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

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

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

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

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

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

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 α and β,

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

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 $J(u) \le J(u + tv)$, i.e.,

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

Hence,

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

This implies

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

Letting $t \to 0$ in both equations, we find that $a(u,v) \ge F(v)$ and $a(u,v) \le F(v)$ 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 $v \in V$,

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

since for all $v \in V$, $a(v,v) \ge 0$. Hence, $J(u) \le J(v)$ for all $v \in V$ and so u is a solution of (2.5). ◻

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:

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

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

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

We next consider what has to be done to solve Problem (2.7). Let ϕ_i, $i = 1, \ldots, M$ be a basis for V_h. Then we can write $u_h = \sum_{j=1}^{M} \alpha_j \phi_j$, for some constants α_j. To determine u_h, we now need only determine the α_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 ϕ_i. Hence, we get that the α_j must satisfy

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

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 α denote the vector with components α_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 $i = 1, \ldots, M$, then for any constants β_i, $i = 1, \ldots, M$, 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 β_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. **Properties of Ritz-Galerkin approximation schemes.** We make the following assumptions about the bilinear form $a(u, v)$ and the linear functional $F(v)$. These can be verified for the particular choices of $a(u, v)$ and $F(v)$ that we are considering.

Lemma 4. *There exist positive constants α, M, and K, such that for all $u, v \in V$,*

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

Note that K will depend on the data f and g and M and α will depend on the coefficients p, q, and γ. 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 $\epsilon > 0$,

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

From the first property, it easily follows that the Galerkin method has a unique solution.

Lemma 5. *If $a(v,v) \ge \alpha\|v\|_1^2$ for all $v \in V$, then the Galerkin approximation scheme has a unique solution.*

Proof. Since the method reduces to a square linear system of equations, we need only show that when $F(v) = 0$, that $u = 0$. Note this corresponds to $f = 0$ and $g = 0$. But then $a(u_h, v) = 0$ for all $v \in V_h$. Choosing $v = u_h$, we get

$$\alpha\|u_h\|_1^2 \le a(u_h, u_h) = 0.$$

Hence $u_h = 0$. □

We also obtain the following error estimate for the Ritz-Galerkin approximation scheme.

Lemma 6. *(Céa's Lemma)*

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

Proof. Recall u and u_h satisfy:

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

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

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

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,

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

and so

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

□