Boris Andreianov
School at Jaca Sept. 2010

Finite volume approximation of the p-laplacian.

The goal of this mini-course is to provide an introduction into finite volume strategies and convergence analysis of the nonlinear diffusion operators of the p-laplacian kind.

Contents

- Leray-Lions kind operators, key properties
- Elliptic problems with Leray-Lions operators: well-posedness, variational nature, regularity of solutions
- meshes of 2D domains, dual meshes, boundary cells, discrete spaces, discrete devergence operator
- strategies for gradient reconstruction, 2D lemmas, discrete duality (covolume scheme, DDFV)
- finite volume schemes with Leray-Lions operators, discrete weak formulation
- discrete solutions: uniqueness, a priori estimates, existence (Brouwer, topological degree), maximum principle
- discrete energy functional, minimization, practical resolution of the discrete equations
- projection of test functions, consistency properties
- discrete compactness (H^1 weak, L^1 strong)
- discrete Poincaré inequality and embedding inequalities
- passage to the limit: Minty trick, strong convergence
- passage to the limit: use of the Young measures
- basic strategy for error estimates for $C^2/W^{2,p}$ solutions
- consistency with Sobolev functions
- superconvergence estimates for $C^4/W^{4,1}$ solutions
- refinements using Bärveff-like arguments
- discrete Besov framework on uniform cartesian meshes
- Chow error estimates in the Besov context, optimality
- some words on different numerical strategies

Lecture 1 (2h.) p.2-8
Lecture 2 (2h) p.9-15

(I) **Def** (Leray-Lions operators)

Let $\vec{a}: \mathbb{R}^d \to \mathbb{R}^d$, continuous, st for some $p \in (1, +\infty)$

- $\vec{a}(\vec{\xi}) \cdot \vec{\xi} \geq \frac{1}{C} |\vec{\xi}|^p$ (coercivity)
- $|\vec{a}(\vec{\xi})|^{p'} \leq C(1 + |\vec{\xi}|)^p$ (growth)
- $(\vec{a}(\vec{\xi}) - \vec{a}(\vec{\eta})) \cdot (\vec{\xi} - \vec{\eta}) \geq 0$ (monotonicity)

Then the operator $A : W^{1,p}(\Omega) \to W^{-1,p'}(\Omega)$

$$u \longmapsto -\operatorname{div} \vec{a}(\vec{\nabla} u) + BC$$

is called Leray-Lions operator.

The prototype of Leray-Lions ops is the p-laplacian: $\vec{a}(\vec{\xi}) = |\vec{\xi}|^{p-2} \vec{\xi}$.
This is also an example of operator deriving from a scalar potential: namely, $\vec{a}(\vec{\xi}) = \nabla_{\vec{\xi}} \Phi(\vec{\xi})$, where $\Phi(\vec{\xi}) = \frac{1}{p} |\vec{\xi}|^p$. In this case, $\vec{a}$ is cyclically monotone.

The p-laplacian corresponds to a strictly monotone $\vec{a}$; moreover, $\vec{a}$ and $\vec{a}^{-1}$ are (locally) Hölder continuous (we'll see later the precise meaning).

Reference Lions "Qqs pbs aux limites non linéaires", Chap. "Compacité + monotonie".

Model problem $\Omega \subset \mathbb{R}^d$ bdd domain

$$(MP_1) \quad \begin{cases} u - \operatorname{div} \vec{a}(\vec{\nabla} u) = f & \text{in } \Omega \\ + \text{homogeneous Neumann (zero-flux) BC}: \ \vec{a}(\vec{\nabla} u) \cdot \vec{n} = 0 \text{ on } \partial\Omega. \end{cases}$$

Mathematical interpretations:

(weak sol') : for $f \in L^{p'}(\Omega)$ (more generally, $f \in W^{-1,p'}(\Omega)$)

$u \in W^{1,p}(\Omega) \cap L^2(\Omega)$

$\forall \varphi \in \mathcal{D}(\Omega)$ ($C^\infty(\bar{\Omega})$ is enough)

$$(W) \qquad \int_\Omega u\varphi + \vec{a}(\vec{\nabla} u) \cdot \vec{\nabla} \varphi = \int_\Omega f \varphi$$

(optimization)

$$u = \operatorname*{arg\,min}_{v \in W^{1,p}(\Omega)} J[v], \qquad J : v \longmapsto \frac{1}{2} \int_\Omega (u - f)^2 + \int_\Omega \Phi(\vec{\nabla} u)$$

Actually, J is a strictly convex lower semi-continuous coercive functional (thus the minimization problem has a unique solution), and (W) is the associated Euler-Lagrange equation for critical points.

It is clear that the minimum is unique; to show uniqueness for (W) take $u, \hat{u}$ two solutions, and take $\varphi = u - \hat{u}$ for the test function.

Then $$\int |u - \hat{u}|^2 \leq \int |u - \hat{u}|^2 + (\vec{a}(\vec{\nabla} u) - \vec{a}(\vec{\nabla} \hat{u})) \cdot (\vec{\nabla} u - \vec{\nabla} \hat{u}) = \int (f - f)(u - \hat{u}) = 0.$$

Existence of a solution will be the byproduct of the numerical method; L^1 stability can be obtained from the above calculations, but is more intricate.

One can note the following properties:

- $\int_\Omega u^2 + \int_\Omega |\nabla u|^p \leq C(\|f\|_{L^{p'}})$ (test function $\varphi = u$ + the Poincaré inequality: $\|u\|_{L^p} \leq C(p,\Omega)\left(\|\nabla u\|_{L^p} + \|u\|_{L^1}\right)$)
- $\int |\nabla u| \leq \int |f|$ (test functions approximating $\mathrm{sign}(u)$), $\int |\nabla u|^q \leq \int |f|^q$ (test fcts approx. $|u|^{q-2}u$, then Young ineq.)
- $\int (u-\hat{u})^+ \leq \int (f-\hat{f})^+$ (idem, with $\mathrm{sign}^+(u-\hat{u})$) $\rightsquigarrow$ L^1 contraction and comparison principle
- maximum principle: $f \leq M$ a.e. $\Rightarrow u \leq M$ a.e.
  (idem, with $\mathrm{sign}^+(u-M)$; comparison with constant solution)

The accurate proof of the three latter properties uses regularization, but in the discrete context things are simpler.

The issue of regularity of solutions, essential for error estimates, is much more delicate than in the linear case. Available results are:

- Hölder regularity (Uhlenbeck, 1988)
- $W^{2,p}$ regularity for $p \leq 2$, and some Besov space $\left(B^{1-\frac{1}{p-1},p}_{\infty}\right)$ regularity for $p > 2$.

(IV) In numerical analysis, one seeks for "good" ways to approach solutions (here, of PDEs). "Good" may mean many things:

- error of approximation as small as possible
- computational complexity as low as possible / qualitative properties of differential / discrete analogues of operators (discrete duality)
- qualitative properties of solutions (e.g. maximum principle, comparison and contraction principles, link with optimization problem) preserved at the discrete level
- possibility to work on "general meshes" (complex geometries, mesh refinement, ...)

For a theoretical analysis, a minimal requirement is: consistency + stability of the method. Convergence is sometimes more difficult to obtain.

Here, we will explore methods that are

- convergent
- preserve many of the structure properties of operators
- work on general meshes, and preserve comparison properties on "orthogonal" meshes.

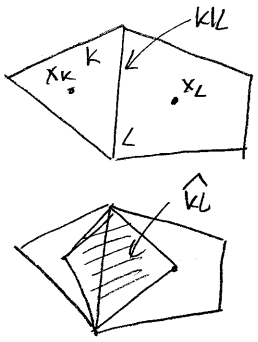
The discrete problem includes a finite number of real values (called degrees of freedom). They may have different interpretations (coefficients of a finite element basis, spectral coefficients); in finite volumes context the degrees of freedom u_K are associated to volumes K (eventually, via the "centers" x_K of the volumes).

Therefore the discretization starts with a mesh of Ω.

Dim. 1 is too easy, dim. 3 is too hard, take dimension 2.

We assume that $\Omega \subset \mathbb{R}^2$ is polygonal, and look on polygonal meshes.

Description



$K \in \mathcal{M}$: volumes

x_K : center of K (barycenter: inside K if K convex
circumcenter: may lie outside K)

K|L : interface between neighbours ($L \in \mathcal{N}(K)$), $d_{KL} = |x_K - x_L|$

$\widehat{K|L}$: "diamond"

the mesh is orthogonal if $\overrightarrow{x_K x_L} \perp K|L$

Important: $\overrightarrow{x_K x_L}$ should form an acute angle with $\vec{n}_K$.

Approx. of $u - \operatorname{div} \vec{F} = f$

we need one unknown u_K per volume K, which is interpreted (≡ which approximates) both as the mean value of u on K and as the point value of u at x_K.

One "integrates" the PDE on K and uses the Green-Gauss formula:

$$m_K u_K - \int_{\partial K} \vec{F} \cdot \vec{n}_K = m_K f_K \quad \text{where } f_K = \fint_K f \ (\text{or } f_K = f(x_K), \text{ anyway quadratures are needed to evaluate } f_K)$$

Thus we need to approximate

$$\int_{\partial K} \vec{F} \cdot \vec{n} = \sum_{L \in \mathcal{N}(K)} m_{KL} \left(\fint_{K|L} \vec{F} \cdot n_K \right) \ ; \text{ one needs the values } F_{K|L} = \fint_{K|L} \vec{F} \cdot \vec{n}_K .$$

In our context, $\vec{F} = \vec{a}(\nabla u)$; for a discrete counterpart, we need to relate $F_{K|L}$ to $(u_K)_K$. There are different strategies for doing this:

- (mixed formulations) put more equations keeping $F_{K|L}$ as unknowns
- (finite volume schemes) reconstruct $F_{K|L}$ "by hands" from the values u_K, u_L and may be some neighbour values.

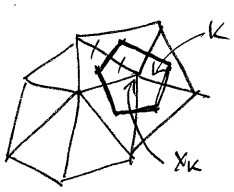
Case $\vec{a}(\cdot) = \mathrm{Id}$, orthogonal mesh $\quad F_{K|L} = \dfrac{u_K - u_L}{d_{KL}}$

Case of the p-laplacian Even on an orthogonal mesh, one seems to need the whole vector $\vec{F}_{K|L}$ on K|L...

III Covolume scheme on triangle mesh

Quite often, we are given a partition $\mathcal{T}$ of Ω onto triangles satisfying the so-called Delaunay condition. With x_{K^*} = circumcenter of K^*, we get an orthogonal mesh... Instead of using this for FV discretization, we will put the unknowns into the vertices of this mesh, and consider $\mathcal{T}$ is the dual to the mesh $\mathcal{M}$ defined as follows (the Voronoi mesh):

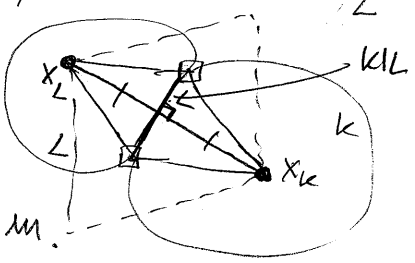
for x_K a vertice of $\mathcal{T}$, $K \in \mathcal{M}$ is the set of $x \in \Omega$ st $|x - x_K| = \min_L |x - x_L|$.



This mesh is also orthogonal,

making zoom on a diamond we have:

- ∘ : vertices of $\mathcal{T}$ ≡ centers of $\mathcal{M}$
- □ : (circumcenters of $\mathcal{T}$ ≡ vertices of $\mathcal{M}$.



Thus our unknowns are located at the vertices of the triangles constituting the mesh $\mathcal{T}$; then, as in the finite element context, it is tempting to use affine / triangle interpolation of these values to produce a function on Ω. Its gradient is constant per triangle; in particular, on each interface KL it takes two values.

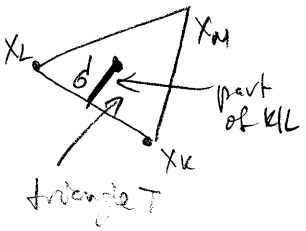
Then we can use this to produce the following scheme:

NB Formally, at $\sigma \in \partial K \cap \partial\Omega$ we set the flux to zero; thus the BC is taken into account

DE9) $\forall K \in \mathcal{M}$, $\quad m_K u_K - \sum_{\sigma \in \partial K \cap \partial\Omega} m_\sigma \vec{a}(\vec{\nabla}_{T(\sigma)} u_h) \cdot \vec{n_K} = m_K f_K$

Here σ's are half-interfaces included in ∂K, and $T = T(\sigma)$ is the triangle containing σ; $u_h = (u_K)_K$ is the set of all unknowns; $m_K = \text{meas}_2(K)$, $m_\sigma = \text{meas}_1(\sigma)$; and

$\nabla_T u_h$ = the gradient of the unique affine function obtained by interpolation of the values u_K, u_L, u_M at the vertices x_K, x_L, x_M of the triangle $T(\sigma)$, for $\sigma \subset KL$.



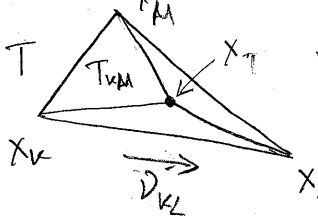
This gives a set of $\#K$ eqns on $\#K$ unknowns, which looks reasonable:

- u_K interpreted as values at (and near) x_K
- starting from $(u_K)_K$, the approximation of ∇u (thus of $\vec{a}(\vec{\nabla u})$) is produced in an "intuitively consistent" way
- starting from the approximate fluxes replacing $\vec{a}(\vec{\nabla u})$ the approximation of $\operatorname{div} \vec{a}(\vec{\nabla u})$ is consistent both mathematically and physically (balance law respected by construction; in particular, the scheme is "locally conservative").

To write the system explicitly, one needs

- geometric quantities from the mesh ($m_K, m_\sigma, \vec{n_K}$)
- pointwise values of quadratures of f (to get f_K)
- an explicit formula for $\vec{\nabla}_T u_h$ in terms of u_K, u_L, u_M and x_K, x_L, x_M.

Lemma Rec1 $\quad \vec{\nabla}_T u_h = \frac{2}{m_T}\left(m_{T_{KL}} \frac{u_L - u_K}{d_{KL}} \vec{\nu}_{KL} + m_{T_{LM}} \frac{u_M - u_L}{d_{LM}} \vec{\nu}_{LM} + m_{T_{MK}} \frac{u_K - u_M}{d_{MK}} \vec{\nu}_{MK} \right)$



where T_{KL}, T_{LM}, T_{MK} are triangles with vertice x_T = circumcenter of T and the opposite side KL, LM, MK, respectively

and $m_{T_{KL}}, \ldots$ are the signed measures of these triangles

Proof: can be obtained by elementary tools (sine theorem, etc.)

- generalization and proof will be given later.

Heuristic explanation: If $\vec{r} = \vec{\nabla}_T u_h$, then $\frac{u_L - u_K}{d_{KL}} \vec{\nu}_{KL} = \mathrm{Proj}_{\vec{\nu}_{KL}} \vec{r}$.

The formula reads $Id = 2\left(\alpha\, \mathrm{Proj}_{\vec{\nu}_{KL}} + \dots + \gamma\, \mathrm{Proj}_{\vec{\nu}_{MK}}\right)$, $\alpha+\beta+\gamma = 1$. (*)

Because projection takes "in the mean" one direction out of two, the role of the coefficient 2 is clear (say, if $\vec{r}$ is random, then $\mathbb{E}[\vec{r}] = 2\,\mathbb{E}[\alpha \mathrm{Proj}_{\vec{\nu}_1}\vec{r} + \dots + \gamma \mathrm{Proj}_{\vec{\nu}_3}\vec{r}]$. The fact that we have the identity (*) is due to the choice of the point x_T = the circumcenter of T... yes?? we'll see.

(IV) Now we have written the scheme; to analyse it, we need much more. For the analysis of convergence, we have to place u (a function on Ω, in $W^{1,p}(\Omega) \cap \dots$) and $u_h = (u_K)_K \in \mathbb{R}^{\#K}$ at the same level; that is, either "project" u on the mesh, or else "lift" $(u_K)_K$ to a function on Ω (cf. the way $(f_K)_K$ were obtained from f) $\{ (\nabla_T u_h)_T$ — " — " — .

If we project u, we still need some $\mathcal{M}$-independent (or rather $\mathcal{M}$-uniform) way to measure the difference $u_h - \mathbb{P}^{\mathcal{M}} u$ (actually, this approach is used for error estimates); thus we still need to interpret discrete functions in some $\mathcal{M}$-independent way.

Therefore lifting operators seem necessary ...

Def L • Given $u_h = (u_K)_{K\in\mathcal{M}}$, $\mathcal{L}_h u_h = \left[\begin{array}{ccc} \Omega & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \sum_K u_K \mathbb{1}_K(x) \end{array}\right]$

In the sequel, $(\mathcal{L}_h u_h)(x)$ will be simply (=abusively) denoted by $u_h(x)$.

• Given $\vec{\mathcal{F}}_h = (\vec{\mathcal{F}}_T)_{T\in\mathcal{T}}$, $\mathcal{L}_h \vec{\mathcal{F}}_h = \left[\begin{array}{ccc} \Omega & \longrightarrow & \mathbb{R}^2 \\ x & \longmapsto & \sum_T \vec{\mathcal{F}}_T \mathbb{1}_T(x) \end{array}\right]$

$(\mathcal{L}_h \vec{\mathcal{F}}_h)(x)$ will be denoted $\vec{\mathcal{F}}_h(x)$.

In particular, if $\vec{\mathcal{F}}_T = \vec{\nabla}_T u_h$, we write $(\vec{\nabla}_h u_h)(x)$.

The analysis of the continuous equation been based on multiplication of the equation by test functions (in some duality framework), we'll probably need the same in the discrete setting. But because everything is finite-dimensional here, we can always use the "L^2-L^2" multiplication. Thus, what we need is a discrete scalar product related to $\int_\Omega uv$ or to $\int_\Omega \vec{F}\cdot\vec{G}$. As we have seen, in our discrete framework scalars and vectors "live" on different meshes: $u_h = (u_K)_K \in \mathbb{R}^{\#K}$, $\vec{\mathcal{F}}_h = (\vec{\mathcal{F}}_T)_T \in (\mathbb{R}^2)^{\#\mathcal{T}}$. Then we can use the above definition to set

Def SP

$$[\![u_h, v_h]\!]_h = \sum_K m_K u_K v_K \quad \left(= \int_\Omega u_h(x)\, v_h(x)\, dx\right)$$

$$\{ \vec{\mathcal{F}}_h, \vec{\mathcal{G}}_h \}_h = \sum_T m_T \vec{\mathcal{F}}_T \cdot \vec{\mathcal{G}}_T \quad \left(= \int_\Omega \vec{\mathcal{F}}_h(x)\cdot\vec{\mathcal{G}}_h(x)\, dx\right).$$

Now, recall our discrete equations: $\forall K \quad u_K |\omega_K| - \sum_{\substack{\sigma \subset \partial K \\ \sigma \not\subset \partial\Omega}} m_\sigma \vec{a}(\vec{\nabla}_{T(\sigma)} u_h) \cdot \vec{n}_K = |\omega_K| f_K$

Take some $v_h = (v_K)_K$ and multiply the eqns term by term:

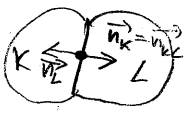
$$\sum_K |\omega_K| u_K v_K + \text{truc} = \sum_K |\omega_K| f_K v_K \text{, i.e., } [\![u_h, v_h]\!]_h + \text{truc} = [\![f_h, v_h]\!]_h.$$

Nice! What have we done with "truc" in the continuous framework? We integrated by parts! There is an analogous procedure in the discrete setting: it is called "gathering by edges" (here, by half-edges σ).

Indeed, each half-edge σ is included in some interface $K|L$; thus in "truc" it appears twice (the boundary edges are excluded from summation, they never appear). Moreover, while $\vec{a}(\vec{\nabla}_{T(\sigma)} u_h)$ only depends on T (i.e. it is the same for K and for L), the normals are reversed: $\vec{n}_L = -\vec{n}_K$. Therefore (with $\vec{F}_{T(\sigma)} = \vec{a}(\vec{\nabla}_{T(\sigma)} u_h)$)

$$\text{truc} := -\sum_K \Big(\sum_{\sigma \subset \partial K \cap \partial\Omega} m_\sigma \vec{F}_{T(\sigma)} \cdot \vec{n}_K \Big) v_K = -\sum_\sigma m_\sigma \vec{F}_{T(\sigma)} (v_K - v_L) \cdot \vec{n}_{KL}$$

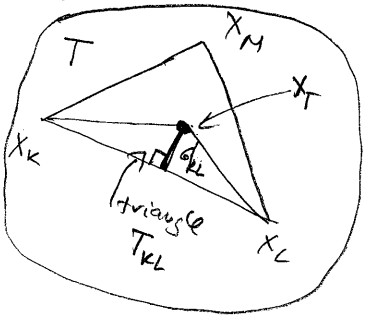
$$= +\sum_\sigma m_\sigma (\vec{F}_{T(\sigma)} \cdot \vec{n}_{KL})(v_L - v_K) \quad \text{where } \sigma \subset K|L.$$



Here comes a wonderful thing... let us continue the calculation, gathering by T:

$$\text{truc} = \sum_T \vec{F}_T \cdot \Big\{ \underbrace{m_{\sigma_{KL}} (v_L - v_K) \vec{n}_{KL} + \dots + m_{\sigma_{MK}} (v_K - v_M) \vec{n}_{MK}}_{\text{with the notation in } T} \Big\}$$

$$= \sum_T \vec{F}_T \cdot 2 \Big\{ \underbrace{\tfrac{1}{2} m_{\sigma_{KL}} d_{KL}}_{= m_{T_{KL}}} \frac{v_L - v_K}{d_{KL}} \vec{n}_{KL} + \dots + \underbrace{\tfrac{1}{2} m_{\sigma_{MK}} d_{MK}}_{= m_{T_{MK}}} \frac{v_K - v_M}{d_{MK}} \vec{n}_{MK} \Big\}$$



$$= \sum_T \vec{F}_T \cdot \vec{\nabla}_T v_h$$ miracle!

Rq Thanks to the orthogonality of the mesh, we have $\vec{\nabla}_{KL} = \vec{n}_{KL}$.

We have found that $\text{truc} = \{\!| \vec{a}(\vec{\nabla}_h u_h), \vec{\nabla}_h v_h |\!\}_h$. We have shown one implication in the following proposition:

Proposition WF The scheme is equivalent to its "weak" formulation: find $u_h = (u_K)_K \in \mathbb{R}^{\#K}$ such that

$$\forall v_h = (v_K)_K \in \mathbb{R}^{\#K} \quad [\![u_h, v_h]\!]_h + \{\!| \vec{a}(\vec{\nabla}_h u_h), \vec{\nabla}_h v_h |\!\}_h = [\![f_h, v_h]\!]_h. \qquad \text{(WDEq)}$$

(The proof of the implication "$\Leftarrow$" is obtained by making the same calculation in the inverse sense, and taking $v_{K_0} = 1$, $v_L = 0$ for all $L \neq K_0$).

The miracle of the above calculation should be examined more closely (actually, we will need the same calculation, with different $\vec{F}_h$, for the convergence analysis). Let's formalize the calculation in terms of discrete operators.

Def. DO

- $\nabla_h : \mathbb{R}^{\#K} \to (\mathbb{R}^2)^{\#T}$, $u_h = (u_K)_K \mapsto (\vec{\nabla}_T u_h)_T$ with $\vec{\nabla}_T u_h = \frac{2}{m_T} \Big\{ m_{T_{KL}} \frac{u_L - u_K}{d_{KL}} \vec{n}_{KL} + \dots \Big\}$
- $\mathrm{div}_h : (\mathbb{R}^2)^{\#T} \to \mathbb{R}^{\#K}$, $\vec{F}_h = (\vec{F}_T)_T \mapsto (\mathrm{div}_K \vec{F}_h)_K$ with $\mathrm{div}_K \vec{F}_h = \frac{1}{|\omega_K|} \sum_{\sigma \subset K \cap \partial\Omega} m_\sigma \vec{F}_{T(\sigma)} \cdot \vec{n}_K$.

Rq This is "div + zero-flux BC"

Proposition VII (The discrete duality property)

The operators $-\mathrm{div}_h \cdot$ and $\nabla_h \cdot$ are dual for the products $[\![\cdot,\cdot]\!]_h$, $\{\!\{\cdot,\cdot\}\!\}_h$, i.e.

$$\forall v_h \in \mathbb{R}^{\#\mathcal{K}},\ \forall \vec{\mathcal{F}}_h \in (\mathbb{R}^2)^{\#\mathcal{T}} \quad [\![-\mathrm{div}_h \vec{\mathcal{F}}_h, v_h]\!]_h = \{\!\{\vec{\mathcal{F}}_h, \vec{\nabla}_h v_h\}\!\}_h .$$

Proof: Was given in the "truc" calculation.

V) Properties of the scheme, straightforward from (WDEq):

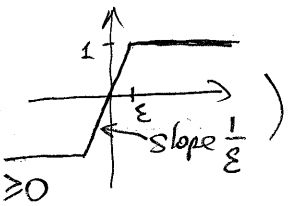
- uniqueness: as in the continuous setting, take $v_h = u_h - \hat{u}_h$ with $u_h, \hat{u}_h$ two solutions.
  Then (WDEq) yields $[\![u_h-\hat{u}_h, u_h-\hat{u}_h]\!]_h + \{\!\{\vec{a}(\vec{\nabla}_h u_h)-\vec{a}(\vec{\nabla}_h \hat{u}_h), \vec{\nabla}_h u_h - \vec{\nabla}_h \hat{u}_h\}\!\}_h = 0$.
  Using lifting operators: $\int_\Omega (u_h-\hat{u}_h)^2(x)\,dx$ = $[\![u_h-\hat{u}_h, u_h-\hat{u}_h]\!]_h$ and $\{\!\{\vec{a}(\vec{\nabla}_h u_h)-\vec{a}(\vec{\nabla}_h \hat{u}_h), \vec{\nabla}_h u_h - \vec{\nabla}_h \hat{u}_h\}\!\}_h$ = $\int_\Omega (\vec{a}(\vec{\mathcal{F}})-\vec{a}(\hat{\vec{\mathcal{F}}}))\cdot(\vec{\mathcal{F}}-\hat{\vec{\mathcal{F}}})(x)\,dx \geq 0$
  Hence $u_h(\cdot) \equiv \hat{u}_h(\cdot)$ a.e, which means $u_K = \hat{u}_K$ for all K, i.e. $u_h = \hat{u}_h$.

- L^q estimates, $1 \leq q \leq \infty$.
  - $q=1$: Take $v_h := \mathrm{sign}\, u_h$ (component per component, i.e., $v_K := \mathrm{sign}\, u_K\ \forall K$)
    NB No regularization is needed! At this point, discrete setting is a bit simpler than the continuous setting.

    Then $[\![u_h, \mathrm{sign}\, u_h]\!]_h + \{\!\{\vec{a}(\vec{\nabla}_h u_h), \vec{\nabla}_h(\mathrm{sign}\, u_h)\}\!\}_h \leq [\![f_h, \mathrm{sign}\, u_h]\!]_h$

    $\int_\Omega u_h(x)\,\mathrm{sign}\, u_h(x)\,dx = \int |u_h(x)|\,dx = [\![|u_h|, 1]\!]_h$ ≥ 0 (?)

    $\int_\Omega f_h(x)\,\mathrm{sign}\, u_h(x)\,dx \leq \int_\Omega |f_h(x)|\,dx = [\![|f_h|, 1]\!]_h$

    NB In general, let S be a Lipschitz non-decreasing function (e.g. a regularization of $z \mapsto \mathrm{sign}\, z$, eg. H_ε: 

)
    Then in the continuous setting, $\vec{a}(\vec{\nabla} u)\cdot\vec{\nabla} S(u) \overset{\text{a.e.}}{=} S'(u)\,\vec{a}(\vec{\nabla} u)\cdot\vec{\nabla} u \geq 0$.
    (chain rule for Lip o W^{1,p} functions!)

    But in the discrete setting, it is easy to see that the chain rule is lost:

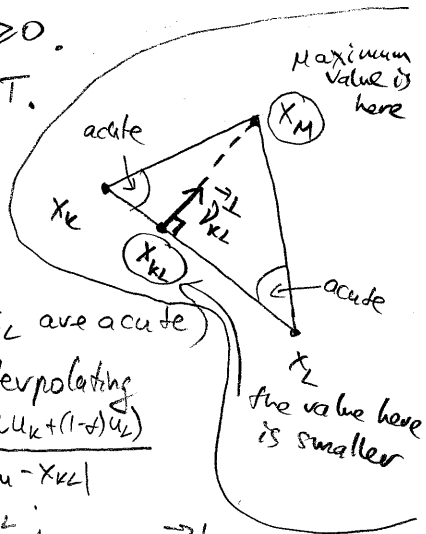
    $\vec{\nabla}_T S(u_h) = \frac{2}{m_T}\left(m_{T_{KL}} \frac{S(u_L)-S(u_K)}{d_{KL}} \vec{n}_{KL} + \dots\right) = \frac{2}{m_T}\left(S'(\xi_{KL}) m_{T_{KL}} \frac{u_L-u_K}{d_{KL}} \vec{D}_{KL} + \dots + S'(\xi_{MK}) m_{T_{MK}} \frac{u_M-u_K}{d_{MK}} \vec{D}_{MK}\right)$
    (mean value theorem)

    The values $S'(\xi_{KL})$, $S'(\xi_{LM})$, $S'(\xi_{MK})$ are, in general, all different, and we cannot write $\vec{\nabla}_T S(u_h) = \Theta\, \vec{\nabla}_T u_h$.
    In particular, the orientation of $\vec{\nabla}_T S(u_h)$ is not necessarily the one of $\vec{\nabla}_T u_h$, and it is delicate (and sometimes false!) to conclude that $\vec{a}(\vec{\nabla}_h u_h)\cdot\vec{\nabla}_h S(u_h) \geq 0$.
    At this point, discrete setting is much more difficult than the continuous one.

  We make two additional assumptions:

  Restrict. Str
  - Structure of $\vec{a}$: $\vec{a}(\vec{\xi}) = k(|\vec{\xi}|)\vec{\xi}$, $k \geq 0$ (verified by the p-laplacian)
  - Structure of the triangulation $\mathcal{T}$: all triangles have angles $\leq 90°$.
    NB In practice, there is no algorithm for creating such partitions of general Ω!

Lemma Under the reftrichous Str, for all non-decreasing (even discontinuous) $S(\cdot)$ we have for all u_h, for all T $\quad \vec{a}(\vec{\nabla}_T u_h)\cdot\vec{\nabla}_T S(u_h) \geq 0$.



Proof Let u_K, u_L, u_M be the values at the vertices x_K, x_L, x_M of T.
Assume u_M is the greatest value. Then also $S(u_M)$ is the greatest among $S(u_K)$, $S(u_L)$, $S(u_M)$.
Take $x_{KL} \in [x_K, x_L]$ such that $\overrightarrow{x_{KL}x_M}$ is aligned with $\vec{\nu}_{KL}^{\perp}$
(this is possible because the angles of T at x_K and at x_L are acute)
Because $\vec{\nabla}_T u_h$ is the gradient of the affine function interpolating the vertices, we have in particular $\vec{\nabla}_T u_h \cdot \vec{\nu}_{KL}^{\perp} = \dfrac{u_M - (\alpha u_K + (1-\alpha)u_L)}{|x_M - x_{KL}|}$
where $\alpha u_K + (1-\alpha)u_L$ is the value of the interpolation at x_{KL};
it is crucial that $\alpha \in [0,1]$, thus $u_M \geq \alpha u_K + (1-\alpha)u_L$ and $\vec{\nabla}_T u_h \cdot \vec{\nu}_{KL}^{\perp} \geq 0$.
In the same way, we see that $\vec{\nabla}_T S(u_h)\cdot\vec{\nu}_{KL}^{\perp} \geq 0$. Thus each of the vectors $\vec{\nabla}_T u_h$ and $\vec{\nabla}_T S(u_h)$ forms an acute angle with $\vec{\nu}_{KL}^{\perp}$; and $\vec{a}(\vec{\nabla}_T u_h)$ is aligned with $\vec{\nabla}_T u_h$ by reftrichous Str.
Now we write both vectors in the basis $(\vec{\nu}_{KL}, \vec{\nu}_{KL}^{\perp})$; we have

$$\vec{a}(\vec{\nabla}_T u_h)\cdot\vec{\nabla}_T S(u_h) = k(|\vec{\nabla}_h u_h|)\left\{ \underbrace{\frac{u_L-u_K}{d_{KL}}\cdot\frac{S(u_L)-S(u_K)}{d_{KL}}}_{\text{the components of the projection on } \vec{\nu}_{KL} \text{ are known!}} + \underbrace{(\vec{\nabla}_T u_h\cdot\vec{\nu}_{KL}^{\perp})}_{\geq 0}\underbrace{(\vec{\nabla}_T S(u_h)\cdot\vec{\nu}_{KL}^{\perp})}_{\geq 0 \text{ as we have just shown}}\right\} \geq 0$$

because $(t-s)(S(t)-S(s)) \geq 0$. ∎

Remark What we have not seen in this proof, is that the orthogonality of the mesh is very important.

Conclusion: we do have $[\![|\nabla u_h|, 1]\!]_h \leq [\![|\nabla f_h|, 1]\!]_h$

- $q \in (1,\infty)$: same reasoning, but use $S(z) = \begin{cases} |z|^{q-2}z, & z \neq 0 \\ 0, & z = 0, \end{cases}$ we get $[\![|\nabla u_h|^q, 1]\!]_h \leq [\![|\nabla f_h|^q, 1]\!]_h$.
- $q = \infty$: see the previous relation as $\left(\int_\Omega |\nabla u_h|^q(x)\right)^{1/q} \leq \left(\int_\Omega |\nabla f_h|^q(x)\right)^{1/q}$, and pass to the limit $q \to \infty$;
we get $\|\nabla u_h(\cdot)\|_\infty \leq \|\nabla f_h(\cdot)\|_\infty$, thus $\max_K |\nabla u_K| \leq \max_K |\nabla f_K|$.

- maximum principle:
almost the same proof; e.g. if $M = \max_K f_K$, substract M to all the values of u_h, f_h and use the test function $S(u_h) = \mathrm{sign}^+(u_h - M)$; we get
$$[\![u_h - M, \mathrm{sign}^+(u_h - M)]\!]_h + (\text{terms} \geq 0) = [\![\underbrace{f_h - M}_{\leq 0}, \underbrace{\mathrm{sign}^+(u_h - M)}_{\geq 0}]\!],$$
thus $\int_\Omega (u_h - M)^+(x)\,dx \leq 0$, i.e. $u^h(x) \leq M$ a.e., i.e. $\forall K \; u_K \leq M$,
i.e. $\max_K u_K \leq \max_K f_K$.

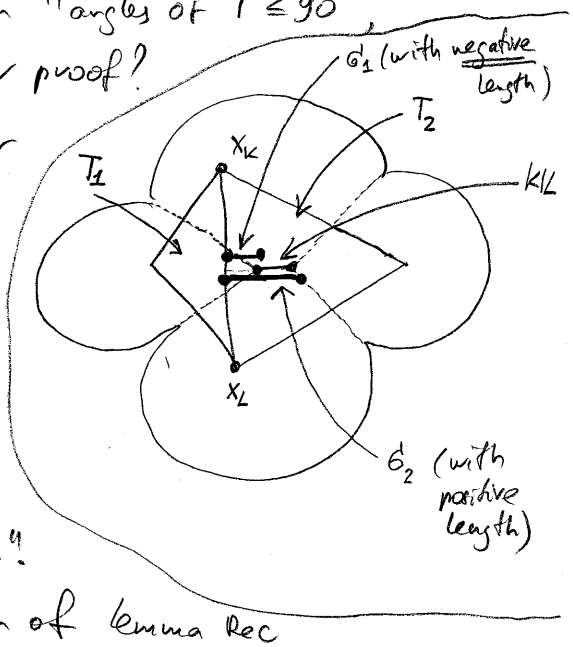
- maximum principle bis:
Let us give another proof, using the standard technique for classical solutions of continuous problem.
NB At this point, discrete setting is much simpler than the continuous one.
Namely, assume $f_h \leq M$ and assume, by contradiction, that $u_K > M$ for some K.
Then we can take the K for which $u_K - M$ attains its (positive) maximum.
Look at the equation for the volume K:
$$\underbrace{m_K u_K}_{\text{term 1}} - \underbrace{\sum_{\text{edges}} m_\sigma k(|\vec{\nabla}_{T(\sigma)} u_h|)\frac{u_L - u_K}{d_{KL}}}_{\text{term 2}} = \underbrace{m_K f_K}_{\text{term 3}}$$

Term 1 is $> m_K M$, term 2 is ≥ 0 because $-(u_L - u_K) \geq 0$ and $k(\cdot) \geq 0$,
and term 3 is $\leq m_K M$. We get $m_K M <$ term 1 + term 2 = term 3 $\leq m_K M$, contradiction.

- Can we prove the comparison principle in the same way?
Let's try! Assume $f_K \leq \hat{f}_K$ for all K, and (by contradiction) take the volume K where $u_K - \hat{u}_K$ attains its positive maximum. The corresponding discrete equation yields:

$$\underbrace{m_K(u_K - \hat{u}_K)}_{\geq 0} - \sum_{\sigma \subset \partial K} m_\sigma \underbrace{\left\{ k\left(|\vec{\nabla}_{T(\sigma)} u_h|\right) \frac{u_L - u_K}{d_{KL}} - k\left(|\vec{\nabla}_{T(\sigma)} \hat{u}_h|\right) \frac{\hat{u}_L - \hat{u}_K}{d_{KL}} \right\}}_{\text{is it } \geq 0 \,??} = m_K (f_K - \hat{f}_K) \leq 0$$

In the linear case ($K \equiv$ const) we get $-\left((u_L - \hat{u}_L) - (u_K - \hat{u}_K)\right) \geq 0$ by the choice of K.
Looks ok? But where have we used the restriction "angles of $T \leq 90°$"
that seemed to be useful for another proof?



Remark In the above scheme, one should consider "negative half-edges σ" except for the case "angles of $T \leq 90°$". Example:

here we have not
$K|L = \sigma_1 \cup \sigma_2$, but
$K|L = \sigma_2 \setminus \sigma_1$.

Thus we assign m_{σ_2} as usual, but we take
for m_{σ_1} the length of σ_1 with the sign "minus"!

NB This convention is compatible with the convention of Lemma Rec
(there we have taken the signed triangle areas $m_{T_{KL}}$, etc).
Proposition DD is true if we stick to the sign convention for both σ's and T_{KL}'s.

Conclusion: the above proof of comparison principle for $k \equiv$ const
works under the same restriction Sfr on the triangulation.

What's for general $k(\cdot)$?

Guess Not true in general.

True if angles of T are all $\leq 90° - \Theta$ where $\Theta = \Theta(\rho)$,
with $\Theta(2) = 0$ and $\lim_{\rho \to 1} \Theta(\rho) = 90° = \lim_{\rho \to \infty} \Theta(\rho)$.

- Should we look at the contraction principle?
No! It is enough to use the celebrated Crandall-Tartar lemma.

Lemma CT [Proc. AMS '80] Assume that $C \subset L^1(\Omega)$ verifies $[f, g \in C \Rightarrow \max\{f, g\} \in C]$
and $T: C \to L^1(\Omega)$ satisfies $\int_\Omega T(f) \leq \int_\Omega f$.
Then (for $f, g \in C$) the following properties are equivalent

- $f \leq g \Rightarrow T(f) \leq T(g)$ (order preservation)
- $\int_\Omega |T(f) - T(g)| \leq \int_\Omega |f - g|$ (L^1 contraction)
- $\int_\Omega (T(f) - T(g))^+ \leq \int_\Omega (f - g)^+$ (L^1 T-contraction)

Application: $C = \left\{ \sum_K u_K \mathbb{1}_K(\cdot),\ u_K \in \mathbb{R} \right\}$; $T: f_h \mapsto u_h$ the solution of (DEq) (OV we need existence first.)
We have $\int_\Omega T(f_h) = \int_\Omega u_h = \sum_K u_K |K| = \sum_K u_K |K| + \sum_{\sigma \in \mathcal{E}_{int}} \vec{\sigma} \cdot \vec{\nabla}_\sigma u_h$, $\vec{\nabla}_\sigma \mathbb{1}_h = 0 \Rightarrow \int_\Omega \mathbb{1}_h = \int_\Omega f_h$ (use (VDEq) with $v_h = 1$).

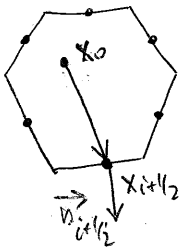
Thus whenever we have the comparison principle, we get contraction on L^1!
Yet, in general, both are false.

(VI) Yet many other important properties of the continuous problem are preserved at the discrete level, thanks to the discrete duality property.

So, let's generalize and prove Lemma Rec 1

(numbered counterclockwise)

Lemma Rec 2 Let T be a polygon, $x_0 \in \mathbb{R}^2$. Let $(x_i)_{i=1}^{\ell}$ be the vertices (we set $x_{\ell+1}=x_1$) and let $x_{i,i+1/2} := \frac{x_i+x_{i+1}}{2}$ (the barycenter of the edge). Let $d_{i,i+1}=|x_{i+1}-x_i|$.



Then for all $\vec{r}\in\mathbb{R}^2$, $\vec{r} = \frac{1}{m_T}\sum_{i=1}^{\ell} d_{i,i+1}(\vec{r}\cdot\overrightarrow{x_0x_{i+1/2}})\vec{n}_{i+1/2}$

Proof Consider $u:\mathbb{R}^2\to\mathbb{R}$ defined by $u(x)=\vec{r}\cdot\overrightarrow{x_0x}$; we have $\vec{\nabla}u=\vec{r}$.

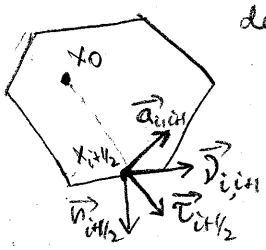
Using the Green-Gauss formula, $m_T r_1 = \int_T \partial_1 u = \int_T dv(u,0) = \int_{\partial T}(u,0)\cdot\vec{n} = \int_{\partial T} u\, n_1$.

In the same way, $m_T r_2 = \int_{\partial T} u\, n_2$, thus $\vec{r} = \frac{1}{m_T}\sum_{i=1}^{\ell} d_{i,i+1}\int_{[x_i,x_{i+1}]}(\vec{r}\cdot\overrightarrow{x_0x})\vec{n}_{i+1/2}$

Because u is affine and $x_{i+1/2}$ is the barycenter of $[x_i,x_{i+1}]$, the result follows ∎

Lemma Rec 3 With the notation of Lemma Rec 2, denote in addition

$\vec{\tau}_{i+1/2}$ = unit vector from x_0 to $x_{i+1/2}$
$\vec{\nu}_{i,i+1}$ = unit vector from x_i to x_{i+1}
$\vec{a}_{i,i+1}$ = unit vector orthogonal to $\vec{\tau}_{i+1/2}$, oriented in the direct sense.



Then for all $\vec{r}\in\mathbb{R}^2$, $\vec{r} = \frac{2}{m}\sum_{i=1}^{\ell}\frac{m_{i,i+1}}{\cos\alpha_{i,i+1}}(\vec{r}\cdot\vec{\nu}_{i,i+1})\vec{a}_{i,i+1}$

with $m_{i,i+1}$ = (signed) area of the triangle $x_0x_ix_{i+1}$
$\alpha_{i,i+1}$ = angle between $\vec{n}_{i+1/2}$ and $\vec{\tau}_{i+1/2}$, $\cos\alpha_{i,i+1}=(\vec{n}_{i+1/2},\vec{\tau}_{i+1/2})=(\vec{\nu}_{i,i+1},\vec{a}_{i,i+1})$

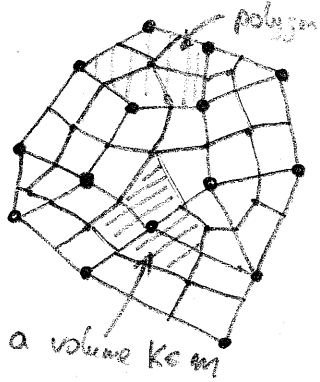
Proof: Several lines starting from Lemma Rec 2 and the identity $\vec{r} = \frac{1}{\cos\alpha_{i,i+1}}\left\{(\vec{r}\cdot\vec{\tau}_{i+1/2})\vec{n}_{i+1/2} + (\vec{r}\cdot\vec{\nu}_{i,i+1})\vec{a}_{i,i+1}\right\}$

(The identity is justified taking scalar product with $\vec{\tau}_{i+1/2}$, $\vec{\nu}_{i,i+1}$ and using orthogonality.) ∎

NB Lemma Rec 1 is a particular case of Lemma Rec 3 (corresponding to T = triangle, x_0 = circumcenter)

Generalization of the scheme

Take an arbitrary partition $\mathcal{T}$ of Ω into polygons, with one point per polygon; then create the "median dual mesh" (= "Donald mesh") $\mathcal{M}$ as shown on the picture.



Define dv_h on the mesh $\mathcal{M}$ as previously.

Define ∇_h piecewise constant per $T\in\mathcal{T}$, using the formula of Lemma Rec 3 ($\vec{r}\cdot\vec{\nu}_{i,i+1}$ substituted by $\frac{u_{i+1}-u_i}{d_{i,i+1}}$).

Proposition DDbis The discrete duality property holds for the so defined scheme.

III) Now, let's embark on analysis of the equations (DEq) (more exactly under the form (WDEq))

A priori estimates

Remind: (WDEq) $[\![u_h, v_h]\!]_h + \{\!\{ \vec{a}(\vec{\nabla}_h u_h), \vec{\nabla}_h v_h \}\!\}_h = [\![f_h, v_h]\!]_h$ for all v_h

Take $v_h = u_h$; we get $\int_\Omega u_h^2(x)dx + \int_\Omega \vec{a}(\vec{\nabla}_h u_h)(x) \cdot (\vec{\nabla}_h u_h)(x) dx = [\![u_h, u_h]\!]_h + \{\!\{ \,, \}\!\}_h = [\![f_h, u_h]\!]_h$

$$= \int_\Omega f_h(x) u_h(x) dx \le \frac{1}{2}\int_\Omega f_h^2(x)dx + \frac{1}{2}\int_\Omega u_h^2(x)dx$$

thus
- $\|u_h(\cdot)\|_{L^2} \le \|f_h(\cdot)\|_{L^2}$
- (by coercivity of $\vec{a}(\cdot)$) $\|(\vec{\nabla}_h u_h)(\cdot)\|_{L^p} \le C(\|f_h(\cdot)\|_{L^2})$ (with tacit convention that the dependence is non-decreasing!)

Moreover, if $f_k := \fint_k f(x)dx$, then from the Jensen inequality, $f_k^2 \le \fint_k f^2(x)dx$;

then $\int_\Omega f_h^2(x)dx = \sum m_k f_k^2 \le \sum m_k \fint_k f^2(x)dx = \int_\Omega f^2(x)dx$,

thus $\|f_h(\cdot)\|_{L^2} \le \|f(\cdot)\|_{L^2}$, for all mesh.

We have shown

Proposition AEst Let f be given. Then for any mesh, any solution of (DEq) verifies $\|u_h(\cdot)\|_{L^2} \le C$, $\|(\vec{\nabla}_h u_h)(\cdot)\|_{L^p} \le C$ with some constant C that only depends on $f, \Omega, \vec{a}$.

Variational interpretation

Here we assume $\vec{a} = \nabla \Phi$ (this is the case of the p-laplacian, $\Phi(\vec{\xi}) = \frac{1}{p}|\vec{\xi}|^p$) with strictly convex coercive functional $\Phi: \mathbb{R}^2 \to \mathbb{R}$

Set $\mathcal{J}_h : \mathbb{R}^{*K} \to \mathbb{R}$

$$v_h \mapsto \frac{1}{2}\int_\Omega (u_h - f_h)^2(x)dx + \int_\Omega \Phi(\vec{\nabla}_h u_h)(x)dx = \frac{1}{2}\sum_K m_K (u_K - f_K)^2 + \sum_T m_T \Phi(\vec{\nabla}_T u_h)$$

Proposition VarII (i) $\mathcal{J}_h$ admits a unique minimizer on $\mathbb{R}^{*K}$

(ii) This minimiser is a solution of (WDEq) (the unique solution)

Proof (i) One easily checks that $\mathcal{J}_h$ is a strictly convex continuous (in any norm of $\mathbb{R}^{*K}$) functional, it is coercive e.g. for $\|u_h\|_{\mathbb{R}^{*K}} := \sqrt{\sum_K m_K u_K^2} + \sqrt[p]{\sum_T m_T |\vec{\nabla}_T u_h|^p}$.

Thus it admits a unique minimizer

(ii) It is sufficient to write the Euler-Lagrange equation: $(d\mathcal{J}_h|_{u_h})[v_h] = 0 \iff$

$$\iff \lim_{t\to 0} \frac{\mathcal{J}_h[u_h + t v_h] - \mathcal{J}_h[u_h]}{t} = \lim_{t\to 0}\left(\int_\Omega \frac{(u_h + t v_h - f_h)^2 - (u_h - f_h)^2}{2t} + \int_\Omega \frac{\Phi(\vec{\nabla}_h u_h + t\vec{\nabla}_h v_h) - \Phi(\vec{\nabla}_h u_h)}{t} \right)$$

($\vec{\nabla}_h \cdot$ is linear)

$$= \lim_{t\to 0}\left(\int_\Omega \frac{t v_h (2u_h - 2f_h + t v_h)}{2t} + \int_\Omega \frac{\vec{a}(\vec{\nabla}_h u_h)\cdot(t\vec{\nabla}_h v_h) + \vec{o}(t)}{t} \right)$$

$$= \int_\Omega (u_h - f_h) v_h + \int_\Omega \vec{a}(\vec{\nabla}_h u_h)\cdot\vec{\nabla}_h v_h \equiv [\![u_h - f_h, v_h]\!]_h + \{\!\{ \vec{a}(\vec{\nabla}_h u_h), \vec{\nabla}_h u_h \}\!\}_h$$

Thus (WDEq) is the Euler-Lagrange equation for the minimization of $\mathcal{J}_h$. □

Existence, general case If $\vec{a}$ does not derive from a scalar potential Φ, we have to use topological existence results (a version of the Brouwer theorem or the topological degree arguments) in order to get existence.

For the application of the Brouwer theorem: see [Lions '69, Lemma 4.3].

Let us give the topological degree argument. Reference: J. Droniou (visit his personal homepage)

Let X be a finite-dimensional space.

Idea Let $\mathcal{H}: [0,1] \times X \longrightarrow X$ be a continuous map such that

- $\mathcal{H}(0,\cdot) = 0$ admits a solution
- for all solution x of $\mathcal{H}(t,\cdot)=0$, one has $\|x\|_X \leq R$

Then for all t, $\mathcal{H}(t,\cdot)=0$ has at least one solution, and in particular $\mathcal{H}(1,\cdot)=0$ admits a solution.

(Tools: homotopy, notion of topological degree and its invariance per homotopy)

Application: • $X = \mathbb{R}^{*K}$ (any norm can be taken on X), $\mathcal{H}: (t, u_h) \longmapsto \left(m_K(u_K - t f_K) - \mathrm{div}_K^{\vec{a}}(\nabla_h u_h) \right)_K$

$[0,1] \times \mathbb{R}^{*K} \longrightarrow \mathbb{R}^{*K}$

- The continuity of $\mathcal{H}$ is evident ($\mathcal{H}(t,(u_K)_K)$ is produced by a finite number of arithmetic operations and a composition by the continuous map $\vec{a}(\cdot)$)
- For $t=0$, $u_h = 0$ is the unique solution
- Applying Proposition AEst with f_h replaced with $t f_h$, we get the uniform bound on possible solutions $u_h^{(t)}$ of $\mathcal{H}(t,\cdot)=0$ ($\Leftrightarrow$ the scheme (DEq) with source $t f_h$), eg. in the norm $\|u_h\|_X := \sqrt{[\![u_h, u_h]\!]_h}$.

Conclusion. There exists a unique solution to the scheme (DEq).

- In the variational case ($\vec{a} = \nabla \mathcal{P}$), it can be obtained by minimization of the discrete energy $\mathcal{J}_h$. In practice, this allows for the use of "descent" iteration methods: conjugate gradient, Polak-Ribière ...
- In the non-variational and in the variational case, one could use the Newton method to solve the nonlinear system; but in practice, it is difficult to pick the initial guess for which the algorithm would converge. Newton works better for the evolution problem associated with $-\mathrm{div}\,\vec{a}(\cdot)$.
- In the non-variational and in the variational case, the "coordination-decomposition" algorithm of Glowinski and Mavrocco (see the recent book of Glowinski) appears to be quite efficient.

We do not discuss more the efficiency of the solution of the nonlinear system (DEq).

Now, the questions are:

- in which sense the discrete solutions could converge to the continuous ones?
- how to prove such convergence?
- at what minimal rate the convergence takes place (a priori error estimates)?
- what rate of convergence do we observe in practice (numerical tests)?
- * can we use the observed numerical solutions to improve the convergence rate estimate (a posteriori error estimates)?
- * can we use the observed numerical solution in order to refine the mesh adaptively and thus improve the ratio approximation/computational effort?

*: will not be answered.

Because we have lifted u_h and $\vec{\nabla}_h u_h$ to functions on Ω, it is natural to state convergence in the following terms: $u_h(\cdot) \to u(\cdot)$, $(\vec{\nabla}_h u_h)(\cdot) \to \nabla u(\cdot)$, but in which sense?

We have the following "asymptotic compactness" result

Proposition Cmp Let $\mathcal{T}_h, \mathcal{M}_h$ be a family of dual/primal meshes and u_h, the associated discrete functions satisfying the estimates of Prop. AEst. Then there exists a sequence $(h_m)_{m\in\mathbb{N}}$ [in the sequel we drop the subscript m] and a function $u \in W^{1,p}(\Omega) \cap L^2(\Omega)$ such that

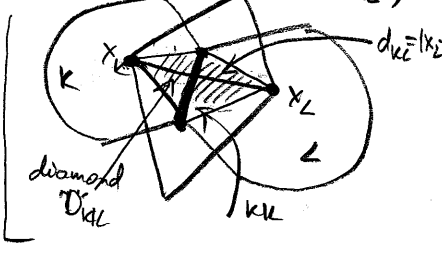
($\mathcal{T}_h$ should be triangulations)

- $u_h(\cdot) \longrightarrow u(\cdot)$ in $L^2(\Omega)$ (weakly)
- $(\vec{\nabla}_h u_h)(\cdot) \longrightarrow \vec{\nabla} u(\cdot)$ in $L^p(\Omega)$ (weakly)
- moreover, $u_h(\cdot) \longrightarrow u(\cdot)$ in $L^1(\Omega)$ (strongly).

Remark This is the simplest statement; using discrete Poincaré embeddings (that may require some uniform regularity of the meshes) one can improve the last statement to a strong convergence in $L^q(\Omega)$ for all $q < p^*$ (the critical Sobolev exponent)

The proof is split into several steps

- Basic weak compactness: the a priori estimates of Proposition AEst yield:
  there exists $u \in L^2(\Omega)$, $\vec{w} \in L^2(\Omega;\mathbb{R}^d)$ such that (for a subsequence)
  $u_h(\cdot) \to u(\cdot)$ in $L^2(\Omega)$
  $(\vec{\nabla}_h u_h)(\cdot) \to \vec{w}(\cdot)$ in $L^p(\Omega)$
  (both convergences are weak)
- Estimate of L^1 translates:
  - It is enough to show the $L^1_{loc}(\Omega)$ compactness. Indeed, if $(\omega_n)_n$ is a sequence of compacts such that $\omega_n \uparrow \Omega$, if $(u_h)_h$ is relatively compact in $L^1(\omega_n)$ for all n, then using the diagonal procedure we get a subsequence $(u_h)_h$ and a $u \in L^1_{loc}(\Omega)$ such that $u_h(\cdot) \to u(\cdot)$ in $L^1(\omega_n)$ for all n. The limit $u(\cdot)$ is the same as in the previous step, because weak $L^2(\Omega)$ convergence and strong $L^1(\omega_n)$ convergence both imply $\mathcal{D}'(\omega_n)$ convergence, and the $\mathcal{D}'$ limit is unique [we work with some fixed convergent extracted subsequence].
    Moreover, $\|u_h(\cdot) - u(\cdot)\|_{L^1(\Omega)} \le \underbrace{\|u_h(\cdot) - u(\cdot)\|_{L^1(\omega_n)}}_{\to 0} + \underbrace{\sqrt{|\Omega \setminus \omega_n|}}_{\to 0} \underbrace{\|u_h(\cdot) - u(\cdot)\|_{L^2(\Omega)}}_{\le 2\cdot \text{const by Prop. AEst.}}$
  - Fix $\omega \in \Omega$, take $\Delta \in \mathbb{R}^2$ such that $|\Delta| < \mathrm{dist}(\omega, \partial\Omega)$. For all $x \in \Omega$ and all interface $K|L$, define $\theta_{K|L}(x) := \begin{cases} 1, & \text{if } [x, x+\Delta] \cap (K|L) \neq \emptyset \\ 0, & \text{otherwise.} \end{cases}$
    Consider $I(\Delta) = \int_\omega |u_h(x+\Delta) - u_h(x)|\,dx$; we notice that
    $$|u_h(x+\Delta) - u_h(x)| \le \sum_{K|L:\, [x,x+\Delta]\cap(K|L)\neq\emptyset} |u_L - u_K| = \sum_{K|L} \theta_{K|L}(x)\,|u_L - u_K|.$$
    Then $I(\Delta) \le \sum_{K|L} \left(\int_\omega \theta_{K|L}(x)\,dx\right) |u_L - u_K|$. But $\mathrm{meas}\{x \mid [x, x+\Delta] \cap (K|L) \neq \emptyset\} \le \Delta \cdot m_{K|L}$,
    thus $I(\Delta) \le \Delta \sum_{K|L} m_{K|L} d_{K|L} \dfrac{|u_L - u_K|}{d_{K|L}}$. Further, introducing "diamonds" $\mathcal{D}_{K|L}$ as in the picture, we have
    $$I(\Delta) \le 2\Delta \sum_{K|L} m_{\mathcal{D}_{K|L}} \left(\frac{|u_L - u_K|}{d_{K|L}}\right) \le 2\Delta \int_\Omega |\vec{\nabla}_h u_h|(x)\,dx \le C\Delta$$
    $\le |(\vec{\nabla}_h u_h)(\cdot)|$ on $\mathcal{D}_{K|L}$; $\le C\,\|\vec{\nabla}_h u_h\|_{L^p(\Omega)}$



Therefore the space translates of the functions $u_h(\cdot)$ on ω are uniformly bounded; by the Kolmogorov compactness theorem, one can extract an $L^1(\omega)$-convergent subsequence.

Question where have we used the assumption that T are triangles??

Answer Only in this case, the divided difference $\frac{|u_L - u_K|}{d_{KL}}$ is $\leq |\nabla_T u_h|$ where T is an element of $\mathcal{T}$ with vertices x_K, x_L.

Rq Actually we have shown that the translates of $u_h(\cdot)$ are estimated by const. Δ, which means (see e.g. [Brezis]) that $u \in BV(\omega)$; moreover, $u_h(\cdot) \longrightarrow u(\cdot)$ on $BV(\omega)$. For this result, an L^1 estimate of $\vec{\nabla}_h u_h(\cdot)$ would be enough.

- Consistency for test functions It remains to show that $\vec{w} = \vec{\nabla} u$, in the $W^{1,q}(\Omega)$ sense. This will be done in indirect way (through a weak formulation)

First we need a first consistency result.

Def. Proj For a scalar or vector function on $L^1(\Omega)$, one defines the projection operators

- $P_h : f \mapsto P_h f = \left(\int_K f(x)dx \right)_K \in \mathbb{R}^{\#K}$
- $\vec{P}_h : \vec{\Psi} \mapsto \vec{P}_h \vec{\Psi} = \left(\int_T \vec{\Psi}(x)dx \right)_T \in (\mathbb{R}^2)^{\#T}$

Proposition Consist 1 Let $\Psi \in C_0^1(\Omega)$. Assume that all $T \in \mathcal{T}_h$ are triangles, and moreover, $\cos d_{i,i+1} \geq \delta > 0$ uniformly (see Lemma Rec 3).

Then for all $V \in \mathbb{R}^{\#K}$,

$$\left| \left[\left[P_h'(dv\,\vec{\Psi}) - dv_h(\vec{P}_h \vec{\Psi}),\ V_h \right] \right]_h \right| \leq C(\Psi)\, h\, \| \vec{\nabla}_h V_h \|_{L^1(\Omega)},$$

where $h \simeq \max_{T \in \mathcal{T}_h} (\operatorname{diam} T)$ (the parameter h has the sense of the size of the mesh).

Rq In the framework of covolume schemes, the operators do not commute and moreover, the consistency property should be stated in the weak sense above.

Proof (sketched)

We introduce for all half-edge σ, the value $\vec{\Psi}_\sigma = \int_\sigma \vec{\Psi}(x)dx$ (in addition to the values $\vec{\Psi}_T = \int_T \vec{\Psi}(x)dx$ taken by $\vec{P}_h \vec{\Psi}$).

We can write $dv_K(\vec{P}_h \vec{\Psi}) = \frac{1}{m_K} \sum_{\sigma \subset \partial K} m_\sigma \vec{\Psi}_{T(\sigma)} \cdot \vec{n}_K$

and $P_K(dv\,\vec{\Psi}) = \frac{1}{m_K} \sum_{\sigma \subset \partial K} m_\sigma \vec{\Psi}_\sigma \cdot \vec{n}_K$.

Gathering by triangles T and proceeding as in the "discrete duality calculation" we find the expression for $\left[\left[P_h \circ dv\,\vec{\Psi} - dv_h \circ \vec{P}_h \vec{\Psi},\ V_h \right] \right]_h$:

$$\sum_T m_T \frac{2}{m_T} \left(\frac{m_{T_{KL}}}{\cos d_{KL}} \frac{V_L - V_K}{d_{KL}} \left(\vec{\Psi}_{d_{KL}} - \vec{\Psi}_T \right) + \dots \right).$$

But $|\vec{\Psi}_{d_{KL}} - \vec{\Psi}_T| \leq C \|\vec{\nabla}\Psi\|_\infty h$ and $\left| \frac{V_L - V_K}{d_{KL}} \right| \leq |\vec{\nabla}_T V_h|$ (because T is a triangle).

Using the lower bound on $\cos d_{i,i+1}$, we end up with the bound $\sum_T m_T\, Ch\, |\vec{\nabla}_T V_h|$, which ends the proof. □

Proposition Consist2 Let $\vec{\Psi} \in C(\Omega)$. Then $\|(\vec{P}_h \vec{\Psi})(\cdot) - \vec{\Psi}(\cdot)\|_{L^\infty(\Omega)} \le C h$.

Proof: Exercise.

- **Use of discrete duality** Now we are ready to conclude.

Let $\vec{\Psi} \in C_0^\infty(\Omega; \mathbb{R}^2)$. Set $\vec{\Psi}_h = \vec{P}_h \vec{\Psi}$; then by Proposition DD,

$$[\![-\mathrm{div}_h \vec{\Psi}_h, u_h]\!]_h = \{\!| \vec{\Psi}_h, \vec{\nabla}_h u_h |\!\}_h.$$

Moreover, by the commutation property of Prop. Consist1,

$$[\![-P_h(\mathrm{div}\vec{\Psi}), u_h]\!]_h = \{\!| \vec{\Psi}_h, \vec{\nabla}_h u_h |\!\}_h + \underset{h\to 0}{\overline{\overline{o}}}(1) \quad \text{(because } \|\vec{\nabla}_h u_h\|_{L^2(\Omega)} \le \text{const.!)}.$$

Rewriting this under the form of integrals, we get

$$-\sum_K m_K \frac{1}{m_K}\Big(\int_K (\mathrm{div}\vec{\Psi})(x)\,dx\Big) u_K = \sum_T m_T \frac{1}{m_T}\Big(\int_T \vec{\Psi}(x)\,dx\Big)\cdot \vec{\nabla}_T u_h + \underset{h\to 0}{\overline{\overline{o}}}(1), \text{ thus}$$

$$-\int_\Omega u_h(x)\,(\mathrm{div}\vec{\Psi})(x)\,dx = \int_\Omega \vec{\Psi}(x)\cdot(\vec{\nabla}_h u_h)(x)\,dx + \underset{h\to 0}{\overline{\overline{o}}}(1).$$

By the definition of $u(\cdot)$ and $\vec{w}(\cdot)$, we get at the limit

$$-\int_\Omega u\,\mathrm{div}\vec{\Psi} = \int_\Omega \vec{\Psi}\cdot\vec{w}, \quad \text{for all } \Psi \in C_0^\infty(\Omega).$$

Recall that $u \in L^2(\Omega)$, $\vec{w} \in L^p(\Omega)$. This means exactly that $u \in W^{1,p}(\Omega) \cap L^2(\Omega)$ and $\vec{w} = \vec{\nabla} u$ in the sense $W^{1,p}(\Omega)$. □

(IX) - Now we want to prove that (for any subsequence of $(u_h)_h$ convergent in the sense of Proposition Cmp, the limit u is a solution to (MP_1). This gives existence! Since uniqueness for (MP_1) is known, this implies that the whole sequence $(u_h)_h$ converges to the unique solution of (MP_1) [argument: by contradiction, extract a subsequence that remains ε-far from u; by Proposition Cmp, it admits a convergent subsequence and if we show that the limit solves (MP_1), we know that the limit is u; this is absurd].

Because our notion of solution is weak sol., we try to pass to the limit into the weak formulation (WEqD) of the discrete system (EqD).

Theorem G Let $\mathcal{T}_h, \mathcal{M}_h$ be a family of meshes: $\mathcal{T}_h$ a triangulation of maximal size h, $\mathcal{M}_h$ the associated mesh of the covolume method,

Assume the centers x_T of triangles $T \in \mathcal{T}_h$ are chosen so that $\mathrm{cos}d_{KL} \ge \delta > 0$ uniformly in h [this kind of uniform regularity assumption on meshes often appears].

Then for all h, there exists a unique solution u_h of the covolume scheme (DEq) for problem (MP_1). Moreover, the associated lifted functions $u_h(\cdot)$ and the associated lifted gradients $(\vec{\nabla}_h u_h)(\cdot)$ converge to $u, \vec{\nabla} u$ in the following sense:

- $u_h(\cdot) \longrightarrow u(\cdot)$ in $L^2(\Omega)$ strongly [even in $L^q(\Omega)$ strongly for $q < p^*$, if we can use optimal discrete Sobolev embedding]
- $(\vec{\nabla}_h u_h)(\cdot) \longrightarrow \vec{\nabla} u(\cdot)$ in $L^p(\Omega)$ weakly, and also strongly if $\vec{a}$ is strictly monotone

where u is the unique solution of (MP_1).

Remark The strong convergences are a byproduct of the proof: they do not follow from Prop. Cmp.
In fact, Prop. Cmp only uses boundedness of u_h is discrete L^2 and $W^{1,p}$ spaces, but Theorem G uses in addition the fact that u_h (resp., u) satisfy the discrete (resp., continuous) "differential relations"; thus we have an effect of the "compensated compactness" kind.

Proof We take $\varphi \in C^\infty(\bar\Omega)$ and use $\varphi_h = (\varphi(x_K))_K$ (not $P_h\varphi$!) as the test function in (wDEq). ($= P_h^C\varphi$ ← new projection.)

we have $[\![u_h - f_h, \varphi_h]\!]_h + \{\!\{ \vec a(\vec\nabla_h u_h), \vec\nabla_h \varphi_h \}\!\}_h = 0$. We will write this as integrals...

As in Prop. Consist2, we can say that $\varphi_h(\cdot) \to \varphi(\cdot)$ in $L^2(\Omega)$ strongly.

We need three more convergences: • $f_h(\cdot) \to f(\cdot)$ in $L^2(\Omega)$ weakly

• $\vec\nabla_h(P_h^C\varphi) \to \vec\nabla\varphi(\cdot)$ in $L^p(\Omega)$ strongly.

this part will be extremely delicate, because $\vec a$ is nonlinear! → [• $\vec a(\vec\nabla_h u_h)(\cdot) \to \vec a(\vec\nabla u)(\cdot)$ in $L^p(\Omega)$ weakly

Proposition Consist3 Let $f \in L^2(\Omega)$ and $f_h = P_h f$. Then $f_h(\cdot) \xrightarrow[h\to 0]{} f(\cdot)$ in $L^2(\Omega)$ strongly.

Proof From the Jensen inequality, we know that $\|f_h(\cdot)\|_{L^2} \le \|f(\cdot)\|_{L^2}$; in the abstract way, we can write that the linear operator $\mathcal{G}_h \circ P_h : L^2(\Omega) \to L^2(\Omega)$ is of norm ≤ 1.

Now, it is sufficient to approximate f by a sequence $(f^m)_m$ of $C(\Omega)$ functions.

By Prop. Consist2, we find $\|f_h^m(\cdot) - f^m(\cdot)\|_{L^2} \to 0$ as $h \to 0$.

Then $\|f_h(\cdot) - f(\cdot)\|_{L^2} \le \|f_h(\cdot) - f_h^m(\cdot)\|_{L^2} + \|f_h^m(\cdot) - f^m(\cdot)\|_{L^2} + \|f^m(\cdot) - f(\cdot)\|_{L^2} \le 2\,\overline{0}(1) + \overline{\overline{0}}^m(1)$ ($m\to\infty$, $h\to 0$; uniform in h!)

$\|\mathcal{G}_h \circ P_h (f - f^m)\|_{L^2} \le \|f - f^m\|_{L^2}$

Idea: we used abstract tools to avoid doing estimates "by hands": the price to pay is that we ignore the convergence rate

Taking $\varepsilon > 0$, then $m = m(\varepsilon)$, then $h = h(\varepsilon, m(\varepsilon))$, we conclude. □

Proposition Consist4 Let $\varphi \in C_0^2(\bar\Omega)$ and $\varphi_h = (\varphi(x_K))_K$. ($\varphi_h = P_h^C\varphi$ ← a new projection) Then $\|\vec\nabla_h \varphi_h(\cdot) - \vec\nabla\varphi(\cdot)\|_{L^\infty} \le C(\varphi, \delta)\, h$

where δ is the mesh regularity constant: $\cos\theta_{KL} \ge \delta > 0$, $\frac{d_{KL}}{\mathrm{diam}(T)} \ge \delta > 0$.

Proof We argue separately in each triangle T. Let w be the affine Taylor polynomial of φ at some fixed point of T. Then from the C^2 regularity of φ, because $\mathrm{diam}(T) \le h$, we get $\|\vec\nabla\varphi - \vec\nabla w\|_{L^\infty(T)} \le C(\varphi)\,h$ and $|\varphi_K - w_K|, |\varphi_L - w_L|, |\varphi_M - w_M| \le C(\varphi)(\mathrm{diam}\,T)^2$

Now, from Lemma Rec3 we know that $\vec\nabla_h$ is exact for affine functions, thus if we reconstruct $\vec\nabla_T w$ from w_K, w_L, w_M, we get $\vec\nabla_T w = \nabla w$ on T.

It remains to estimate $\|\vec\nabla_T \varphi_h - \vec\nabla_T w\|_\infty$; looking at the reconstruction formula, $= \vec\nabla_T(\varphi_h - w)$, because $\vec\nabla_T$ is linear

$$\vec\nabla_T(\varphi_h - w) = \frac{2}{m_T}\left\{ \frac{m_{T_{KL}}}{\cos\theta_{KL}} \frac{(\varphi_L - w_L) - (\varphi_K - w_K)}{d_{KL}} + \dots \right\}$$

($\frac{m_{T_{KL}}}{\cos\theta_{KL}} \le 1/\delta$)

$$\le C(\varphi)\frac{(\mathrm{diam}\,T)^2}{d_{KL}} \le C(\varphi)\frac{1}{\delta}\,\mathrm{diam}\,T \le C(\varphi,\delta)\,h.$$

here it is important that we take $\varphi_K = \varphi(x_K)$ and not $\frac{1}{|K|}\int_K \varphi(x)\,dx$, except if x_K is the barycenter of K

□

Remark Here we see that further proportionality assumptions on the meshes appear: the points x_K, x_L should not be too close with respect to the "typical length" in T.

This only forbids "too flat" triangles of the kind

(this restriction is much weaker than the "angle condition $\le 90°$" that was used for the maximum principle on Voronoi meshes)

Now, from the uniform bound of $\|\vec{\nabla}_h u_h(\cdot)\|_{L^p}$ and the growth assumption on $\vec{a}(\cdot)$, we get the uniform bound on $\|\vec{a}(\vec{\nabla}_h u_h)(\cdot)\|_{L^{p'}}$, and [extracting a further subsequence] we can say that $\vec{a}(\vec{\nabla}_h u_h)(\cdot) \longrightarrow \vec{\chi}(\cdot)$ in $L^{p'}(\Omega)$ as $h \to 0$.

Main question of the theory: why would we have $\vec{\chi} = \vec{a}(\vec{\nabla} u)$???

From the function-analytic arguments, nothing follows (or almost nothing: using the Young measures to describe weak convergence, we can establish some relation between $\vec{\chi}$ and ∇u, because both are calculated via the Young measure... we will see it later, for the $p(x)$-laplacian).

Therefore the only hope is that the PDE itself brings additional information. Here is the information:

Lemma Anti-weak With the notation above, we have $\int \vec{\chi} \cdot \vec{\nabla} u \geq \overline{\lim_{h \to 0}} \{\vec{a}(\vec{\nabla}_h u_h), \vec{\nabla}_h u_h\}_h$

UB. We have the inequality with $\overline{\lim}$, but $\lim$ would be enough.

- In the sequel we will show that the limit exists, and the equality holds (this is related to upgrading the weak convergence of $\vec{\nabla}_h u_h$ to the strong one).
- If $\vec{a} = \mathrm{Id}$, the inequality yields "$\|\vec{\nabla} u\|_{L^2} \geq \overline{\lim_{h \to 0}} \|(\vec{\nabla}_h u_h)(\cdot)\|_{L^2}$", which is known as a sufficient condition to upgrade the weak convergence "$\vec{\nabla}_h u_h \longrightarrow \vec{\nabla} u$ in L^2" into the strong L^2 convergence; in this sense, the inequality of the lemma "contradicts" weak (= non strong) convergence.

Proof The idea is: - take u for the test fct. in the limit of the scheme (involving $\vec{\chi}$)
- take u_h —"— —"— in (WDEq)
- pass to the limit where you can do so, and compare what remains.

So, we have proved so far that

$\forall \varphi \in C^\infty(\bar{\Omega})$ $\int u\varphi + \vec{\chi} \cdot \vec{\nabla}\varphi = \int f\varphi$; by the density of $C^\infty(\bar{\Omega})$ in $W^{1,p}(\Omega) \cap L^2(\Omega)$, we can take $\varphi = u$ for the test function;

($u \in L^2$, $\vec{\chi} \in L^{p'}$, $f \in L^2$)

then $\int_\Omega \vec{\chi} \cdot \vec{\nabla} u = \int_\Omega fu - \int_\Omega u^2$.

At the discrete level, we denote $\vec{\chi}_h := \vec{a}(\vec{\nabla}_h u_h)(\cdot)$ and from (WDEq) with $v_h = u_h$ we get $\int_\Omega \vec{\chi}_h(x) \cdot \vec{\nabla}_h u_h(x)\,dx = \int_\Omega f_h(x) u_h(x)\,dx - \int_\Omega u_h^2(x)\,dx$.

We can "almost" pass to the limit in the right-hand side:

- $\int u h = \lim_{h \to 0} \int f_h u_h$ by the properties ($f_h \to f$ in L^2 strongly (Prop. Consist 3); $u_h \to u$ in L^2 weakly (Prop. Cmp))
- $\int u^2 \leq \lim_{h \to 0} \int u_h^2$ by the weak convergence (or by the Fatou lemma; recall that we have shown $u_h \to u$ in $L^1(\Omega)$, which also gives "a.e." upon extraction of a further subseq.)

This is "fortunately" the good sign!

Comparing the two relations, we get the result. □

Now, it's time to exploit the monotonicity of $\vec{a}(\cdot)$ which is the key property for identification of $\vec{\chi}$ to $\vec{a}(\vec{\nabla} u)$.

We have the following "miracle" lemma (known as "Minty trick")

Lemma Minty Assume $(v_n)_n \subset V$ and $(w_n)_n \subset V^*$ (V a Banach space, V^* its dual, $\langle \cdot, \cdot \rangle$ the duality pairing) such that

- $v_n \to v$ in V weakly
- $w_n \overset{*}{\rightharpoonup} w$ in V^* weakly-$*$
- $w_n = T v_n$ where $T: V \to V^*$ is a monotone [nonlinear] operator

$$\forall v, \hat{v} \quad \langle T(v) - T(\hat{v}), v - \hat{v} \rangle \geq 0$$

and hemicontinuous: $T(v + \lambda \hat{v}) \overset{*}{\rightharpoonup} T(v)$ in V^* weakly-$*$ as $\lambda \to 0$, $\forall v, \hat{v} \in V$

Then $w = Tv$ provided we have $\langle w, v \rangle \geq \varlimsup_{n \to \infty} \langle w_n, v_n \rangle$.

Proof For $\lambda \in \mathbb{R}$ and $\xi \in V$, consider $\varphi := v - \lambda \xi$. We have

$$\langle w, v - \varphi \rangle \geq \varlimsup \langle w_n, v_n - \varphi \rangle = \varlimsup \langle T(\varphi) + (T(v_n) - T(\varphi)), v_n - \varphi \rangle$$

$$\overset{\text{assumption}}{\geq} \varlimsup \langle T(\varphi), v_n - \varphi \rangle = \varlimsup \langle T(v - \lambda\xi), (v_n - v) + \lambda \xi \rangle = \lambda \langle T(v - \lambda\xi), \xi \rangle$$

$\overset{\text{monotonicity of } T}{\geq} \lambda \langle w, \xi \rangle$.

Now,

- take $\lambda > 0$, divide by λ, send λ to zero

$$\Rightarrow \quad \langle w, \xi \rangle \overset{(\geq)}{} \varlimsup_{\lambda \downarrow 0} \langle T(v - \lambda\xi), \xi \rangle = \langle T(v), \xi \rangle$$

(↑ hemicontinuity of T)

- take $\lambda < 0$, divide by λ — change the sign "$\geq$" to "$\leq$"! — and get

$$\langle w, \xi \rangle \overset{(\leq)}{} \langle Tv, \xi \rangle.$$

Hence $\forall \xi \in V \quad \langle w, \xi \rangle = \langle T(v), \xi \rangle$, which yields $w = T(v)$. □

Announcement €100 to the person who convinces me that the proof is a natural one...

Application: $V = L^p(\Omega)$, $V^* = L^{p'}(\Omega)$, $T: v \longmapsto \vec{a}(\vec{\nabla} v)$ maps V into V^* thanks to the growth assumption.

Hemicontinuity comes from the dominated convergence theorem + growth assumption (exercise).

In conclusion, we establish $\overrightarrow{\chi} = \vec{a}(\vec{\nabla} u)$.

It remains to prove strong convergence claims.

Upgrading to strong convergence (if $\forall \xi, \eta \quad (\vec{a}(\xi) - \vec{a}(\eta)) \cdot (\xi - \eta) > 0$ whenever $\xi \neq \eta$)

Coming back to the calculation of Lemma Anti-weak, we get

$$\left(\int u^2 - \int u_h^2 \right) + \left(\int \vec{a}(\vec{\nabla} u) \cdot \vec{\nabla} u - \int \vec{a}(\vec{\nabla}_h u_h) \cdot \vec{\nabla}_h u_h \right) \xrightarrow[h \to 0]{} 0.$$

Notice that $\vec{a}(\vec{\nabla} u) \cdot \vec{\nabla} u - \vec{a}(\vec{\nabla}_h u_h) \cdot \vec{\nabla}_h u_h = \left(\vec{a}(\vec{\nabla} u) - \vec{a}(\vec{\nabla}_h u_h) \right) \cdot (\vec{\nabla} u - \vec{\nabla}_h u_h)$

$$+ \underbrace{\vec{a}(\vec{\nabla}_h u_h)}_{\overrightarrow{\chi_h}} \cdot \vec{\nabla} u - \underbrace{\vec{a}(au)}_{\overrightarrow{\chi}} \cdot \vec{\nabla}_h u_h$$

We have $\int \overrightarrow{\chi_h} \cdot \vec{\nabla} u \to \int \overrightarrow{\chi} \cdot \vec{\nabla} u$

$$\int \vec{a}(\vec{\nabla} u) \cdot \vec{\nabla}_h u_h \to \int \vec{a}(\vec{\nabla} u) \cdot \vec{\nabla} u = \int \overrightarrow{\chi} \cdot \vec{\nabla} u.$$

Hence $\underbrace{\varliminf_{h \to 0} \left(\int u^2 - \int u_h^2 \right)}_{\geq 0, \text{ by the weak convergence}} + \varliminf_{h \to 0} \int \underbrace{\left(\vec{a}(\vec{\nabla} u) - \vec{a}(\vec{\nabla}_h u_h) \right) \cdot (\vec{\nabla} u - \vec{\nabla}_h u_h)}_{\geq 0, \text{ by the monotonicity of } \vec{a}(\cdot)} \geq 0$

From the weak convergence of u_h to u in L^2 we know that $\|u\|_{L^2} \leq \varliminf \|u_h\|_{L^2}$; thus we deduce $\|u\|_{L^2} = \lim \|u_h\|_{L^2}$ (and the limit exists). Uniform convexity $\Rightarrow$ the cv is strong.

We also have from the convergence to zero in $L^1(\Omega)$ of the nonnegative function

$$\left(\vec{a}(\vec{\nabla}u) - \vec{a}(\vec{\nabla}_h u_h)\right)\cdot\left(\vec{\nabla}u - \vec{\nabla}_h u_h\right) \quad \text{(up to a subsequence)}.$$

Thus we also have the a.e. convergence to zero of this quantity.

Now we can reason pointwise: set $\vec{\xi} = \vec{\nabla}u(x)$ and $\vec{\eta}_h = (\vec{\nabla}_h u_h)(x)$.

If $\vec{\eta}_h \not\to \vec{\xi}$, we have $\left(\vec{a}(\vec{\xi}) - \vec{a}(\vec{\eta}_h)\right)\cdot(\vec{\xi} - \vec{\eta}_h) \not\to 0$ by the strong monotonicity of $\vec{a}(\cdot)$.

Thus $\vec{\nabla}_h u_h \to \vec{\nabla}u$ a.e. on Ω.

Then we have the following situation:

setting $F := \vec{a}(\vec{\nabla}u)\cdot\vec{\nabla}u$

$F_h := \vec{a}(\vec{\nabla}_h u_h)\cdot\vec{\nabla}_h u$,

$\left.\begin{array}{l} F_h \to F \text{ a.e. on } \Omega \\ F_h \geq 0 \\ \text{and moreover, } \int F = \lim \int F_h \end{array}\right\}$ Fatou lemma: $\int F \leq \underline{\lim} \int F_h$

(i.e. we are in the critical case of the Fatou lemma)

Lemma Sch (known as the Schaeffle lemma)

If in the assumptions of the Fatou lemma we have in addition $\int F = \lim \int F_h$, then $F \to F_h$ in L^1 strongly.

Proof $\lim \int |F - F_h| = \lim \left(\int (F - F_h)^+ + \int (F - F_h)^- \right)$, provided the limit exist.

The nice thing about $(F - F_h)^+$ is that it is dominated by F, because $F_h \geq 0$; therefore since $F \in L^1$, the dominated convergence can be used: since $F_h - F \to 0$, we get $\lim \int (F - F_h)^+ = 0$.

This trick cannot be used on $\int (F - F_h)^-$ (not dominated), but we use the additional assumption:

$0 = \lim \int (F - F_h) = \lim \left(\int (F - F_h)^+ - \int (F - F_h)^- \right) = 0 - \lim \int (F - F_h)^-$; the proof is finished □

Now we know that F_h converge in L^1, therefore it's

equi-integrable: $\forall E \subset \Omega$ (E measurable) $\quad \forall h \quad \int_E |F_h| \leq \omega(\text{meas}(E))$ where $\omega(t) \to 0$ as $t \to 0$.

We use the following refinement of the dominated convergence theorem:

Theorem (Vitali) Assume G_h are equi-integrable and $G_h \to G$ a.e.

Then $G \in L^1(\Omega)$ and $G_h \to G$ in $L^1(\Omega)$ strongly.

We use this theorem on the quantity $|\vec{\nabla}_h u_h|^p =: G_h$; it is equi-integrable because $\vec{a}(\vec{\xi})\cdot\vec{\xi} \geq \frac{1}{C}|\vec{\xi}|^p$ (coercivity) and $\underbrace{\vec{a}(\vec{\nabla}_h u_h)\cdot\vec{\nabla}_h u_h}_{= F_h}$ is equi-integrable.

Thus we have $\|(\vec{\nabla}_h u_h)(\cdot)\|_{L^p} \to \|\vec{\nabla}u\|_{L^p}$.

The norm convergence permits to upgrade the weak to the strong convergence.

NB The combination of the Minty trick with the Fatou-Schaeffle-equicint.-Vitaly arguments is called "the Minty-Browder argument".

In the sequel, we'll give a "more straightforward" version of the argument, using Young measures to characterize weak convergence.

Rq It follows that the energies converge: $\mathcal{J}_h[u_h] \to \mathcal{J}[u]$. Actually, Γ-convergence arguments could have been used for the variational case.

This ends the convergence proof. □

X Now let us look at a slightly different equation and at a different finite volume scheme.

p(x) laplacian with Dirichlet BC

Let $p : \Omega \mapsto [p_-, p_+] \subset (1, \infty)$ be a sufficiently regular (continuous, with log-Hölder modulus of continuity) variable exponent

Consider

$$(MP_2) \quad \begin{cases} -\operatorname{div}\left(\underbrace{|\vec{\nabla} u|^{p(x)-2}\vec{\nabla} u}_{\text{denoted } \vec{a}(x,\vec{\nabla}u)}\right) = f \\ u|_{\partial\Omega} = 0 \end{cases} \quad \text{with } f \in (L^{p_-})^* \text{ (for simplicity)}$$

Weak sol A weak solution of (MP_2) is $u \in W^{1,p(\cdot)}_0(\Omega)$ such that for all $v \in W^{1,p(\cdot)}_0(\Omega)$ ($C^\infty_c(\Omega)$ is enough) one has

$$(W_2) \quad \int_\Omega \vec{a}(x,\vec{\nabla}u)\cdot\vec{\nabla}v = \int_\Omega fv$$

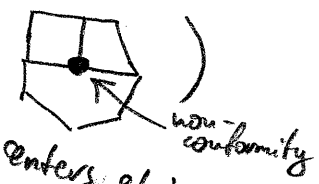
Optimization

$$u = \underset{v \in W^{1,p(\cdot)}_0(\Omega)}{\operatorname{argmin}} J[v], \quad J : v \mapsto \int_\Omega \frac{1}{p(x)}|\vec{\nabla} v|^{p(x)} dx - \int_\Omega fv$$

There exists a unique weak solution, which is the unique minimizer of the energy J.

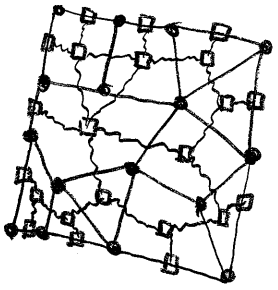
DDFV meshing and operators

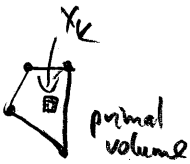
Take a possibly non-conformal unstructured mesh with centers located inside cells (barycenters, etc)



Create the dual mesh

(primal vertices ≡ dual centers
primal centers ≡ dual vertices)





NB At the boundary: put "degenerate primal volumes" that are flat

They are called boundary volumes, and no unknown is associated with them (Dirichlet BC is used instead).

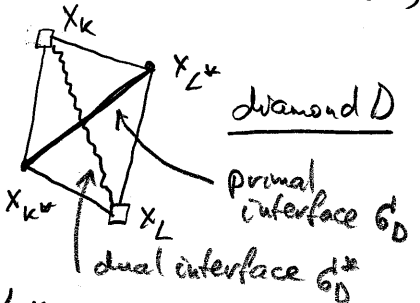
We also have boundary dual volumes, they have centers located on $\partial\Omega$; the Dirichlet BC is also used in these volumes.

Unknowns:

u_K at points x_K ◻ that are not on $\partial\Omega$

u_{K^*} at points x_{K^*} • that are not on $\partial\Omega$.

Diamonds: isolate interaction between two couples of neighbours and dual neighbours:



Discrete spaces: $\left((u_K)_K, (u_{K^*})_{K^*}\right) \in \mathbb{R}^{\#K} \times \mathbb{R}^{\#K^*}$ (denoted u_h)

$(\vec{F}_D)_D \in (\mathbb{R}^2)^{\#D}$ (denoted $\vec{F}_h$)

Scalar products: $[\![u_h, v_h]\!]_h := \frac{1}{2}\sum_K m_K u_K v_K + \frac{1}{2}\sum_{K^*} m_{K^*} u_{K^*} v_{K^*}$

$\{\!\{\vec{F}_h, \vec{G}_h\}\!\}_h := \sum_D m_D \vec{F}_D \cdot \vec{G}_D$

Discrete divergence: Standard FV approach. Eg. $\operatorname{div}_{K^*}\vec{F}_h := \frac{1}{m_{K^*}}\sum_{D : D\cap K^* \neq 0} m_{\sigma^*_D} \vec{F}_D \cdot \vec{n}_K$,

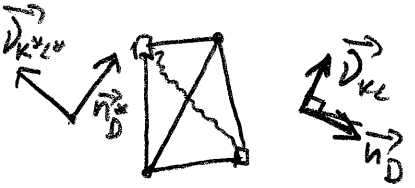
Discrete gradient This is the key idea (Herbin, Domelevo, Omnès).

We reconstruct the gradient per diamond (according to the definition of the space of discrete vectors) as follows: $\vec{\nabla}_D u$ = the unique vector of $\mathbb{R}^2$ s.t.

$$\begin{cases} \mathrm{Proj}_{\vec{D}_{KL}}(\vec{\nabla}_D u) = \dfrac{u_L - u_K}{d_{KL}} \vec{D}_{KL} \\ \mathrm{Proj}_{\vec{D}_{K^*L^*}}(\vec{\nabla}_D u) = \dfrac{u_{L^*} - u_{K^*}}{d_{K^*L^*}} \vec{D}_{K^*L^*} \end{cases}$$

Explicit formula: this was calculated in the proof of Lemma Rec 3:

$$(DGr) \quad \vec{\nabla}_D u = \frac{1}{\cos \alpha_D} \left(\frac{u_L - u_K}{d_{KL}} \vec{n}_D^* + \frac{u_{L^*} - u_{K^*}}{d_{K^*L^*}} \vec{n}_D \right)$$



α_D is the angle between $\vec{D}_{KL}$ and $\vec{D}_{K^*L^*}$.

Important particular case! $\alpha_D = 90°$ (orthogonal mesh)

In this case, we can prove the maximum principle for the scheme.

Discrete duality Holds true!!

$$[\![-\mathrm{div} \vec{F}_h, v_h]\!]_h = \{ \vec{F}_h, \vec{\nabla}_h v_h \}$$

for all $\vec{F}_h$ and for all v_h such that the boundary values of v_h are zero.

Scheme: pick a value p_D per diamond D

(e.g. $p_D = \fint_D p(x)\,dx$, or $p_D = p(x_D)$, $x_D \in D$, or $p_D = \sup_D p(\cdot)$ or $p_D = \inf_D p(\cdot)$)

Define $\vec{a}_D(\vec{\xi}) := |\vec{\xi}|^{p_D - 2} \vec{\xi}$, write $\vec{a}_h(\cdot)$ for $(\vec{a}_D(\cdot))_D$,

then
$$\begin{cases} -\mathrm{div}_h [\vec{a}_h(\vec{\nabla}_h u_h)] = f_h & f_h = P_h f \\ \text{with } u_h = 0 \text{ in boundary volumes.} \end{cases}$$

We can again write, thanks to the discrete duality,

$(WDEq_2)$ for all v_h which is zero in the boundary volumes, $\{ \vec{a}_h(\vec{\nabla}_h u_h), \vec{\nabla}_h v_h \}_h = [\![f_h, v_h]\!]_h$.

Uniqueness immediate (same proof as for (WDEq))

Variational nature immediate, for $J[v_h] = \sum_D m_D \dfrac{|\vec{\nabla}_h v_h|^{p_D}}{p_D}$.

Existence From the connection to the minimization problem (a priori estimates + topological degree are also ok)

A priori estimate: requires the discrete Poincaré inequality. Here we use the fixed $W_0^{1,p}$- discrete Poincaré (proof: Eymard, Gallouët, Herbin IMAJNA, Sushi scheme) and the assumption $f \in L^{(p_-)'}$

The proof may use translations (cf. the proof of Proposition Cmp) in L^{p_-}, but the Gallouët idea is much nicer (L^1 translates + the Nirenberg idea as in the continuous case)

The estimate obtained is: $\sum_D m_D |\vec{\nabla}_h u_h|^{p_D} \leq C(f)$ uniformly in h,

i.e., $\int_\Omega |(\vec{\nabla}_h u_h)(x)|^{p_h(x)} dx \leq C$.

Discrete compactness Uses again discrete Poincaré.

But it is essentially Prop. Cmp, provided we do it on $W_0^{1,p_-}(\Omega)$.

Thus: • $u_h(\cdot) \longrightarrow u(\cdot)$ strongly in L^1 (and in L^p, from discrete compact Sobolev embedding)

• $(\vec{\nabla}_h u_h)(\cdot) \longrightarrow \vec{\nabla} u(\cdot)$ weakly in $L^{p_-}(\Omega)$.

Consistency lemmas Roughly speaking, the same as for the covolume scheme

Proof of convergence The approach remains the same:

- write (WDEq_2) with a test function $\varphi_h = \left((\varphi(x_K))_K, (\varphi(x_{K^*}))_{K^*}\right)$ where φ is a test function on Ω
- pass to the limit using compactness; a "weak limit" $\vec{\chi}$ of $\vec{a}_h(\vec{\nabla}_h u_h)$ appears
- identify $\vec{\chi}(\cdot)$ to $\vec{a}(\cdot, \vec{\nabla} u(\cdot))$.

Two arguments: · creation of $\vec{\chi}$
· its identification
become more difficult.

Indeed, there seems to be no universal space on which all the discrete norms could be calculated! The exponent $p_h(\cdot)$ depends on h.
We can manage to put all the functions $(\vec{\nabla}_h u_h)(\cdot)$ into $L^{p(\cdot)}(\Omega)$ if we pick $p_D = \sup_D p(\cdot)$.
But then $\vec{a}_h(\vec{\nabla}_h u)(\cdot)$ is not in $L^{p'(\cdot)}(\Omega)$! So we loose duality
more exactly, the duality framework should move with h.

XI) A way out of the trouble is "to pull everything down to L^1 convergence"!
The key fact is the fundamental theorem for Young measures:

Theorem FTYM (J. Ball, –, N. Hungerbühler)

Assume Ω is a bounded domain, and $(V_n)_n$ is an equi-integrable sequence of functions with values in $\mathbb{R}^d$. (⟺ weakly compact in $L^1(\Omega)$ by the Dunford-Pettis criterion, see e.g. [Brézis])

Then, upon extraction of a subsequence, there exists a family $(\nu_x(\cdot))_{x\in\Omega}$ of probability measures on $\mathbb{R}^d$, measurable in the ad hoc sense, such that for all Carathéodory function F on $\Omega \times \mathbb{R}^d$,

int) $$\left[F(\cdot, V_n(\cdot)) \text{ equi-integrable}\right] \Rightarrow \left[\exists \lim_{n\to\infty} \int_\Omega F(x, v_n(x))dx = \int_\Omega \left(\int_{\mathbb{R}^d} F(x,\vec{\lambda}) d\nu_x(\vec{\lambda})\right) dx\right]$$

Moreover, we have

$$\left[\nu_x(\cdot) = \delta(\cdot - v(x)) \text{ a.e.}\right] \Leftrightarrow \left[V_n \text{ converge to } v \text{ in measure on } \Omega\right].$$

Finally, if • $V_n = (W_n, Z_n)$ and $(\nu_x(\cdot))_x$ is the Young measure for V_n
• $Z_n \to Z$ in measure (so that $(\delta(\cdot - z(x)))_x$ is the Young measure for Z_n)
then $\nu_x = \mu_x \otimes \delta(\cdot - z(x))$ where $(\mu_x(\cdot))_x$ is the Young measure for W_n.

Here is the way we use Theorem FTYM.

- $(\vec{\nabla}_h u_h)_h$ is a sequence bounded in $L^{p-}(\Omega)$, thus it is equi-integrable on Ω. (Exercice)
  We pass to a subsequence and get a Young measure $(\nu_x^{(\cdot)})_x$
- We first use (Int_{YM}) on $F(x,\vec{\xi}) = \vec{\xi}\, g(x)$ with $g \in L^\infty(\Omega)$. Equiintegrability is clear.
  The outcome is:
  $$\int_\Omega (\vec{\nabla}_h u_h)(x)\, g(x)\, dx = \int_\Omega g(x) \left(\int_{\mathbb{R}^2} \vec{\lambda}\, d\nu_x(\vec{\lambda}) \right) dx, \quad \forall g \in L^\infty(\Omega).$$
  This means that $\vec{\nabla}_h u_h$ weakly converse to $\int_{\mathbb{R}^2} \vec{\lambda}\, d\nu_x(\vec{\lambda})$ in $L^1(\Omega)$;
  because we already know that $\vec{\nabla}_h u_h \to \vec{\nabla} u$ in $L^{p-}(\Omega)$ (from compactness proposition),
  we get
  $$\boxed{(\vec{\nabla} u)(x) = \int_{\mathbb{R}^2} \vec{\lambda}\, d\nu_x(\vec{\lambda}) \quad \text{a.e. on } \Omega} \qquad (YEq1)$$
- Then we use Theorem FTYM on the couples $(\vec{\nabla}_h u_h, p_h)$ and use the fact that $p_h \to p$ on Ω (convergence in measure is enough)
  [If $p(\cdot)$ is continuous, any choice of p_D leads to the L^∞ convergence ...]
  Then the associated Young measure is $\nu_x(\cdot) \otimes \delta(\cdot - p(x))$.
  We would like to use (Int_{YM}) with
  $$F(x,\vec{\xi},p) = |\vec{\xi}|^{p-2}\vec{\xi}\, g(x), \quad g \in L^\infty(\Omega);$$
  but we have to check the equi-integrability of $F(\cdot, (\vec{\nabla}_h u_h)(\cdot), p_h(\cdot))$.
  This follows from the estimate $\int_\Omega |\vec{\nabla}_h u_h(x)|^{p_h(x)} dx \le \text{Const}$
  and the variable exponent Hölder inequality. The proof is a bit technical and left to a reader with motivation $\ge \delta > 0$.
  Then the outcome is, making g vary in $L^\infty(\Omega)$:
  $$\boxed{\vec{a}_h(\vec{\nabla}_h u_h)(\cdot) \to \vec{\chi}(\cdot) := \int_{\mathbb{R}^2} |\vec{\lambda}|^{p(x)-2} d\nu_x(\vec{\lambda}) \text{ in } L^1(\Omega) \text{ weakly}} \qquad (YEq2)$$

Two things remain: • identify $\vec{\chi}(\cdot)$ to $\vec{a}(\cdot, \vec{\nabla} u(\cdot))$
• prove that $u \in W_0^{1,p(\cdot)}(\Omega)$ (for the time being, we only have $u \in W_0^{1,p-}(\Omega)$)

- We prove that $\int_\Omega |\vec{\nabla} u|^{p(x)} dx < \infty$;
  then one can use the fact [p log-Hölder continuous] $\Rightarrow$ [$u \in W_0^{1,p}(\Omega)$ iff $u \in W_0^{1,1}(\Omega)$ and $\vec{\nabla} u \in L^{p(\cdot)}(\Omega)$]. a fortiori, $W_0^{1,p-}(\Omega)$ is ok here
  ⚠ log-Hölder condition is important here
  The estimate follows by using $F_k(\vec{\xi},p) = \begin{cases} |\vec{\xi}|^p, & |\vec{\xi}| \le k \\ k^p, & \text{otherwise} \end{cases}$ (x-independant)
  It is L^∞, thus $F_k(\vec{\nabla}_h u_h(\cdot), p_h(\cdot))$ are equi-integrable.
  Hence $\int_\Omega \int_{\mathbb{R}^2} F_k(\lambda, p(x))\, d\nu_x(\vec{\lambda}) = \lim_{h\to 0} \int_\Omega F_k(\vec{\nabla}_h u_h, p_h) \le \int_\Omega |\vec{\nabla}_h u_h|^{p_h} \le \text{const}$.
  But F_k is increasing in k; and the limit is $F(\vec{\xi},p) = |\vec{\xi}|^p$. Thus by Beppo Levi theorem,
  $\int_\Omega \int_{\mathbb{R}^2} |\vec{\lambda}|^{p(x)} d\nu_x(\vec{\lambda}) \le \text{const}$.

It remains to use the Jensen inequality for the probability measure $\nu_x(\cdot)$:

for a.e. x, $|\nabla u(x)|^{p(x)} = \left| \int_{\mathbb{R}^2} \vec{\lambda}\, d\nu_x(\vec{\lambda}) \right|^{p(x)} \leq \int_{\mathbb{R}^2} |\vec{\lambda}|^{p(x)} d\nu_x(\vec{\lambda})$.

Hence $\int_\Omega |\nabla u(x)|^{p(x)} dx \leq \text{const} < \infty$; we have therefore established that $u \in W_0^{1,p(\cdot)}(\Omega)$.

- Now we would like to prove the following inequality:

("DvRot") $$\int_\Omega \underbrace{\left(\int_{\mathbb{R}^2} |\vec{\lambda}|^{p(x)-2} \vec{\lambda}\, d\nu_x(\vec{\lambda}) \right)}_{=\vec{\chi}} \underbrace{\left(\int_{\mathbb{R}^2} \vec{\lambda}\, d\nu_x(\vec{\lambda}) \right)}_{=\vec{\nabla} u} dx \geq \int_\Omega \left(\int_{\mathbb{R}^2} |\vec{\lambda}|^{p(x)} d\nu_x(\vec{\lambda}) \right)$$

This is reminiscent of the inequality $\int_\Omega \vec{\chi} \cdot \vec{\nabla} u \geq \underline{\lim} \int_\Omega \vec{a}(\vec{\nabla}_h u) \cdot \vec{\nabla}_h u$ that was the starting point of the tricky Minty-Browder argument.

We have $\int_\Omega \vec{\chi} \cdot \vec{\nabla} u = \int_\Omega f u$ because we were able to pass to the limit in the $(WDEq_2)$ and then put $v = u$ (recall that $u \in W_0^{1,p(\cdot)}(\Omega)$ and $C_c^\infty(\Omega)$ is dense in $W_0^{1,p(\cdot)}(\Omega)$).

Moreover, $\int_\Omega f u = \lim_{n \to \infty} \int_\Omega f_n u_n = \lim \left\{ a_n(\vec{\nabla}_h u_n), \vec{\nabla}_h u_n \right\}_h \equiv \lim \int_\Omega |\vec{\nabla}_h u_n|^{p_n(x)} dx$.

Thus it would be sufficient to show that

$$\int_\Omega \int_{\mathbb{R}^2} |\vec{\lambda}|^{p(x)} d\nu_x(\vec{\lambda}) \leq \lim \int_\Omega |\vec{\nabla}_h u_n|^{p_n(x)} dx.$$

But we have just shown this (in a weaker form) in the previous argument!

Rq Why have we used truncations?
It's because we could not take directly $F(\vec{\xi}, p) = |\vec{\xi}|^p$:
the equi-integrability of $|\vec{\nabla}_h u_n(\cdot)|^{p_n(\cdot)}$ was missing, only an L^1 bound was available!

- To conclude, we show that ("DvRot") is a compactification relation; this is because $\vec{a}(x, \cdot) = |\cdot|^{p(x)-2} \cdot$ is monotone!

We calculate

$$\Delta := \int_\Omega \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \left(\vec{a}(x, \vec{\lambda}) - \vec{a}(x, \vec{\mu}) \right) \cdot (\vec{\lambda} - \vec{\mu})\, d\nu_x(\vec{\lambda}) d\nu_x(\vec{\mu}) =$$

$$= \int_\Omega \underbrace{\left(\int_{\mathbb{R}^d} d\nu_x(\vec{\mu}) \right)}_{=1} \left(\int_{\mathbb{R}^d} \vec{a}(x, \vec{\lambda}) \cdot \vec{\lambda}\, d\nu_x(\vec{\lambda}) \right) + \text{identical term with } \lambda, \mu \text{ exchanged}$$

$$- \int_\Omega \left(\int_{\mathbb{R}^d} \vec{a}(x, \vec{\lambda}) d\nu_x(\vec{\lambda}) \right) \left(\int_{\mathbb{R}^d} \vec{\mu}\, d\nu_x(\vec{\mu}) \right) + \text{identical term with } \lambda, \mu \text{ exchanged}$$

(here we can now substitute $\vec{\mu}$ by $\vec{\lambda}$)

$$\leq 0 \text{ from ("DvRot")}.$$

But $\Delta \geq 0$ because the integrand is ≥ 0 by monotonicity of $\vec{a}$! Thus $\Delta = 0$, moreover, the integrand is zero $d\nu_x(\vec{\lambda}) d\nu_x(\vec{\mu}) dx$ - a.e.

By the strict monotonicity of $\vec{a}(x,\cdot)$, this means that the support of $d\nu_x(\cdot)$ should be reduced to one point (exercise). This means, for a.e. $x\in\Omega$ $\nu_x(\cdot)$ is a Dirac measure concentrated at some $\vec{V}(x)$. By Theorem RTYM, we deduck that $\vec{\nabla}_h u_h(\cdot) \to \vec{V}(\cdot)$ in measure, and also weakly in $L^{p(\cdot)}(\Omega)$. From the Proposition on compactness, we already know that $\vec{\nabla}_h u_h(\cdot) \to \vec{\nabla} u(\cdot)$ in $L^1(\Omega)$.

We conclude that $\nu_x(\cdot) = \delta(\cdot - (\vec{\nabla}u)(x))$ a.e. on Ω.

Now from the representation formula for $\vec{\chi}$,

$$\vec{\chi}(x) = \int_{\mathbb{R}^2} |\vec{\lambda}|^{p(x)-2}\vec{\lambda}\, d\delta(\vec{\lambda} - \vec{\nabla}u(x)) = |\vec{\nabla}u(x)|^{p(x)-2}\vec{\nabla}u(x).$$

The identification is complete! We conclude that $u = \lim_h u_h$ solves (MP_2). ∎

In addition, upon extraction of a subsequence, $\vec{\nabla}_h u_h(\cdot) \to \vec{\nabla}u$ a.e. Then we can upgrade the weak convergence a bit; we get in particular

$$\int_\Omega |\vec{\nabla}_h u_h|^{p_h} \to \int_\Omega |\vec{\nabla}u|^p$$ (convergence of the energy $\mathcal{J}_h[u_h]$ to $\mathcal{J}[u]$ follows).

NB We cannot state that $\vec{\nabla}_h u_h$ converge to $\vec{\nabla}u$ in $L^{p(\cdot)}(\Omega)$ because $p_h(\cdot)$ is different from $p(\cdot)$; but, interpolating the uniform bound on $\int |\vec{\nabla}_h u_h|^{p_h}$ with the a.e. convergence of $\vec{\nabla}_h u_h(\cdot)$ to $\vec{\nabla}u(\cdot)$, we get strong convergence in $L^q(\Omega)$ with $q<p$ (recall that $p_h \to p$ in the L^∞ norm).

(XII) Let's give essential ideas for a priori error estimates.

Properties of $\vec{a}: \vec{\xi} \to |\vec{\xi}|^{p-2}\vec{\xi}$ One has to distinguish between $p>2$ and $1<p<2$.

① $\begin{cases} |\vec{a}(\vec{\xi}) - \vec{a}(\vec{\eta})| \le C|\vec{\xi}-\vec{\eta}|^{p-1}, & 1<p\le 2 : \vec{a}(\cdot) \text{ is } C^{0,\alpha} \text{ Hölder}, \ \alpha = p-1 \\ |\vec{a}(\vec{\xi}) - \vec{a}(\vec{\eta})| \le C|\vec{\xi}-\vec{\eta}|\left(|\vec{\xi}|^p + |\vec{\eta}|^p\right)^{\frac{p-2}{p}}, & p\ge 2 : \vec{a}(\cdot) \text{ is locally Lipschitz} \end{cases}$

Remark: If we denote $\vec{\Xi} = |\vec{\xi}|^{p-2}\vec{\xi}$, then $\vec{\xi} = |\vec{\Xi}|^{p'-2}\vec{\Xi}$ (exercise)

Therefore the inverse of $\vec{a}_p(\cdot)$ is $\vec{a}_{p'}(\cdot)$.

Consequence: $1<p\le 2$, $\vec{a}^{-1}(\cdot)$ is locally Lipschitz

$p\ge 2$, $\vec{a}^{-1}(\cdot)$ is $C^{0,\beta}$ Hölder, $\beta = p'-1 = \frac{1}{p-1}$

The Hölder continuity results are the basis of most straightforward error estimates:
roughly speaking, from the order of consistency (e.g., h) we first extract an estimate on consistency of fluxes $\vec{F} = \vec{a}(\cdot)$; then it becomes of order h^α, $\alpha = \begin{cases} p-1, & p\le 2 \\ 1, & p\ge 2 \end{cases}$.
We get an estimate on $|\vec{a}(\vec{\nabla}_h u_h) - \vec{a}(\vec{\nabla}_h \Pi_h u)|$ in h^α, and inverting $\vec{a}(\cdot)$, we deduce for $|\vec{\nabla}_h u_h - \vec{\nabla}_h \Pi_h u|$ in $(h^\alpha)^\beta$ where $\beta = \begin{cases} 1, & p\le 2 \\ \frac{1}{p-1}, & p\ge 2 \end{cases}$. Finally, the convergence order is $h^{\min\{p-1, \frac{1}{p-1}\}}$.

② To be precise, instead of the Hölder continuity of $\vec{a}^{-1}(\cdot)$ we rather use

$$\begin{cases} (\vec{a}(\vec{\xi})-\vec{a}(\vec{\eta}))\cdot(\vec{\xi}-\vec{\eta}) \geq C|\vec{\xi}-\vec{\eta}|^2/(|\vec{\xi}|+|\vec{\eta}|)^{\frac{p-2}{p}} & , 1<p\leq 2 \\ (\vec{a}(\vec{\xi})-\vec{a}(\vec{\eta}))\cdot(\vec{\xi}-\vec{\eta}) \geq C|\vec{\xi}-\vec{\eta}|^p & , p\geq 2. \end{cases}$$

③ Everything comes from the following general inequalities:

$$\forall p\geq 1 \quad \forall \delta>0 \quad \exists C_1, C_2 \text{ depending on } p,\delta \text{ s.t.}$$

$$|\vec{a}(\vec{\xi})-\vec{a}(\vec{\eta})| \leq C_1|\vec{\xi}-\vec{\eta}|^{1-\delta}\left(|\vec{\xi}|+|\vec{\eta}|\right)^{p-2+\delta}$$

$$(\vec{a}(\vec{\xi})-\vec{a}(\vec{\eta}))\cdot(\vec{\xi}-\vec{\eta}) \geq C_2|\vec{\xi}-\vec{\eta}|^{2+\delta}\left(|\vec{\xi}|+|\vec{\eta}|\right)^{p-2-\delta}$$

These inequalities allow to improve the Hölder continuity exponents when the solutions have gradients far enough from zero or infinity, thus to improve error estimates for solutions for which some additional information is available.

Reference Barrett, Liu '1993.

How to get consistency orders? A typical estimate that is needed looks as follows:

(on covolume mesh) compare $\int_\sigma \vec{a}(\vec{\nabla}u_e) - \vec{a}\left(\vec{\nabla}_{T(\sigma)} I_h u_e\right)$,

where u_e is the exact solution.

Using Hölder continuity of $\vec{a}$, we reduce the question to comparison of $(\vec{\nabla}u_e)(x)$ for $x\in\sigma$ and of $\vec{\nabla}_{T(\sigma)} I_h u_e$. Recall that $I_h u_e = (\int_K u_e)_K$ or $(u_e(x_K))_K$ (if $u_e\in C(\Omega)$!)

- If one assumes $u_e \in C^{s\text{thg}}$, one can use Taylor expansions to prove that for C^2 solutions, the difference is order h. Moreover, on meshes possessing symmetry properties and in case u_e is smooth and $\vec{a}$ is smooth ($p\geq s\text{thg}\geq 2$ or $p=2$), summing on edges we observe compensations that may lead to better convergence orders (cancellation in Taylor expansions).
- If one assumes $u_e\in W^{2,p}$, which seems more realistic, then the essential tool is the so-called Hadamard formula: say if $\begin{cases} x \text{ runs over } K \\ y \text{ runs over } L \end{cases}$

$$u_L-u_K = \int_L u_e(x)dx - \int_K u_e(y)dy = \frac{1}{m_K m_L}\int_K\int_L (u_e(x)-u_e(y))\,dxdy$$

$$= \frac{1}{m_K m_L}\int_K\int_L \left(\int_0^1 \vec{\nabla}u_e\underbrace{(tx+(1-t)y)}_{z}\,dt\right)\cdot(x-y)\,dxdy$$

Then a change of variables $(x,y,t) \leftrightsquigarrow (x,z,t)$ brings an estimate in terms of $\int_{K\cup L}\vec{\nabla}u_e(z)dz$.

We end up with $\|\vec{\nabla}_h u_h(u)-\vec{\nabla}u_e\|_{L^p} \leq C\,h^{\min\{p-1,\frac{1}{p-1}\}}$.

Rq
- If we only assume the minimal ($W^{1,p}(\Omega)$) regularity of u_e, no convergence order can be obtained: we have shown convergence, but the rate of convergence can be as slow as desired.
- Assuming $u_e\in W^{2,p}(\Omega)$ is nice (standard in the context of the laplacian) but we can prove it only for $p\leq 2$ (with $f\in L^p(\Omega)$, which is more restrictive than $f\in L^{p'}$ for which we have shown convergence).

Rq • For $p>2$, no condition on f is known to ensure $u \in W^{2,p}(\Omega)$, so, the result is not satisfactory.

• Optimal regularity for $p>2$ and $f \in L^p$ was proved by Y. Simon '81: this is the so-called Besov $B^{1+\frac{1}{p-1},p}_{\infty}(\Omega)$ regularity, in case the solution admits an extension outside Ω (on a rectangle with Dirichlet homogeneous BC, we achieve such extension by reflexions and periodization).

• How to use this regularity? The approach described above would lead to $\left(\left(h^{\frac{1}{p-1}}\right)^{\alpha}\right)^{\beta}$ convergence order, i.e., $h^{\frac{1}{(p-1)^2}}$ (for $p>2$).

• There is a very smart approach due to Tyukhtin '82, popularized by S.S. Chow '89: it uses the minimization properties for both discrete and continuous energies, and the outcome is the $h^{\frac{2}{(p-1)^2}}$ convergence order.

• Numerically, we observe that for u_e which has precisely the $B^{1+\frac{1}{p-1},p}_{\infty}$ regularity (and not more), the order $h^{\frac{2}{(p-1)^2}}$ is what happens in practice. Thus, this estimate is optimal!

Some details on error estimates are given in the "Annexes" pages (in French) at the end of this document.

References

• Existence for p-laplacian & Co:
- J.L. Lions, 1969 "Quelques problèmes aux limites non linéaires", Chapter II. Dunod, Paris

• Finite element approximation of the p-laplacian:
- S.S. Chow, 1989, Numerische Mathematik
- J.W. Barrett, W.B. Liu, 1993, J. Math. Anal. Appl.
- " " " " 1993, Math. Comp.

• Covolume Finite Volume schemes
- Walkington 1996, SIAM J. Numer. Anal.
- Afif, Amaziane 2002, Comput. Meth. Appl. Mech. Eng.
- Handlovičová, Mikula, Sgallari 2003, Numer. Math.
- Handlovičová, Mikula 2008, Appl. Math.

• DDFV Finite Volume schemes
- Hermeline 1998, CRAS; 2000, J. Comput. Phys.; 2009, Comp. Meth. Appl. Mech. Eng.
- Domelevo, Omnès 2005, M2AN
- Andreianov, Boyer, Hubert 2007, Num. Meth. PDE
- Andreianov, Bendahmane, Hubert, preprint to appear on HAL

• Schemes on cartesian meshes and error estimates
- Andreianov, Boyer, Hubert, 2004, M2AN; 2005, Numer. Math.; 2006, IMA J. Num. Anal.

• Other FV schemes
- Eymard, Gallouët, Herbin "Finite Volume methods" 2000, personal page of R. Herbin
- Coudière, Vila, Villedieu 1999, M2AN
- Andreianov, Gutnic, Wittbold 2004, SIAM J. Num. Anal.
- Droniou, 2006, M2AN
- Eymard, Gallouët, Herbin 2009, IMA J. Num. Anal.
- Eymard, Gallouët, Herbin HAL preprint, January 2009

• Mimetic schemes
Shashkov, Lipnikov, Brezzi, Manzini 2005-...

Méthodes de volumes finis pour les problèmes elliptiques non linéaires

B. Andreianov (Besançon) & F. Boyer, F. Hubert (Marseille)

Problème à discrétiser : $\begin{cases} -\operatorname{div} \varphi(x, \nabla u_e) = f(x), \quad x=(x,y) \in \Omega \subset \mathbb{R}^2 \\ + \text{CB sur } u_e \text{ (Dirichlet homogène / inhomogène)} \end{cases}$ u_e : solution exacte

Prototype : le p-laplacien, avec $\varphi(x,\xi) = |\xi|^{p-2}\xi$, $1<p<+\infty$.

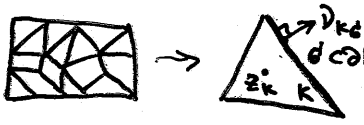
Propriétés structurelles de φ :

pour existence, unicité, convergence :
- monotonie de φ en ξ
- coercivité : $\varphi(\xi)\cdot\xi \geq c|\xi|^p$
- hypothèse de croissance tq $u \mapsto -\operatorname{div}\varphi(\nabla u)$ agisse de $W_0^{1,p}$ dans $W^{-1,p'}$
- (cas particulier) structure variationnelle : $\varphi(\xi) = \nabla \Phi(\xi)$

pour l'analyse quantitative de l'erreur :
- $\varphi : \xi \to \varphi(\xi)$ est (localement) Hölderienne d'ordre $\begin{cases} 1, & p \geq 2 \\ p-1, & 1<p\leq 2 \end{cases}$
- φ^{-1} est (localement) Hölderienne d'ordre $\begin{cases} \frac{1}{p-1}, & p \geq 2 \\ 1, & 1<p\leq 2 \end{cases}$.

Méthodes de volumes finis : Pour le problème linéaire sur maillages conformes :
cf. Eymard, Gallouët, Herbin "Finite Volume Methods" '05

Principe :
- Partitionner Ω en volumes de contrôle, notés K ($K \in \mathcal{T}$)



- On cherche une "solution discrète", notée $u^{\mathcal{T}} = (u_K)_{K\in\mathcal{T}}$, constante par volume
- On "intègre" l'équation sur chaque volume :

$$\forall K \in \mathcal{T}, \quad a_K[u^{\mathcal{T}}] = \sum_{\text{arêtes } \sigma \subset \partial K} \int_\sigma (\varphi \cdot \nu_{K\sigma})^{\mathcal{T}}[u^{\mathcal{T}}] = m_K f_K$$

- On fournit $(\varphi\cdot\nu_{K\sigma})^{\mathcal{T}}[u^{\mathcal{T}}]$, qui représente une reconstruction du flux normal discret via l'arête σ, à partir des valeurs $(u_K)_{K\in\mathcal{T}}$.

Cas linéaire : Si on peut munir chaque volume K d'un centre x_K tq $x_K x_L \perp \sigma$, ($\varphi(x,\xi) = k(x)\xi$) où $\sigma = \partial K \cap \partial L$ (i.e. maillage conforme), alors $\frac{u_L - u_K}{|x_L - x_K|}$ fournit la reconstruction "naturelle" de $(\varphi\cdot\nu_{K\sigma})^{\mathcal{T}}[u^{\mathcal{T}}]$.

Cette reconstruction préserve, au niveau discret, les propriétés structurelles du problème continu (citées ci-dessus).

Objet de l'exposé :
- présenter les schémas VF pour le problème nonlinéaire qui préservent également ces propriétés structurelles
- analyser leur convergence (en se servant desdites propriétés).

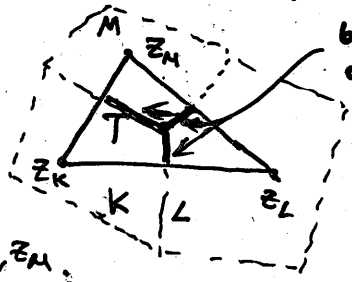
Schémas disponibles

- **B.A. & M.Gutnic, P.Wittbold SIAM JNA'04**

  marche par reconstruction affine/morceaux sur des maillages duaux aux maillages triangulaires.

  Dans chaque triangle $T = z_K z_L z_M$ on prend sur $\sigma \cap T$, $(\varphi \cdot \nu_{K\sigma})^{\mathcal{T}}[u^{\mathcal{T}}] = \varphi(\nabla_T u^{\mathcal{T}}) \cdot \nu_{K\sigma}$

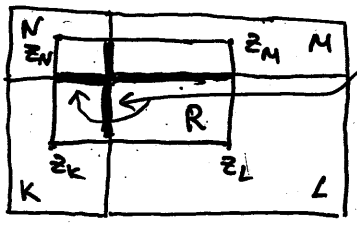
  où $\nabla_T u^{\mathcal{T}}$ est l'interpolation affine des valeurs (celle de u_K, u_L, u_M aux points z_K, z_L, z_M).



- **B.A. & F.Boyer, F.Hubert M2AN'04 ; Num. Math.'05 ; IMAJNA'06?**

  marche pour les problèmes variationnels sur maillages cartésiens

  Dans chaque rectangle $R = z_K z_L z_M z_N$ on prend pour $|\nabla_R u^{\mathcal{T}}|$ une valeur reconstruite à partir de u_K, u_L, u_M, u_N. Il n'y a pas de reconstruction "naturelle", mais la structure rigide (on veut que le schéma discret dérive d'un potentiel) fait qu'il n'y a qu'une famille à un paramètre de reconstructions possibles de $|\nabla_R u^{\mathcal{T}}|$ (les formules sont compliquées, sauf pour maillages uniformes). On observe que l'équation d'Euler-Lagrange pour minimisation de la fonctionnelle discrète $\mathcal{J}_{\mathcal{T}}[u^{\mathcal{T}}] = \sum_R m_R \Phi(|\nabla_R u^{\mathcal{T}}|)$ est exactement un schéma de volumes finis.



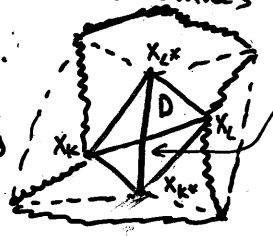
- **B.A & F.Boyer, F.Hubert Num.PDE'06? , proc. FVCA-4'05**

  marche sur maillages "en duplex" : deux maillages recouvrent, chacun, le domaine, on a deux jeux d'inconnues et d'équations discrètes.

  Gain : approximation du gradient discret simple et très efficace, structure de maillages très générale (raffinement local etc.)

  Idée : Coudière-Vila-Villedieu / Hermeline / Domélévo-Omnès '99/2000/'05

  On partitionne Ω en "diamants" $D = z_K z_{K^*} z_L z_{L^*}$. Les points $(z_K)_K$ et $(z_{K^*})_{K^*}$ engendrent deux maillages de Ω. On prend le double jeu d'inconnues $u^{\mathcal{T}} = \{(u_K)_K, (u_{K^*})_{K^*}\}$ et on écrit le schéma de volumes finis (une équation pour chaque K et une pour chaque K^*). Dans chaque diamant D, on reconstruit un vecteur $\nabla_D u^{\mathcal{T}}$ et on prend



$$(\varphi \cdot \nu_{K\sigma})^{\mathcal{T}}[u^{\mathcal{T}}] = \varphi(\nabla_D u^{\mathcal{T}}) \cdot \nu_{K\sigma}$$

  et idem pour K^*, σ^*.

- **J.Droniou '06? (cf. le cas linéaire Eymard, Droniou '05?)**

  marche sur maillages presque arbitraires, grâce à la relaxation du lien flux-gradient. On garde les flux φ_σ via chaque arête comme inconnues et on les relie aux inconnues $u^{\mathcal{T}}$ par une relation bien choisie qui complète le système discret.

Dans tous les cas, la structure est héritée du pb. continu grâce à la

dualité des opérateurs -div et grad discrets.

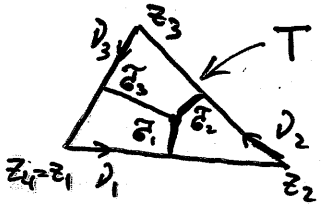
Exemple (1er schéma) en continu : $\boxed{\int_\Omega -\operatorname{div}\varphi(\nabla u_\varepsilon)\,v = \int_\Omega \varphi(\nabla u_\varepsilon)\cdot\nabla v} \quad \forall v \in W_0^{1,p}$.

• en discret : v^τ nulle sur les mailles de bord

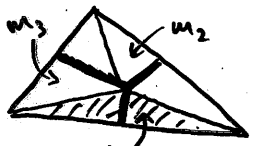
$$\boxed{\sum_k a_k[u^\tau]\,v^\tau = \sum_K \sum_{\sigma\in\partial K} \int_\sigma (\varphi\cdot\nu_{K\sigma})^\tau[u^\tau]\,v^\tau} = \left\{\begin{array}{l}\text{on regroupe les termes}\\ \text{par demi-arête } \tilde\sigma\\ \text{dans } T \text{ et on fait}\\ \text{la sommation par}\\ \text{parties}\end{array}\right.$$

$$= \sum_T \sum_{i=1}^3 m_{\sigma_i}\,\varphi(\nabla_T u^\tau)\cdot\nu_i\,(v_{i+1}-v_i) =$$

$$= \sum_T \varphi(\nabla_T u^\tau)\cdot \sum_{i=1}^3 m_{\sigma_i} d_i\,(\nabla_T v^\tau\cdot\nu_i)\,\nu_i = \boxed{\sum_T m_T\,\varphi(\nabla_T u^\tau)\cdot\nabla_T v^\tau,}$$



où la dernière inégalité vient du calcul suivant :



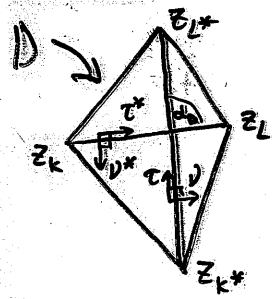
Lemme $\dfrac{2}{m_T}\sum_{i=1}^3 m_i\,(\xi\cdot\nu_i)\,\nu_i = \xi \quad \forall \xi\in\mathbb{R}^2$.

De la dualité, on déduit par ex. la coercivité :

$$\sum_k a_k[v^\tau]\,v^\tau = \sum_T m_T\,\varphi(\nabla_T v^\tau)\cdot\nabla v^\tau \geq C\sum_T |\nabla_T v^\tau|^p,$$

mais aussi la monotonie, les dépendances Höldériennes.

Exemple (schémas duplex) Dans chaque diamant D, on reconstruit le gradient :



$$\boxed{\nabla_D u^\tau = \frac{1}{\sin\theta_D}\left(\frac{u_L-u_K}{|z_L-z_K|}\,\nu + \frac{u_{L^*}-u_{K^*}}{|z_{L^*}-z_{K^*}|}\,\nu^*\right)}$$

Lemme $\dfrac{1}{\sin\theta_D}\left((\xi\cdot\tau^*)\,\nu + (\xi\cdot\tau)\,\nu^*\right) = \xi \quad \forall \xi\in\mathbb{R}^2$.

Corollaire (div-grad dualité discrète)

On a $\frac12\sum_K a_K[u^\tau]\,v_K + \frac12\sum_{K^*} a_{K^*}[u^\tau]\,v_{K^*} = \sum_D m_D\,\varphi(\nabla_D u^\tau)\cdot\nabla_D v^\tau$,

pour toutes u^τ, v^τ fonctions duplex discrètes avec v^τ nulle dans les volumes de bord.

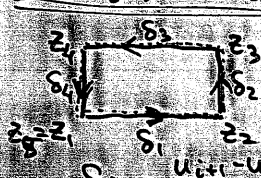
Propriétés du schéma duplex ||
- inégalités de Poincaré, injections, interpolations discrètes ;
- existence et unicité de solution discrète, estimation $W^{1,p,\tau}$ à priori ;
- si φ dérive d'un potentiel Φ, le problème discret garde la structure variationnelle : $a_K[u^\tau] = \dfrac{\partial}{\partial u_K}\sum_D m_D\,\Phi(\nabla_D u^\tau)$.

Ces propriétés sont quantifiées par une constante reg(h) (où h = taille du maillage) ; reg(h) est censée rester bornée lorsque $h\to 0$.

Schémas sur maillages cartésiens || Si maillage uniforme : $|\nabla_K u^\tau|^2 = \frac12\sum_{i=1}^4 \delta_i^2 + \alpha(\delta_1+\delta_3)^2$,

où α (un paramètre de "torsion") peut varier dans certaines limites. En général, on obtient une formule (avec paramètre α) via les propositions de R.

Ces schémas ont aussi la propriété de dualité div-grad.



$\delta_i = \dfrac{u_{i+1}-u_i}{|z_{i+1}-z_i|}$

Conclusion Pour les schémas "duplex" (les mêmes résultats sont/peuvent être obtenus sur maillages cartésiens et sur maillages duaux aux triangulaires)

- convergence, pour tout terme source $f \in L^{p'}$ (et $W^{-1,p'}$ si on fait entrer la partie singulière dans le flux)
- estimations d'erreur en normes $W^{1,p,\tau}$ et L^p en $\begin{cases} h^{\frac{1}{p-1}}, & p \geq 2 \\ h^{p-1}, & 1<p\leq 2 \end{cases}$ à condition que f donne lieu à une solution $u_e \in W^{2,p}$.

→ Numériquement, ces taux sont pessimistes (en particulier, les ordres de convergence obtenus tendent vers zéro lorsque $p \to +\infty$ ou $p \to 1$).

→ Sauf pour $1<p\leq 2$, on ne dispose pas de résultats de régularité déjà pour le p-laplacien.

On va alors chercher à améliorer ces résultats en deux directions

① → Tant qu'à faire, on suppose la régularité aussi élevée que nécessaire pour avoir un meilleur taux de convergence : B.A, F.Boyer, F.Hubert INAWA'06

② → On considère les solutions génériques mais on n'utilise que leur régularité naturelle : $u_e \in \begin{cases} B^{1+\frac{1}{p-1},p}_{\infty}, & p>2 \\ W^{2,p}, & 1<p<2 \end{cases}$ (espace de Besov)

B.A, F.Boyer, F.Hubert Num.Math'05

Idée de la preuve d'est. d'erreur de base

Le schéma : $a_K[u^\tau] = m_K f_K$

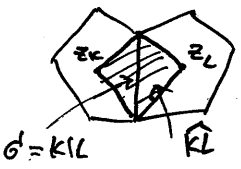
L'éq. continue : $\sum_{\sigma \in \partial K} \int_\sigma \varphi(\nabla u_e)\cdot \nu_{K\sigma} = m_K f_K$

On retranche et multiplie par $u^\tau - u_e^\tau$:

$$\sum_K \left(a_K[u^\tau] - a_K[u_e^\tau]\right)(u_K - u_{eK}) = \sum_K \sum_{\sigma\in\partial K} m_\sigma (\Delta f)_\sigma (u_K - u_{eK})$$

$$= \sum_{\sigma = K|L} m_\sigma |z_L - z_K| (\Delta f)_\sigma \frac{(u_K - u_{eK}) - (u_L - u_{eL})}{|z_K - z_L|}$$

ici on retrouve la mesure de la partie de Ω entourant l'arête $\sigma = K|L$ qui sépare K et L ; on la note $m_{\widehat{KL}}$



Ainsi, tout se passe comme si $\Delta f := \sum_{K|L} (\Delta f)_{K|L} \mathbb{1}_{\widehat{KL}}$ était une perturbation du terme source.

$$a_K[u_e^\tau] + \sum_{\sigma\in\partial K} m_\sigma (\Delta f)_\sigma$$

ou le note $(\Delta f)_\sigma^\tau$

$$= a_K[u_e^\tau] + \sum_{\sigma\in\partial K} m_\sigma \int_\sigma \varphi(\nabla u_e)\cdot \nu_{K\sigma} - (\varphi\cdot\nu)_{K\sigma}[u_e^\tau]$$

↑ ça joue le rôle de perturbation du terme source f

Or, on a :

- $|\nabla^\tau u_e^\tau - \nabla u_e| \sim h$ lorsque u_e a deux dérivées
- la continuité Hölder de φ fait que $\|\Delta f\|_{L^{p'}} \sim \begin{cases} h, & p\geq 2 \\ h^{p-1}, & 1<p\leq 2 \end{cases}$
- la continuité Hölder de φ^{-1} fait que

$$\|\nabla^\tau e^\tau\|_{L^p} \sim \begin{cases} \|\Delta f\|^{\frac{1}{p-1}}, & p\geq 2 \\ \|\Delta f\|, & 1<p\leq 2 \end{cases} \sim \begin{cases} (h^1)^{\frac{1}{p-1}}, & p\geq 2 \\ (h^{p-1})^1, & 1<p\leq 2. \end{cases}$$

Améliorations

① Sur maillages cartésiens uniformes, on a $(\Delta f)_\sigma \sim h$, pour $p \geq 2$, mais $\sum_{\sigma \subset K} m_\sigma^K (\Delta f)_\sigma \sim h^4$ pour $p \geq 4$ (ou $p=2$) et des solutions $u_e \in W^{4,q}_{\text{pch}}$

Alors $\sum_K \sum_{\sigma \subset K} m_\sigma^K (\Delta f)_\sigma (u_K - u_{eK}) = \sum_K m_K (\widetilde{\Delta f})_K (u_K - u_{eK})$ avec $(\widetilde{\Delta f})_K = \frac{1}{m(K)} \sum_{\sigma \subset K} m_\sigma^K (\Delta f)_\sigma$.

On peut considérer $\widetilde{\Delta f} = \sum_K (\widetilde{\Delta f})_K \mathbb{1}_K$ comme une perturbation de terme source ; cette perturbation est maintenant de l'ordre h^2 et on double le taux de convergence : $\boxed{\|\nabla^\tau e^\tau\|_{L^p} \sim \begin{cases} h^{\frac{2}{p-1}}, & p \geq 4 \text{ ou } p=2 \\ h^{\frac{p-2}{p-1}}, & 3 < p < 4 \end{cases}}$

On peut faire encore mieux (idée de Bazrett-Liu '93 + techniques d'interpolation) si on remarque que la continuité Hölder de φ^{-1} s'améliore lorsque l'on contrôle les petites valeurs de $|\nabla u_e|$: loin de zéro, φ^{-1} est Lipschitz !
Alors en fonction de ν tq $\int |\nabla u_e|^{-\nu} < +\infty$, on peut améliorer le taux.

Cas extrêmes : • $\nu = 0$, alors les taux sont donnés ci-dessus ($h^{\frac{2}{p-1}}$, $p \geq 4$; $p=2$)
• $\nu = +\infty$ (i.e., il s'agit de solutions $W^{4,q}_{\text{pch}}$ sans points critiques)

numériquement, cet ordre se confirme comme optimal → ordre h^2 en normes L^∞ et $W^{1,2,\tau}$; $h^{\frac{4}{p}}$ en norme $W^{1,p,\tau}$ } pour $p \geq 2$

→ ordre h^2 en norme $W^{1,p,\tau}$; $h^{\frac{3p-2}{p}}$ en norme L^∞ } pour $1 < p < 2$.

Pour $0 < \nu < +\infty$ (on a notamment $\nu = \gamma(p-2)$ si $|f|^{-\gamma} \in L^1$), on montre des ordres intermédiaires.

② Sur maillages cartésiens uniformes (ou gentiment zaffinés - work in progress)
On adapte au cadre EF l'idée de Tyukhtin '82, utilisée par Chow '89 et on fait l'usage de la régularité naturelle Besov $B^{1+\frac{1}{p-1},p}_\infty$ de u_e.
On se place dans le cadre variationnel : e.g. pour le p-laplacien,

$$a_K = \frac{\partial}{\partial u_K} \mathcal{J}_\tau[u^\tau] \quad \text{où} \quad \mathcal{J}_\tau[u^\tau] = \frac{1}{p} \sum_R m_R |\nabla_R u^\tau|^p - \sum_K m_K f_K u_K,$$

la fonctionnelle $\mathcal{J}_\tau$ approche la fonctionnelle continue $\mathcal{J}[u] = \frac{1}{p} \int_\Omega |\nabla u|^p - \int_\Omega f u$.

Rq Numériquement, on peut alors résoudre le schéma par des méthodes simples du type descente ; la nonlinéarité du schéma est alors facile à gérer !

L'estimation d'erreur vient de la convexité de $\mathcal{J}_\tau$, qui peut s'exprimer, de manière équivalente, comme

$$\|\nabla^\tau u^\tau - \nabla^\tau v^\tau\|^p_{L^p} \leq \left[(\nabla \mathcal{J}_\tau[u^\tau] - \nabla \mathcal{J}_\tau[v^\tau]) \cdot (u^\tau - v^\tau)\right.$$
$$\text{(cas } p \geq 2\text{)} \quad \leq \left[\mathcal{J}_\tau[v^\tau] - \mathcal{J}_\tau[u^\tau] - \nabla \mathcal{J}_\tau[u^\tau] \cdot (v^\tau - u^\tau)\right.$$

- La première forme utilise l'optimalité de u^τ pour $\mathcal{J}_\tau$ et donne la preuve de l'estimation de base de tout à l'heure.
- La deuxième permet d'utiliser également l'optimalité de u_e pour $\mathcal{J}$:

$$\|\nabla^\tau u^\tau - \nabla^\tau u_e^\tau\|_{L^p}^p \le \mathcal{J}_\tau[u_e^\tau] - \mathcal{J}_\tau[u^\tau] - \underbrace{\nabla \mathcal{J}_\tau[u^\tau](u_e^\tau - u^\tau)}_{\le 0 \text{ car } u^\tau \text{ est optimale pour } \mathcal{J}_\tau} \le$$

$$\le \mathcal{J}_\tau[u_e^\tau] - \mathcal{J}[\pi_\tau u_e] + \mathcal{J}[\pi_\tau u_e] - \mathcal{J}[u_e] + \underbrace{\mathcal{J}[u_e] - \mathcal{J}[\pi_\tau u^\tau]}_{\le 0 \text{ car } u \text{ est optimale pour } \mathcal{J}} + \mathcal{J}[\pi_\tau u^\tau] - \mathcal{J}_\tau[u^\tau]$$

où π_τ est une interpolation affine/morceaux sur maillage τ telle que $\pi_\tau u_e = \pi_\tau u_e^\tau$.

Il faut alors contrôler trois termes, pour lesquelles on a :

- $|\mathcal{J}[u_e] - \mathcal{J}[\pi_\tau u_e]| \le Ch^{2\alpha}$ lorsque $u_e \in B_\infty^{1+\alpha,p}$
- $|\mathcal{J}_\tau[u_e^\tau] - \mathcal{J}[\pi_\tau u_e^\tau]| \le Ch^{2\alpha}$ lorsque $u_e^\tau, u^\tau \in B_\infty^{1+\alpha,p}$-discret
  $|\mathcal{J}_\tau[u^\tau] - \mathcal{J}[\pi_\tau u^\tau]|$

On en déduit $\|\nabla^\tau e^\tau\|_{L^p}^p \le Ch^{2\alpha} \Rightarrow$ **l'ordre $W^{1,p,\tau}$ en $h^{\frac{2\alpha}{p}}$**.

Or, on connaît

Th (Y. Simon '78) Pour $p \ge 2$, Ω rectangle + cond. Dirichlet homogène, pour tout $f \in L^{p'}(\Omega)$ la solution u_e de $-\Delta_p u_e = f$ est dans $B_\infty^{1+\frac{1}{p-1},p}$.

(et cette régularité est optimale dans les Sobolev fractionnaires).

On montre alors

Proposition
- si $u_e \in B_\infty^{1+\alpha,p}$, alors $u_e^\tau \in B_\infty^{1+\alpha,p}$-discret
- toute solution u^τ du schéma VF sur maillage cartésien uniforme pour $-\Delta_p u = f$, $f \in L^{p'}$, est dans $B_\infty^{1+\frac{1}{p-1},p}$-discret.
  (autrement dit, le schéma VF hérite de la régularité naturelle du problème continu !)

Corollaire Pour tout $f \in L^{p'}$, sans hypothèse a priori sur u_e, on trouve en normes $W^{1,p,\tau}$ et L^p l'ordre de convergence en $h^{\frac{2}{p(p-1)}}$ pour $p \ge 2$.

Rq Numériquement, cet ordre se confirme comme optimal.

— // — // —

Conclusion : on sait refaire la théorie "variationnelle" (fonction test u) du cas continu dans le cadre VF.
La clé : la propriété de dualité div-grad.

[?] Problème ouvert : la théorie L^1 (pour le pb. d'évolution) : fct. test sig(u-k)