Chapter 3

Discretization of Optimality Systems

3.1 Introduction

In this chapter, the analysis of discretization of PDE optimality systems is presented in the case of finite difference discretization. Elliptic and parabolic control problems are considered with second-order and higher-order discretization. The case of an optimal control problem governed by a Fredholm integral equation is also discussed. These classes of problems are especially of academic interest, and finite differences are preferred because they allow easy implementation, which makes this choice convenient for teaching purposes. On the other hand, the theoretical framework that we use is valid in a finite element or a finite volume context. In fact, other discretization schemes and other problem classes are discussed in later chapters. In the chapter dedicated to multigrid methods, in Section 5.7.3 we discuss the solution of an elliptic optimality system with FEM discretization. In Section 7.3.3, we discuss a modified Crank–Nicolson scheme for time-dependent bilinear control problems and address the importance of the choice of the optimization space for an efficient numerical optimization process. Further, in Section 7.4, we illustrate the finite difference time domain discretization of the Maxwell equations and of a related optimization problem.

It is clear that a PDE-based optimization problem is naturally defined with infinite-dimensional functional spaces that are Hilbert spaces that concern the PDE model and the formulation of the objective to be optimized. However, the solution of these problems becomes possible after an approximation procedure which involves both the PDE equation and the objective. Furthermore, we know that the solution of a PDE optimization problem is characterized as the solution of the first-order optimality conditions that consist of the PDE model, the corresponding adjoint PDE equation, and the condition of optimality. Now, the discretization step can be applied at the level of the PDE model and of the objective and then followed by the derivation of the optimality condition of the resulting discrete constrained problem, or one applies the discretization step later after the formulation of the optimality system. These two possible discretization procedures are named as follows. The former is called the discretize-before-optimize (DBO) approach, while the latter is called the optimize-before-discretize (OBD) approach. We remark that there are advantages and disadvantages with both approaches which we mention in the following.

On the one hand, the OBD approach has the advantage of allowing a substantially easier formulation of the optimality system which is considered at the continuous level and then each equation of the system is discretized with the preferred method, allowing any different choice of the discretization schemes for the forward and adjoint equations, and the optimality conditions. In particular, one could use approximations of different accuracy [48] and even different grids for the different PDE equations involved.

On the other hand, the DBO approach guarantees a consistent evaluation of the gradient and of the Hessian independently of the discretization used. In fact, the drawback of the OBD approach is the possible inconsistency between the discretized objective and the reduced gradient given by the discrete optimality condition; see [16, 99, 157]. This means discrepancy between the directional derivative $(\nabla \tilde{J}(u), \phi)_{\mathcal{H}}$ and its approximation $\frac{\tilde{J}(u+\epsilon\phi)-\tilde{J}(u-\epsilon\phi)}{2\epsilon}$. The other disadvantage of the OBD approach is that the Hessian may not be symmetric; see, e.g., [29]. This problem cannot arise in the DBO approach. However, it is reasonably clear but less investigated that the differences between the DBO and OBD frameworks can be controlled at the cost of increasing accuracy of discretization by using, e.g., finer meshes and higher-order schemes; see [99]. Another possibility to have the advantages of the OBD scheme with the consistency of the DBO scheme is to discretize the state and adjoint equations with appropriate schemes; see, e.g., [16, 29]. Later in the applications chapter, a dipole quantum control problem is discussed, where the DBO approach becomes advantageous.

For detailed discussions and further results and references on discretization schemes for PDE optimization problems, see [9, 29, 57, 58, 61, 99, 245, 254, 289, 356] and, in particular, [339].

3.2 Discretization of Elliptic Optimization Problems

Consider the following elliptic optimization problem

$$\begin{cases} \min J(y,u) & := & \frac{1}{2}\|y-z\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u\|^2_{L^2(\Omega)}, & \\ -\Delta y & = & u+g & \text{in } \Omega, \\ y & = & 0 & \text{on } \partial\Omega. \end{cases} \tag{3.1}$$

Existence of a unique solution to (3.1) and its characterization are well known [235]. This solution is characterized by the following optimality system

$$\begin{array}{rcll} -\Delta y & = & \frac{1}{\nu}p+g & \text{in } \Omega, \\ y & = & 0 & \text{on } \partial\Omega, \\ -\Delta p & = & -(y-z) & \text{in } \Omega, \\ p & = & 0 & \text{on } \partial\Omega, \\ \nu u - p & = & 0 & \text{in } \Omega. \end{array} \tag{3.2}$$

We have that if $z, g \in L^2(\Omega)$, then $(y,u,p) \in (H^1_0(\Omega) \cap H^2(\Omega))^3$. The fact that u attains the same regularity as the Lagrange multiplier p is due to the special form of the optimality condition. Next, we consider the finite difference discretization of (3.2) with $\nu > 0$.

Consider a sequence of grids $\{\Omega_h\}_{h>0}$ given by

$$\Omega_h = \{\mathbf{x} \in \mathbb{R}^2 : x_i = s_i h, s_i \in \mathbb{Z}\} \cap \Omega.$$

We assume that Ω is a rectangular domain and that the values of the mesh size h are chosen such that the boundaries of Ω coincide with grid lines. For grid functions v_h and w_h defined on Ω_h we introduce the discrete L^2-scalar product

$$(v_h, w_h)_{L_h^2} = h^2 \sum_{\mathbf{x}\in\Omega_h} v_h(\mathbf{x})\, w_h(\mathbf{x}),$$

with associated norm $|v_h|_0 = (v_h, v_h)_{L_h^2}^{1/2}$. We also need $|v_h|_\infty = \max_{\mathbf{x}\in\Omega_h} |v_h(\mathbf{x})|$.

We introduce first-order backward and forward partial space derivatives of v_h in the x_i direction that are denoted by ∂_i^- and ∂_i^+, respectively, and are given by

$$\partial_i^- v_h(\mathbf{x}) = \frac{v_h(\mathbf{x}) - v_h(\mathbf{x} - \hat{\imath}\, h)}{h} \qquad \text{and} \qquad \partial_i^+ v_h(\mathbf{x}) = \frac{v_h(\mathbf{x} + \hat{\imath}\, h) - v_h(\mathbf{x})}{h},$$

where $\hat{\imath}$ denotes the i coordinate direction vector and v_h is extended by 0 on grid points outside of Ω; see [174, 165, 327]. In this framework, the discrete H^1-norm is given by

$$|v_h|_1 = \left(|v_h|_0^2 + \sum_{i=1}^{2} |\partial_i^- v_h|_0^2 \right)^{1/2}.$$

The spaces L_h^2 and H_h^1 consist of the sets of grid functions v_h endowed with $|v_h|_0$, respectively, $|v_h|_1$, as norm. For the definition of H_h^2 we refer the reader to [174] as well. We have the inverse property $|v_h|_2 \le c h^{-1} |v_h|_1$.

Functions in $L^2(\Omega)$ and $H^2(\Omega)$ are approximated by grid functions defined through their mean values with respect to elementary cells $[x_1 - \frac{h}{2}, x_1 + \frac{h}{2}] \times [x_2 - \frac{h}{2}, x_2 + \frac{h}{2}]$. This gives rise to the restriction operators $\tilde{R}_h : L^2(\Omega) \to L_h^2$ and $R_h : H_0^1(\Omega) \cap H^2(\Omega) \to L_h^2$ defined in [174]. For the definition of H_h^2 we refer the reader to [174] as well. Further, we define $\tilde{R}_h^2 : L^2(\Omega) \times L^2(\Omega) \to L_h^2 \times L_h^2$ by $\tilde{R}_h^2 = (\tilde{R}_h, \tilde{R}_h)$ and analogously $R_h^2 = (R_h, R_h)$. For continuous functions $v \in C^k(\bar{\Omega})$, $k = 0, 1, \ldots$, we denote with $(R_h v)(x) = v(x)$ the restriction operator on $\bar{\Omega}_h$; that is, continuous functions are represented by their grid values.

The following property can be proved

$$|\tilde{R}_h v - R_h v|_0 \le c\, h^2\, |v|_{H^2(\Omega)} \quad \forall v \in H^2(\Omega). \tag{3.3}$$

Here and below, c denotes a positive constant which does not depend on the discretization parameters. We denote with V_h the vector space of nodal functions v_h defined on Ω_h which are zero on the boundary. The system of nodal functions (v_h, w_h) is denoted by $\mathcal{V}_h = V_h \times V_h$.

We need the following lemma [327, 328].

Lemma 3.1 (Poincaré–Friedrichs inequality for finite differences). *For any grid function v_h, there exists a constant c_*, independent of $v_h \in V_h$ and h, such that*

$$|v_h|_0^2 \le c_* \sum_{i=1}^{2} |\partial_i^- v_h|_0^2. \tag{3.4}$$

In particular, for $\Omega = (a,b)\times(c,d)$ we have that

$$c_* = \left(\frac{2}{(b-a)^2} + \frac{2}{(d-c)^2}\right)^{-1};$$

see, e.g., [165, 166, 327].

The second-order five-point approximation to the Laplacian with homogeneous Dirichlet boundary conditions is defined by

$$\Delta_h = \partial_1^+\partial_1^- + \partial_2^+\partial_2^-.$$

We have the following consistency result

$$|\Delta_h R_h v - \tilde{R}_h \Delta v|_\infty \le c\, h^2 \|v\|_{C^4(\bar{\Omega})}; \tag{3.5}$$

see, e.g., [174].

Next, we consider an elliptic control problem where the control u has been eliminated using the optimality condition. After discretization, we obtain

$$\left\{\begin{array}{rcl} -\nu\,\Delta_h y_h - p_h &=& \nu\,\tilde{R}_h g, \\ -\Delta_h p_h + y_h &=& \tilde{R}_h z. \end{array}\right. \tag{3.6}$$

To investigate the convergence of the solution of (3.6) to the solution of the continuous optimality system as $h \to 0^+$, we introduce the family of operators [61]

$$\mathcal{A}_h = \begin{pmatrix} -\nu\,\Delta_h & -I_h \\ I_h & -\Delta_h \end{pmatrix}, \tag{3.7}$$

where I_h is the identity operator on grid functions v_h. The operators $\mathcal{A}_h$ are defined between product spaces of grid functions. Here, the cases $\mathcal{A}_h : H_h^1 \times H_h^1 \to H_h^{-1} \times H_h^{-1}$ and $\mathcal{A}_h : H_h^2 \times H_h^2 \to L_h^2 \times L_h^2$ are important. Here H_h^{-1} denotes the dual space of H_h^1 with L_h^2 as pivot space.

The family $\{\mathcal{A}_h\}_{h>0}$ is called H_h^1-regular if $\mathcal{A}_h$ is invertible and there exists a constant C_1 independent of h such that

$$||\mathcal{A}_h^{-1}||_{\mathcal{L}(H_h^{-1}\times H_h^{-1}, H_h^1 \times H_h^1)} \le C_1,$$

and analogously it is called H_h^2-regular if

$$||\mathcal{A}_h^{-1}||_{\mathcal{L}(L_h^2\times L_h^2, H_h^2 \times H_h^2)} \le C_2$$

for C_2 independent of h.

Lemma 3.2. *The family of operators $\{\mathcal{A}_h\}_{h>0}$, with h such that the boundaries of Ω are grid lines, is H_h^1-regular.*

Proof. Let $(v_h, w_h) \in \mathcal{V}_h$ be a pair of grid functions. Then

$$\begin{aligned}(\mathcal{A}_h(v_h,w_h),(v_h,w_h))_{L_h^2\times L_h^2} &= \nu(-\Delta_h v_h, v_h)_{L_h^2} + (-\Delta_h w_h, w_h)_{L_h^2} \\ &\ge \min(\nu,1)\, C \sum_{i=1}^{2} (|\partial_i^- v_h|_0^2 + |\partial_i^- w_h|_0^2),\end{aligned} \tag{3.8}$$

where C is independent of h and arises from the coercivity estimate for $-\Delta_h$, i.e.,

$$(-\Delta_h v_h, v_h)_{L_h^2} \geq C \sum_{i=1}^{2} |\partial_i^- v_h|_0^2 \quad \forall v_h; \tag{3.9}$$

see, e.g., [174]. Using the Poincaré inequality in (3.8) results in

$$(\mathcal{A}_h(v_h, w_h), (v_h, w_h))_{L_h^2 \times L_h^2} \geq C_1^{-2} \, |(v_h, w_h)|^2_{H_h^1 \times H_h^1} \quad \forall (v_h, w_h) \in L_h^2 \times L_h^2,$$

with $C_1^{-2} = \min(\nu, 1) C \, c_0$. Due to the Lax–Milgram lemma, $\mathcal{A}_h$ is invertible. Moreover

$$\|\mathcal{A}_h^{-1}\|_{\mathcal{L}(H_h^{-1} \times H_h^{-1}, H_h^1 \times H_h^1)} \leq C_1 \quad \forall h. \quad \square$$

The infinite-dimensional analogue of $\mathcal{A}_h$ is the operator

$$\mathcal{A} = \begin{pmatrix} -\nu \Delta & -I \\ I & -\Delta \end{pmatrix}, \tag{3.10}$$

where Δ is understood with homogeneous Dirichlet boundary conditions. It is well defined from $H_0^1(\Omega) \times H_0^1(\Omega)$ to $H^{-1}(\Omega) \times H^{-1}(\Omega)$ as well as from $(H^2(\Omega) \cap H_0^1(\Omega)) \times (H^2(\Omega) \cap H_0^1(\Omega))$ to $L^2(\Omega) \times L^2(\Omega)$. We have the following consistency result.

Lemma 3.3. *There exists a constant C_K independent of h such that*

$$\|\mathcal{A}_h R_h^2 - \tilde{R}_h^2 \mathcal{A}\|_{\mathcal{L}((H^2 \cap H_0^1)^2, (H_h^{-1} \times H_h^{-1}))} \leq C_K \, h.$$

Proof. Let $(v, w) \in (H^2(\Omega) \cap H_0^1(\Omega))^2$ and note that due to the consistency property of $-\Delta_h$ as discretization of $-\Delta$ we have

$$\begin{aligned} &|\mathcal{A}_h R_h^2(v, w) - \tilde{R}_h^2 \mathcal{A}(v, w)|^2_{H_h^{-1} \times H_h^{-1}} \\ &\quad \leq \nu \, |(-\Delta_h) R_h v - \tilde{R}_h (-\Delta) v|^2_{H_h^{-1}} + |(-\Delta_h) R_h w - \tilde{R}_h (-\Delta) w|^2_{H_h^{-1}} \\ &\qquad + |R_h v - \tilde{R}_h v|^2_{H_h^{-1}} + |R_h w - \tilde{R}_h w|^2_{H_h^{-1}} \\ &\qquad\qquad \leq C_K^2 \, h^2 \, |(v, w)|_{H^2(\Omega)^2}; \end{aligned}$$

see [174, p. 232]. $\square$

We have the following result [61].

Theorem 3.4. *There exists a constant K_1, depending on Ω, g, z, and independent of h, such that*

$$|y_h - R_h y|_1 + |u_h - R_h u|_1 + |p_h - R_h p|_1 \leq K_1 \, h.$$

In case of a general convex domain and finite differences, attention must be paid to the discretization of $-\Delta$ along the boundary. The literature offers several options. For the Shortley–Weller discretization [174], $-\Delta_h$ is H_h^1-regular and consistent with $-\Delta$ from $H^2(\Omega)$ to H_h^{-1}. Using these facts the generalization of Theorem 3.4 to convex domains is straightforward.

In the following result [61] the assumption that the boundaries of Ω coincide with grid lines is used.

Theorem 3.5. *There exists a constant K_2, depending on Ω, g, z, and independent of h, such that*

$$|y_h - R_h y|_0 + |u_h - R_h u|_0 + |p_h - R_h p|_0 \leq K_2 h^2.$$

Notice that this result remains valid in the case of a linear FEM discretization on regular triangulation as illustrated in Section 5.7.3; see [230, 339] for further details and references.

Next, we discuss higher-order approximations of the elliptic optimality system. As mentioned in [174], the five-point formula above is optimal in the sense that there is no compact nine-point formula which provides an order of accuracy higher than two. Indeed, Collatz's compact nine-point scheme provides fourth-order accuracy when used in combination with a five-point representation of the right-hand side. In the case of optimality systems the use of the compact nine-point scheme requires a five-point representation of the control $u \in C^2(\Omega)$ in the state equation and of y in the adjoint equation. However, in the cases where constraints on the control are active, the control u may result not sufficiently smooth to allow Collatz's approach, and in these cases the accuracy of the compact nine-point optimal control solution collapses to second order. For this reason, it is preferable to use the extended nine-point schema considered in [77, 299] which is suitable for less smooth controls. We use the following fourth-order nine-point approximation to the Laplacian

$$\Delta_h = \left(1 - \frac{h^2}{12}\partial_1^+\partial_1^-\right)\partial_1^+\partial_1^- + \left(1 - \frac{h^2}{12}\partial_2^+\partial_2^-\right)\partial_2^+\partial_2^-. \tag{3.11}$$

For more insight, the one-dimensional expanded form of this operator is given by

$$\left(1 - \frac{h^2}{12}\partial_1^+\partial_1^-\right)\partial_1^+\partial_1^- v(x) = \frac{1}{12h^2}(-v(x-2h) + 16\,v(x-h) - 30\,v(x) + 16\,v(x+h) - v(x+2h)).$$

In the following we denote with $\tilde{\Delta}_h$ the second-order five-point Laplacian.

At grid points with distance h from the boundary, the operator Δ_h must be modified. In [77] it is shown that the approximation of the Laplacian near the boundary needs to be only $\mathcal{O}(h^2)$ without destroying the overall fourth-order accuracy of the scheme. Thus the five-point Laplace operator could be used (in particular to implement boundary controls as in [58]). However, on coarse grids it turns out [48] that the following asymmetric fourth-order approximation of the second partial derivative is more accurate. For $\mathbf{x} \in \Omega_h$ next to the left-hand side boundary we have [299]

$$\frac{\partial^2 v}{\partial x_1^2} \approx \frac{1}{12h^2}(10\,v(x_1-h) - 15\,v(x_1) - 4\,v(x_1+h) + 14\,v(x_1+2h) - 6\,v(x_1+3h) + v(x_1+4h)). \tag{3.12}$$

The scheme (3.11)–(3.12) results in a matrix of coefficients which is neither diagonally dominant nor of nonnegative type [77]. Nevertheless, in the case that $\tilde{\Delta}_h$ is used close to the boundary, it is proved in [77] that the resulting problem satisfies a maximum

principle. The same proof applies with few modifications to the case when (3.12) is used instead. Further we have that

$$|\Delta_h R_h v - \tilde{R}_h \Delta v|_\infty \leq c\, h^4 \|v\|_{C^6(\bar{\Omega})}, \tag{3.13}$$

where c is independent of v and h.

Next, an a priori estimate of the accuracy of solutions to the optimality system (3.2) is illustrated. After discretization, we have the following discrete optimality system

$$-\Delta_h y_h - p_h/\nu = \tilde{g}_h, \tag{3.14}$$

$$-\Delta_h p_h + y_h = \tilde{z}_h, \tag{3.15}$$

where $\tilde{g}_h = \tilde{R}_h g$ and $\tilde{z}_h = \tilde{R}_h z$.

Now consider the inner product of (3.14) by $\nu\, y_h$ and of (3.15) by p_h and take the sum of the two resulting equations. We obtain

$$\nu(-\Delta_h y_h, y_h)_{L_h^2} + (-\Delta_h p_h, p_h)_{L_h^2} = \nu(\tilde{g}_h, y_h)_{L_h^2} + (\tilde{z}_h, p_h)_{L_h^2},$$

which implies that

$$\nu(-\Delta_h y_h, y_h)_{L_h^2} + (-\Delta_h p_h, p_h)_{L_h^2} \leq \nu\,|(\tilde{g}_h, y_h)_{L_h^2}| + |(\tilde{z}_h, p_h)_{L_h^2}|.$$

By construction of (3.11), we have that $(-\Delta_h v_h, v_h)_{L_h^2} \geq (-\tilde{\Delta}_h v_h, v_h)_{L_h^2}$ for all functions v_h. Because $(-\tilde{\Delta}_h v_h, v_h)_{L_h^2} = \sum_{i=1}^2 |\partial_i^- v_h|_0^2$ [327] and using Lemma 3.1, we obtain

$$\nu\,|y_h|_0^2 + |p_h|_0^2 \leq c_*\,\nu\,|(\tilde{g}_h, y_h)_{L_h^2}| + c_*\,|(\tilde{z}_h, p_h)_{L_h^2}|.$$

Applying the Cauchy–Schwarz and Cauchy inequalities on the right-hand side of this expression results in

$$\nu\,|y_h|_0^2 + |p_h|_0^2 \leq c\,(\nu\,|\tilde{g}_h|_0^2 + |\tilde{z}_h|_0^2), \tag{3.16}$$

where $c = c_*/(2 - c_*)$. We remark that the same inequality is obtained if we use the five-point Laplacian in place of Δ_h in (3.14) and/or (3.15).

Using (3.16), we are now able to determine the degree of accuracy of the optimal solution. For this purpose, notice that (3.14)–(3.15) hold true with y_h and p_h replaced by their respective error functions, and with $\tilde{g}_h$ and $\tilde{z}_h$ replaced by the truncation error for Δ_h estimated by (3.13) (resp., by (3.5)). Further notice that dividing (3.16) by ν and recalling that $u_h = p_h/\nu$, we obtain the estimate for the control from $|y_h|_0^2 + \nu\,|u_h|_0^2 \leq c\,(|\tilde{f}_h|_0^2 + |\tilde{z}_h|_0^2/\nu)$. These statements are summarized in the following theorem.

Theorem 3.6. *Let $y \in C^{k+2}(\bar{\Omega})$, $k = 2,4$, and $p \in C^{l+2}(\bar{\Omega})$, $l = 2,4$, be solutions to* (3.2) *and let y_h and p_h be solutions to* (3.14)–(3.15)*. Then there exists a constant c, depending on Ω, and independent of h, such that*

$$|y_h - R_h y|_0^2 + \frac{1}{\nu}\,|p_h - R_h p|_0^2 \leq c\left(h^{2k}\,\|y\|_{C^{k+2}(\bar{\Omega})}^2 + h^{2l}\,\frac{1}{\nu}\,\|p\|_{C^{l+2}(\bar{\Omega})}^2\right).$$

This estimate holds for optimality systems with a linear control mechanism. For results concerning a priori error estimates for elliptic optimal control problems with a bilinear control structure see [225] and the references given therein.

3.3 Discretization of Parabolic Optimization Problems

In this section, we discuss discretization issues concerning the following parabolic optimal control problem

$$\left\{\begin{array}{rcll} \min J(y,u) & := & \frac{1}{2}\|y-z\|^2_{L^2(Q)} + \frac{\nu}{2}\|u\|^2_{L^2(Q)}, & \\ -\partial_t y + \Delta y & = & u & \text{in } Q = \Omega \times (0,T), \\ y(\mathbf{x},0) & = & y_0(\mathbf{x}) & \text{in } \Omega \text{ at } t = 0, \\ y(\mathbf{x},t) & = & 0 & \text{on } \Sigma = \partial\Omega \times (0,T), \end{array}\right. \tag{3.17}$$

where we take $y_0(\mathbf{x}) \in H^1_0(\Omega)$. Here, $\nu > 0$ is the weight of the cost of the control and $z \in L^2(Q)$ denotes the desired state. Then there exists a unique solution to the optimal control problem above; see [235, 339]. Corresponding to our setting we have $y^*(u^*) \in H^{2,1}(Q)$, where $H^{2,1}(Q) = L^2(0,T;H^2(\Omega)\cap H^1_0(\Omega)) \cap H^1(0,T;L^2(\Omega))$.

The solution to (3.17) is characterized by the following optimality system

$$\begin{aligned} -\partial_t y + \Delta y &= u, \\ \partial_t p + \Delta p + (y - z) &= 0, \\ \nu u - p &= 0, \end{aligned} \tag{3.18}$$

with initial condition $y(\mathbf{x},0) = y_0(\mathbf{x})$ for the state equation (evolving forward in time) and terminal condition $p(\mathbf{x},T) = 0$ for the adjoint equation (evolving backward in time). From (3.18) we have $p, u \in H^{2,1}(Q)$.

In the following, the numerical solution of the optimality system (3.18) in the framework of finite differences and backward Euler schemes is considered. We call $\overline{\Omega}_h$ the mesh, Ω_h is the set of interior mesh-points, and Γ_h is the set of boundary mesh-points. We consider the negative Laplacian with homogeneous Dirichlet boundary conditions approximated by the five-point stencil.

Let $\delta t = T/N_t$ be the time step size. Define

$$Q_{h,\delta t} = \{(\mathbf{x}, t_m) : \mathbf{x} \in \Omega_h, t_m = (m-1)\delta t, 1 \le m \le N_t + 1\}.$$

On this grid, y_h^m and p_h^m denote grid functions at time level m. The action of the one-step backward and forward time-discretization operator on these functions is defined as follows

$$\partial^+ y_h^m := \frac{y_h^m - y_h^{m-1}}{\delta t} \quad \text{and} \quad \partial^- p_h^m := -\frac{p_h^m - p_h^{m+1}}{\delta t}.$$

For grid functions defined on $Q_{h,\delta t}$ we use the discrete $L^2(Q)$ scalar product with norm $\|v_{h,\delta t}\| = (v_{h,\delta t}, v_{h,\delta t})^{1/2}_{L^2_{h,\delta t}(Q_{h,\delta t})}$.

For convenience, it is assumed that there exist positive constants $c_1 \le c_2$ such that $c_1 h^2 \le \delta t \le c_2 h^2$. Hence h can be considered as the only discretization parameter. Therefore, in the following, the subscript δt is omitted.

On the cylinder Q_h define the family of functions piecewise constant on intervals $[t_m, t_{m+1})$ as follows

$$V_h = \{v_h \,|\, v_h(t) = v_h(t_m) \text{ for } t \in [t_m, t_{m+1}), v_h(t_m) \in L^2_h(\Omega_h)\}.$$

The space-time extension of the operators $\tilde{R}_h$ and R_h is denoted by

$$\tilde{R}_{h,Q} : L^2(Q) \to V_h \quad \text{and} \quad R_{h,Q} : H^{2,1}(Q) \to V_h.$$

Condition (3.3) implies

$$||\tilde{R}_{h,Q}\, v - R_{h,Q}\, v|| \le c\, h^2\, |v|_{H^{2,1}(Q)}. \tag{3.19}$$

With this preparation, we can formulate the discrete version of the parabolic optimality system (3.18). We have

$$\begin{aligned} -\partial_t^+ y_h + \Delta_h y_h &= u_h, \\ \partial_t^- p_h + \Delta_h p_h &= -(y_h - \tilde{R}_{h,Q} z), \\ \nu u_h - p_h &= 0. \end{aligned} \tag{3.20}$$

Next we eliminate u_h from this system. In expanded form, we obtain

$$\begin{aligned} -\big[1+4\gamma\big] y_{i\,j\,m} + \gamma\big[y_{i+1\,j\,m} + y_{i-1\,j\,m} + y_{i\,j+1\,m} + y_{i\,j-1\,m}\big] + y_{i\,j\,m-1} \\ = \frac{\delta t}{\nu}\, p_{i\,j\,m}, \quad 2 \le m \le N_t + 1, \end{aligned} \tag{3.21}$$

$$\begin{aligned} -\big[1+4\gamma\big] p_{i\,j\,m} + \gamma\big[p_{i+1\,j\,m} + p_{i-1\,j\,m} + p_{i\,j+1\,m} + p_{i\,j-1\,m}\big] + p_{i\,j\,m+1} \\ + \delta t\,(y_{i\,j\,m} - \tilde{z}_{i\,j\,m}) = 0, \quad 1 \le m \le N_t, \end{aligned} \tag{3.22}$$

where $\gamma = \frac{\delta t}{h^2}$, $2 \le i, j \le N_{\mathbf{x}}$ index the internal grid points and $\tilde{z} = \tilde{R}_{h,Q} z$. The implementation of the boundary conditions on Σ, of the initial condition at $t = 0$, and of the terminal condition at $t = T$ should be clear.

In [50], the theory of [245] is elaborated on to prove that the solution of (3.20) is second-order accurate. For this purpose, in [50] the approach of [245] is extended to the present finite difference framework. The following estimates are obtained

$$\|u_h - R_{h,Q}\, u\| \le c\, h^2 \tag{3.23}$$

and

$$\|y_h - R_{h,Q}\, y\| \le c\, h^2 \quad \text{and} \quad \|p_h - R_{h,Q}\, p\| \le c\, h^2. \tag{3.24}$$

In [50], results of numerical experiments are reported to validate this accuracy estimate.

Next, we discuss the case of second-order time discretization. For this purpose, we follow [150] and consider the second-order backward differentiation formula (BDF2) together with the Crank–Nicolson (CN) method in order to obtain a second-order time-discretization scheme. Notice that the techniques discussed here can be generalized to higher-order BDF schemes. We remark that CN schemes are strictly nondissipative but easily oscillatory in contrast to BDF schemes that introduce numerical dissipation and thus are more appropriate in a multigrid framework. Therefore, we use the CN scheme only as a second-order one-step method for the purpose of initialization. For a detailed discussion of the BDF and CN schemes, see [13, 121].

To illustrate the BDF2 approach, we use the framework in [170, 174, 337] and assume that Ω is a square domain. The action of the BDF2 time-difference operators on these functions is as follows

$$\partial^+_{BD} y_h^m := \frac{3y_h^m - 4y_h^{m-1} + y_h^{m-2}}{2\delta t} \quad \text{and} \quad \partial^-_{BD} p_h^m := -\frac{3p_h^m - 4p_h^{m+1} + p_h^{m+2}}{2\delta t}.$$

The coefficients in the last two expressions above are given by the classical BDF2 formula (see, e.g., [13]), while the minus sign in the second operator allows us to discretize the adjoint variable taking into account its backward evolution in time.

With this setting, the following discrete optimality system is obtained

$$\begin{aligned} -\partial^+_{BD} y_h^m + \sigma \Delta_h y_h^m &= f_h^m + u_h^m, \\ \partial^-_{BD} p_h^m + \sigma \Delta_h p_h^m + \alpha(y_h^m - y_{dh}^m) &= 0, \\ \nu u_h^m - p_h^m &= 0, \end{aligned} \tag{3.25}$$

where we assume sufficient regularity of the data, y_d, y_T, and f, such that these functions are properly approximated by their values at grid points.

Notice that the BDF2 scheme is a multistep method, and therefore we need to combine it with a second-order one-step scheme for initialization of the state and adjoint equations at their initial and terminal time steps, respectively. For this purpose, we use the CN method.

Therefore, at $t = \delta t$ ($m = 2$), the optimality system results in the following

$$\begin{aligned} -\partial^+ y_h^m &= \tfrac{1}{2}\Big[\left(-\sigma \Delta_h y_h^m + u_h^m + f_h^m\right) \\ &\qquad + \left(-\sigma \Delta_h y_h^{m-1} + u_h^{m-1} + f_h^{m-1}\right)\Big], \\ \partial^-_{BD} p_h^m + \sigma \Delta_h p_h^m + \alpha(y_h^m - y_{dh}^m) &= 0, \\ \nu u_h^m - p_h^m &= 0. \end{aligned} \tag{3.26}$$

On the other hand, at $t = T - \delta t$ ($m = N_t$), the optimality system results in the following

$$\begin{aligned} -\partial^+_{BD} y_h^m + \sigma \Delta_h y_h^m &= f_h^m + u_h^m, \\ \partial^- p_h^m &= \tfrac{1}{2}\Big[\left(-\sigma \Delta_h p_h^{m+1} - \alpha(y_h^{m+1} - y_{dh}^{m+1})\right) \\ &\qquad + \left(-\sigma \Delta_h p_h^m - \alpha(y_h^m - y_{dh}^m)\right)\Big], \\ \nu u_h^m - p_h^m &= 0. \end{aligned} \tag{3.27}$$

Now, using the theory in [245] and the BDF2 estimates theory in [121] one can prove that the scheme above guarantees a second-order accurate approximation [150]. We have

$$\max_{2 \le m \le N_{t+1}} |y(t_m) - y_h^m|^2 \le c\left(|y(t_0) - y_0|^2 + |y(\delta t) - y_h^1|^2 + O(\delta t^6)\right). \tag{3.28}$$

Similarly, for the adjoint variable we have

$$\max_{N_{t+1} \le m \le 2} |p(t_m) - p_h^m|^2 \le c\left(|p(T) - p_h^{Nt+1}|^2 + |p(t_{Nt}) - p_h^{Nt}|^2 + O(\delta t^6)\right). \tag{3.29}$$

The two estimates (3.28) and (3.29) allow us to state that if the initial and terminal conditions, the first-step initial approximations of the state, and the adjoint variables are second-order accurate, then the proposed approach guarantees an optimal-order accuracy.

3.4 Discretization of Optimization Problems with Integral Equations

In this section, we discuss optimization with a model given by integral equations. A motivating reference for this particular class of optimization problems is [338], where optimal control problems are considered with PDE models in the integral formulation. Here, we follow [6] to discuss an optimal control problem with a Fredholm integral equation of the second kind.

Consider a Fredholm integral equation of the second kind with a linear distributed control mechanism. With a tracking objective, the purpose of the control is to determine a control function such that the resulting state $y \in L^2(\Omega)$ tracks as closely as possible a desired target configuration $z \in L^2(\Omega)$, where Ω is the domain. The corresponding optimal control problem is formulated as the minimization of a cost functional J subject to the constraint given by the integral equation. We have

$$\min J(y,u) := \frac{1}{2}\|y-z\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u\|^2_{L^2(\Omega)}, \tag{3.30}$$

$$y = f(y) + u + g \qquad \text{in } \Omega. \tag{3.31}$$

Here $g \in L^2(\Omega)$ is given, $u \in L^2(\Omega)$ is the control function, and the optimization parameter $\nu > 0$ is the weight of the cost of the control. The term $f(y)$ represents the integral operator, and it is given by

$$f(y)(x) = \int_\Omega K(x,t)\, y(t)\, dt. \tag{3.32}$$

Regarding the governing model, we assume an integral equation of the second kind, where the kernel K satisfies the conditions of the Fredholm alternative theorem [224] such that existence and uniqueness of solution for a given u is guaranteed. In particular, we consider a symmetric integral operator $f(\cdot) = f^T(\cdot)$, where $f^T(y)(x) := \int_\Omega K(t,x)\, y(t)\, dt$, and we require that

$$\|K\|_{L^2(\Omega\times\Omega)} = \int\!\!\int_{\Omega\times\Omega} |K(x,t)|^2\, dx\, dt < 1. \tag{3.33}$$

This condition itself is sufficient to prove existence and uniqueness of solution to the integral equation [224], and it can be easily verified in application. We have [6] the following.

Theorem 3.7. *Let K be such that existence and uniqueness of solution for* (3.31) *are guaranteed and define the gradient $\nabla \hat{J}(u) := \nu u - p$, where p is the solution to the integral equation $p = f^T(p) - y + z$ in Ω. Then the control problem* (3.30)–(3.31) *has a unique solution in $L^2(\Omega)$ if and only if $\nabla \hat{J}(u) = 0$. Therefore, the optimal solution is characterized as the solution of the following first-order optimality system*

$$\left\{ \begin{array}{rcl} y - f(y) - u &=& g, \\ p - f^T(p) + y &=& z, \\ \nu u - p &=& 0. \end{array} \right. \tag{3.34}$$

Using the scalar equation $\nu u - p = 0$, we can replace $u = p/\nu$ in the state equation and obtain the following equivalent system:

$$\left\{ \begin{array}{rcl} y - f(y) - p/\nu & = & g, \\ p - f(p) + y & = & z. \end{array} \right. \tag{3.35}$$

Notice that system (3.35) corresponds to two coupled integral equations that can be recast as a unique integral equation system, as follows

$$\left(\begin{array}{cc} 1 & -1/\nu \\ 1 & 1 \end{array} \right) \left(\begin{array}{c} y \\ p \end{array} \right) = \left(\begin{array}{c} f(y)+g \\ f^T(p)+z \end{array} \right),$$

that is,

$$\left(\begin{array}{c} y \\ p \end{array} \right) = \frac{1}{1+\nu} \left(\begin{array}{cc} \nu & 1 \\ -\nu & \nu \end{array} \right) \left(\begin{array}{c} f(y)+g \\ f^T(p)+z \end{array} \right). \tag{3.36}$$

The advantage of this formulation is the possibility to prove existence and uniqueness of solutions to (3.35) using condition (3.33). Notice that the coefficient matrix is never singular for $\nu > 0$. Following [224], we have that, in general, the integral system

$$\phi_i(x) = \lambda \sum_{j=1}^{n} \int_{\Omega} \tilde{K}_{ij}(x,t)\phi_j(t)\,dt + g_i(x), \qquad i = 1,\ldots,n,$$

where $\Omega = (a,b)$, $g_i \in L^2(\Omega)$, and $\tilde{K} \in L^2(\Omega \times \Omega)$, has a unique solution $\phi_i \in L^2(\Omega)$ provided that $|\lambda| < 1/C$, where $C^2 = \sum_{i,j=1}^{n} \int\int_{\Omega\times\Omega} |\tilde{K}_{ij}(x,t)|^2\,dx\,dt$. In our case, we have

$$C^2 = \frac{1}{(1+\nu)^2} \int\int_{\Omega\times\Omega} (2\nu^2|K(x,t)|^2 + (1+\nu^2)|K(x,t)|^2)\,dx\,dt.$$

Therefore, applying (3.33) to the optimality system, we find that this system admits a unique solution provided that

$$\int\int_{\Omega\times\Omega} |K(x,t)|^2\,dx\,dt < \frac{(1+\nu)^2}{1+3\nu^2}, \tag{3.37}$$

which is less restrictive than (3.33) when ν is sufficiently small. This result shows that controlled solutions may exist under weaker conditions than those required for the uncontrolled problem. A similar result is obtained in [57], in the case of singular elliptic control problems.

Now, we discuss the discretization of the Fredholm optimality system by the Nyström method. In [6], the optimality system (3.35) is discretized on a finite difference grid using direct quadrature (DQ) with the Nyström method [14, 176]. Take $x \in \Omega = (-D, D)$ and set the grid points

$$\Omega_h := \{x_i = i\,h, i = -N, -N+1, \ldots, 0, 1, \ldots, N-1, N\},$$

where $h = D/N$. On this grid, the following semidiscrete version of (3.35) is considered

$$\left\{ \begin{array}{rcl} y_N(x) - f_N(y_N)(x) - p_N(x)/\nu & = & g(x), \\ p_N(x) - f_N(p_N)(x) + y_N(x) & = & z(x), \end{array} \right. \tag{3.38}$$

where $x \in \Omega$. Here $y_N(x), p_N(x)$ are approximations to the solutions $y(x), p(x)$, for $x \in \Omega$, and $f_N(y_N)$ is the approximation to the integral with a direct quadrature: $f(y)(x) \approx f_N(y_N)(x) = h\sum_{j=-N}^{N} v_j\, K(x,t_j)\, y_N(t_j)$, where the v_j are the weights of a DQ of order q.

Following the Nyström method, the full discretization of (3.38) is obtained setting $y_i = y_N(x_i)$, $p_i = p_N(x_i)$, having suppressed the evidence of N for an easier notation. Thus, one obtains the following discrete optimality system

$$\begin{aligned} y_i - h\sum_{j=-N}^{N} w_{ij}\, y_j - p_i/\nu &= g_i, \\ p_i - h\sum_{j=-N}^{N} w_{ij}\, p_j + y_i &= z_i, \end{aligned} \tag{3.39}$$

where $i = -N,\dots,N$ and $w_{ij} = v_j\, K(x_i,t_j)$, with v_j given by the quadrature rule. In the following, we denote $y_h = \{y_{-N},\dots,y_N\}$ and $p_h = \{p_{-N},\dots,p_N\}$. The solution of (3.39) gives the approximate solution (y_h, p_h) of (3.35) at the mesh-points. Assuming that $K(x,t)\,y(t)$ is q-time continuously differentiable in t, and uniformly differentiable in x [14, 176], and assuming g is continuous, the following solution error estimate is obtained:

$$\|y - y_h\|_\infty + \|p - p_h\|_\infty \le O(h^q).$$

In a semidiscrete setting, approximation formulas for $y_N(x)$ and $p_N(x)$, $x \in \Omega$, can be found from (3.38), i.e., by the Nyström interpolation formula

$$\begin{cases} y_N = \frac{1}{1+\nu}\big[z + f_N(p_N) + \nu(g + f_N(y_N))\big], \\ p_N = \frac{\nu}{1+\nu}\big[z + f_N(p_N) - (g + f_N(y_N))\big]. \end{cases} \tag{3.40}$$

Under the same regularity condition on the kernel as given above, uniqueness of the solution of (3.39) results from the uniqueness of solution to (3.40). For these functions, one obtains a q-order convergence $\|y - y_N\|_\infty + \|p - p_N\|_\infty \le O(h^q)$.