Chapitre 5

Les méthodes d'approximation variationnelle

Les formulations variationnelles se prêtent très naturellement à la définition de méthodes d'approximation, c'est-à-dire de réduction à une suite de problèmes en dimension finie que l'on peut effectivement résoudre sur ordinateur. L'objet de ce bref chapitre est d'établir les propriétés communes à toutes ces méthodes du point de vue abstrait.

5.1 Définition et premières propriétés

On considère donc un problème variationnel abstrait bien posé, à savoir on se donne un espace de Hilbert V, a une forme bilinéaire continue coercive et ℓ une forme linéaire continue sur V et on cherche $u \in V$ tel que

$$\text{(PVA)} \qquad a(u,v) = \ell(v), \qquad v \in V$$

Définition 5.1.1 *Une méthode d'approximation variationnelle consiste en la donnée d'un sous-espace vectoriel $V_n \subset V$ de dimension finie. On cherche alors $u_n \in V_n$ solution du problème discret*

$$a(u_n, v_n) = \ell(v_n), \qquad v_n \in V_n. \tag{5.1}$$

On parle également de *méthode de Galerkin*. Comme $V_n \subset V$, on dit qu'il s'agit d'une approximation *conforme*. On peut définir aussi des méthodes d'approximation non conformes pour lesquelles $V_n \not\subset V$ (nous n'en rencontrerons pas ici). Comme on le montre plus loin, l'intérêt de cette méthode est que le calcul de u_n se ramène à la résolution d'un système linéaire de taille $d_n \times d_n$ où on a posé

$$d_n := \dim(V_n).$$

La méthode de Galerkin a donc pour objectif d'approcher la solution exacte $u \in V$ par la solution calculable $u_n \in V_n$. Afin d'améliorer la précision, on considère typiquement une suite d'espaces $(V_n)_{n\geq 1}$ dont la dimension d_n augmente avec n. Dans certains cas, les sous-espaces de dimension finie seront emboîtés, $V_n \subset V_m$ pour tout $m \geq n$. Dans d'autres cas, cette propriété ne sera pas forcément vérifiée. On pourra aussi avoir exactement $d_n = n$, mais cela ne sera pas toujours le cas.

Théorème 5.1.1 *Le problème discret (5.1) admet une solution u_n et une seule.*

Démonstration. Tout sous-espace vectoriel de dimension finie est complet, donc ici V_n est un espace de Hilbert. Les formes bilinéaire a et linéaire ℓ sont naturellement continues et a reste coercive sur V_n. On peut appliquer le théorème de Lax-Milgram dans V_n. □

On sait exactement quand une méthode d'approximation variationnelle est convergente. On peut même préciser quantitativement l'erreur d'approximation. C'est en fait *l'estimation fondamentale* des méthodes d'approximation variationnelle, exprimé par le résultat fondamental suivant.

Théorème 5.1.2 *(Lemme de Cea) Il existe une constante C, indépendante de n et de u, telle que*

$$\|u-u_n\|_V \leq C \min_{v_n \in V_n} \|u-v_n\|_V, \tag{5.2}$$

avec $C = C_a/\alpha$, où C_a et α sont les constantes de continuité et de coercivité de la forme bilinéaire a. Dans le cas où la forme bilinéaire a est symétrique, on obtient le même résultat avec $C = (C_a/\alpha)^{1/2}$.

Démonstration. La formulation variationelle indique en particulier que $a(u,v_n) = \ell(v_n)$ pour tout $v_n \in V_n$, et par soustraction avec l'équation $a(u_n,v_n) = \ell(v_n)$, on trouve

$$a(u-u_n,v_n) = 0.$$

Par ailleurs, d'après la V-ellipticité et la continuité de a, on a

$$\begin{aligned} \alpha\|u-u_n\|_V^2 &\leq a(u-u_n,u-u_n) = a(u-u_n,u-v_n+v_n-u_n) \\ &= a(u-u_n,u-v_n) + a(u-u_n,v_n-u_n) \\ &= a(u-u_n,u-v_n) \leq C_a\|u-u_n\|_V\|u-v_n\|_V, \end{aligned}$$

car $v_n - u_n \in V_n$. D'où en divisant par $\|u-u_n\|_V$ (dans le cas où cette quantité n'est pas nulle, sinon il n'y a rien à démontrer), il vient

$$\forall v_n \in V_n, \qquad \|u-u_n\|_V \leq \frac{C_a}{\alpha}\|u-v_n\|_V$$

d'où l'inégalité désirée, avec $C = C_a/\alpha$.

Dans le cas où a est symétrique, la propriété

$$a(u-u_n,v_n) = 0, \qquad v_n \in V_n$$

s'interprète en disant que u_n est la projection orthogonale de u sur le sous-espace vectoriel fermé V_n pour le produit scalaire $(\cdot,\cdot)_a = a(\cdot,\cdot)$. On a par conséquent, pour tout $v_n \in V_n$,

$$\alpha\|u-u_n\|_V^2 \leq \|u-u_n\|_a^2 \leq \|u-v_n\|_a^2 \leq C_a\|u-v_n\|_V^2,$$

ce qui entraîne l'estimation (5.2) avec la constante $C = (C_a/\alpha)^{1/2}$ □

Remarque 5.1.1 *La quantité $u-u_n$ est l'*erreur *commise par la méthode d'approximation variationnelle. Cette erreur en norme V est majorée par une constante qui ne dépend que du problème variationnel abstrait, et non pas du choix de méthode, multipliée par la distance de la solution u au sous-espace V_n. On se ramène donc à estimer cette distance. Il s'agit alors d'un problème d'approximation. Les estimations de ces erreurs d'approximation vont elles bien sûr dépendre fortement du choix de méthode, et c'est là que celles-ci vont commencer à se distinguer entre elles.*

Remarque 5.1.2 *Puisque $u_n \in V_n$ on a aussi trivialement l'inégalité*

$$\min_{v_n \in V_n} \|u-v_n\|_V \leq \|u-u_n\|_V,$$

ce qui montre que les quantités $\|u-u_n\|_V$ et $\min_{v_n\in V_n}\|u-v_n\|_V$ sont équivalentes, indépendamment du choix de l'espace V_n.

On obtient ainsi comme conséquence immédiate une condition sur les espaces V_n pour la convergence de la méthode de Galerkin.

Corollaire 5.1.3 *Soit $u \in V$ la solution du problème variationnel abstrait, et u_n celle du problème discret. Alors $u_n \to u$ dans V si et seulement si pour tout $v \in V$, il existe une suite (w_n) de V telle que $w_n \in V_n$ pour tout $n \in \mathbb{N}$ et telle que $w_n \to v$ dans V quand $n \to \infty$.*

Une méthode d'approximation variationnelle est donc sûre de converger si et seulement si les sous-espaces V_n deviennent «assez gros» quand n tend vers l'infini, au sens où l'on peut approcher tout élément de V arbitrairement près au sens de la norme de V par un élément de V_n pour n assez grand.

Dans l'exemple qui nous intéresse $V = H_0^1(\Omega)$, $c \in L^\infty(0,1)$, $c \geq 0$ et $f \in L^2(0,1)$,

$$a(u,v) = \int_\Omega (u'v' + cuv), \quad \ell(v) = \int_\Omega fv,$$

on peut donner deux exemples traditionnels de sous-espace de dimension finie V_n.

Le premier exemple, que nous allons développer dans le chapitre suivant, conduit à la méthode des éléments finis $\mathbb{P}_1$. On prend une subdivision $0 = x_0^{(n)} < x_1^{(n)} < \ldots < x_n^{(n)} < x_{n+1}^{(n)} = 1$ de l'intervalle $[0,1]$, et on prend pour V_n les fonctions continues, affines sur chaque intervalle $[x_i^{(n)}, x_{i+1}^{(n)}]$, nulles en 0 et 1. C'est un sous espace de $H_0^1(\Omega)$, de dimension finie $d_n = n$. On pose $h_n = \max_i(x_{i+1}^{(n)} - x_i^{(n)})$, on suppose que $h_n \to 0$ quand $n \to \infty$. Dans ce cas, les espaces ne sont pas nécessairement emboités (les points $x_i^{(n+1)}$ de la subdivision d'indice $n+1$ ne contiennent pas nécessairement les points $x_j^{(n)}$).

Dans le deuxième exemple, à l'opposé, on prend pour V_n les fonctions p_n globalement polynômiales de degré inférieur ou égal à n, sur $[0,1]$, nulles en 0 et 1. Un tel polynôme s'écrit $p_n = (1-x)xq_{n-2}$ où q_{n-2} est un polynôme de degré inférieur ou égal à $n-2$, et l'espace correspondant V_n est de dimension $n-1$. On a dans ce cas $V_n \subset V_{n+1}$. Ce choix conduit aux *méthodes spectrales* qui seront abordées dans le chapitre 7 dans le cadre général de méthodes fondées sur des bases hilbertiennes.

5.2 Forme matricielle de la méthode de Galerkin

En vue de l'implémentation effective sur ordinateur de la méthode de Galerkin, il convient de réinterpréter celle-ci en termes de systèmes linéaires. Pour cela, on doit d'abord se donner une base de V_n, soit w_i, $i = 1, \ldots, d_n$ où on a noté $\dim V_n = d_n$. La solution du système discret se décompose sur cette base sous la forme $u_n = \sum_{i=1}^{d_n} \lambda_i w_i$, et calculer u_n est équivalent à calculer ses composantes λ_i dans la base choisie w_i. On notera

$$U_n = (\lambda_1, \ldots, \lambda_{d_n})^t$$

le vecteur colonne de $\mathbb{R}^{d_n}$ correspondant.

Théorème 5.2.1 *Le vecteur $U_n \in \mathbb{R}^{d_n}$ est l'unique solution du système linéaire*

$$A_n U_n = F_n$$

avec A_n la matrice de coefficients $(A_n)_{ij} = a(w_j, w_i)$ et F_n le vecteur de composantes $(F_n)_i = \ell(w_i)$.

Démonstration. On remarque que la solution discrete u_n est caractérisée par les n équations

$$a(u_n, w_i) = \ell(w_i), \quad i = 1, \ldots, d_n,$$

car pour tout $v_n \in V_n$, on peut écrire $v_n = \sum_{i=1}^{d_n} \mu_i w_i$. En sommant les équations multipliées par μ_i on obtient ainsi $a(u_n, v_n) = \ell(v_n)$.

L'équation i se développe suivant

$$\sum_{j=1}^{d_n} \lambda_j a(w_i, w_j) = \ell(w_i) = (F_n)_i,$$

et ceci montre que U_n est solution du système annoncé. □

Une fois choisie une base de V_n, on se ramène donc finalement à la construction de la matrice A_n, parfois appelée *matrice de rigidité* et à celle du second membre F_n, puis à la résolution effective du système linéaire obtenu. Cette matrice a automatiquement de bonnes propriétés.

Proposition 5.2.1 *La matrice A_n est définie positive (et par conséquent inversible).*

Démonstration. Soit $v_n = \sum_{i=1}^{d_n} \mu_i w_i$ un élément quelconque de V_n et $X_n = (\mu_1, \ldots, \mu_{d_n})^t$ le vecteur de $\mathbb{R}^{d_n}$ associé. On a

$$X_n^T A_n X_n = \sum_{i,j=1}^{d_n} \mu_i \mu_j a(w_j, w_i) = a(v_n, v_n) \geq \alpha \|v_n\|_V^2.$$

Donc $X_n^T A_n X_n \geq 0$, et $X_n^T A_n X_n = 0$ implique $v_n = 0$, donc bien sûr $X_n = 0$. □

Remarque 5.2.1 *Si de plus la forme bilinéaire a est symétrique, alors la matrice A_n est aussi symétrique. On pourra dans ce cas utiliser des méthodes de résolution de systèmes linéaires adaptées aux matrices symétriques, définies positives (comme la méthode de Cholesky, par exemple). Ceci est à comparer aux matrices produites par la méthode des différences finies qui n'avaient ces même bonnes propriétés que par accident finalement, et encore pour un bon choix de numérotation des nœuds. Rien de tel ici.*

Notons enfin une dernière propriété des méthodes d'approximation variationnelles dans le cas symétrique.

Proposition 5.2.2 *Lorsque la forme bilinéaire a est de plus symétrique, le problème variationnel discret est équivalent au problème de minimisation quadratique : minimiser la fonction*

$$J(X_n) = \frac{1}{2} X_n^T A_n X_n - F_n^T X_n$$

sur $\mathbb{R}^{d_n}$.

Démonstration. Évident (le faire !). On a $\nabla J(X_n) = A_n X_n - F_n$. □

Pour la résolution effective du système linéaire, on pourra donc également appliquer des méthodes de descente : gradient à pas optimal, mais surtout gradient conjugué et ses multiples variantes. Ces méthodes calculent la solution du système avec une erreur qui décroit au cours des itérations et qu'il faut prendre en compte en plus de l'erreur d'approximation $\|u - u_n\|_V$ de la méthode de Galerkin.

Remarque 5.2.2 *On rappelle en particulier que la rapidité de convergence des méthode itératives est intimement liée au nombre de conditionnement, qui dans le cas d'une matrice A symétrique définie positive est donné par*

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)},$$

où $\lambda_{\max}(A)$ et $\lambda_{\min}(A)$ désignent la plus grande et plus petite valeur propre. Prenons par exemple la descente de gradient à pas fixe τ appliquée à J. L'itération $k \to k+1$ a la forme

$$U_n^{k+1} = U_n^k + \tau(F_n - A_n U_n^k).$$

L'erreur $E_n^k := U_n - U_n^k$ vérifie $E_n^k = (I - \tau A_n)E_n^{k-1} = \cdots = (I - \tau A_n)^k E_n^0$. On pourra vérifier (exercice) que le pas qui minimise la norme matricielle $\|I - \tau A_n\|_2$ est $\tau = 2(\lambda_{\min}(A_n) + \lambda_{\max}(A_n))^{-1}$, et que la vitesse de convergence de l'erreur est alors en $O(\rho^k)$, où $\rho = \frac{\kappa(A_n)-1}{\kappa(A_n)+1}$. On voit ainsi que cette vitesse se détériore lorsque $\kappa(A_n)$ est très grand.

5.3 Perturbation d'une approximation variationelle

On a vu que l'assemblage du système linéaire nécessite le calcul des éléments de matrice $a(w_j, w_i)$ et des fonctionnelles $\ell(w_i)$. En pratique ces éléments sont donnés par des intégrales : par exemple dans le cas du problème aux limites (P), on a

$$a(w_j, w_i) = \int_\Omega (w_j'(x) w_i'(x) + c(x) w_j(x) w_i(x) dx \quad \text{et} \quad \ell(w_i) = \int_\Omega f(x) w_i(x) dx.$$

Dans certains cas, ces intégrales sont calculables de manière exacte. Le plus souvent, il est nécessaire de les approcher par de formules de quadratures (rectangles, trapezes, Simpson, Gauss-Legendre...).

De façon générale, de telles quadratures approchent pour toute fonction v une intégrale $I(w) = \int_\Omega w(x)dx$ par une formule discrète

$$Q(w) = \sum_{j=1}^{p} \omega_j w(y_j),$$

où les y_j sont des points fixés sur Ω et les ω_j sont des poids réels fixés. Le choix des points et des poids est important pour assurer la précision de la quadrature, c'est à dire la petitesse de l'erreur $|I(w) - Q(w)|$ pour certaines classes de fonctions w que l'on souhaite intégrer.

Notre objectif n'est pas ici de rentrer dans les détails de telles méthodes mais de comprendre ce que l'erreur de quadrature induit sur le calcul de la solution discrète. Une manière de faire cette analyse est de remarquer que toutes les méthodes de quadrature employées reviennent à *perturber* les formes a et ℓ et les remplacer par des approximations $\tilde{a}$ et $\tilde{\ell}$. Par exemple, avec une quadrature du type ci-dessus, on a

$$\tilde{\ell}(v) = \sum_{j=1}^{p} \omega_j v(y_j) f(y_j).$$

Cela signifie que l'on résoud en fait le probleme discret perturbé suivant : trouver $\tilde{u}_n \in V_n$ tel que

$$\tilde{a}(\tilde{u}_n, v_n) = \tilde{\ell}(v_n), \quad v_n \in V_n. \tag{5.3}$$

La solution $\tilde{u}_n$ sera différente de la solution de Galerkin u_n, et il n'est même pas aquis que cette nouvelle solution soit bien définie. Afin de comprendre cela, on fait les hypothèse suivantes sur les perturbations induites par la formule de quadrature, et on obtient un contrôle de la déviation entre u_n et $\tilde{u}_n$. C'est l'objet du résultat suivant, appelé parfois *Lemme de Strang*.

Théorème 5.3.1 *On suppose qu'il existe deux constantes positives ε_1 et ε_2 telles que*

$$|\ell(v_n) - \tilde{\ell}(v_n)| \leq \varepsilon_1 \|v_n\|_V, \quad v_n \in V_n, \tag{5.4}$$

et

$$|a(v_n, w_n) - \tilde{a}(v_n, w_n)| \leq \varepsilon_2 \|v_n\|_V \|w_n\|_V, \quad v_n, w_n \in V_n. \tag{5.5}$$

On suppose aussi que $\varepsilon_2 < \alpha$ où α est la constante de coercivité de a. Alors, il existe une unique solution $\tilde{u}_n \in V_n$ de (5.3)*, et celle-ci vérifie*

$$\|u_n - \tilde{u}_n\|_V \leq \frac{\varepsilon_1 + \varepsilon_2 \frac{C_\ell}{\alpha}}{\alpha - \varepsilon_2} \tag{5.6}$$

où la constante C dépend de C_1, C_2, et α.

Démonstration. Par l'inégalité triangulaire, les hypothèses (5.4) et (5.5) entrainent

$$|\tilde{\ell}(v_n)| \leq (C_\ell + \varepsilon_1)\|v_n\|_V, \quad v_n \in V_n,$$

ainsi que

$$|\tilde{a}(v_n, w_n)| \leq (C_a + \varepsilon_2)\|v_n\|_V \|w_n\|_V, \quad v_n, w_n \in V_n.$$

On a aussi

$$|\tilde{a}(v_n, v_n)| \geq (\alpha - \varepsilon_2)\|v_n\|_V^2, \quad v_n \in V_n.$$

Les hypothèse de Lax-Milgram sont donc vérifiées sur V_n avec les constantes de continuité et de coercivité perturbées $C_{\tilde{\ell}} = C_\ell + \varepsilon_1$, $C_{\tilde{a}} = C_A + \varepsilon_2$ et $\tilde{\alpha} = \alpha - \varepsilon_2 > 0$, ce qui assure l'existence et l'unicité de la solution $\tilde{u}_n$.

Pour controler la distance entre u_n et $\tilde{u}_n$ on fait la différence des équations qui les définissent et on obtient, pour tout $v_n \in V_n$,

$$a(u_n, v_n) - \tilde{a}(\tilde{u}_n, v_n) = \ell(v_n) - \tilde{\ell}(v_n).$$

Ceci peut aussi s'écrire

$$\tilde{a}(u_n - \tilde{u}_n, v_n) = \ell(v_n) - \tilde{\ell}(v_n) + \tilde{a}(u_n, v_n) - a(u_n, v_n).$$

Les hypothèse de perturbations permettent de majorer les termes de droite, ce qui donne

$$\tilde{a}(u_n - \tilde{u}_n, v_n) \leq \varepsilon_1 \|v_n\|_V + \varepsilon_2 \|u_n\|_V \|v_n\|_V.$$

En prenant $v_n = u_n - \tilde{u}_n$ et en utilisant l'ellipticité de $\tilde{a}$ on obtient

$$\tilde{\alpha}\|u_n - \tilde{u}_n\|_V \leq \varepsilon_1 + \varepsilon_2 \|u_n\|_V \leq \varepsilon_1 + \varepsilon_2 \frac{C_\ell}{\alpha},$$

où la deuxième inégalité provient de l'estimation a priori sur u_n. On obtient ainsi l'estimation (5.6) annoncée □

En résumé, on peut dire que toutes les méthodes d'approximation variationnelles (conformes) partagent la même structure abstraite, et notamment l'estimation d'erreur fondamentale (5.2). Tout l'art ensuite va résider dans le choix des espaces de dimension finie, qui devront approcher le mieux possible la solution u, puis dans le choix d'une base de V_n qui produise des matrices faciles à calculer et à résoudre (par exemple des matrices aussi creuses que possible). A l'erreur d'approximation entre u et u_n viennent éventuellement s'ajouter les erreurs supplémentaires liées aux perturbations dues aux quadratures, ainsi que l'erreur d'itération lorsqu'on ne résoud pas le système linéaire de manière exacte.