
A Generalized Lax Equivalence Theorem: Consistency, Stability and Convergence

Consistency, stability and convergence are basic concepts for almost all discretization methods. These concepts are mostly commonly encountered in numerical methods for partial differential equations. In the finite difference methods, the best known result is the theory of P. Lax [?] on the equivalence of stability and convergence for certain class of consistent finite difference schemes. In the discretization methods based on variational principle such as Petrov-Galerkin methods (including finite element and finite volume methods), there are the fundamental theories by Babuška [?] for the standard Galerkin method and Brezzi [?, ?] for the mixed Galerkin method.

Despite of these existing fundamental theories, the concepts of consistency, stability and convergence and their relationship are not always transparent in different applications. One notable example is that for certain finite element and finite volume schemes on some special grids, the resulting finite difference scheme is inconsistent (in the classic sense) at all the grid points, but nevertheless, these methods are known to be optimally convergent. It was somewhat a surprise that an inconsistent scheme would actually converge! This phenomenon was then known as “supraconvergence” and there was a sizeable literature devoted to this subject.

The surprise was rooted to the different understanding of the concept of consistency. We will show that a convergent scheme has to be consistent, namely consistency is necessary for convergence. A more nontrivial part of the Lax theorem is that stability is necessary for convergence for a consistent scheme. As a result, stability analysis becomes a central issue in finite difference theory. Von Neumann analysis is a general technique.

The so-called *Lax Equivalence Theorem* states that a given consistent (finite difference) scheme (for partial differential equations) is convergent if and only if it is stable. This is perhaps one of the most basic and also most popular theorem in numerical methods for partial differential equations. In some way, this is also the most mis-conceived theorem. The mis-conception of this theorem mainly lie in two categories. The original Lax theorem was stated for a specific class of finite difference method for a specific class of initial boundary value problems. But this theorem has been quoted in many different contexts which are much more general than the original result. One part in the Lax theorem, namely stability implies convergence, is most useful in practical applications, but its proof is most straightforward and trivial. The other part, namely convergence actually also implies stability, is much more nontrivial in mathematical theory but much less useful in practical applications. This part of the theorem is actually often misleading from a practical point of view. This part of the theorem literally says that a unstable method can not be convergent, and, naturally, can not be used. But this is a misconception on the theorem.

The importance of stability, however, may be easily over-emphasized. The necessity of the stability in the Lax theorem is only valid when the convergence is considered for the most general data for which the underlying partial differential equations are well-posed. In practice, the worst scenario rarely occur, a scheme that is unstable in a strict sense may still lead to a reasonable (though not optimal) convergence in many practical situations. We will give a brief discussion on this subject in this chapter.

The motivation is to set-up a right framework so that the following statement is valid

- **consistency + stability \Leftrightarrow convergence.**

for a wide range of discretization methods including

1. numerical quadrature
2. finite difference method
3. finite volume method
4. finite element method
5. discontinuous Galerkin method

9.1 Consistency, stability, and convergence

9.1.1 Continuous and discrete problems

We assume that V (the space of solutions) and F (the space of data) are two normed Banach spaces equipped with the norms $\|\cdot\|_V$ and $\|\cdot\|_F$ respectively. We consider a linear operator $\mathcal{L} : V \mapsto F$.

We consider the equation

$$(9.1) \quad \mathcal{L}u = f.$$

We assume that (9.1) is well-posed, namely \mathcal{L} is an isomorphism from V to F . The well-posedness of (9.1) implies that there exists a constant c_L such that

$$\|u\|_V \leq c_L \|f\|_F, \quad \text{or} \quad \|\mathcal{L}^{-1}f\|_V \leq c_L \|f\|_F \quad f \in F.$$

Let now V_h and F_h are two “discrete” vector spaces with distance functions given by $\|\cdot\|_{V_h}$ and $\|\cdot\|_{F_h}$ respectively. Consider a family of discrete problems

$$(9.2) \quad \mathcal{L}_h u_h = R_h f.$$

where $\mathcal{L}_h : V_h \mapsto F_h$ is an isomorphism under the norms $\|\cdot\|_{V_h}$ and $\|\cdot\|_{F_h}$. The space F_h is related to F by a restriction operator: $R_h : F \mapsto F_h$. In applications, the space V_h or F_h is often finite dimensional space but in theory the finite dimensionality is not necessary. V_h (or F_h) may or may not be a subspace V (or F).

The discrete spaces V_h and F_h should be related to the original space V and F in some specific way. Let us first discuss how V_h is connected with V . We make the following assumptions.

Assumption 9.3 *There is an extended distance functions $\|\cdot\|_h$ that is well defined on both $V + V_h$ satisfying the following two properties*

$$(9.4) \quad \|v_h\|_h = \|v_h\|_{V_h}, \quad \forall v_h \in V_h,$$

and

$$(9.5) \quad c_V^{-1} \|v\|_V \leq \|v\|_h \leq c_V \|v\|_V, \quad \forall v \in V.$$

Implicitly, we can imagine there is a larger space, say \bar{V} , such that $V + V_h \subset \bar{V}$. But we do not need to use this space explicitly, only its norm.

Assumption 9.6 R_h is assumed to satisfy:

1. *Surjective and uniformly bounded*

$$(9.7) \quad \|R_h g\|_{F_h} \leq c_R \|g\|_F \quad g \in F.$$

2. *Uniformly bounded right inverse. Namely, for any $f_h \in F_h$, there exists $f := R_h^\dagger f_h \in F$ such that*

$$(9.8) \quad R_h f = f_h, \quad \|f\|_F \leq c_F \|f_h\|_{F_h}.$$

It is known that a surjective continuous linear map R_h has a right inverse iff $\ker(R_h)$ is complemented. We note that c_R and c_F are independent of h .

9.1.2 Consistency, stability and convergence

Consistency, stability and convergence are the three basic concepts of a discrete method. For any given h , we introduce

1. Error: $\|u - u_h\|_h$;
2. Stability constant

$$K_h := \|\mathcal{L}_h^{-1}\|_{\mathcal{L}(F_h, V_h)} = \sup_{f_h \in F_h} \frac{\|\mathcal{L}_h^{-1} f_h\|_{V_h}}{\|f_h\|_{F_h}};$$

3. Consistency error

$$E_h^c(u) := \inf_{v_h \in V_h} (\|u - v_h\|_h + \|\mathcal{L}_h v_h - R_h \mathcal{L} u\|_{F_h}).$$

Definition 10. 1. We say that the discrete method is convergent if

$$(9.9) \quad \lim_{h \rightarrow 0} \|u - u_h\|_h = 0 \quad \forall u \in \mathcal{L}^{-1}(F).$$

2. We say that the discretization (9.2) is uniformly stable if

$$(9.10) \quad K \equiv \sup_{h>0} K_h < \infty.$$

3. We say that the discretization (9.2) is consistent at u if

$$(9.11) \quad \lim_{h \rightarrow 0} E_h^c(u) = 0.$$

If further (9.11) holds for any $u \in \mathcal{L}^{-1}(F)$, then the discretization is consistent. ^{c0}

Lemma 64.

$$(9.12) \quad E_h^c(u) \leq \text{err}_h(u) \leq \max(1, K_h) E_h^c(u).$$

Proof. By taking $v_h = u_h$ in the definition of $E_h^c(u)$, we see immediately that $E_h^c(u) \leq \|u - u_h\|_h$. Now we write $u - u_h = u - v_h + v_h - \mathcal{L}_h^{-1} R_h \mathcal{L} u$ and obtain by triangle inequality that

$$\begin{aligned} \|u - u_h\|_h &\leq \|u - v_h\|_h + \|v_h - \mathcal{L}_h^{-1} R_h \mathcal{L} u\|_h \\ &= \|u - v_h\|_h + \|\mathcal{L}_h^{-1} \mathcal{L}_h v_h - \mathcal{L}_h^{-1} R_h \mathcal{L} u\|_{V_h} \\ &\leq \|u - v_h\|_h + K_h \|\mathcal{L}_h v_h - R_h \mathcal{L} u\|_{F_h} \\ &\leq \max(1, K_h) E_h^c(u). \end{aligned}$$

This completes the proof. \square

The well-known Lax Theorem states that equivalence between consistent and convergence under the stability. We hope to obtain the stability from the convergence. To this end, we need one more assumption besides (9.8).

Assumption 9.13 There exists an operator $\Pi_h^c : V_h \mapsto V$ such that

$$(9.14) \quad c_D^{-1} \|v_h\|_h \leq \|\Pi_h^c v_h\|_h \leq c_D \|v_h\|_h \quad \forall v_h \in V_h.$$

^{c0} According to our definition, when we verify a scheme is consistent, we can not just verify for any u , but rather we need to verify those special $u = \mathcal{L}^{-1} f$ with $f \in F$.

If V_h is a subspace of V , the I_h^c is the trivial inclusion operator and $c_I = 1$. For the nonconforming methods, the above assumption relates to the conforming relatives [?, ?] in the finite element methods. We are now in the position to state the generalized Lax Theorem.

Theorem 57. *The following statements hold:*

1. A discretization that is convergent must be consistent.
2. A discretization that is consistent and stable must be convergent.
3. If the restriction operator R_h has a uniformly bounded right inverse and Assumption 9.13 holds, then convergence implies stability.

Proof. The first and second statements follow directly from Lemma 64. To prove the third statement, if the scheme is convergent, for any $f \in F$, there exists a constant $c(f)$ such that

$$\|\mathcal{L}^{-1}f - \mathcal{L}_h^{-1}R_h f\|_h \leq c(f) \quad \forall h.$$

Thus

$$\|\mathcal{L}_h^{-1}R_h f\|_h \leq c(f) + \|\mathcal{L}^{-1}f\|_h \leq c(f) + c_V \|\mathcal{L}^{-1}f\|_V := \tilde{c}(f) \quad \forall h.$$

By the Assumption 9.13,

$$\tilde{c}(f) \geq \|\mathcal{L}_h^{-1}R_h f\|_h \geq c_I^{-1} \|I_h^c \mathcal{L}_h^{-1}R_h f\|_h \geq c_I^{-1} c_V^{-1} \|I_h^c \mathcal{L}_h^{-1}R_h f\|_V.$$

which implies that

$$\|I_h^c \mathcal{L}_h^{-1}R_h f\|_V \leq c_I c_V \tilde{c}(f) \quad \forall h.$$

By the uniform boundedness principle, we have

$$\|I_h^c \mathcal{L}_h^{-1}R_h f\|_V \leq C_1 \|f\|_F \quad \forall f \in F.$$

For any $f_h \in F_h$, by the Assumption 9.6, there exists $f = R_h^\dagger f_h$ such that $R_h f = f_h$ and $\|R_h^\dagger f_h\|_F \leq c_F \|f_h\|_{F_h}$. Then,

$$\begin{aligned} \|I_h^c \mathcal{L}_h^{-1} f_h\|_h &\leq c_V \|I_h^c \mathcal{L}_h^{-1} f_h\|_V = c_V \|I_h^c \mathcal{L}_h^{-1} R_h f\|_V \\ &\leq c_V C_1 \|f\|_F = c_V C_1 \|R_h^\dagger f_h\|_F \leq c_V c_F C_1 \|f_h\|_{F_h}. \end{aligned}$$

along with the Assumption 9.13, we obtain

$$\|\mathcal{L}_h^{-1} f_h\|_h \leq c_I \|I_h^c \mathcal{L}_h^{-1} f_h\|_h \leq c_I c_V c_F C_1 \|f_h\|_{F_h}.$$

which implies the stability. \square

Remark 15. Comparing what are in the literature, one major difference in the theorem is that we made an additional assumption that the original problem (9.1) is also stable.

Let us now attempt to write some general theorem on the necessity of stability for convergency. We first start with some special cases.

9.2 An example — 2nd order elliptic boundary value problems

We first start with some special cases. Let us look at a simple example of 2nd order elliptic boundary value problems on a bounded planar domain Ω :

$$(9.15) \quad -\Delta u = f, x \in \Omega, \quad u = 0, x \in \partial\Omega.$$

We assume that Ω is a polygonal domain in \mathbb{R}^n with $1 \leq n \leq 3$. Let us try to identify spaces V and F in two different ways.

9.2.1 Stability with respect to different pairs of spaces

The most natural setting is that $-\mathcal{A} : H_0^1(\Omega) \mapsto H^{-1}(\Omega)$ is an isomorphism:

$$(9.16) \quad \|u\|_1 \lesssim \|f\|_{-1}.$$

We may also use other pairs of spaces between which $-\mathcal{A}$ is an isomorphism, but none of these pairs of “ \mathcal{A} -isomorphic” spaces are appropriate for finite difference methods. Instead, we can use the following two types of stability results for finite difference method analysis. The first stability is rooted to the well-known the maximum principle:

$$(9.17) \quad \|u\|_{C(\bar{\Omega})} \lesssim \|f\|_{C(\bar{\Omega})}.$$

This is the basic stability that underlies the classic convergence analysis for finite difference methods. But obviously, $-\mathcal{A}$ does not map $C(\bar{\Omega})$ to $C(\bar{\Omega})$.

We note that the stability property for the continuous problem is crucial for Lax type of theory whereas the continuous isomorphism property is less relevant.

Choice A

: $V = H_0^1(\Omega)$ and $F = H^{-1}(\Omega)$. In this setting, the Poisson equation (9.15) can be cast into a variational formulation: Given $f \in H^{-1}(\Omega)$, find $u \in H_0^1(\Omega)$ such that

$$(9.18) \quad (\nabla u, \nabla v) = \langle f, v \rangle, \quad v \in H_0^1(\Omega).$$

In this case, $\mathcal{L} = -\mathcal{A} : V \mapsto V$ is an isomorphism (in the weak sense, or the distributional sense), and

$$\|u\|_{H_0^1(\Omega)} \lesssim \|f\|_{H^{-1}(\Omega)}.$$

Choice B

: $V = C(\bar{\Omega})$ and $F = C(\bar{\Omega})$. In this setting, we use classic theory of PDE to conclude that for any $f \in C(\bar{\Omega})$, there exists a unique $u \in C(\bar{\Omega})$. Unlike the Choice A, the operator $\mathcal{L} = -\mathcal{A}$ does not map $C(\bar{\Omega})$ to $C(\bar{\Omega})$ in any reasonable sense. But its inverse $\mathcal{L}^{-1} : C(\bar{\Omega}) \mapsto C(\bar{\Omega})$ is a well-defined and bounded operator:

$$\|(-\mathcal{A})^{-1}f\|_{C(\bar{\Omega})} \lesssim \|f\|_{C(\bar{\Omega})}, \quad \text{or} \quad \|u\|_{C(\bar{\Omega})} \lesssim \|f\|_{C(\bar{\Omega})}.$$

9.2.2 Conforming finite element method

In the finite element setting, we choose $V = H_0^1(\Omega)$ and $F = H^{-1}(\Omega) = V'$. For the discretization, the choice of V_h is natural, namely the finite element space $V_h \subset H_0^1(\Omega)$ with the same H^1 -norm. The assumption (9.5) and (9.4) are trivially satisfied with (9.5) being equality with $c_V = 1$.

The choice of F_h is slightly less obvious. But we choose $F_h = V_h'$ with the norm given by the discrete H^{-1} norm:

$$\|f_h\|_{-1,h} := \sup_{v_h \in V_h} \frac{(f_h, v_h)}{\|v_h\|_1}.$$

Now the discrete operator $\mathcal{L}_h : V_h \mapsto F_h$ is defined by

$$\langle \mathcal{L}_h u_h, v_h \rangle = (\nabla u_h, \nabla v_h) \quad \forall v_h \in V_h.$$

Let $i_h : V_h \mapsto V$ is the inclusion. Then, $R_h := i_h' : V' \mapsto V_h'$ satisfies

$$\langle R_h f, v_h \rangle = \langle f, i_h v_h \rangle = \langle f, v_h \rangle \quad \forall v_h \in V_h.$$

For any linear functional f_h on $V_h \subset V$, define $p(v_h) := \|f_h\|_{F_h} \|v_h\|_V$, then $\langle f_h, v_h \rangle \leq p(v_h)$ for all $v_h \in V_h$. By Hahn-Banach theorem, there exist a linear functional $f \in V'$ such that

1. $\langle f_h, v_h \rangle = \langle f, v_h \rangle$, for all $v_h \in V_h$
2. $|\langle f, v \rangle| \leq \|f_h\|_{F_h} \|v\|_V$, for all $v \in V$.

which gives (9.7) by $R_h^\dagger f_h := f$ with $c_R = 1$.

We first analyze the stability of \mathcal{L}_h^{-1} . The operator \mathcal{L}_h is a bijection from V_h to F_h . And by Poincaré inequality:

$$\|u_h\|_{1,\Omega} \leq C \|\nabla u_h\|_{0,\Omega} = C \sup_{v_h \in V_h} \frac{(\nabla u_h, \nabla v_h)}{\|\nabla v_h\|} = C \sup_{v_h \in V_h} \frac{(\mathcal{L}_h u_h, v_h)}{\|\nabla v_h\|} \leq C \sup_{v_h \in V_h} \frac{(\mathcal{L}_h u_h, v_h)}{\|v_h\|_V} = C \|\mathcal{L}_h u_h\|_{-1,h},$$

which implies

$$\|\mathcal{L}_h^{-1} f_h\|_V \leq C \|f_h\|_{-1,h}.$$

Next we consider the consistency

$$E_h^c(u) = \inf_{v_h \in V_h} \|u - v_h\|_1 + \|\mathcal{L}_h v_h - R_h f\|_{-1,h}.$$

By the definition of $\|\cdot\|_{-1,h}$, we have

$$\begin{aligned} \|\mathcal{L}_h v_h - R_h f\|_{-1,h} &= \sup_{w_h \in V_h} \frac{(\mathcal{L}_h v_h, w_h) - (R_h f, w_h)}{\|w_h\|_1} \\ &= \sup_{w_h \in V_h} \frac{(\nabla v_h, \nabla w_h) - (f, w_h)}{\|w_h\|_1} \\ &= \sup_{w_h \in V_h} \frac{(\nabla v_h - \nabla u, \nabla w_h)}{\|w_h\|_1} \leq \|u - v_h\|_1. \end{aligned}$$

Therefore, the consistency is equivalent to the approximability, i.e.

$$\inf_{v_h \in V_h} \|u - v_h\|_1 \leq E_h^c(u) \lesssim \inf_{v_h \in V_h} \|u - v_h\|_1.$$

9.2.3 Finite difference method

Let us assume that Ω is the unit square in the plane. We consider the finite difference method (using 5-point) stencil for the Laplacian:

$$\mathcal{L}_h u_h = R_h f,$$

and

$$\mathcal{L}_h u_h(x_{ij}) := \frac{4u_{ij} - (u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1})}{h^2}, \quad R_h f(x_{ij}) := f_h(x_{ij}) = f(x_{ij}).$$

The function spaces $V = F = C^0(\Omega)$ with the $\|\cdot\|_{C(\Omega)}$ norm, and $V_h = F_h = \mathbb{R}^N$ with the $\|\cdot\|_{l^\infty}$ norm.

Lemma 65. *If $L_h u_h \leq 0$ on Ω_h , then $\max_{\Omega_h} u_h \leq \max_{\Gamma_h} u_h$. Furthermore, $\max_{\Omega_h} u_h = \max_{\Gamma_h} u_h$ if and only if u_h is a constant on $\Omega_h \cup \Gamma_h$.*

Proof. Suppose $\max_{\Omega_h} u_h > \max_{\Gamma_h} u_h$. Then

$$4u_h(x_0) = 4 \max_{\Omega_h \cup \Gamma_h} u_h = h^2 L_h u_h(x_0) + \sum_{i=1}^4 u_h(x_i) \leq \sum_{i=1}^4 u_h(x_i),$$

where $u_h(x_i)$, $i = 1, 2, 3, 4$, are four nearest neighbors. This implies $u_h(x_0) = u_h(x_i)$, $i = 1, 2, 3, 4$. And run this argument through the domain, we get u_h is a constant on $\Omega_h \cup \Gamma_h$, which is a contradiction. \square

Theorem 58. For $u_h \in V_h$ solves the difference equation, there holds

$$\|u_h\|_{l^\infty} \lesssim \|\mathcal{L}_h u_h\|_{l^\infty}.$$

Proof. Introduce an auxiliary function $g(x, y) = \frac{(x-\frac{1}{2})^2 + (y-\frac{1}{2})^2}{4}$, and $\mathcal{L}_h g = -1$. Then we have

$$\mathcal{L}_h(u_h + \|\mathcal{L}_h u_h\|_{l^\infty} g) = \mathcal{L}_h u_h + \|\mathcal{L}_h u_h\|_{l^\infty} \mathcal{L}_h g = \mathcal{L}_h u_h - \|\mathcal{L}_h u_h\|_{l^\infty} \leq 0.$$

By maximum theorem,

$$\max_{\bar{\Omega}_h} u_h \leq \max_{\bar{\Omega}_h} (u_h + \|\mathcal{L}_h u_h\|_{l^\infty} g) = \max_{\Gamma_h} (u_h + \|\mathcal{L}_h u_h\|_{l^\infty} g) \leq \max_{\Gamma_h} u_h + \|\mathcal{L}_h u_h\|_{l^\infty} \max_{\Gamma_h} g \leq C \|\mathcal{L}_h u_h\|_{l^\infty}.$$

To apply the similar argument to $-u_h$, we can prove the theorem. \square

Theorem 58 means \mathcal{L}_h^{-1} is uniform bounded as the operator from F_h to V_h .

Theorem 59. Suppose $u \in C^4(\bar{\Omega})$, then there holds

$$\|u - u_h\|_{l^\infty} \leq Ch^2.$$

Proof. By

$$\mathcal{L}_h(u - u_h)(x_i) = \mathcal{L}_h u(x_i) - \mathcal{L}u(x_i) + \mathcal{L}u(x_i) - \mathcal{L}_h u_h(x_i) = (\mathcal{L}_h - \mathcal{L})u(x_i),$$

Taylor expansion and above theorem, we have

$$\|u - u_h\|_{l^\infty} \leq C \|\mathcal{L}_h u - \mathcal{L}u\|_{l^\infty} \leq Ch^2.$$

\square

9.2.4 Nonconforming finite element method

We still consider $-A : H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$. The finite element space V_h is defined on the partition \mathcal{T}_h with the norm $\|\cdot\|_{1,h}$. The $\|\cdot\|_{1,h}$ is the piecewise H^1 norm. Define the bilinear form on $V_h \times V_h$:

$$a_h(w_h, v_h) = \sum_{T \in \mathcal{T}_h} \int_T \nabla w_h \cdot \nabla v_h,$$

By Necas inequality, any $f \in H^{-1}(\Omega)$ can be represented by $f = f_0 - \sum_i \partial_i f_i$. For any $f \in H^{-1}(\Omega)$, we define the ‘‘action’’ on v_h by

$$\langle f, v_h \rangle_h := \sum_{T \in \mathcal{T}_h} \sum_{|\alpha| \leq 1} (f_\alpha, \partial^\alpha v_h)_T.$$

The norm $\|\cdot\|_{F_h}$ can be firstly defined on $H^{-1}(\Omega)$:

$$\|f\|_{F_h} = \|f\|_{-1,h} := \sup_{v_h \in V_h} \frac{\langle f, v_h \rangle_h}{\|v_h\|_{1,h}}.$$

Then, the space F_h is the completion of $H^{-1}(\Omega)$ under the norm $\|\cdot\|_{F_h}$. By the coercivity of a_h and Riesz representation theorem, F_h is isomorphism to V_h and we can define

$$\langle \mathcal{L}_h w_h, v_h \rangle_h := a_h(w_h, v_h) \quad \forall v_h \in V_h.$$

We first show the stability of \mathcal{L}_h^{-1} . By the coercivity of $a_h(\cdot, \cdot)$,

$$\|\mathcal{L}_h u_h\|_{-1,h} \|u_h\|_{1,h} \geq \langle \mathcal{L}_h u_h, u_h \rangle = a_h(u_h, u_h) \geq \alpha_0 \|u_h\|_{1,h}^2,$$

The uniformly stability is obtained: $K = \alpha_0^{-1}$. From Theorem 57, convergence and consistency is equivalent for the nonconforming finite element methods.

We then study the consistency. Recall that

$$E_h^c(u) = \inf_{w_h \in V_h} (\|u - w_h\|_{1,h} + \|\mathcal{L}_h w_h - R_h \mathcal{L}u\|_{-1,h}).$$

And

$$\begin{aligned} \|\mathcal{L}_h w_h - R_h \mathcal{L}u\|_{-1,h} &= \sup_{v_h \in V_h} \frac{\langle \mathcal{L}_h w_h, v_h \rangle_h - \langle R_h \mathcal{L}u, v_h \rangle_h}{\|v_h\|_{1,h}} \\ &= \sup_{v_h \in V_h} \frac{a_h(w_h, v_h) - \langle f, v_h \rangle_h}{\|v_h\|_{1,h}} \\ &\leq \|u - w_h\|_{1,h} + \sup_{v_h \in V_h} \frac{a_h(u, v_h) - \langle f, v_h \rangle_h}{\|v_h\|_{1,h}} \end{aligned}$$

Thus,

$$(9.19) \quad \|u - u_h\|_{1,h} \cong \inf_{w_h \in V_h} \|u - w_h\|_{1,h} + \sup_{v_h \in V_h} \frac{a_h(u, v_h) - \langle f, v_h \rangle_h}{\|v_h\|_{1,h}},$$

which is the Strang Lemma.

Verify Assumption 9.13 by conforming relatives

Let V_h^c be the \mathcal{P}_2 Lagrange space. Consider the Crouzeix-Raviart element, we will construct the operator $\Pi_h^c : V_h \mapsto V_h^c$ (called *enriching operator* in literature). Π_h^c is defined by

$$N(\Pi_h^c v) = \frac{1}{|\mathcal{T}_p|} \sum_{T \in \mathcal{T}_p} N(v_T) \quad \forall v \in V_h,$$

where p is any nodal point for V_h^c , N is any nodal variable at p , and \mathcal{T}_p is the set of triangles in \mathcal{T}_h whose closure share the nodal point p .

1. The stability of Π_h^c follows from the following argument. The standard scaling argument gives

$$\begin{aligned} \|v - \Pi_h^c v\|_{L^2(T)}^2 &\approx \sum_N h_T^2 (N(v - \Pi_h^c v))^2 \\ &= \sum_{\text{vertex dof } N} h_T^2 (N(v - \Pi_h^c v))^2 \\ &\lesssim \sum_{e \ni a, a \text{ is vertex of } T} h_T^2 \max[v]^2 \\ &\lesssim \sum_{e \ni a, a \text{ is vertex of } T} h_T^3 |v|_{1,e}^2 \\ &\lesssim h_T^2 |v|_{1,\omega_T}^2. \end{aligned}$$

Then,

$$\|v - \Pi_h^c v\|_{L^2}^2 \lesssim h^2 |v|_{1,h}^2.$$

By inverse inequality,

$$\|\Pi_h^c v\|_{1,h}^2 \lesssim h^{-2} \|v - \Pi_h^c v\|_{L^2}^2 + \|v\|_{1,h}^2 \lesssim |v|_{1,h}^2.$$

2. Now we try to prove $\|v\|_{1,h} \lesssim \|I_h^c v\|_{1,h}$ for all $v \in V_h$. Since the d.o.f of Crouzeix-Raviart is contained in \mathcal{P}_2 Lagrange, it is straightforward that I_h^c is one-to-one. The standard scaling argument gives the lower bound of I_h^c .

9.3 Petrov-Galerkin methods based on variational formulation

In this section, we apply the generalized Lax Theorem 57 to the Petrov-Galerkin variation formulation. We assume that V and Q are Hilbert spaces, and $F = Q'$ is the dual space of Q .

The equation (9.1) is given from a variational problem: Find $u \in V$, such that

$$(9.20) \quad a(u, q) = \langle f, q \rangle \quad \forall q \in Q.$$

This equation is equivalent to (9.1) if we define

$$\langle \mathcal{L}u, q \rangle = a(u, q) \quad u \in V, q \in Q.$$

The well-posedness of (9.1) is guaranteed by the following conditions:

1. Boundedness of $a(\cdot, \cdot)$:

$$(9.21) \quad a(u, q) \leq M \|u\|_V \|q\|_Q \quad \forall u \in V, q \in Q.$$

2. Inf-sup condition

$$(9.22) \quad \inf_{u \in V} \sup_{q \in Q} \frac{a(u, q)}{\|u\|_V \|q\|_Q} = \inf_{q \in Q} \sup_{u \in V} \frac{a(u, q)}{\|u\|_V \|q\|_Q} = \beta > 0.$$

Given the discrete spaces V_h and Q_h , a general Petrov-Galerkin method can be defined as: Find $u_h \in V_h$, such that

$$(9.23) \quad a_h(u_h, q_h) = \langle f, q_h \rangle_h \quad \forall q_h \in Q_h.$$

Here $\langle f, q_h \rangle_h$ represents certain approximate evaluation of $\langle f, q_h \rangle$, such as numerical quadrature. The solution u_h of this problem is often known as the Galerkin (or Petrov-Galerkin) approximation of u . Usually in applications V_h and Q_h are finite dimensional and the subscript h is related to certain discretization parameters (such as grid size and polynomial degree).

We introduce the following operator $\mathcal{L}_h : V_h \mapsto Q_h'$:

$$\langle \mathcal{L}_h u_h, q_h \rangle_h = a_h(u_h, q_h), \quad u_h \in V_h, q_h \in Q_h.$$

The restriction operator $R_h : F \mapsto F_h$ is given by

$$\langle R_h f, q_h \rangle_h = \langle f, q_h \rangle_h.$$

According to (9.22), we have that the problem (9.23) is uniquely solvable if and only if the following conditions hold:

$$(9.24) \quad \inf_{u_h \in V_h} \sup_{q_h \in Q_h} \frac{a_h(u_h, q_h)}{\|u_h\|_V \|q_h\|_Q} = \inf_{q_h \in Q_h} \sup_{u_h \in V_h} \frac{a_h(u_h, q_h)}{\|u_h\|_V \|q_h\|_Q} = \beta_h > 0.$$

9.3.1 Babuska-Brezzi theory

Let us review the theory of Babuska and Brezzi. A Petrov-Galerkin method (9.23) is called variationally exact if $V_h \subset V$, $Q_h \subset Q$, and (9.23) turns out to be

$$(9.25) \quad a(u_h, q_h) = \langle f, q_h \rangle \quad \forall q_h \in Q_h.$$

A fundamental result for Galerkin approximation (by Babuska-Brezzi) is stated as follows.

Theorem 60. *If the discrete variational problem (9.25) is variationally exact and well-posed, then*

$$\|u - u_h\|_V \leq \left(\frac{M}{\beta_h} + 1 \right) \inf_{v_h \in V_h} \|u - v_h\|_V.$$

Proof. We define $P_h : V \mapsto V_h$ s.t.

$$a(P_h u, q_h) = a(u, q_h) \quad \forall q_h \in Q_h.$$

Trivially, $P_h^2 = P_h$. By the inf-sup condition (12.26),

$$\beta_h \|P_h u\|_V \leq \sup_{q_h \in Q_h} \frac{a(P_h u, q_h)}{\|q_h\|_Q} = \sup_{q_h \in Q_h} \frac{a(u, q_h)}{\|q_h\|_Q} \leq M \|u\|_V \quad \forall u \in V.$$

Therefore, $\|P_h\| \leq M/\beta_h$.

Finally, we have for any $v_h \in V_h$,

$$\begin{aligned} \|u - u_h\|_V &= \|u - P_h u\|_V \\ &= \|(I - P_h)(u - v_h)\|_V \leq \|I - P_h\| \|u - v_h\|_V \\ &\leq \left(1 + \frac{M}{\beta_h} \right) \|u - v_h\|_V. \end{aligned}$$

This finishes our proof. \square

Remark 16. In case $\beta_h > \beta_0 > 0$, then the constant $1 + M/\beta_h$ is bounded above independent of h . This corresponds to the case of quasi-optimal approximation.

Remark 17. In case that V_h is Hilbert space and $\{0\} \subsetneq V_h \subsetneq V$, then $\|I - P_h\| = \|P_h\|$ (Xu and Zikatanov [?]). We have the improved result:

$$(9.26) \quad \|u - u_h\|_V \leq \frac{M}{\beta_h} \inf_{v_h \in V_h} \|u - v_h\|_V.$$

Lax theory further implies that if the Petrov-Galerkin method converges for all $f \in V'$, then the discrete inf-sup condition should be satisfied uniformly.

We note that, in our derivation, we have not assumed that the original problem is well-posed. Based on the above theory, if the Petrov-Galerkin method converges for any $f \in V'$, then the original problem has to be well-posed, namely \mathcal{L}^{-1} is bounded. We also did not assume that the bilinear form is bounded.

9.3.2 Applying Lax Theorem to Petro-Galerkin variational problems

In order to apply the generalized Lax Theorem 57, we first verify the Assumption (9.8) for the variationally exact case.

Lemma 66. *For the variationally exact case, $\{R_h\}$ has uniformly bounded right inverse.*

Proof. Since Q_h is closed subspace of Q , we have

$$Q = Q_h \oplus Q_h^\perp.$$

Define $R_h^\dagger : Q_h \mapsto Q$ such that $\langle R_h^\dagger f_h, q \rangle = \langle f_h, q \rangle$ if $q \in Q_h$, and $\langle R_h^\dagger f_h, q \rangle = 0$ if $q \in Q_h^\perp$. It is straightforward that $\|R_h^\dagger f_h\|_F = \|f_h\|_{F_h}$. In fact, for any $q \in Q$, we have $q = q_1 + q_2$ and $q_1 \in Q_h, q_2 \in Q_h^\perp$. Then,

$$\|R_h^\dagger f_h\|_F = \sup_{q \in Q} \frac{\langle R_h^\dagger f_h, q \rangle}{\|q\|_Q} = \sup_{q \in Q} \frac{\langle f_h, q_1 \rangle}{\|q\|_Q} = \sup_{q_1 \in Q_h} \frac{\langle f_h, q_1 \rangle}{\|q_1\|_Q} = \|f_h\|_{F_h}.$$

This completes the proof. \square

We have the following theorem when applying Theorem 57 to Petrov-Galerkin problem (9.25).

Theorem 61. *Let (9.25) be a discretization of (9.20). Then,*

1. *If the discrete inf-sup conditions (12.26) hold uniformly, namely $\beta_h > \beta_0 > 0$ and discretization is consistent, then the discretization is convergent, and*

$$\|u - u_h\|_V \leq \frac{M}{\beta_h} \inf_{v_h \in V_h} \|u - v_h\|_V.$$

2. *If the discretization is convergent, then it must be consistent, namely $\inf_{v_h \in V_h} \|u - v_h\|_V \rightarrow 0$ as $h \rightarrow 0$; and the discrete inf-sup conditions (12.26) hold uniformly.*

General Petrov-Galerkin variational formulation

We go back to the general Petrov-Galerkin formulation (9.23). We introduce a linear operator $P_h : V + V_h \mapsto V_h$ such that

$$a_h(P_h u, q_h) = a_h(u, q_h) \quad \forall q_h \in Q_h.$$

Theorem 62. *We have the following results:*

1. *A Petrov-Galerkin method is consistent if and only if*

$$(9.27) \quad \inf_{w_h \in V_h} \left(\|u - w_h\|_{V_h} + \sup_{v_h \in V_h} \frac{a_h(w_h, v_h) - \langle f, v_h \rangle_h}{\|v_h\|_{V_h}} \right) \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

2. *We have the estimate*

$$(9.28) \quad \|u - u_h\|_{V_h} \leq \inf_{w_h \in V_h} \left(\|u - w_h\|_{V_h} + \beta_0^{-1} \sup_{v_h \in V_h} \frac{a_h(w_h, v_h) - \langle f, v_h \rangle_h}{\|v_h\|_{V_h}} \right).$$

Variationally exact Petrov-Galerkin methods

A Petrov-Galerkin method (9.23) is called variationally exact if

$$(9.29) \quad a_h(u, q_h) = \langle f, q_h \rangle_h, \quad \langle f_h, q_h \rangle_h = \langle f, q_h \rangle_h, \quad q_h \in Q_h.$$

For a variationally exact Petrov-Galerkin method, we have the following relation:

Lemma 67. *A bounded and variationally exact Petro-Galerkin is consistent.*

Lemma 68. *If V_h is a Hilbert space. Then a bounded and variationally exact Galerkin method satisfies*

$$(9.30) \quad \|u - u_h\|_{V_h} \leq \frac{M_h}{\beta_h} \inf_{w_h \in V_h} \|u - w_h\|_{V_h}.$$

Proof. In this case, the P_h operator is idempotent, namely $P_h^2 = P_h$ and as a result

$$\|I - P_h\|_{U_h} = \|P_h\|_{U_h}.$$

This completes the proof. \square

9.4 Finite element method: consistency and superconvergence

Consider a linear finite element method of the Poisson equation on the uniform criss-cross grid on unit square. This scheme is equivalent to a finite difference scheme with its finite difference stencil at grid point i given by

$$(\nabla u_h, \nabla \phi_i) = (f, \phi_i).$$

This scheme is known to be inconsistent in the classic sense, namely

$$(\nabla u_I, \nabla \phi_i) - (f, \phi_i) \neq o(h^2)$$

in general.

But it is well-known that this finite element or finite difference scheme is perfectly convergent. From the classic convergency theory of finite difference method, it is perhaps a little bit surprising that an inconsistent finite difference scheme is actually convergent. This phenomenon has been known as *supraconvergence* in the literature, especially in the context of finite volume method (which is a Petro-Galerkin method).

But judging within a variational framework in which the method is actually defined, this method is perfectly consistent. The convergency of the method is natural, not really “supra”, since the scheme is obviously stable.

It is interesting to note that if the variational exact is actually consistent in the classic sense, some really supra-convergence phenomenon may indeed occur.

Theorem 63. *Consider a linear finite element discretization for the Poisson equation on a quasi-uniform grid on a convex polygonal domain. If the resulting finite difference scheme is consistent in the classic sense, then the finite element method has a superconvergence property. More specifically, if*

$$\frac{1}{h^2}((\nabla u_I, \nabla \phi_i) - (f, \phi_i)) = O(h^\delta), \delta > 0,$$

then

$$\|\nabla(u_h - u_I)\| = O(h^{1+\sigma/2}).$$

Then, with $e_h = u_h - u_I$, we have, for any $v_h \in V_h$

$$\begin{aligned} a(e_h, v_h) &= \sum_i v_h(x_i) a(e_h, \phi_i) \\ &= \sum_i v_h(x_i) ((f, \phi_i) - a(u_I, \phi_i)) \\ &= \sum_i v_h(x_i) O(h^4) \\ &= O(h^\delta) \sum_i h^2 |v_h(x_i)| \\ &= O(h^\delta) \|v_h\|_{L^1(\Omega)}. \end{aligned}$$

This is related to some well-known superconvergence property. In fact, if we take $v_h = e_h$, we then

$$(9.31) \quad \|\nabla(u_h - u_I)\| = O(h^{1+\delta/2}).$$

The above theorem and its proof are quite simple, but nevertheless it indicate something quite interesting. Based on what we know about superconvergence of finite element methods, we have the following remarks:

1. A variationally exact scheme does not need to be pointwise consistent but it may naturally admit optimal order of convergence. But this is not a supraconvergence phenomenon.
2. If a variationally exact scheme happens to lead to a finite difference scheme that has some positive order of truncation error, the approximation will indeed exhibit superconvergence phenomenon.

Another remark we would like to make is that all known proofs of finite element superconvergence are rather elaborate, our new analysis offer an extremely simple approach to superconvergence analysis. But our analysis does not lead sharp result. For example, let us consider the uniform grid on a unique square. it is easy to see that $\delta = 1$ in this case. In fact, for general f ,

$$\begin{aligned} & \frac{1}{h^2}((\nabla u_I, \nabla \phi_i) - (f, \phi_i)) \\ &= \frac{1}{h^2}(\nabla u_I, \nabla \phi_i) - f(x_i) + \frac{1}{h^2} \int_{\Omega} (f(x_i) - f(x)) \phi_i(x) \\ &= O(h^2) + O(h) = O(h). \end{aligned}$$

9.5 On the L^∞ error estimate for finite element method

The L^∞ error estimate is one of the most difficult error estimates to get in the finite element method. We can not use the techniques in finite difference methods because of at least two reasons:

1. Maximal principle is hard to be established.
2. The local truncation error may not approach to zero.

Let us recall roughly how an L^∞ estimate may be obtained in a finite element method. We use the approach of regularized Green function which is defined to be, for any $z \in \Omega$,

$$(9.32) \quad v_h(z) = a(v_h, g_z^h), \forall v_h \in V_h.$$

The choice of g_z^h is not unique. It is possible to choose g_z^h such that

$$(9.33) \quad \max_z \|g_z^h\|_{2,1} \leq C |\log h|.$$

The ordinary Green's function has the singularity like the fundamental solution $\log(x - z)$ and it just miss the space $W^{2,1}$ in two dimensions. But in order that (9.32) to be satisfied only in the finite element space, we can regularize the Green's function that (??) can be established. With the estimate (9.33) at our disposal, we can easily establish the stability estimate for finite element method:

$$(9.34) \quad \|u_h\|_{L^\infty(\Omega)} \leq C |\log h| \|f_h\|_{W_*^{-2,\infty}(\Omega)}$$

where

$$\|f_h\|_{W_*^{-2,\infty}(\Omega)} = \sup_{v \in W^{2,1}(\Omega)} \frac{(f_h, v)}{\|v\|_{W^{2,1}(\Omega)}}$$

As we can see that the finite element stability (9.34) is much stronger than the counter part in the finite difference estimate. The variational property of the finite element method makes it possible to establish such a strong stability result. This stability estimate can be used to establish the nearly optimal error estimate in the L^∞ norm even though the finite element element local truncation error does not go to zero. In fact

$$\begin{aligned} (u_h - u_I)(z) &= a(u_h - u_I, g_z^h) = a(u - u_I, P_h g_z^h) \\ &= a(u - u_I, (P_h - I)g_z^h) + a(u - u_I, g_z^h) \\ &\leq |u - u_I|_{1,\infty} |(I - P_h)g_z^h|_{1,1} + (u - u_I, Ag_z^h) \\ &\leq h^2 |u|_{2,\infty} |g_z^h|_{2,1} + \|u - u_I\|_{0,\infty} |g_z^h|_{2,1} \\ &\leq h^2 |\log h| |u|_{2,\infty}. \end{aligned}$$

We roughly have obtained the following well-known error estimate

$$(9.35) \quad \|u - u_h\|_{0,\infty,\Omega} \leq Ch^2 |\log h| |u|_{2,\infty}.$$

9.6 Stability of convection-diffusion problems

Oftentimes partial differential equations come with parameters. It is sometimes important to design numerical algorithms that are uniformly bounded with respect to the parameters. Let us discuss this phenomenon using a simple example of convection diffusion problem in one dimension.

We consider the model problem:

$$(9.36) \quad -\epsilon u'' + u' = f(x), \quad x \in (0, 1), \quad u(0) = u(1) = 1.$$

What would be the right concept of stability of this problem?

The simplest stability result follows from the maximal principle:

$$\|u\|_{L^\infty} \leq C \|f\|_{L^\infty}$$

where C is a constant independent of ϵ .

Now if we discretize the equation (9.36), what can we say about its stability?

When ϵ is small, it is well-known that a standard finite element method or central difference scheme on a uniform grid is not stable. The instability can be seen from the oscillations in the numerical solution.

The upwinding scheme, on the other hand, is known to be stable. In fact, by a discrete maximum principle, we have

$$\|u_h\|_{L^\infty} \leq C \|f_h\|_{L^\infty}.$$

But since the truncation error is of first order, the upwinding scheme is only first order.