A Coq Formal Proof of the Lax–Milgram theorem

Sylvie Boldo, François Clément, Florian Faissole, Vincent Martin, Micaela Mayero

A Coq Formal Proof of the Lax–Milgram theorem

Sylvie Boldo
Inria
LRI, Université Paris-Sud & CNRS,
Université Paris-Saclay
sylvie.boldo@inria.fr

François Clément
Inria
2 rue Simone Iff, CS 42112,
FR-75589 Paris cedex 12
francois.clement@inria.fr

Florian Faissole
Inria
LRI, Université Paris-Sud & CNRS,
Université Paris-Saclay
florian.faissole@inria.fr

Vincent Martin
LMAC, UTC, BP 20529,
FR-60205 Compiègne, France.
vincent.martin@utc.fr

Micaela Mayero
LIPN, Université Paris 13, CNRS UMR 7030,
Villetaneuse, F-93430
mayero@lipn.univ-paris13.fr

Abstract

The Finite Element Method is a widely-used method to solve numerical problems coming for instance from physics or biology. To obtain the highest confidence on the correction of numerical simulation programs implementing the Finite Element Method, one has to formalize the mathematical notions and results that allow to establish the soundness of the method. The Lax–Milgram theorem may be seen as one of those theoretical cornerstones: under some completeness and coercivity assumptions, it states existence and uniqueness of the solution to the weak formulation of some boundary value problems. This article presents the full formal proof of the Lax–Milgram theorem in Coq. It requires many results from linear algebra, geometry, functional analysis, and Hilbert spaces.

Keywords formal proof, Coq, finite element method, functional analysis, Lax-Milgram theorem

1. Introduction

The Finite Element Method is now widely used to solve Partial Differential Equations (PDEs) written in weak formulation. It can be applied to simulate numerically many problems arising for instance from physics, biology, or mechanics. Its success is partly due to its sound mathematical ground, see for instance (Ciarlet 2002; Ern and Guermond 2004; Zienkiewicz et al. 2013) among a huge literature.

To vouch the Finite Element Method, the Lax–Milgram theorem is one of the main results: it is a way to establish existence and uniqueness of the solution to the continuous problem (i.e. defined on an infinite dimensional functional space) and of its discrete approximation (defined on a finite dimensional space); it is valid for coercive bilinear mappings defined on Hilbert spaces (complete normed vector spaces endowed with an inner product). The corollary known as Céa's lemma provides a bound on the error between the computed approximation and the unknown solution.

The present work is a mandatory step toward the correctness of programs implementing the method, for instance built above the FELiScE[1] library written in C++. Such a formal verification would increase the trust in both the library and all the programs relying on it. This proof will represent an advance at the same time for the Finite Element Method, but also for the safety of critical software using complex numerical algorithms.

Other mathematical results can be used to prove this kind of existence and uniqueness results, such as the more general Banach–Nečas–Babuška theorem. But the choice was made to focus on the Lax–Milgram theorem because it is sufficient to solve a wide range of PDEs using standard approximations, and there exists a proof using fairly elementary tools. Nonetheless, although elementary in its setting, the proof remains quite intricate and spans large parts of various domains of mathematics, such as linear and bilinear algebra, analysis (including the Banach fixed-point theorem and completeness), geometry (orthogonal projection). Thus, the formalization of the Lax–Milgram theorem was interesting by itself. The description of this formalization is the subject

[1] Finite Elements for Life Sciences and Engineering
`https://gforge.inria.fr/projects/felisce/`

of the present article, based on a very detailed unpublished pen-and-paper proof (Clément and Martin 2016).

Let $(E, \langle \cdot, \cdot \rangle_E)$ be a real Hilbert space, and let $\|\cdot\|_E$ be the associated norm. Let E' be its topological dual, i.e. the space of continuous linear forms from E to $\mathbb{R}$ (see also Section 2.3). Let E_h be a closed vector subspace of E (in practice, E_h is finite-dimensional: it is the finite element approximation space). Given a bounded bilinear form a and a continuous linear form f, let the general problems in weak formulation be defined as

$$\text{find } u \in E \text{ s.t.} \quad \forall v \in E,\ a(u,v) = f(v); \tag{1}$$

$$\text{find } u_h \in E_h \text{ s.t.} \quad \forall v_h \in E_h,\ a(u_h,v_h) = f(v_h). \tag{2}$$

With the preceding hypotheses on a and f, the Lax–Milgram theorem can be stated as

Textbook Theorem 1 (Lax–Milgram).
Assume that a is coercive with constant $\alpha > 0$. Then, there exists a unique $u \in E$ solution to Problem (1). *Moreover,*

$$\|u\|_E \leq \frac{1}{\alpha} \|f\|_{E'}.$$

Boundedness and coercivity are defined in Definitions 1 and 4 (see Sections 2.3.3 and 3.2.4). Note that theorems from textbooks are framed, while formally proved theorems are not. All the textbooks theorems presented here have a formal counterpart. For instance, Textbook Theorem 1 corresponds to Theorem 7 (see Section 3.2.4).

The Lax–Milgram theorem also applies on the closed subspace E_h of E, proving existence and uniqueness for Problem (2). Moreover, under the same boundedness and coercivity hypotheses, there holds

Textbook Theorem 2 (Céa's lemma).
Let $C \geq 0$ be a continuity constant of the bounded bilinear form a. Let $u \in E$ be the unique solution to Problem (1). *Let $u_h \in E_h$ be the unique solution to Problem* (2). *Then,*

$$\forall v_h \in E_h, \quad \|u - u_h\|_E \leq \frac{C}{\alpha} \|u - v_h\|_E.$$

Its formal counterpart is Theorem 8 (see also Section 3.2.4).

An important intermediate result is the Riesz–Fréchet representation theorem. It states that any continuous linear form can be uniquely mapped onto a vector of the Hilbert space:

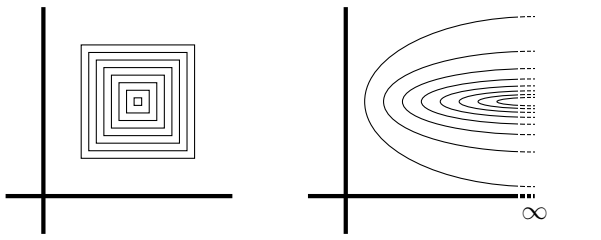

Figure 1. Examples of filters, toward a finite point, and a point at infinity.

Textbook Theorem 3 (Riesz–Fréchet).
Let $\varphi \in E'$ be a continuous linear form on E. Then, there exists a unique vector $u \in E$ such that for all $v \in E$, $\varphi(v) = \langle u, v \rangle_E$. Moreover, the mapping $\tau = (\varphi \mapsto u)$ is a continuous isometric isomorphism from E' onto E.

Its formal counterpart is split into Theorem 5, Definition 3, and Lemma 6 (see Sections 3.2.2 and 3.2.3).

Interactive theorem proving has long been more interested in formalizing parts of algebra rather than analysis. It is now changing as more and more interest goes into real and numerical analysis. Real-life applications, such as hybrid systems or cyber-physical systems are critical and rely on advanced analysis results. In particular, Isabelle has a large library of results, including many results about ODEs (Immler and Hölzl 2012; Immler 2014; Immler and Traut 2016). Our work is especially close to the recent work by Immler and Traut (Immler and Traut 2016) (this will be highlighted in Section 2.3).

This work relies on the Coq proof assistant. Previous work about analysis in Coq include the full formalization of the discretization of the wave equation (Boldo et al. 2013) and the formalization of Picard's operator for ODEs (Makarov and Spitters 2013). We rely on the Coquelicot library (Boldo et al. 2015; Lelay 2015b,a). Coquelicot is a conservative extension of Coq's standard library, with total functions for limit, derivative, and integrals. Its topology is defined using filters. Filters are sets of neighborhoods, with a few assumptions. They can represent converging sets such as shown in Figure 1. Moreover, Coquelicot includes a hierarchy of algebraic structures: group, ring, module (`ModuleSpace`), normed module (`NormedModule`), and so on inspired from (Hölzl et al. 2013) and based on canonical structures.

A difference between Isabelle developments and ours is that Coq has an intuitionistic logic. Our choice is to stay intuitionistic as much as possible. Yet, this work is not fully intuitionistic for several reasons: first, we rely on the standard Coq library that axiomatizes real numbers. Second, we require several decidability hypothesis (see Section 3.3). Moreover, we have several occurrences of $\neg\neg\varphi$, instead of

φ, due to our decision to require as less decidable hypothesis as possible, in particular we do not assume the decidability of φ by default.

The Coq code is available at the following address: `http://www.lri.fr/~sboldo/elfic/index.html` with both the code and a web interface.

The paper is organized as follows. Section 2 is devoted to the formalization of algebraic structures, substructures, and continuous linear mappings. The main steps of the proof of the Lax–Milgram theorem, including geometry results, are detailed in Section 3. Section 4 concludes and gives some perspectives.

2. Spaces and subspaces

Céa's lemma (Textbook Theorem 2) is based on the application of Textbook Theorem 1 on both a Hilbert space E and a closed subspace E_h. Of course, E_h is also a Hilbert space itself, but we found more convenient to formally prove a generalization of the Lax–Milgram theorem stated on a (closed) subspace of an ambient Hilbert space. Therefore, most of the theorems presented here consider a given subset, with properties such as completeness. How to define a subset in Coq is known to be difficult and several possibilities are offered, and used in this development.

This section is organized as follows. Section 2.1 presents the Banach fixed-point theorem on a subset. Section 2.2 presents the (canonical) structures used for defining pre-Hilbert and Hilbert spaces. Section 2.3 presents the continuous and linear mappings, as a subspace of a functional space. Section 2.4 comes back to the various substructure definitions to compare them.

2.1 Banach fixed-point theorem on a subset

A fixed-point theorem is not a new stuff. The interesting point here is that it applies on a subset, described by a characteristic function. Assuming the subset is nonempty and complete, the function is a contraction map on the subset, and the input point is in the subset, then the iterates of the function are converging in the subset toward the fixed-point.

Another subtle point is the assumptions on the space: considering a `UniformSpace` in Coquelicot, the topology is only described by balls. Then, being Lipschitz is defined by:

```
Definition is_Lipschitz (f: X → Y) (k:R) := 0 <= k ∧
  forall x1 x2 r, 0 < r →
      ball_x x1 r x2 → ball_y (f x1) (k*r) (f x2).
```

A contraction is a Lipschitz map with $k < 1$.

However, working with balls instead of a distance has drawbacks. In particular, there may be points which are not comparable. We require that any two points *in the subset* are at a certain distance. More precisely, for any x and y in the subset, there exists a positive real M such that `ball` x M y. Then this distance is multiplied by the contracting factor at each step, providing a convergent sequence. As topology is based on filters in Coquelicot, we consider the following proper and Cauchy filter:

```
F := (fun P ⇒ eventually (fun n ⇒ P (iter f n x)))
```

with `x` being any element of the subset. Then, we have the following fixed-point theorem on the subset `phi`:

```
Context {X : CompleteSpace}.
Definition is_eq : X → X → Prop := fun a b
      ⇒ forall eps:posreal, ball a eps b.
[...]
Hypothesis phi_distanceable: forall (x y:X),
  phi x → phi y→ exists M, 0 <= M ∧ ball x M y.
Hypothesis phi_complete: my_complete phi.

Theorem FixedPoint_C: is_contraction f →
  exists a:X, phi a ∧ is_eq (f a) a
    ∧ (forall b, phi b → is_eq (f b) b → is_eq b a)
    ∧ forall x, phi x → is_eq
   (lim (fun P ⇒ eventually (fun n ⇒ P (iter f n x)))) a.
```

It states existence and uniqueness of the limit. Moreover, it states that any initial point gives the same limit.

Note that in practice, we rely on a version of this theorem on `CompleteNormedModule`, where `is_eq` can be replaced by the native equality and the `phi_distanceable` hypothesis can be removed. But we prove and present here the most general version.

A last point in this proof is the first encounter with decidability issues, that will go on during the whole article. The subset `phi` must be complete. Coquelicot provides the canonical structure `CompleteSpace`, but we need here to specify completeness for a subset. Thus, we use the following definition:

```
Context {X : CompleteSpace}.
Definition my_complete (phi : X → Prop) :=
  forall F, ProperFilter F → cauchy F →
    (forall P, F P → ¬¬(exists x, P x ∧ phi x))
        → phi (lim F).
```

Its meaning is: given a proper and Cauchy filter such that it always intersects the subset, then the limit belongs to the subset. In fact, it is not "always intersects", but "does never not intersect", which is weaker. It corresponds to the mathematical meaning: "a Cauchy sequence of the subset converges into the subset" expressed with filters.

This definition of completeness is linked to the closedness (defined in Coquelicot as the complement of open subsets). More precisely, we were able to prove they are quasi-equivalent:

```
closed phi ↔ my_complete (fun x ⇒ ¬¬phi x).
```

The first implication holds on a `CompleteSpace`, but we need a `CompleteNormedModule` for the second one. Indeed, it is known from textbooks that the second implication is valid on separated spaces (or Hausdorff spaces), that metric spaces are separated, and that normed vector

spaces are metric spaces. In our case, the separation of a `CompleteNormedModule` is expressed as:

```
forall x y : X, is_eq x y → x = y.
```

2.2 Functional spaces, pre-Hilbert and Hilbert spaces

Canonical structures are used intensively in this work. In particular, functional spaces can be seen as spaces with properties. For instance, we prove that, if F is a `ModuleSpace`, then $E \to F$ is also a `ModuleSpace`. An immediate advantage is that we may write for $(\texttt{f},\texttt{g} : E \to F)$ and $(\texttt{x},\texttt{y} : E)$:

```
(plus f g) (plus x y) = zero
```

with Coq guessing the correct `plus` operators: `(plus f g)` is in $E \to F$ so it guesses the addition of functions; `(plus x y)` is in E so it guesses $+_E$; Coq also guesses `zero` as being 0_F.

The Lax–Milgram theorem is stated on a Hilbert space. A Hilbert space is a structure built upon a series of structures. See Figure 2 for a sketch of the hierarchy.

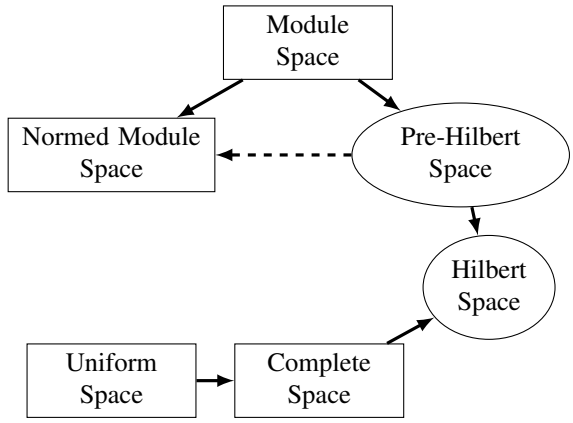

Figure 2. Hierarchy of structures. Rectangles are Coquelicot structures and ovals are the structures we add. Plain arrows mean that the structure is built from another. The dashed arrow means we prove that the structure is an instance of another.

A pre-Hilbert space (or inner product space) is a module with an inner product:

```
Record mixin_of (E : ModuleSpace R_Ring) := Mixin {
  inner : E → E → R ;
  ax1 : forall (x y : E), inner x y = inner y x ;
  ax2 : forall (x : E), 0 <= inner x x ;
  ax3 : forall (x : E), inner x x = 0 → x = zero ;
  ax4 : forall (x y : E) (l:R),
            inner (scal l x) y = l * inner x y ;
  ax5 : forall (x y z : E),
            inner (plus x y) z = inner x z + inner y z
}.
```

where `Mixin` is the common constructor for canonical structures (Mahboubi and Tassi 2013).

Given the inner product, we define a norm and prove its properties. In particular, we prove that a pre-Hilbert space is (an instance of) a `NormedModule`, so that we may use the corresponding properties. It corresponds to the dashed arrow of Figure 2.

A Hilbert space is a complete pre-Hilbert space, with exactly the Coquelicot definition of completeness based on filters:

```
Record mixin_of (E : PreHilbert) := Mixin {
  lim : ((E → Prop) → Prop) → E ;
  ax1 : forall F, ProperFilter F → cauchy F →
        forall eps : posreal, F (ball (lim F) eps)
}.
```

2.3 Continuous and linear mappings

The Lax–Milgram theorem involves continuous linear mappings on a Hilbert space. While the linearity is not really difficult to formalize, there are many ways to define the continuity of a linear mapping. Unsurprisingly, our definition of continuous and linear mappings looks very alike that of (Immler and Traut 2016), but this was unknown to us at that moment as both formal developments were done in parallel. A difference is that their instantiations of continuous and linear mappings as groups, modules, and so on is simplified by the use of the Isabelle/HOL lifting and transfer package (Huffman and Kunčar 2013), while we do it by hand.

2.3.1 Linear and bilinear mappings

When E, F : `ModuleSpace`, we define the predicate `is_linear_mapping` on $E \to F$ to characterize linear mappings. We then prove `is_linear_mapping` is compliant both with the group structure and the module structure, so that it defines a subgroup and a sub-module (see Section 2.4).

Furthermore, we define bilinear mappings, i.e. linear in each argument separately.

2.3.2 Operator norm

In the sequel, we consider E, F : `NormedModule`. Given $f : E \to F$ and $\varphi : E \to$ `Prop`, we define the operator norm of f on the subset φ, noted $|||f|||_\varphi$. It is an element of $\overline{\mathbb{R}}$ defined by $|||f|||_\varphi = \sup_{u \neq 0_E \wedge \varphi(u)} \frac{\|f(u)\|_F}{\|u\|_E}$. As we want to define it as a total function, we need to distinguish the case where φ is $\{0_E\}$: then the set is empty and its supremum is not the value we expect. The emptiness of a subset of $\mathbb{R}$ is decidable thanks to Coquelicot and the axioms defining the reals in Coq. Therefore, we are able to decide whether a subset of E is $\{0_E\}$, by deciding if the following subset of $\mathbb{R}$ is empty:

```
fun x ⇒ exists u:E, u <> zero ∧ phi u ∧ x = norm u.
```

Below, the notation `Is_only_zero_set_dec_phi` refers to this decidability property, which allows to define the operator norm on φ:

```
Definition operator_norm_phi (f:E→ F) : Rbar :=
 match Is_only_zero_set_dec_phi E with
  | left _ ⇒ Lub_Rbar (fun x ⇒ exists u:E, u <> zero ∧
                        phi u ∧  x = norm (f u) / norm u)
  | right _ ⇒ 0
 end.
```

2.3.3 Continuous linear mappings

We prove the equivalences between eight different definitions of the continuity of linear mappings of $E \to F$. In particular, the continuity of a linear mapping is equivalent to the finiteness of the associated operator norm on E.

Continuity is defined in (Immler and Traut 2016) as $\exists K, \forall y, |f(y)| \leq K \cdot |y|$ and the operator norm is $\max\{|f(y)|, \text{such that } |y| \leq 1\}$. It is quite similar, except that we take into account subsets to have an operator norm on φ.

We define the type of continuous linear mappings as a depend record with the proofs of linearity and continuity, and with a coercion between the continuous linear mapping and the mapping itself:

```
Record clm := Clm {
  m:> E→ F ;
  Lf: is_linear_mapping m;
  Cf: is_finite (operator_norm m) }.
```

Below we note $\texttt{clm}(E, F)$ the space of continuous linear mappings from E to F. The $\texttt{clm}(E, F)$ set is proved to be a `ModuleSpace` and a `UniformSpace`. As the operator norm is finite, it is even a `NormedModule` with the operator norm as norm. Note that the use of a depend record implies the need of the `ProofIrrelevance` property to show equalities between elements of $\texttt{clm}(E, F)$ and thus that it is a `ModuleSpace`. We also rely on the `FunctionalExtensionality`: we want continuous linear mappings that are point-wise equal to be equal, as in mathematics.

Another known mathematical property of $\texttt{clm}(E, F)$ is its completeness. To be comprehensive, we also want to have this property. It turned out to be more challenging than expected due to the dependent types. The Coquelicot definition of the completeness is a total function `lim` that takes a filter and produces an element, and a property that characterizes the limit, provided the filter is a proper Cauchy filter (see the definition in the definition of a Hilbert space in Section 2.2). This means that, given a set of neighborhoods of continuous linear mappings, we have to produce a continuous linear mapping, even when the filter is not a converging one, and the good one in the case of convergence. To circumvent this problem, we create a new dependent completeness, where we produce an element only if the filter is a proper Cauchy one. This makes the statements unpleasant as dependent types are creeping in. But we are able to define some dependent equivalent of the iota operator. Finally, we prove that $\texttt{clm}(E, F)$ is a complete space in the dependent meaning. Note that this was useless for Lax-Milgram theorem, and we are unsure this will be used due to the impracticality of the approach.

The specific case of $F = \mathbb{R}$ is of paramount importance for the proof of the Lax–Milgram theorem. The `NormedModule` of continuous linear forms on E is denoted $E' = \texttt{clm}(E, \mathbb{R})$; it is called the *topological dual* of E.

A first needed definition is that of a bounded mapping:

Definition 1 (boundedness).
A function $f : E \to F \to \mathbb{R}$ is bounded iff

$$\exists C \in \mathbb{R},\ \forall x \in E,\ \forall y \in F, \quad \|f(x, y)\|_F \leq C \|x\|_E \|y\|_E .$$

Now, we may state our first simple lemma: the application of the bilinear form to its first argument can be represented by a continuous linear form that depends on this first argument, as done in (1). More precisely, we have

Lemma 1 (representation of bilinear forms).
Let $a : E \to E \to \mathbb{R}$. If a is bilinear and bounded, then there exists a unique $\mathcal{A}$ in $\texttt{clm}(E, E')$ such that

$$\forall u, v \in E, \quad a(u, v) = (\mathcal{A}(u))(v).$$

Note that to be closer to mathematical notations, total applications of bilinear mappings are displayed under their uncurried form.

Lemma 1 can be seen as defining a partial application, but the proof lies in the type of $\mathcal{A}$ that should give an element in E', therefore a dependent record that represents a continuous linear mapping. We just have to prove that any partial application is both linear and continuous.

2.4 Subspaces

Contrarily to other systems such as PVS, subtyping is hardly developed in Coq. Therefore, defining a substructure is known to be a difficult problem (Mahboubi 2013). A possibility was to rely on a decidability predicate, such as in the proof of the Feit–Thompson theorem, or odd order theorem, (Gonthier et al. 2013). As said before, we want to stay as intuitionistic as possible, therefore we choose not to be able to decide if an element belongs to the subgroup or not. Another difference is that we consider functions over generic spaces, for example the subgroup of linear mappings. Deciding if a function is linear over $\mathbb{R}$ (or over $\mathbb{R} \to \mathbb{R}$) is less straightforward than belonging to a finite group.

Our problem can be easily explained by considering the subgroup of linear mappings. It is a subgroup (and a submodule, and so on), but we would like to use the same operations on a linear mapping as on a function (addition, application, and so on). Therefore, we would like to have both the properties of the subgroup and the operations of the overgroup. This is more general, as we consider not only groups, but also modules, pre-Hilbert and Hilbert spaces.

Two possibilities are available in Coq. Surprisingly, we use both. The first one is the use of a depend record as in the

definition of continuous linear mappings in Section 2.3. We define a record with an element of the structure and a proof that it belongs to the substructure. There is left to define the structure operations and prove that this set is indeed the same structure as its over-structure. This does not fully respect our wish as the operations on the substructure are not that on the over-structure. This is partially solved by defining a coercion from the record to the over-structure element, so that it can most of the time be used as an over-structure element.

The second possibility is the most used one in this development, see for example Sections 2.1 and 3.2.1. It is to consider the characteristic function of the substructure. Then, we have to provide a proof that it has several properties. For example, a subgroup needs to include the zero and to be compliant with the addition. The characteristic function is then called `AbelianGroup`-compatible. When it is also compliant with the scalar multiplication, the characteristic function is called `ModuleSpace`-compatible.

For the sake of completeness, we mix the two approaches. More precisely, we prove that, given a characteristic function having the compliance properties, we are able to construct a dependent record and to prove it is also a group or a module. For that, we rely on the `ProofIrrelevance` axiom.

3. Main steps of the proof of the Lax–Milgram theorem

Below we present several main steps of the proof of the Lax–Milgram theorem. First, we exhibit some geometrical results on subspaces of pre-Hilbert spaces in Section 3.1. Then, in Section 3.2, we detail the proofs of the Riesz–Fréchet theorem, which is the main intermediate theorem, and the Lax–Milgram theorem.

In the following statements we will suppose some decidability hypotheses. We consider a subspace `phi` defined as a characteristic function (see Section 2.4) and we define

```
H1 (phi : E → Prop) :
    forall u : E, forall eps : posreal,
    decidable (exists w : E, phi w ∧ norm (minus u w) < eps).
H2 (phi : E → Prop) (f : topo_dual E) :
    decidable (exists u, ¬¬phi u ∧ f u <> 0).
```

In Section 3.3, we discuss these decidability hypotheses.

3.1 Geometrical results

In this section, we define geometrical notions as the orthogonal projection of a vector and the orthogonal complement of a subspace. Until further notice, E is a pre-Hilbert space.

3.1.1 Orthogonal projection

Theorem 2 (orthogonal projection onto a nonempty complete convex subset). *Let $\varphi : E \to$ `Prop`, φ nonempty, complete, convex. Suppose $\mathcal{H}_1(\varphi)$. Then, for all $u \in E$, there exists a unique $P_\varphi(u) \in E$ such that*

$$\varphi(P_\varphi(u)) \quad \wedge \quad \|u - P_\varphi(u)\|_E = \min_{w \in E \wedge \varphi(w)} \|u - w\|_E .$$

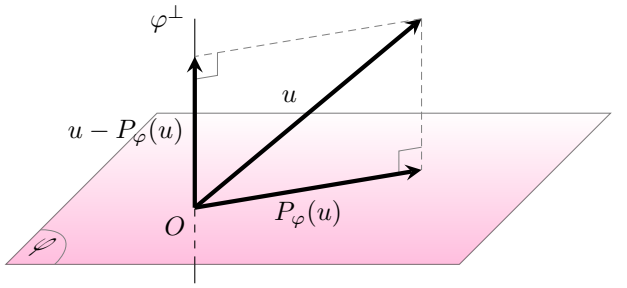

Figure 3. Orthogonal projections on a subspace φ and on its orthogonal complement $\varphi^\perp$.

$P_\varphi(u)$ is called the *orthogonal projection* of u onto φ, as seen in Figure 3.

The proof of this theorem is adapted from the pen-and-paper proof (Clément and Martin 2016) as our topological definitions rely on filters and as the existence of the orthogonal projection is a consequence of the completeness of φ.

Let $\delta = \min_{w \in E \wedge \varphi(w)} \|u - w\|_E$. We build the orthogonal projection as the limit of the filter defined by:

```
F := fun (V:E→ Prop) ⇒ exists eps:posreal, forall x,
       phi x → norm (minus u x) < delta + eps → V x.
```

We need to prove that it is a proper and Cauchy filter, then to show we can approach δ by a real sequence, showing

$$\forall \varepsilon > 0,\ \exists x \in E, \quad \varphi(x) \ \wedge \ \|u - x\|_E < \delta + \varepsilon.$$

However, it is not constructively possible (see Section 3.3.2). Note that the existence proof is technical due to the use of filters whereas the uniqueness proof is a pure manipulation of mathematical expressions.

We want to characterize whether an element is the orthogonal projection of another onto a subspace φ. We prove the following characterization lemma which is valid for any `ModuleSpace`-compatible subset φ of E:

Lemma 3 (characterization of the orthogonal projection onto a subspace). *Let $\varphi : E \to$ `Prop`, φ `ModuleSpace`-compatible. Suppose $\mathcal{H}_1(\varphi)$. Let $u, v \in E$ with $\varphi(v)$. Then,*

$$\left(\|u - v\|_E = \inf_{w \in E \wedge \varphi(w)} \|u - w\|_E\right) \iff \\ (\forall w \in E,\ \varphi(w) \implies \langle v, w\rangle_E = \langle u, w\rangle_E) .$$

Moreover, when φ is complete, this lemma characterizes the relation between u and its orthogonal projection $P_\varphi(u) = v$ onto φ.

3.1.2 Orthogonal complement and direct sum

Definition 2 (orthogonal complement).
Let $\varphi : E \to$ `Prop`, φ `ModuleSpace`-compatible. Its orthogonal complement is the subset $\varphi^\perp$ of E defined as

```
Definition orth_compl :=
    fun x : E ⇒ forall (y : E), phi y → inner x y = 0.
```

When φ is complete, any vector $u \in E$ is the sum of two vectors respectively in φ and $\varphi^\perp$. Moreover, this decomposition is unique: we say that E is the *direct sum* of φ and $\varphi^\perp$, and we note $E = \varphi \oplus \varphi^\perp$. Furthermore, it is possible to exhibit this unique decomposition: $u = P_\varphi(u) + (u - P_\varphi(u))$ (see Figure 3). The uniqueness of the decomposition comes from the uniqueness of the orthogonal projection provided by Theorem 2. To formalize the direct sum $\varphi \oplus \varphi'$, we provide several equivalent definitions.

3.2 Riesz–Fréchet and Lax–Milgram theorems

In this section except Section 3.2.1, we suppose E : `Hilbert`. We first define the kernel of a linear mapping in Section 3.2.1. In Section 3.2.2, we prove an intermediate result called the Riesz–Fréchet theorem and define the corresponding function in Section 3.2.3. This permits — with the Banach fixed-point theorem — to prove the Lax–Milgram theorem in Section 3.2.4.

3.2.1 Kernel of a linear mapping

Let $f : E \to F$ (with E, F : `ModuleSpace`). The kernel of f is the subset $\ker(f)$ of E defined by

```
Definition ker (f : E → F) := fun x:E ⇒ f x = zero.
```

When E : `Hilbert` and $F = \mathbb{R}$, the most useful property is

Lemma 4. *If f is in E', then $\ker(f)$ is a `ModuleSpace`-compatible and complete subset of E.*

While the compatibility result is easy, the completeness of $\ker(f)$ is a little more technical. The straightforward result is `closed`$(\ker(f))$, hence, `my_complete`$(\neg\neg \ker(f))$ (see Section 2.1). To be able to remove the double negation, and prove `my_complete`$(\ker(f))$, we need the decidability of $\ker(f)$. This means that for any x in E, we have to decide whether $f(x)$ is zero or not. Fortunately, the linear form f takes its values in $\mathbb{R}$ and the equality of real numbers is decidable in the Coq real standard library. Therefore, we have $\forall x,$ `Decidable`$(\ker(f)(x))$. And thus, we have `my_complete`$(\ker(f))$. Note that this is true if and only if E is complete, and so the result is false in incomplete pre-Hilbert spaces.

3.2.2 Riesz–Fréchet theorem

The Riesz–Fréchet theorem is the main intermediate result to prove the Lax–Milgram theorem. It involves most of the previously defined mathematical concepts. This representation theorem exhibits an element equivalent to a given continuous linear form.

Theorem 5 (Riesz–Fréchet).
Let $f \in E'$. Let $\varphi : E \to$ `Prop`, φ `ModuleSpace`-compatible and complete.
Suppose $\forall f \in E',\ \mathcal{H}_1(\ker(f) \wedge \neg\neg\varphi) \wedge \mathcal{H}_2(\varphi, f)$.
Then, there exists a unique $u \in E$ such that

$$\neg\neg\varphi(u) \quad \wedge \quad \forall v \in E,\ \neg\neg\varphi(v) \implies f(v) = \langle u, v\rangle_E\,.$$

It is the formal counterpart of the main part of Textbook Theorem 3 (from the introduction), see also Definition 3 and Lemma 6 below.

While it is simple to prove uniqueness, the existence proof presents several difficulties. The paper proof is classical because we have to distinguish two cases, one where f is the null-point function, and the other where there exists an element u_0 such that $f(u_0) \neq 0$ (see also Section 3.3.1).

Moreover, there are many real number manipulations, which could be partially left to the reader in a paper proof, but not in a Coq development. For instance, when f is a non null-point function, the proof consists mainly in building a correct solution by a complex formula which involves the projection of u_0 on $\ker(f)^\perp$: $u_0 - P_{\ker(f)}(u_0)$ (see Figure 3). As the norm of this vector appears in denominators, we prove it is a nonzero vector. After that, we show that the built candidate is a solution. This was unexpectedly tedious as the canonical form of two akin terms including fractions are not always akin.

3.2.3 Riesz representation function

For each continuous linear form $f \in E'$, the Riesz–Fréchet theorem ensures the existence of a unique solution to a given problem. Even if we prove there exists such a solution, the theorem does not give a function to calculate this solution. Mathematicians define a function $\tau : E' \to E$ which takes f and returns the associated solution:

Definition 3 (Riesz representation function).
According to Theorem 5, for any $f \in E'$, there exists a unique $u_f \in E$ such that $\forall v \in E,\ f(v) = \langle u_f, v\rangle_E$. The Riesz representation function is the function $\tau = (f \mapsto u_f)$.

We want to define τ in a constructive way, as a total function which returns the expected element. A first solution is to use the Hilbert's operator `epsilon`, and more precisely the Hilbert's operator `iota`, both provided by the Coq standard library, whose existence needs Hilbert's axiom `Epsilon`.

However we want to be as intuitionistic as possible. When E : `CompleteSpace`, the Coquelicot library provides — without any additional axiom — an `iota` operator with the same behavior as the one from the standard library. In this paper, the `iota` notation shall refer to Coquelicot's operator with which we define τ as a total function:

```
Definition tau := fun (f:topo_dual E) ⇒
  (iota (fun u : E ⇒ forall v : E, f v = inner u v)).
```

The proof of the Lax–Milgram theorem relies mainly on two properties of τ:

Lemma 6 (Riesz representation function properties).
τ is a linear mapping and $\forall f \in E',\ |||f||| = \|\tau(f)\|_E$.

It is the formal counterpart of the last part of Textbook Theorem 3 (from the introduction), see also Theorem 5 and Definition 3 above.

The first part (linearity) is easy, whereas the second part of the lemma (isometry) is harder. It has subcases whether φ

is $\{0_E\}$, and has to deal with a greatest lower bound which takes values in $\overline{\mathbb{R}}$ by default.

3.2.4 Lax–Milgram theorem

The Lax–Milgram theorem is based on two definitions, bounded (see Definition 1) and coercive.
C is called a *continuity constant* of f.

Definition 4 (coercivity).
A function $f : E \to E \to \mathbb{R}$ is α-coercive with $0 < \alpha$ iff

$$\forall x \in E, \quad \alpha \|x\|_E^2 \leq f(x, x).$$

We can now state the Lax–Milgram theorem, that is our formal variant of Textbook Theorem 1 (from the introduction).

Theorem 7 (Lax–Milgram).
Let $f \in E'$, $0 < \alpha$. Let $\varphi : E \to$ `Prop`, φ `ModuleSpace`-compatible and complete. Let a be a bilinear form on E, bounded and α-coercive.
Suppose $\forall f \in E'$, $\mathcal{H}_1(\ker(f)) \wedge \neg\neg\varphi) \wedge \mathcal{H}_2(\varphi, f)$.
Then, there exists a unique $u \in E$ such that

$$\begin{array}{c} \neg\neg\varphi(u) \quad \wedge \\ \forall v \in E, \quad \neg\neg\varphi(v) \implies f(v) = a(u, v) \quad \wedge \\ \|u\|_E \leq \alpha \cdot |||f|||_\varphi. \end{array}$$

Both the detailed pen-and-paper proof and the formal proof of the Riesz–Fréchet theorem are already substantial, but this is getting more complicated for the Lax–Milgram theorem. There are a lot of intermediate calculations ensuring mathematical inequalities or equalities which are useful in the proof. Our choice is to treat them aside as a bunch of auxiliary lemmas.

With regard to the proof itself, we first show that existence and uniqueness of the solution for a simpler problem implies the Lax–Milgram theorem. After that, we prove that the solution of the simpler problem is actually the unique fixed point of a given function.

According to Lemma 1, let $\mathcal{A}$ be the unique continuous linear mapping from E to its topological dual E' such that

$$\forall u, v \in E, \ (\mathcal{A}(u))(v) = a(u, v).$$

Then, using the Riesz–Fréchet theorem twice, we can replace the general problem (1) through the implication

$$\begin{array}{r} \exists! u \in E, \quad \tau(\mathcal{A}(u)) = \tau(f) \quad \implies \\ \exists! u \in E, \ \forall v \in E, \quad a(u, v) = f(v). \end{array} \quad (3)$$

Then, we exhibit the function $g : E \to E$ defined by

$$\forall x \in E, \ g(x) = x - \rho \cdot \tau(A(x)) + \rho \cdot \tau(f),$$

and we replace the problem by a fixed-point problem through

$$g(u) = u \quad \implies \quad \tau(\mathcal{A}(u)) = \tau(f). \quad (4)$$

We prove that g is k-Lipschitz for $k = \sqrt{1 - 2\rho\alpha + \rho^2 C^2}$, which is obviously cumbersome, and we verify that for some suitable ρ, we have $0 < k < 1$, making g a contraction. Finally, applying the Banach fixed-point theorem (Section 2.1), we get existence and uniqueness of the fixed point of g, and thus that of the general problem (1) via (4) and (3).

Applying the Lax–Milgram theorem both on a Hilbert space E and on a `ModuleSpace`-compatible and closed subset of E, we obtain a bound on the difference between the two solutions provided by the theorem. This is our formal variant of Céa's lemma, stated as Textbook Theorem 2 in the introduction.

Theorem 8 (Céa's lemma).
Let $f \in E'$, $0 < \alpha$. Let $\varphi : E \to$ `Prop`, φ `ModuleSpace`-compatible and closed. Let a be a bilinear form on E, bounded by $C > 0$ and α-coercive.
Suppose $\forall f \in E'$, $\mathcal{H}_1(\ker(f)) \wedge \neg\neg\varphi) \wedge \mathcal{H}_2(\varphi, f)$.
Be u and u_φ the two solutions given by the Lax–Milgram theorem respectively on E and on the subspace φ. Then,

$$\forall v_\varphi \in E, \quad \varphi(v_\varphi) \implies \|u - u_\varphi\|_E \leq \frac{C}{\alpha} \|u - v_\varphi\|_E .$$

This lemma is a basic block to prove that the finite element method gives reasonable approximations of PDEs. Note that the proof of Céa's lemma is not difficult itself, but makes sense only within the existence of the solutions given by the Lax–Milgram theorem.

3.3 Decidability hypotheses

The previous theorems involve a set of decidability hypotheses, weaker than the excluded-middle axiom. In this section, we discuss these hypotheses.

3.3.1 Case of the null-point function

We first look into the hypothesis $\mathcal{H}_2$ defined at the beginning of Section 3. It decides whether any continuous linear mapping is zero on $\neg\neg\varphi$, for a complete and `ModuleSpace`-compatible subset φ.

In the proof of the Riesz–Fréchet theorem, we distinguish two cases, one for the null-point function, and the other for a non null-point function f, using a witness u_0 such that $f(u_0) \neq 0$. However, it is not possible to decide constructively whether a function is null-point or not, because it would need to test the equality to 0 for each point of φ, and thus possibly an infinite set of points. This is possible using the axioms defining Coq real numbers, but this does not provide us with the witness u_0 we need.

3.3.2 Subspace and greatest lower bound

We now look into the hypothesis $\mathcal{H}_1$, which first appears in the formal proof of Theorem 2. It decides whether there exists a point in $\neg\neg\varphi$ near enough a given point in E, for a complete and `ModuleSpace`-compatible subset φ.

In the proof of Theorem 2, we have to exhibit, for a given $u \in E$ and an arbitrary small positive real ε, an element x such that

$$\neg\neg\varphi(x) \quad \wedge \quad \|u - x\|_E < \min_{w \in E \wedge \varphi(w)} \|u - w\|_E + \varepsilon.$$

It seems natural to have such a property, which corresponds to the existence of a convergent sequence in $\mathbb{R}$ whose limit is the greatest lower bound of $\{\|u - w\|_E \mid w \in E \wedge \varphi(w)\}$. Nonetheless, due to Coquelicot's definition of lower bound, we can prove "it is false that such an element does not exist", but we cannot prove the existence itself. That is why we suppose the decidability hypotheses $\mathcal{H}_1$ which is a way to have this property without importing the excluded-middle axiom.

4. Conclusion and perspectives

The Coq formal proof of the Lax–Milgram theorem is achieved. Its dependency graph is detailed in Figure 4. The formal development is about 7000-line long, as is the LaTeX source file of the detailed pen-and-paper proof (Clément and Martin 2016). To put it into perspective, as we have already noticed in a previous work (Boldo et al. 2013), in-depth pen-and-paper proofs can be an order of magnitude longer than usual proofs from textbooks, and the lengths of formal and detailed pen-and-paper proofs are similar.

In this development, several choices can be emphasized. First, the use of the `FunctionalExtensionality` and `ProofIrrelevance` axioms. Then, we want to stay as intuitionistic as possible, but we did not choose to import the classical library, just the necessary decidability properties. And it sometimes implies the surprising use of the double negation. This approach explains why some Coq statements are different than the paper-proof references.

Another difference is the tedious use of sub-structures. This is entirely hidden in pen-and-paper proofs, while we have to exhibit our substructure and its properties. A more elegant subtyping approach may have been helpful.

The use of the Coquelicot library has an advantage: most of the basic results from real analysis we needed are already there, therefore we were encouraged to use canonical structures, that were of great help. The statements are quite smooth as they really look like mathematical statements. Debugging canonical structures can be tricky, but their use, and the existing Coquelicot hierarchy and library has surely lightened our proof burden.

The main drawback of the Coquelicot library in our usage is its choices about topology: in most textbooks, limits are expressed in a metric space in terms of distances, whereas Coquelicot is based on the more general notion of filters in a uniform space.

As for the perspectives, in order to establish the soundness of the Finite Element Method, Céa's lemma has to be applied on the finite element approximation subspace,

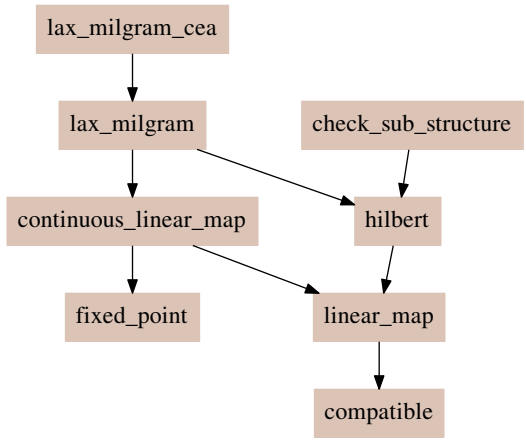

Figure 4. Dependency graph of the Coq development.

which is finite dimensional. Thus, to be able to apply Theorem 8, we have to prove that a finite dimensional subspace is always `ModuleSpace`-compatible and closed. This work is currently under progress, and although `ModuleSpace`-compatibility is obvious, closedness seems to be more challenging. On one hand, the general result of closedness relies on the equivalence of norms in a finite dimensional subspace, which needs a compactness argument on a continuous function. For instance, see some Lecture Notes in undergraduate mathematics such as (Gostiaux 1993, Th 6.28 pp. 192–3). On the other hand, since we only consider subspaces of a Hilbert space, a weaker result using the inner product is proposed in (Clément and Martin 2016). But, the proof involves to study the behavior of sequences built from other sequences. And since we consider here more general topological reasonings, such proof requires to build the corresponding correct filters and to prove their properties from those of other filters, and this is getting us pretty far from the pen-and-paper proof.

As our final objective is the formal proof of a scientific computing program implementing the Finite Element Method, several additional important steps are necessary. From the mathematical standpoint, one needs to formalize large parts of integration and distribution theories to define Sobolev spaces, such as $L^2(\Omega)$, $H^1(\Omega)$ and $H^1_0(\Omega)$ for some bounded and regular domain Ω of $\mathbb{R}^d$ with $d = 1$, 2, or 3. For instance, for the Laplace equation, one needs to apply the Lax–Milgram theorem with $E = H^1_0(\Omega)$ (e.g. see (Ern and Guermond 2004)), thus one needs to prove that the latter is a Hilbert space.

Moreover, many results of interpolation and approximation theory will be required to formalize the Finite Element spaces, to define the Finite Element Method itself as an algorithm. Finally, the verification of (at least parts of) a realistic scientific computing code in C++ will need first its speci-

fication. For each function, the properties it requires and it ensures have to be formally expressed. This is a heavy work as the library is large, and the genericity is high due to a heavy use of objects. We may hope to use both abstract interpretation and deductive methods in order to lighten both the specification and the verification. As for the proofs, automation is improving; nonetheless, we will have to use interactive provers, if only to link with the mathematical theorems presented here.

Acknowledgments

We are grateful to Assia Mahboubi, Guillaume Melquiond, and Pierre Weis for fruitful discussions, mainly about canonical structures.

This work was partly supported by GT ELFIC from Labex DigiCosme.

References

S. Boldo, F. Clément, J.-C. Filliâtre, M. Mayero, G. Melquiond, and P. Weis. Wave Equation Numerical Resolution: a Comprehensive Mechanized Proof of a C Program. *Journal of Automated Reasoning*, 50(4):423–456, Apr. 2013. URL `http://hal.inria.fr/hal-00649240/en/`.

S. Boldo, C. Lelay, and G. Melquiond. Coquelicot: A user-friendly library of real analysis for coq. *Mathematics in Computer Science*, 9(1):41–62, 2015. ISSN 1661-8270. URL `http://dx.doi.org/10.1007/s11786-014-0181-1`.

P. G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. ISBN 0-89871-514-8. doi: 10.1137/1.9780898719208. URL `http://dx.doi.org/10.1137/1.9780898719208`. Reprint of the 1978 original [North-Holland, Amsterdam; MR0520174 (58 #25001)].

F. Clément and V. Martin. The Lax-Milgram Theorem. A detailed proof to be formalized in Coq. Research Report RR-8934, Inria Paris, July 2016. URL `https://hal.inria.fr/hal-01344090`.

A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004. ISBN 0-387-20574-8. doi: 10.1007/978-1-4757-4355-5. URL `http://dx.doi.org/10.1007/978-1-4757-4355-5`.

G. Gonthier, A. Asperti, J. Avigad, Y. Bertot, C. Cohen, F. Garillot, S. Le Roux, A. Mahboubi, R. O'Connor, S. Ould Biha, I. Pasca, L. Rideau, A. Solovyev, E. Tassi, and L. Théry. A Machine-Checked Proof of the Odd Order Theorem. In S. Blazy, C. Paulin, and D. Pichardie, editors, *ITP 2013, 4th Conference on Interactive Theorem Proving*, volume 7998 of *LNCS*, pages 163–179, Rennes, France, July 2013. Springer. URL `https://hal.inria.fr/hal-00816699`.

B. Gostiaux. *Cours de mathématiques spéciales - Tome 2 [Lecture notes in Special Mathematics - Tome 2]*. Mathématiques [Mathematics]. Presses Universitaires de France, Paris, 1993. ISBN 2-13-045836-X. Topologie, analyse réelle [Topology, real analysis]. In French.

J. Hölzl, F. Immler, and B. Huffman. Type classes and filters for mathematical analysis in Isabelle/HOL. 7998:279–294, 2013.

B. Huffman and O. Kunčar. *Lifting and Transfer: A Modular Design for Quotients in Isabelle/HOL*, pages 131–146. Springer International Publishing, Cham, 2013. ISBN 978-3-319-03545-1. doi: 10.1007/978-3-319-03545-1_9. URL `http://dx.doi.org/10.1007/978-3-319-03545-1_9`.

F. Immler. Formally verified computation of enclosures of solutions of ordinary differential equations. In J. M. Badger and K. Y. Rozier, editors, *NASA Formal Methods - 6th International Symposium, NFM 2014, Houston, TX, USA, April 29 - May 1, 2014. Proceedings*, volume 8430 of *Lecture Notes in Computer Science*, pages 113–127. Springer, 2014.

F. Immler and J. Hölzl. Numerical analysis of ordinary differential equations in isabelle/hol. In L. Beringer and A. P. Felty, editors, *Interactive Theorem Proving - Third International Conference, ITP 2012, Princeton, NJ, USA, August 13-15, 2012. Proceedings*, volume 7406 of *Lecture Notes in Computer Science*, pages 377–392. Springer, 2012.

F. Immler and C. Traut. The Flow of ODEs. In C. J. Blanchette and S. Merz, editors, *Proceedings of the 7th International Conference on Interactive Theorem Proving*, pages 184–199, Nancy, France, 8 2016. Springer International Publishing. ISBN 978-3-319-43144-4.

C. Lelay. *Repenser la bibliothèque réelle de Coq : vers une formalisation de l'analyse classique mieux adaptée*. Thèse de doctorat, Université Paris-Sud, June 2015a.

C. Lelay. How to express convergence for analysis in Coq. In *The 7th Coq Workshop*, Sophia Antipolis, France, June 2015b.

A. Mahboubi. The Rooster and the Butterflies. In J. Carette, D. Aspinal, C. Lange, P. Sojka, and W. Windsteiger, editors, *CICM 2013 - Conference on Intelligent Computer Mathematics - 2013*, volume 7961 of *Lecture Notes in Artificial Intelligence*, pages 1–18, Bath, United Kingdom, July 2013. Springer. URL `https://hal.inria.fr/hal-00825074`.

A. Mahboubi and E. Tassi. Canonical structures for the working coq user. In *Interactive Theorem Proving - 4th International Conference, ITP 2013, Rennes, France, July 22-26, 2013. Proceedings*, pages 19–34, 2013. doi: 10.1007/978-3-642-39634-2_5.

E. Makarov and B. Spitters. The Picard algorithm for ordinary differential equations in Coq. In S. Blazy, C. Paulin, and D. Pichardie, editors, *ITP 2013, 4th Conference on Interactive Theorem Proving*, volume 7998 of *LNCS*, pages 463–468, Rennes, France, July 2013. Springer.

O. C. Zienkiewicz, R. L. Taylor, and J. Z. Zhu. *The finite element method: its basis and fundamentals*. Elsevier/Butterworth Heinemann, Amsterdam, seventh edition, 2013. ISBN 978-1-85617-633-0. doi: 10.1016/B978-1-85617-633-0.00001-0. URL `http://dx.doi.org/10.1016/B978-1-85617-633-0.00001-0`.