial condition is also immediate, thanks to the continuity of the temporal trace operator $u \in W^{1,2,2}(I;H^1_0(\Omega), H^{-1}(\Omega)) \mapsto u(0) \in L^2(\Omega)$.

The reason why sometimes people try to avoid this argument is that the stability bound (4.23) is not valid for arbitrary families of meshes $\mathcal{T}_n$ (in a finite element context). However, it will be satisfied for most mesh refinement strategies used in practice [18]. Moreover, this bound is a necessary condition to guarantee that the norms $\|\cdot\|_{V_n^*}$ and $\|\cdot\|_{H^{-1}(\Omega)}$ are equivalent, which is in turn a necessary ingredient in proving quasi-optimality of the space-time formulation (4.11). In summary, assuming the Sobolev stability of the L^2-projector is arguably not the end of the world, and it makes life substantially easier.

18: Try to prove this assuming a quasiuniform family of meshes.

4.2 Exercises

Exercise 4.1 (Strong convergence II) In the proof of Theorem 4.1.1, we saw that the weakly converging discrete approximations $u_{k,n} \rightharpoonup u$ satisfy:

$$\limsup_{k,n\to\infty} \langle A_k(u_{k,n}), u_{k,n}\rangle_{L^2(I;H_0^1(\Omega))} \leq \langle \overline{A}, u\rangle_{L^2(I;H_0^1(\Omega))}, \tag{4.25}$$

where $A_k \in L^2(I;H_0^1(\Omega)) \to L^2(I;H^{-1}(\Omega))$ is defined for all $v, w \in L^2(I;H_0^1(\Omega))$ through $\langle A_k(v), w\rangle_{L^2(I;H_0^1(\Omega))} := \int_Q \underline{a}_k \nabla v \cdot \nabla w$, and $\overline{A}$ is the weak* limit of $A_k(u_{k,n})$ in $L^2(I;H^{-1}(\Omega))$[19].

Prove that (4.25) implies that $u_{k,n} \to u$ strongly in $L^2(I;H_0^1(\Omega))$, assuming that A_k is (uniformly in k) strongly monotone and Lipschitz [20]. Conclude that $\overline{A} = A(u)$.

19: In Theorem 4.1.1, since the operators were linear and bounded, it was clear that $\overline{A} = A(u)$.

20: Why does the weak* limit exist in the first place?

Exercise 4.2 (Strong measurability of semidiscrete functions) Let V_n be a finite-dimensional subspace of a Banach space V with a basis $\{\varphi_j\}_{j=1}^n$, and let $\{\psi_j\}_{j=1}^n$ be functions in $L^1(I)$ with $I \subset \mathbb{R}$ an interval. Prove that the function $v\colon I \to V$ defined as

$$v(t) := \sum_{j=1}^{n} \psi_j(t)\varphi_j, \tag{4.26}$$

is strongly measurable.

Exercise 4.3 (General forcing terms) In the analysis of the fully discrete approximation (4.10), we assumed that the load $f \in C(I;H^{-1}(\Omega))$ was continuous in time. For general loads in the dual of the energy space $f \in L^2(I;H^{-1}(\Omega))$, we could define the piecewise-constant (in time) approximations $f_k\colon I \to H^{-1}(\Omega)$ through:

$$f_k(t) := f_k^j := \frac{1}{\tau}\int_{I_j} f(s)\,\mathrm{d}s \qquad \text{for } t \in I_j := (t_{j-1}, t_j],\ j \in \{1,\dots,m_k\}, \tag{4.27}$$

and use f_k instead of f; this is well-defined, since point values of f_k are well-defined (as opposed to those of f). Provide a justification for why basically the same proof would yield convergence to a weak solution with load f (similar remarks can be made about $A(t)$).

Exercise 4.4 (Summation-by-parts) Take two finite sequences $\{v_j\}_{j=0}^k, \{w_j\}_{j=0}^k$ of elements of a Hilbert space H, and denote the backward difference quotient (with a given time step $\tau > 0$) by

$$\mathrm{d}_\tau v_j := \frac{v_j - v_{j-1}}{\tau} \qquad \text{for } j \in \{1,\dots,k\}. \tag{4.28}$$

(a) Prove the discrete product rule $\mathrm{d}_\tau(v_j, w_j)_H = (\mathrm{d}_\tau v_j, w_j)_H + (v_{j-1}, \mathrm{d}_\tau w_j)$, for $j \in \{1,\dots,k\}$.

(b) Prove the Summation-by-parts formula:

$$\tau\sum_{j=1}^{k}\left[(\mathrm{d}_\tau v_j, w_j)_H + (v_{j-1}, \mathrm{d}_\tau w_j)_H\right] = (v_k, w_k)_H - (v_0, w_0). \tag{4.29}$$