1.3 Variational Numerical Methods

The abstract variational equation (1.61) inspires the formulation of a class of **numerical methods**. However, before describing them, we will discuss what a numerical method is.

What is a numerical method? A numerical method for a problem such as Problem 1.1 is a definition of a sequence of functions $\{u_{h1}, u_{h2}, \dots u_{hn}, \dots\}$ that *can be computed* and such that they *approximate* the exact solution u as $n \to \infty$.

The phrase "can be computed" means that functions $u_{h1}, u_{h2}, \dots$ are not implicitly defined but their values can be explicitly computed, limited only by available computational resources.

The most important requirement, however, is that the sequence $u_{h1}, u_{h2}, \dots u_{hn}, \dots$ approximates the solution of the problem to any degree of accuracy provided we choose n large enough. For example, the method could guarantee that for some value n the maximum difference between the u and u_{hn} is smaller than a desired tolerance everywhere in the domain.

In the following, we will be concerned with defining finite element methods and their implementation, and postpone the discussion of their approximation properties to §**??**. In other words, we will first discuss *what* finite element methods are, and then we will discuss *how* finite element methods provide approximations.

1.3.1 Variational Methods

A variational numerical numerical method defines approximations by using the following observation. Consider a solution u of a problem that satisfies a variational equation of the form

$$F(u, v) = 0 \qquad \forall\, v \in \mathcal{V}$$

for some function F and test space $\mathcal{V}$.

A variational method is defined by finite dimensional spaces function spaces $\mathcal{V}_h$ and $\mathcal{S}_h$. It defines an approximation u_h of u by finding $u_h \in \mathcal{S}_h$ that satisfies that

$$F(u_h, v_h) = 0 \qquad \forall\, v \in \mathcal{V}_h. \tag{1.65}$$

It is also possible to consider a variational equation that changes with h, i.e. a function F_h. We will see examples of this later.

The space $\mathcal{S}_h$ is called the **trial space**, it is an *affine space* where u_h is sought. The space $\mathcal{V}_h$ is a *vector space* that acts as the test space for variational equation (1.65). By prescribing a way to select $\mathcal{V}_h$, $\mathcal{S}_h$ for different values of h, the method can define a sequence $\{u_{h1}, u_{h2}, \dots\}$ that approximates u.

The simplest scenario, and the one we will be concerned with, is that in which F defines a linear variational equation and the test space is the direction of the trial space. Such methods define the approximation u_h as the solution of a problem of the following type.

Problem 1.2 (Variational Method)**.** *Let $\mathcal{W}_h$ be a finite-dimensional vector space, $a\colon \mathcal{W}_h \times \mathcal{W}_h \to \mathbb{R}$ be a bilinear form, $\ell\colon \mathcal{W}_h \to \mathbb{R}$ be a linear functional, $\mathcal{S}_h \subseteq \mathcal{W}_h$ and $\mathcal{V}_h \subseteq \mathcal{W}_h$, with $\mathcal{V}_h$ the direction of $\mathcal{S}_h$.*

$$\text{Find } u_h \in \mathcal{S}_h \text{ such that } a(u_h, v_h) = \ell(v_h) \text{ for all } v_h \in \mathcal{V}_h. \tag{1.66}$$

The variational equation F and the trial and test spaces need to be selected so that they can provide an approximation to the solution of the problem of interest.

Different choices of a, ℓ, and $\mathcal{S}_h$ lead to different variational methods.
Finite element methods are a type of variational numerical methods.

Let's look at a simple example of such problem and its solution.

Example 1.51 Consider the linear variational equation in Example 1.49 with $\Omega = [0,1]$, $k(x) = 1$, $b(x) = c(x) = 0$, $f(x) = 1$ and $d_L = 0$.

Let

$$\mathcal{W}_h = \operatorname{span}(1, x, x^2, x^3)$$

so that if $w_h \in \mathcal{W}_h$, then $w_h = w_0 \cdot 1 + w_1 x + w_2 x^2 + w_3 x^3$. For this example, we are going to seek a function u_h that satisfy a Dirichlet boundary condition at $x = 0$, in this case $u_h(0) = 2$, so we will set the trial space to

$$\begin{aligned}\mathcal{S}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = 2\} \\ &= \{w_h = 2 + w_1 x + w_2 x^2 + w_3 x^3 \mid (w_1, w_2, w_3) \in \mathbb{R}^3\}.\end{aligned} \tag{1.67}$$

For $\mathcal{V}_h$, we find the direction of $\mathcal{S}_h$ to get

$$\begin{aligned}\mathcal{V}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = 0\} \\ &= \operatorname{span}(x, x^2, x^3).\end{aligned}$$

The method consists in finding $u_h \in \mathcal{S}_h$ such that

$$a(u_h, v_h) \int_0^1 u_h' v_h' \, dx = \int_0^1 v_h \, dx = \ell(v_h) \tag{1.68}$$

for all $v_h \in \mathcal{V}_h$.

We can now find $u_h = 2 + u_1 x + u_2 x^2 + u_3 x^3$. To this end, we will find u_1, u_2 and u_3 by testing the variational equation with each function in the basis for $\mathcal{V}_h$. To wit,

$$\begin{aligned}\int_0^1 (u_1 + 2u_2 x + 3u_3 x^2)\, dx &= \int_0^1 x \, dx \\ \int_0^1 (u_1 + 2u_2 x + 3u_3 x^2) 2x \, dx &= \int_0^1 x^2 \, dx \\ \int_0^1 (u_1 + 2u_2 x + 3u_3 x^2) 3x^2 \, dx &= \int_0^1 x^3 \, dx.\end{aligned}$$

Evaluating the integrals, we obtain

$$\begin{aligned} u_1 + u_2 + u_3 &= \frac{1}{2} \\ u_1 + \frac{4}{3}u_2 + \frac{3}{2}u_3 &= \frac{1}{3} \\ u_1 + \frac{3}{2}u_2 + \frac{9}{5}u_3 &= \frac{1}{4}. \end{aligned}$$

This defines a system of 3 equations with 3 unknonws, u_1, u_2 and u_3. The solution is $u_1 = 1$, $u_2 = -1/2$ and $u_3 = 0$. Therefore,

$$u_h = 2 + x - x^2/2.$$

1.3.1.1 Choice of Trial and Test Spaces

In §1.1.2 we showed that the solution u of a problem such as Problem 1.1 satisfies different variational equations. Each variational equation separates the boundary conditions in Problem 1.1 into two classes, essential and natural boundary conditions. The same boundary condition can be essential for one variational equation, and natural for another. For example, the Dirichlet boundary condition (1.12a) is an essential boundary condition for variational equation (1.9a) that we obtain from the recipe in §1.1.2.3, but it is a natural boundary condition for the variational equation in Nitsche's method, (1.31). Therefore, boundary conditions are not intrinsically essential or natural, but they acquire that role for a given variational equation.

When approximating the solution u of a problem such as Problem 1.1 with a variational numerical method, we will first need to identify whether the boundary conditions of the problem are essential or natural for the variational equation defined by F (or by a and ℓ in Problem 1.2). Then, any essential boundary condition will need to be enforced through the choice of the trial space $\mathcal{S}_h$. The simplest way to enforce them is to require that any function that belongs to $\mathcal{S}_h$ satisfy the essential boundary conditions. Natural boundary conditions are going to be enforced by the variational equation.

Later in §YYY, we will see that in many situations we will need to require functions in $\mathcal{S}_h$ to satisfy essential boundary conditions only approximately, since it is not possible to exactly enforce them.

Given a variational equation $F(u, v) = 0$ for all $v \in \mathcal{V}$ that u satisfies, a variational method for it is:

$$\text{Find } u_h \in \mathcal{S}_h \text{ such that } F(u_h, v_h) = 0 \text{ for all } v_h \in \mathcal{V}_h.$$

In this case, it is convenient to also require that $\mathcal{V}_h \subseteq \mathcal{V}$, so that it also holds that

$$F(u, v_h) = 0 \qquad \forall v_h \in \mathcal{V}_h. \tag{1.69}$$

In other words, the exact solution satisfies the variational equations used in the numerical method. A method that satisfies (1.69) is said to be **consistent**, and (1.69) is called a **consistency condition**. This condition will play a crucial role to

guarantee that we can approximate u with the method, as we will the have the chance to discuss in §**??**.

Summarizing,

- $\mathcal{S}_h = \{u_h \in \mathcal{W}_h \mid u_h \text{ satisfies essential boundary conditions for } F\}$
- $\mathcal{V}_h$ = Direction of $\mathcal{S}_h$
- For consistency, we require $\mathcal{V}_h \subseteq \mathcal{V}$.

We will illustrate these ideas by considering three different methods for Problem 1.1 with $k(x) = 1$, $b(x) = c(x) = 0$. That is, the problem is to find $u\colon [0, L] \to \mathbb{R}$ that satisfies that

$$-u''(x) = f(x) \qquad x \in (0, L) \tag{1.70a}$$
$$u(0) = g_0 \tag{1.70b}$$
$$u'(L) = d_L. \tag{1.70c}$$

For concreteness, we set

$$\mathcal{W}_h = \operatorname{span}(1, x, x^2, x^3),$$

so that we can provide explicit expressions for $\mathcal{S}_h$ and $\mathcal{V}_h$. This is the same space we adopted in Example 1.51. In the three example, the method reads:

$$\text{Find } u_h \in \mathcal{S}_h \text{ such that } a_h(u_h, v_h) = \ell(v_h) \text{ for all } v_h \in \mathcal{V}_h.$$

We indicate choices for a_h, ℓ_h, $\mathcal{V}_h$ and $\mathcal{S}_h$.

Examples:

1.52 The most common variational method adopts variational equation (1.16), so

$$a(u_h, v_h) = \int_0^L u_h' v_h' \, dx$$
$$\ell(v_h) = \int_0^L f v_h \, dx + d_L v_h(L).$$

For this variational equation, the Dirichlet boundary condition is essential, so we set

$$\begin{aligned}\mathcal{S}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = g_0\} \\ &= \{g_0 + c_1 x + c_2 x^2 + c_3 x^3 \mid (c_1, c_2, c_3) \in \mathbb{R}^3\}\end{aligned}$$

and its direction is

$$\begin{aligned}\mathcal{V}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = 0\} \\ &= \{c_1 x + c_2 x^2 + c_3 x^3 \mid (c_1, c_2, c_3) \in \mathbb{R}^3\}.\end{aligned}$$

Because functions in $\mathcal{V}_h$ are smooth, $\mathcal{V}_h \subset \mathcal{V} = \{w\colon [0,L] \to \mathbb{R} \text{ smooth } \mid w(0) = 0\}$, and the method is consistent. This is the method we used in Example 1.51.

1.53 **Nitsche's Method.** For this also the method we adopt variational equation (1.31), so for $\mu > 0$,

$$a(u_h, v_h) = \int_0^L u_h' v_h' \, dx + u_h'(0)\, v_h(0) - u_h(0)\, v_h'(0) + \mu u_h(0)\, v_h(0)$$
$$\ell(v_h) = \int_0^L f v_h \, dx + d_L v_h(L) - g_0 v_h'(0) + \mu g_0 v_h(0).$$

All boundary conditions are natural, so we can set

$$\mathcal{S}_h = \mathcal{V}_h = \mathcal{W}_h = \{c_0 + c_1 x + c_2 x^2 + c_3 x^3 \mid (c_0, c_1, c_2, c_3) \in \mathbb{R}^4\}.$$

Since $\mathcal{W}_h$ is a vector space, it is also its own direction. Since $\mathcal{V}_h \subset \mathcal{V} = \{w\colon [0,L] \to \mathbb{R} \text{ smooth}\}$, the method is consistent.

1.54 In this case, we adopt variational equation (1.18), so we set

$$a(u_h, v_h) = -\int_0^L u'' v_h \, dx$$
$$\ell(v_h) = \int_0^L f v_h \, dx.$$

As mentioned in Example 1.12, for this variational equation 1.18 both boundary conditions are essential. Therefore, we need to require them in the definition of $\mathcal{S}_h$:

$$\begin{aligned}\mathcal{S}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = g_0, w_h'(L) = d_L\} \\ &= \{g_0 + c_1 x + c_2 x^2 + c_3 x^3 \mid c_1 + 2c_2 L + 3c_3 L^2 = d_L, (c_1, c_2, c_3) \in \mathbb{R}^3\}.\end{aligned}$$

Its direction is

$$\begin{aligned}\mathcal{V}_h &= \{w_h \in \mathcal{W}_h \mid w_h(0) = 0, w_h'(L) = 0\} \\ &= \{c_1 x + c_2 x^2 + c_3 x^3 \mid c_1 + 2c_2 L + 3c_3 L^2 = 0, (c_1, c_2, c_3) \in \mathbb{R}^3\}.\end{aligned}$$

These two spaces have a non-trivial constraint in their definition, but it is simple to solve for c_1 in both cases and replace in the expression for the functions.

The test space of (1.18) is

$$\mathcal{V} = \{w\colon [0,L] \to \mathbb{R}\},$$

so $\mathcal{V}_h \subset \mathcal{V}$, and the method is consistent.

1.3.2 Solution to a Variational Method

We next describe the general procedure to find the solution of (1.66), regardless of the way we construct the discrete space $\mathcal{W}_h$.

Let $\{N_a\}_{a=1,\dots,m} = \{N_1, \dots N_m\}$, $m \in \mathbb{N}$, be a basis for $\mathcal{W}_h$. Then, the approximate solution $u_h \in \mathcal{S}_h$ of Problem 1.2 and any test function $v_h \in \mathcal{V}_h$ can be written as

$$u_h(x) = \sum_{b=1}^{m} u_b N_b(x)$$
$$v_h(x) = \sum_{a=1}^{m} v_a N_a(x).$$

Additionally, we will assume that the subset of basis functions $\{N_a\}_{a=1,\dots,n}$ with $n \le m$ is a basis for $\mathcal{V}_h$. Graphically,

$$\underbrace{\overbrace{N_1, \dots, N_n}^{\text{Basis for } \mathcal{V}_h}, \dots, N_m}_{\text{Basis for } \mathcal{W}_h}. \tag{1.71}$$

This automatically means that $v_a = 0$ for $n < a \le m$.

The solution u_h satisfies (1.66) for any $v_h \in \mathcal{V}_h$. To find u_h, we will take advantage that we can choose the test functions v_h we can "test" with, and of the fact that u_h belongs to $\mathcal{S}_h$. If we choose enough test functions, we will get enough equations to define u_h completely. We can then show that such u_h satisfies (1.66) for any $v_h \in \mathcal{V}_h$, not only for those chosen as particular test functions.

For this plan, we will select each basis function of $\mathcal{V}_h$ as a test function, namely,

$$\ell(N_a) = a(u_h, N_a) \qquad a = 1, \dots, n. \tag{1.72a}$$

This gives us n algebraic equations, for the m unknown components $\{u_1, \dots, u_m\}$ of u in the basis $\{N_a\}_{a=1,\dots,m}$. The remaining $n-m$ equations follow from the fact that $u_h \in \mathcal{S}_h$. Typically, this means that the remaining equations come from the boundary conditions. To impose them, it is enough to select *any* element $\overline{u}_h$ of $\mathcal{S}_h$, write

$$\overline{u}_h = \underbrace{\overline{u}_1 N_1 + \dots + \overline{u}_n N_n}_{\in \mathcal{V}_h} + \underbrace{\dots + \overline{u}_m N_m}_{\notin \mathcal{V}_h}$$

and set

$$u_b = \overline{u}_b \qquad n < b \le m, \tag{1.72b}$$

which provide the remaining $n-m$ equations needed to completely determine the m components $u_1, \dots, u_m$ of u_h in the basis $\{N_1, \dots, N_m\}$.

The solution to (1.72) amounts to the solution of a linear system of equations. To see this, we first expand u_h in components inside (1.72a) and use the bilinearity of a to get:

$$\ell(N_a) = a(u_h, N_a) = a\left(\sum_{b=1}^{m} u_b N_b, N_a\right) = \sum_{b=1}^{m} a(N_b, N_a) u_b \qquad a = 1, \dots, n. \tag{1.73}$$

We then label

$$F_a = \ell(N_a), \qquad K_{ab} = a(N_b, N_a), \qquad 1 \le a \le n,\ 1 \le b \le m$$

and from (1.72b),

$$F_a = \overline{u}_a, \qquad K_{ab} = \delta_{ab}, \qquad n < a \le m,\ 1 \le b \le m$$

where δ_{ab} is called the **Kronecker Delta** [3], and arrange them in a matrix and two columns vectors

$$K = \begin{bmatrix} K_{11} & K_{12} & \dots & K_{1m} \\ K_{21} & K_{22} & \dots & K_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ K_{m1} & K_{m2} & \dots & K_{mm} \end{bmatrix}, \quad F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}, \text{ and } U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}. \tag{1.74}$$

☞ We could be a little bit more specific about K, and use that we know the values for $a, b > n$. Namely,

$$K = \begin{bmatrix} K_{11} & \dots & K_{1(n+1)} & \dots & K_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ K_{n1} & \dots & K_{n(n+1)} & \dots & K_{nm} \\ 0 & \dots & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 1 \end{bmatrix}$$

Then, (1.72) is expressed as the linear system of equations

$$KU = F. \tag{1.75}$$

The matrix K is often called the **stiffness matrix** and the vector F is often called the **load vector**, for their origins in mechanical problems.

Solving the linear system (1.75) for U defines the components $u_1, \dots, u_m$ needed to construct the function $u_h = u_1 N_1 + \dots + u_m N_m$, the solution to the variational method.

Example 1.55 Let's revisit example 1.51, to see that we have solved it exactly as we outlined here. The spaces in the example are:

$$\mathcal{W}_h = \text{span}(1, x, x^2, x^3)$$
$$\mathcal{V}_h = \text{span}(x, x^2, x^3)$$
$$\mathcal{S}_h = \{2 + v_h \mid v_h \in \mathcal{V}_h\}.$$

Hence, we have $m = 4$ and $n = 3$, and index the basis functions in $\mathcal{W}_h$ so that indices 1 to 3 form a basis for $\mathcal{V}_h$. To wit, we set $N_1(x) = x$, $N_2(x) = x^2$, $N_3(x) = x^3$ and $N_4(x) = 1$. We also need to choose one element $\overline{u}_h$ of $\mathcal{S}_h$. Among the infinite options we have here, one is $\overline{u}_h(x) = 2N_4(x) = 2$, and another one is $\overline{u}_h(x) = 3N_1(x) + 2N_4(x)$. Notice that regardless of what we choose for $\overline{u}_h$, all of them will have $\overline{u}_4 = 2$, since this is the only way to construct the constant function 2 needed to belong to $\mathcal{S}_h$.

With these choices, the equations imposed by the variational method, (1.72a), are

$$a(u_h, N_1) = \ell(N_1)$$
$$a(u_h, N_2) = \ell(N_2)$$
$$a(u_h, N_3) = \ell(N_3)$$

[3] It is defined as $\delta_{ab} = \begin{cases} 1 & a = b, \\ 0 & a \neq b. \end{cases}$

while the equations that impose that $u_h \in \mathscr{S}_h$, (1.72b), is

$$u_4 = 2.$$

Replacing, the load vector is

$$F = \begin{bmatrix} \ell(N_1) \\ \ell(N_2) \\ \ell(N_3) \\ u_4 \end{bmatrix} = \begin{bmatrix} \int_0^1 x\,dx \\ \int_0^1 x^2\,dx \\ \int_0^1 x^3\,dx \\ \overline{u}_4 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/3 \\ 1/4 \\ 2 \end{bmatrix}.$$

The stiffness matrix is

$$\begin{aligned} K &= \begin{bmatrix} a(N_1,N_1) & a(N_2,N_1) & a(N_3,N_1) & a(N_4,N_1) \\ a(N_1,N_2) & a(N_2,N_2) & a(N_3,N_2) & a(N_4,N_2) \\ a(N_1,N_3) & a(N_2,N_3) & a(N_3,N_3) & a(N_4,N_3) \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \int_0^1 1\cdot 1\,dx & \int_0^1 2x\cdot 1\,dx & \int_0^1 3x^2\cdot 1\,dx & \int_0^1 0\cdot 1\,dx \\ \int_0^1 1\cdot 2x\,dx & \int_0^1 2x\cdot 2x\,dx & \int_0^1 3x^2\cdot 2x\,dx & \int_0^1 0\cdot 2x\,dx \\ \int_0^1 1\cdot 3x^2\,dx & \int_0^1 2x\cdot 3x^2\,dx & \int_0^1 3x^2\cdot 3x^2\,dx & \int_0^1 0\cdot 3x^2\,dx \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 4/3 & 3/2 & 0 \\ 1 & 3/2 & 9/5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

The components of u_h are then

$$U = K^{-1}F = \begin{bmatrix} 1 \\ -1/2 \\ 0 \\ 2 \end{bmatrix}, \tag{1.76}$$

from where

$$u_h(x) = 1.N_1(x) - 1/2N_2(x) + 0.N_3(x) + 2N_4(x) = 2 + x - \frac{x^2}{2}. \tag{1.77}$$

This happens to be the *exact* solution of the problem, whose strong form consists of the following three equations:

$$\begin{aligned} -u''(x) &= 1 \qquad x \in (0,1) \\ u(0) &= 2 \\ u'(1) &= 0. \end{aligned}$$

This can be understood because in this case $\mathscr{S}_h$ *contains* the exact solution of the problem. A *consistent* variational method will *always* find the exact solution if it belongs to $\mathscr{S}_h$ and it is the sole solution of Problem 1.2. In general, however, this will not be the case.

1.3.2.1 Why does this solution procedure work?

We complete the last section by answering two questions about this solution procedure. The first question is:

Why do we test with the basis functions only, if variational equation (1.66) *should hold for* all *test functions?*

The answer is that if the variational equation is satisfied for every function in a basis for the test space $\mathcal{V}_h$, it is satisfied for *every* function in the test space.

The proof is simple, and worth reading, and it takes advantage of the bilinearity of a and the linearity of ℓ:

$$\begin{aligned} a(u_h, v_h) &= a\left(u_h, \sum_{b=1}^{n} v_b N_b\right) \\ &= \sum_{b=1}^{n} v_b a(u_h, N_b) \qquad \text{bilinearity of } a \\ &= \sum_{b=1}^{n} v_b \ell(N_b) \qquad \text{use of (1.72a)} \\ &= \ell\left(\sum_{b=1}^{n} v_b N_b\right) \qquad \text{linearity of } \ell \\ &= \ell(v_h). \end{aligned}$$

So, for $u_h \in \mathcal{S}_h$, (1.72a) implies (1.66). The converse is trivially true, namely, if (1.66) is satisfied for any $v_h \in \mathcal{V}_h$, it is satisfied for any basis function $N_b \in \mathcal{V}_h$ in particular, and hence it implies (1.72a). In summary, if $u_h \in \mathcal{S}_h$,

$$u_h \text{ is a solution of (1.66)} \iff u_h \text{ is a solution of (1.72a).}$$

In words, this implies that the solution of the variational method is a solution of the linear system of equations defined by the basis functions of the test space, and conversely[4].

The second question we answer is:

Why does the solution u_h belong to $\mathcal{S}_h$, and why is it independent of our choice of $\overline{u}_h \in \mathcal{S}_h$?

The answer to this relies on the fact that $\mathcal{S}_h$ is an affine subspace of $\mathcal{W}_h$ and $\mathcal{V}_h$ is its direction. To see that $u_h \in \mathcal{S}_h$, notice that since $u_b = \overline{u}_b$ for $n < b \le m$, then

$$\Delta u_h = u_h - \overline{u}_h = (u_1 - \overline{u}_1)N_1 + \ldots + (u_n - \overline{u}_n)N_n,$$

from where we conclude that $\Delta u_h \in \mathcal{V}_h$, or $u_h = \overline{u}_h + \Delta u_h$, and hence it follows from (1.47) that $u_h \in \mathcal{S}_h$. To see that the choice of $\overline{u}_h$ does not affect the u_h we

[4] It is possible to regard (1.72a) as the Euler-Lagrange equations of (1.66).

compute, consider another function $\overline{w}_h \in \mathcal{S}_h$. Then, by the definition of affine subspace, $\overline{u}_h - \overline{w}_h \in \mathcal{V}_h$, or in terms of the basis for $\mathcal{W}_h$,

$$\overline{u}_h = \overline{w}_h + \sum_{b=1}^{n} v_b N_b.$$

So, $\overline{u}_h$ and $\overline{w}_h$ can only differ in the values of the components $v_1, \ldots, v_n$, but need to have precisely the same values for the components v_b with $n < b \le m$. Since the latter are the only components that participate in (1.72b), the solution u_h does not change if we choose $\overline{w}_h$ instead. In other words, u_h does not depend on our choice of $\overline{u}_h$.

1.3.2.2 Solution to a variational method with an arbitrarily-ordered basis

In general, an ordered basis as in (1.71) is not readily available, nor is it necessary. We discuss next how to proceed in the case in which the basis functions for $\mathcal{W}_h$ and $\mathcal{V}_h$ are not neatly ordered as in the earlier discussion.

Again, let $\{N_a\}_{a=1,\ldots,m}$ be a basis for $\mathcal{W}_h$, and again we will assume that a subset of $n \le m$ of these basis functions is a basis for $\mathcal{V}_h$. However, the basis for $\mathcal{V}_h$ need *not* be the set $\{N_a\}_{a=1,\ldots,n}$. To indicate the basis for $\mathcal{V}_h$, it is convenient to introduce three sets of indices, or **index sets**. First, we denote by $\eta = \{1, \ldots, m\}$ the set of indices of all basis functions in $\mathcal{W}_h$. The basis functions for $\mathcal{V}_h$ can be indicated by a subset of η. The set of indices of basis functions for $\mathcal{V}_h$ is denoted $\eta_a \subset \eta$; it is called the set of **active indices**, and we can write

$$\mathcal{V}_h = \text{span}\left(\bigcup_{a \in \eta_a} \{N_a\} \right)$$

or

$$w_h \in \mathcal{V}_h \Longleftrightarrow w_h = \sum_{a \in \eta_a} w_a N_a.$$

The remaining indices in η, those that are *not* in η_a, is denoted $\eta_g = \eta \setminus \eta_a$; it is called the set of **constrained indices**.

We next rewrite the equations to solve the variational method using these index sets. First, testing with each basis function in $\mathcal{V}_h$, (1.72a), is restated as

$$\ell(N_a) = a(u_h, N_a) \qquad a \in \eta_a \tag{1.78a}$$

The arbitrary element $\overline{u}_h \in \mathcal{S}_h$ used to impose the fact that $u_h \in \mathcal{S}_h$ is still written as $\overline{u}_h = \overline{u}_1 N_1 + \ldots + \overline{u}_m N_m$, but (1.72b) is restated as

$$u_b = \overline{u}_b \qquad b \in \eta_g. \tag{1.78b}$$

We then label

$$\begin{aligned} F_a &= \ell(N_a), \quad & K_{ab} &= a(N_b, N_a) \quad & a \in \eta_a, b \in \eta \\ F_a &= \overline{u}_a, & K_{ab} &= \delta_{ab} & a \in \eta_g, b \in \eta \end{aligned} \tag{1.78c}$$

which define the stiffness matrix K and load vector F.

To illustrate these ideas, let's consider Example 1.55 again.

Example 1.56 Consider Example 1.55 again, but in this case we set $N_1(x) = x$, $N_2(x) = 1$, $N_3(x) = x^2$ and $N_4(x) = x^3$. Therefore, the basis for $\mathcal{V}_h$ is $\{N_1, N_3, N_4\}$, and the index sets are $\eta = \{1,2,3,4\}$, $\eta_a = \{1,3,4\}$, and $\eta_g = \{2\}$. We can then set $\overline{u}_h = 2N_2(x) = 2$.

The stiffness matrix and load vector in this case are

$$F = \begin{bmatrix} \int_0^1 x\,dx \\ \overline{u}_2 \\ \int_0^1 x^2\,dx \\ \int_0^1 x^3\,dx \end{bmatrix} = \begin{bmatrix} 1/2 \\ 2 \\ 1/3 \\ 1/4 \end{bmatrix}.$$

The stiffness matrix is

$$K = \begin{bmatrix} \int_0^1 1\cdot 1\,dx & \int_0^1 0\cdot 1\,dx & \int_0^1 2x\cdot 1\,dx & \int_0^1 3x^2\cdot 1\,dx \\ 0 & 1 & 0 & 0 \\ \int_0^1 1\cdot 2x\,dx & \int_0^1 0\cdot 2x\,dx & \int_0^1 2x\cdot 2x\,dx & \int_0^1 3x^2\cdot 2x\,dx \\ \int_0^1 1\cdot 3x^2\,dx & \int_0^1 0\cdot 3x^2\,dx & \int_0^1 2x\cdot 3x^2\,dx & \int_0^1 3x^2\cdot 3x^2\,dx \end{bmatrix}$$
$$= \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 4/3 & 3/2 \\ 1 & 0 & 3/2 & 9/5 \end{bmatrix}.$$

The components of the solution are

$$U = \begin{bmatrix} 1 \\ 2 \\ -1/2 \\ 0 \end{bmatrix},$$

and the solution is

$$u_h(x) = 1.N_1(x) + 2N_2(x) - 1/2N_3(x) + 0N_4(x) = 2 + x - \frac{x^2}{2}, \qquad (1.79)$$

which is exactly the same function we obtained in Example 1.55.

Comparing the stiffness matrix and load vector in Examples 1.55 and 1.56, we notice that they have the same entries, but reordered: the last row and column in Example 1.55 were moved to be the second row and column in Example 1.56. The solution U in the former has the last row moved to the be the second row in the latter. Of course, the solution u_h is the same in both cases, since the entries in U are multiplied by the reordered basis functions as well.

To conclude this discussion, notice that reordering the basis functions does not change the spaces $\mathcal{V}_h$, $\mathcal{S}_h$ and $\mathcal{W}_h$, and hence it should not change the solution to the variational method.

Example 1.57 Let's look at another example of a variational method, in this case with a basis of trigonometric functions. To this end, we will revisit Example 1.8 in a domain $\Omega = [0, \pi/2]$. The problem is given by

$$-u''(x) + u'(x) + u(x) = -5\exp(-2x), \qquad \forall\, x \in \Omega, \tag{1.80a}$$

$$u(0) = 1, \tag{1.80b}$$

$$u'(\pi/2) = -2\exp(-\pi). \tag{1.80c}$$

and the exact solution is $u(x) = \exp(-2x)$. A variational equation that the solution u satisfies is

$$\int_0^{\pi/2} u'(x)\,v'(x) + u'(x)\,v(x) + u(x)\,v(x)\,dx + 2\exp(-\pi)\,v(\pi/2) = -\int_0^{\pi/2} 5\exp(-2x)\,v(x)\,dx, \tag{1.80d}$$

where

$$\mathcal{V} = \{w\colon [0,1] \to \mathbb{R} \text{ smooth } \mid w(0) = 0\}.$$

The bilinear form and linear functional here are

$$a(u,v) = \int_0^{\pi/2} u'(x)\,v'(x) + u'(x)\,v(x) + u(x)\,v(x)\,dx \tag{1.80e}$$

$$\ell(v) = -\int_0^{\pi/2} 5\exp(-2x)\,v(x)\,dx - 2\exp(-\pi)\,v(\pi/2). \tag{1.80f}$$

The bilinear form is not symmetric. In this variational equation, boundary condition (1.80b) is essential and (1.80c) is natural.

To formulate a variational method, we consider the space

$$\mathcal{W}_h^n = \operatorname{span}(1, \sin x, \ldots, \sin nx)$$

for $n \in \mathbb{N}$. We included a dependence on n for generality, but we will proceed with $n = 2$ next. To this end, we will label $N_1(x) = 1, N_2(x) = \sin x, N_3(x) = \sin 2x$.

Let's find spaces $\mathcal{S}_h$ and $\mathcal{V}_h$. For $w_h \in \mathcal{W}_h^2$, we can write

$$\begin{aligned} w_h(x) &= w_1 N_1(x) + w_2 N_2(x) + w_3 N_3(x) \\ &= w_1.1 + w_2 \sin x + w_3 \sin 2x. \end{aligned}$$

The space $\mathcal{S}_h$ follows by requiring essential boundary condition (1.80b) to be satisfied by the functions in it, namely,

$$\begin{aligned} \mathcal{S}_h &= \{w_h \in W_h^2 \mid w_h(0) = 1\} \\ &= \{1 + w_2 \sin x + w_3 \sin 2x \mid (w_2, w_3) \in \mathbb{R}^2\} \\ \mathcal{V}_h &= \{w_h \in W_h^2 \mid w_h(0) = 0\} \\ &= \{w_2 \sin x + w_3 \sin 2x \mid (w_2, w_3) \in \mathbb{R}^2\}. \end{aligned}$$

Here $\mathcal{V}_h$ is the direction of $\mathcal{S}_h$, and $\mathcal{V}_h \subset \mathcal{V}$, so the method is consistent.

To proceed, we need to identify active and constrained indices. In this case, $\eta_c = \{1\}$ and $\eta_a = \{2,3\}$. The stiffness matrix is then (careful because this is a non-symmetric bilinear form):

$$K = \begin{bmatrix} 1 & 0 & 0 \\ a(N_1,N_2) & a(N_2,N_2) & a(N_3,N_2) \\ a(N_1,N_3) & a(N_2,N_3) & a(N_3,N_3) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & (1+\pi)/2 & 2/3 \\ 1 & 2 & 5\pi/4 \end{bmatrix}.$$

We set $\overline{u}_h(x) = 1$, so that $\overline{u}_h \in \mathcal{S}_h$, and $u_1 = 1$. The load vector is

$$F = \begin{bmatrix} 1 \\ \ell(N_2) \\ \ell(N_3) \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ -5(1+\exp(-\pi))/4 \end{bmatrix}.$$

The components of the solution are obtained from $U = K^{-1}F$, or

$$U = \begin{bmatrix} 1 \\ -\frac{36+20\exp(-\pi)-60\pi}{32-15\pi-15\pi^2} \\ -\frac{3\exp(-\pi)\left(5(1+\pi)+\exp(\pi)(9\pi-23)\right)}{-32+15\pi+15\pi^2} \end{bmatrix} \approx \begin{bmatrix} 1 \\ -0.93 \\ -0.11 \end{bmatrix}.$$

Hence,

$$u_h(x) \approx 1 - 0.93\sin x - 0.11\sin 2x.$$

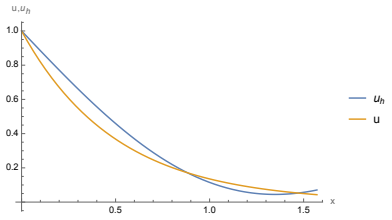

Figure 1.5

A plot of the exact versus the approximate solutions is shown in Fig. 1.5. By selecting a larger value of n, a better approximation is obtained. You can check that.

Example 1.58 Let's consider a twist of Example 1.57 to illustrate the effect of additional boundary conditions on the method. To this end, we change the problem in that example to have a Dirichlet boundary condition at $x = \pi/2$ as well, keeping the same exact solution. The problem is

$$-u''(x) + u'(x) + u(x) = -5\exp(-2x), \qquad \forall\, x \in \Omega, \tag{1.81a}$$
$$u(0) = 1, \tag{1.81b}$$
$$u(\pi/2) = \exp(-\pi). \tag{1.81c}$$

and the exact solution is still $u(x) = \exp(-2x)$. The variational equation that the solution u of this problem satisfies is:

$$\int_0^{\pi/2} u'(x)\,v'(x) + u'(x)\,v(x) + u(x)\,v(x)\,dx = -\int_0^{\pi/2} 5\exp(-2x)\,v(x)\,dx \qquad \forall\, v \in \mathcal{V}, \tag{1.81d}$$

where

$$\mathcal{V} = \{w\colon [0,1] \to \mathbb{R} \text{ smooth } \mid w(0) = 0, w(\pi/2) = 0\}.$$

Notice that there is no longer a term that appears from the natural boundary condition. Both boundary conditions are now essential.

The bilinear form and linear functional here are

$$a(u,v) = \int_0^{\pi/2} u'(x)v'(x) + u'(x)v(x) + u(x)v(x)\,dx \tag{1.81e}$$

$$\ell(v) = -\int_0^{\pi/2} 5\exp(-2x)v(x)\,dx. \tag{1.81f}$$

The bilinear form is not symmetric.

In this case, we consider the space

$$\mathcal{W}_h = \text{span}(1, \sin x, \sin 2x, \sin 4x).$$

We will label $N_1(x) = 1, N_2(x) = \sin x, N_3(x) = \sin 2x,\ N_4(x) = \sin 4x$. Notice that we did not include the function $\sin 3x$.

Let's find spaces $\mathcal{S}_h$ and $\mathcal{V}_h$ next. For $w_h \in \mathcal{W}_h$, we can write

$$\begin{aligned} w_h(x) &= w_1 N_1(x) + w_2 N_2(x) + w_3 N_3(x) + w_4 N_4(x) \\ &= w_1.1 + w_2 \sin x + w_3 \sin 2x + w_4 \sin 4x. \end{aligned}$$

For $w_h \in \mathcal{S}_h$, we need $1 = w_h(0) = w_1$ and $\exp(-\pi) = w_h(\pi/2) = w_1 + w_2$, or $w_1 = 1$ and $w_2 = \exp(-\pi) - 1$. Therefore,

$$\begin{aligned} \mathcal{S}_h &= \{w_h \in W_h \mid w_h(0) = 1, w_h(\pi/2) = \exp(-\pi)\} \\ &= \{1 + (\exp(-\pi) - 1)\sin x + w_3 \sin 2x + w_4 \sin 4x \mid (w_3, w_4) \in \mathbb{R}^2\} \\ \mathcal{V}_h &= \{w_h \in W_h \mid w_h(0) = w_h(\pi/2) = 0\} \\ &= \{w_3 \sin 2x + w_4 \sin 4x \mid (w_3, w_4) \in \mathbb{R}^2\}. \end{aligned}$$

Had we included a term with $\sin 3x$, the characterization of $\mathcal{S}_h$ and $\mathcal{V}_h$ would have been somewhat more complicated, because we would have had a total of 3 functions that are non-zero at $x = \pi/2$.

The active and constrained indices are $\eta_c = \{1,2\}$ and $\eta_a = \{3,4\}$. The stiffness matrix is then (careful because this is a non-symmetric bilinear form):

$$\begin{aligned} K &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ a(N_1,N_3) & a(N_2,N_3) & a(N_3,N_3) & a(N_4,N_3) \\ a(N_1,N_4) & a(N_2,N_4) & a(N_3,N_4) & a(N_4,N_4) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 2 & 5\pi/4 & -4/3 \\ 0 & -4/15 & 4/3 & 17\pi/4 \end{bmatrix}. \end{aligned}$$

We set $\overline{u}_h(x) = 1 + (\exp(-\pi) - 1)\sin x$, so that $\overline{u}_h \in \mathscr{S}_h$, $u_1 = 1$ and $u_2 = \exp(-\pi) - 1$. The load vector is

$$F = \begin{bmatrix} 1 \\ \exp(-\pi) - 1 \\ \ell(N_3) \\ \ell(N_4) \end{bmatrix} = \begin{bmatrix} 1 \\ \exp(-\pi) - 1 \\ 5(1 + \exp(-\pi))/4 \\ \exp(-\pi) - 1 \end{bmatrix}.$$

The components of the solution are obtained from $U = K^{-1}F$, or

$$U = \begin{bmatrix} 1 \\ \exp(-\pi) - 1 \\ \frac{\exp(-\pi)(-e^{\pi}(1216+765\pi)-9945\pi+1216)}{5(256+765\pi^2)} \\ \frac{12\exp(-\pi)(e^{\pi}(4-19\pi)+19\pi+52)}{256+765\pi^2} \end{bmatrix} \approx \begin{bmatrix} 1 \\ -0.96 \\ -0.13 \\ -0.08 \end{bmatrix}.$$

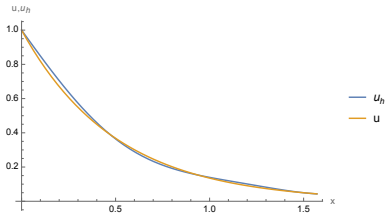

Figure 1.6

The solution is then

$$1 - 0.96\sin x - 0.13\sin 2x - 0.08\sin 4x,$$

and it is plotted in Fig. 1.6.