

Chapter 3

Mathematical Foundations of Least-Squares Finite Element Methods

In Section 2.2, we introduced many of the ideas that form the core of modern least-squares finite element methods (LSFEMs). In this chapter, we develop a mathematical theory that makes precise the key ideas and provides a rigorous framework for the application of least-squares principles. At the center of our framework is an abstract least-squares theory for solving operator equations in Hilbert spaces. The specialization of this framework to partial differential equation (PDE) problems provides a template for LSFEMs that is used throughout the book.

Our theory emphasizes the use of least-squares principles as external variational formulations that replace the naturally occurring formulation of a given PDE problem. Consequently, we focus on methods for which the least-squares minimization step precedes the discretization step.¹ The fact that this approach is by far the most popular in practice further justifies our choice of emphasis.

The basic framework for solving operator equations by residual minimization is developed in Section 3.1, using the results collected in Appendix C. We apply the framework to an abstract PDE problem in Section 3.2 and, in Section 3.2.2, obtain the external, synthetic energy minimization principles that constitute the continuous least-squares principle (CLSP) class for the PDE problem.

The remainder of the chapter is devoted to the formulation and analysis of the discrete least-squares principles (DLSPs) that define least-squares finite element approximations of the solution of the PDE problem. We present the theory in two stages. The first stage (see Section 3.3) examines what can be expected from a discrete residual minimization principle if *no connection* to a CLSP class is assumed. This stage not only helps to explain the remarkable robustness of LSFEMs, but also reveals the limitations of this very general setting. In the second stage, the analysis is extended to include DLSPs obtained from a CLSP class associated with the PDE

¹ Methods for which discretization precedes least-squares minimization can also fit into our framework by linking them to a companion continuous least-squares principle. For example, a discrete least-squares principle based on collocating the PDE at Gauss points and then applying an algebraic least-squares principle to the resulting discrete equations can be viewed as resulting from the approximation of integrals in some CLSP by a quadrature rule. This process can be realized in many different ways; examples of such methods are provided in Section 12.4.

problem. In Section 3.4, we examine the transformation of a CLSP into a DLSP and show that this process consists of choosing approximate *norm-generating* and *differential* operators.² In Section 3.5, the three basic types of DLSPs, e.g., *compliant*, *norm-equivalent*, and *quasi-norm-equivalent*, are shown to result from specific approximation choices. Using the link between CLSPs and DLSPs, we develop there an approximation theory for least-squares finite element approximations of solutions of PDE problems.

Throughout this chapter, we use the notation established in Appendix C.

3.1 Least-Squares Principles for Linear Operator Equations in Hilbert Spaces

Given two Hilbert spaces X and Y , a Fredholm operator³ $\mathcal{Q} \in L(X, Y)$, and a function $f \in Y$, consider the operator equation

$$\text{find } u \in X \quad \text{such that} \quad \mathcal{Q}u = f. \quad (3.1)$$

In this section, we develop and analyze variational methods that “solve” (3.1) by minimizing some norm of the *residual* $\mathcal{R}v = \mathcal{Q}v - f$. The quotation marks indicate that, depending on the deficiency and the nullity of \mathcal{Q} , the problem (3.1) may or may not be solvable for arbitrary $f \in Y$ or may have multiple solutions. On the other hand, it turns out that a properly formulated residual minimization principle, i.e., a least-squares based-method, always has a unique minimizer so that the sense in which this minimizer “solves” (3.1) must be made clear. To this end, we show that a properly formulated least-squares principle recovers the *exact* solution of one of the “nearby” *auxiliary* operator equations defined in Appendix C.⁴

The key to constructing a Rayleigh–Ritz setting for (3.1) is to find a solution space S and a data space H such that the energy balance (C.2) holds, possibly after a redefinition of the operator \mathcal{Q} . Then, the least-squares functional

$$J(v; f) = \|\mathcal{R}v\|_H^2 = \|\mathcal{Q}v - f\|_H^2 \quad (3.2)$$

is *norm equivalent*, i.e., it satisfies

$$C_1 \|v\|_S \leq J(v; 0) \leq C_2 \|v\|_S \quad \forall v \in S \quad (3.3)$$

for some positive constants C_1 and C_2 having values independent of v . Thus, we see that the estimates in (C.2) for the operator equation (3.1) constitute the fundamental prerequisites for the norm-equivalence of the functional $J(\cdot; \cdot)$.

² Concrete examples of discrete operators and norms that are useful in LSFEMs are provided in Sections B.3.1 and B.4.

³ See Appendix C for the definition of a Fredholm operator and of $L(X, Y)$.

⁴ From a regularization perspective, this means that least-squares principles provide *regularization by selection*; see [337].

Positive nullity and positive deficiency have different impacts on least-squares principles. The former affects the norm equivalence of least-squares functionals and can be dealt with by changing the solution and data spaces to something other than the standard choices. The latter requires us to find a nearby problem whose solution coincides with the minimizer of (3.2). However, in both cases, we are led to consider auxiliary operator equations in lieu of (3.1). The energy balances developed in Section C.2 for those equations provide us with all the necessary tools to formulate least-squares principles for (3.1) and to interpret their minimizers as “solutions” of that abstract operator equation. We first consider problems with trivial null spaces and then proceed to the general case of problems with positive nullity.

3.1.1 Problems with Zero Nullity

If $\text{null}(\mathcal{Q}) = 0$, then the spaces X^N , X^C , and X^\perp defined in Section C.1 coincide with the standard space X and (C.9) defaults to the original setting for \mathcal{Q} . For simplicity, we write X instead of \tilde{X} . In this case, the energy balance required to formulate a least-squares principle for (3.1) is given by (C.13). This balance gives rise to the external, residual energy functional

$$J(u; f) = \|\mathcal{Q}u - f\|_Y^2 \quad (3.4)$$

and the least-squares principle for (3.1)

$$\min_{u \in X} J(u; f). \quad (3.5)$$

With (3.4), we associate the “energy” inner product

$$((u, v)) = (\mathcal{Q}u, \mathcal{Q}v)_Y \quad (3.6)$$

and “energy” norm

$$\|v\| = J(v; 0)^{1/2} = (\mathcal{Q}v, \mathcal{Q}v)_Y^{1/2}. \quad (3.7)$$

The following theorem shows that (3.5) is a well-posed problem.

Theorem 3.1 *Assume that $\text{null}(\mathcal{Q}) = 0$. Then, the least-squares minimization problem (3.5) has a unique minimizer $u_{LS} \in X$ for any $f \in Y$.*

Proof. The first-order necessary condition for (3.5) is the variational problem

$$\text{seek } u \in X \text{ such that } \mathcal{Q}(u; v) = F(v) \quad \forall v \in X, \quad (3.8)$$

where

$$\mathcal{Q}(u; v) = ((u, v)) \quad \text{and} \quad F(v) = (\mathcal{Q}v, f)_Y.$$

From (C.13), it follows that the least-squares functional (3.4) is norm equivalent and that (3.6) is an equivalent inner product on X . Therefore, $\mathcal{Q}(\cdot; \cdot)$ is a continuous and

strongly coercive bilinear form on $X \times X$; it is also not difficult to show that $F(\cdot)$ is a continuous linear functional. As a result, the unique solvability of the problem (3.8) follows from Corollary 1.4 (the Lax–Milgram lemma). \square

If, in addition to having a trivial null space, \mathcal{Q} also has zero deficiency, then it is clear that u_{LS} is also a solution of the operator equation (3.1). However, if \mathcal{Q} has positive deficiency, then (3.1) is not solvable unless $f \in R(\mathcal{Q})$ whereas, according to Theorem 3.1, the least-squares principle (3.5) has a unique minimizer u_{LS} for *any* $f \in Y$.

The following theorem shows that u_{LS} solves the auxiliary equation (C.16). This makes it possible to interpret u_{LS} as a “solution” of (3.1), even when f does not belong to the range of \mathcal{Q} .

Theorem 3.2 *Assume that $\text{null}(\mathcal{Q}) = 0$ and let $u_{LS} \in X$ denote the unique minimizer of (3.5). Furthermore, for a given $f \in Y$, let⁵ $\vec{a} = (f, \vec{v})_Y$. Then, the pair $\{u_{LS}, \vec{a}\}$ is the unique solution of the modified problem (C.16).*

Proof. From Lemma C.8, we know that (C.16) has a unique solution $\{u, \vec{a}\}$, where \vec{a} is as in the statement of that lemma. Consider now a least-squares principle for (C.16) derived from the auxiliary operator (C.15). The energy balance for this operator, given by (C.19), gives rise to the minimization problem

$$\min_{\{u, \vec{a}\} \in X \times \mathbb{R}^{K^*}} \|\mathcal{Q}u + \vec{v} \cdot \vec{a} - f\|_Y^2. \quad (3.9)$$

Thanks to (C.19), the functional in (3.9) is norm equivalent. Therefore, the arguments from the proof of Theorem 3.1 can be invoked to show that (3.9) has a unique minimizer. But the unique solution $\{u, \vec{a}\}$ of the auxiliary problem is also a minimizer of (3.9) so that they must coincide. Now that we have established this fact, the theorem follows if we can also prove that u that solves (3.9) is also solution of the minimization problem (3.5), i.e., that $u = u_{LS}$.

To this end, consider the first-order optimality condition for (3.9): seek $\{u, \vec{a}\} \in X \times \mathbb{R}^{K^*}$ such that

$$\mathcal{Q}(\{u, \vec{a}\}; \{w, \vec{c}\}) = F(\{w, \vec{c}\}) \quad \forall \{w, \vec{c}\} \in X \times \mathbb{R}^{K^*}, \quad (3.10)$$

where

$$\mathcal{Q}(\{u, \vec{a}\}; \{w, \vec{c}\}) = (\mathcal{Q}u + \vec{v} \cdot \vec{a}, \mathcal{Q}w + \vec{v} \cdot \vec{c})_Y \quad \text{and} \quad F(\{w, \vec{c}\}) = (f, \mathcal{Q}w + \vec{v} \cdot \vec{c})_Y.$$

The unique solution of (3.10) is, of course, the unique solution $\{u, \vec{a}\}$ of the auxiliary problem. Because $\{v_i\}_{i=1}^{K^*}$ is a (finite) basis for the co-range, we have that $(\mathcal{Q}w, v_i)_Y = 0$ for $i = 1, \dots, K^*$ and any $w \in X$. This simplifies (3.10) to

$$(\mathcal{Q}u, \mathcal{Q}w)_Y + (\vec{v} \cdot \vec{a}, \vec{v} \cdot \vec{c})_Y = (f, \mathcal{Q}w + \vec{v} \cdot \vec{c})_Y.$$

Moving all co-range terms to the right-hand side yields

⁵ $\vec{v} = (v_1, \dots, v_{K^*})^T$ is a basis for the co-range of \mathcal{Q} ; see Section C.1.

$$\begin{aligned}
(\mathcal{Q}u, \mathcal{Q}w)_Y &= (f, \mathcal{Q}w + \vec{v} \cdot \vec{c})_Y - (\vec{v} \cdot \vec{a}, \vec{v} \cdot \vec{c})_Y \\
&= (f, \mathcal{Q}w)_Y + (f - \vec{v} \cdot \vec{a}, \vec{v} \cdot \vec{c})_Y = (f, \mathcal{Q}w)_Y \quad \forall w \in X.
\end{aligned}$$

The last identity follows from the definition of \vec{a} that implies that $f - \vec{v} \cdot \vec{a} \in R(\mathcal{Q})$. Therefore, (3.10) is reduced to exactly the same variational equation as in (3.8) and we can conclude that $u = u_{LS}$. \square

Corollary 3.3 *Assume that $\text{null}(\mathcal{Q}) = 0$ and let $u_{LS} \in X$ denote the unique minimizer of (3.4).*

1. *If $f \in R(\mathcal{Q})$, then u_{LS} is also a solution of the operator equation (3.1).*
2. *If the data belong to the co-range, i.e., if $f \in \text{span}\{v_i\}_{i=1}^{K^*}$, then $u_{LS} \equiv 0$.*
3. *If the data are neither entirely in the range nor the co-range, i.e., $f \notin \text{span}\{v_i\}_{i=1}^{K^*}$ and $|\vec{a}| \neq 0$, then the pair $\{u_{LS}, \vec{a}\}$ solves (C.16).* \square

This corollary shows that least-squares principles for (3.1) always select the “best” possible solution for any given right-hand side f in the sense that u_{LS} coincides with the unique solution of the equation

$$\mathcal{Q}u = f^\perp,$$

where f^\perp is the orthogonal projection of f onto $R(\mathcal{Q})$. Thus, in the ideal case for which $f \in R(\mathcal{Q})$, the least-squares principle simply recovers the solution of (3.1). In the extreme case when f is in the co-range, minimization of (3.5) returns zero. In intermediate situations, the least-squares solution u_{LS} is a function that is mapped to the part of f that belongs to $R(\mathcal{Q})$.⁶

The fact that least-squares principles have a unique, well-defined minimizer, even when the data fail to be compatible, is undoubtedly a valuable computational advantage of the approach. Note that least-squares principles offer this advantage “for free” because the computation of u_{LS} does not require either knowledge of the co-range basis \vec{v} or the use of the auxiliary problem (C.16).

3.1.2 Problems with Positive Nullity

When \mathcal{Q} has positive nullity, the setting of (C.9) involves non-standard spaces that are more difficult to approximate. In this case, it is preferable to work in the bijective setting provided by (C.10) which uses only the standard space X . The relevant energy balance for (C.10) is given by (C.14) and leads to the following residual energy functional:

⁶ The results of Corollary 3.3 also mean that least-squares principles can handle, in a natural way, data errors introduced by discretization. For instance, a perturbation of f in (3.1) caused by, e.g., approximation, may turn a formerly solvable equation into one with incompatible data f^h ; see [67] for an example. Some variational methods break down under these circumstances; however, least-squares principles automatically generate a solution $\{u_{LS}, \vec{a}\}$ in which u_{LS} matches the part of f^h that is in the range of \mathcal{Q} and \vec{a} is proportional to the approximation error $f - f^h$.

$$J(u; f, \vec{c}) = \|\mathcal{Q}u - f\|_Y^2 + |\vec{\ell}(u) - \vec{c}|^2 \quad (3.11)$$

and the least-squares principle for (3.1)

$$\min_{u \in X} J(u; f, \vec{c}). \quad (3.12)$$

The “energy” inner product and norm associated with (3.11) are

$$((u, w)) = (\mathcal{Q}u, \mathcal{Q}w)_Y + \vec{\ell}(u) \cdot \vec{\ell}(w) \quad (3.13)$$

and

$$\|u\| = J(u; 0, \vec{0})^{1/2} = ((u, u))^{1/2}, \quad (3.14)$$

respectively. It follows from the energy balance (C.14) that the functional (3.11) is norm equivalent; as a result, (3.13) is an equivalent inner product on X . The following theorem extends the results of Theorem 3.1 to the present case.

Theorem 3.4 *The least-squares minimization problem (3.12) has a unique minimizer $u_{LS} \in X$ for any $f \in Y$ and $\vec{c} \in \mathbb{R}^K$. Moreover, $\vec{\ell}(u_{LS}) = \vec{c}$.*

Proof. The existence and uniqueness of the least-squares minimizer u_{LS} follows from the Lax–Milgram lemma (Corollary 1.4) along the same lines as in the proof of Theorem 3.1.

To prove the second part of the theorem, note that u_{LS} satisfies the first-order optimality condition

$$(\mathcal{Q}u_{LS}, \mathcal{Q}w)_Y + \vec{\ell}(u_{LS}) \cdot \vec{\ell}(w) = (f, \mathcal{Q}w)_Y + \vec{c} \cdot \vec{\ell}(w) \quad \forall w \in X. \quad (3.15)$$

Testing with the basis $\{u_k\}_{k=1}^K$ of the null space $N(\mathcal{Q})$ reduces (3.15) to the algebraic equation

$$\vec{\ell}(u_{LS}) \cdot \vec{\ell}(u_k) = \vec{c} \cdot \vec{\ell}(u_k) \quad \text{for } k = 1, \dots, K$$

that, in terms of the matrix L defined in (C.6), takes the form

$$L^T \vec{\ell}(u_{LS}) = L^T \vec{c}.$$

By assumption, the ℓ_k ’s are such that L is nonsingular so that $\vec{\ell}(u_{LS}) = \vec{c}$. □

The interpretation of u_{LS} as a solution to (3.1) can be derived from the second auxiliary problem (C.18).

Theorem 3.5 *Let $f \in Y$ and $\vec{c} \in \mathbb{R}^K$ be given and let u_{LS} be the unique minimizer of (3.12). Then, the pair $\{u_{LS}, \vec{a}\}$, where $\vec{a} = (f, \vec{v})_Y$, solves the auxiliary problem (C.18).*

Proof. According to Lemma C.9, the problem (C.18) has a unique solution $\{u, \vec{a}\}$ for any $f \in Y$ and $\vec{c} \in \mathbb{R}^K$. In this solution, \vec{a} is the same as in the statement of the theorem and $\vec{\ell}(u) = \vec{c}$. From Theorem 3.4, we already know that $\vec{\ell}(u_{LS}) = \vec{c}$, i.e., u_{LS} satisfies the second equation in the auxiliary problem. Thus, it remains to show

that $\{u_{LS}, \vec{a}\}$ also satisfies the first equation in (C.18). With slight modifications, the proof in Theorem 3.2 can be used to this end. \square

3.2 Application to Partial Differential Equations

In this section, we specialize the abstract least-squares theory given in Section 3.1 to operator equations (3.1) that represent PDE problems. In what follows, $\mathcal{L}(\mathbf{x}, D)$ is a linear differential operator that acts on functions u , defined on a bounded open region $\Omega \subset \mathbb{R}^d$, and $\mathcal{B}(\mathbf{x}, D)$ is a linear operator acting on functions u defined on the boundary $\partial\Omega$ of Ω .

For simplicity, we often write $\mathcal{L}u$ and $\mathcal{B}u$ whenever the meaning of these symbols is clear from the context. It is also convenient to separate the data space into the two spaces $Y = Y(\Omega)$ and $B = B(\partial\Omega)$ corresponding to the data for the PDE and boundary condition, respectively. Then, (3.1) specializes to the following boundary value problem: given $f \in Y(\Omega)$ and $g \in B(\partial\Omega)$, find $u \in X = X(\Omega)$ such that

$$\begin{cases} \mathcal{L}u = f & \text{in } Y(\Omega) \\ \mathcal{B}u = g & \text{in } B(\partial\Omega). \end{cases} \quad (3.16)$$

Remark 3.6 The concrete forms of \mathcal{L} and \mathcal{B} depend on the arrangement of the dependent variables in u . One possibility is to divide u into scalar and/or vector fields corresponding to various physical quantities modeled by the variables, e.g., currents, fluxes, concentrations, or potentials. Alternatively, one can view u as a vector comprising of the scalar coordinate functions of the physical fields relative to some coordinate system. Of course, this only changes the appearance of the PDE problem (3.16) but not the problem itself, nor its solution.

For an example, consider the first-order Poisson system (1.55) in \mathbb{R}^2 endowed with Cartesian coordinates $\{x, y\}$. The variables in that system can be arranged as a pair $u = \{\phi, \mathbf{v}\}$ of a scalar field and a vector field, or as a triple $u = \{\phi, v_1, v_2\}$ of scalar functions, where v_1 and v_2 denote the components of \mathbf{v} . The forms of $\mathcal{L}(\mathbf{x}, D)$ corresponding to these two ways of expressing the variables are given by⁷

$$\mathcal{L} = \begin{pmatrix} 0 & \nabla \cdot \\ \nabla & \mathcal{I} \end{pmatrix} \quad \text{and} \quad \mathcal{L} = \begin{pmatrix} 0 & \partial_x & \partial_y \\ \partial_x & 1 & 0 \\ \partial_y & 0 & 1 \end{pmatrix},$$

respectively. In the first case \mathcal{L} is a matrix of coordinate-independent differential operators, whereas in the second case \mathcal{L} is a matrix of partial derivatives with respect to the assumed Cartesian coordinate system. We have more to say about these two viewpoints in the subsequent chapters. \square

⁷ ∂_x denotes $\partial/\partial x$.

To formulate least-squares principles for (3.16) and interpret their minimizers as “solutions” of that boundary value problem, we apply the template developed for the abstract equation (3.1). Recall that the energy balances in Section C.2 are the key to defining well-posed residual minimization problems and that the auxiliary problems in Section C.1 are the key to interpreting their minimizers as solutions of (3.1). Therefore, we begin by specializing the results of those sections to (3.16).

To apply the abstract theory of Section 3.1, it is necessary to make the following assumption about \mathcal{L} and \mathcal{B} which we tacitly assume holds throughout this chapter.

Assumption 3.7 *There exist Hilbert spaces $X = X(\Omega)$, $Y = Y(\Omega)$, and $B = B(\partial\Omega)$ such that the mapping $\mathcal{Q} : X \mapsto Y \times B$ defined by $u \mapsto \{\mathcal{L}u, \mathcal{B}u\}$ is a Fredholm operator.⁸* \square

We retain the same notation for the finite-dimensional basis of the null space so that we write $N\{\mathcal{L}, \mathcal{B}\} = \text{span}\{u_1, u_2, \dots, u_K\}$, where $u_k \in X$. If $\mathcal{Q} = \{\mathcal{L}, \mathcal{B}\}$ has a nontrivial co-range, a basis for the co-range consists of K^* linearly independent functions $v_i = \{r_i, b_i\} \in Y \times B$ such that

$$(\{\mathcal{L}u, \mathcal{B}u\}, \{r_i, b_i\})_{Y \times B} = (\mathcal{L}u, r_i)_Y + (\mathcal{B}u, b_i)_B = 0 \quad \text{for } i = 1, 2, \dots, K^*.$$

We set $\vec{v} = (\vec{r}, \vec{b})^\top \in \mathbb{R}^{2K^*}$, where $\vec{r} = (r_1, \dots, r_{K^*})^\top \in \mathbb{R}^{K^*}$ and $\vec{b} = (b_1, \dots, b_{K^*})^\top \in \mathbb{R}^{K^*}$.

3.2.1 Energy Balances

Whenever the boundary value problem (3.16) has a unique solution, i.e., the null space of $\{\mathcal{L}, \mathcal{B}\}$ is trivial, well-posed least-squares principles can be defined without any modifications to the spaces and operators. When $\{\mathcal{L}, \mathcal{B}\}$ has positive nullity, we use the bijective setting provided by the augmented operator

$$\{\mathcal{L}, \mathcal{B}, \vec{\ell}\} : X \mapsto Y \times B \times \mathbb{R}^K; \quad \{\mathcal{L}, \mathcal{B}, \vec{\ell}\}u = \begin{pmatrix} \mathcal{L}u \\ \mathcal{B}u \\ \vec{\ell}(u) \end{pmatrix}. \quad (3.17)$$

The following theorem is a direct consequence of Theorem C.7.⁹

Theorem 3.8 *If $\{\mathcal{L}, \mathcal{B}\}$ has zero nullity, there exist positive constants C_1 and C_2 such that*

$$C_1\|u\|_X \leq \|\mathcal{L}u\|_Y + \|\mathcal{B}u\|_B \leq C_2\|u\|_X \quad \forall u \in X. \quad (3.18)$$

If $\{\mathcal{L}, \mathcal{B}\}$ has finite nullity K , then

$$C_1\|u\|_X \leq \|\mathcal{L}u\|_Y + \|\mathcal{B}u\|_B + |\vec{\ell}(u)| \leq C_2\|u\|_X \quad \forall u \in X. \quad \square \quad (3.19)$$

⁸ In the sequel we simply write X , Y , and B instead of $X(\Omega)$, $Y(\Omega)$, and $B(\partial\Omega)$.

⁹ Recall that throughout this chapter, we assume that Assumption 3.7 holds.

Specialization of the auxiliary problems of Section C.1 to (3.16) is straightforward. For operators with trivial null spaces, the auxiliary problem is

$$\begin{cases} \mathcal{L}u + \vec{r} \cdot \vec{a} = f & \text{in } \Omega \\ \mathcal{B}u + \vec{b} \cdot \vec{a} = g & \text{on } \partial\Omega \end{cases} \quad (3.20)$$

and for operators with positive nullity, that problem is

$$\begin{cases} \mathcal{L}u + \vec{r} \cdot \vec{a} = f & \text{in } \Omega \\ \mathcal{B}u + \vec{b} \cdot \vec{a} = g & \text{on } \partial\Omega \\ \vec{\ell}(u) = \vec{c} & \text{in } \mathbb{R}^K. \end{cases} \quad (3.21)$$

The following theorem is a direct consequence of Lemma C.9 and Theorem C.10.

Theorem 3.9 *Let $K^* = \text{def}\{\mathcal{L}, \mathcal{B}\}$ and $K = \text{null}\{\mathcal{L}, \mathcal{B}\}$. The problem (3.20) has a unique solution $\{u, \vec{a}\} \in X \times \mathbb{R}^{K^*}$ for any $\{f, g\} \in Y \times B$. The problem (3.21) has a unique solution $\{u, \vec{a}\} \in X \times \mathbb{R}^{K^*}$ for any $\{f, g, \vec{c}\} \in Y \times B \times \mathbb{R}^K$. Moreover, there exist positive constants C_1 and C_2 such that*

$$C_1(\|u\|_X + |\vec{a}|) \leq \|\mathcal{L}u + \vec{r} \cdot \vec{a}\|_Y + \|\mathcal{B}u + \vec{b} \cdot \vec{a}\|_B \leq C_2(\|u\|_X + |\vec{a}|) \quad (3.22)$$

for (3.20) and

$$C_1(\|u\|_X + |\vec{a}|) \leq \|\mathcal{L}u + \vec{r} \cdot \vec{a}\|_Y + \|\mathcal{B}u + \vec{b} \cdot \vec{a}\|_B + |\vec{\ell}(u)| \leq C_2(\|u\|_X + |\vec{a}|) \quad (3.23)$$

for (3.21). \square

3.2.2 Continuous Least-Squares Principles

In this section, we use residual minimization to develop external, *least-squares* variational formulations for (3.16). These formulations replace the naturally occurring and/or formal Galerkin variational principles as the basis for developing finite element methods for this boundary value problem. Because the least-squares principles considered in this section operate in infinite dimensional Hilbert spaces, we refer to them as *continuous least-squares principles* (CLSPs), a terminology that is first encountered in Section 2.3.

For boundary value problems with zero nullity, the fundamental energy balance is given by (3.18) which leads to the residual energy functional

$$J(u; f, g) = \|\mathcal{L}u - f\|_Y^2 + \|\mathcal{B}u - g\|_B^2 \quad (3.24)$$

and the corresponding CLSP

$$\min_{u \in X} J(u; f, g). \quad (3.25)$$

If $\{\mathcal{L}, \mathcal{B}\}$ has positive nullity, then the relevant energy balance is given by (3.19), the residual energy functional is given by

$$J(u; f, g, \vec{c}) = \|\mathcal{L}u - f\|_Y^2 + \|\mathcal{B}u - g\|_B^2 + |\vec{\ell}(u) - \vec{c}|^2, \quad (3.26)$$

and the CLSP is given by

$$\min_{u \in X} J(u; f, g, \vec{c}). \quad (3.27)$$

The energy inner products corresponding to (3.25) and (3.27) are given by

$$((u, w)) = (\mathcal{L}u, \mathcal{L}w)_Y + (\mathcal{B}u, \mathcal{B}w)_B \quad (3.28)$$

and

$$((u, w)) = (\mathcal{L}u, \mathcal{L}w)_Y + (\mathcal{B}u, \mathcal{B}w)_B + \vec{\ell}(u) \cdot \vec{\ell}(w), \quad (3.29)$$

respectively. In both cases, the energy norm is

$$\|u\| = J(u; 0, \dots, 0)^{1/2} = ((u, u))^{1/2},$$

where J is one of the functionals (3.24) or (3.26). The following theorem reveals the relationship between the minimizers of problems (3.25) and (3.27) and the solutions of the auxiliary boundary value problems (3.20) and (3.21), respectively.

Theorem 3.10 *Assume that $\{\mathcal{L}, \mathcal{B}\}$ has a trivial kernel. Then, (3.25) has a unique minimizer $u_{LS} \in X$ for any $\{f, g\} \in Y \times B$. The pair $\{u_{LS}, \vec{a}\}$, where $\vec{a} = (f, \vec{r})_Y + (g, \vec{b})_B$, solves the auxiliary problem (3.20). If $\{\mathcal{L}, \mathcal{B}\}$ has positive nullity, (3.27) has a unique minimizer u_{LS} for any $\{f, g, \vec{c}\} \in Y \times B \times \mathbb{R}^K$. The pair $\{u_{LS}, \vec{a}\}$ solves the auxiliary problem (3.21). In particular, $\vec{\ell}(u_{LS}) = \vec{c}$. \square*

We omit the proof of this theorem as it is basically a repetition of the arguments found in the proofs of the abstract results in Section 3.1. Specialization of the statement of Corollary 3.3 to least-squares principles for the boundary value problem (3.16) is also straightforward. The key point here is that the main message of Section 3.1 carries over unchanged to the present setting: if (3.16) has a unique solution $u \in X$, then it is recovered by the least-squares principle; otherwise, the unique least-squares minimizer solves an auxiliary problem.

Remark 3.11 We have just described a formal procedure for defining external variational formulations for the boundary value problem (3.16) that provide a Rayleigh–Ritz-like variational setting for a finite element method. These formulations are based on least-squares residual minimization and are unrelated and independent of any naturally occurring variational formulations for (3.16). \square

Remark 3.12 Because least-squares principles are external to the PDE problem, the associated weak problem

$$\text{seek } u \in X \text{ such that } Q(u; v) = F(v) \quad \forall v \in X \quad (3.30)$$

is not a Galerkin weak formulation of (3.16). For example, assuming that (3.30) is the Euler–Lagrange equation corresponding to (3.24) and that \mathcal{L} is such that the Green’s formula¹⁰

$$(u, \mathcal{L}v)_Y - \langle \mathcal{L}_Y^* u, v \rangle_\Omega = \langle \mathcal{B}_Y^* u, v \rangle_{\partial\Omega} \quad (3.31)$$

holds, where $\langle \cdot, \cdot \rangle$ denotes an appropriate duality pairing, then smooth solutions of (3.30) are not directly solutions of (3.16); instead, they solve the *strong* PDE

$$\mathcal{L}_Y^* \mathcal{L}u = \mathcal{L}^* f \quad \text{in } \Omega \quad (3.32)$$

along with the essential boundary condition

$$\mathcal{B}u = g \quad \text{on } \partial\Omega \quad (3.33)$$

and the natural boundary condition

$$\mathcal{B}_Y^* \mathcal{L}u = \mathcal{B}_Y^* f \quad \text{on } \partial\Omega. \quad (3.34)$$

The system (3.32)–(3.34) forms the boundary value problem for which the least-squares functional (3.24) is the naturally occurring convex quadratic, energy functional that provides the Rayleigh–Ritz setting.¹¹ In other words, the strong problem (3.32)–(3.34) is the PDE problem whose weak Galerkin formulation coincides with the *least-squares* variational problem (3.30). Thus, it is conceivable to develop a least-squares principle for (3.16) by immersion¹² of these equations into the appropriate strong least-squares PDE problem followed by a standard Galerkin procedure. Of course, this is hardly the most efficient or lucid way to proceed. \square

Remark 3.13 To provide a final bird’s-eye view of the least-squares framework, let us revert again to formal operator notation with the understanding that \mathcal{Q} is one of $\{\mathcal{L}, \mathcal{B}\}$ or $\{\mathcal{L}, \mathcal{B}, \tilde{\ell}\}$. By $D(\mathcal{Q})$ and $R(\mathcal{Q})$ we denote the appropriate domain and range, respectively. The starting point in the development of continuous least-squares principles is the *energy balance*

$$C_1 \|u\|_{D(\mathcal{Q})} \leq \|\mathcal{Q}u\|_{R(\mathcal{Q})} \leq C_2 \|u\|_{D(\mathcal{Q})} \quad (3.35)$$

that gave rise to the residual energy functional

$$J(u; f) = \|\mathcal{Q}u - f\|_{R(\mathcal{Q})}^2, \quad (3.36)$$

the least-squares minimization principle

$$\min_{u \in D(\mathcal{Q})} J(u; f), \quad (3.37)$$

¹⁰ Note that \mathcal{L}_Y^* coincides with the standard adjoint of \mathcal{L} only if $Y \equiv L^2(\Omega)$; in general, \mathcal{L}_Y^* is the (Hilbert space) adjoint of \mathcal{L} with respect to the inner product on Y .

¹¹ The system (3.32)–(3.34) can be viewed as the normal equations for the original system (3.16).

¹² In general, the problem (3.32)–(3.34) can be determined from (3.16) thorough differentiation and linear combinations that account for the norm structure of Y .

the energy norm

$$\|u\| = J(u; 0)^{1/2}, \quad (3.38)$$

and the energy inner product

$$((u, v)) = (Qu, Qv)_{R(Q)}. \quad (3.39)$$

The fact that (3.36) was derived from the energy balance (3.35) made certain that (3.38) and (3.39) are an equivalent norm and inner product, respectively. This was essential in proving that (3.37) has a unique minimizer because the bilinear form $Q(\cdot; \cdot)$ appearing in

$$\text{seek } u \in D(Q) \text{ such that } Q(u; w) = F(w) \quad \forall w \in D(Q) \quad (3.40)$$

that is the Euler–Lagrange equation corresponding to (3.37) turned out to be identical to the energy inner product (3.39). It is clear that the process described so far establishes

a mathematical framework that associates a well-posed unconstrained minimization problem with any linear PDE problem that satisfies Assumption 3.7.

Minimization problems constructed through this process are completely defined by the *function space* X and the *norm-equivalent functional* $J(\cdot)$ being minimized over that space, i.e., the pair $\{J, X\}$. The set of all such pairs that can be associated with a given PDE problem constitutes the class of its *continuous least-squares principles* (CLSPs).¹³ \square

Remark 3.14 As a final note, before we move on to DLSPs, let us mention that the abstract CLSPs developed in this section did not assume any specific connection between the PDE and the spaces X, Y , and B , except that they verify Assumption 3.7. In Section 12.2, we briefly describe an approach, referred to as LL* least-squares methods, in which least-squares principles are defined using norms induced by the differential operator \mathcal{L} itself. \square

3.3 General Discrete Least-Squares Principles

CLSPs¹⁴ $\{J, X\}$ offer an alternative, external variational formulation of (3.16). By using $\{J, X\}$ instead of the naturally occurring and/or formal Galerkin formula-

¹³ As we have already mentioned in Section 2.2, for some PDE problems, stability properties are more naturally stated using norms in Banach spaces. Formal application of residual minimization in a Banach space setting gives rise to unconstrained optimization problems with non-differentiable functionals. Examples of such “least-squares” formulations and strategies for their solution are considered in Chapter 10.

¹⁴ $\{J, X\}$ is shorthand for the principle $\min_{u \in X} J(u)$. In fact, because the pair $\{J, X\}$ completely defines the least-square principle, from now on we state least-squares principles by providing this pair and not writing, in every instance, $\min_{u \in X} J(u)$.

tions for (3.16), we can develop finite element methods for this problem using a Rayleigh–Ritz like setting *even if the partial differential equation problem itself is not associated with an unconstrained minimization problem for a convex quadratic functional*.

LSFEMs for which the least-squares step precedes the discretization step have the necessary prerequisite of identifying a CLSP class for the given boundary value problem, i.e., to find the data and solution spaces that verify (3.35). For some problems, this task may be far from trivial and, as we saw in Section 2.2, the CLSP class can include impractical norms and inner products. These issues, critical to the formulation of LSFEMs, are dealt with in Section 3.4; here, we take some time to explore what happens if one decides not to bother with CLSPs and instead defines *discrete least-squares principles* (DLSPs) directly.

A general DLSP for (3.16) is a parameterized family of unconstrained minimization problems

$$\text{seek } u^h \text{ in } X^h \text{ such that } J^h(u^h; f, g) \leq J^h(w^h; f, g) \quad \forall w^h \in X^h, \quad (3.41)$$

where X^h is a finite dimensional space, parameterized by h , whose dimension is proportional to h^{-d} and $J^h : X^h \mapsto \mathbb{R}$ is a convex functional. Following the established shorthand notation, we denote DLSPs by $\{J^h, X^h\}$.

To guarantee that $\{J^h, X^h\}$ has a unique minimizer that, in some sense, approximates solutions of (3.16), it is necessary to make the following two nonrestrictive assumptions.

Assumption 3.15 There exists a *discrete energy inner product*

$$((\cdot, \cdot))_h : X^h \times X^h \mapsto \mathbb{R} \quad (3.42)$$

and a *discrete energy norm* $\|\cdot\|_h = ((\cdot, \cdot))_h^{1/2}$ such that

$$J^h(u^h; 0, 0) = ((u^h, u^h))_h = \|u^h\|_h^2 \quad \forall u^h \in X^h. \quad \square \quad (3.43)$$

Assumption 3.16 The discrete energy norm can be extended to all smooth functions $u \in X$. Furthermore, there exist positive semi-definite bilinear forms $e(\cdot, \cdot)$ and¹⁵ $\varepsilon(\cdot, \cdot)$ such that, for all $u^h \in X^h$ and for all smooth functions $u \in X$,

$$J^h(u^h; \mathcal{L}u, \mathcal{B}u) = \frac{1}{2} \left(((u^h, u^h))_h + ((u, u))_h + \varepsilon(u, u) \right) - ((u, u^h))_h - e(u, u^h). \quad \square \quad (3.44)$$

The first assumption guarantees that $\{J^h, X^h\}$ is a well-posed minimization problem with a unique minimizer out of X^h . The second assumption links the DLSP with the PDE problem that we are trying to solve.

¹⁵ The term $\varepsilon(u, u)$ in (3.44) has no effect on the minimization of $J^h(u^h; \cdot, \cdot)$; it merely adjusts the minimum value of the functional. It is introduced so that the definitions of discrete least-squares functionals are of a relatively simple form.

3.3.1 Error Analysis

The following theorem shows that if the value of the *truncation error* $e(u, u^h)$ is “small,” then the minimization of J^h selects a function $u^h \in X^h$ that is “close” to the exact solution u of the PDE problem.

Theorem 3.17 *Assume that Assumptions 3.15 and 3.16 hold for the DLSP $\{J^h, X^h\}$ and let $u \in X$ denote a sufficiently smooth solution of (3.16). Then, the discrete problem (3.41) has a unique minimizer $u^h \in X^h$. Moreover, that minimizer satisfies the error estimate*

$$\|u - u^h\|_h \leq \inf_{w^h \in X^h} \|u - w^h\|_h + \sup_{w^h \in X^h} \frac{e(u, w^h)}{\|w^h\|_h}, \quad (3.45)$$

i.e., up to a truncation error term, u^h is the orthogonal projection of u with respect to the discrete “energy” inner product (3.42).

Proof. Using (3.42) and (3.44), it is easy to see that the first-order necessary condition that minimizers of (3.41) satisfy is given by

$$\text{seek } u^h \text{ in } X^h \quad \text{such that} \quad Q^h(u^h; w^h) = F^h(w^h) \quad \forall w^h \in X^h, \quad (3.46)$$

where

$$Q^h(\cdot; \cdot) = ((\cdot, \cdot))_h \quad \text{and} \quad F^h(\cdot) = ((u, \cdot))_h + e(u, \cdot).$$

To show the unique solvability of (3.46), let $\{\phi_j^h\}_{j=1}^J$ denote a basis for X^h and $\vec{w} = (w_1, \dots, w_J)^T$ denote the coefficient vector of a function $w^h \in X^h$ with respect to this basis so that $w^h = \sum_{j=1}^J \vec{w}_j \phi_j^h$. Define the matrix A and the vector \vec{f} by

$$A_{ij} = ((\phi_j^h, \phi_i^h))_h \quad \text{and} \quad \vec{f}_i = ((u, \phi_i^h))_h + e(u, \phi_i^h), \quad (3.47)$$

respectively. Then, (3.46) is equivalent to the linear system

$$A\vec{u} = \vec{f} \quad (3.48)$$

of algebraic equations for the unknown coefficient vector \vec{u} corresponding to u^h . From Assumption 3.15, it follows that A is a Gramm matrix corresponding to the basis $\{\phi_j^h\}_{j=1}^J$ relative to $((\cdot, \cdot))_h$ so that A is symmetric and positive definite, the system (3.48) has a unique solution \vec{u} , and (3.41) has a unique minimizer $u^h = \sum_{j=1}^J \vec{u}_j \phi_j^h$.

To prove the error estimate (3.45), let u_\perp^h denote the orthogonal projection of u onto X^h with respect to the discrete energy inner product $((\cdot, \cdot))_h$. Using the triangle inequality, we obtain

$$\|u^h - u\|_h \leq \|u_\perp^h - u\|_h + \|u^h - u_\perp^h\|_h \leq \inf_{w^h \in X^h} \|w^h - u\|_h + \|u^h - u_\perp^h\|_h.$$

To complete the proof, it is necessary to show that $\|u^h - u_\perp^h\|_h$ is bounded by the truncation error. Because u^h solves (3.46),

$$((u^h - u, w^h))_h = e(u, w^h) \quad \forall w^h \in X^h$$

and because u_\perp^h is an orthogonal projection,

$$((u_\perp^h - u, w^h))_h = 0 \quad \forall w^h \in X^h.$$

Subtracting the last two equations yields

$$((u_\perp^h - u^h, w^h))_h = ((u - u^h, w^h))_h = e(u, w^h) \quad \forall w^h \in X^h. \quad (3.49)$$

The error estimate (3.45) then follows from

$$\|u_\perp^h - u^h\|_h = \sup_{w^h \in X^h} \frac{((u_\perp^h - u^h, w^h))_h}{\|w^h\|_h} = \sup_{w^h \in X^h} \frac{e(u, w^h)}{\|w^h\|_h}. \quad \square$$

Definition 3.18 The discrete least-squares functional J^h is *order r -consistent*, or simply *r -consistent*, if there exists a positive number r such that for all sufficiently smooth functions $u \in X$,

$$\sup_{w^h \in X^h} \frac{e(u, w^h)}{\|w^h\|_h} \leq C(u)h^r, \quad (3.50)$$

where $C(u)$ is a positive number whose value may depend on u but not on h . A discrete least-squares functional is called *consistent* if the truncation error is zero, i.e., (3.44) holds with $e(\cdot, \cdot) \equiv 0$.¹⁶ \square

We then have the following result.

Corollary 3.19 *Let the hypotheses of Theorem 3.17 hold. Let $u^h \in X^h$ denote the unique solution of the DLSP $\{J^h, X^h\}$, i.e., of (3.41). If J^h is order r -consistent, then*

$$\|u - u^h\|_h \leq \inf_{w^h \in X^h} \|u - w^h\|_h + C(u)h^r. \quad (3.51)$$

If $\{J^h, X^h\}$ is consistent, then its solution u^h minimizes the discrete energy norm error, i.e.,

¹⁶ If J^h is a consistent least-squares functional,

$$J^h(u, \mathcal{L}u, \mathcal{B}u) = \frac{1}{2} \varepsilon(u, u)$$

for any sufficiently smooth solution u of (3.16). On the other hand, given that $X^h \subset X$, it is clear that

$$J^h(u, f, g) \leq J^h(u^h, f, g).$$

Therefore, if $\varepsilon(u, u) = 0$, consistent least-squares functionals provide natural error monitors that can be used in mesh refinement algorithms. See Section 12.14.

$$\|u - u^h\|_h = \inf_{w^h \in X^h} \|u - w^h\|_h. \quad (3.52)$$

Proof. The proof follows from (3.45) by using (3.50) in the r -consistent case and by noting that $e(\cdot, \cdot) \equiv 0$ for a consistent principle. \square

Corollary 3.19 implies that minimizers of r -consistent functionals are orthogonal projections of the exact solutions up to $O(h^r)$ terms, and that minimizers of consistent functionals are true orthogonal projections of these solutions.

3.3.2 The Need for Continuous Least-Squares Principles

Theorem 3.17 shows that reasonable DLSPs can be defined under a minimal set of assumptions. This explains why LSFEMs tend to be much more robust than their mixed and formal Galerkin brethren and why even naively defined least-squares principles rarely fail in a spectacular way. Indeed, chances are that a naively defined mixed or a formal Galerkin method almost surely violate the delicate variational compatibility conditions in (1.41) and (1.42) or (1.33) and (1.34). On the other hand, one must be extremely unlucky to stumble upon a pair $\{J^h, X^h\}$ that does not satisfy Assumptions 3.15 and 3.16.

However, the fact that least-squares principles can produce reasonable results assuming only a loose connection with the PDE problem leads to the erroneous (and precarious) conclusion that LSFEMs can be safely developed using Assumptions 3.15 and 3.16 as the sole guidelines. The fallacy of this thinking is that neither Theorem 3.17 nor Corollary 3.19 can give us any further information about the asymptotic behavior of the error or the condition number of the least-squares algebraic problem (3.48). Let us conduct a “thought experiment” that shows how things can go wrong if we do not establish a tighter bond between a DLSP $\{J^h, X^h\}$ and a well-posed artificial energy principle $\{J, X\}$.

Let $\{J, X\}$ denote a well-posed CLSP for (3.16) and assume that $\{J^h, X^h\}$ is a DLSP for the same problem that satisfies Assumptions 3.15 and 3.16. Assume that both the continuous energy norm $\|\cdot\|$ and the norm $\|\cdot\|_X$ are meaningful for $u^h \in X^h$ so that their restrictions are well-defined norms on X^h . The discrete energy norm $\|\cdot\|_h$ is another norm on this finite-dimensional space so that it must be equivalent to the restrictions of $\|\cdot\|$ and $\|\cdot\|_X$ to X^h . As a result, for every fixed $h > 0$, there exist constants $C_1(h)$ and $C_2(h)$ such that

$$C_1(h)\|u^h\|_X \leq \|u^h\|_h \leq C_2(h)\|u^h\|_X.$$

This explains the remarkable robustness of least-squares principles; *a version of the correct energy balance continues to hold for any fixed h* . Likewise, there are two other constants $\delta_1(h)$ and $\delta_2(h)$ such that

$$\delta_1(h)|\bar{u}|^2 \leq \bar{u}^T A \bar{u} \leq \delta_2(h)|\bar{u}|^2$$

for the coefficient vector \vec{u} of u^h . This is the whole story as far as an individual, fixed minimization problem from $\{J^h, X^h\}$ is concerned.

However, for a *parameterized family* of problems, what matters is the asymptotic behavior of these “constants” that depends entirely on the relation between the *discrete* and *continuous* energy norms; neither Assumptions 3.15 or 3.16 can control the growth (or decay) of $C_i(h)$ and $\delta_i(h)$. As a result, although each individual minimization problem may appear perfectly adequate, the family as a whole may experience a decline in convergence rates and/or sharp increase in the condition number of A as h becomes smaller and smaller.

A LSFEM for which these ratios deteriorate asymptotically may continue to perform well and without apparent indication of a failure. One reason is that realistic values of h used in practice are often not “small” enough for the deterioration of $C_2(h)/C_1(h)$ to be noticeable. Deterioration of $\delta_2(h)/\delta_1(h)$, on the other hand, tends to be more noticeable, especially when iterative methods are used to solve the linear systems. Users of such LSFEMs who have tried to switch from direct to iterative solvers have experienced a steep growth of iterations as the problem size increases.

The conclusion from this thought experiment is unambiguous: despite the initial appeal of simple conditions such as Assumptions 3.15 and 3.16, they alone cannot guarantee robust and efficient LSFEMs. To accomplish this, a DLSP $\{J^h, X^h\}$ must “mimic” a well-posed CLSP $\{J, X\}$ for (3.16). The level of association between discrete and continuous principles and how it affects the properties of $\{J^h, X^h\}$ are the subject of the next section.

3.4 Binding Discrete Least-Squares Principles to Partial Differential Equations

In defining a DLSP for a given PDE problem, two steps are taken; one must define a least-squares principle and one must discretize the problem, i.e., transform it from a infinite-dimensional to a finite-dimensional problem. The order in which one takes the discretization and least-squares steps corresponds to two fundamentally different ways of associating PDEs with DLSPs.

In the *discretize and then minimize* approach, the PDE is first discretized, i.e., replaced by a possibly overdetermined system of algebraic equations. This system is then solved by an algebraic least-squares method.¹⁷ The lack of a clear connection with a well-posed residual energy principle for (3.16) is a serious drawback of the discretize and then minimize strategy. In general, LSFEMs obtained by this approach are not easily amenable to stability and error analysis beyond that which is already established in Theorem 3.17 and Corollary 3.19.¹⁸ For this reason, the

¹⁷ A typical example is collocation LSFEMs in which the initial reduction of the PDE to an overdetermined linear system is accomplished by collocation. This and other examples are considered in Section 12.4.

¹⁸ In some special cases, e.g., when collocation points coincide with quadrature nodes, a connection can be developed and exploited in the analysis.

discretize and then minimize approach is not pursued in this book, other than the brief discussion in Section 12.4.

Instead, our main focus continues to be on methods where the least-squares step precedes the discretization step. Such *minimize and then discretize* approaches maintain a close connection between CLSPs and DLSPs and result in a much more satisfying mathematical structure and also in methods that are more easily amenable to rigorous stability and error analyses. The main prerequisite for the success of these methods is the careful execution of the discretization step, i.e., the transition from a CLSP $\{J, X\}$ to a DLSP $\{J^h, X^h\}$.

3.4.1 Transformations from Continuous to Discrete Least-Squares Principles

The transformation of a CSLP $\{J, X\}$ into a DLSP $\{J^h, X^h\}$ should ultimately be guided by practicality considerations. In an ideal situation, that transformation would only entail restriction of the minimization process to a subspace X^h of X , i.e., we would thus obtain the DLSP $\{J, X^h\}$.¹⁹ In less ideal situations, a practical DLSP $\{J^h, X^h\}$ may differ substantially from its continuous prototype. In either case, properties, both attractive and unattractive, of DLSPs depend on their deviation from the energy balance prescribed by $\{J, X\}$. Thus, to evaluate the outcome of a particular transition to a discrete principle, we must be able to assess how well J^h mimics the norm equivalence of J .

To facilitate this task, we rewrite the energy balance of (3.16) using norm-generating operators for the data and the solution spaces. Then, the transition to a discrete principle can be regarded as a process wherein norm-generating and problem-defining operators are replaced by discrete approximations. This viewpoint allows one to easily track the impact that different approximation choices for the original operators have on the resulting DLSP. In particular, it is shown below that depending what choice one makes, the transformation of CLSPs to DLSPs can take three distinct routes that lead to three different categories of DLSPs.

To formalize the discussion, we focus on problems with trivial null spaces. This setting is sufficient because norm equivalence is unaffected by the terms used to remove the null spaces. We consider the problem (3.16) and assume that the energy balance (3.18) holds for some spaces X , Y , and B .

Let $\mathcal{S}_{(*)}$ for $* \in \{X, Y, B\}$ denote norm-generating operators for X , Y , and B , respectively, with $L^2(\Omega)$ acting as a pivot space, i.e., we have that

$$\|u\|_X = \|\mathcal{S}_X u\|_0, \quad \|w\|_Y = \|\mathcal{S}_Y w\|_0, \quad \text{and} \quad \|b\|_B = \|\mathcal{S}_B b\|_{0, \partial\Omega}.$$

Using these operators, the energy balance (3.18) takes the form

$$C_1 \|\mathcal{S}_X u\|_0 \leq \|\mathcal{S}_Y \circ \mathcal{L}u\|_0 + \|\mathcal{S}_B \circ \mathcal{B}u\|_{0, \partial\Omega} \leq C_2 \|\mathcal{S}_X u\|_0. \quad (3.53)$$

¹⁹ This is exactly the “straightforward” LSFEM discussed in Section 2.2.1.

Norm-generating operators²⁰ depend on the spaces involved in (3.18). Because (3.16) may admit multiple energy balances, we treat $\mathcal{S}_{(*)}$ as a family of operators that represent all admissible space combinations in (3.18). In the setting relevant to LSFEMs, X , Y , and B belong to Hilbert scales induced by $L^2(\Omega)$ or $L^2(\partial\Omega)$ and a generating, self-adjoint, positive definite operator; see, e.g., [249]. As a result, each family $\mathcal{S}_{(*)}$ can be identified with selected powers of the generating operators for these Hilbert scales. Consequently, $\mathcal{S}_{(*)}$ is a family of self-adjoint, positive definite operators.

Armed with the operator form (3.53) of the energy balance (3.18), we are now prepared to talk about transitions to DLSPs. To make the discussion as straightforward as possible, it is convenient to make the following assumption.

Assumption 3.20 A practical subspace²¹ X^h of X that satisfies the *approximability assumption* (B.8) exists for the CLSP $\{J, X\}$.

This assumption means that the rest of this chapter targets primarily *conforming* LSFEMs, i.e., we have that $X^h \subset X$.²²

In terms of norm-generating operators, Assumption 3.20 means that \mathcal{S}_X is such that its domain $D(\mathcal{S}_X)$ contains “practical” discrete subspaces. As a result,

the practicality of the CSLP $\{J, X\}$ depends solely on the effort required to compute $\mathcal{S}_Y \circ \mathcal{L}u^h$ and $\mathcal{S}_B \circ \mathcal{B}u^h$.

If this effort is deemed reasonable, the original energy norm

$$\|u\| = \|\mathcal{S}_Y \circ \mathcal{L}u\|_0 + \|\mathcal{S}_B \circ \mathcal{B}u\|_{0, \partial\Omega}$$

can be retained and the transition process is complete. Otherwise, we proceed to replace the composite operators $\mathcal{S}_Y \circ \mathcal{L}$ and $\mathcal{S}_B \circ \mathcal{B}$ by computable discrete approximations $\mathcal{S}_Y^h \circ \mathcal{L}^h$ and $\mathcal{S}_B^h \circ \mathcal{B}^h$, respectively. Finally, we may need *projection* operators π_Ω^h and $\pi_{\partial\Omega}^h$ that act on the data f and g so as to place them in the domains of \mathcal{S}_Y^h and \mathcal{S}_B^h , respectively. In both cases, the conversion process and the key properties of the resulting DSLP can be encoded by the *transition* diagram

²⁰ Concrete examples of norm-generating operators are given at the end of this section.

²¹ Recall that a practical subspace is one for which basis functions are easily constructed, i.e., with no more difficulty than one would encounter for Galerkin and mixed-Galerkin finite element methods for the same problem.

²² Conformity of LSFEMs is not a crippling restriction because, for a large number of PDEs, Assumption 3.20 can be fulfilled and conforming least-squares methods can be defined through a judicious choice of $\{J, X\}$, including a possible reformulation of (3.16) into an equivalent problem, e.g., a first-order system. One notable exception are PDEs whose solution space is the intersection of the spaces $\mathbf{C}(\Omega)$ and $\mathbf{D}(\Omega)$ (see Appendix B for definition of these spaces). In Section B.2.2, we explain why, in this case, conforming finite element subspaces are unsatisfactory and why non-conforming approximations are preferable. Typical representatives of such PDEs are the div-curl systems that are considered in Section 6.3. There, and also in Section 7.7, we encounter examples of *non-conforming* LSFEMs for which $X^h \not\subset X$. Analysis of these methods relies on specific properties of X^h which makes it less amenable to generalizations. This is why the bulk of the abstract least-squares theory developed in this chapter deals with conforming LSFEMs.

$$\begin{array}{ccccccc}
J(u; f, g) & = & \|S_Y \circ (\mathcal{L}u - f)\|_0^2 & + & \|S_B \circ (\mathcal{B}u - g)\|_0^2 & \rightarrow & \|u\| \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
J^h(u^h; f, g) & = & \|S_Y^h \circ (\mathcal{L}^h u^h - \pi_\Omega^h f)\|_0 & + & \|S_B^h \circ (\mathcal{B}^h u^h - \pi_{\partial\Omega}^h g)\|_0 & \rightarrow & \|u^h\|_h
\end{array} \quad (3.54)$$

and the companion *norm-equivalence* diagram^{23,24}

$$\begin{array}{ccccc}
C_1 \|u\|_X & \leq & \|u\| & \leq & C_2 \|u\|_X \\
\downarrow & & \downarrow & & \downarrow \\
C_1(h) \|u^h\|_X & \leq & \|u^h\|_h & \leq & C_2(h) \|u^h\|_X.
\end{array} \quad (3.55)$$

Remark 3.21 The scope of (3.54) obviously extends to non-conforming LSFEMs for which $X^h \not\subset X$. However, for such methods nothing much can be said without further information about the discrete spaces and operators involved. In particular, for non-conforming LSFEMs, we do not have the companion diagram (3.55) because $\|\cdot\|_X$ may not be defined for $u^h \in X^h$. The absence of such a diagram makes it more difficult to estimate the level of “norm equivalence” of non-conforming methods because they lack the clear reference point provided by the upper row in (3.55) that is available in the conforming case. \square

The truncation error of discrete least-squares functionals obtained by (3.54) is easily quantifiable.

Lemma 3.22 Assume that J^h is transformed from J according to (3.54). Then, for all sufficiently smooth $u \in X$,

$$\begin{aligned}
e(u, w^h) &= - (S_Y^h \circ (\pi_\Omega^h \mathcal{L} - \mathcal{L}^h)u, S_Y^h \circ \mathcal{L}^h w^h)_{0,\Omega} \\
&\quad - (S_B^h \circ (\pi_{\partial\Omega}^h \mathcal{B} - \mathcal{B}^h)u, S_B^h \circ \mathcal{B}^h w^h)_{0,\partial\Omega}
\end{aligned} \quad (3.56)$$

and

$$\varepsilon(u, u) = \frac{1}{2} \|S_Y^h \circ (\pi_\Omega^h \mathcal{L} - \mathcal{L}^h)u\|_{0,\Omega}^2 + \frac{1}{2} \|S_B^h \circ (\pi_{\partial\Omega}^h \mathcal{B} - \mathcal{B}^h)u\|_{0,\partial\Omega}^2. \quad \square \quad (3.57)$$

²³ Together, (3.54) and (3.55) represent a useful and important tool for understanding LSFEMs. In addition to providing a systematic process for *deriving* DLSPs, they can also be used to establish a reverse *association* between a given DLSP and a continuous prototype. As a tool for the design of LSFEMs, (3.54) and (3.55) highlight (in combination with (3.53)) the different roles played by \mathcal{L} , \mathcal{B} , and S_Y , S_B , provide guidelines for choosing approximations compatible with these roles, and allow one to assess the impact of these choices. On the other hand, the possibility to always relate DLSPs to well-posed CLSPs is convenient when trying to assess the qualities of DLSPs defined in an ad hoc manner.

²⁴ In principle, a reverse association with a well-posed CLSP $\{J, X\}$ may be sought for any DLSP, including those obtained by the discretize and then minimize approach, so as to enable their variational analysis. However, that process entails many more ambiguities compared to reverse associations for DLSPs obtained by the minimize and then discretize approach.

The proof of this lemma is straightforward.

\mathcal{L} and \mathcal{B} define the problem that is being solved so that the main objective is to choose \mathcal{L}^h and \mathcal{B}^h that bind $\{J^h, X^h\}$ to this problem, i.e., that make J^h as close as possible to J for the exact solution of (3.16). An appropriate choice is to use operators that lead to truncation errors of order r in (3.44), i.e., \mathcal{L}^h and \mathcal{B}^h are such that (3.50) holds for some positive r .

On the other hand, S_Y and S_B define the energy balance of (3.16), i.e., the proper scaling between data and solution. As a result, the main objective in the choice of S_Y^h and S_B^h is to ensure that the scaling induced by J^h is as close as possible to (3.18), i.e., to “bind” $\{J^h, X^h\}$ to the energy balance of $\{J, X\}$.

To obtain robust and efficient DLSPs, approximations of the problem-defining operators and the norm-generating operators must work in concert to provide good accuracy and reasonable norm equivalence for $\{J^h, X^h\}$. The scaling between discrete data and solution spaces

$$C_1(h)\|u^h\|_X \leq \|u^h\|_h \leq C_2(h)\|u^h\|_X \quad (3.58)$$

induced by $\{J^h, X^h\}$ depends on the discrete energy norm

$$\|u^h\|_h = \left(\|S_Y^h \circ \mathcal{L}^h u^h\|_0^2 + \|S_B^h \circ \mathcal{B}^h u^h\|_{0,\partial\Omega}^2 \right)^{1/2}$$

which in turn depends on the choices made for S_Y^h , S_B^h , \mathcal{L}^h , and \mathcal{B}^h . In general, the upper and lower bounds in (3.58) are mesh-dependent and this may affect performance of DLSPs in several ways; see Section 3.3.2. The severity of the performance drop as measured by the decrease in convergence rates and/or increase in condition numbers can serve as a precise measure for the deviation of $\{J^h, X^h\}$ from the optimal energy balance. Based on how well $\|\cdot\|_h$ represents this balance, DLSPs can be divided into three distinct categories. These categories, their properties, and their formal error analysis are presented in Section 3.5.

Examples of norm-generating operators

The following examples illustrate the notion of norm-generating operators.

Example 3.23 The Laplace operator $-\Delta$ with homogeneous boundary conditions is self-adjoint and positive definite. As a result, all powers of this operator are defined and we have that

$$\|\phi\|_{-1} = \|(-\Delta)^{-1/2}\phi\|_0, \quad \|\phi\|_0 = \|(-\Delta)^0\phi\|_0, \quad \text{and} \quad \|\phi\|_1 = \|(-\Delta)^{1/2}\phi\|_0.$$

For proof of the first identity, see Theorem A.1. \square

Example 3.24 The first-order Poisson equation (1.55) admits the following energy balance²⁵ (see Theorem 5.9):

$$C(\|\phi\|_1 + \|\mathbf{v}\|_0) \leq \|\nabla \cdot \mathbf{v}\|_{-1} + \|\nabla \phi + \mathbf{v}\|_0.$$

Using the characterizations from Example 3.23, we can rewrite this balance as

$$\begin{aligned} C \left((\|(-\Delta)^0 \phi\|_0^2 + \|(-\Delta)^{-1/2} \phi\|_0^2)^{1/2} + \|(-\Delta)^0 \mathbf{v}\|_0 \right) \\ \leq \|(-\Delta)^{-1/2} \nabla \cdot \mathbf{v}\|_0 + \|(-\Delta)^0 (\nabla \phi + \mathbf{v})\|_0. \end{aligned}$$

Therefore, the operator S_X which acts on the solution $\{\phi, \mathbf{v}\}$ and the operator S_Y which acts on the data can be identified with

$$S_X = \left((-\Delta)^0 + (-\Delta)^{1/2}, (-\Delta)^0 \right)^\top \quad \text{and} \quad S_Y = \left((-\Delta)^{-1/2}, (-\Delta)^0 \right)^\top,$$

respectively.

3.5 Taxonomy of Conforming Discrete Least-Squares Principles and their Analysis

Broadly speaking, there are two fundamental types of LSFEMs. LSFEMs for which Assumption 3.20 is satisfied belong to the class of conforming finite element methods. For such methods, restriction of their parent CLSP to the conforming discrete subspace $X^h \subset X$ (which may or may not be practical), provides a natural reference point that allows one to assess the level of deviation of conforming LSFEMs from the ideal Rayleigh–Ritz setting. This leads to the three distinct subclasses of conforming LSFEMs discussed in this section for which it is possible to develop an abstract approximation theory under a reasonable set of assumptions.

The second fundamental type consists of non-conforming LSFEMs, i.e., methods for which $X^h \not\subset X$. In this case, solution norms from X cannot be applied to functions in X^h so that we lack an obvious reference point to estimate the deviation from the ideal Rayleigh–Ritz setting; see Remark 3.21. For this reason, further subdivision of non-conforming LSFEMs, based on their “norm equivalence,” and an abstract approximation theory is not attempted; we deal with non-conforming LSFEMs on a case by case basis.

Apart from the ideal *compliant* class for which LSFEMs reproduce the classical Rayleigh–Ritz principle, there are two other kinds of DLSPs that gradually drift away from this setting, primarily by *simplifying the approximations* of the norm-generating operators. The *norm-equivalent* class retains virtually all attractive prop-

²⁵ Throughout the book we often state only the lower bound in the energy balance because, as a rule, it is much harder to prove than the upper bound. The latter in most cases follows directly from the triangle inequality.

erties of the Rayleigh–Ritz setting, including identical convergence rates and matrix condition numbers. The *quasi-norm-equivalent* class admits the broadest range of DLSPs.²⁶ However, the generality of this class also makes its analysis much more complex and involved, and, for this reason, we present it last.

In what follows, we assume that the dependent variables are arranged as $u = \{u_1, u_2, \dots, u_M\}$ and that \mathcal{L} and \mathcal{B} are in a form that corresponds to that arrangement (see Remark 3.6), i.e., \mathcal{L} and \mathcal{B} are given by

$$\mathcal{L} = \begin{pmatrix} \mathcal{L}_{11} & \cdots & \mathcal{L}_{1M} \\ \vdots & \ddots & \vdots \\ \mathcal{L}_{M1} & \cdots & \mathcal{L}_{MM} \end{pmatrix} \quad \text{and} \quad \mathcal{B} = \begin{pmatrix} \mathcal{B}_{11} & \cdots & \mathcal{B}_{1M} \\ \vdots & \ddots & \vdots \\ \mathcal{B}_{L1} & \cdots & \mathcal{B}_{LM} \end{pmatrix},$$

where \mathcal{L}_{ij} and \mathcal{B}_{ij} are differential operators acting on Ω and $\partial\Omega$, respectively. We use X_i to denote the component space for the i th element²⁷ of u , i.e., $X = X_1 \times \cdots \times X_M$. Likewise, $Y = Y_1 \times \cdots \times Y_M$ and $B = B_1 \times \cdots \times B_L$, where Y_j and B_l are the data spaces for the j th differential equation and the l th boundary condition in (3.16).

Some additional hypotheses about X^h are necessary to discuss the least-squares classes. First, X^h is expected to satisfy the following inverse assumptions.

Assumption 3.25 There exist non-negative weights $\omega_1, \dots, \omega_M$ such that, for all components $u_i^h \in X_i^h$, $i = 1, \dots, M$, of $u^h \in X^h = X_1^h \times X_2^h \times \cdots \times X_M^h$, either

$$\|u_i^h\|_0 \leq \|u_i^h\|_{X_i} \leq C_I h^{-\omega_i} \|u_i^h\|_0 \quad (3.59)$$

or

$$\|u_i^h\|_0 \geq \|u_i^h\|_{X_i} \geq C_I h^{\omega_i} \|u_i^h\|_0. \quad \square \quad (3.60)$$

We write (3.59) and (3.60) in the compact forms

$$\|u^h\|_0 \leq \|u^h\|_X \leq C_I h^{-\bar{\omega}} \|u^h\|_0$$

and

$$\|u^h\|_0 \geq \|u^h\|_X \geq C_I h^{\bar{\omega}} \|u^h\|_0,$$

respectively, where $\bar{\omega} = (\omega_1, \dots, \omega_M)^\top$ and $h^{\pm \bar{\omega}} = (h^{\pm \omega_1}, \dots, h^{\pm \omega_M})^\top$. We define $\omega_{\max} = \max\{\omega_j\}$.

The second assumption deals with the relation between the $L^2(\Omega)$ norm of a finite element function u^h and the Euclidean norm of the corresponding vector of coefficients \vec{u} .

Assumption 3.26 There exist positive constants μ_1 and μ_2 such that, for all $u^h \in X^h$,

$$\mu_1 h^d |\vec{u}|^2 \leq \|u^h\|_0^2 \leq \mu_2 h^d |\vec{u}|^2, \quad (3.61)$$

²⁶ In fact, the quasi-norm-equivalent class contains the compliant and norm-equivalent classes as special cases.

²⁷ Recall that each element u_i can be a scalar or a vector field.

where d denotes the space dimension. \square

Both (3.59) and (3.61) hold for a wide range of finite element spaces under some mild restrictions on the partition of the domain Ω into finite elements; see Theorems B.26 and B.27 in Section B.3.4.

In what follows, we use u , u^h , and w^h to denote a solution of (3.16), its least-squares approximation out of X^h , and an arbitrary element of X^h , respectively. For simplicity, we continue to restrict attention to problems with trivial null spaces for which the relevant energy balance is (3.18). Also for simplicity, we consider only homogeneous boundary conditions and assume that X and X^h are constrained by the boundary condition $\mathcal{B}u = 0$.

Our last assumption deals with the possibility that the discrete operator $S_Y^h \circ \mathcal{L}^h$ may not necessarily be defined for all functions in X .

Assumption 3.27 There exist subspaces $\tilde{X} \subseteq X$ and $\tilde{Y} \subseteq Y$ such that $S_Y^h \circ \mathcal{L}^h u$ is defined for every $u \in \tilde{X}$ and the energy balance (3.18) holds with \tilde{X} and \tilde{Y} . \square

3.5.1 Compliant Discrete Least-Squares Principles

We say that²⁸ $\{J^h, X^h\}$ is a *compliant* DLSP for $\{J, X\}$ if

- X^h is a finite dimensional subspace of X
- $J^h \equiv J$.

LSFEMs corresponding to compliant least-squares principles are exactly the straightforward LSFEMs discussed in Section 2.2.1. A compliant DLSP is obtained by restricting the minimization process in $\{J, X\}$ to a finite dimensional subspace of X , i.e., the transition process is completed by simply choosing $X^h \subset X$. As a result, $\{J^h, X^h\}$ inherits²⁹ the energy norm and the energy balance of its continuous counterpart, i.e. $\|\cdot\|_h = \|\cdot\|$ and

$$C_1 \|w^h\|_X \leq \|w^h\|_h \leq C_2 \|w^h\|_X \quad \forall w^h \in X^h \quad (3.62)$$

with the same C_1 and C_2 as in (3.18).

Theorem 3.28 Assume that $\{J, X\}$ is a well-posed least-squares principle for (3.16). Every compliant DLSP $\{J^h, X^h\}$ has a unique minimizer u^h . Moreover, that minimizer satisfies the error estimate

$$\|u - u^h\|_X \leq \frac{C_2}{C_1} \inf_{w^h \in X^h} \|u - w^h\|_X. \quad (3.63)$$

²⁸ Recall that the notation $\{J^h, X^h\}$ implies the variational principle $\min_{X^h} J^h$.

²⁹ This justifies the use of the term *compliant* as a way to indicate that approximation in such DLSPs is confined to the space X^h and that otherwise they are indistinguishable from their parent CLSP.

Proof. Given that $\|\cdot\|_h$ is simply the energy norm $\|\cdot\|$, Assumption 3.15 holds trivially for $\{J^h, X^h\}$. For the same reason, Assumption 3.16 holds with³⁰ $e(\cdot, \cdot) \equiv 0$. From (3.52) in Corollary 3.19,

$$\|u - u^h\|_h \leq \|u - w^h\|_h \quad \forall w^h \in X^h.$$

Using again $\|\cdot\|_h = \|\cdot\|$, we see that in actuality, (3.62) holds not only on X^h but on all of X . Therefore,

$$C_1 \|u - u^h\|_X \leq \|u - w^h\|_h \leq C_2 \|u - w^h\|_X.$$

The theorem follows by taking the infimum over X^h . \square

To estimate the condition numbers of resulting linear systems, we use the inverse inequalities (3.59) and (3.60) as well as (3.61).

Theorem 3.29 *Let $\{J^h, X^h\}$ be a compliant DLSP and A the matrix of the associated linear system of algebraic equations (3.48). Let Assumptions 3.25 and 3.26 hold. Then,³¹*

$$\text{cond}(A) \leq Ch^{-2\omega_{\max}}. \quad (3.64)$$

Proof. Assume that (3.59) holds; the proof in the case of (3.60) is similar. Using (3.59), (3.61), (3.62), and the definition (3.47) of A ,

$$\begin{aligned} h^d |\vec{w}|^2 &\leq \frac{1}{\mu_1} \|w^h\|_0^2 \leq \frac{1}{\mu_1} \|w^h\|_X^2 \leq \frac{1}{\mu_1 C_1} \|w^h\|_h^2 \\ &= \frac{1}{\mu_1 C_1} ((w^h, w^h))_h = \frac{1}{\mu_1 C_1} \vec{w}^T A \vec{w}. \end{aligned}$$

Therefore,

$$\lambda_{\min} \geq h^d \mu_1 C_1.$$

Using the upper bounds in the same inequalities,

$$\begin{aligned} \vec{w}^T A \vec{w} &= ((w^h, w^h))_h = \|w^h\|_h^2 \leq C_2 \|w^h\|_X^2 \\ &\leq C_2 C_I^2 h^{-2\omega_{\max}} \|w^h\|_0^2 \leq \mu_2 C_2 C_I^2 h^{-2\omega_{\max}} h^d |\vec{w}|^2. \end{aligned}$$

Therefore,

$$\lambda_{\max} \leq \mu_2 C_2 C_I^2 h^{-2\omega_{\max} + d}$$

and thus

$$\text{cond}(A) = \frac{\lambda_{\max}}{\lambda_{\min}} \leq \frac{\mu_2 C_2 C_I^2}{\mu_1 C_1} h^{-2\omega_{\max}}. \quad \square$$

³⁰ As a result, we see that compliant least-squares principles are consistent.

³¹ We denote the condition number of the matrix A by $\text{cond}(A)$.

3.5.2 Norm-Equivalent Discrete Least-Squares Principles

We say that $\{J^h, X^h\}$ is a *norm-equivalent* DLSP for $\{J, X\}$ if

- X^h is a finite dimensional subspace of X
- J^h is such that

$$\widehat{C}_1 \|w^h\|_X \leq \|w^h\|_h \leq \widehat{C}_2 \|w^h\|_X \quad \forall w^h \in X^h \quad (3.65)$$

with positive constants \widehat{C}_1 and \widehat{C}_2 independent of h .

Every compliant DLSP is trivially norm equivalent because (3.62) holds for *all* functions in X , including functions from its proper subspace X^h . The converse is not true because (3.65) is only required to hold for discrete functions. This difference turns out to be far less important for the analysis than the requirement that \widehat{C}_1 and \widehat{C}_2 are independent of h .

From (3.18) and (3.65), it is easy to show that $\|\cdot\|_h$ and $\|\cdot\|$ are equivalent norms on X^h .

Lemma 3.30 *Assume that (3.65) holds for $\{J^h, X^h\}$. Then,*

$$\frac{C_1}{\widehat{C}_2} \|w^h\|_h \leq \|w^h\| \leq \frac{C_2}{\widehat{C}_1} \|w^h\|_h \quad \forall w^h \in X^h. \quad \square \quad (3.66)$$

This result implies that norm-equivalent functionals preserve, uniformly in h , the correct energy balance between the discrete data and solution spaces; thus the terminology *norm-equivalent* is fully justified. The following theorem shows that this is enough to guarantee that norm-equivalent functionals enjoy virtually the same computational properties, with regard to accuracy and matrix conditioning, as their compliant cousins.

Theorem 3.31 *Let $\{J^h, X^h\}$ be a norm-equivalent DLSP associated with a well-posed CLSP $\{J, X\}$ for (3.16). Assume that the minimizer u of $\{J, X\}$ belongs to the space \widetilde{X} from Assumption 3.27. Then, for every $h > 0$, $\{J^h, X^h\}$ has a unique minimizer u^h . Moreover, that minimizer satisfies the error estimate*

$$\begin{aligned} \|u - u^h\|_X &\leq \frac{1}{C_1} \inf_{w^h \in X^h} \|u - w^h\| \\ &+ \frac{C_2}{C_1 \widehat{C}_1} \left(2 \inf_{w^h \in X^h} \|u - w^h\|_h + \sup_{w^h \in X^h} \frac{e(u, w^h)}{\|w^h\|_h} \right). \end{aligned} \quad (3.67)$$

Proof. A norm-equivalent functional trivially satisfies Assumption 3.15 so that, as a result, the existence and uniqueness of the discrete least-squares solution follows from Theorem 3.17.

Regarding the error estimate, there are two factors that prevent us from reusing the proof of Theorem 3.28. The first is that the error $u - u^h$ does not belong to X^h and

thus (3.65) is not applicable to this function. The second factor is that u^h minimizes the *discrete* rather than the continuous energy norm. This means that the inequality

$$\|u - u^h\| \leq \|u - w^h\|$$

does not hold for all $w^h \in X^h$.

So, instead, we begin by bounding the error from above by using the continuous energy balance (3.18):

$$\|u - u^h\|_X \leq \frac{1}{C_1} \|u - u^h\|. \quad (3.68)$$

A repeated use of the triangle inequality, application of the norm equivalence (3.66), and the error estimate (3.45) from Theorem 3.17 yield the following estimate of the continuous energy norm: for all $w^h \in X^h$,

$$\begin{aligned} \|u - u^h\| &\leq \|u - w^h\| + \|w^h - u^h\| \\ &\leq \|u - w^h\| + \frac{C_2}{C_1} \|w^h - u^h\|_h \\ &\leq \|u - w^h\| + \frac{C_2}{C_1} (\|u - u^h\|_h + \|u - w^h\|_h) \\ &\leq \|u - w^h\| + \frac{C_2}{C_1} \|u - w^h\|_h \\ &\quad + \frac{C_2}{C_1} \left(\inf_{w^h \in X^h} \|u - w^h\|_h + \sup_{w^h \in X^h} \frac{e(u, w^h)}{\|w^h\|_h} \right). \end{aligned}$$

Because the last inequality holds for an arbitrary $w^h \in X^h$, it follows that

$$\|u - u^h\| \leq \inf_{w^h \in X^h} \|u - w^h\| + \frac{C_2}{C_1} \left(2 \inf_{w^h \in X^h} \|u - w^h\|_h + \sup_{w^h \in X^h} \frac{e(u, w^h)}{\|w^h\|_h} \right).$$

The theorem follows by combining the last bound and (3.68). \square

The next result shows that, for a given CLSP $\{J, X\}$, the conditioning of the systems generated by compliant and norm-equivalent discrete principles is essentially the same.

Theorem 3.32 *Let $\{J^h, X^h\}$ be a norm-equivalent DLSP and A the matrix of the associated linear system of algebraic equations (3.48). Let Assumptions 3.25 and 3.26 hold. Then,*

$$\text{cond}(A) \leq Ch^{-2\omega_{\max}}. \quad (3.69)$$

Proof. In contrast to Theorem 3.31, the proof of Theorem 3.29 carries over to the norm-equivalent case without modifications. The reason for this is that estimates of the largest and the smallest eigenvalues of A require only the discrete energy balance (3.65), the inverse inequalities (3.59) and (3.60), and assumption (3.61). \square

Before moving on to the next class of discrete principles, let us stress again that norm-equivalent DLSPs *are not compliant principles*. In particular, a norm-equivalent functional J^h is *different* from J , may not be meaningful outside X^h , and its construction can be rather complicated.

3.5.3 Quasi-Norm-Equivalent Discrete Least-Squares Principles

We say that $\{J^h, X^h\}$ is a *quasi-norm-equivalent* DLSP for $\{J, X\}$ if

- X^h is a finite dimensional subspace of X ;
- J^h is such that

$$\widehat{C}_1(h) \|w^h\|_X \leq \|w^h\|_h \leq \widehat{C}_2(h) \|w^h\|_X \quad \forall w^h \in X^h, \quad (3.70)$$

where $\widehat{C}_1(h) > 0$ and $\widehat{C}_2(h) > 0$ for all $h > 0$.

Virtually every conforming DLSP obtained by the *minimize and then discretize* approach, including compliant and norm-equivalent DLSPs, falls within this definition and thus can be deemed quasi-norm-equivalent. However, (3.70) admits DLSPs that are neither compliant nor norm-equivalent. This necessitates some fundamental changes in the way one approaches the analysis of quasi-norm-equivalent principles.

Because the lower and the upper bounds in (3.70) involve mesh-dependent comparability constants, standard elliptic arguments based on continuity and coercivity inevitably lead to suboptimal error estimates. This does not necessarily reflect the true behavior of quasi-norm-equivalent DLSPs; as seen below, such methods *can* produce optimally accurate approximations despite the mesh dependence of the constants in (3.70). However, to reveal such behavior, error estimates must be based on carefully constructed duality arguments. This, of course, requires additional assumptions about the regularity of solutions of (3.16). Before stating the assumptions, recall that $Y = Y_1 \times \cdots \times Y_M$ and $X = X_1 \times \cdots \times X_M$. Let Y_i^* denote the dual of Y_i with $L^2(\Omega)$ acting as a pivot space, and d_i the number of scalar components of $y \in Y_i$.

Assumption 3.33 For all $i = 1, \dots, M$, $Y_i^* \subseteq [L^2(\Omega)]^{d_i} \subseteq Y_i$. □

Assumption 3.34 There exist spaces $X^{(i)} \subset X$ and $Y^{(i)} \subset Y$, $i = 1, \dots, M$, such that³²

$$Y^{(i)} = Y_1^{(i)} \times \cdots \times Y_i^* \times \cdots \times Y_M^{(i)}$$

and, for all $u \in X^{(i)}$, we have the “stronger” energy balance

$$C_1 \|u\|_{X^{(i)}} \leq \|\mathcal{L}u\|_{Y^{(i)}} \leq C_2 \|u\|_{X^{(i)}}. \quad \square \quad (3.71)$$

³² Note that the i th component space of $Y^{(i)}$ is required to coincide with the i th component of the dual space Y^* . However, the remaining components of $Y^{(i)}$ are in no way restricted to be components of Y^* or the original space Y .

Assumption 3.35 For all $i = 1, \dots, M$, the discrete norm-generating operators $\mathcal{S}_{Y_i}^h$ can be extended to symmetric, positive definite operators $Y_i^* \mapsto Y_i^*$. \square

Assumption 3.36 For all $i = 1, \dots, M$, the problem (3.16) is $X^{(i)}$ -regular in the sense that, for every $f_i \in Y_i^*$, the system

$$\begin{cases} \mathcal{L}v = f^{(i)} & \text{in } \Omega \\ \mathcal{B}v = 0 & \text{on } \partial\Omega, \end{cases} \quad (3.72)$$

where $f^{(i)} = (0, \dots, 0, f_i, 0, \dots, 0)^\top \in Y^{(i)}$, has a solution $v^{(i)} \in X^{(i)}$ that satisfies (3.71). \square

As a consequence of (3.71) and (3.72), we have that

$$C_1 \|v^{(i)}\|_{X^{(i)}} \leq \|f^{(i)}\|_{Y_i} = \|f\|_{Y_i^*}. \quad (3.73)$$

The key to the analysis of quasi-norm-equivalent principles is the ability to bound the continuous energy norm associated with the CLSP $\{J, X\}$ by the discrete energy norm associated with the quasi-norm-equivalent DLSP. The following lemma establishes a general result of this type that is valid under the additional assumptions made above.

Lemma 3.37 *Let $w \in \tilde{X}$ and let Assumptions 3.15, 3.16, and 3.33–3.36 hold. Then,*

$$\begin{aligned} \|w\| &\leq \|(\mathcal{L} - \mathcal{L}^h)w\|_Y \\ &+ \sum_{i=1}^M \sup_{f_i \in Y_i^*} \left\{ \frac{\sum_{k=1}^M \left(\mathcal{L}_k^h w, (\mathcal{S}_{Y_k}^h)^2 \circ (\mathcal{L}_k - \mathcal{L}_k^h) v^{(i)} \right)}{\|f_i\|_{Y_i^*}} \right\} \\ &+ \sum_{i=1}^M \sup_{f_i \in Y_i^*} \frac{((w, v^{(i)}))_h}{\|f_i\|_{Y_i^*}}, \end{aligned} \quad (3.74)$$

where, for $i = 1, \dots, M$, the functions $v^{(i)} \in Y^{(i)}$ solve the problems

$$\begin{cases} \mathcal{L}v^{(i)} = f_S^{(i)} & \text{in } \Omega \\ \mathcal{B}v^{(i)} = 0 & \text{on } \partial\Omega \end{cases} \quad (3.75)$$

for

$$f_S^{(i)} = (0, \dots, 0, (\mathcal{S}_{Y_i}^h)^{-2} f_i, 0, \dots, 0) \in Y^{(i)}.$$

Proof. For the sake of brevity, we use $\mathcal{L}_k u$ to denote $\sum_{j=1}^M \mathcal{L}_{kj} u_j$. Recall that

$$\|w\| = \|\mathcal{L}w\|_Y = \sum_{k=1}^M \|\mathcal{L}_k w\|_{Y_k}$$

and

$$\|w\|_h = \|\mathcal{S}_Y^h \circ \mathcal{L}^h w\|_0 = \sum_{k=1}^M \|\mathcal{S}_{Y_k}^h \circ \mathcal{L}_k^h w\|_0.$$

Let $i = 1, \dots, M$. Using duality,

$$\begin{aligned} \|\mathcal{L}_i w\|_{Y_i} &= \sup_{f_i \in Y_i^*} \frac{(\mathcal{L}_i w, f_i)}{\|f_i\|_{Y_i^*}} = \sup_{f_i \in Y_i^*} \frac{(\mathcal{L}_i^h w, f_i) + ((\mathcal{L}_i - \mathcal{L}_i^h)w, f_i)}{\|f_i\|_{Y_i^*}} \\ &\leq \sup_{f_i \in Y_i^*} \frac{(\mathcal{L}_i^h w, f_i)}{\|f_i\|_{Y_i^*}} + \|(\mathcal{L}_i - \mathcal{L}_i^h)w\|_{Y_i}. \end{aligned}$$

According to (3.75), we have that $f_i = (\mathcal{S}_{Y_i}^h)^2 \circ \mathcal{L}_i v^{(i)}$ and $\mathcal{L}_k v^{(i)} = 0$ for $k \neq i$. Using this and the definition of the discrete inner product $((\cdot, \cdot))_h$ gives

$$\begin{aligned} \frac{(\mathcal{L}_i^h w, f_i)}{\|f_i\|_{Y_i^*}} &= \frac{(\mathcal{L}_i^h w, (\mathcal{S}_{Y_i}^h)^2 \circ \mathcal{L}_i v^{(i)})}{\|f_i\|_{Y_i^*}} = \frac{\sum_{k=1}^M (\mathcal{L}_k^h w, (\mathcal{S}_{Y_k}^h)^2 \circ \mathcal{L}_k v^{(i)})}{\|f_i\|_{Y_i^*}} \\ &= \frac{\sum_{k=1}^M (\mathcal{S}_{Y_k}^h \circ \mathcal{L}_k^h w, \mathcal{S}_{Y_k}^h \circ \mathcal{L}_k v^{(i)})}{\|f_i\|_{Y_i^*}} + \frac{\sum_{k=1}^M (\mathcal{L}_k^h w, (\mathcal{S}_{Y_k}^h)^2 \circ (\mathcal{L}_k - \mathcal{L}_k^h) v^{(i)})}{\|f_i\|_{Y_i^*}} \\ &= \frac{((w, v^{(i)}))_h}{\|f_i\|_{Y_i^*}} + \frac{\sum_{k=1}^M (\mathcal{L}_k^h w, (\mathcal{S}_{Y_k}^h)^2 \circ (\mathcal{L}_k - \mathcal{L}_k^h) v^{(i)})}{\|f_i\|_{Y_i^*}}. \end{aligned}$$

Combining with the previous inequality,

$$\begin{aligned} \|\mathcal{L}_i w\|_{Y_i} &\leq \|(\mathcal{L}_i - \mathcal{L}_i^h)w\|_{Y_i} + \sup_{f_i \in Y_i^*} \frac{((w, v^{(i)}))_h}{\|f_i\|_{Y_i^*}} \\ &\quad + \sup_{f_i \in Y_i^*} \left\{ \frac{\sum_{k=1}^M (\mathcal{L}_k^h w, (\mathcal{S}_{Y_k}^h)^2 \circ (\mathcal{L}_k - \mathcal{L}_k^h) v^{(i)})}{\|f_i\|_{Y_i^*}} \right\}. \end{aligned}$$

The lemma follows by summing the bounds for $i = 1, \dots, M$. \square

The first two terms in the upper bound (3.74) represent consistency errors. For consistent least-squares functionals, Lemma 3.37 can be substantially simplified.

Corollary 3.38 *Assume that $\{J^h, X^h\}$ is consistent. For every $w \in \tilde{X}$*

$$\|w\| \leq \sum_{i=1}^M \sup_{f_i \in Y_i^*} \frac{((w, v^{(i)}))_h}{\|f_i\|_{Y_i^*}}, \quad (3.76)$$

where the $v^{(i)}$ are defined by (3.75). \square

The key idea in the analysis of quasi-norm-equivalent DSLPs is to use bounds such as (3.74) or (3.76) instead of the mesh-dependent inequalities in (3.70) to estimate the error in the norm of X . For consistent discrete functionals, this approach is illustrated in the next theorem. The general case differs by the presence of truncation error terms and, for brevity, is not considered.

Theorem 3.39 *Assume that $\{J^h, X^h\}$ is a consistent quasi-norm-equivalent DLSP associated with a well-posed CLSP $\{J, X\}$ for (3.16). Let Assumptions 3.15, 3.16, and 3.33–3.36 hold and assume that the minimizer u of $\{J, X\}$ belongs³³ to \tilde{X} . Then, for every $h > 0$, $\{J^h, X^h\}$ has a unique minimizer u^h . Moreover, that minimizer satisfies the error estimate*

$$\|u - u^h\|_X \leq \frac{1}{C_1} \inf_{w^h \in X^h} \|u - w^h\|_h \sum_{i=1}^M \sup_{f_i \in Y_i^*} \left\{ \frac{\inf_{w^h \in X^h} \|v^{(i)} - w^h\|_h}{\|f_i\|_{Y_i^*}} \right\}. \quad (3.77)$$

Proof. Because every quasi-norm-equivalent DLSP satisfies Assumptions 3.15 and 3.16, the existence and uniqueness of u^h follow from the general result in Theorem 3.17. Using the continuous energy balance (3.18) and (3.76), we bound the error in the norm of X as follows:

$$\|u - u^h\|_X \leq \frac{1}{C_1} \|u - u^h\| \leq \frac{1}{C_1} \sum_{i=1}^M \sup_{f_i \in Y_i^*} \frac{((u - u^h, v^{(i)}))_h}{\|f_i\|_{Y_i^*}}.$$

By assumption, $\{J^h, X^h\}$ is consistent so that

$$((u - u^h, v^{(i)}))_h = ((u - u^h, v^{(i)} - w^h))_h \quad \forall w^h \in X^h.$$

As a result, using the Cauchy inequality,

$$((u - u^h, v^{(i)}))_h \leq \|u - u^h\|_h \|v^{(i)} - w^h\|_h \quad \forall w^h \in X^h.$$

Because the last inequality holds for all w^h , it is easy to see that

$$\frac{((u - u^h, v^{(i)}))_h}{\|f_i\|_{Y_i^*}} \leq \|u - u^h\|_h \frac{\inf_{w^h \in X^h} \|v^{(i)} - w^h\|_h}{\|f_i\|_{Y_i^*}}.$$

The theorem follows by taking supremum over Y_i^* and using (3.52). \square

Remark 3.40 First, we note that the discrete energy balance (3.70) is not used at all in the proof of the error estimates. As a result, the mesh-dependence of its lower and upper bounds is prevented from entering explicitly into the error bounds. This, however, does not automatically imply that (3.77) is optimal with respect to the norm of X . Whether this is true ultimately depends on the structure of the discrete

³³ See Assumption 3.27 for the definition of the space \tilde{X} .

energy norm and the order of the approximation error for *a set of sufficiently regular functions* as measured by this norm. In other words, the quality of the quasi-norm-equivalent least-squares approximation is strongly affected by the properties of the term

$$\sum_{i=1}^M \sup_{f_i \in Y_i^*} \frac{\inf_{w^h \in X^h} \|v^{(i)} - w^h\|_h}{\|f_i\|_{Y_i^*}}. \quad (3.78)$$

In Chapter 4, we consider examples of quasi-norm-equivalent methods for first-order elliptic systems that illustrate the importance of this term in the error analysis. There, we also discuss in more detail the conditioning of the algebraic systems engendered by quasi-norm-equivalent methods, as this requires more detailed knowledge of the spaces and operators involved in setting up the method.

In particular, in Section 4.5, we present two methods that have identical mesh-dependent energy balances but differ tremendously in their asymptotic order of accuracy. One of them uses trivial approximations of norm-generating operators that causes (3.78) to contribute negative powers of h , thereby reducing the order of accuracy in (3.77). By using a judicious choice of discrete operators \mathcal{S}_Y^h , the other method simultaneously keeps this term of $O(1)$ and ensures that $\|u - u^h\|_h$ is of optimal-order of accuracy for all sufficiently smooth functions u . As a result, for this method, (3.77) is optimal. \square

3.5.4 Summary Review of Discrete Least-Squares Principles

This chapter provides the necessary tools to imbue the informal least-squares classification scheme given in Figure 2.1 with precise mathematical meaning. In particular, analyses of compliant, norm-equivalent, and quasi-norm-equivalent DLSPs show that recovery of a Rayleigh–Ritz-like setting for the finite element method becomes increasingly difficult as DLSPs deviate more and more from the compliant setting of Section 3.5.1. This observation underscores the key role of norm-equivalence in LSFEMs and validates our decision to base the classification scheme in Figure 2.1 on this key property. A concise summary of the properties of the three basic types of conforming DLSPs is presented in Table 3.1.

DLSP→ Property↓	Conforming			Non-conforming
	compliant	norm-equivalent	quasi-norm-equivalent	
Energy balance	independent of h	independent of h	dependent on h	varies
Error estimate	standard elliptic argument	uses norm-equivalence	uses duality argument	varies
Condition number	$h^{-2\omega_{\max}}$	$h^{-2\omega_{\max}}$	varies	varies

Table 3.1 Comparison of different classes of discrete least-squares principles.



<http://www.springer.com/978-0-387-30888-3>

Least-Squares Finite Element Methods

Bochev, P.B.; Gunzburger, M.D.

2009, XXII, 660 p., Hardcover

ISBN: 978-0-387-30888-3