

**2.2. Common abstract formulation.** Both these variational formulations may be thought of as special cases of a common abstract formulation. To see this, we define

$$a(u, v) = \int_{\Omega} (p \nabla u \cdot \nabla v + quv) dx \left( + \int_{\partial\Omega} \gamma uv ds \right), \quad F(v) = \int_{\Omega} fv dx \left( + \int_{\partial\Omega} gv ds \right),$$

where the terms in parenthesis are only needed for (2.3). We then define a space  $V = \dot{H}^1(\Omega)$  for boundary condition (i) and  $V = H^1(\Omega)$  for boundary condition (ii). Then, both variational formulations have the common form:

$$(2.4) \quad \text{Find } u \in V \text{ such that } a(u, v) = F(v), \quad \text{for all } v \in V.$$

**2.3. Formulation as a minimization problem.** In the problem we are considering, the bilinear form  $a(u, v)$  is symmetric, i.e.,  $a(u, v) = a(v, u)$ . In such a case, we also can formulate problem (2.4) as a minimization problem, i.e., defining  $J(v) = \frac{1}{2}a(v, v) - F(v)$ , we consider

$$(2.5) \quad \text{Find } u \in V \text{ such that } J(u) \leq J(v) \quad \text{for all } v \in V.$$

Then we have the following result.

**Lemma 3.** *If  $u$  is a solution of (2.4), then it is a solution of (2.5) and if  $u$  is a solution of (2.5), then it is a solution of (2.4).*

*Proof.* We use the fact that  $a(u, v)$  is a bilinear form on  $V \times V$ , i.e., for all  $u, v, w \in V$  and all constants  $\alpha$  and  $\beta$ ,

$$a(\alpha u + \beta w, v) = \alpha a(u, v) + \beta a(w, v), \quad a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w),$$

and  $F$  is a linear functional on  $V$ , i.e.,  $F(\alpha u + \beta w) = \alpha F(u) + \beta F(w)$ . First suppose that  $u$  is a solution of (2.5). Since for all  $v \in V$  and constants  $t$ ,  $u + tv \in V$ , we have  $J(u) \leq J(u + tv)$ , i.e.,

$$\frac{1}{2}a(u, u) - F(u) \leq \frac{1}{2}a(u + tv, u + tv) - F(u + tv) = \frac{1}{2}a(u, u) + ta(u, v) + \frac{1}{2}t^2a(v, v) - F(u) - tF(v).$$

Hence,

$$ta(u, v) + \frac{1}{2}t^2a(v, v) - tF(v) \geq 0.$$

This implies

$$a(u, v) + \frac{t}{2}a(v, v) \geq F(v) \quad (t > 0) \quad \text{and} \quad a(u, v) + \frac{t}{2}a(v, v) \leq F(v) \quad (t < 0).$$

Letting  $t \rightarrow 0$  in both equations, we find that  $a(u, v) \geq F(v)$  and  $a(u, v) \leq F(v)$  and so  $a(u, v) = F(v)$  for all  $v \in V$ . Hence  $u$  is a solution of (2.4). Next, let  $u$  be a solution of (2.4). Then for all  $v \in V$ ,

$$\begin{aligned} J(u) - J(v) &= \frac{1}{2}a(u, u) - F(u) - \frac{1}{2}a(v, v) + F(v) \\ &= a(u, u - v) - F(u - v) - \frac{1}{2}a(u, u) + a(u, v) - \frac{1}{2}a(v, v) = 0 - \frac{1}{2}a(u - v, u - v) \leq 0, \end{aligned}$$

since for all  $v \in V$ ,  $a(v, v) \geq 0$ . Hence,  $J(u) \leq J(v)$  for all  $v \in V$  and so  $u$  is a solution of (2.5).  $\square$

**2.4. Ritz-Galerkin approximation schemes.** Let  $V_h$  be a finite dimensional subspace of  $V$  (in practice, we will use piecewise polynomials to construct this subspace). Then, a natural approximation scheme based on formulation (2.5) is:

$$(2.6) \quad \text{Find } u_h \in V_h \text{ such that } J(u_h) \leq J(v_h) \quad \text{for all } v_h \in V_h.$$

In the same way as in the continuous problem, this is equivalent to the method:

$$(2.7) \quad \text{Find } u_h \in V_h \text{ such that } a(u_h, v_h) = F(v_h) \quad \text{for all } v_h \in V_h.$$

We next consider what has to be done to solve Problem (2.7). Let  $\phi_i$ ,  $i = 1, \dots, M$  be a basis for  $V_h$ . Then we can write  $u_h = \sum_{j=1}^M \alpha_j \phi_j$ , for some constants  $\alpha_j$ . To determine  $u_h$ , we now need only determine the  $\alpha_j$ . Since the variation in (2.7) holds for all  $v_h \in V_h$ , it holds when  $v_h$  is chosen to be any of the basis functions  $\phi_i$ . Hence, we get that the  $\alpha_j$  must satisfy

$$(2.8) \quad \sum_{j=1}^M \alpha_j a(\phi_j, \phi_i) = F(\phi_i), \quad i = 1, \dots, M.$$

Next define a matrix  $A = (A_{ij})$  where  $A_{ij} = a(\phi_j, \phi_i)$  and a vector  $b = (b_i)$  by  $b_i = F(\phi_i)$ . If we let  $\alpha$  denote the vector with components  $\alpha_j$ , then our problem reduces to the solution of the linear system of equations  $A\alpha = b$ . Note that it is enough to require that the variation only hold for the basis functions, since if  $a(u_h, \phi_i) = F(\phi_i)$  for  $i = 1, \dots, M$ , then for any constants  $\beta_i$ ,  $i = 1, \dots, M$ , we have

$$a(u_h, \sum_i \beta_i \phi_i) = \sum_i \beta_i a(u_h, \phi_i) = \sum_i \beta_i F(\phi_i) = F(\sum_i \beta_i \phi_i).$$

But any  $v_h \in V_h$  can be written as  $\sum_i \beta_i \phi_i$  for some choice of the constants  $\beta_i$ . Hence, we obtain  $a(u_h, v_h) = F(v_h)$  for all  $v_h \in V_h$ .

The finite element method is a special case of the Ritz-Galerkin method in which we choose the space  $V_h$  to consist of piecewise polynomials.

**2.5. A simple error analysis of Ritz-Galerkin approximation schemes.** To derive error estimates for the Ritz-Galerkin approximation scheme, we use the following properties of the bilinear form  $a(u, v)$  and the linear functional  $F(v)$ .

**Lemma 4.** *There exist positive constants  $\alpha$ ,  $M$ , and  $K$ , such that*

$$a(v, v) \geq \alpha \|v\|_1^2, \quad |a(u, v)| \leq M \|u\|_1 \|v\|_1, \quad |F(v)| \leq K \|v\|_1,$$

Note that  $K$  will depend on the data  $f$  and  $g$  and  $M$  and  $\alpha$  will depend on the coefficients  $p$ ,  $q$ , and  $\gamma$ . When  $p = q = 1$  and  $\gamma = 0$ , the first two inequalities are simple, since then  $a(v, v) = \|v\|_1^2$  and the second follows directly from the Cauchy-Schwarz inequality. In general, one needs estimates such as the following: There exists a positive constant  $C$  such

that for all  $\epsilon > 0$ ,

$$\begin{aligned} \int_{\Omega} u^2 dx &\leq C \left( \int_{\Omega} |\nabla u|^2 dx + \int_{\partial\Omega} u^2 ds \right), \\ \int_{\partial\Omega} u^2 ds &\leq \epsilon \int_{\Omega} |\nabla u|^2 dx + \left[ \frac{C^2}{4\epsilon} + C \right] \int_{\Omega} u^2 dx. \end{aligned}$$

**Theorem 6.** (*Céa's Lemma*)

$$\|u - u_h\|_1 \leq \frac{M}{\alpha} \|u - v_h\|_1, \quad \text{for all } v \in V_h.$$

*Proof.* Recall  $u$  and  $u_h$  satisfy:

$$a(u, v) = F(v), \quad v \in V, \quad a(u_h, v_h) = F(v_h), \quad v_h \in V_h.$$

Since  $V_h \subset V$ , choosing  $v = v_h$  and subtracting equations, we get

$$a(u - u_h, v_h) = 0, \quad v \in V_h \quad (\text{Galerkin orthogonality}).$$

Using Galerkin orthogonality, we get

$$a(u - u_h, u - u_h) = a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) = a(u - u_h, u - v_h),$$

since if  $u_h, v_h \in V_h$ , then  $v_h - u_h \in V_h$ . Hence,

$$\alpha \|u - u_h\|_1^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \leq M \|u - u_h\|_1 \|u - v_h\|_1,$$

and so

$$\|u - u_h\|_1 \leq \frac{M}{\alpha} \|u - v_h\|_1.$$

□