



**Universidade do Estado do Rio de Janeiro**  
Centro de Ciência e Tecnologia  
Faculdade de Engenharia


Marlon Michael López Flores

**Final-State Approximate Control for the Heat Equation**

Rio de Janeiro  
2018

Marlon Michael López Flores

**Final-State Approximate Control for the Heat Equation**



Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Engenharia Mecânica, da Universidade do Estado do Rio de Janeiro. Área de concentração: Fenômenos de Transporte.

Orientador: Prof. Dr. Rogério Martins Saldanha da Gama

Orientador: Prof. Dr. Gilberto Oliveira Corrêa

Rio de Janeiro

2018

CATALOGAÇÃO NA FONTE  
UERJ / REDE SIRIUS / BIBLIOTECA CTC/B

F634 Flores, Marlon Michael López.  
Final-state approximate control for the heat equation / Marlon  
Michael López Flores. – 2018.  
124f.

Orientadores: Rogério Martins Saldanha da Gama, Gilberto  
Oliveira Corrêa.

Tese - (Doutorado) – Universidade do Estado do Rio de  
Janeiro, Faculdade de Engenharia.

1. Engenharia mecânica - Teses. 2. Equações diferenciais  
lineares - Teses. 3. Controle de temperatura - Teses. 4. Galerkin,  
Métodos de - Teses. 5. Modelos matemáticos - Teses. I. Gama,  
Rogério Martins Saldanha da. II. Corrêa, Gilberto Oliveira. III.  
Universidade do Estado do Rio de Janeiro, Faculdade de  
Engenharia. IV. Título.

CDU 517.977

Bibliotecária: Júlia Vieira – CRB7/6022

Autorizo, apenas para fins acadêmicos e científicos, a reprodução total ou  
parcial desta tese, desde que citada a fonte.

---

Assinatura

---

Data

Marlon Michael López Flores

**Final-State Approximate Control for the Heat Equation**

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Engenharia Mecânica, da Universidade do Estado do Rio de Janeiro. Área de concentração: Fenômenos de Transporte.

Aprovado em: 03 julho de 2018.

Banca examinadora:

---

Prof. Dr. Rogério Martins Saldanha da Gama (Orientador)  
Faculdade de Engenharia–UERJ

---

Prof. Dr. Gilberto Oliveira Corrêa (Orientador)  
Laboratório Nacional de Computação Científica–LNCC

---

Prof. Dr. José Júlio Pedrosa Filho  
Instituto de Matemática e Estatística–UERJ

---

Profa. Dra. Maria Laura Martins-Costa  
Departamento de Engenharia Mecânica–UFF

---

Prof. Dr. Alexandre Loureiro Madureira  
Laboratório Nacional de Computação Científica–LNCC

---

Prof. Dr. Paulo César Marques Vieira  
Laboratório Nacional de Computação Científica–LNCC

Rio de Janeiro

2018

## DEDICATÓRIA

Dedico este trabalho aos meus avós, Margarita de Jesús Flores Figueroa (*in memoriam*) e Rodolfo López Durón (*in memoriam*). Por terem me mostrado o caminho e o valor da educação.

## AGRADECIMENTOS

Quero agradecer a DEUS, por ter me permitido completar um desafio a mais na vida e me proporcionar o conhecimento, a sabedoria e o apoio para continuar novos desafios que enfrentarei antes de voltar para ELE.

Aos meus orientadores, os professores Gilberto e Rogério, pelo apoio durante todo o processo para completar este trabalho e terem me introduzido em uma linha de pesquisa muito interessante, a qual continuarei na minha vida acadêmica.

À Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro–FAPERJ pelo apoio financeiro dado em forma de uma bolsa de estudo para o doutorado.

Ao Programa de Pós-Graduação em Engenharia Mecânica da UERJ, que me proporcionou a oportunidade de desenvolver este projeto.

Aos professores Alexandre Madureira, Paulo César Vieira e Marcelo Fragoso do Laboratório Nacional de Computação Científica–LNCC pelo apoio dado ao longo do desenvolvimento deste trabalho.

Ao Instituto Nacional de Matemática Pura e Aplicada–IMPA, por abrir suas instalações e me permitir acesso ao acervo da sua biblioteca. Ao professor Elon Lages Lima (*in memoriam*) por toda sua motivação. Aos funcionários da biblioteca, sempre dispostos para ajudar, especialmente a Carolina e Cecília por todo o apoio e amizade. Aos meus amigos da copiadora, com os quais posso sempre contar, Antônio Carlos, Miguel e Valdo.

Aos meus formadores e amigos da Universidade Nacional Autônoma de Honduras–UNAH, em particular aos meus orientadores, Concepción Ferrufino e Eduardo Bravo de las Casas. Às professoras Rosibel Pacheco, Silvia Alcerro e Raquel Angúlo e aos professores Francisco Figeac, Jorge Destephen, Adalid Gutiérrez, Oscar Montes, Salvador Llopis (*in memoriam*) e Gustavo Pérez por terem me fornecido as ferramentas intelectuais com as quais cheguei até aqui. À amiga, Rocio Somarriba, a prezada secretaria da escola de matemática da UNAH, pela sua amizade e colaboração em todo momento.

À todos os meus amigos que sempre tem me apoiado, sendo impossível mencioná-los todos aqui. Quero agradecer especialmente ao meu amigo e irmão Hugo Martín Montes de Oca Velásquez pela sua amizade incondicional. Aos amigos Marcos Mendes, Marly Damasceno, Victoria Damasceno e Laura Damasceno pelo apoio e alegre companhia. Também quero agradecer aos senhores Marlos Viana, Marvin Taylor Dormond, Eduardo Quadros Spinola, Honório Cambeche, Sergio Romana, Wanderson Costa, Cassio Jardim, Luiz Fernando Souza, José Roberto Schuabb, Daniel Blanquicett, Clarena Arrieta, Rodrigo Matos, Rafael Alves, Angel Cano, Guilherme Welter, Elizabeth Vásquez, Suélen Sgorla, Tiecheng Xu, Xu Yang e Xiao-Chuan Liu. MUITO OBRIGADO À TODOS!

Only those who have the patience to do simple things perfectly will acquire the skill to do difficult things easily.

*Johann Christoph Friedrich von Schiller. (1759-1805)*

## RESUMO

FLORES, Marlon Michael López. *Controle Aproximado do Estado Final para a Equação de Calor*. Brasil. 2018. 124 f. Tese (Doutorado em Engenharia Mecânica) Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

Neste trabalho, dois tipos de problemas de controle em malha aberta são abordados em conexão com a equação linear de calor em domínios retangulares com condições de contorno tipo Dirichlet na qual a função de controle (dependendo apenas do tempo) constitui um termo de fonte. Em ambos os casos, o objetivo principal é impor um estado prescrito (distribuição de temperatura) no instante final de um dado intervalo de tempo. Sinais de controle serão selecionados com base em dois problemas de otimização, um sem restrições e outro envolvendo restrições nas magnitudes máximas dos valores obtidos pelos sinais de controle no intervalo de tempo em questão. Ambos os problemas têm o mesmo custo-funcional quadrático. Aproximações para os sinais de controle ótimo são obtidos com base na aproximação de Galerkin, de dimensão finita, para a equação linear de calor. Como consequência, os sinais de controle ótimos resultantes podem ser calculados de forma eficaz. Resultados numéricos para as equações de calor lineares 1D e 2D são apresentados para ilustrar os resultados mencionados acima. Com base nos resultados obtidos para a equação de calor linear, um esquema de linearização heurística é introduzido para tratar problemas de controle de estado final para a equação de calor não-linear. Este esquema baseia-se numa linearização por partes das ODEs não-lineares de dimensão finita correspondentes às aproximações de Galerkin da equação de calor não-linear. Alguns resultados numéricos também são apresentados para ilustrar este esquema de linearização heurística para a equação de calor não-linear 1D.

**Palavras-chave:** Controle ótimo; equações diferenciais parciais; soluções aproximadas.



## ABSTRACT

FLORES, Marlon Michael López. *Final-State Approximate Control for the Heat Equation*. Brasil. 2018. 124 f. Tese (Doutorado em Engenharia Mecânica) Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

In this work, two types of open-loop control problems are addressed in connection with the linear heat equation in rectangular domains with Dirichlet type boundary conditions in which the control function (depending only on time) constitutes a source term. In both cases, the main objective is to impose a prescribed state (temperature distribution) at the final instant of a given time-interval. Control signals are to be selected on the basis of two optimization problems, one unconstrained and the other one involving constraints on the maximum magnitudes of the values taken by the control signals on the time-interval in question. Both problems have the same quadratic cost-functional. Approximations for the optimal control signals are obtained on the basis of finite-dimensional Galerkin approximation for the linear heat equation. As a consequence, the resulting optimal control signals can be effectively computed. Numerical results for the 1D and 2D linear heat equations are presented to illustrate the results mentioned above. On the basis of the results obtained for the linear heat equation, a heuristic linearization scheme is introduced to address final-state control problems for the non-linear heat equation. This scheme rests on a piecewise linearization of the finite-dimensional, non-linear ODEs corresponding to Galerkin approximations of the non-linear heat equation. Some numerical results are also presented to illustrate this heuristic linearization scheme for the 1D non-linear heat equation.

**Keywords:** Optimal control; partial differential equations; approximate solutions.

## LIST OF FIGURES

|             |   |    |
|-------------|---|----|
| Figure 1 –  | Example 1. $\theta_r$ : target final state. . . . .   | 68 |
| Figure 2 –  | Example 1. $\beta_S$ : control-to-state actuator. . . . .   | 68 |
| Figure 3 –  | Example 1. Control signals $\mathbf{u}_K$ (blue dashed), $\mathbf{u}_K^R$ (red solid) for $\rho_F = 2000$ . . . . . | 69 |
| Figure 4 –  | Example 1. Approximations to target final state for $\rho_F = 2000$ . . . . .                                       | 70 |
| Figure 5 –  | Example 1. Control signals $\mathbf{u}_K$ (blue dashed), $\mathbf{u}_K^R$ (red solid) for $\rho_F = 4000$ . . . . . | 70 |
| Figure 6 –  | Example 1. Approximations to target final state for $\rho_F = 4000$ . . . . .                                       | 71 |
| Figure 7 –  | Example 2. $\theta_r$ : target final state. . . . .   | 71 |
| Figure 8 –  | Example 2. $\beta_S$ : control-to-state actuator. . . . .   | 72 |
| Figure 9 –  | Example 2. Control signals $\mathbf{u}_K$ (blue dashed), $\mathbf{u}_K^R$ (red solid) for $\rho_F = 2000$ . . . . . | 72 |
| Figure 10 – | Example 2. Approximations to target final state for $\rho_F = 2000$ . . . . .                                       | 73 |
| Figure 11 – | Example 2. Control signals $\mathbf{u}_K$ (blue dashed), $\mathbf{u}_K^R$ (red solid) for $\rho_F = 4000$ . . . . . | 73 |
| Figure 12 – | Example 2. Approximations to target final state for $\rho_F = 4000$ . . . . .                                       | 74 |
| Figure 13 – | Example 2. Approximations to target final state for $\rho_F = 4000$ , $\ell_x = 3/10$ . . . . .                     | 74 |
| Figure 14 – | Example 2. Approximations to target final state for $\rho_F = 4000$ , $\ell_x = 1$ . . . . .                        | 75 |
| Figure 15 – | Example 2. Approximations to target final state for $\rho_F = 4000$ , $\ell_x = 2 - 3/10$ . . . . .                 | 76 |
| Figure 16 – | Transversal section of $\mathcal{T}_\theta^K[\mathbf{u}_c^K]$ at $\ell_x = \ell_y$ for $\rho_F = 8000$ . . . . .    | 79 |
| Figure 17 – | Transversal section of $\mathcal{T}_\theta^K[\mathbf{u}_c^K]$ at $\ell_x = \ell_y$ for $\rho_F = 20000$ . . . . .   | 79 |
| Figure 18 – | Transversal section of $\theta_{ro}^K$ . . . . .  | 80 |
| Figure 19 – | Graphs of $\mathbf{u}_K$ and $\mathbf{u}_c^K$ for $\rho_F = 8000$ . . . . .   | 80 |
| Figure 20 – | Graphs of $\mathbf{u}_K$ and $\mathbf{u}_c^K$ for $\rho_F = 20000$ . . . . .  | 81 |
| Figure 21 – | $\rho_F = 10000$ , $\mu_u = 30$ , $\alpha_1 = 0.1$ , $\alpha_2 = 0.9$ , $\gamma_2 = 1$ . . . . .                    | 89 |
| Figure 22 – | $\rho_F = 40000$ , $\mu_u = 30$ , $\alpha_1 = 0.1$ , $\alpha_2 = 0.9$ , $\gamma_2 = 1$ . . . . .                    | 89 |
| Figure 23 – | $\rho_F = 10000$ , $\mu_u = 70$ , $\alpha_1 = 0.1$ , $\alpha_2 = 0.9$ , $\gamma_2 = 1$ . . . . .                    | 90 |
| Figure 24 – | $\rho_F = 40000$ , $\mu_u = 70$ , $\alpha_1 = 0.1$ , $\alpha_2 = 0.9$ , $\gamma_2 = 1$ . . . . .                    | 90 |
| Figure 25 – | $\rho_F = 10000$ , $\mu_u = 30$ , $\alpha_1 = 0.1$ , $\alpha_2 = 0.9$ , $\gamma_2 = 10$ . . . . .                   | 91 |
| Figure 26 – | $\rho_F = 40000$ , $\mu_u = 30$ , $\alpha_1 = 0.1$ , $\alpha_2 = 0.9$ , $\gamma_2 = 10$ . . . . .                   | 91 |
| Figure 27 – | $\rho_F = 10000$ , $\mu_u = 70$ , $\alpha_1 = 0.1$ , $\alpha_2 = 0.9$ , $\gamma_2 = 10$ . . . . .                   | 92 |

|             |   |     |
|-------------|---|-----|
| Figure 28 – | $\rho_F = 40000, \mu_{\mathbf{u}} = 70, \alpha_1 = 0.1, \alpha_2 = 0.9, \gamma_2 = 10.$   | 92  |
| Figure 29 – | $\rho_F = 10000, \mu_{\mathbf{u}} = 30, \alpha_1 = 0.8, \alpha_2 = 0.2, \gamma_2 = 1.$  | 93  |
| Figure 30 – | $\rho_F = 40000, \mu_{\mathbf{u}} = 30, \alpha_1 = 0.8, \alpha_2 = 0.2, \gamma_2 = 1.$  | 93  |
| Figure 31 – | $\rho_F = 10000, \mu_{\mathbf{u}} = 70, \alpha_1 = 0.8, \alpha_2 = 0.2, \gamma_2 = 1.$  | 94  |
| Figure 32 – | $\rho_F = 40000, \mu_{\mathbf{u}} = 70, \alpha_1 = 0.8, \alpha_2 = 0.2, \gamma_2 = 1.$  | 94  |
| Figure 33 – | $\rho_F = 10000, \mu_{\mathbf{u}} = 30, \alpha_1 = 0.2, \alpha_2 = 0.8, \gamma_2 = 1.$  | 95  |
| Figure 34 – | $\rho_F = 40000, \mu_{\mathbf{u}} = 30, \alpha_1 = 0.2, \alpha_2 = 0.8, \gamma_2 = 1.$  | 95  |
| Figure 35 – | $\rho_F = 10000, \mu_{\mathbf{u}} = 70, \alpha_1 = 0.2, \alpha_2 = 0.8, \gamma_2 = 1.$  | 96  |
| Figure 36 – | $\rho_F = 40000, \mu_{\mathbf{u}} = 70, \alpha_1 = 0.2, \alpha_2 = 0.8, \gamma_2 = 1.$  | 96  |
| Figure 37 – | $\rho_F = 10000, \mu_{\mathbf{u}} = 30, \alpha_1 = 0.5, \alpha_2 = 0.5, \gamma_2 = 1.$  | 97  |
| Figure 38 – | $\rho_F = 40000, \mu_{\mathbf{u}} = 30, \alpha_1 = 0.5, \alpha_2 = 0.5, \gamma_2 = 1.$  | 97  |
| Figure 39 – | $\rho_F = 10000, \mu_{\mathbf{u}} = 70, \alpha_1 = 0.5, \alpha_2 = 0.5, \gamma_2 = 1.$  | 98  |
| Figure 40 – | $\rho_F = 40000, \mu_{\mathbf{u}} = 70, \alpha_1 = 0.5, \alpha_2 = 0.5, \gamma_2 = 1.$  | 98  |
| Figure 41 – | $\rho_F = 10000, \mu_{\mathbf{u}} = 30, \alpha_1 = 0.8, \alpha_2 = 0.2, \gamma_2 = 1.$  | 99  |
| Figure 42 – | $\rho_F = 40000, \mu_{\mathbf{u}} = 30, \alpha_1 = 0.8, \alpha_2 = 0.2, \gamma_2 = 1.$  | 99  |
| Figure 43 – | $\rho_F = 10000, \mu_{\mathbf{u}} = 70, \alpha_1 = 0.8, \alpha_2 = 0.2, \gamma_2 = 1.$  | 100 |
| Figure 44 – | $\rho_F = 40000, \mu_{\mathbf{u}} = 70, \alpha_1 = 0.8, \alpha_2 = 0.2, \gamma_2 = 1.$  | 100 |
| Figure 45 – | Block diagram of a generic open-loop control system.  | 117 |
| Figure 46 – | Open-loop system representing a process modeled by the heat equation where the control function, $\mathbf{u}$ , depends only of time with a disturbance signal, $f_S$ , entering the process which gives us a desired output signal $\theta_{\text{cont.}}$ | 118 |

## LIST OF TABLES

|            |   |    |
|------------|---|----|
| Table 1 –  | Unconstrained problem for the first pair $(\theta_r, \beta_S)$ , $\rho_F = 2000$ . . . .  | 65 |
| Table 2 –  | Constrained problem for the first pair $(\theta_r, \beta_S)$ , $\rho_F = 2000$ . . . . .  | 65 |
| Table 3 –  | Unconstrained problem for the first pair $(\theta_r, \beta_S)$ , $\rho_F = 4000$ . . . .  | 66 |
| Table 4 –  | Constrained problem for the first pair $(\theta_r, \beta_S)$ , $\rho_F = 4000$ . . . . .  | 66 |
| Table 5 –  | Unconstrained problem for the second pair $(\theta_r, \beta_S)$ , $\rho_F = 2000$ . . .   | 66 |
| Table 6 –  | Constrained problem for the second pair $(\theta_r, \beta_S)$ , $\rho_F = 2000$ . . . .   | 67 |
| Table 7 –  | Unconstrained problem for the second pair $(\theta_r, \beta_S)$ , $\rho_F = 4000$ . . .   | 67 |
| Table 8 –  | Constrained problem for the second pair $(\theta_r, \beta_S)$ , $\rho_F = 4000$ . . . .   | 67 |
| Table 9 –  | Unconstrained problem with $\rho_F = 8000$ . . . . .  | 76 |
| Table 10 – | Constrained problem with $\rho_F = 8000$ . . . . .  | 76 |
| Table 11 – | Unconstrained problem with $\rho_F = 20000$ . . . . .   | 77 |
| Table 12 – | Constrained problem with $\rho_F = 20000$ . . . . .   | 77 |
| Table 13 – | Final state for $\theta_{r1}$ . Approximation error norms on $S_K$ for Figures<br>21 – 24 ( $\alpha_1 = 0.1$ , $\alpha_2 = 0.9$ and $\gamma_2 = 1$ ). . . . .   | 89 |
| Table 14 – | Final state for $\theta_{r1}$ . Approximation error norms on $S_K$ for Figures<br>25 – 28 ( $\alpha_1 = 0.1$ , $\alpha_2 = 0.9$ and $\gamma_2 = 10$ ) . . . . . | 91 |
| Table 15 – | Final state for $\theta_{r1}$ . Approximation error norms on $S_K$ for Figures<br>29 – 32 ( $\alpha_1 = 0.8$ , $\alpha_2 = 0.2$ and $\gamma_2 = 1$ ). . . . .   | 93 |
| Table 16 – | Final state for $\theta_{r2}$ . Approximation error norms on $S_K$ for Figures<br>33 – 36 ( $\alpha_1 = 0.2$ , $\alpha_2 = 0.8$ and $\gamma_2 = 1$ ). . . . .   | 95 |
| Table 17 – | Final state for $\theta_{r2}$ . Approximation error norms on $S_K$ for Figures<br>37 – 40 ( $\alpha_1 = 0.5$ , $\alpha_2 = 0.5$ and $\gamma_2 = 1$ ). . . . .   | 97 |
| Table 18 – | Final state for $\theta_{r2}$ . Approximation error norms on $S_K$ for Figures<br>41 – 44 ( $\alpha_1 = 0.8$ , $\alpha_2 = 0.2$ and $\gamma_2 = 1$ ). . . . .   | 99 |

## LIST OF SYMBOLS

|                    |  |
|--------------------|--|
| $\alpha$           | Thermal diffusivity  |
| $e_K$              | Error of approximation   |
| $\phi$             | Test function  |
| $\mathcal{J}$      | Cost functional  |
| $\mathcal{J}_K$    | Approximate cost functional                                      |
| $\kappa$           | Thermal conductivity   |
| $L_2(\Omega)$      | Space of square integrable real functions defined over $\Omega$  |
| $L_\infty(\Omega)$ | Space of all essentially bounded functions defined over $\Omega$ |
| $n_K$              | Dimension of subspace $S_K$                                      |
| $\Omega$           | Bounded open set in $\mathbb{R}^d$                               |
| $S_K$              | Finite-dimensional subspace of $L_2(\Omega)$                     |
| $\theta$           | Temperature function   |
| $\theta_r$         | Objective function   |
| $\mathbf{u}$       | Control signal   |
| $\mathbf{u}_K$     | Approximate control signal                                       |

## LIST OF PROBLEMS

- Prob. I* Infinite-dimensional unconstrained optimization problem with a unique solution given by  $\mathbf{u} \in L_2(0, t_F)^m$
- Prob. I<sub>K</sub>* Finite-dimensional unconstrained optimization problem with a unique solution given by  $\mathbf{u}_K \in L_2(0, t_F)^m$
- Prob. II* Infinite-dimensional constrained optimization problem with a unique solution given by  $\mathbf{u}_c \in L_2(0, t_F)^m$  where the values of  $\mathbf{u}_c$  are limited by a fixed interval
- Prob. II<sub>K</sub>* Finite-dimensional constrained optimization problem with a unique solution given by  $\mathbf{u}_c^K$  where the values of  $\mathbf{u}_c^K$  are limited in a fixed interval
- Prob. II<sub>D<sub>K</sub></sub>* Finite-dimensional constrained optimization dual problem with a unique solution given by  $\boldsymbol{\lambda} \in S_\lambda$  where the values of  $\boldsymbol{\lambda}$  are nonnegative
- Prob. D<sub>K $\gamma$</sub>*  Finite-dimensional constrained optimization auxiliary dual problem with a unique solution given by  $(\underline{\gamma}_a^o, \underline{\gamma}_b^o)$  where the values of  $\underline{\gamma}_a^o$  and  $\underline{\gamma}_b^o$  are nonnegative

## SUMMARY

|     |   |     |
|-----|---|-----|
|     | <b>INTRODUCTION</b> . . . . .   | 14  |
| 1   | <b>THE NON-LINEAR HEAT EQUATION AND GALERKIN APPROX-<br/>IMATIONS</b> . . . . .   | 18  |
| 2   | <b>FINITE-STATE APROXIMATE CONTROL FOR THE LINEAR HEAT<br/>EQUATION</b> . . . . .   | 23  |
| 2.1 | <b>Final State Positioning with Source Control</b> . . . . .  | 26  |
| 2.2 | <b>Approximate Solutions</b> . . . . .  | 29  |
| 2.3 | <b>APPENDIX – PROOFS FROM CHAPTER 2</b> . . . . .   | 37  |
| 3   | <b>PEAK-VALUE CONSTRAINTS ON CONTROL SIGNALS AND AC-<br/>TUATOR LOCATION</b> . . . . .  | 42  |
| 3.1 | <b>Peak-value Constraints on Control Signals</b> . . . . .  | 42  |
| 3.2 | <b>Actuator Location</b> . . . . .  | 49  |
| 3.3 | <b>Sample Size for Random Search</b> . . . . .  | 52  |
| 3.4 | <b>APPENDIX – PROOFS FROM CHAPTER 3</b> . . . . .   | 53  |
| 4   | <b>EXAMPLES AND NUMERICAL RESULTS FOR THE LHE<sub>q</sub></b> . . . . .   | 60  |
| 4.1 | <b>A One-Dimensional Example</b> . . . . .  | 60  |
| 4.2 | <b>Numerical Results for the One-dimensional Example</b> . . . . .  | 65  |
| 4.3 | <b>A Two-Dimensional Example</b> . . . . .  | 75  |
| 4.4 | <b>Actuator Location</b> . . . . .  | 80  |
| 5   | <b>FINITE-STATE, APPROXIMATE CONTROL OF THE NLHE<sub>q</sub>: A<br/>HEURISTIC SCHEME BASED ON LINEARIZATION</b> . . . . .         | 82  |
| 5.1 | <b>Numerical Examples</b> . . . . .   | 87  |
| 6   | <b>CONCLUDING REMARKS</b> . . . . .   | 100 |
|     | <b>REFERENCES</b> . . . . .   | 102 |
|     | <b>APPENDIX A – ELEMENTS OF CONTINUUM MECHANICS AND THE MATHE-<br/>MATICAL DESCRIPTION OF HEAT CONDUCTION IN SOLIDS</b> . . . . . | 108 |
|     | <b>APPENDIX B – MATERIAL ON SYSTEMS AND CONTROL THEORY</b> . . . . .  | 116 |
|     | <b>APPENDIX C – MATERIAL ON SEMIGROUP THEORY AND THE 4TH-ORDER<br/>RUNGE–KUTTA METHOD</b> . . . . .                               | 120 |

## INTRODUCTION

The design of equipment and software for temperature control is an important engineering endeavor for the solution of several technological problems. Amongst them, the following could be mentioned:

- (*i*) Ambient temperature control for comfort purposes or to ensure adequate operating conditions for sensitive equipment, cf. (SHAIKH et al, 2018; MAHESH, 2018).
- (*ii*) Heat dissipation or cooling of specific pieces of hardware, cf. (NDAO; PELES; JENSEN, 2009; AHMED, 2018).
- (*iii*) Thermal processing in the metal industry, cf. (LOTOV, 2005; BLECK et al., 2014).
- (*iv*) Temperature control to ensure satisfactory results of chemical processes, cf. (SLYADNEV et al., 2001; ALVAREZ-RAMIREZ; ALVAREZ, 2005).
- (*v*) Temperature control in the glass and ceramics industry, cf. (CLEVER; LANG, 2012; MOROZKIN; TKACHEV, 2016).
- (*vi*) Food safety and the pharmaceutical industry, cf. (MERCIER et al., 2017; KUMAR; JHA, 2017).
- (*vii*) Biological tissue management and conservation, cf. (HOFFMANN; BOTKIN; TUR-OVA, 2011, 2014).

It is worth noting that such a range of potential applications involves different levels of intended accuracy in the control objectives as well as distinct levels of difficulties in the way of obtaining “good-enough” mathematical models of the corresponding physical processes which could be used for the design of control schemes.

Modelling the process of energy transfer in (supposedly) continuous bodies is a long-standing topic in the scientific literature (at least in the last 200 years). Nevertheless, several questions have not been satisfactorily answered yet such as, for instance, the fact that heat is assumed in most models to propagate with infinite speed. Indeed, a realistic hyperbolic model that yields good accuracy is not available yet.

The major aim of the vast majority of works on heat transfer is to simulate the behavior of the temperature in the interior of a body on the basis of informations about inner heat sources and interactions with the (external) environment – the latter being described by boundary conditions. In other words, to characterize the temperature as a function of position and time is the most frequent aim of the works in this area.



In this work, a somewhat more complex and ambitious objective is pursued. Methods are sought to locally control the temperature of a body so that it is kept within prescribed limits. Actually, although more complex, this is perhaps a greater motivation in various application areas. In other words, the main interest here is not only to characterize the temperature distribution in a body, but also to devise “what to do” to ensure that the temperature in question is kept within admissible limits.

To illustrate the importance of this topic, the effect of temperature variation on semi-conductors (e.g., a computer chip) may be considered. Such devices, usually made of silicon, suffer internal damage when the temperature exceeds  $95^{\circ}C$ . It is therefore necessary to find some kind of control process to avoid such damage, especially when replacement is not possible (e.g., when such a device is located in an artificial satellite).

A similar problem arises in connection with thermal comfort problems in which temperature should be confined to a relatively narrow range (in this case, humidity should also be controlled).

There are other situations in which temperature control plays an even more vital role such as those involving medicines and foodstuffs. For example, to avoid contamination due to bacterial proliferation. The National Health Surveillance Agency of Brazil (ANVISA, as per its Portuguese acronym) guidelines require that foodstuffs should be subject to one of the following conditions:

- Kept heated above  $60^{\circ}C$  for up to 6 hours,
- Kept cooled below  $5^{\circ}C$  for up to 5 days,
- Kept frozen below  $-18^{\circ}C$  for unspecified periods.

These guidelines stem from the fact that between  $5^{\circ}C - 60^{\circ}C$  in humid environments (high risk zone), pathogenic bacteria reproduce very quickly. In high risk conditions, a simple bacterium may generate 130000 off-springs in only 6 hours. It is indeed very necessary to keep perishable foodstuffs under temperature control.

Similar considerations apply to medicines. However, in addition to the risks of contamination, effectiveness may be lost with exposure to inappropriate conditions. For example, HGH (Human Growth Hormone used in cases of growth deficit) ceases to be effective if it is kept for more than 20 minutes above  $8^{\circ}C$  (at current prices, each mL of HGH can cost up to US\$1000).

It is worth noting that temperature control should be considered as a “local” issue, *i.e.*, not only average temperatures should be kept in a prescribed range but the temperatures at all points of a body should behave likewise.

In this work, models will not take into account convective transport (it would be required by more realistic models for situations involving medicines and foodstuffs). Non opaque bodies (those that allow thermal radiation through) will not be considered either. Being a preliminary study only models for rigid and opaque bodies (at rest) will be handled. Nevertheless, problems will be tackled here for models in which thermal conductivity (and hence, thermal diffusiveness) depends on the temperature – in cases such as the ones involving silicon, this dependency should not be neglected.

Summing up, a simple (idealized) model of heat propagation/conduction (in solids) will be considered here as the basic set-up on the basis of which certain temperature control problems will be studied. The major aim here is to explore computational methods for obtaining approximate solutions to the corresponding mathematical problems.

Accordingly, the so-called heat equation is taken to be the fundamental mathematical model linking heat sources to temperature evolution for which open-loop control problems will be formulated and approximate solutions will be sought. More specifically, for linear (or linearized) versions of the heat equation optimal (open-loop) control problems will be considered here.

Such optimal control problems have been extensively studied – see for example (KOGUT; LEUGERING, 2011; TRÖLTZSCH, 2010; ZUAZUA, 2002) and references therein. Generally speaking, this work has emphasized the possibility of establishing “abstract” optimality conditions (rather than computational schemes to obtain control signals) in very general set-ups – for example, general spatial domains with smooth boundaries, control functions which vary in time and space and control objectives involving approximating a desired state-variable trajectory over the whole of a time-interval. Quite often, results are obtained on the basis of advanced general methods such as the so-called *Hilbert Uniqueness Method* introduced by (LIONS, 1988a,b).

In contrast, the main objective here is to exploit a simpler set-up involving approximation of the final-state of a given time-interval (as the main control objective) and control functions which depend only on time (“point controls”) and whose spatial action is defined by properties of “actuators” reflected in the evolution equations at stake. Restricting attention to such simpler set-ups, the major aim here is to obtain, by elementary means, characterizations of approximate optimal control signals which only involve relatively simple computational tasks and could, therefore, be effectively generated.

The DSc. thesis is organized as follows: Chapter 1 presents basic material about finite-dimensional Galerkin approximations for the one-dimensional heat equation in which the diffusion coefficient depends on the temperature (henceforth, referred to as the “non-

linear heat equation”). In Chapter 2 and 3 attention is restricted to the case in which the diffusion coefficient is taken to be constant (leading to the “linear heat equation”). In this simplified set-up, multi-dimensional spatial domains are considered and quadratic optimal control problems involving multivariable control signals are formulated with (Chapter 3) and without (Chapter 2) constraints on the maximum modulus allowed for the values taken by each scalar control signal. Chapter 3 ends with a brief section on the question of how to choose the location (in the spatial domain) of the “point-controls”. In Chapter 4, numerical examples are presented in which the methods presented in Chapter 2 and 3 for obtaining approximate solutions to constrained and unconstrained optimal control problems are applied to the 1D and 2D linear heat equation. In Chapter 5, a linearization scheme is introduced to obtain control signals for finite-dimensional Galerkin approximations to the non-linear heat equation with the same final objective, namely, to reach approximately a desired final state over a prescribed time-interval – this scheme essentially amounts to using the methods of Chapter 2 and 3 in a “piecewise” basis. Numerical examples are then presented to illustrate the linearization scheme at hand. Chapter 6 presents some concluding remarks and ideas for future works related to the results presented along this thesis.

Finally, with a view to facilitate reading, basic material on continuum mechanics, some topics on systems and control problems, semigroups and some information on numerical methods used in this work is presented in the APPENDIX.

# 1 THE NON-LINEAR HEAT EQUATION AND GALERKIN APPROXIMATIONS

The heat transfer process in a rigid and opaque body is one of the most studied phenomena in Mechanics. Usually, the main objective is to determine the temperature distribution based on a set of given characteristics. This phenomenon, called conduction heat transfer, for a body represented by a bounded open set  $\Omega$ , is mathematically described by the following partial differential equation

$$\rho c \frac{\partial \theta}{\partial t} = \operatorname{div}(\kappa(\theta) \operatorname{grad} \theta) + \dot{q}, \quad \text{in } \Omega \times (0, t_F] \quad (1.1)$$

subject to boundary conditions and to initial data.

In equation (1.1)  $\theta$  represents the temperature (the unknown),  $\rho$  is the mass density,  $c$  is the specific heat,  $\kappa$  is the internal conductivity and  $\dot{q}$  is the thermal heat supply (per unit time and unit volume). All the above parameters may depend on the temperature  $\theta$ . Nevertheless, for real bodies, the most important dependence involves the thermal conductivity  $\kappa$  and the local temperature  $\theta$ .

In this work, we shall restrict our attention to one-dimensional phenomena for this type of equation. In such cases, the above equation reduces to

$$\rho c \frac{\partial \theta}{\partial t} = \frac{\partial}{\partial x} \left[ \kappa(\theta) \frac{\partial \theta}{\partial x} \right] + \dot{q}.$$

In addition, we will assume that the product  $\rho c$  is a constant and rewrite the equation as follows

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial x} \left[ \alpha(\theta) \frac{\partial \theta}{\partial x} \right] + f, \quad \text{with } \alpha(\theta) = \frac{\kappa(\theta)}{\rho c}$$

in which  $\alpha$  represents the thermal diffusivity and  $f$  is a modified heat supply, defined by  $f = \frac{\dot{q}}{\rho c}$  – the resulting PDE is henceforth referred to as the non-linear heat equation (NLHEq).

In this work, the main objective goes beyond the determination of the evolution of the temperature distribution determined by equation (1.1). Starting from a given temperature distribution at the initial time, we look for an adequate source  $\dot{q}$  which ensures that the solution of equation (1.1) approximates a desired one in a prescribed time-horizon. More specifically, for  $\dot{q}$  being an affine function of control signal (time functions whose values can be imposed), the problem of choosing each control signals will be addressed with the aim of approximating a desired, pre-specified temperature distribution at the final instant of a given time-interval. This will be done on the basis

of Galerkin approximation to an initial/boundary condition problem involving equation (1.1).

To this effect, basic results leading to the construction of Galerkin approximations for the one-dimensional (1D) NLHEq with a prescribed initial condition and homogeneous Dirichlet boundary conditions are reviewed in the sequel.

On the basis of these results and the content of the next chapter on final-state, approximate control of the linear heat equation (LHEq), a linearization scheme will be introduced in Chapter 5 for tackling similar control problems for the NLHEq under the conditions mentioned above.

The initial-value/boundary condition problem for the so-called NLHEq is given in classical form as follows:

For  $L_x \in \mathbb{R}_+$ ,  $t_F \in \mathbb{R}_+$ ,  $\alpha : \mathbb{R} \rightarrow \mathbb{R}_+$ ,  $f : (0, L_x) \times (0, t_F) \rightarrow \mathbb{R}$  and  $g : (0, L_x) \rightarrow \mathbb{R}$ , a function  $\theta : [0, L_x] \times [0, t_F] \rightarrow \mathbb{R}$  is sought such that

$$\frac{\partial \theta}{\partial t}(x, t) = \frac{\partial}{\partial x} \left[ \alpha(\theta) \frac{\partial \theta}{\partial x} \right] + f(x, t) \quad \forall x \in (0, L_x), \forall t \in (0, t_F), \quad (1.2)$$

$$\theta(0, t) = \theta(L_x, t) = 0 \quad \forall t \in (0, t_F] \quad (\text{Boundary Conditions}), \quad (1.3)$$

$$\theta(x, 0) = g(x) \quad \forall x \in (0, L_x) \quad (\text{Initial Condition}). \quad (1.4)$$

To avoid excessively restrictive hypotheses on the pair  $(f, g)$  in the process of guaranteeing the existence and uniqueness of solutions for equation (1.2) – (1.4), a version of (1.2) – is considered which involves partial derivatives in the weak sense (EVANS, 2010, pp. 371 – 380).

This leads to the weak version of this problem which is now stated. To this effect, take the space of square integrable functions on a domain  $\Omega$ . In this case take  $\Omega = (0, L_x)$ , *i.e.*,

$$L_2(0, L_x) = \{f : (0, L_x) \rightarrow \mathbb{R} : f \text{ is measurable and } \int_0^{L_x} |f|^2 < \infty\},$$

then consider the Sobolev spaces (for further details see (BREZIS, 2011, pp. 202 – 219))

$$H^1(0, L_x) = \left\{ u \in L_2(0, L_x) : \exists v \in L_2(0, L_x) \text{ such that } \forall \phi \in C^1(0, L_x) \right. \\ \left. \text{with } \phi(0) = \phi(L_x) = 0 \text{ and } \int_0^{L_x} u \phi' = - \int_0^{L_x} v \phi \right\}$$

and

$$H_0^1(0, L_x) = \{h \in H^1(0, L_x) \text{ such that } h(0) = h(L_x) = 0\}$$

– note that if  $\phi \in H^1(0, L_x)$ , then  $\phi$  is continuous. On the basis of test functions  $\phi$  in  $H_0^1(0, L_x)$  (which satisfy the boundary conditions) the following problem is posed:

Find a function  $\underline{\theta} : [0, t_F] \rightarrow H_0^1(0, L_x)$  differentiable such that

$$\forall \phi \in H_0^1(0, L_x), \quad \left\langle \frac{d\underline{\theta}}{dt}(t), \phi \right\rangle = \left\langle \frac{\partial}{\partial x} \left[ \alpha(\underline{\theta}(t)) \frac{\partial \underline{\theta}(t)}{\partial x} \right], \phi \right\rangle + \langle f(\cdot, t), \phi \rangle \quad \text{a.e. in } (0, t_F),$$

Using *Green's first identity* (see APPENDIX A.2) then we obtain,

$$\forall \phi \in H_0^1(0, L_x), \quad \left\langle \frac{d\underline{\theta}}{dt}(t), \phi \right\rangle = - \left\langle \alpha(\underline{\theta}(t)) \frac{\partial \underline{\theta}(t)}{\partial x}, \frac{\partial \phi}{\partial x} \right\rangle + \langle f(\cdot, t), \phi \rangle \quad \text{a.e. in } (0, t_F) \quad (1.5)$$

$$\text{and} \quad \langle \underline{\theta}(0), \phi \rangle = \langle g, \phi \rangle, \quad (1.6)$$

where  $\langle \cdot, \cdot \rangle$  is defined by  $\langle u, v \rangle = \int_0^{L_x} u(x)v(x)dx$ .

The existence and uniqueness of solutions  $\underline{\theta} \in L_2(0, t_F; H_0^1(0, L_x))$  for the problem defined by (1.2) – (1.4) has been ascertained (as pointed out by (BERGAM;BERNARD;MGHAZLI, 2004)) by (LIONS, 1961, pp. 113 – 116) for any  $\underline{f} \in L_2(0, t_F; L_2(0, L_x))$ ,  $g \in L_2(0, L_x)$  and  $\alpha(\cdot)$  a continuously differentiable function from  $\mathbb{R}$  to  $\mathbb{R}$  satisfying

$$\forall \xi \in \mathbb{R}, \quad \alpha_{\min} \leq \alpha(\xi) \leq \alpha_{\max} \quad \text{and} \quad \left| \frac{d\alpha(\xi)}{d\xi} \right| \leq \alpha_{d\max}$$

for some positive constants  $\alpha_{\min}$ ,  $\alpha_{\max}$  and  $\alpha_{d\max}$ .

To compute approximate solutions to the problem defined by (1.5) – (1.6) the so-called Galerkin approximations have considered by (THOMÉE, 2006, and references therein), (DOUGLAS; DUPONT, 1970) and (WHEELER, 1973).

To define Galerkin approximations of this problem, let  $\{S_K\}, K = 1, 2, \dots$  be a family of finite-dimension subspaces  $S_K \subset H_0^1(0, L_x)$  with the approximation property, *i.e.*, such that

$$\forall f \in H_0^1(0, L_x), \quad \lim_{K \rightarrow \infty} \|f - \hat{f}_K\|_{L_2(0, L_x)} = 0,$$

where  $\hat{f}_K$  denotes the  $L_2$ -orthogonal projection of  $f$  on  $S_K$ .

An approximate problem is then formulated as follows:

Find  $\underline{\theta}_K : [0, t_F] \rightarrow S_K$  differentiable such that  $\forall t$  a.e. in  $[0, t_F]$

$$\forall \phi \in S_K, \quad \left\langle \frac{d\underline{\theta}_K}{dt}(t), \phi \right\rangle = - \left\langle \alpha(\underline{\theta}_K(t)) \frac{\partial \underline{\theta}_K(t)}{\partial x}, \frac{\partial \phi}{\partial x} \right\rangle + \langle f(\cdot, t), \phi \rangle \quad \text{a.e. in } (0, t_F) \quad (1.7)$$

$$\text{and} \quad \langle \underline{\theta}_K(0), \phi \rangle = \langle g, \phi \rangle. \quad (1.8)$$

Taking a basis  $\{\phi_1, \dots, \phi_{n_K}\}$  for  $S_K$ , where  $n_K = \dim(S_K)$ ,  $\underline{\theta}_K$  can be written as  $\underline{\theta}_K(t) = \sum_{k=1}^{n_K} c_k(t) \phi_k$  so that (1.7) can be written as

$$\forall \ell = 1, 2, \dots, n_K, \quad \left\langle \sum_{k=1}^{n_K} \dot{c}_k(t) \phi_k, \phi_\ell \right\rangle = - \left\langle \alpha_K(\underline{\mathbf{c}}_K(t)) \sum_{k=1}^{n_K} c_k(t) \frac{\partial \phi_k}{\partial x}, \frac{\partial \phi_\ell}{\partial x} \right\rangle + \langle f(\cdot, t), \phi_\ell \rangle,$$

where  $\dot{c}_k(t) = \frac{dc_k(t)}{dt}$ ,  $\underline{\mathbf{c}}_K(t) \triangleq [c_1(t) \dots c_{n_K}(t)]^T$  and  $\alpha_K(\underline{\mathbf{c}}_K(t)) \triangleq \alpha(\underline{\theta}_K(t))$ . Equivalently,

$$\forall \ell = 1, 2, \dots, n_K, \quad \sum_{k=1}^{n_K} \dot{c}_k(t) \langle \phi_k, \phi_\ell \rangle = - \sum_{k=1}^{n_K} c_k(t) \left\langle \alpha_K(\underline{\mathbf{c}}_K(t)) \frac{\partial \phi_k}{\partial x}, \frac{\partial \phi_\ell}{\partial x} \right\rangle + \langle f(\cdot, t), \phi_\ell \rangle.$$

Thus  $\forall \ell = 1, \dots, n_K$ ,

$$[\langle \phi_1, \phi_\ell \rangle \dots \langle \phi_{n_K}, \phi_\ell \rangle] \dot{\underline{\mathbf{c}}}_K(t) =$$

$$\left[ - \left\langle \alpha_K(\underline{\mathbf{c}}_K(t)) \frac{\partial \phi_1}{\partial x}, \frac{\partial \phi_\ell}{\partial x} \right\rangle \dots - \left\langle \alpha_K(\underline{\mathbf{c}}_K(t)) \frac{\partial \phi_{n_K}}{\partial x}, \frac{\partial \phi_\ell}{\partial x} \right\rangle \right] \underline{\mathbf{c}}_K(t) + \langle f(\cdot, t), \phi_\ell \rangle,$$

*i.e.*,

$$\mathbf{G}_K^\phi \dot{\underline{\mathbf{c}}}_K(t) = \check{\mathbf{R}}_K(\underline{\mathbf{c}}_K(t)) \underline{\mathbf{c}}_K(t) + \check{\mathbf{f}}_K(t),$$

where  $\{\mathbf{G}_K^\phi\}_{\ell k} = \langle \phi_k, \phi_\ell \rangle$ ,  $\{\check{\mathbf{R}}_K(\underline{\mathbf{c}}_K(t))\}_{\ell k} = - \langle \alpha_K(\underline{\mathbf{c}}_K(t)) \frac{\partial \phi_k}{\partial x}, \frac{\partial \phi_\ell}{\partial x} \rangle$  and  $\{\check{\mathbf{f}}_K\}_\ell = \langle f(\cdot, t), \phi_\ell \rangle$ . This can be rewritten as

$$\dot{\underline{\mathbf{c}}}_K(t) = \mathbf{R}_K(\underline{\mathbf{c}}_K(t)) \underline{\mathbf{c}}_K(t) + \mathbf{f}_K(t), \quad (1.9)$$

where  $\mathbf{R}_K(\cdot) = (\mathbf{G}_K^\phi)^{-1} \check{\mathbf{R}}(\cdot)$  and  $\mathbf{f}_K(t) = (\mathbf{G}_K^\phi)^{-1} \check{\mathbf{f}}_K(t)$ .

In turn, (1.8) can be written as

$$\forall \ell = 1, 2, \dots, n_K, \quad \langle \underline{\theta}_K(0), \phi_\ell \rangle = \langle g, \phi_\ell \rangle \Leftrightarrow \sum_{k=1}^{n_K} c_k(0) \langle \phi_k, \phi_\ell \rangle = \langle g, \phi_\ell \rangle.$$

Thus

$$\mathbf{G}_K^\phi \underline{\mathbf{c}}_K(0) = \check{\underline{\mathbf{g}}}_K \Rightarrow \underline{\mathbf{c}}_K(0) = \underline{\mathbf{g}}_K, \quad (1.10)$$

where  $\check{\underline{\mathbf{g}}}_K \triangleq [\langle g, \phi_1 \rangle \cdots \langle g, \phi_{n_K} \rangle]^\top$  and  $\underline{\mathbf{g}}_K \triangleq (\mathbf{G}_K^\phi)^{-1} \check{\underline{\mathbf{g}}}_K$ .

The  $K$ th-order approximate problem is then posed as follows:

Find  $\underline{\mathbf{c}}_K : [0, t_F] \rightarrow \mathbb{R}^{n_K}$  such that

$$\dot{\underline{\mathbf{c}}}_K(t) = \mathbf{R}_K(\underline{\mathbf{c}}_K(t)) \underline{\mathbf{c}}_K(t) + \mathbf{f}_K(t) \quad \text{and} \quad \underline{\mathbf{c}}_K(0) = \underline{\mathbf{g}}_K. \quad (1.11)$$

Once a solution to the problem defined by (1.11) is obtained, the corresponding  $K$ th-order Galerkin, candidate approximation to a solution of (1.5) – (1.6) is given by  $\underline{\theta}_K(t) = \sum_{k=1}^{n_K} c_k(t) \phi_k$ . The first step in this process is, therefore, to establish the existence of a unique solution to (1.11). To this effect, let  $h : \mathbb{R}^{n_K} \rightarrow \mathbb{R}^{n_K}$  be defined as  $h(\mathbf{z}) = \mathbf{R}_K(\mathbf{z})\mathbf{z}$ . As  $\mathbf{R}_K = (\mathbf{G}_K^\phi)^{-1} \check{\mathbf{R}}_K(\mathbf{z})$  and  $\{\check{\mathbf{R}}_K(\mathbf{z})\}_{\ell k} = -\langle \alpha(\sum_{i=1}^{n_K} z_i \phi_i) \frac{\partial \phi_k}{\partial x}, \frac{\partial \phi_\ell}{\partial x} \rangle$  if  $\alpha : \mathbb{R} \rightarrow \mathbb{R}$  is continuously differentiable, then so is  $\check{\mathbf{R}}_K(\cdot)$ ,  $\mathbf{R}_K(\cdot)$  and  $h$ . This together with the fact that  $\mathbf{f}_K : [0, t_F] \rightarrow \mathbb{R}^{n_K}$  is continuous leads to the existence and uniqueness of solution of (1.11), see (KELLEY; PETERSON, Theorem 8.13 – The Picard-Lindelöf theorem, p. 350).

The approximation error  $e_K \triangleq \underline{\theta}_K(t) - \underline{\theta}_o(t)$  is analyzed in (THOMÉE, 2006) for a specific family of finite-element subspaces  $S_K$ . Under the assumption on  $f$ ,  $g$  and  $\alpha(\cdot)$  mentioned above and (implicit) regularity assumptions on the unique solution  $\underline{\theta}_o(\cdot)$  for the problem defined by (1.5) – (1.6), upper bounds are derived in  $L_2(0, L_x)$ -norm of  $e_K(t)$ , see (THOMÉE, 2006, Theorem 13.1, p. 235). As those upper bounds go to zero as  $K \rightarrow \infty$  (making  $h = 1/K$  in the theorem mentioned, where  $h$  is the finite-element mesh size), it is then established that  $\underline{\theta}_K(t) \rightarrow \underline{\theta}_o$  (uniformly in  $t \in (0, t_F)$  in the  $L_2(0, L_x)$ -norm). A similar convergence result also follows from Theorem 3.1 in (DOUGLAS; DUPONT, 1970).



## 2 FINITE-STATE APPROXIMATE CONTROL FOR THE LINEAR HEAT EQUATION

In this chapter, the problem will be considered of choosing a control signal  $\mathbf{u} : [0, t_F] \rightarrow \mathbb{R}^m$ , where  $m \in \mathbb{N}$  and is defined by the number of scalar control signals chosen to approximately steer the solution of a LHEq (for a given initial value and homogeneous boundary Dirichlet condition) towards a prescribed final state. Given the relatively simple nature of the LHEq (vis-à-vis the NLHEq), a more general set-up will be explored, with  $m$  scalar control signals and spatial domains in  $\mathbb{R}^{m_x}$ . Accordingly, in Section 2.1, an optimal control problem is formulated for the LHEq and the optimal solution is characterized by a linear equation on  $L_2(0, t_F)^m$ . Then in Section 2.2, approximations to the optimal control signal are characterized (by means of linear equations in  $\mathbb{R}^n$ ) as optimal solutions to an optimal control problem posed on the basis of a Galerkin approximation (of a given dimension  $n$ ) for the LHEq. This chapter ends with a summary of the computational steps required to obtain the desired control signal.

Linear-quadratic optimal control problems have been extensively studied – see, for example (TRÖLTZSCH, 2010), (ZUAZUA, 2002) and references therein. Very often, general parabolic equations and more general cost-functional involving state values along the whole of  $[0, t_F]$  are considered. To cope with such general set-ups, results tend to concentrate on showing existence of optimal controls and establishing “abstract” optimality conditions (rather than computational schemes to compute control signals). This is often achieved invoking advanced general methods such as the so called *Hilbert Uniqueness Method* (HUM for short) devised by (LIONS, 1988a), (LIONS, 1988b).

In contrast, the main objective here is to exploit a simpler set-up (the LHEq and final-state control) to obtain, by elementary means, explicit characterizations of approximate optimal control signals which would only involve relatively simple computational tasks – with the end result that the desired “approximately-optimal” control signals could be effectively generated, see (CORRÊA; LÓPEZ-FLORES; MADUREIRA, 2012).

To this effect, consider a initial/boundary condition problem for the parabolic equation given (“in its classical form”) by

$$\frac{\partial \theta(\mathbf{x}, t)}{\partial t} = \alpha \sum_{i=1}^{m_x} \frac{\partial^2 \theta(\mathbf{x}, t)}{\partial x_i^2} + f(\mathbf{x}, t) \quad \forall \mathbf{x} \in \Omega, \forall t \in (0, \infty) \quad (2.1)$$

$$\theta(\mathbf{x}, t) = 0 \quad (\text{Boundary Conditions}) \quad \forall t \in (0, \infty), \forall \mathbf{x} \in \partial\Omega \quad (2.2)$$

$$\theta(\mathbf{x}, 0) = g(\mathbf{x}) \quad (\text{Initial Condition}) \quad \forall \mathbf{x} \in \Omega \quad (2.3)$$

where  $\Omega \in \mathbb{R}^{m_x}$  is a bounded, open and connected set,  $f$  and  $g$  are given functions and

$\alpha \in \mathbb{R}_+$ . The “weak” version of this problem is then formulated as follows:

Given  $\alpha \in \mathbb{R}_+$ ,  $\underline{f}(t) = f(\cdot, t) \in L_2(\Omega)$ ,  $g \in L_2(\Omega)$  and  $\Omega \in \mathbb{R}^{m_x}$  open and connected find  $\underline{\theta} : [0, t_F] \rightarrow H_0^1(\Omega)$  such that  $\forall \phi \in H_0^1(\Omega)$  and  $\forall t \in (0, t_F)$

$$\left\langle \frac{d\underline{\theta}}{dt}(t), \phi \right\rangle = -\alpha \sum_{i=1}^{m_x} \left\langle \frac{\partial \underline{\theta}(t)}{\partial x_i}, \frac{\partial \phi}{\partial x_i} \right\rangle + \langle \underline{f}(t), \phi \rangle, \quad (2.4)$$

$$\langle \underline{\theta}(0), \phi \rangle = \langle g, \phi \rangle. \quad (2.5)$$

The existence and uniqueness of solutions to this problem follows from the result of (EVANS, 2010, Theorem 7.1.1, p. 376).

Given that the main interest here is the final-state control problem, the semigroup representation of the solution to (2.1)–(2.3), see (CURTAIN; ZWART, 1995, Chapter 2, pp. 13–52), will be exploited. To bring in such a representation, let the operator  $A : H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$ , the dual space of  $H_0^1$ , be defined by

$$\forall \phi \in H_0^1(\Omega), \forall \psi \in H_0^1(\Omega), \quad \langle A[\phi], \psi \rangle = -\mathbf{B}[\phi, \psi], \quad (2.6)$$

where

$$\mathbf{B}[\phi, \psi] \triangleq \alpha \sum_{i=1}^{m_x} \left\langle \frac{\partial \phi}{\partial x_i}, \frac{\partial \psi}{\partial x_i} \right\rangle, \quad (2.7)$$

where  $\langle \cdot, \cdot \rangle$  denotes the inner product of  $L_2(\Omega)$ .

Then the problem above can be recast as the following Cauchy problem:

Find  $\underline{\theta} : [0, t_F] \rightarrow H_0^1(\Omega)$  such that

$$\dot{\underline{\theta}}(t) = A[\underline{\theta}(t)] + f(t) \quad , \quad t > 0 \quad , \quad \underline{\theta}(0) = g \quad (2.8)$$

where  $g \in L_2(\Omega)$ ,  $f : (0, \infty) \rightarrow L_2(\Omega)$  and  $\underline{\theta} : [0, \infty) \rightarrow L_2(\Omega)$ .

The operator  $A$  so defined is the infinitesimal generator of a  $C_o$ -semigroup  $S_A(t) : L_2(\Omega) \rightarrow L_2(\Omega)$ ,  $t \geq 0$  on the basis of which  $\underline{\theta}(\cdot)$  is given by

$$\underline{\theta}(t; f, g) = S_A(t)[g] + \int_0^t S_A(t - \tau)[f(\tau)]d\tau, \quad \forall t \in [0, t_F], \quad (2.9)$$

see (CURTAIN; ZWART, 1995, Chapter 3, pp. 101–107).

As the operator  $A$  is symmetric and elliptic, the eigenvectors of  $A$  constitute a complete orthonormal set for  $L_2(\Omega)$ , see (EVANS, 2010, Theorem 6.5.1, p. 355), so that  $\forall \phi \in H_0^1(\Omega)$ ,  $A[\phi] = \sum_{i=1}^{\infty} \lambda_i \langle \phi_i, \phi \rangle \phi_i$ , where  $\{\lambda_i\}_{i=1}^{\infty}$  are the corresponding eigenvalues of  $A$ , *i.e.*,  $A[\phi_i] = \lambda_i \phi_i$ . As a result,  $S_A(t)$  is given by

$$\forall \phi \in L_2(\Omega), \quad S_A(t)[\phi] = \sum_{i=1}^{\infty} e^{\lambda_i t} \langle \phi_i, \phi \rangle \phi_i. \quad (2.10)$$

**Remark 2.1.** *In the case of tensorial spatial domains, *i.e.*,  $\Omega = (0, L_{x_1}) \times \dots \times (0, L_{x_{m_x}})$  the eigenfunctions  $\phi_{\mathbf{k}}$ ,  $\mathbf{k} = (k_1, \dots, k_{m_x})$  of the operator  $A$  are given by*

$$\phi_{\mathbf{k}}(\mathbf{x}) = \prod_{i=1}^{m_x} \sqrt{\frac{2}{L_{x_i}}} \sin \left[ \frac{k_i \pi x_i}{L_{x_i}} \right].$$

▽

It is now assumed that  $f(\mathbf{x}, t) = f_S(\mathbf{x}, t) + \boldsymbol{\beta}_S(\mathbf{x})^T \mathbf{u}(t)$ , where  $f_S : \Omega \times [0, t_F] \rightarrow \mathbb{R}$  and  $\boldsymbol{\beta}_S : \Omega \rightarrow \mathbb{R}^m$  are given functions, where  $f_S$  would model “disturbances” (*i.e.*, control-independent heat sources) and  $\mathbf{u} : [0, t_F] \rightarrow \mathbb{R}^m$  is a control signal to be chosen in such a way as to make  $\underline{\theta}(t_F; f, g)$  “close” to a prescribed  $\theta_r \in L_2(\Omega)$ . This source term consists of a given “disturbance” term,  $f_S(\mathbf{x}, t)$ , and the controlling term with  $\boldsymbol{\beta}_S^T(\mathbf{x})$  which represents the spatial effects (and position), of the different controlled sources characterized by  $\boldsymbol{\beta}_{S_i}(\mathbf{x})$ . Thus, the control function can be a vector function  $\mathbf{u}(t) \in \mathbb{R}^m$ , *i.e.*,  $\mathbf{u}(t) = (u_1(t), \dots, u_m(t))$ , where each element of the vector represents a control signal for each of the individual sources given by  $\boldsymbol{\beta}_{S_i}^T(\mathbf{x})$ , *i.e.*,  $\boldsymbol{\beta}_{S_i}(\mathbf{x}) u_i(t)$ ,  $i = 1, \dots, m$ .

Now, let  $\mathbf{u} \in L_2(0, t_F)^m$ ,  $\rho_{\mathbf{u}} \in \mathbb{R}_+$  and define the cost functional

$$\mathcal{J}(\mathbf{u}) \triangleq \|\underline{\theta}(t_F; f, g) - \theta_r\|_{L_2(\Omega)}^2 + \rho_{\mathbf{u}} \|\mathbf{u}\|_{L_2(0, t_F)^m}^2 \quad (2.11)$$

(from now on, the “space” subindices of norms and inner products will be omitted whenever context information makes them redundant). Each of the terms of this functional provides us with very important information about the system represented by the heat equation and how it behaves under the influence of one or more control signals.

The “energy” that the control  $\mathbf{u}$  requires to take the system to the desired final state (objective) in a finite interval of time,  $(0, t_F)$ , can be studied by analyzing the behavior of  $\|\mathbf{u}\|_{L_2(0, t_F)^m}^2$ . This can be done by varying the parameter  $\rho_{\mathbf{u}}$  that penalizes this term. This has a direct influence over  $\underline{\theta}(t_F; f, g)$  when it approximates to the desired final

state,  $\theta_r$ . This information can be useful in the process of designing a real controller. The term  $\|\underline{\theta}(t_F; f, g) - \theta_r\|_{L_2(\Omega)}^2$  provides us with information of the proximity of the system's optimal final state under the effect of the control and the desired state (objective) which we want to approximate.

A control signal is to be chosen on the basis of the optimization problem

$$\underline{\text{Prob. I}}: \min_{\mathbf{u} \in L_2(0, t_F)^m} \mathcal{J}(\mathbf{u}). \quad (2.12)$$

Moreover, the cost functional,  $\mathcal{J}(\mathbf{u})$ , is a *convex, continuous and coercive functional* and with the fact that  $L_2(0, t_F)^m$  is closed and convex guarantees the existence of a function  $\mathbf{u}$  that minimizes it (EKELAND; TÉMAM, 1976, pp. 35–36).

## 2.1 Final State Positioning with Source Control

In this section, optimality conditions are presented for *Prob. I* on the basis of which its solution can be explicitly characterized. To this effect, note first that due to the linearity of  $\underline{\theta}(\cdot; f, g)$  on  $(f, g)$ ,

$$\underline{\theta}(\cdot; f, g) = \underline{\theta}(\cdot; f_S, g) + \underline{\theta}(\cdot; f_u, 0), \text{ where } f_u(t) = \beta_S^T(\cdot)\mathbf{u}(t), \quad (2.13)$$

*i.e.*,

$$\underline{\theta}(\cdot; f, g) = \underline{\theta}(\cdot; f_S, g) + \check{\mathcal{T}}_\theta[\mathbf{u}](\cdot), \quad (2.14)$$

where  $\check{\mathcal{T}}_\theta : L_2(0, t_F)^m \rightarrow \{\underline{h} : [0, t_F] \rightarrow H_0^1(\Omega)\}$

$$\check{\mathcal{T}}_\theta[\mathbf{u}](t) \triangleq \int_0^t S_A(t - \tau)[f_u(\tau)]d\tau. \quad (2.15)$$

From (2.10) we see that

$$\check{\mathcal{T}}_\theta[\mathbf{u}](t) = \int_0^t \sum_{i=1}^{\infty} e^{\lambda_i(t-\tau)} \langle \phi_i, f_u(\tau) \rangle \phi_i d\tau. \quad (2.16)$$

Now  $\mathcal{J}(\mathbf{u})$  can be rewritten as

$$\mathcal{J}(\mathbf{u}) = \|\mathcal{T}_\theta[\mathbf{u}] - \theta_{ro}\|_{L_2(\Omega)}^2 + \rho_u \|\mathbf{u}\|_{L_2(0, t_F)^m}^2, \quad (2.17)$$

where  $\theta_{ro} \triangleq \theta_r - \underline{\theta}(t_F; f, g)$  and  $\mathcal{T}_\theta : L_2(0, t_F)^m \rightarrow L_2(\Omega)$  is defined by  $\mathcal{T}_\theta[\mathbf{u}] = \check{\mathcal{T}}_\theta[\mathbf{u}](t_F)$ .

Exploiting the specific nature of the cost functional, the existence of an optimal solution to *Prob. I* can be ascertained by means of a basic result on minimum-distance problems pertaining to closed convex sets (LUENBERGER, 1969, p. 69), as stated in the next proposition in which the optimal solution is also characterized.

**Proposition 2.1.** *There exists  $\mathbf{u}_o \in L_2(0, t_F)^m$  such that  $\forall \mathbf{u} \in L_2(0, t_F)^m$ ,  $\mathbf{u} \neq \mathbf{u}_o$ ,  $\mathcal{J}(\mathbf{u}_o) < \mathcal{J}(\mathbf{u})$ .*

*Moreover,  $\mathbf{u}_o$  is the unique solution of the linear equation*

$$\rho_{\mathbf{u}} \mathbf{u}_o + \mathcal{T}_{\theta}^* \cdot \mathcal{T}_{\theta}[\mathbf{u}_o] - \mathcal{T}_{\theta}^*[\theta_{ro}] = 0, \quad (2.18)$$

i.e.,

$$\mathbf{u}_o = [\rho_{\mathbf{u}} I + \mathcal{T}_{\theta}^* \cdot \mathcal{T}_{\theta}]^{-1} [\mathcal{T}_{\theta}^*[\theta_{ro}]], \quad (2.19)$$

where  $\mathcal{T}_{\theta}^* : L_2(\Omega) \rightarrow L_2(0, t_F)^m$  is the adjoint of  $\mathcal{T}_{\theta}$ . ∇

*Proof.* Let  $\mathcal{T}_a : L_2(0, t_F)^m \rightarrow L_2(0, t_F)^m \times L_2(\Omega)$  be defined by  $\mathcal{T}_a[\mathbf{u}] \triangleq (\rho_{\mathbf{u}}^{1/2} \mathbf{u}, \mathcal{T}_{\theta}[\mathbf{u}])$ . Then  $\mathcal{J}(\mathbf{u}) = \|\mathcal{T}_a[\mathbf{u}] - (0, \theta_{ro})\|_{X_a}^2$ , where  $X_a \triangleq L_2(0, t_F)^m \times L_2(\Omega)$ , and *Prob. I* is seen as the problem of finding the minimum-distance approximation to  $(0, \theta_{ro}) \in X_a$  in  $\mathcal{T}_a[L_2(0, t_F)^m]$  - note that  $X_a$  is a Hilbert Space with the inner product

$$\langle (v_1, w_1), (v_2, w_2) \rangle_{X_a} = \langle v_1, v_2 \rangle_{L_2(0, t_F)^m} + \langle w_1, w_2 \rangle_{L_2(\Omega)}.$$

Moreover,  $\mathcal{T}_a[L_2(0, t_F)^m]$  is closed. Indeed, if  $\mathcal{T}_a[\mathbf{u}_K] \rightarrow \mathbf{x}_0 = (\hat{\mathbf{u}}_o, \hat{\theta}_{ao})$  or, equivalently,  $(\rho_{\mathbf{u}}^{1/2} \mathbf{u}_K, \mathcal{T}_{\theta}[\mathbf{u}_K]) \rightarrow (\hat{\mathbf{u}}_o, \hat{\theta}_{ao})$  then  $\mathbf{u}_K \rightarrow \rho_{\mathbf{u}}^{-1/2} \hat{\mathbf{u}}_o$  and (since  $\mathcal{T}_{\theta}$  is continuous)  $\mathcal{T}_{\theta}[\mathbf{u}_K] \rightarrow \mathcal{T}_{\theta}[\rho_{\mathbf{u}}^{-1/2} \hat{\mathbf{u}}_o] = \hat{\theta}_{ao}$ . Thus,  $\mathcal{T}_a(\rho_{\mathbf{u}}^{-1/2} \hat{\mathbf{u}}_o) = (\hat{\mathbf{u}}_o, \mathcal{T}_{\theta}[\rho_{\mathbf{u}}^{-1/2} \hat{\mathbf{u}}_o]) = (\hat{\mathbf{u}}_o, \hat{\theta}_{ao}) = \mathbf{x}_0 \Rightarrow \mathbf{x}_0 \in \mathcal{T}_a[L_2(0, t_F)^m]$ .

As  $\mathcal{T}_a[L_2(0, t_F)^m]$  is also convex, it follows from (LUENBERGER, 1969, Theorem 3.12.1, p. 69) that *Prob. I* has a unique solution  $\mathbf{u}_o$  (say).

Note now that  $\mathbf{u}_o$  is a solution to *Prob. I*  $\Leftrightarrow \forall \delta \mathbf{u} \in L_2(0, t_F)^m$ ,

$$\mathcal{J}(\mathbf{u}_o) \leq \mathcal{J}(\mathbf{u}_o + \delta \mathbf{u}) \Leftrightarrow \forall \delta \mathbf{u} \in L_2(0, t_F)^m,$$

$$2\rho_{\mathbf{u}} \langle \mathbf{u}_o, \delta \mathbf{u} \rangle_{L_2(0, t_F)^m} + \rho_{\mathbf{u}} \|\delta \mathbf{u}\|_{L_2(0, t_F)^m}^2 + 2\langle \mathcal{T}_{\theta}[\mathbf{u}_o] - \theta_{ro}, \mathcal{T}_{\theta}[\delta \mathbf{u}] \rangle + \|\mathcal{T}_{\theta}[\delta \mathbf{u}]\|_{L_2(\Omega)}^2 \geq 0$$

$$\Leftrightarrow \forall \delta \mathbf{u} \in L_2(0, t_F)^m, \quad \langle \rho_{\mathbf{u}} \mathbf{u}_o + \mathcal{T}_{\theta}^* \cdot \mathcal{T}_{\theta}[\mathbf{u}_o] - \mathcal{T}_{\theta}^*[\theta_{ro}], \delta \mathbf{u} \rangle_{L_2(0, t_F)^m} \geq 0$$

$$\Leftrightarrow \rho_{\mathbf{u}} \mathbf{u}_o + \mathcal{T}_{\theta}^* \cdot \mathcal{T}_{\theta}[\mathbf{u}_o] - \mathcal{T}_{\theta}^*[\theta_{ro}] = 0.$$

(if  $v_o \triangleq \rho_{\mathbf{u}} \mathbf{u}_o + \mathcal{T}_{\theta}^* \cdot \mathcal{T}_{\theta}[\mathbf{u}_o] - \mathcal{T}_{\theta}^*[\theta_{ro}]$  is such that  $v_o \neq 0$ , then it can be seen that  $\delta \mathbf{u} = -v_o$

violates the optimality condition)

Thus,  $\mathbf{u}_o$  is the unique solution of the linear equation (2.18).  $\square$

**Remark 2.2.** *The final-state error achieved with a given control signal, namely,*

$$\|\underline{\theta}(t_F; f_S + \boldsymbol{\beta}_S^T \mathbf{u}, g) - \theta_r\|_2^2 = \|\mathcal{T}_\theta[\mathbf{u}] - \theta_{r_o}\|_2^2$$

can be written as

$$\|\mathcal{T}_\theta[\mathbf{u}] - \hat{\theta}_{r_o}\|_2^2 + \|\theta_{r_o} - \hat{\theta}_{r_o}\|_2^2,$$

where  $\hat{\theta}_{r_o}$  denotes the  $L_2(\Omega)$ -orthogonal projection of  $\theta_{r_o}$  on the closure of  $\mathcal{T}_\theta[L_2(0, t_F)^m]$  in  $L_2(\Omega)$ . Thus, by appropriately choosing control signals, the final-state error can be made arbitrarily close to

$$\inf \left\{ \|\mathcal{T}_\theta[\mathbf{u}] - \hat{\theta}_{r_o}\|_2^2 : \mathbf{u} \in L_2(0, t_F)^m \right\} + \|\theta_{r_o} - \hat{\theta}_{r_o}\|_2^2 = \|\theta_{r_o} - \hat{\theta}_{r_o}\|_2^2.$$

In fact, this can be done with the optimal  $\mathbf{u}_o(\rho_u)$  of Prob. I, for decreasing values of  $\rho_u$ . Indeed, taking  $\varepsilon > 0$  and  $\mathbf{u}_\varepsilon \in L_2(0, t_F)^m$  such that

$$\|\mathcal{T}_\theta[\mathbf{u}_\varepsilon] - \hat{\theta}_{r_o}\|_2^2 \leq \varepsilon, \text{ the fact that } \mathcal{J}(\mathbf{u}_o(\rho_u); \rho_u) \leq \mathcal{J}(\mathbf{u}_\varepsilon; \rho_u)$$

implies that

$$\rho_u \|\mathbf{u}_o(\rho_u)\|_{L_2(0, t_F)^m}^2 + \|\mathcal{T}_\theta[\mathbf{u}_o(\rho_u)] - \hat{\theta}_{r_o}\|_2^2 \leq \rho_u \|\mathbf{u}_\varepsilon\|_{L_2(0, t_F)^m}^2 + \varepsilon.$$

Thus,

$$\forall \varepsilon > 0, \forall \rho_u > 0, \|\mathcal{T}_\theta[\mathbf{u}_o(\rho_u)] - \hat{\theta}_{r_o}\| \leq \rho_u \|\mathbf{u}_\varepsilon\|_{L_2(0, t_F)^m}^2 + \varepsilon$$

and, hence,  $\lim_{\rho_u \rightarrow 0} \|\mathcal{T}_\theta[\mathbf{u}_o(\rho_u)] - \hat{\theta}_{r_o}\|_2^2 = 0$ .  $\nabla$

Proposition 2.1 above characterizes the optimal solution  $\mathbf{u}_o$  in terms of the linear operators  $\mathcal{T}_\theta$  and  $\mathcal{T}_\theta^*$ . However, computing  $\mathbf{u}_o$  involves finding ways of computing the operator  $(\rho_u \mathbf{I} + \mathcal{T}_\theta^* \circ \mathcal{T}_\theta)^{-1}$  as well as to apply the result to  $\mathcal{T}_\theta^*[\theta_{r_o}]$ . To do so, it is natural to search for explicit approximations to  $\mathbf{u}_o$ , which are to be obtained by considering finite-dimensional approximations to the operator  $\mathcal{T}_\theta$  and  $\mathcal{T}_\theta^*$  and the corresponding version of equation (2.18). This is the theme of the next section.

## 2.2 Approximate Solutions

In this section, a sequence  $\{\mathbf{u}_K\}$  is introduced which is defined on the basis of finite-dimensional approximations to the operator  $\mathcal{T}_\theta$ . It is then shown that under appropriate conditions this sequence converges to  $\mathbf{u}_o$  in the  $L_2(0, t_F)^m$ -norm.

To this effect, let  $\{X_K\}$  be a sequence of finite-dimensional subspaces of  $H_0^1(\Omega)$  with approximability property, *i.e.*, such that  $\forall \psi \in H_0^1(\Omega)$  there exists a sequence  $\{\psi_K\} \subset H_0^1(\Omega)$  such that  $\psi_K \in X_K$  and

$$\lim_{K \rightarrow \infty} \|\psi - \psi_K\|_{H_0^1(\Omega)} = 0. \quad (2.20)$$

Let  $\mathcal{A}_K : X_K \rightarrow X_K$  be such that

$$\forall \phi \in X_K, \forall \psi \in X_K, \quad \langle \mathcal{A}_K[\phi], \psi \rangle = -\mathbf{B}[\phi, \psi]$$

or, equivalently, for an orthonormal basis  $\{\phi_1, \dots, \phi_K\}$  of  $X_K$ ,

$$\forall \phi \in X_K, \quad \mathcal{A}_K[\phi] = -\sum_{k=1}^n \mathbf{B}[\phi, \phi_k] \phi_k \quad \Leftrightarrow \quad \forall \ell = 1, \dots, n, \quad \mathcal{A}_K[\phi_\ell] = -\sum_{k=1}^n \mathbf{B}[\phi_\ell, \phi_k] \phi_k.$$

Let then  $\mathbf{A}_K \in \mathbb{R}^{n \times n}$  be defined by  $\{\mathbf{A}_K\}_{\ell k} = -\mathbf{B}[\phi_\ell, \phi_k]$ , *i.e.*,  $\mathbf{A}_K$  is the matrix representation of  $\mathcal{A}_K$  in the basis  $\{\phi_1, \dots, \phi_K\}$  so that for  $\phi = \sum_{k=1}^K \gamma_k \phi_k$ ,  $\mathcal{A}_K^\ell[\phi] = \sum_{k=1}^K \gamma_k^\ell \phi_k$ , where  $\mathcal{A}_K^\ell[\phi]$  is the  $\ell$ th-power of  $\mathcal{A}_K[\phi]$  and  $\bar{\gamma}_K^\ell = \mathbf{A}_K^\ell \bar{\gamma}$ ,  $\bar{\gamma} = [\gamma_1 \ \dots \ \gamma_K]^\top$  and  $\bar{\gamma}^\ell = [\gamma_1^\ell \ \dots \ \gamma_K^\ell]^\top$ .

**Remark 2.3.** *By way of example, consider the one-dimensional heat equation – in this case  $\Omega = (0, L_x)$  and  $\mathbf{B}[\phi, \psi] = \alpha \langle \frac{\partial \phi}{\partial x}, \frac{\partial \psi}{\partial x} \rangle$  and let  $\phi_k = \sqrt{\frac{2}{L_x}} \sin \left[ \frac{k\pi x}{L_x} \right]$ . Thus,  $\{\mathbf{A}_K\}_{\ell k} = -\alpha \langle \frac{\partial \phi_\ell}{\partial x}, \frac{\partial \phi_k}{\partial x} \rangle$ , *i.e.*,  $\{\mathbf{A}_K\}_{\ell k} = \alpha \left[ \frac{\ell\pi}{L_x} \right] \left[ \frac{k\pi}{L_x} \right] \left\langle \sqrt{\frac{2}{L_x}} \cos \left[ \frac{\ell\pi x}{L_x} \right], \cos \left[ \frac{k\pi x}{L_x} \right] \right\rangle$ , so that  $\mathbf{A}_K = \text{diag} \left( -\alpha \left[ \frac{\pi}{L_x} \right]^2 \ \dots \ -\alpha \left[ \frac{K\pi}{L_x} \right]^2 \right)$ .  $\nabla$*

Let  $P_K$  be the orthogonal projection from  $L_2(\Omega)$  onto  $X_K$  and define

$$\mathcal{T}_\theta^K : L_2(0, t_F)^m \rightarrow X_K \quad \text{by} \quad \mathcal{T}_\theta^K[\mathbf{u}] \triangleq \left[ \int_0^{t_F} S_K(t_F - \tau) [P_K [\boldsymbol{\beta}_S^\top \mathbf{u}(\tau)]] d\tau \right],$$

where  $S_K$  is the semigroup generated by  $\mathcal{A}_K$ , *i.e.*,  $S_K(t) = \sum_{\ell=0}^{\infty} \mathcal{A}_K^\ell t^\ell / \ell!$ . Thus,

$$\begin{aligned} \mathcal{T}_\theta^K[\mathbf{u}] &= \int_0^{t_F} S_K(t_F - \tau) \left[ P_K \left[ \sum_{i=1}^m \beta_{\mathbf{S}_i} u_i(\tau) \right] \right] d\tau \Leftrightarrow \\ \mathcal{T}_\theta^K[\mathbf{u}] &= \int_0^{t_F} S_K(t_F - \tau) \left[ \sum_{i=1}^m P_K[\beta_{\mathbf{S}_i}] u_i(\tau) \right] d\tau \Leftrightarrow \\ \mathcal{T}_\theta^K[\mathbf{u}] &= \int_0^{t_F} \sum_{i=1}^m \{S_K(t_F - \tau) [P_K[\beta_{\mathbf{S}_i}]]\} u_i(\tau) d\tau. \end{aligned} \quad (2.21)$$

Moreover,  $P_K[\beta_{\mathbf{S}_i}] = \sum_{k=1}^K \langle \beta_{\mathbf{S}_i}, \phi_k \rangle \phi_k$ .

Now, for  $\phi = \sum_{k=1}^K \gamma_k \phi_k$ ,

$$\begin{aligned} S_K(t)[\phi] &= \sum_{\ell=0}^{\infty} \frac{t^\ell}{\ell!} \mathcal{A}_K^\ell [\phi] = \sum_{\ell=0}^{\infty} \frac{t^\ell}{\ell!} \sum_{q=1}^K \bar{\gamma}_K^\ell \phi_q = \sum_{\ell=0}^{\infty} \sum_{q=1}^K \frac{t^\ell}{\ell!} \{e_q^T \mathbf{A}_K^\ell \bar{\gamma}\} \phi_q \\ &= \sum_{q=1}^K c_q^S[\phi](t) \phi_q, \end{aligned}$$

where

$$c_q^S[\phi](t) = \sum_{\ell=0}^{\infty} \frac{t^\ell}{\ell!} e_q^T \mathbf{A}_K^\ell \bar{\gamma} = e_q^T \left[ \sum_{\ell=0}^{\infty} \frac{t^\ell}{\ell!} \mathbf{A}_K^\ell \right] \bar{\gamma}$$

so that the vector of coefficients  $\underline{c}_S = [c_1^S[\phi](t) \cdots c_K^S[\phi](t)]^T$  is given by  $\underline{c}_S[\phi](t) = \exp[\mathbf{A}_K t] \bar{\gamma}$ .

It then follows that

$$S_K(t) [P_K[\beta_{\mathbf{S}_i}]] = \sum_{q=1}^K c_q^S [P_K[\beta_{\mathbf{S}_i}]](t) \phi_q,$$

where

$$\underline{c}_S [P_K[\beta_{\mathbf{S}_i}]](t) = \exp[\mathbf{A}_K t] \begin{bmatrix} \langle \beta_{\mathbf{S}_i}, \phi_1 \rangle \\ \vdots \\ \langle \beta_{\mathbf{S}_i}, \phi_K \rangle \end{bmatrix}. \quad (2.22)$$



Thus, taking (2.22) into (2.21) leads to

$$\begin{aligned}\mathcal{T}_\theta^K[\mathbf{u}] &= \int_0^{t_F} \sum_{i=1}^m \sum_{q=1}^K e_q^\top \left\{ \exp[\mathbf{A}_K(t-\tau)] \begin{bmatrix} \langle \boldsymbol{\beta}_{S_i}, \phi_1 \rangle \\ \vdots \\ \langle \boldsymbol{\beta}_{S_i}, \phi_K \rangle \end{bmatrix} u_i(\tau) \right\} \phi_q d\tau \\ &= \sum_{q=1}^K e_q^\top \left\{ \int_0^{t_F} \exp[\mathbf{A}_K(t-\tau)] \mathbf{M}_\beta^K \mathbf{u}(\tau) d\tau \right\} \phi_q\end{aligned}$$

so that  $\mathcal{T}_\theta^K[\mathbf{u}] = \sum_{q=1}^K c_q(t_F; \mathbf{u}) \phi_q$ , where  $\underline{c}_K(t; \mathbf{u}) = [c_1(t; \mathbf{u}), \dots, c_K(t; \mathbf{u})]^\top$  is given by

$$\underline{c}_K(t; \mathbf{u}) = \int_0^t \exp[\mathbf{A}_K(t-\tau)] \mathbf{M}_\beta^K \mathbf{u}(\tau) d\tau, \quad \boldsymbol{\beta}_S^\top = [\boldsymbol{\beta}_{S_1} \cdots \boldsymbol{\beta}_{S_m}] \quad \text{and}$$

$$\mathbf{M}_\beta^K \triangleq \begin{bmatrix} \langle \boldsymbol{\beta}_{S_1}, \phi_1 \rangle & \cdots & \langle \boldsymbol{\beta}_{S_m}, \phi_1 \rangle \\ \vdots & & \vdots \\ \langle \boldsymbol{\beta}_{S_1}, \phi_K \rangle & \cdots & \langle \boldsymbol{\beta}_{S_m}, \phi_K \rangle \end{bmatrix}.$$

The corresponding version of *Prob. I* is then defined by

$$\underline{\text{Prob. } I_K} : \min_{\mathbf{u} \in L_2(0, t_F)^m} \mathcal{J}_K(\mathbf{u}), \quad (2.23)$$

where

$$\mathcal{J}_K[\mathbf{u}] \triangleq \|\mathcal{T}_\theta^K[\mathbf{u}] - \theta_{ro}\|_{L_2(\Omega)}^2 + \rho_u \|\mathbf{u}\|_{L_2(0, t_F)^m}^2.$$

Similarly to what happens in the case of *Prob. I*, *Prob. I<sub>K</sub>* has a unique solution  $\mathbf{u}_K$  which is obtained from the optimality condition

$$\rho_u \mathbf{u}_K + (\mathcal{T}_\theta^K)^* [\mathcal{T}_\theta^K[\mathbf{u}_K] - \theta_{ro}] = 0, \quad (2.24)$$

where the adjoint operator  $(\mathcal{T}_\theta^K)^* : L_2(\Omega) \rightarrow L_2(0, t_F)^m$  is such that

$$\begin{aligned}\forall \mathbf{u} \in L_2(0, t_F)^m, \quad \forall \phi \in L_2(\Omega), \quad \langle \phi, \mathcal{T}_\theta^K[\mathbf{u}] \rangle &= \langle (\mathcal{T}_\theta^K)^*[\phi], \mathbf{u} \rangle \\ \Leftrightarrow \sum_{k=1}^n \langle \phi, \phi_k \rangle c_k(t_F; \mathbf{u}) &= \bar{\boldsymbol{\phi}}_K^\top \underline{c}_K(t_F; \mathbf{u}) = \int_0^{t_F} (\mathbf{F}_K(\tau) \bar{\boldsymbol{\phi}}_K)^\top \mathbf{u}(\tau) d\tau\end{aligned}$$

so that  $(\mathcal{T}_\theta^K)^*[\phi] = \mathbf{F}_K(\tau)\bar{\phi}_K$ , where  $\bar{\phi}_K^T \triangleq [\langle\phi, \phi_1\rangle \cdots \langle\phi, \phi_K\rangle]$  and

$$\mathbf{F}_K(\tau) \triangleq (\mathbf{M}_\beta^K)^T \exp[\mathbf{A}_K^T(t_F - \tau)]. \quad (2.25)$$

To obtain  $\mathbf{u}_K$  note that it follows from (2.24) that  $\mathbf{u}_K$  belongs to the image of  $(\mathcal{T}_\theta^K)^*$ , *i.e.*, there exists

$$\phi \in L_2(\Omega) \text{ such that } \mathbf{u}_K = (\mathcal{T}_\theta^K)^*[\phi] = \mathbf{F}_K\bar{\phi}_K,$$

*i.e.*, there exists  $\bar{\alpha}_K \in \mathbb{R}^n$  such that

$$\mathbf{u}_K = \mathbf{F}_K\bar{\alpha}_K. \quad (2.26)$$

It then follows from (2.24) that

$$\rho_u \mathbf{F}_K \bar{\alpha}_K + \mathbf{F}_K \underline{\mathbf{c}}_K(t_F; \mathbf{F}_K \bar{\alpha}_K) - \mathbf{F}_K \bar{\theta}_{ro}^K = 0 \quad (2.27)$$

a sufficient condition for which being

$$\rho_F \bar{\alpha}_K + \underline{\mathbf{c}}_K(t_F; \mathbf{F}_K \bar{\alpha}_K) - \bar{\theta}_{ro}^K = \mathbf{0}, \quad (2.28)$$

where  $\bar{\theta}_{ro}^K \triangleq [\langle\phi_1, \theta_{ro}\rangle \cdots \langle\phi_{n_K}, \theta_{ro}\rangle]^T$ .

Thus, as  $\underline{\mathbf{c}}_K(t_F; \mathbf{F}_K \bar{\alpha}_K) = \mathbf{G}_K \bar{\alpha}_K$ , where  $\mathbf{G}_K \triangleq \int_0^{t_F} \mathbf{F}_K(\tau)^T \mathbf{F}_K(\tau) d\tau$ , (2.27) can be rewritten as  $\rho_u \bar{\alpha}_K + \mathbf{G}_K \bar{\alpha}_K = \bar{\theta}_{ro}^K$  from which it follows that  $\bar{\alpha}_K = (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} \bar{\theta}_{ro}^K$  and, hence,

$$\mathbf{u}_K(\tau) = \mathbf{F}_K(\tau) (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} \bar{\theta}_{ro}^K, \quad \tau \in [0, t_F]. \quad (2.29)$$

**Remark 2.4.** *It is interesting to notice that  $\mathbf{G}_K$  can be computed from a linear equation in  $\mathbb{R}^{n_K \times n_K}$ . Indeed from (2.27) it can be seen that  $\mathbf{G}_K$  can be expressed as*

$\mathbf{G}_K = \int_0^{t_F} \exp[\mathbf{A}_K(t_F - \tau)] \mathbf{M}_\beta^K (\exp[\mathbf{A}_K(t_F - \tau)] \mathbf{M}_\beta^K)^T d\tau$ . If we define  $\omega = t_F - \tau$  we have that  $\mathbf{G}_K = \int_0^{t_F} \check{\mathbf{H}}_K(\omega) d\omega$ , where  $\check{\mathbf{H}}_K(\omega) \triangleq \exp[\mathbf{A}_K(\omega)] \mathbf{M}_\beta^K (\exp[\mathbf{A}_K(\omega)] \mathbf{M}_\beta^K)^T$ .

We see that

$$\begin{aligned} \frac{d}{d\omega} \check{\mathbf{H}}_K(\omega) &= \mathbf{A}_K \{ \exp[\mathbf{A}_K(\omega)] \mathbf{M}_\beta^K (\exp[\mathbf{A}_K(\omega)] \mathbf{M}_\beta^K)^T \} \\ &\quad + \{ \exp[\mathbf{A}_K(\omega)] \mathbf{M}_\beta^K (\exp[\mathbf{A}_K(\omega)] \mathbf{M}_\beta^K)^T \} \mathbf{A}_K^T = \mathbf{A}_K \check{\mathbf{H}}_K(\omega) + \check{\mathbf{H}}_K(\omega) \mathbf{A}_K^T. \end{aligned}$$

By integrating both sides from 0 to  $\omega$  we have an expression for all  $\omega \in [0, t_F]$  given by

$$\check{\mathbf{H}}(\omega) - \check{\mathbf{H}}(0) = \int_0^\omega \{\mathbf{A}_K \check{\mathbf{H}}_K(\sigma) + \check{\mathbf{H}}_K(\sigma) \mathbf{A}_K^T\} d\sigma = \mathbf{A}_K \int_0^\omega \check{\mathbf{H}}_K(\sigma) d\sigma + \int_0^\omega \check{\mathbf{H}}_K(\sigma) d\sigma \mathbf{A}_K^T.$$

Then, for  $\omega = t_F$  we have  $\mathbf{A}_K \mathbf{G}_K + \mathbf{G}_K \mathbf{A}_K^T = \check{\mathbf{M}}_K$ , where  $\check{\mathbf{M}}_K = \check{\mathbf{H}}(t_F) - \check{\mathbf{H}}(0) = \exp[\mathbf{A}_K t_F] \mathbf{M}_\beta^K (\exp[\mathbf{A}_K t_F] \mathbf{M}_\beta^K)^T - \mathbf{M}_\beta^K (\mathbf{M}_\beta^K)^T$ . Therefore,  $\mathbf{G}_K$  can be obtained as the unique solution of the Lyapunov equation, see (LAUB, 2005, pp. 144 – 148), (ZHOU; DOYLE; GLOVER, 1996, pp.71 – 72).  $\nabla$

**Remark 2.5.** Note that  $\mathbf{u}_K : [0, t_F] \rightarrow \mathbb{R}^m$  is explicitly given by (2.29) in terms of  $\exp[\mathbf{A}_K^T(t_F - \tau)]$ . Note also that  $\mathbf{u}_K$  can be obtained from the solution of the linear ordinary differential equation  $\dot{\mathbf{x}}_u(\tau) = -\mathbf{A}_K^T \mathbf{x}_u(\tau)$ ,  $\tau \geq 0$  with the initial condition  $\mathbf{x}_u(0) = \exp[\mathbf{A}_K^T t_F] (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} \bar{\boldsymbol{\theta}}_{r_o}^K$ , i.e.,  $\mathbf{u}(\tau) = (\mathbf{M}_\beta^K)^T \mathbf{x}_u(\tau)$ .  $\nabla$

The next step is to analyze the question of whether the sequence  $\{\mathbf{u}_K\}$  of approximate solutions to the optimal control problem converges to the solution  $\mathbf{u}_o$  of the original problem. To this effect, consider the following proposition (which is proved in the APPENDIX at the end of this chapter).

**Proposition 2.2.** *There exists a real sequence  $\{\eta_{\mathcal{T}}^K : K \in \mathbb{Z}_+\}$  such that*

$$(a) \forall \mathbf{u} \in L_2(0, t_F)^m, \quad \|\mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}]\|_{L_2(\Omega)} \leq \eta_{\mathcal{T}}^K \|\mathbf{u}\|_{L_2(0, t_F)^m}.$$

(b)  $\{\eta_{\mathcal{T}}^K\}$  converges to zero.  $\nabla$

$$\text{Note now that } \mathcal{J}_K(\mathbf{u}) = \rho_u \|\mathbf{u}\|_{L_2(0, t_F)}^2 + \|\mathcal{T}_\theta[\mathbf{u}] - \theta_{r_o} - (\mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}])\|_2^2 \iff \mathcal{J}_K(\mathbf{u}) = \mathcal{J}(\mathbf{u}) + \|\mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}]\|_2^2 - 2\langle \mathcal{T}_\theta[\mathbf{u}] - \theta_{r_o}, \mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}] \rangle.$$

As a result, with  $E_{\mathcal{J}}^K(\mathbf{u}) \triangleq \mathcal{J}(\mathbf{u}) - \mathcal{J}_K(\mathbf{u})$ , it follows from Proposition 2.2 that

$$|E_{\mathcal{J}}^K(\mathbf{u})| \leq (\eta_{\mathcal{T}}^K)^2 \|\mathbf{u}\|_{L_2(0, t_F)^m}^2 + 2\|\mathcal{T}_\theta[\mathbf{u}] - \theta_{r_o}\|_2 (\eta_{\mathcal{T}}^K) \|\mathbf{u}\|_{L_2(0, t_F)^m}. \quad (2.30)$$

On the other hand,

$$\begin{aligned} \mathcal{J}_K(\mathbf{u}_K) &\leq \mathcal{J}_K(\mathbf{u}_o) = \mathcal{J}(\mathbf{u}_o) - E_{\mathcal{J}}^K(\mathbf{u}_o) \iff \mathcal{J}(\mathbf{u}_K) - E_{\mathcal{J}}^K(\mathbf{u}_K) \leq \mathcal{J}(\mathbf{u}_o) - E_{\mathcal{J}}^K(\mathbf{u}_o) \\ &\implies \mathcal{J}(\mathbf{u}_K) \leq \mathcal{J}(\mathbf{u}_o) - E_{\mathcal{J}}^K(\mathbf{u}_o) + E_{\mathcal{J}}^K(\mathbf{u}_K) \implies \end{aligned}$$

$$\mathcal{J}(\mathbf{u}_K) \leq \mathcal{J}(\mathbf{u}_o) + |E_{\mathcal{J}}^K(\mathbf{u}_o)| + |E_{\mathcal{J}}^K(\mathbf{u}_K)|$$

$\implies$  (since  $\mathcal{J}(\mathbf{u}_K) \geq \mathcal{J}(\mathbf{u}_o)$ )

$$0 \leq \mathcal{J}(\mathbf{u}_K) - \mathcal{J}(\mathbf{u}_o) \leq |E_{\mathcal{J}}^K(\mathbf{u}_o)| + |E_{\mathcal{J}}^K(\mathbf{u}_K)|. \quad (2.31)$$

Note also that, as  $\eta_{\mathcal{J}}^K \rightarrow 0$  (Proposition 2.2(b)), it follows from (2.30) that  $|E_{\mathcal{J}}^K(\mathbf{u}_o)| \rightarrow 0$ . Moreover,  $\{\mathbf{u}_K\}$  is a bounded sequence – indeed,  $\|\mathbf{u}_K\|_{L_2(0,t_F)^m}^2 \leq \|\boldsymbol{\theta}_{ro}\|_{L_2(\Omega)}^2 \rho_{\mathbf{u}}^{-1}$  for, if  $\|\mathbf{u}_K\|^2 > \rho_{\mathbf{u}}^{-1} \|\boldsymbol{\theta}_{ro}\|_{L_2(\Omega)}^2$  then  $\mathcal{J}_K(\mathbf{u}_K) > \|\boldsymbol{\theta}_{ro}\|_{L_2(\Omega)}^2 = \mathcal{J}_K(0)$  in which case  $\mathbf{u}_K$  would not be optimal for *Prob. I*<sub>K</sub>. Thus, as

$\mathcal{T}_{\theta}[\mathbf{u}] = \int_0^{t_F} S_A(t_F - \tau) \left\{ \sum_{i=1}^m \beta_{S_i} \mathbf{u}(\tau) \right\} d\tau$ ,  $\{\mathcal{T}_{\theta}[\mathbf{u}_K]\}$  is also bounded and, hence, it follows from (2.30) that (as  $\eta_{\mathcal{J}}^K \rightarrow 0$ )  $E_{\mathcal{J}}^K(\mathbf{u}_K) \rightarrow 0$ . Thus,

$$\{|E_{\mathcal{J}}^K(\mathbf{u}_o)| + |E_{\mathcal{J}}^K(\mathbf{u}_K)|\} \rightarrow 0 \quad (2.32)$$

which together with (2.31) implies that  $\mathcal{J}(\mathbf{u}_K) \rightarrow \mathcal{J}(\mathbf{u}_o)$ . Thus, the following corollary of Proposition 2.2 has been established.

**Corollary 2.1:**  $\mathcal{J}(\mathbf{u}_K) \rightarrow \mathcal{J}(\mathbf{u}_o)$ . ∇

Moreover, as  $\{\mathbf{u}_K\}$  is bounded and  $\mathcal{J}(\mathbf{u}_K) \rightarrow \mathcal{J}(\mathbf{u}_o)$ , the desired convergence of the approximate solutions  $\{\mathbf{u}_K\}$  can be established, as stated in the following proposition.

**Proposition 2.3.** *The sequence  $\{\mathbf{u}_K : K \in \mathbb{Z}_+\}$  of solutions to the approximate problems *Prob. I*<sub>K</sub> converges to the solution  $\mathbf{u}_o$  of *Prob. I* in the sense of the  $L_2(0, t_F)^m$ -norm. ∇*

*Proof.* Note first that (since  $\mathbf{u}_o$  is an optimal solution of *Prob. I*)

$$\mathcal{J}(\mathbf{u}_K) = \mathcal{J}(\mathbf{u}_o + (\mathbf{u}_K - \mathbf{u}_o)) = \mathcal{J}(\mathbf{u}_o) + \rho_{\mathbf{u}} \|\mathbf{u}_K - \mathbf{u}_o\|_{L_2(0,t_F)^m}^2 + \|\mathcal{T}_{\theta}[(\mathbf{u}_K - \mathbf{u}_o)]\|_{L_2(\Omega)}^2.$$

It then follows from (2.31) that

$$\rho_{\mathbf{u}} \|\mathbf{u}_K - \mathbf{u}_o\|_{L_2(0,t_F)^m}^2 + \|\mathcal{T}_{\theta}[(\mathbf{u}_K - \mathbf{u}_o)]\|_{L_2(\Omega)}^2 \leq |E_{\mathcal{J}}^K(\mathbf{u}_o)| + |E_{\mathcal{J}}^K(\mathbf{u}_K)| \Rightarrow$$

$$\rho_{\mathbf{u}} \|\mathbf{u}_K - \mathbf{u}_o\|_{L_2(0,t_F)^m}^2 \leq |E_{\mathcal{J}}^K(\mathbf{u}_o)| + |E_{\mathcal{J}}^K(\mathbf{u}_K)|.$$

Thus, in the light of (2.32),  $\mathbf{u}_K \rightarrow \mathbf{u}_o$  in  $L_2(0, t_F)^m$ . □

To conclude this chapter, a summary is presented of the steps required to compute the approximate solution  $\mathbf{u}_K$  for the problem  $\min_{\mathbf{u} \in L_2(0, t_F)^m} \mathcal{J}(\mathbf{u})$  where  $\mathcal{J}(\mathbf{u})$  is given by (2.11).

Given the problem data  $(f, g, \theta_r, \rho_u)$  and a family  $\{S_K\}$  of subspaces each with an orthonormal basis  $\{\phi_1 \dots \phi_{n_K}\}$ :

- (1) Compute  $\bar{\boldsymbol{\theta}}_{r_o}^K = [\langle \theta_{r_o}, \phi_1 \rangle \dots \langle \theta_{r_o}, \phi_{n_K} \rangle]^T$ , where  $\theta_{r_o} = \theta_r - \underline{\theta}(t_F; f, g)$  and  $\underline{\theta}(t_F; f, g)$  is given by (2.9).
- (2) Compute  $\mathbf{M}_\beta^K \in \mathbb{R}^{n_K \times m}$ , where  $\{\mathbf{M}_\beta^K\}_{ki} = \langle \boldsymbol{\beta}_{S_i}, \phi_k \rangle$ .
- (3) For  $\mathbf{A}_K \in \mathbb{R}^{n_K}$  such that  $\{\mathbf{A}_K\}_{\ell k} = -\alpha \sum_{i=1}^{m_x} \left\langle \frac{\partial \phi_\ell}{\partial x_i}, \frac{\partial \phi_k}{\partial x_i} \right\rangle$  compute  $\mathbf{G}_K$  solving the Lyapunov equation  $\mathbf{A}_K \mathbf{G}_K + \mathbf{G}_K \mathbf{A}_K^T = \check{\mathbf{M}}_K$ , where  $\check{\mathbf{M}}_K = \exp[\mathbf{A}_K t_F] \mathbf{M}_\beta^K (\exp[\mathbf{A}_K t_F] \mathbf{M}_\beta^K)^T - \mathbf{M}_\beta^K (\mathbf{M}_\beta^K)^T$ .
- (4)  $\mathbf{u}_K$  can then be obtained from (2.29) (see also Remark 2.5).

In the case of primary interest here, *i.e.*, with  $\Omega = (0, L_{x_1}) \times \dots \times (0, L_{x_{m_x}})$  and  $S_K$  the span of the eigenfunctions of  $A$ ,  $\{\phi_{\mathbf{k}}(x_1 \dots x_{m_x}) : \mathbf{k} = (k_1, \dots, k_{m_x}), k_i \leq K\}$  where  $\phi_{\mathbf{k}}$  are as in Remark 2.1, *i.e.*,  $\langle \theta_{r_o}, \phi_{\mathbf{k}} \rangle = \langle \theta_r, \phi_{\mathbf{k}} \rangle - \langle S_A(t)[g], \phi_{\mathbf{k}} \rangle - \left\langle \int_0^{t_F} S_A(t_F - \tau) [f(\tau)] d\tau, \phi_{\mathbf{k}} \right\rangle \Rightarrow$  (in the light of (2.10) that)

$$\langle \theta_{r_o}, \phi_{\mathbf{k}} \rangle = \langle \theta_r, \phi_{\mathbf{k}} \rangle - e^{\lambda_{\mathbf{k}} t_F} \langle g, \phi_{\mathbf{k}} \rangle - \int_0^{t_F} e^{\lambda_{\mathbf{k}}(t_F - \tau)} \langle \underline{f}(\tau), \phi_{\mathbf{k}} \rangle d\tau, \quad (2.33)$$

where  $\lambda_{\mathbf{k}} = -\alpha \sum_{i=1}^{m_x} \left[ \frac{k_i \pi}{L_{x_i}} \right]^2$ .

Moreover, in the case  $\mathbf{A}_K \in \mathbb{R}^{n_K}$  is diagonal – in the one-dimensional case ( $m_x = 1$ ),  $n_K = K$  and  $\{\mathbf{A}_K\}_{kk} = -\alpha \left[ \frac{k\pi}{L_x} \right]^2$ . This allows for the Lyapunov equation to be solved term by term,

$$\{\mathbf{A}_K \mathbf{G}_K\}_{kl} + \{\mathbf{G}_K \mathbf{A}_K^T\}_{kl} = \{\check{\mathbf{M}}_K\}_{kl} \Leftrightarrow$$

$$\{\mathbf{A}_K\}_{kk} \{\mathbf{G}_K\}_{kl} + \{\mathbf{G}_K\}_{kl} \{\mathbf{A}_K\}_{\ell\ell} = \{\check{\mathbf{M}}_K\}_{kl} \Leftrightarrow$$

$$\{\mathbf{G}_K\}_{kl} = \{\check{\mathbf{M}}_K\}_{kl} / (\{\mathbf{A}_K\}_{kk} + \{\mathbf{A}_K\}_{\ell\ell}) \Leftrightarrow$$

$$\{\mathbf{G}_K\}_{\ell\ell} = \frac{1}{\{\mathbf{A}_K\}_{kk} + \{\mathbf{A}_K\}_{\ell\ell}} [1 - \exp\{(\{\mathbf{A}_K\}_{kk} + \{\mathbf{A}_K\}_{\ell\ell}) t_F\}] \{\mathbf{M}_\beta^K (\mathbf{M}_\beta^K)^T\}_{kl}.$$

Thus, in this case, the computations required to obtain  $\mathbf{u}_K$  amount to the numerical evaluation of the integrals  $\langle \theta_r, \phi_{\mathbf{k}} \rangle$ ,  $\langle g, \phi_{\mathbf{k}} \rangle_{\mathbf{k}}$  and  $\langle \beta_{\mathbf{S}i}, \phi_{\mathbf{k}} \rangle$  over the spatial domain  $\Omega$ , of the (scalar) exponential function over the time-interval  $(0, t_F)$  and the last integral in (2.33) over both time and space – note that if  $f(x, t) = \sum_{i=1}^m \beta_{\mathbf{S}i}(x) u_i(t)$ , then

$$\int_0^{t_F} e^{\lambda_{\mathbf{k}}(t_F - \tau)} \langle \underline{f}(\tau), \phi_{\mathbf{k}} \rangle d\tau = \sum_{i=1}^m \left\{ \int_0^{t_F} e^{\lambda_{\mathbf{k}}(t_F - \tau)} u_i(\tau) d\tau \langle \beta_{\mathbf{S}i}, \phi_{\mathbf{k}} \rangle \right\}.$$

So that computing the last integral in (2.33) is reduced to computing  $\langle \beta_{\mathbf{S}i}, \phi_{\mathbf{k}} \rangle$  and  $\int_0^{t_F} e^{\lambda_{\mathbf{k}}(t_F - \tau)} u_i(\tau) d\tau$ .

**Remark 2.6.** *It is often necessary to make  $\rho_F = \rho_{\mathbf{u}}^{-1}$  large in order to achieve an acceptably small value for the final-state error norm (see Remark 2.2). This might make the conditioning number (with respect to inversion) of the matrix  $(\mathbf{I} + \rho_F \mathbf{G}_K)$  very large which, in-turn, could give rise to numerical difficulties in the process of computing its inverse. To cope with this potential problem using the available tools from MATLAB<sup>®</sup>, the symmetric structure of  $(\mathbf{I} + \rho_F \mathbf{G}_K)$  was exploited in the following way:*

- (a) Choose  $\delta_S > 0$  and put  $\check{\mathbf{G}}_{K\rho} = (1 + \delta_S)\mathbf{I} + \rho_F \mathbf{G}_K$  (Note that the eigenvalues of  $\check{\mathbf{G}}_{K\rho}$  are not smaller than  $1 + \delta_S$ ).
- (b) Take a SVD decomposition  $\check{\mathbf{G}}_{K\rho} = \mathbf{V}_{K\rho} \check{\Sigma}_{K\rho} \mathbf{U}_{K\rho}^T$  of  $\check{\mathbf{G}}_{K\rho}$  (note that as  $\check{\mathbf{G}}_{K\rho}$  is symmetric and positive  $\mathbf{V}_{K\rho} = \mathbf{U}_{K\rho}$ ,  $\mathbf{V}_{K\rho}$  and  $\check{\Sigma}_{K\rho}$  are the eigenvector and eigenvalue matrices of  $\check{\mathbf{G}}_{K\rho}$ , respectively).
- (c) As  $\check{\mathbf{G}}_{K\rho} = \mathbf{V}_{K\rho} \check{\Sigma}_{K\rho} \mathbf{V}_{K\rho}^T$ ,  $\mathbf{G}_{K\rho} = \check{\mathbf{G}}_{K\rho} - \delta_S \mathbf{I} = \mathbf{V}_{K\rho} [\check{\Sigma}_{K\rho} - \delta_S \mathbf{I}] \mathbf{V}_{K\rho}^T$ .
- (d) Put  $\mathbf{G}_{K\rho}^{-1} = \mathbf{V}_{K\rho} (\check{\Sigma}_{K\rho} - \delta_S \mathbf{I})^{-1} \mathbf{V}_{K\rho}^T$  (Note that as  $(\check{\Sigma}_{K\rho} - \delta_S \mathbf{I})$  is diagonal computing its inverse is numerically straightforward).

▽

### 2.3 APPENDIX – PROOFS FROM CHAPTER 2

**Proof of Proposition 2.2(a):** Proposition 2.2(a) is an immediate consequence of the following auxiliary propositions

**Auxiliary Proposition 1:**  $\mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}] = E_S^K[\mathbf{u}] + E_{\mathcal{T}}^K[\mathbf{u}]$  where

$$E_S^K[\mathbf{u}] \triangleq \int_0^{t_F} \sum_{i=1}^m (S_A(t_F - \tau) - S_K(t_F - \tau)) [P_K[\boldsymbol{\beta}_{\mathbf{S}_i}]] u_i(\tau) d\tau \quad \text{and}$$

$$E_{\mathcal{T}}^K[\mathbf{u}] \triangleq \int_0^{t_F} \sum_{i=1}^m S_A(t_F - \tau) [(\mathbf{I} - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i}]] u_i(\tau) d\tau. \quad \nabla$$

**Auxiliary Proposition 2:**  $\|E_{\mathcal{T}}^K[\mathbf{u}]\|_{L_2(\Omega)} \leq \eta_{\mathcal{T}\mathbf{f}}^K \|\mathbf{u}\|_{L_2(0,t_F)^m}$  and

$$\|E_S^K[\mathbf{u}]\|_{L_2(\Omega)} \leq \eta_{\mathcal{T}\mathbf{g}}^K \|\mathbf{u}\|_{L_2(0,t_F)^m},$$

where

$$\eta_{\mathcal{T}\mathbf{f}}^K \triangleq \left\{ \sum_{i=1}^m \|f_i^K(t_F - \cdot)\|_{L_2(0,t_F)}^2 \right\}^{1/2}, \quad \eta_{\mathcal{T}\mathbf{g}}^K \triangleq \left\{ \sum_{i=1}^m \|g_i^K(t_F - \cdot)\|_{L_2(0,t_F)}^2 \right\}^{1/2}$$

$$f_i^K(t_F - \sigma) \triangleq \|S_A(t_F - \sigma) [(\mathbf{I} - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i}]]\|_{L_2(\Omega)} \quad \text{and}$$

$$g_i^K(t_F - \sigma) \triangleq \|(S_A(t_F - \sigma) - S_K(t_F - \sigma)) [P_K[\boldsymbol{\beta}_{\mathbf{S}_i}]]\|_{L_2(\Omega)}. \quad \nabla$$

Proposition 2.2(a) follows immediately from the two statements above, since bringing the second one to bear on the first leads to

$$\|\mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}]\|_{L_2(\Omega)} \leq (\eta_{\mathcal{T}\mathbf{f}}^K + \eta_{\mathcal{T}\mathbf{g}}^K) \|\mathbf{u}\|_{L_2(0,t_F)^m} \quad (\text{i.e., } \eta_\theta^K = \eta_{\mathcal{T}\mathbf{f}}^K + \eta_{\mathcal{T}\mathbf{g}}^K).$$

□

**Proof of Auxiliary Proposition 1:** Recall that

$$\mathcal{T}_\theta^K[\mathbf{u}] = \int_0^{t_F} S_K(t_F - \tau) [P_K[\boldsymbol{\beta}_{\mathbf{S}}^T \mathbf{u}(\tau)]] d\tau = \int_0^{t_F} \sum_{i=1}^m (S_K(t_F - \tau) [P_K[\boldsymbol{\beta}_{\mathbf{S}_i} u_i(\tau)]]) d\tau \quad (2.A.1)$$

Note now that

$$\mathcal{T}_\theta[\mathbf{u}] = \int_0^{t_F} S_A(t_F - \tau) [\boldsymbol{\beta}_{\mathbf{S}}^T \mathbf{u}(\tau)] d\tau = \int_0^{t_F} \sum_{i=1}^m S_A(t_F - \tau) [\boldsymbol{\beta}_{\mathbf{S}_i}] u_i(\tau) d\tau.$$

This, taking the orthogonal projections of  $\boldsymbol{\beta}_{\mathbf{S}_i}$  on  $S_K$  and its orthogonal complement (in  $L_2(\Omega)$ ), *i.e.*,  $\boldsymbol{\beta}_{\mathbf{S}_i} = P_K[\boldsymbol{\beta}_{\mathbf{S}_i}] + (\mathbf{I} - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i}]$  it follows that

$$\mathcal{T}_\theta[\mathbf{u}] = \int_0^{t_F} \sum_{i=1}^m S_A(t_F - \tau) [P_K[\boldsymbol{\beta}_{\mathbf{S}_i}] u_i] d\tau,$$

where  $E_{\mathcal{T}}^K[\mathbf{u}] \triangleq \int_0^{t_F} \sum_{i=1}^m S_A(t_F - \tau) [(\mathbf{I} - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i} u_i(\tau)]] d\tau.$

As a result,

$$\mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}] = \int_0^{t_F} \sum_{i=1}^m S_A(t_F - \tau) [P_K[\boldsymbol{\beta}_{\mathbf{S}_i} u_i(\tau)]] d\tau + E_{\mathcal{T}}^K[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}]$$

$$\Rightarrow \text{(in the light of (2.A.1)) } \mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}] = E_S^K[\mathbf{u}] + E_{\mathcal{T}}^K[\mathbf{u}],$$

where  $E_S^K[\mathbf{u}]$  is defined above (in the statement of Auxiliary Proposition 1). □



**Proof of Auxiliary Proposition 2:** Note first that

$$\begin{aligned}
\|E_{\mathcal{T}}^K[\mathbf{u}]\|_{L_2(\Omega)} &\leq \int_0^{t_F} \left\| \sum_{i=1}^m S_A(t_F - \tau) [(\mathbf{I} - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i}]] u_i(\tau) \right\|_{L_2(\Omega)} d\tau \Rightarrow \\
\|E_{\mathcal{T}}^K[\mathbf{u}]\|_{L_2(\Omega)} &\leq \int_0^{t_F} \sum_{i=1}^m \|S_A(t_F - \tau) [(\mathbf{I} - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i}]] u_i(\tau)\|_{L_2(\Omega)} d\tau \Rightarrow \\
\|E_{\mathcal{T}}^K[\mathbf{u}]\|_{L_2(\Omega)} &\leq \sum_{i=1}^m \int_0^{t_F} \|S_A(t_F - \tau) [(I - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i}]] u_i(\tau)\|_{L_2(\Omega)} d\tau \Rightarrow \\
\|E_{\mathcal{T}}^K[\mathbf{u}]\|_{L_2(\Omega)} &\leq \sum_{i=1}^m \int_0^{t_F} \|S_A(t_F - \tau) [(I - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i}]] \| |u_i(\tau)| d\tau \Rightarrow \\
\|E_{\mathcal{T}}^K[\mathbf{u}]\|_{L_2(\Omega)} &\leq \sum_{i=1}^m \int_0^{t_F} f_i^K(t_F - \tau) |u_i(\tau)| d\tau \Rightarrow \\
\|E_{\mathcal{T}}^K[\mathbf{u}]\|_{L_2(\Omega)} &\leq \sum_{i=1}^m \|f_i^K(t_F - \cdot)\|_{L_2(0, t_F)} \|\mathbf{u}\|_{L_2(0, t_F)},
\end{aligned}$$

where  $f_i^K(t_F - \tau) \triangleq \|S_A(t_F - \tau) [(\mathbf{I} - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i}]]\|_{L_2(\Omega)}$ . Note that  $\int_0^{t_F} f_i^K(t_F - \tau) |u_i(\tau)| d\tau$  is the inner product in  $L_2(0, t_F)$  of the function  $f_i^K : (t_F - \cdot) : [0, t_F] \rightarrow \mathbb{R}$  and  $|u_i(\cdot)| : \mathbb{R} \rightarrow \mathbb{R}$  so that, in the light of the Cauchy-Schwarz (CS) inequality

$$\int_0^{t_F} f_i^K(t_F - \tau) |u_i(\tau)| d\tau \leq \|f_i^K(t_F - \cdot)\|_{L_2(0, t_F)} \|\mathbf{u}\|_{L_2(0, t_F)}.$$

As a result,  $\|E_{\mathcal{T}}^K[\mathbf{u}]\|_{L_2(\Omega)} \leq \sum_{i=1}^m \|f_i^K(t_F - \cdot)\|_{L_2(0, t_F)} \|\mathbf{u}\|_{L_2(0, t_F)}^m$  so that (applying the CS inequality for  $\mathbb{R}^n$ )

$$\|E_{\mathcal{T}}^K[\mathbf{u}]\|_{L_2(\Omega)} \leq \left\{ \sum_{i=1}^m \|f_i^K(t_F - \cdot)\|_{L_2(0, t_F)}^2 \right\}^{1/2} \|\mathbf{u}\|_{L_2(0, t_F)}^m = \eta_{\mathcal{T}\mathbf{f}} \|\mathbf{u}\|_{L_2(0, t_F)}^m.$$

Proceeding along the same lines, it follows that

$$\begin{aligned} \|E_S^K[\mathbf{u}]\|_{L_2(\Omega)} &\leq \int_0^{t_F} \left\| \sum_{i=1}^m (S_A(t_F - \tau) - S_K(t_F - \tau)) [P_K[\boldsymbol{\beta}_{\mathbf{S}_i}]] u_i(\tau) \right\|_{L_2(\Omega)} d\tau \Rightarrow \\ \|E_S^K[\mathbf{u}]\|_{L_2(\Omega)} &\leq \sum_{i=1}^m \int_0^{t_F} \|(S_A(t_F - \tau) - S_K(t_F - \tau)) [P_K[\boldsymbol{\beta}_{\mathbf{S}_i}]] u_i(\tau)\|_{L_2(\Omega)} d\tau \Rightarrow \\ \|E_S^K[\mathbf{u}]\|_{L_2(\Omega)} &\leq \sum_{i=1}^m \int_0^{t_F} g_i^K(t_F - \tau) |u_i(\tau)| d\tau \end{aligned}$$

so that

$$\|E_S^K[\mathbf{u}]\|_{L_2(\Omega)} \leq \left\{ \sum_{i=1}^m \|g_i^K(t_F - \cdot)\|_{L_2(0,t_F)}^2 \right\}^{1/2} \|\mathbf{u}\|_{L_2(0,t_F)^m} = \eta_{\mathcal{T}_g}^K \|\mathbf{u}\|_{L_2(0,t_F)^m}.$$

□

**Proof of Proposition 2.2(b):** Note first that it follows from (CURTAIN; ZWART, 1995, Theorem 2.1.6, p. 18) that there exists  $\mu_A > 0$  and  $\sigma_A \in \mathbb{R}$  such that  $\forall t \geq 0$ ,  $\|S_A(t)\| \leq \mu_A e^{\sigma_A t}$ . Hence,

$$\begin{aligned} f_i^K(t_F - \tau) &\leq \mu_A e^{\sigma_A(t_F - \tau)} \|(I - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i}]\|_{L_2(\Omega)} \Rightarrow \\ \|f_i^K(t_F - \cdot)\|_{L_2(0,t_F)}^2 &= \int_0^{t_F} [f_i^K(t_F - \tau)]^2 d\tau \\ &\leq \|(I - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i}]\|_{L_2(\Omega)}^2 \int_0^{t_F} \{\mu_A e^{\sigma_A(t_F - \tau)}\}^2 d\tau \Rightarrow \\ \|f_i^K(t_F - \cdot)\|_{L_2(0,t_F)} &= \mu_A \|(I - P_K)[\boldsymbol{\beta}_{\mathbf{S}_i}]\|_{L_2(\Omega)} \|e^{\sigma_A(t_F - \cdot)}\|_{L_2(0,t_F)}. \end{aligned}$$

Thus it follows from the approximation property of  $\{S_K\}$ , *i.e.*, (2.20), that  $\|f_i^K(t_F - \cdot)\|_{L_2(0,t_F)} \rightarrow 0$  as  $K \rightarrow \infty$  and, hence,  $\eta_{\mathcal{T}_f}^K \rightarrow 0$  as  $K \rightarrow \infty$ .

With respect to  $\{\eta_{\mathcal{T}_g}^K\}$  note that under the ‘‘assumption’’ that  $\mathbf{B}[\phi, \psi]$  satisfies Garding’s inequality, see (EVANS, 2010, Theorem 6.2.2, p. 318), it follows from (2.6) and

(MORRIS, 1994, Theorem 5.2) that

$$\forall \theta \in L_2(\Omega), \quad \|S_K(t)[\theta] - S_A(t)[\theta]\|_{L_2(\Omega)}$$

converges uniformly on  $[0, t_F]$  to zero as  $K \rightarrow \infty$ . Hence,

$$\|g_i^K(t_F - \cdot)\|_{L_2(0, t_F)}^2 \rightarrow 0 \Rightarrow \eta_{Tg}^K \rightarrow 0 \text{ as } K \rightarrow \infty.$$

It then follows that  $\eta_T^K = \eta_{Tf}^K + \eta_{Tg}^K \rightarrow 0$  as  $K \rightarrow \infty$ . □

### 3 PEAK-VALUE CONSTRAINTS ON CONTROL SIGNALS AND ACTUATOR LOCATION

#### 3.1 Peak-value Constraints on Control Signals

In this chapter, the main concern is that upper bounds on the magnitudes of the control signals  $\mathbf{u}_i$  have to be imposed in connection with potential application to engineering problems. Thus, although setting the coefficient  $\rho_{\mathbf{u}}$  at different values may indirectly contribute to such an objective, it is natural to directly impose upper bound constraints on the optimal control problem at stake. Accordingly, a constrained optimization problem is formulated in (3.1) for which optimality conditions are then presented. Then a truncated version is introduced in (3.2) to generate approximate solutions to the original constrained problem. The latter is then tackled on the basis of the duality results in (3.3). To obtain approximate solutions to the dual problem, a class of piecewise-linear continuous Lagrange multipliers is introduced. The dual functional is explicitly written as a quadratic functional of the “free” parameters of this class of multipliers which are their values at a grid on  $[0, t_F]$ . To obtain approximate solutions to the dual problem is then reduced to maximizing this quadratic functional under non-negativeness constraints.

A summary is then provided of the computational steps required to obtain the desired control signals which satisfy the prescribed peak-value constraints.

Finally, actuator location is discussed in (3.14). Initially, a version of *Prob. I* with pointwise (with respect to  $t$ ) constraints is formulated as follows

$$\begin{aligned} \underline{\text{Prob. II}}: \quad & \min_{\mathbf{u} \in L_2(0, t_F)^m} \mathcal{J}(\mathbf{u}) \\ & \text{subject to: } \forall i = 1, \dots, m, \forall t \text{ a.e. in } [0, t_F], -\mu_i \leq \mathbf{u}_i(t) \leq \mu_i, \end{aligned} \quad (3.1)$$

where  $\mu_i \in \mathbb{R}_+$ .

The existence of an optimal solution to *Prob. II* can be ascertained by means of an argument entirely similar to the one used in connection with *Prob. I*. This leads to the next proposition.

**Proposition 3.1.** *Let  $\mathcal{I}_{F_i}(t) \triangleq [-\mu_i, \mu_i]$  and*

$$S_{\mathbf{u}F} \triangleq \{\mathbf{u} \in L_2(0, t_F)^m : \forall i = 1, \dots, m, \forall t \text{ a.e. in } [0, t_F], \mathbf{u}_i(t) \in \mathcal{I}_{F_i}(t)\}$$

*There exists  $\mathbf{u}_c \in S_{\mathbf{u}F}$  such that  $\forall \mathbf{u} \in S_{\mathbf{u}F}, \mathbf{u} \neq \mathbf{u}_c, \mathcal{J}(\mathbf{u}_c) < \mathcal{J}(\mathbf{u})$ .*

Moreover,  $\mathbf{u}_c$  satisfies the following optimality condition:

$\forall \delta \mathbf{u}$  such that  $(\mathbf{u}_c + \delta \mathbf{u}) \in S_{\mathbf{u}F}$ ,  $\langle \rho_{\mathbf{u}} \mathbf{u}_c + \mathcal{T}_{\theta}^*[\mathcal{T}_{\theta} - \theta_{ro}] , \delta \mathbf{u} \rangle_{L_2(0, t_F)^m} \geq 0$  or, equivalently,  
 $\forall \mathbf{u}$  such that  $(\mathbf{u}_c + \delta \mathbf{u}) \in S_{\mathbf{u}F}$ ,  $\forall t$  a.e. in  $[0, t_F]$ ,  $\{(\rho_{\mathbf{u}} \mathbf{u}_c + \mathcal{T}_{\theta}^*[\mathcal{T}_{\theta}[\mathbf{u}_c] - \theta_{ro}])^T \delta \mathbf{u}\}(t) \geq 0$ .

▽

Following (TRÖLTZSCH, 2010, p. 16), a simpler characterization of the optimal solution is now presented for which the following “saturation” operators are required: for  $i = 1, \dots, m$  define  $P_{Ii} : L_2(0, t_F) \rightarrow L_2(0, t_F)$

$$\begin{aligned} P_{Ii}[v](t) &= v(t) & \text{if } v(t) \in \mathcal{I}_{F_i}(t) \\ P_{Ii}[v](t) &= -\mu_i & \text{if } v(t) < -\mu_i \\ P_{Ii}[v](t) &= \mu_i & \text{if } v(t) > \mu_i \end{aligned}$$

**Proposition 3.2.** Let  $Z_a[\mathbf{u}] \triangleq \mathcal{T}_{\theta}^*[\mathcal{T}_{\theta}[\mathbf{u}] - \theta_{ro}]$ . For  $i = 1, \dots, m$   
 $\{\mathbf{u}_c\}_i = P_{Ii}[-(1/\rho_{\mathbf{u}})\{Z_a[\mathbf{u}_c]\}_i]$  a.e. in  $[0, t_F]$ .

▽

In the light of Proposition 3.2, the problem of computing (approximations to)  $\mathbf{u}_c$  is reduced to (approximately) solving a “system” of equations in  $L_2(0, t_F)$  with respect to  $\mathbf{u} \in L_2(0, t_F)^m$ , the  $i$  –  $th$  one of which is based on  $P_{Ii}$ . Following the approach pursued in connection with the unconstrained problem  $\mathcal{T}_{\theta}$  is going to be replaced by an approximation  $\mathcal{T}_{\theta}^K$ .

To this effect, let  $\mathcal{J}_K(\mathbf{u}) \triangleq \rho_{\mathbf{u}} \|\mathbf{u}\|_{L_2(0, t_F)^m}^2 + \|\mathcal{T}_{\theta}^K[\mathbf{u}] - \theta_{ro}\|_2^2$  and consider

$$\underline{\text{Prob. II}_K} : \min_{\mathbf{u} \in S_{\mathbf{u}F}} \mathcal{J}_K(\mathbf{u}). \quad (3.2)$$

Approximate solutions to *Prob. II* can be obtained on the basis of *Prob. II<sub>K</sub>*, as stated in the following proposition.

**Proposition 3.3. (a)**  $\forall K \in \mathbb{Z}_+$  there exists  $\mathbf{u}_c^K \in S_{\mathbf{u}F}$  such that  $\forall \mathbf{u} \in S_{\mathbf{u}F}$ ,  $\mathbf{u} \neq \mathbf{u}_c^K$ ,  $\mathcal{J}_K(\mathbf{u}_c^K) < \mathcal{J}_K(\mathbf{u})$ .

**(b)**  $\mathbf{u}_c^K \rightarrow \mathbf{u}_c$  in  $L_2(0, t_F)^m$ , as  $K \rightarrow \infty$ .

▽

The solution  $\mathbf{u}_c^K$  of *Prob. II<sub>K</sub>* can be characterized along the lines of Proposition 3.1, i.e.,

$$\{\mathbf{u}_c^K\}_i = P_{Ii}[-(1/\rho_{\mathbf{u}})\{Z_a^K[\mathbf{u}]\}_i] \quad (3.3)$$

where  $Z_a^K[\mathbf{u}] \triangleq (\mathcal{T}_{\theta}^K)^*[\mathcal{T}_{\theta}^K[\mathbf{u}] - \theta_{ro}]$ . However, computing  $\mathbf{u}_c^K$  on the basis of (3.3) is a difficult problem. Thus, in order to obtain approximations to  $\mathbf{u}_c^K$ , duality considerations pertaining to *Prob. II<sub>K</sub>* are now introduced.

To this effect, note first that  $\mathbf{u}_i(t) \in \mathcal{I}_{F_i}(t) \Leftrightarrow \mu_i - \mathbf{u}_i(t) \leq 0$  and  $\mathbf{u}_i(t) - \mu_i \leq 0$ , so that a Lagrangian functional for *Prob. II<sub>K</sub>* can be defined by

$$Lag_K(\mathbf{u}, \boldsymbol{\lambda}) \triangleq \mathcal{J}_K(\mathbf{u}) + 2\langle \boldsymbol{\lambda}_a, \mathbf{u}_a - \mathbf{u} \rangle_{L_2(0, t_F)^m} + 2\langle \boldsymbol{\lambda}_b, \mathbf{u} - \mathbf{u}_b \rangle_{L_2(0, t_F)^m}, \quad (3.4)$$

where  $\mathbf{u}_a = -\mu_i[1 \dots 1]^T$  and  $\mathbf{u}_b = \mu_i[1 \dots 1]^T$ ,  $\boldsymbol{\lambda}_a \in L_2(0, t_F)^m$ ,  $\boldsymbol{\lambda}_b \in L_2(0, t_F)^m$  and  $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b)$ .

The corresponding dual functional and dual problem are given by

$$\varphi_{DK}(\boldsymbol{\lambda}) \triangleq \min\{Lag_K(\mathbf{u}, \boldsymbol{\lambda}) : \mathbf{u} \in L_2(0, t_F)^m\} \quad (3.5)$$

and  $\underline{\text{Prob. II}}_{DK} : \max_{\boldsymbol{\lambda} \in S_\lambda} \varphi_{DK}(\boldsymbol{\lambda})$ ,

where

$$S_\lambda \triangleq \{(\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b) : \boldsymbol{\lambda}_a \in L_2(0, t_F)^m, \boldsymbol{\lambda}_b \in L_2(0, t_F)^m, \forall t \text{ a.e. in } [0, t_F], \boldsymbol{\lambda}_{ai}(t) \geq 0, \boldsymbol{\lambda}_{bi}(t) \geq 0\}.$$

The use of duality considerations in order to obtain  $\mathbf{u}_c^K$  consists of solving the dual problem *Prob. II*<sub>DK</sub> thereby obtaining  $\boldsymbol{\lambda}_K$  (say) and then compute the optimal solution of the unconstrained problem  $\min_{\mathbf{u}} Lag_K(\mathbf{u}; \boldsymbol{\lambda}_K)$ . This is stated in the following proposition which is a direct consequence of (LUENBERGER, 1969, p. 224).

**Proposition 3.4.** (a)  $\sup\{\varphi_{DK}(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in S_\lambda\} = \min\{\mathcal{J}_K(\mathbf{u}) : \mathbf{u} \in S_{u_F}\}$ .

(b) Let  $\mathbf{u}_c^K(\boldsymbol{\lambda})$  be the unique solution of  $\min_{\mathbf{u} \in L_2(0, t_F)^m} Lag_K(\mathbf{u}, \boldsymbol{\lambda})$ . Then  $\mathbf{u}_c^K = \mathbf{u}_c^K(\boldsymbol{\lambda}_K)$ , where  $\boldsymbol{\lambda}_K = \arg \max_{\boldsymbol{\lambda} \in S_\lambda} \varphi_{DK}(\boldsymbol{\lambda})$ .  $\nabla$

Note that for a given  $\boldsymbol{\lambda}$  the minimization of  $Lag_K(\mathbf{u}, \boldsymbol{\lambda})$  with respect to  $\mathbf{u} \in L_2(0, t_F)^m$  is a linear-quadratic problem similar to *Prob. I<sub>K</sub>* and can thus be solved in a similar way. The more difficult part in the use of duality to solve *Prob. II<sub>K</sub>* is to solve the dual problem. Thus, to rely on Proposition 3.4 to obtain approximate solutions to *Prob. II<sub>K</sub>* explicit characterizations of both  $\mathbf{u}_c^K(\boldsymbol{\lambda})$  and  $\varphi_{DK}(\boldsymbol{\lambda})$  are presented in the next proposition.

**Proposition 3.5.** For any  $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b) \in S_\lambda$

$$\mathbf{u}_c^K[\boldsymbol{\lambda}] = \mathbf{u}_K - \mathbf{F}_K(\rho_u^{-1}\mathbf{I} - (\rho_u\mathbf{I} + \mathbf{G}_K)^{-1})\bar{\boldsymbol{\alpha}}_\lambda^K + \rho_u^{-1}(\boldsymbol{\lambda}_a - \boldsymbol{\lambda}_b), \text{ and}$$

$$\varphi_{DK}(\boldsymbol{\lambda}) = \|\theta_{ro}\|_{L_2(\Omega)}^2 + \langle \mathcal{T}_\theta^K[\mathbf{u}_K], -\theta_{ro} \rangle + \hat{\varphi}_{DK}(\boldsymbol{\lambda}),$$

where

$$\hat{\varphi}_{DK}(\boldsymbol{\lambda}) \triangleq -\rho_u^{-1}\langle \boldsymbol{\lambda}_{ab}, \boldsymbol{\lambda}_{ab} \rangle + \rho_u^{-1}\langle \boldsymbol{\xi}_\lambda^K, (\rho_u\mathbf{I} + \mathbf{G}_K)^{-1}\boldsymbol{\xi}_\lambda^K \rangle - 2\langle \boldsymbol{\xi}_\lambda^K, \bar{\boldsymbol{\alpha}}_K \rangle + 2\langle \boldsymbol{\lambda}_a, \mathbf{u}_a \rangle - 2\langle \boldsymbol{\lambda}_b, \mathbf{u}_b \rangle,$$

$$\boldsymbol{\lambda}_{ab} = \boldsymbol{\lambda}_a - \boldsymbol{\lambda}_b, \quad \boldsymbol{\xi}_\lambda^K \triangleq \int_0^{t_F} \mathbf{F}_k^T(\tau)(\boldsymbol{\lambda}_a(\tau) - \boldsymbol{\lambda}_b(\tau))d\tau \quad \text{and} \quad \mathbf{G}_K \bar{\boldsymbol{\alpha}}_\lambda^K = \boldsymbol{\xi}_\lambda^K. \quad \nabla$$

**Remark 3.1.** Recall that  $\mathbf{u}_K = \mathbf{F}_K (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} \bar{\boldsymbol{\theta}}_{r_o}^K$  and note that (since  $\mathbf{M}(\mathbf{I} + \mathbf{M})^{-1} = \mathbf{I} - (\mathbf{I} + \mathbf{M})^{-1} = (\mathbf{I} + \mathbf{M})^{-1} \mathbf{M}$ )

$$(\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} = \rho_u^{-1} (\mathbf{I} + \rho_u^{-1} \mathbf{G}_K)^{-1} = \rho_u^{-1} \{ \mathbf{I} - (\mathbf{I} + \rho_u^{-1} \mathbf{G}_K)^{-1} \rho_u^{-1} \mathbf{G}_K \}$$

so that

$$\begin{aligned} \mathbf{u}_c^K[\boldsymbol{\lambda}] &= \mathbf{u}_K - \mathbf{F}_K \rho_u^{-1} (\mathbf{I} + \rho_u^{-1} \mathbf{G}_K)^{-1} \rho_u^{-1} \mathbf{G}_K \bar{\boldsymbol{\alpha}}_\lambda^K + \rho_u^{-1} (\boldsymbol{\lambda}_a - \boldsymbol{\lambda}_b) \\ \Leftrightarrow \mathbf{u}_c^K[\boldsymbol{\lambda}] &= \mathbf{u}_K - \mathbf{F}_K \rho_u^{-1} (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} \mathbf{G}_K \bar{\boldsymbol{\alpha}}_\lambda^K + \rho_u^{-1} (\boldsymbol{\lambda}_a - \boldsymbol{\lambda}_b) \\ \Leftrightarrow \mathbf{u}_c^K[\boldsymbol{\lambda}] &= \mathbf{F}_K (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} \{ \bar{\boldsymbol{\theta}}_{r_o}^K - \rho_u^{-1} \boldsymbol{\xi}_\lambda^K \} + \rho_u^{-1} (\boldsymbol{\lambda}_a - \boldsymbol{\lambda}_b) \\ \mathbf{u}_c^K[\boldsymbol{\lambda}] &= \mathbf{F}_K (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} \bar{\boldsymbol{\theta}}_r^K(\boldsymbol{\lambda}) + \rho_u^{-1} (\boldsymbol{\lambda}_a - \boldsymbol{\lambda}_b), \end{aligned}$$

where  $\bar{\boldsymbol{\theta}}_r^K(\boldsymbol{\lambda}) = \bar{\boldsymbol{\theta}}_{r_o}^K - \rho_u^{-1} \boldsymbol{\xi}_\lambda^K$ . It can thus be seen the optimal solution  $\mathbf{u}_c^K[\boldsymbol{\lambda}_K]$  of the constrained problem is obtained by adding a ‘‘correction term’’  $\rho_u^{-1} (\boldsymbol{\lambda}_a^K - \boldsymbol{\lambda}_b^K)$  to the output of a linear autonomous system, i.e.,  $\mathbf{u}_c^K[\boldsymbol{\lambda}](\tau) = (\mathbf{M}_\beta^K)^T \mathbf{x}_u^c(\tau)$  where  $\mathbf{x}_u^c$  is solution of the linear ordinary differential equation

$$\dot{\mathbf{x}}_u^c(\tau) = -\mathbf{A}_K^T \mathbf{x}_u^c(\tau), \quad \tau \geq 0 \quad \text{with initial condition} \quad \mathbf{x}_u^c(0) = (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} \bar{\boldsymbol{\theta}}_r^K(\boldsymbol{\lambda}_K).$$

▽

It follows from Proposition 3.5 that  $\boldsymbol{\lambda}_K = \arg \max_{\boldsymbol{\lambda} \in S_\lambda} \hat{\varphi}_{DK}(\boldsymbol{\lambda})$  is the solution of a quadratic problem in  $L_2(0, t_F)^m$  with non-negativeness constraints on the values  $\boldsymbol{\lambda}(t)$  (a.e. in  $[0, t_F]$ ). This suggests that approximate solutions  $\hat{\boldsymbol{\lambda}}_K$  for this problem should be sought on the basis of which the corresponding approximate solutions  $\mathbf{u}_c^K[\hat{\boldsymbol{\lambda}}_K]$  can be readily obtained in the light of Proposition 3.5.

The most difficult step in the duality approach described above is the computation of an approximate solution  $\hat{\boldsymbol{\lambda}}_K$  for the dual problem *Prob. D<sub>K</sub>*. To accomplish this task, piecewise-linear continuous classes of Lagrange multipliers are now introduced: let  $N_\lambda$  be a positive integer,  $\delta_t \triangleq t_F/N_\lambda$  and take  $N_\lambda$  subintervals  $\mathcal{I}_k$  of  $[0, t_F]$  where  $\mathcal{I}_k = [(k-1)\delta_t, k\delta_t]$ . Piecewise-linear multipliers are then defined by

$$\forall k = 1, \dots, N_\lambda, \quad \forall t \in \mathcal{I}_k, \quad \forall i = 1, \dots, m, \quad \lambda_i(t, \underline{\boldsymbol{\gamma}}_i) = \gamma_{ik} + (1/\delta_t) \Delta t_k (\gamma_{i(k+1)} - \gamma_{ik}),$$

where  $\underline{\gamma}_i \triangleq [\gamma_{i1} \dots \gamma_{i(N_\lambda+1)}]^\top$  and  $\Delta t_k \triangleq t - ((k-1)\delta_t)$  – note that  $\gamma_{ik}$  and  $\gamma_{i(k+1)}$  are the values of  $\lambda_i(\cdot; \underline{\gamma}_i)$  at the extreme points of  $\mathcal{I}_k$ . Thus,  $\forall t \in \mathcal{I}_k$ ,  $\lambda_i(\cdot; \underline{\gamma}_i) = \underline{\mathbf{h}}_k^\top \mathbf{E}_k \underline{\gamma}_i$  where  $\mathbf{E}_k^\top = [\mathbf{e}_k(N_\lambda + 1) : \mathbf{e}_{k+1}(N_\lambda + 1)]$  and  $\mathbf{e}_k(q)$  is the  $q$ th-vector in the canonical basis for  $\mathbb{R}^q$ ,  $\underline{\mathbf{h}}_k^\top \triangleq [1 - h_{kb}(t) : h_{kb}(t)]$  where  $h_{kb}(t) \triangleq (1/\delta_t)(t - (k-1)\delta_t)$ . As a result,  $\boldsymbol{\lambda}(t; \underline{\gamma}) = [\lambda_1(t; \underline{\gamma}_1) \dots \lambda_m(t; \underline{\gamma}_m)]^\top$  can be written as a linear function of  $\underline{\gamma} \triangleq [\underline{\gamma}_1^\top \dots \underline{\gamma}_m^\top]$  as

$$\forall k = 1, \dots, N_\lambda, \quad \forall t \in \mathcal{I}_k, \quad \boldsymbol{\lambda}(t; \underline{\gamma}) = \mathbf{E}_{\mathbf{h}k}(t) \underline{\gamma}, \quad (3.6)$$

where  $\mathbf{E}_{\mathbf{h}k}(t) \triangleq \text{diag}(\mathbf{h}_k^\top(t) \mathbf{E}_k, \dots, \mathbf{h}_k^\top(t) \mathbf{E}_k)$ .

It then follows that, for  $\boldsymbol{\lambda}_a$  and  $\boldsymbol{\lambda}_b$  piecewise-linear and defined by the parameters  $\underline{\gamma}_a$  and  $\underline{\gamma}_b$ , the dual functional can be written in terms of  $\underline{\gamma}_a$  and  $\underline{\gamma}_b$ , since (see Proposition 3.5)

$$\begin{aligned} \hat{\varphi}_{DK}(\boldsymbol{\lambda}) &= \sum_{k=1}^{N_\lambda} \int_{\mathcal{I}_k} \left\{ -\rho_{\mathbf{u}}^{-1} \underline{\gamma}_{ab}^\top [\mathbf{E}_{\mathbf{h}k}(t)]^\top \mathbf{E}_{\mathbf{h}k}(t) \underline{\gamma}_{ab} - 2\mu_{\mathbf{u}} \mathbf{1}_m^\top \mathbf{E}_{\mathbf{h}k}(t) \underline{\gamma}_a - 2\mu_{\mathbf{u}} \mathbf{1}_m^\top \mathbf{E}_{\mathbf{h}k}(t) \underline{\gamma}_b \right\} dt \\ &\quad + \rho_{\mathbf{u}}^{-1} \left\langle \boldsymbol{\xi}_\lambda^K(\underline{\gamma}_{ab}), (\rho_{\mathbf{u}} \mathbf{I} + \mathbf{G}_K)^{-1} \boldsymbol{\xi}_\lambda^K(\underline{\gamma}_{ab}) \right\rangle_E - 2\bar{\boldsymbol{\alpha}}_K^\top \boldsymbol{\xi}_\lambda^K(\underline{\gamma}_{ab}), \end{aligned}$$

or, equivalently,

$$\hat{\varphi}_{DK}(\underline{\gamma}_a, \underline{\gamma}_b) = -\rho_{\mathbf{u}}^{-1} \underline{\gamma}_{ab}^\top \mathbf{G}_{\mathbf{E}h} \underline{\gamma}_{ab} - 2\mu_{\mathbf{u}} \mathbf{1}_m^\top \bar{\mathbf{E}}_h (\underline{\gamma}_a + \underline{\gamma}_b) \quad (3.7)$$

$$(3.8)$$

$$+ \rho_{\mathbf{u}}^{-1} \boldsymbol{\xi}_\lambda^K(\underline{\gamma}_{ab})^\top (\rho_{\mathbf{u}} \mathbf{I} + \mathbf{G}_K)^{-1} \boldsymbol{\xi}_\lambda^K(\underline{\gamma}_{ab}) - 2\bar{\boldsymbol{\alpha}}^\top \boldsymbol{\xi}_\lambda^K(\underline{\gamma}_{ab}), \quad (3.9)$$

where

$$\underline{\gamma}_{ab} \triangleq \underline{\gamma}_a - \underline{\gamma}_b, \quad \boldsymbol{\xi}_\lambda^K(\underline{\gamma}_{ab}) \triangleq \left\{ \sum_{k=1}^{N_\lambda} \int_{\mathcal{I}_k} [\mathbf{F}_K(\tau)]^\top \mathbf{E}_{\mathbf{h}k}(\tau) d\tau \right\} \underline{\gamma}_{ab}, \quad \mathbf{1}_m = [1 \dots 1]^\top \in \mathbb{R}^m. \quad (3.10)$$

The quadratic functional  $\hat{\varphi}_{DK}$  (on  $(\underline{\gamma}_a, \underline{\gamma}_b)$ ) is then written as

$$\hat{\varphi}_{DK}(\underline{\gamma}_a, \underline{\gamma}_b) = -\rho_{\mathbf{u}}^{-1} \underline{\gamma}_{ab}^\top (\mathbf{G}_{\mathbf{E}h} + \mathbf{G}_{\mathbf{F}E}) \underline{\gamma}_{ab} - 2\bar{\boldsymbol{\alpha}}_K^\top \mathbf{F}_\xi^\top \underline{\gamma}_{ab} - 2\mu_{\mathbf{u}} \mathbf{1}_m^\top \bar{\mathbf{E}}_h (\underline{\gamma}_a + \underline{\gamma}_b), \quad (3.11)$$

where

$$\mathbf{G}_{\mathbf{E}h} \triangleq \sum_{k=1}^{N_\lambda} \left\{ \int_{\mathcal{I}_k} [\mathbf{E}_{\mathbf{h}k}(t)]^\top \mathbf{E}_{\mathbf{h}k}(t) dt \right\}, \quad \bar{\mathbf{E}}_h \triangleq \sum_{k=1}^{N_\lambda} \left\{ \int_{\mathcal{I}_k} \mathbf{1}_m^\top \mathbf{E}_{\mathbf{h}k}(t) dt \right\}, \quad (3.12)$$



$$\mathbf{F}_\xi^T \triangleq \sum_{k=1}^{N_\lambda} \left\{ \int_{\mathcal{I}_k} [\mathbf{F}_K(\tau)]^T \mathbf{E}_{hk}(\tau) d\tau \right\} \quad \text{and} \quad \mathbf{G}_{FE} \triangleq \mathbf{F}_\xi (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} \mathbf{F}_\xi^T. \quad (3.13)$$

To maximize the dual functional over this class of piecewise-linear multipliers amounts to solving the following optimization problem

$$\underline{\text{Prob. } D_{K\gamma}}: \quad \max_{\substack{\underline{\gamma}_a \in \mathbb{R}^{m(N_\lambda+1)}, \\ \underline{\gamma}_b \in \mathbb{R}^{m(N_\lambda+1)}}} \hat{\varphi}_{DK}(\underline{\gamma}_a, \underline{\gamma}_b) \quad \text{subject to: } \underline{\gamma}_a \geq 0, \quad \underline{\gamma}_b \geq 0$$

(note that  $\forall t \in [0, t_F], \quad \boldsymbol{\lambda}_i(t; \underline{\gamma}_i) \geq 0 \Leftrightarrow \underline{\gamma}_i \geq 0$ ).

**Remark 3.2.** *Prob.  $D_{K\gamma}$  is a finite-dimensional optimization problem with a quadratic cost functional and only non-negativity constraints on the decision variables which can be (numerically) solved in efficient ways.*  $\nabla$

Once a solution to *Prob.  $D_{K\gamma}$*  is obtained, say  $(\underline{\gamma}_a^o, \underline{\gamma}_b^o)$ , the approximate solution  $\hat{\boldsymbol{\lambda}}_K$  from the dual problem is given by (3.6) from which  $\mathbf{u}_c^K[\hat{\boldsymbol{\lambda}}_K]$  can be obtained from Proposition 3.5 and Remark 3.1.

**Remark 3.3.** *Although  $\mathbf{u}_c^K[\hat{\boldsymbol{\lambda}}_K]$  may fail to be in the feasible set  $S_{uF}$ , a closely-related feasible solution  $\mathbf{u}_K^R[\hat{\boldsymbol{\lambda}}_K]$  can also be obtained on the the basis of Proposition 3.5. To this effect, let  $\hat{\mathbf{u}}_c^K[\boldsymbol{\lambda}] = \mathbf{u}_K - \mathbf{F}_K(\mathbf{I} - (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1}) \bar{\boldsymbol{\alpha}}_\lambda^K$  so that  $\mathbf{u}_c^K[\boldsymbol{\lambda}] = \hat{\mathbf{u}}_c^K[\boldsymbol{\lambda}] + (\boldsymbol{\lambda}_a - \boldsymbol{\lambda}_b)$ ; define  $\mathbf{u}_K^R[\boldsymbol{\lambda}]$  by:*

$$\begin{aligned} \forall t \in [0, t_F] \quad \text{such that} \quad \{\hat{\mathbf{u}}_c^K[\boldsymbol{\lambda}(t)]\}_i &\in I_{F_i}(t), \quad \{\mathbf{u}_K^R[\boldsymbol{\lambda}(t)]\}_i = \{\hat{\mathbf{u}}_c^K[\boldsymbol{\lambda}(t)]\}_i, \\ \forall t \in [0, t_F] \quad \text{such that} \quad \{\hat{\mathbf{u}}_c^K[\boldsymbol{\lambda}(t)]\}_i &< -\mu_i, \quad \{\mathbf{u}_K^R[\boldsymbol{\lambda}(t)]\}_i = -\mu_i, \\ \forall t \in [0, t_F] \quad \text{such that} \quad \{\hat{\mathbf{u}}_c^K[\boldsymbol{\lambda}(t)]\}_i &> \mu_i, \quad \{\mathbf{u}_K^R[\boldsymbol{\lambda}(t)]\}_i = \mu_i. \end{aligned}$$

For any  $\boldsymbol{\lambda}$ ,  $\mathbf{u}_K^R[\boldsymbol{\lambda}] \in S_{uF}$ ; moreover, due to the so-called KKT optimality conditions for *Prob.  $II_K$* ,  $\mathbf{u}_K^R[\boldsymbol{\lambda}_K] = \mathbf{u}_c^K[\boldsymbol{\lambda}_K]$  (i.e., at  $\boldsymbol{\lambda} = \boldsymbol{\lambda}_K$ ,  $\mathbf{u}_K^R$  equals the optimal solution of *Prob.  $II_K$* ). In addition, given  $\hat{\boldsymbol{\lambda}}_K$ , the assessment of  $\mathbf{u}_K^R[\hat{\boldsymbol{\lambda}}_K]$  as an approximate solution to *Prob.  $II_K$*  can be carried out on the basis of the inequality  $\varphi_{DK}(\hat{\boldsymbol{\lambda}}_K) \leq \mathcal{J}_K(\mathbf{u}_c) \leq \mathcal{J}(\mathbf{u}_K^R[\hat{\boldsymbol{\lambda}}_K])$  so that whenever  $\varphi_{DK}(\hat{\boldsymbol{\lambda}}_K)$  and  $\mathcal{J}(\mathbf{u}_K^R[\hat{\boldsymbol{\lambda}}_K])$  are “close”  $\mathbf{u}_K^R[\hat{\boldsymbol{\lambda}}_K]$  can be regarded as an “approximate” solution to *Prob.  $II_K$* . Note also that  $\mathbf{u}_K^o$  is simply the output of an autonomous linear dynamic system ( $\hat{\mathbf{u}}_c^K[\boldsymbol{\lambda}_K]$ ) followed by a “saturation” operation which replaces its value by the prescribed limits  $-\mu_i$  and  $\mu_i$  whenever necessary. In Chapter 5 this approach is illustrated in two simple numerical examples.  $\nabla$

To conclude this chapter, a summary is presented of the steps required to compute the approximate solutions,  $\mathbf{u}_c^K[\boldsymbol{\lambda}_K]$  and  $\mathbf{u}_K^R[\boldsymbol{\lambda}_K]$  for the optimal control problem with peak-value constraints, namely

$$\min_{\mathbf{u} \in L_2(0, t_F)^m} \mathcal{J}_K(\mathbf{u}) \quad \text{subject to: } \forall i = 1, \dots, m, \quad -\mu_i \leq \mathbf{u}_i(t) \leq \mu_i \quad \forall t \text{ a.e. in } [0, t_F],$$

where  $\mathcal{J}(\mathbf{u})$  is given by (2.11).

Given the problem data  $(f, g, \theta_r, \rho_u, \{\mu_i : i = 1, \dots, m\})$ ,  $S_K$  and  $\{\phi_1, \dots, \phi_{n(K)}\}$  (as before)  $N_\lambda$  and subintervals  $\mathcal{I}_k = [(k-1)\delta_t, k\delta_t]$ ,  $k = 1, \dots, N_\lambda$  and  $\delta_t = t_F/N_\lambda$ .

- (1) Compute  $\bar{\boldsymbol{\theta}}_{ro}^K$ ,  $\mathbf{M}_\beta^K$ ,  $\mathbf{A}_K$ ,  $\mathbf{G}_K$  and  $(\mathbf{I} + \rho_u^{-1}\mathbf{G}_K)^{-1}$  as indicated in the summary of the computational steps required in the unconstrained problem (see final part of Chapter 2).
- (2) Compute  $\bar{\boldsymbol{\alpha}}_K = (\mathbf{I} + \rho_u^{-1}\mathbf{G}_K)^{-1}\rho_u^{-1}\bar{\boldsymbol{\theta}}_{ro}^K$ , compute  $\mathbf{G}_{Eh}$ ,  $\bar{\mathbf{E}}_h$  and  $\mathbf{F}_\xi$  given by (3.12) – (3.13) numerically solving the corresponding integrals over the sub-intervals  $\mathcal{I}_k$  and compute  $\mathbf{G}_{FE} = \mathbf{F}_\xi(\mathbf{I} + \rho_u^{-1}\mathbf{G}_K)^{-1}\rho_u^{-1}\mathbf{F}_\xi^T$ .
- (3) Compute an approximate solution  $(\underline{\boldsymbol{\gamma}}_a^K, \underline{\boldsymbol{\gamma}}_b^K)$  to *Prob. D<sub>K</sub>γ*.
- (4) Putting  $\boldsymbol{\lambda}_a^K(t) = \mathbf{E}_{hk}(t)\underline{\boldsymbol{\gamma}}_a^K$ ,  $\boldsymbol{\lambda}_b^K(t) = \mathbf{E}_{hk}(t)\underline{\boldsymbol{\gamma}}_b^K$  and  $\boldsymbol{\xi}_\lambda^K = \int_0^{t_F} \mathbf{F}_K(\tau)^T [\boldsymbol{\lambda}_a^K(\tau) - \boldsymbol{\lambda}_b^K(\tau)] d\tau$  and  $\hat{\mathbf{u}}_c^K[\boldsymbol{\lambda}_K]$  is obtained from  $\hat{\mathbf{u}}_c^K[\boldsymbol{\lambda}_K] = \mathbf{F}_K(\mathbf{I} + \rho_u^{-1}\mathbf{G}_K)^{-1}\rho_u^{-1}\bar{\boldsymbol{\theta}}_{r\lambda}^K$ , where  $\bar{\boldsymbol{\theta}}_{r\lambda}^K = \bar{\boldsymbol{\theta}}_{ro}^K - \rho_u^{-1}\boldsymbol{\xi}_\lambda^K$ .
- (5)  $\mathbf{u}_c^K[\boldsymbol{\lambda}_K]$  and  $\mathbf{u}_K^R[\boldsymbol{\lambda}_K]$  are obtained as:  $\mathbf{u}_c^K[\boldsymbol{\lambda}_K] = \hat{\mathbf{u}}_c^K[\boldsymbol{\lambda}_K] + \rho_u^{-1}(\boldsymbol{\lambda}_a^K - \boldsymbol{\lambda}_b^K)$  and  $\mathbf{u}_K^R[\boldsymbol{\lambda}_K]$  as in Remark 3.3.

**Remark 3.4.** *The major computations task in steps (1) – (5) above corresponds to numerically obtaining a solution to Prob.  $D_{K\gamma}$ . As the number of decision variables in it  $2m(N_\lambda + 1)$  may be relatively large a change of variables was considered to make possible calculations simpler and less susceptible to numerical problems. This goes as follows: first  $(\underline{\gamma}_a, \underline{\gamma}_b)$  are replaced by  $(\underline{\sigma}_a, \underline{\sigma}_b)$  where*

$$\begin{bmatrix} \underline{\sigma}_a \\ \underline{\sigma}_b \end{bmatrix} = \begin{bmatrix} \underline{\gamma}_a - \underline{\gamma}_b \\ \underline{\gamma}_a + \underline{\gamma}_b \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} \\ \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \underline{\gamma}_a \\ \underline{\gamma}_b \end{bmatrix} \quad \left( \Leftrightarrow \begin{bmatrix} \underline{\gamma}_a \\ \underline{\gamma}_b \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \mathbf{I} & \mathbf{I} \\ -\mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \underline{\sigma}_a \\ \underline{\sigma}_b \end{bmatrix} \right)$$

so that  $\hat{\varphi}_{DK}(\cdot, \cdot)$  in (3.11) can be written in terms of  $\underline{\sigma}_a$  and  $\underline{\sigma}_b$  as

$$\hat{\varphi}_{DK}(\underline{\sigma}_a, \underline{\sigma}_b) = -\rho_{\mathbf{u}}^{-1} \underline{\sigma}_a^T \mathbf{G}_{FH} \underline{\sigma}_a - 2\bar{\alpha}_K^T \mathbf{F}_\xi \underline{\sigma}_a - 2\mu_{\mathbf{u}} \mathbf{1}_m \bar{\mathbf{E}}_h \underline{\sigma}_b.$$

Then (exploiting the fact that  $\mathbf{G}_{FH}$  is square and symmetric) an eigenvector decomposition of  $\mathbf{G}_{FH}$  (where  $\mathbf{G}_{FH} = \mathbf{G}_{Eh} + \mathbf{G}_{FE}$ ) is obtained, i.e.,  $\mathbf{G}_{FH} = \mathbf{V}_{Fh} \boldsymbol{\Lambda}_{Fh} \mathbf{V}_{Fh}^T$  and a new decision variable is defined, namely  $\check{\underline{\sigma}}_a = \mathbf{V}_{Fh}^T \underline{\sigma}_a$  ( $\Leftrightarrow \underline{\sigma}_a = \mathbf{V}_{Fh} \check{\underline{\sigma}}_a$ ). The dual functional is then written in terms of  $\check{\underline{\sigma}}_a$  and  $\underline{\sigma}_b$  as

$$\hat{\varphi}_{DK}(\check{\underline{\sigma}}_a, \underline{\sigma}_b) = -\rho_{\mathbf{u}}^{-1} \check{\underline{\sigma}}_a^T \boldsymbol{\Lambda}_{Fh} \check{\underline{\sigma}}_a - 2\bar{\alpha}_K^T (\mathbf{F}_\xi^T \mathbf{V}_{Fh}) \check{\underline{\sigma}}_a - 2\mu_{\mathbf{u}} \mathbf{1}_m \bar{\mathbf{E}}_h \underline{\sigma}_b$$

so that the matrix defining the quadratic form above is diagonal. The dual problem to be solved is then given by

$$\max_{\check{\underline{\sigma}}_a, \underline{\sigma}_b} \hat{\varphi}_{DK}(\check{\underline{\sigma}}_a, \underline{\sigma}_b) \quad \text{subject to:} \quad \begin{bmatrix} \mathbf{V}_{Fh} & \mathbf{I} \\ -\mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \check{\underline{\sigma}}_a \\ \underline{\sigma}_b \end{bmatrix} \geq 0.$$

▽

### 3.2 Actuator Location

Recall that the control signals  $\mathbf{u}_i : [0, t_F] \rightarrow \mathbb{R}$  appear in the “source” terms  $\beta_{S_i}(\mathbf{x}) \mathbf{u}_i(t)$  of the heat equation (2.1). Thus, the spatial effect of the control signals  $\{\mathbf{u}_i\}$  depend on the functions  $\{\beta_{S_i}\}$  which may be viewed as spreading over the spatial domain the action of the control signals (which, in turn, depend solely on  $t$ ). As a result, it is often the case that the spatial effect of the control signals  $\mathbf{u}_i$ ,  $i = 1, \dots, m$ , have a local character due to the functions  $\beta_{S_i}$  only having non-zero value on “small” subsets of the spatial domain  $\Omega$ . In such cases, the “location” of each  $\mathbf{u}_i$  (i.e., the “centre”

of the support of  $\beta_{\mathcal{S}_i}$ ) may have significant effects on the magnitude of the final-state approximation error attained with the optimal  $\mathbf{u}$ .

The case will now be considered of each  $\beta_{\mathcal{S}_i}$  being a spatially-displaced version of a simple function  $\beta_{\mathcal{S}}$  which models the way the control action is spatially distributed. More specifically, assume that  $\Omega$  is symmetric with respect to  $x_{\mathbf{a}} \in \Omega$ , *i.e.*,  $\forall x = x_{\mathbf{a}} + (x - x_{\mathbf{a}}) \in \Omega$ ,  $x_{\mathbf{a}} - (x - x_{\mathbf{a}}) \in \Omega$  and let  $\Omega_{\beta} \subset \Omega$  be an open and connected set also centred on  $x_{\mathbf{a}}$ . Let  $\beta_{\mathbf{a}} : \Omega \rightarrow \mathbb{R}$  be such that  $\forall x \in \Omega - \Omega_{\beta}$ ,  $\beta_{\mathbf{a}}(x) = 0$  (*i.e.*,  $\Omega_{\beta}$  is the support of  $\beta_{\mathbf{a}}$ ) and for a list  $\underline{\mathcal{X}}$  of locations  $\mathcal{X}_i$ ,  $\underline{\mathcal{X}} = (\mathcal{X}_1, \dots, \mathcal{X}_m)$ ,  $\mathcal{X}_i \in \Omega$  and such that  $\Omega_{\beta} + (\mathcal{X}_i - x_{\mathbf{a}}) \subset \Omega$ , define  $\beta_{\mathcal{S}_i}(\cdot; \mathcal{X}_i) : \Omega \rightarrow \mathbb{R}$  by  $\forall x \in \Omega$ ,  $\beta_{\mathcal{S}_i}(x; \mathcal{X}_i) \triangleq \beta_{\mathbf{a}}(x - (\mathcal{X}_i - x_{\mathbf{a}}))$  – note that  $\Omega_{\beta} + (\mathcal{X}_i - x_{\mathbf{a}})$  is the support of  $\beta_{\mathcal{S}_i}(\cdot; \mathcal{X}_i)$ .

Recall that the approximation error magnitude is given by

$$\|\mathcal{T}_{\theta}^K[\mathbf{u}_K] - \theta_{ro}^K\|_2 = \|\underline{\mathbf{c}}_K(t_F; \mathbf{u}_K) - \bar{\theta}_{ro}^K\|_2 \text{ where } \mathbf{u}_K = \mathbf{F}_K \bar{\alpha}_K, \underline{\mathbf{c}}_K(t_F; \mathbf{u}_K) = \mathbf{G}_K \bar{\alpha}_K \text{ and } \bar{\alpha}_K = (\rho_{\mathbf{u}} \mathbf{I} + \mathbf{G}_K)^{-1} \bar{\theta}_{ro}^K.$$

$$\text{Thus, } \|\mathcal{T}_{\theta}^K[\mathbf{u}_K] - \bar{\theta}_{ro}^K\|_2 = \|\{\mathbf{G}_K(\rho_{\mathbf{u}} \mathbf{I} + \mathbf{G}_K)^{-1} - \mathbf{I}\} \bar{\theta}_{ro}^K\|_2 = \|(\mathbf{I} + \rho_{\mathbf{u}}^{-1} \mathbf{G}_K)^{-1} \bar{\theta}_{ro}^K\|_2, \text{ since } \mathbf{G}_K(\rho_{\mathbf{u}} \mathbf{I} + \mathbf{G}_K)^{-1} = \rho_{\mathbf{u}}^{-1} \mathbf{G}_K(\mathbf{I} + \rho_{\mathbf{u}}^{-1} \mathbf{G}_K)^{-1} = \mathbf{I} - (\mathbf{I} + \rho_{\mathbf{u}}^{-1} \mathbf{G}_K)^{-1}.$$

Thus, to choose actuator locations with the purpose of obtaining a good final-state approximation, a natural formulation for the actuator location problem would be:

$$\text{Prob. Loc.: } \min_{\substack{\underline{\mathcal{X}} = (\mathcal{X}_1, \dots, \mathcal{X}_m), \\ \Omega_{\beta} + (\mathcal{X}_i - x_{\mathbf{a}}) \subset \Omega}} \nu(\underline{\mathcal{X}}), \quad (3.14)$$

where  $\nu(\underline{\mathcal{X}}) \triangleq \|\{\mathbf{I} + \rho_{\mathbf{u}}^{-1} \mathbf{G}_K(\mathbf{M}_{\beta}^K(\underline{\mathcal{X}}))\}^{-1} \bar{\theta}_{ro}^K\|_2^2$ ,  $\mathbf{G}_K(\mathbf{M}) \triangleq \int_0^{t_F} \exp[\mathbf{A}_K t] \mathbf{M} \mathbf{M}^T \exp[\mathbf{A}_K^T t] dt$ ,

$$\mathbf{M}_{\beta}^K(\underline{\mathcal{X}}) = \begin{bmatrix} \langle \beta_{\mathcal{S}_1}(\mathcal{X}_1), \phi_1 \rangle & \cdots & \langle \beta_{\mathcal{S}_m}(\mathcal{X}_m), \phi_1 \rangle \\ \vdots & & \vdots \\ \langle \beta_{\mathcal{S}_1}(\mathcal{X}_1), \phi_K \rangle & \cdots & \langle \beta_{\mathcal{S}_m}(\mathcal{X}_m), \phi_K \rangle \end{bmatrix}.$$

**Remark 3.5.** *The problem formulation above hinges upon the approximation error attained with the optimal, unconstrained control signal  $\mathbf{u}_k$ . It is also natural to focus on the constrained optimal control signal  $\mathbf{u}_K^c$ , in which case  $\nu(\cdot)$  would be replaced by  $\nu_c(\cdot)$  in the formulation of Prob. Loc. by  $\nu_c(\underline{\mathcal{X}}) = \|\underline{\mathbf{c}}_K(t_F; \mathbf{u}_K^c) - \bar{\theta}_{ro}^K\|_2$ .  $\nabla$*

**Remark 3.6.** *Those two choices of cost functional for the actuator location problem are “tuned” to a given final-state target  $\bar{\theta}_{ro}^K$ . Alternatively, if any final-state in a “broad” class may be targeted with the same actuator-location arrangement, a natural choice for the cost-functional of Prob. Loc. would be  $\nu_s(\underline{\mathcal{X}}) \triangleq \|(\mathbf{I} + \rho_{\mathbf{u}}^{-1} \mathbf{G}_K(\mathbf{M}_{\beta}^K(\underline{\mathcal{X}})))^{-1}\|_s$ . This would be relevant for both  $\mathbf{u}_K$  and  $\mathbf{u}_K^c$  for, in the case of  $\mathbf{u}_K$ , it yields an upper bound on  $\nu(\underline{\mathcal{X}})$  for*

any  $\bar{\theta}_{r_o}^K$  with euclidean norm smaller or equal to a pre-specified value; whereas, in the case of  $\mathbf{u}_K^c$ , as  $\mathbf{u}_K(\tau) = \mathbf{F}_K(\tau)\rho_{\mathbf{u}}^{-1}(\mathbf{I} + \rho_{\mathbf{u}}^{-1}\mathbf{G}_K(\mathbf{M}_{\beta}^K(\underline{\mathcal{X}})))^{-1}\bar{\theta}_{r_o}^K$ , making  $\nu_s(\underline{\mathcal{X}})$  “small” tends to make the values of  $\mathbf{u}(\cdot)$  smaller thereby mitigating the increase in the approximation error magnitude due to the enforcement of peak-value constraints.  $\nabla$

The possible effect of actuator locations on the controlled final state is illustrated in Figures 13, 14 and 15 of Chapter 4 for the case of the one-dimensional heat equation with one scalar control signal (*i.e.*,  $\mathbf{u}(t) \in \mathbb{R}$ ). Three locations are considered: a central one and two others symmetrically situated with respect to the centre of  $\Omega = (0, L_x)$  (*i.e.*,  $x = L_x/2$ ) and close to the boundary  $\partial\Omega$ . In this case, with the desired final state also symmetric with respect to  $x = L_x/2$  and for the approximating subspaces  $S_K = \text{span}\{\sqrt{2/L}\sin((\pi/L_x)x), \dots, \sqrt{2/L}\sin((K\pi/L_x)x)\}$ , it can be shown that  $\mathcal{X}_0 = L_x/2$  is a local extremum for  $\nu(\cdot)$ . It can be observed that the central location yields significantly better approximations for the desired final state than those provided by the two other locations taken into account – this is the case for both  $\mathbf{u}_K(\mathcal{X})$  and  $\mathbf{u}_c^K(\mathcal{X})$ .

In general, solving *Prob. Loc.* (even for the cost-functional  $\nu(\cdot)$ ) is a difficult task as global optimization techniques are required to obtain a solution on  $\Omega^m$  and  $\nu(\cdot)$  depends on  $\underline{\mathcal{X}}$  in an intricate manner (through the inverse of  $(\mathbf{I} + \rho_{\mathbf{u}}^{-1}\mathbf{G}_K(\mathbf{M}_{\beta}^K(\underline{\mathcal{X}})))$  with  $\mathbf{G}_K(\mathbf{M})$  depending on  $\mathbf{M}\mathbf{M}^T$  and  $\{\mathbf{M}_{\beta}^K(\underline{\mathcal{X}})\}_{\ell k} = \langle \boldsymbol{\beta}_{\mathcal{S}_{\ell}}(\mathcal{X}_{\ell}), \phi_k \rangle$ ). Although a grid search would seem feasible in the physically motivated cases of  $n$ -dimensional spatial domains with  $n = 1, 2, 3$ , it is noted that with  $N_g$  points along each dimension, the number of possible actuator locations arrangement would be  $(N_g^n)^m$ .

To perform a less demanding search, optimization objectives may be weakened so that a randomly-generated sample of possible actuator-location arrangements is examined with the sample size being specified on the basis of probabilistic considerations – this approach has attracted considerable attention in the control literature, see (TEMPO; ISHII, 2007 and references therein). In this case, a number of actuator locations  $\{\underline{\mathcal{X}}_i, i = 1, \dots, N_S\}$  would be randomly generated and a location  $\underline{\mathcal{X}}_o$  would be chosen so that  $\underline{\mathcal{X}}_o$  minimizes  $\nu$  (or  $\nu_c$ ) on the sample  $\{\underline{\mathcal{X}}_i\}, i = 1, \dots, N_S$ . The only “parameter” to be chosen in this approach is the sample size  $N_S$ , whose value is determined on the basis of probabilistic considerations – roughly speaking requiring that with a “high probability” the chosen location  $\underline{\mathcal{X}}_o$  is better than “most” possible ones. The sample size calculations of interest here are presented below.

### 3.3 Sample Size for Random Search

Let  $x$  be a continuous,  $n$ -dimensional random variable with probability density function (pdf)  $p_x$  the support of which is denoted by  $S_x \subset \mathbb{R}^n$ . Let  $f : S_x \rightarrow \mathbb{R}_+^c$  be continuous and such that  $\forall w \in f(S_x)$  the set  $\{x \in S_x : f(x) = w\}$  has zero Lebesgue measure. Let  $f_*$  be defined as  $f_* = \inf\{f(x) : x \in S_x\}$ . For a given  $\varepsilon \in (0, 1)$  define  $\delta_\varepsilon > 0$  by  $Pr\{x \in S_x : f(x) \geq f_* + \delta_\varepsilon\} = 1 - \varepsilon$ . Note that  $\delta_\varepsilon \rightarrow 0$  as  $\varepsilon \rightarrow 0$ .

Let  $\{x_i : i = 1, \dots, N\}$  be (a sample of) independent and identically distributed random variables with pdf  $p_x$  and define  $f_*^N \triangleq \min\{f(x_i) : i = 1, \dots, N\}$ . For a given  $\alpha \in (0, 1)$ ,  $N$  is to be chosen so that

$$Pr\{f_*^N < f_* + \delta_\varepsilon\} \geq 1 - \alpha. \quad (3.15)$$

To this effect, note that

$$\begin{aligned} Pr\{f_*^N < f_* + \delta_\varepsilon\} &= 1 - Pr\{f_*^N \geq f_* + \delta_\varepsilon\} = 1 - Pr\left\{\bigcap_{i=1}^N \{x_i \in S_x : f(x_i) \geq f_* + \delta_\varepsilon\}\right\} \Leftrightarrow \\ &= 1 - \prod_{i=1}^N Pr\{x_i \in S_x : f(x_i) \geq f_* + \delta_\varepsilon\} \\ &= 1 - \{Pr\{x \in S_x : f(x) \geq f_* + \delta_\varepsilon\}\}^N \end{aligned}$$

$$\Leftrightarrow Pr\{f_*^N < f_* + \delta_\varepsilon\} = 1 - (1 - \varepsilon)^N.$$

Thus, (3.15) holds if and only if

$$\begin{aligned} 1 - (1 - \varepsilon)^N \geq 1 - \alpha &\Leftrightarrow \alpha \geq (1 - \varepsilon)^N \Leftrightarrow \log \alpha \geq N \log(1 - \varepsilon) \\ \Leftrightarrow N \geq N_{\alpha\varepsilon} \triangleq \log \alpha / \log(1 - \varepsilon) &= \frac{\log(1/\alpha)}{\log(1/(1 - \varepsilon))}. \end{aligned}$$

Thus, roughly speaking, in the case of a uniform pdf on  $S_x$ , for  $N \geq N_{\alpha\varepsilon}$  the probability that  $f_*^N$  is smaller than “the values of  $f(x)$  on  $(1 - \varepsilon) \times 100\%$  of  $S_x$ ” is greater than  $(1 - \alpha)$ .

In Section 4.3, an example is presented to illustrate the potential of such a random search to choose the locations of two “actuators” in connection with the heat equation on a two-dimensional spatial domain.

### 3.4 APPENDIX – PROOFS FROM CHAPTER 3

**Proof of Proposition 3.1:** Once it is established that  $S_{\mathbf{u}_F}$  is convex and closed, the argument employed in the proof of Proposition 2.1 also proves Proposition 3.1.

To show that  $S_{\mathbf{u}_F}$  is convex let  $\mathbf{u}_j \in S_{\mathbf{u}_F}$ ,  $j = 1, 2$  and define  $\mathbf{u}(t; \sigma) = \sigma \mathbf{u}_1(t) + (1 - \sigma) \mathbf{u}_2(t)$ ,  $\sigma \in [0, 1]$ . Then  $\forall i = 1, \dots, m$ ,  $\forall t \in [0, t_F]$  a.e.  $u_i(t, \sigma) = \sigma u_{1i}(t) + (1 - \sigma) u_{2i}(t) \in \mathcal{I}_{F_i}(t)$  (since  $u_{1i}(t) \in \mathcal{I}_{F_i}(t)$ ,  $u_{2i}(t) \in \mathcal{I}_{F_i}(t)$  and  $\mathcal{I}_{F_i}(t)$  is an interval).

To show that  $S_{\mathbf{u}_F}$  is closed, let  $\mathbf{u}^\ell \in S_{\mathbf{u}_F}$  be such that  $\mathbf{u}^\ell \rightarrow \mathbf{u}$  in the sense of the  $L_2(0, t_F)^m$ -norm. Then  $\forall i = 1, \dots, m$ ,  $|u_i - u_i^\ell| \rightarrow 0$  and, hence,

$$\forall t \text{ a.e. in } [0, t_F], |u_i(t) - u_i^\ell(t)| \rightarrow 0. \quad (3.A.1)$$

Now  $\forall i, \forall \ell$ ,  $\mathbf{u}^\ell \in S_{\mathbf{u}_F} \Rightarrow \forall t \text{ a.e. in } [0, t_F]$ ,

$$\begin{aligned} |u_i^\ell(t)| &\leq \mu_i \text{ and} \\ |u_i(t) - u_i^\ell(t)| &\geq |u_i(t)| - |u_i^\ell(t)|. \end{aligned}$$

Thus  $\forall t \text{ a.e. in } [0, t_F]$ ,  $|u_i(t) - u_i^\ell(t)| \geq |u_i(t)| - \mu_i \Rightarrow \forall i = 1, \dots, m$ ,  $\forall \ell \in \mathbb{Z}_+$

$$|u_i(t)| \leq \mu_i + |u_i(t) - u_i^\ell(t)|.$$

Thus in the light of (3.A.1),  $\forall t \text{ a.e. in } [0, t_F]$ ,  $\forall i = 1, \dots, m$ ,  $|u_i(t)| \leq \mu_i \Rightarrow \mathbf{u} \in S_{\mathbf{u}_F}$ .

With respect to the optimality condition, note that

$$\mathcal{J}(\mathbf{u} + \Delta \mathbf{u}) = \mathcal{J}(\mathbf{u}) + 2\rho_{\mathbf{u}} \langle \mathbf{u}, \Delta \mathbf{u} \rangle + \rho_{\mathbf{u}} \|\Delta \mathbf{u}\|_{L_2(0, t_F)^m}^2 + 2\langle \mathcal{T}_\theta[\mathbf{u}] - \boldsymbol{\theta}_{ro}, \mathcal{T}_\theta[\Delta \mathbf{u}] \rangle + \|\mathcal{T}_\theta[\Delta \mathbf{u}]\|_2^2$$

$$\iff \mathcal{J}(\mathbf{u} + \Delta \mathbf{u}) = \mathcal{J}(\mathbf{u}) + 2\langle \rho_{\mathbf{u}} \mathbf{u} + Z_a[\mathbf{u}], \Delta \mathbf{u} \rangle + (\rho_{\mathbf{u}} \|\Delta \mathbf{u}\|_2^2 + \|\mathcal{T}_\theta[\Delta \mathbf{u}]\|_{L_2(0, t_F)^m}^2),$$

where  $Z_a[\mathbf{u}] \triangleq \mathcal{T}_\theta^*[\mathcal{T}_\theta[\mathbf{u}] - \boldsymbol{\theta}_{ro}]$ .

Thus  $\mathbf{u}_c \in S_{\mathbf{u}_F}$  is optimal if and only if  $\forall \Delta \mathbf{u} \in L_2(0, t_F)^m$  such that  $(\mathbf{u}_c + \Delta \mathbf{u}) \in S_{\mathbf{u}_F}$ ,  $\langle \rho_{\mathbf{u}} \mathbf{u}_c + Z_a[\mathbf{u}_c], \Delta \mathbf{u} \rangle \geq 0$ .

Note now that since

$$\langle \rho_{\mathbf{u}} \mathbf{u} + Z_a[\mathbf{u}], \Delta \mathbf{u} \rangle = \int_0^{t_F} (\rho_{\mathbf{u}} \mathbf{u}(t) + Z_a[\mathbf{u}](t))^T \Delta \mathbf{u}(t) dt ,$$

the condition

“ $\forall \Delta \mathbf{u}$  such that  $(\mathbf{u}_c + \Delta \mathbf{u}) \in S_{\mathbf{u}F}$ ,  $\forall t$  a.e. in  $[0, t_F]$ ,  $(\rho_{\mathbf{u}} \mathbf{u}_c(t) + Z_a[\mathbf{u}_c](t))^T \Delta \mathbf{u}(t) \geq 0$ ” is sufficient for  $\mathbf{u}_c$  to be optimal. To see that it is also necessary, suppose that there exists  $\Delta \mathbf{u} \in L_2(0, t_F)^m$  such that  $(\mathbf{u}_c + \Delta \mathbf{u}) \in S_{\mathbf{u}F}$  and for some subset  $S_a$  of  $[0, t_F]$  with non-zero measure,  $(\rho_{\mathbf{u}} \mathbf{u}_c(t) + Z_a[\mathbf{u}_c](t))^T \Delta \mathbf{u}(t) < 0$  for any  $t \in S_a$ . Then, defining  $\hat{\Delta} \mathbf{u}(t) = \Delta \mathbf{u}(t)$  for  $t \in S_a$  and  $\hat{\Delta} \mathbf{u}(t) = 0$  otherwise,

$$(\mathbf{u}_c + \hat{\Delta} \mathbf{u}) \in S_{\mathbf{u}F} \text{ and } \langle \rho_{\mathbf{u}} \mathbf{u}_c + Z_a[\mathbf{u}_c], \hat{\Delta} \mathbf{u} \rangle = \int_{S_a} (\rho_{\mathbf{u}} \mathbf{u}_c(t) + Z_a[\mathbf{u}_c](t))^T \Delta \mathbf{u}(t) dt < 0$$

so that  $\mathbf{u}_c$  cannot be optimal. □

**Proof of Proposition 3.2:** Consider the following optimization problem for  $t \in [0, t_F]$  :

$$\min_{\mathbf{v} \in \mathbb{R}^m} \|\rho_{\mathbf{u}} \mathbf{v} + Z_a[\mathbf{u}_c](t)\|_2^2 \text{ subject to } \forall i = 1, \dots, m \ v_i \in \mathcal{I}_{F_i}(t).$$

As  $\|\rho_{\mathbf{u}}(\mathbf{v} + \Delta \mathbf{v}) + Z_a[\mathbf{u}_c](t)\|_2^2 = \|\rho_{\mathbf{u}} \mathbf{v} + Z_a[\mathbf{u}_c](t)\|_2^2 + 2\langle \rho_{\mathbf{u}} \mathbf{v} + Z_a[\mathbf{u}_c](t), \rho_{\mathbf{u}} \Delta \mathbf{v} \rangle + \|\rho_{\mathbf{u}} \Delta \mathbf{v}\|_2^2$   $\mathbf{v}_t$  is optimal if and only if  $\mathbf{v}_{t_i} \in \mathcal{I}_{F_i}(t)$  and  $\forall \Delta \mathbf{v}$  such that  $\mathbf{v}_{t_i} + \Delta \mathbf{v}_i \in \mathcal{I}_{F_i}(t)$

$$\langle \rho_{\mathbf{u}} \mathbf{v}_t + Z_a[\mathbf{u}_c](t), \rho_{\mathbf{u}} \Delta \mathbf{v} \rangle \geq 0 \Leftrightarrow \langle \rho_{\mathbf{u}} \mathbf{v}_t + Z_a[\mathbf{u}_c](t), \Delta \mathbf{v} \rangle \geq 0. \quad (3.A.2)$$

As the solution of both this problem and of *Prob. II* are unique it follows from (3.A.1) and (3.A.2) that  $\forall t$  a.e. in  $[0, t_F]$ ,  $\mathbf{u}_c(t) = \mathbf{v}_t(t)$ .

Now, the problem above is equivalent to the problem

$$\min_{\substack{\mathbf{v}_i \in \mathbb{R}, \\ i=1, \dots, m}} \sum_{i=1}^m (\rho_{\mathbf{u}} v_i + \{Z_a[\mathbf{u}_c](t)\}_i)^2 \text{ subject to } \forall i = 1, \dots, m, \ v_i \in \mathcal{I}_{F_i}(t)$$

which breaks down into  $m$  problems (for  $i = 1, \dots, m$ )

$$\min_{\mathbf{v}_i \in \mathbb{R}} (\mathbf{v}_i - (1/\rho_{\mathbf{u}}) \{-Z_a[\mathbf{u}_c](t)\}_i)^2 \text{ subject to } \mathbf{v}_i \in \mathcal{I}_{F_i}(t)$$



the solution of which is given by

$$\begin{aligned} v_i &= -(1/\rho_{\mathbf{u}}) \{Z_a[\mathbf{u}_c(t)]\}_i \quad \text{if } -(1/\rho_{\mathbf{u}}) \{Z_a[\mathbf{u}_c(t)]\}_i \in \mathcal{I}_{F_i}(t) \\ v_i &= \mu_i \quad \text{if } -(1/\rho_{\mathbf{u}}) \{Z_a[\mathbf{u}_c(t)]\}_i > \mu_i \\ v_i &= -\mu_i \quad \text{if } -(1/\rho_{\mathbf{u}}) \{Z_a[\mathbf{u}_c(t)]\}_i < -\mu_i. \end{aligned}$$

□

**Proof of Proposition 3.3: (a)** It was established in the proof of Proposition 3.1 that

$S_{\mathbf{u}_F}$  is convex and closed. Then, as done in the proof of Proposition 2.1, *Prob. II<sub>K</sub>* is cast as a minimum distance problem to a convex and closed set so that the existence of  $\mathbf{u}_c^K$  follows from (LUENBERGER, 1969, Theorem 3.12.1, p. 69).

**(b)** Proceeding as in the proof of Proposition 2.3, write

$$\begin{aligned} \mathcal{J}(\mathbf{u}_c^K) &= \mathcal{J}(\mathbf{u}_c + (\mathbf{u}_c^K - \mathbf{u}_c)) = \mathcal{J}(\mathbf{u}_c) + 2\langle \rho_{\mathbf{u}}\mathbf{u}_c + Z_a[\mathbf{u}_c], (\mathbf{u}_c^K - \mathbf{u}_c) \rangle \\ &\quad + \|\rho_{\mathbf{u}}(\mathbf{u}_c^K - \mathbf{u}_c)\|_2^2 + \|\mathcal{T}_{\theta}[\mathbf{u}_c^K - \mathbf{u}_c]\| \end{aligned} \quad (3.A.3)$$

and note that (as in the derivation of (2.30))

$$\mathcal{J}_K(\mathbf{u}_c^K) \leq \mathcal{J}_K(\mathbf{u}_c) = \mathcal{J}(\mathbf{u}_c) - E_{\mathcal{J}}^K(\mathbf{u}_c) \Leftrightarrow$$

$$\mathcal{J}(\mathbf{u}_c^K) - E_{\mathcal{J}}^K(\mathbf{u}_c^K) \leq \mathcal{J}(\mathbf{u}_c) - E_{\mathcal{J}}^K(\mathbf{u}_c) \Rightarrow \quad (3.A.4)$$

$$\mathcal{J}(\mathbf{u}_c^K) \leq \mathcal{J}(\mathbf{u}_c) - E_{\mathcal{J}}^K(\mathbf{u}_c) + E_{\mathcal{J}}^K(\mathbf{u}_c^K) \Rightarrow \quad (3.A.5)$$

$$\mathcal{J}(\mathbf{u}_c^K) \leq \mathcal{J}(\mathbf{u}_c) + |E_{\mathcal{J}}^K(\mathbf{u}_c)| + |E_{\mathcal{J}}^K(\mathbf{u}_c^K)|. \quad (3.A.6)$$

Combining (3.A.3) and (3.A.6) leads to

$$\begin{aligned} \|\rho_{\mathbf{u}}(\mathbf{u}_c^K - \mathbf{u}_c)\|_2^2 + \|\mathcal{T}_{\theta}[\mathbf{u}_c^K - \mathbf{u}_c]\|_2^2 &+ 2\langle \rho_{\mathbf{u}}\mathbf{u}_c + Z_a[\mathbf{u}_c], (\mathbf{u}_c^K - \mathbf{u}_c) \rangle \\ &\leq |E_{\mathcal{J}}^K(\mathbf{u}_c)| + |E_{\mathcal{J}}^K(\mathbf{u}_c^K)| \end{aligned}$$

$\Rightarrow$  (in the light of the optimality condition of Proposition 3.1)

$$\rho_{\mathbf{u}}\|\mathbf{u}_c^K - \mathbf{u}_c\|_2^2 \leq |E_{\mathcal{J}}^K(\mathbf{u}_c)| + |E_{\mathcal{J}}^K(\mathbf{u}_c^K)|.$$

Now it follows from (2.30) and the fact that  $\eta_{\mathcal{J}}^K \rightarrow 0$  as  $K \rightarrow \infty$  (Proposition 2.2(b)) that  $|E_{\mathcal{J}}^K(\mathbf{u}_c)| \rightarrow 0$  as  $K \rightarrow \infty$ . Moreover, as  $\mathbf{u}_c^K$  is bounded (since

$\mathbf{u}_c^K \in S_{\mathbf{u}F}$  and hence  $\|\mathbf{u}_c^K\|_{L_2(0,t_F)^m} \leq (\sum_{i=1}^m \mu_i^2)^{1/2} t_F$ , (2.30) and “ $\eta_T^K \rightarrow 0$ ” also imply that  $|E_{\mathcal{J}}^K(\mathbf{u}_c^K)| \rightarrow 0$  as  $K \rightarrow \infty$ . Hence,  $\|\mathbf{u}_c^K - \mathbf{u}_c\|_2 \rightarrow 0$  as  $K \rightarrow \infty$ .  $\square$

**Proof of Proposition 3.5:** The optimality condition satisfied by  $\mathbf{u}_c^K(\boldsymbol{\lambda})$  is given by

$$\begin{aligned} \forall \delta_{\mathbf{u}} \in L_2(0, t_F)^m, \text{Lag}_K(\mathbf{u}, \boldsymbol{\lambda}) \leq \text{Lag}_K(\mathbf{u}_c^K + \delta_{\mathbf{u}}, \boldsymbol{\lambda}) &\Leftrightarrow \\ \forall \delta_{\mathbf{u}} \in L_2(0, t_F)^m, \langle \rho_{\mathbf{u}} \mathbf{u}_c^K, \delta_{\mathbf{u}} \rangle + \langle \mathcal{T}_{\theta}^K[\mathbf{u}_c^K] - \theta_{ro}, \mathcal{T}_{\theta}^K[\delta_{\mathbf{u}}] \rangle + \langle \boldsymbol{\lambda}_a, -\delta_{\mathbf{u}} \rangle + \langle \boldsymbol{\lambda}_b, \delta_{\mathbf{u}} \rangle = 0 &\Leftrightarrow \\ \rho_{\mathbf{u}} \mathbf{u} + (\mathcal{T}_{\theta}^K)^*[\mathcal{T}_{\theta}^K[\mathbf{u}] - \boldsymbol{\theta}_{ro}] + (\boldsymbol{\lambda}_b - \boldsymbol{\lambda}_a) = 0 &\quad (3.A.7) \end{aligned}$$

or, equivalently, taking orthogonal projections  $\mathbf{u}^1$  and  $\mathbf{u}^2$  of  $\mathbf{u}$  on  $(\mathcal{T}_{\theta}^K)^*[L_2(\Omega)]$  and on its orthogonal complement,

$$\rho_{\mathbf{u}} \mathbf{u}^1 + (\mathcal{T}_{\theta}^K)^*[\mathcal{T}_{\theta}^K[\mathbf{u}^1 + \mathbf{u}^2]] - (\mathcal{T}_{\theta}^K)^*[\boldsymbol{\theta}_{ro}] - \boldsymbol{\lambda}_{ab}^1 = 0$$

and  $\rho_{\mathbf{u}} \mathbf{u}^2 - \boldsymbol{\lambda}_{ab}^2 = 0$  where  $\boldsymbol{\lambda}_{ab} \triangleq \boldsymbol{\lambda}_a - \boldsymbol{\lambda}_b$ ,  $\boldsymbol{\lambda}_{ab}^1$  and  $\boldsymbol{\lambda}_{ab}^2$  are the corresponding projections of  $\boldsymbol{\lambda}_{ab}$ .

Noting further that  $\mathcal{T}_{\theta}^K[\mathbf{u}^2] = 0$  ( $\mathbf{u}^2$  is orthogonal to the range space of  $(\mathcal{T}_{\theta}^K)^*$  and hence is in the null space of  $\mathcal{T}_{\theta}^K$ ) the equations above can be rewritten as

$$\rho_{\mathbf{u}} \mathbf{u}^1 + (\mathcal{T}_{\theta}^K)^*[\mathcal{T}_{\theta}^K[\mathbf{u}^1]] - (\mathcal{T}_{\theta}^K)^*[\boldsymbol{\theta}_{ro}] - \boldsymbol{\lambda}_{ab}^1 = 0$$

and  $\rho_{\mathbf{u}} \mathbf{u}^2 = \boldsymbol{\lambda}_{ab} - \boldsymbol{\lambda}_{ab}^1$ .

Now,  $\mathcal{T}_{\theta}^K[\mathbf{u}] = \sum_{k=1}^K c_k(t_F; \mathbf{u}) \phi_k$  and  $(\mathcal{T}_{\theta}^K)^*[\mathbf{w}](\tau) = \mathbf{F}_K(\tau) \bar{\mathbf{w}}^K$ , where  $\{\phi_k; k = 1, \dots, n(K)\}$  is an orthogonal basis for  $X_K$ ,  $c_k(t_F; \mathbf{u}) \triangleq \mathbf{e}_k(n(K))^T \int_0^{t_F} \mathbf{F}_K(\tau)^T \mathbf{u}(\tau) d\tau$ , where  $\mathbf{F}_K(\tau) \triangleq (\mathbf{M}_{\beta}^K)^T \exp[\mathbf{A}_K^T(t_F - \tau)]$ ,  $\bar{\mathbf{w}}^K \triangleq [\langle \mathbf{w}, \phi_1, \rangle \cdots \langle \mathbf{w}, \phi_{n(K)}, \rangle]$  and

$$\mathbf{M}_{\beta}^K \triangleq \begin{bmatrix} \langle \boldsymbol{\beta}_{S1}, \phi_1 \rangle & \cdots & \langle \boldsymbol{\beta}_{Sm}, \phi_1 \rangle \\ \vdots & & \vdots \\ \langle \boldsymbol{\beta}_{S1}, \phi_K \rangle & \cdots & \langle \boldsymbol{\beta}_{Sm}, \phi_{n(K)} \rangle \end{bmatrix}.$$

It follows that  $\mathbf{u}^1 = \mathbf{F}_K \bar{\boldsymbol{\alpha}}_c^K$  and  $\boldsymbol{\lambda}_{ab}^1 = \mathbf{F}_K \bar{\boldsymbol{\alpha}}_{\lambda}^K$  and, hence, the equation involving  $\mathbf{u}^1$  above can be written as

$$\mathbf{F}_K \left\{ \rho_{\mathbf{u}} \bar{\boldsymbol{\alpha}}_c^K + \bar{\mathbf{w}}_a^K [\bar{\boldsymbol{\alpha}}_c^K] - \bar{\boldsymbol{\theta}}_{ro}^K - \bar{\boldsymbol{\alpha}}_{\lambda}^K \right\} = 0, \quad (3.A.8)$$

where  $\bar{\boldsymbol{\theta}}_{ro}^K \triangleq [\langle \theta_{ro}, \theta_1 \rangle \cdots \langle \theta_{ro}, \theta_{n(K)} \rangle]^T$  and

$$\bar{\boldsymbol{w}}_a^K[\bar{\boldsymbol{\alpha}}_c^K] \triangleq [\langle \mathcal{T}_\theta^K[\mathbf{u}^1], \phi_1 \rangle \cdots \langle \mathcal{T}_\theta^K[\mathbf{u}^1], \phi_{n(K)} \rangle]^T$$

$$\begin{aligned} i.e., \bar{\boldsymbol{w}}[\bar{\boldsymbol{\alpha}}_c^K] &= [c_1(t_F; \mathbf{u}^1) \cdots c_{n(K)}(t_F; \mathbf{u}^1)]^T = \int_0^{t_F} \mathbf{F}_K(\tau)^T \mathbf{u}^1(\tau) d\tau \\ &= \left\{ \int_0^{t_F} \mathbf{F}_K(\tau)^T(\tau) \mathbf{F}_K(\tau) \right\} \bar{\boldsymbol{\alpha}}_c^K \Leftrightarrow \bar{\boldsymbol{w}}_a^K[\bar{\boldsymbol{\alpha}}_c^K] = \mathbf{G}_K \bar{\boldsymbol{\alpha}}_c^K \text{ and } \mathbf{G}_K \triangleq \int_0^{t_F} \mathbf{F}_K(\tau)^T \mathbf{F}_K(\tau) d\tau. \end{aligned}$$

A sufficient condition for (3.A.8) to be satisfied is then given by

$$\rho_u \bar{\boldsymbol{\alpha}}_c^K + \mathbf{G}_K \bar{\boldsymbol{\alpha}}_c^K = \bar{\boldsymbol{\theta}}_{ro}^K + \bar{\boldsymbol{\alpha}}_\lambda^K \Leftrightarrow \bar{\boldsymbol{\alpha}}_c^K = (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} (\bar{\boldsymbol{\theta}}_{ro}^K + \bar{\boldsymbol{\alpha}}_\lambda^K).$$

It then follows that  $\mathbf{u}_c^K[\boldsymbol{\lambda}]$  is given by (since  $\rho_u \mathbf{u}^2 = \boldsymbol{\lambda}_{ab}^2$ )

$$\begin{aligned} \mathbf{u}_c^K[\boldsymbol{\lambda}] &= \mathbf{F}_K \bar{\boldsymbol{\alpha}}_c^K + \rho_u^{-1} (\boldsymbol{\lambda}_{ab} - \mathbf{F}_K \bar{\boldsymbol{\alpha}}_\lambda^K) \Leftrightarrow \\ \mathbf{u}_c^K[\boldsymbol{\lambda}](\tau) &= \mathbf{F}_K(\tau) (\bar{\boldsymbol{\alpha}}_c^K - \rho_u^{-1} \bar{\boldsymbol{\alpha}}_\lambda^K) + \rho_u^{-1} \boldsymbol{\lambda}_{ab}(\tau) \Leftrightarrow \\ \mathbf{u}_c^K[\boldsymbol{\lambda}](\tau) &= \mathbf{u}_K(\tau) + \mathbf{F}_K(\tau) \{ (\rho_u \mathbf{I} + \mathbf{G}_K)^{-1} - \rho_u^{-1} \mathbf{I} \} \bar{\boldsymbol{\alpha}}_\lambda^K + \rho_u^{-1} \boldsymbol{\lambda}_{ab}(\tau). \end{aligned}$$

With respect to the dual functional  $\varphi_{DK}(\boldsymbol{\lambda})$ , rewrite  $Lag_K$  as

$$\begin{aligned} Lag_K(\mathbf{u}, \boldsymbol{\lambda}) &= \langle \rho_u \mathbf{u} + (\mathcal{T}_\theta^K)^* [\mathcal{T}_\theta^K[\mathbf{u}] - \boldsymbol{\theta}_{ro}] + (\boldsymbol{\lambda}_b - \boldsymbol{\lambda}_a), \mathbf{u} \rangle + \langle \mathcal{T}_\theta^K[\mathbf{u}] - \boldsymbol{\theta}_{ro}, -\boldsymbol{\theta}_{ro} \rangle \\ &\quad + \langle \boldsymbol{\lambda}_b - \boldsymbol{\lambda}_a, \mathbf{u} \rangle + 2\langle \boldsymbol{\lambda}_a, \mathbf{u}_a \rangle - 2\langle \boldsymbol{\lambda}_b, \mathbf{u}_b \rangle. \end{aligned} \quad (3.A.9)$$

Thus, as  $\varphi_{DK}(\boldsymbol{\lambda}) = Lag_K(\mathbf{u}_c^K[\boldsymbol{\lambda}], \boldsymbol{\lambda})$ , it follows from (3.A.7) and (3.A.9) that

$$\varphi_{DK}(\boldsymbol{\lambda}) = \langle \mathcal{T}_\theta^K[\mathbf{u}_c^K[\boldsymbol{\lambda}]] - \boldsymbol{\theta}_{ro}, -\boldsymbol{\theta}_{ro} \rangle + \langle \boldsymbol{\lambda}_b - \boldsymbol{\lambda}_a, \mathbf{u}_c^K[\boldsymbol{\lambda}] \rangle + 2\langle \boldsymbol{\lambda}_a, \mathbf{u}_a \rangle - 2\langle \boldsymbol{\lambda}_b, \mathbf{u}_b \rangle$$

or equivalently, since

$$\mathbf{u}_c^K[\boldsymbol{\lambda}] = \mathbf{u}_K - \mathbf{u}_K^\xi + \rho_u^{-1} \boldsymbol{\lambda}_{ab}, \quad \varphi_{DK}(\boldsymbol{\lambda}) = \|\boldsymbol{\theta}_{ro}\|_2^2 + \langle \mathcal{T}_\theta^K[\mathbf{u}_K], -\boldsymbol{\theta}_{ro} \rangle + \hat{\varphi}_{DK}(\boldsymbol{\lambda}),$$

where

$$\hat{\varphi}_{DK}(\boldsymbol{\lambda}) \triangleq \langle \mathcal{T}_\theta^K[\rho_u^{-1} \boldsymbol{\lambda}_{ab} - \mathbf{u}_K^\xi], -\boldsymbol{\theta}_{ro} \rangle - \langle \boldsymbol{\lambda}_{ab}, \mathbf{u}_K - \mathbf{u}_K^\xi + \rho_u^{-1} \boldsymbol{\lambda}_{ab} \rangle + 2\langle \boldsymbol{\lambda}_a, \mathbf{u}_a \rangle - 2\langle \boldsymbol{\lambda}_b, \mathbf{u}_b \rangle$$

i.e.,

$$\begin{aligned}\hat{\varphi}_{DK}(\boldsymbol{\lambda}) &= -\rho_{\mathbf{u}}^{-1}\langle\boldsymbol{\lambda}_{ab}, \boldsymbol{\lambda}_{ab}\rangle + \langle\boldsymbol{\lambda}_{ab}, \mathbf{u}_K^\xi - \mathbf{u}_K - \rho_{\mathbf{u}}^{-1}(\mathcal{T}_\theta^K)^*[\boldsymbol{\theta}_{ro}]\rangle + \langle\mathbf{u}_K^\xi, (\mathcal{T}_\theta^K)^*[\boldsymbol{\theta}_{ro}]\rangle \\ &\quad + 2\langle\boldsymbol{\lambda}_a, \mathbf{u}_a\rangle - 2\langle\boldsymbol{\lambda}_b, \mathbf{u}_b\rangle,\end{aligned}$$

and  $\mathbf{u}_K^\xi[\boldsymbol{\lambda}] \triangleq \mathbf{F}_K \{\rho_{\mathbf{u}}^{-1}\mathbf{I} - (\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\} \bar{\boldsymbol{\alpha}}_\lambda^K$  or, equivalently

$$\begin{aligned}(\text{as } \rho_{\mathbf{u}}^{-1}\mathbf{I} - (\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1} &= \rho_{\mathbf{u}}^{-1} \{\mathbf{I} - \rho_{\mathbf{u}}(\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\} = \rho_{\mathbf{u}}^{-1} \{\mathbf{I} - (\mathbf{I} + \rho_{\mathbf{u}}^{-1}\mathbf{G}_K)^{-1}\} \\ &= \rho_{\mathbf{u}}^{-1} \{\rho_{\mathbf{u}}^{-1}\mathbf{G}_K(\mathbf{I} + \rho_{\mathbf{u}}^{-1}\mathbf{G}_K)^{-1}\} = \rho_{\mathbf{u}}^{-1}(\mathbf{I} + \rho_{\mathbf{u}}^{-1}\mathbf{G}_K)^{-1}\rho_{\mathbf{u}}^{-1}\mathbf{G}_K)\end{aligned}$$

$$\mathbf{u}_K^\xi[\boldsymbol{\lambda}] = \mathbf{F}_K \rho_{\mathbf{u}}^{-1}(\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\mathbf{G}_K \bar{\boldsymbol{\alpha}}_\lambda^K.$$

Finally, as  $(\mathcal{T}_\theta^K)^*[\boldsymbol{\theta}_{ro}] = \mathbf{F}_K \bar{\boldsymbol{\theta}}_{ro}^K$  and  $\mathbf{u}_K = \mathbf{F}_K \bar{\boldsymbol{\alpha}}_K$ ,  
 $\langle\mathbf{u}_K^\xi[\boldsymbol{\lambda}], (\mathcal{T}_\theta^K)^*[\boldsymbol{\theta}_{ro}]\rangle = \langle\rho_{\mathbf{u}}^{-1}(\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\mathbf{G}_K \bar{\boldsymbol{\alpha}}_\lambda^K, \mathbf{G}_K \bar{\boldsymbol{\theta}}_{ro}^K\rangle_E$  and  
 $\langle\boldsymbol{\lambda}_{ab}, \mathbf{u}_K^\xi - \mathbf{u}_K - \rho_{\mathbf{u}}^{-1}(\mathcal{T}_\theta^K)^*[\boldsymbol{\theta}_{ro}]\rangle = \langle\boldsymbol{\xi}_\lambda^K, \rho_{\mathbf{u}}^{-1}(\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\mathbf{G}_K \bar{\boldsymbol{\alpha}}_\lambda^K - \bar{\boldsymbol{\alpha}}_K - \rho_{\mathbf{u}}^{-1}\bar{\boldsymbol{\theta}}_{ro}^K\rangle_E$ ,  
 where  $\bar{\boldsymbol{\alpha}}_K = (\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\bar{\boldsymbol{\theta}}_{ro}^K$ ,  $\boldsymbol{\xi}_\lambda^K \triangleq \int_0^{t_F} \mathbf{F}_K(\tau)^\top \boldsymbol{\lambda}_{ab}(\tau) d\tau$  ( $\boldsymbol{\xi}_\lambda^K = \mathbf{G}_K \bar{\boldsymbol{\alpha}}_\lambda^K$ ).

As a result,  $\hat{\varphi}_{DK}(\boldsymbol{\lambda})$  is given by

$$\begin{aligned}\hat{\varphi}_{DK}(\boldsymbol{\lambda}) &= -\rho_{\mathbf{u}}^{-1}\langle\boldsymbol{\lambda}_{ab}, \boldsymbol{\lambda}_{ab}\rangle + \rho_{\mathbf{u}}^{-1}\langle\boldsymbol{\xi}_\lambda^K, (\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\mathbf{G}_K \bar{\boldsymbol{\alpha}}_\lambda^K\rangle - \langle\boldsymbol{\xi}_\lambda^K, \bar{\boldsymbol{\alpha}}_K + \rho_{\mathbf{u}}^{-1}\bar{\boldsymbol{\theta}}_{ro}^K\rangle \\ &\quad + \rho_{\mathbf{u}}^{-1}\langle(\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\bar{\boldsymbol{\alpha}}_\lambda^K, \mathbf{G}_K \bar{\boldsymbol{\theta}}_{ro}^K\rangle + 2\langle\boldsymbol{\lambda}_a, \mathbf{u}_a\rangle - 2\langle\boldsymbol{\lambda}_b, \mathbf{u}_b\rangle.\end{aligned}$$

Now,  $\langle(\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\mathbf{G}_K \bar{\boldsymbol{\alpha}}_\lambda^K, \mathbf{G}_K \bar{\boldsymbol{\theta}}_{ro}^K\rangle = \langle\boldsymbol{\xi}_\lambda^K, (\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\mathbf{G}_K \bar{\boldsymbol{\theta}}_{ro}^K\rangle \Rightarrow$

$$\begin{aligned}\hat{\varphi}_{DK}(\boldsymbol{\lambda}) &= -\rho_{\mathbf{u}}^{-1}\langle\boldsymbol{\lambda}_{ab}, \boldsymbol{\lambda}_{ab}\rangle + \rho_{\mathbf{u}}^{-1}\langle\boldsymbol{\xi}_\lambda^K, (\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\boldsymbol{\xi}_\lambda^K\rangle - \langle\boldsymbol{\xi}_\lambda^K, \bar{\boldsymbol{\alpha}}_K\rangle \\ &\quad + \rho_{\mathbf{u}}^{-1}\langle\boldsymbol{\xi}_\lambda^K, \{(\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\mathbf{G}_K - \mathbf{I}\} \bar{\boldsymbol{\theta}}_{ro}^K\rangle + 2\langle\boldsymbol{\lambda}_a, \mathbf{u}_a\rangle - 2\langle\boldsymbol{\lambda}_b, \mathbf{u}_b\rangle.\end{aligned}$$

Moreover, as  $(\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\mathbf{G}_K = (\mathbf{I} + \rho_{\mathbf{u}}^{-1}\mathbf{G}_K)^{-1}\rho_{\mathbf{u}}^{-1}\mathbf{G}_K = \mathbf{I} - (\mathbf{I} + \rho_{\mathbf{u}}^{-1}\mathbf{G}_K)^{-1}$   
 so that

$$\begin{aligned}\hat{\varphi}_{DK}(\boldsymbol{\lambda}) &= -\rho_{\mathbf{u}}^{-1}\langle\boldsymbol{\lambda}_{ab}, \boldsymbol{\lambda}_{ab}\rangle + \rho_{\mathbf{u}}^{-1}\langle\boldsymbol{\xi}_\lambda^K, (\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\boldsymbol{\xi}_\lambda^K\rangle - \langle\boldsymbol{\xi}_\lambda^K, \bar{\boldsymbol{\alpha}}_K\rangle \\ &\quad - \rho_{\mathbf{u}}^{-1}\langle\boldsymbol{\xi}_\lambda^K, (\mathbf{I} + \rho_{\mathbf{u}}^{-1}\mathbf{G}_K)^{-1}\bar{\boldsymbol{\theta}}_{ro}^K\rangle + 2\langle\boldsymbol{\lambda}_a, \mathbf{u}_a\rangle - 2\langle\boldsymbol{\lambda}_b, \mathbf{u}_b\rangle.\end{aligned}$$

Note now that  $\rho_{\mathbf{u}}^{-1}(\mathbf{I} + \rho_{\mathbf{u}}^{-1}\mathbf{G}_K)^{-1}\bar{\boldsymbol{\theta}}_{ro}^K = (\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\bar{\boldsymbol{\theta}}_{ro}^K = \bar{\boldsymbol{\alpha}}_K$ . Thus,

$$\begin{aligned}\hat{\varphi}_{DK}(\boldsymbol{\lambda}) &= -\rho_{\mathbf{u}}^{-1}\langle \boldsymbol{\lambda}_{ab}, \boldsymbol{\lambda}_{ab} \rangle + \rho_{\mathbf{u}}^{-1}\langle \boldsymbol{\xi}_{\lambda}^K, (\rho_{\mathbf{u}}\mathbf{I} + \mathbf{G}_K)^{-1}\boldsymbol{\xi}_{\lambda}^K \rangle - 2\langle \boldsymbol{\xi}_{\lambda}^K, \bar{\boldsymbol{\alpha}}_K \rangle \\ &\quad + 2\langle \boldsymbol{\lambda}_a, \mathbf{u}_a \rangle - 2\langle \boldsymbol{\lambda}_b, \mathbf{u}_b \rangle.\end{aligned}$$

□

## 4 EXAMPLES AND NUMERICAL RESULTS FOR THE LHEQ

In this chapter, two simple numerical examples are presented to illustrate the way the results above can be used to characterize control signals which aim at steering a solution of a PDEE over a given interval  $[0, t_F]$  towards a prescribed final state. It is of particular interest here to illustrate the role of the coefficient  $\rho_F$  in improving final-state approximation, the effect on imposing a peak-value constraint on the control signals (vis-à-vis the unconstrained one) and the way piecewise-linear multipliers yield approximation to the optimal control signals under peak-value constraints.

In Section 4.1, the one-dimensional LHEq is considered under the action of a single scalar control signal (*i.e.*,  $m = 1$ ). To facilitate reading (and for concreteness) some of the relevant symbol definitions ( $\mathbf{A}_K, \bar{\boldsymbol{\beta}}_{\mathbf{S}K}^T, \boldsymbol{\theta}_{ro}^K$ ) are re-stated now for the basis functions  $\left\{ \sqrt{\frac{2}{L_x}} \sin \left[ \frac{k\pi x}{L_x} \right] \right\}$ ,  $k = 1, \dots, K$ . Exploiting the simple case at hand, an explicit upper bound is presented on the  $L_2$ -norm of the approximation error to the final state of the LHEq as a function of the correcting error in the truncated (ODE in  $\mathbb{R}^K$ ) problem. Some of notation introduced for the dual problem (Chapter 3) is also reproduced in Section 3.1 to facilitate reading. In Section 4.2, numerical results are presented for the one-dimensional example of Section 4.1 with two distinct temperature distribution taken as desired final states and one actuator located at the mid-point of the interval  $(0, L_x)$ . In Section 4.3, numerical results are presented for the two-dimensional LHEq with one scalar control signal; for one desired final state, numerical experiments were carried with two different values of  $\rho_F$ .

Finally in Section 4.4, numerical experiments to “locate” actuator are reported for the two-dimensional LHEq under the action of two and three scalar control signals ( $m = 2$  and  $m = 3$ ).

### 4.1 A One-Dimensional Example

Let  $\Omega = (0, L_x)$  and consider the one-dimensional heat equation with homogeneous Dirichlet boundary conditions and single-point control  $\mathbf{u} : [0, t_F] \rightarrow \mathbb{R}$ , *i.e.*,

$$\begin{aligned} \frac{\partial \theta}{\partial t}(x, t) &= k_\alpha \frac{\partial^2 \theta}{\partial x^2}(x, t) + \boldsymbol{\beta}_{\mathbf{S}}(x) \mathbf{u}(t) && \forall t \in (0, \infty), \forall x \in \Omega, \\ \theta(x, 0) &= 0 && \text{(zero initial condition)} \quad \forall x \in \Omega, \\ \theta(0, t) &= \theta(L_x, t) = 0 && \text{(boundary conditions)} \quad \forall t \in (0, \infty) \end{aligned}$$

with the corresponding weak version given by

$$\forall k = 1, 2, \dots, K, \quad \left\langle \frac{\partial \theta}{\partial t}(\cdot, t), \phi_k \right\rangle = -k_\alpha \left\langle \frac{\partial \theta}{\partial x}(\cdot, t), \frac{\partial \phi_k}{\partial x} \right\rangle + \langle \boldsymbol{\beta}_S, \phi_k \rangle \mathbf{u}(t)$$

$$\langle \theta(\cdot, 0), \phi_k \rangle = 0,$$

where  $\phi_k : [0, L_x] \rightarrow \mathbb{R}$  is given by  $\phi_k(x) = \sqrt{\frac{2}{L_x}} \sin \left[ \frac{k\pi x}{L_x} \right]$ .

Approximate solutions  $\mathbf{u}_K$  and  $\mathbf{u}_c^K$  are sought to the problems

$$\underline{\text{Prob. I}}: \min_{\mathbf{u} \in L_2(0, t_F)} \check{\mathcal{J}}(\mathbf{u}; \rho_F) \quad \text{or} \quad \underline{\text{Prob. I}_c}: \min_{\mathbf{u} \in S_{\mathbf{u}_F}} \check{\mathcal{J}}(\mathbf{u}; \rho_F),$$

where  $\check{\mathcal{J}}(\mathbf{u}; \rho_F) = \|\mathbf{u}\|_{L_2(0, t_F)}^2 + \rho_F \|\mathcal{T}_\theta[\mathbf{u}] - \theta_{ro}\|_2^2$ ,  $\theta_{ro}$  is the final state to be approximately reached,  $\rho_F = \rho_{\mathbf{u}}^{-1}$  and

$$S_{\mathbf{u}_F} = \{ \mathbf{u} \in L_\infty(0, t_F) : \|\mathbf{u}\|_{L_\infty(0, t_F)} \leq \mu_{\mathbf{u}} \}.$$

In this case,  $\{\mathbf{A}_K\}_{kl} = -k_\alpha \left\langle \sqrt{\frac{2}{L_x}} \left[ -\frac{k\pi}{L_x} \right] \cos \left[ \frac{k\pi(\cdot)}{L_x} \right], \left[ \sqrt{\frac{2}{L_x}} \left[ -\frac{\ell\pi}{L_x} \right] \cos \left[ \frac{\ell\pi(\cdot)}{L_x} \right] \right\rangle$ , i.e.,

$$\mathbf{A}_K = \text{diag} \left\{ -k_\alpha \left[ \frac{k\pi}{L_x} \right]^2 \right\}$$

and

$$\bar{\boldsymbol{\beta}}_{SK}^T = \left[ \left\langle \boldsymbol{\beta}_S, \sqrt{\frac{2}{L_x}} \sin \left[ \frac{1\pi(\cdot)}{L_x} \right] \right\rangle \cdots \left\langle \boldsymbol{\beta}_S, \sqrt{\frac{2}{L_x}} \sin \left[ \frac{K\pi(\cdot)}{L_x} \right] \right\rangle \right].$$

The optimal solution of *Prob. I* is given by,  $\forall \tau \in [0, t_F]$

$$\mathbf{u}(\tau) = \bar{\boldsymbol{\beta}}_{SK}^T \exp\{\mathbf{A}_K^T(t_F - \tau)\} \bar{\boldsymbol{\alpha}}_K,$$

where  $\bar{\boldsymbol{\alpha}}_K = (\mathbf{I} + \rho_F \mathbf{G}_K)^{-1} \rho_F \bar{\boldsymbol{\theta}}_{ro}^K$ ,

$$(\bar{\boldsymbol{\theta}}_{ro}^K)^T = \left[ \left\langle \theta_{ro}, \sqrt{\frac{2}{L_x}} \sin \left[ \frac{1\pi(\cdot)}{L_x} \right] \right\rangle \cdots \left\langle \theta_{ro}, \sqrt{\frac{2}{L_x}} \sin \left[ \frac{K\pi(\cdot)}{L_x} \right] \right\rangle \right]$$

and  $\mathbf{G}_K = \int_0^{t_F} \exp[\mathbf{A}_K t] \bar{\boldsymbol{\beta}}_{S_K} \bar{\boldsymbol{\beta}}_{S_K}^T \exp[\mathbf{A}_K t]^T dt$ , *i.e.*,  $\mathbf{G}_K$  is the unique solution of

$$\mathbf{A}_K \mathbf{G}_K + \mathbf{G}_K \mathbf{A}_K^T = \exp[\mathbf{A}_K t_F] \bar{\boldsymbol{\beta}}_{S_K} \bar{\boldsymbol{\beta}}_{S_K}^T \exp[\mathbf{A}_K t_F]^T - \bar{\boldsymbol{\beta}}_{S_K} \bar{\boldsymbol{\beta}}_{S_K}^T.$$

The approximation error on the final state for a given control signal  $\mathbf{u}$  is given by  $\mathcal{T}_\theta[\mathbf{u}] - \theta_{ro} = \mathbf{e}_K[\mathbf{u}] + \check{\mathbf{e}}_K[\mathbf{u}]$  where  $\mathbf{e}_K[\mathbf{u}] \triangleq \mathcal{T}_\theta^K[\mathbf{u}] - \theta_{ro}^K$  (error projection on  $\text{span}\{\phi_1, \dots, \phi_K\}$ ) and  $\check{\mathbf{e}}_K[\mathbf{u}] = \{\mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}]\} - \{\theta_{ro} - \theta_{ro}^K\}$ .

To get an upper bound on  $\|\mathcal{T}_\theta[\mathbf{u}] - \theta_{ro}\|_2$  note that

$$\|\mathcal{T}_\theta[\mathbf{u}] - \theta_{ro}\|_2^2 = \|\mathbf{e}_K[\mathbf{u}]\|_2^2 + \|\check{\mathbf{e}}_K[\mathbf{u}]\|_2^2, \quad (4.1)$$

$$\|\mathbf{e}_K[\mathbf{u}]\|_2^2 = \|\bar{\mathbf{c}}_K(t_F; \mathbf{u}) - \bar{\boldsymbol{\theta}}_{ro}^K\|_E^2, \quad (4.2)$$

$$\|\check{\mathbf{e}}_K[\mathbf{u}]\|_2 \leq \|\mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}]\|_2 + \|\theta_{ro} - \theta_{ro}^K\|_2. \quad (4.3)$$

Note also that  $\|\mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}]\|_2^2 = \left\| \sum_{k=K+1}^{\infty} \mathbf{c}_k(t_F; \mathbf{u}) \phi_k \right\|_2^2 = \sum_{k=K+1}^{\infty} \mathbf{c}_k(t_F; \mathbf{u})^2$ , and

$\mathbf{c}_k(t_F, \mathbf{u}) = \int_0^{t_F} \exp\left[-k_\alpha \left[\frac{k\pi}{L_x}\right]^2 (t_F - \tau)\right] \boldsymbol{\beta}_{S_k} \mathbf{u}(\tau) d\tau$ , where  $\boldsymbol{\beta}_{S_k} \triangleq \langle \boldsymbol{\beta}_S, \phi_k \rangle$ , so that (in the light of Cauchy-Schwarz inequality)

$$\begin{aligned} \Rightarrow \mathbf{c}_k(t_F; \mathbf{u})^2 &\leq |\boldsymbol{\beta}_{S_k}|^2 \left\| \exp\left[-k_\alpha \left[\frac{k\pi}{L_x}\right]^2 (t_F - \cdot)\right] \right\|_{L_2(0, t_F)}^2 \|\mathbf{u}\|_{L_2(0, t_F)}^2 \\ \Rightarrow \mathbf{c}_k(t_F; \mathbf{u})^2 &\leq |\boldsymbol{\beta}_{S_k}|^2 \frac{1}{k_\alpha \left[\frac{k\pi}{L_x}\right]^2} \left\{ 1 - \exp\left[-k_\alpha \left[\frac{k\pi}{L_x}\right]^2 t_F\right] \right\} \|\mathbf{u}\|_{L_2(0, t_F)}^2 \\ &\leq |\boldsymbol{\beta}_{S_k}|^2 \frac{1}{k_\alpha \left[\frac{k\pi}{L_x}\right]^2} \|\mathbf{u}\|_{L_2(0, t_F)}^2. \end{aligned}$$

It then follows that

$$\|\mathcal{T}_\theta[\mathbf{u}] - \mathcal{T}_\theta^K[\mathbf{u}]\|_2^2 \leq \|\boldsymbol{\beta}_S - \hat{\boldsymbol{\beta}}_{S_K}\|_2^2 \frac{1}{k_\alpha \left\{ (K+1) \frac{\pi}{L_x} \right\}^2} \|\mathbf{u}\|_{L_2(0, t_F)}^2, \quad (4.4)$$

where  $\hat{\boldsymbol{\beta}}_{S_K} \triangleq \sum_{k=1}^K \boldsymbol{\beta}_{S_k} \phi_k$ .



Thus, combining (4.1) -(4.4) gives an upper bound on  $\|\mathcal{T}_\theta[\mathbf{u}] - \theta_{ro}\|_2^2$  namely,

$$\|\mathcal{T}_\theta[\mathbf{u}] - \theta_{ro}\|_2^2 \leq \|\bar{\mathbf{c}}_K(t_F; \mathbf{u}) - \bar{\boldsymbol{\theta}}_{ro}^K\|_E^2 + \left\{ \frac{\|\boldsymbol{\beta}_S - \hat{\boldsymbol{\beta}}_{SK}\|_2}{\sqrt{k_\alpha(K+1)} \frac{\pi}{L_x}} \|\mathbf{u}\|_{L_2(0,t_F)} + \|\theta_{ro} - \theta_{ro}^K\|_2 \right\}^2.$$

Thus, as the approximately property of  $\text{span} \left\{ \sqrt{\frac{2}{L_x}} \sin \left[ \frac{k\pi x}{L_x} \right] : k = 1, \dots, K \right\}$ ,  $K \geq 1$  ensures that  $\|\boldsymbol{\beta}_S - \hat{\boldsymbol{\beta}}_{SK}\|_2 \rightarrow 0$  and  $\|\theta_{ro} - \theta_{ro}^K\|_2 \rightarrow 0$  as  $K \rightarrow \infty$ , that  $L_2$ -norm of the approximation error for the final state for the LHEq (*i.e.*,  $\|\mathcal{T}_\theta[\mathbf{u}] - \theta_{ro}\|_2$  approaches the corresponding error for the  $K$ -dimensional ODE, *i.e.*,  $\|\bar{\mathbf{c}}_K(t_F; \mathbf{u}) - \bar{\boldsymbol{\theta}}_{ro}^K\|_E$ ).

For the optimal solution of *Prob. I<sub>K</sub>*, the latter is given by

$$\|\mathbf{e}_K[\mathbf{u}_K]\|_2^2 = \|\bar{\mathbf{c}}_K(t_F; \mathbf{u}_K) - \bar{\boldsymbol{\theta}}_{ro}^K\|_2^2 \quad \text{and since}$$

$$\begin{aligned} \bar{\mathbf{c}}_K(t_F; \mathbf{u}_K) &= \int_0^{t_F} \mathbf{H}_K(t_F - \tau) \mathbf{u}_K(\tau) d\tau = \int_0^{t_F} \mathbf{H}_K(t_F - \tau) \mathbf{H}_K(t_F - \tau)^T \bar{\boldsymbol{\alpha}}_K d\tau, \text{ where} \\ \mathbf{H}_K(t) &= \exp[\mathbf{A}_K t] \boldsymbol{\beta}_{SK}, \quad \bar{\mathbf{c}}_K(t_F; \mathbf{u}_K) = \mathbf{G}_K \bar{\boldsymbol{\alpha}}_K \quad \Leftrightarrow \\ \bar{\mathbf{c}}_K(t_F; \mathbf{u}_K) &= \mathbf{G}_K (\mathbf{I} + \rho_F \mathbf{G}_K)^{-1} \rho_F \bar{\boldsymbol{\theta}}_{ro}^K = \{\mathbf{I} - (\mathbf{I} + \rho_F \mathbf{G}_K)^{-1}\} \bar{\boldsymbol{\theta}}_{ro}^K \text{ it follows that} \end{aligned}$$

$$\|\mathbf{e}_K[\mathbf{u}_K]\|_2^2 = \|(\mathbf{I} + \rho_F \mathbf{G}_K)^{-1} \bar{\boldsymbol{\theta}}_{ro}^K\|_2^2. \quad (4.5)$$

–it can thus be seen that whenever  $\mathbf{G}_K$  is nonsingular  $\|\mathbf{e}_K[\mathbf{u}_K]\|_2 \rightarrow 0$  as  $\rho_F \rightarrow \infty$ .

To compute approximate solutions to *Prob. I<sub>c</sub>*, consider the truncated problem

*Prob. I<sub>cK</sub>* :  $\min_{\mathbf{u} \in S_{uF}} \tilde{\mathcal{J}}_K(\mathbf{u}; \rho_F)$  and the corresponding dual problem,

*Prob. D<sub>K</sub>* :  $\max_{\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b} \varphi_D^K(\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b; \rho_F)$  subject to  $\forall t$  a.e. in  $(0, t_F)$ ,  $\boldsymbol{\lambda}_a \geq 0$ ,  $\boldsymbol{\lambda}_b \geq 0$ ,

where  $\varphi_D^K(\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b) = \inf\{\text{Lag}_K(\mathbf{u}; \boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b) : \mathbf{u} \in L_2(0, t_F)\}$ ,

$\text{Lag}_K(\mathbf{u}; \boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b) = \tilde{\mathcal{J}}_K(\mathbf{u}; \rho_F) + 2\langle \boldsymbol{\lambda}_a, \mathbf{u}_a - \mathbf{u} \rangle + 2\langle \boldsymbol{\lambda}_b, \mathbf{u} - \mathbf{u}_b \rangle$  and  $\mathbf{u}_b = \mu_{\mathbf{u}}$  and  $\mathbf{u}_a = -\mu_{\mathbf{u}}$ ,

and  $S_{uF} = \{\mathbf{u} \in L_2(0, t_F) : \forall t \text{ a.e. in } (0, t_F), -\mu_{\mathbf{u}} \leq \mathbf{u}(t) \leq \mu_{\mathbf{u}}\}$ .

The unique solution to the problem  $\min_{\mathbf{u} \in L_2(0, t_F)} \text{Lag}_K(\mathbf{u}; \boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b)$  is given by

$$\mathbf{u}_c^K[\boldsymbol{\lambda}] = \hat{\mathbf{u}}_c^K + \boldsymbol{\lambda}_{ab}, \quad \text{where} \quad \hat{\mathbf{u}}_c^K[\boldsymbol{\lambda}](\tau) = \mathbf{H}_K^T(t_F - \tau) \left\{ \bar{\boldsymbol{\alpha}}_K - (\mathbf{I} + \rho_F \mathbf{G}_K)^{-1} \rho_F \boldsymbol{\xi}_\lambda^K \right\},$$

$$\boldsymbol{\lambda} = (\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b), \quad \boldsymbol{\lambda}_{ab} = \boldsymbol{\lambda}_a - \boldsymbol{\lambda}_b \text{ and } \boldsymbol{\xi}_\lambda^K = \int_0^{t_F} \mathbf{H}_K(t_F - \tau) \boldsymbol{\lambda}_{ab}(\tau) d\tau.$$

The corresponding value for the dual functional is given by

$$\varphi_D^K(\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b) = \text{Lag}_K(\mathbf{u}_c^K[\boldsymbol{\lambda}]; \boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b) = \rho_F \|\theta_{ro}\|_2^2 + \rho_F \langle \mathcal{T}_\theta^K[\mathbf{u}_c^K], -\theta_{ro} \rangle + \hat{\varphi}_D^K(\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b),$$

where

$$\begin{aligned}\hat{\varphi}_D^K(\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b) &= -\langle \boldsymbol{\lambda}_{ab}, \boldsymbol{\lambda}_{ab} \rangle + \rho_F \langle (\mathbf{I} + \rho_F \mathbf{G}_K)^{-1} \boldsymbol{\xi}_\lambda^K, \boldsymbol{\xi}_\lambda^K \rangle_E - 2\langle \boldsymbol{\xi}_\lambda^K, \bar{\boldsymbol{\alpha}}_K \rangle_E + 2\langle \boldsymbol{\lambda}_a, \mathbf{u}_a \rangle \\ &\quad - 2\langle \boldsymbol{\lambda}_b, \mathbf{u}_b \rangle.\end{aligned}$$

Note that for any non-negative  $\boldsymbol{\lambda}_a$  and  $\boldsymbol{\lambda}_b$ ,  $\varphi_D^K(\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b)$  is a lower bound for the optimal value of  $Prob. I_{cK}$ . If  $(\boldsymbol{\lambda}_a^o, \boldsymbol{\lambda}_b^o)$  is optimal  $\mathbf{u}_c^K \in S_{uF}$ . Moreover,  $\boldsymbol{\lambda}_a^o(\tau) = 0$  and  $\boldsymbol{\lambda}_b^o(\tau) = 0$  (hence,  $\boldsymbol{\lambda}_{ab}^o(\tau) = 0$ ) whenever  $\mathbf{u}_c^K[\boldsymbol{\lambda}^o](\tau) \in (\mathbf{u}_a, \mathbf{u}_b)$  so that, in this case,  $\hat{\mathbf{u}}_K^c[\boldsymbol{\lambda}^o](\tau)$  also belongs to  $(\mathbf{u}_a, \mathbf{u}_b)$ . When  $\boldsymbol{\lambda}_a^o(\tau) \neq 0$  (respectively  $\boldsymbol{\lambda}_b^o(\tau) \neq 0$ )  $\mathbf{u}_c^K(\tau) = \mathbf{u}_a$  and  $\hat{\mathbf{u}}_K^c[\boldsymbol{\lambda}^o](\tau) < \mathbf{u}_a$  (respectively,  $\mathbf{u}_c^K[\boldsymbol{\lambda}^o](\tau) = \mathbf{u}_b$  and  $\hat{\mathbf{u}}_K^c[\boldsymbol{\lambda}^o](\tau) > \mathbf{u}_a$ ). This suggests a heuristic way of obtaining a feasible  $\mathbf{u}_K^R[\boldsymbol{\lambda}]$ , namely,  $\mathbf{u}_K^R[\boldsymbol{\lambda}](\tau) = \hat{\mathbf{u}}_K^c[\boldsymbol{\lambda}](\tau)$  if  $\hat{\mathbf{u}}_K^c[\boldsymbol{\lambda}] \in (\mathbf{u}_a, \mathbf{u}_b)$ ,  $\mathbf{u}_K^R[\boldsymbol{\lambda}](\tau) = \mathbf{u}_a$  if  $\hat{\mathbf{u}}_K^c[\boldsymbol{\lambda}](\tau) \leq \mathbf{u}_a$  and  $\mathbf{u}_K^R[\boldsymbol{\lambda}](\tau) = \mathbf{u}_b$  if  $\hat{\mathbf{u}}_K^c[\boldsymbol{\lambda}](\tau) \geq \mathbf{u}_b$ .

To obtain approximate solutions to  $Prob. D_K$ , piecewise linear classes of multipliers are considered, *i.e.*, let  $N_\lambda \in \mathbb{Z}_+$ ,  $\delta_t = t_F/N_\lambda$ ,  $\mathcal{I}_k = [(k-1)\delta_t, k\delta_t]$ ,  $\boldsymbol{\gamma} = [\gamma_1 \cdots \gamma_{N_\lambda+1}]$  and define  $\forall k = 1, \dots, N_\lambda$ ,  $\forall t \in \mathcal{I}_k$ ,  $\boldsymbol{\lambda}(t; \boldsymbol{\gamma}) = \boldsymbol{\gamma}_k + (1/\delta_t)(\boldsymbol{\gamma}_{k+1} - \boldsymbol{\gamma}_k)\Delta t_k$ , where  $\Delta t_k = t - (k-1)\delta_t$  (note that  $\boldsymbol{\gamma}_k$  and  $\boldsymbol{\gamma}_{k+1}$  are respectively the values of  $\boldsymbol{\lambda}(t, \boldsymbol{\lambda})$  at the lower and upper extreme points of the interval  $\mathcal{I}_k$ ). Such multipliers can then be written as a function of  $\boldsymbol{\gamma}$  as follows:

$$\forall t \in \mathcal{I}_k, \quad \boldsymbol{\lambda}(t; \boldsymbol{\gamma}) = \mathbf{h}_{kab}^\top(t) \mathbf{E}_k \boldsymbol{\gamma},$$

where  $\mathbf{h}_{kab}^\top(t) = [h_{ka}(t) \ : \ h_{kb}(t)]$ ,  $\mathbf{E}_k^\top = [e_k(m_\gamma) \ : \ e_{k+1}(m_\gamma)]$ ,  $m_\gamma = N_\lambda + 1$ ,  $h_{ka} : \mathcal{I}_k \rightarrow \mathbb{R}$ ,  $h_{ka}(t) = 1 - h_{kb}(t)$ ,  $h_{kb} : \mathcal{I}_k \rightarrow \mathbb{R}$ ,  $h_{kb}(t) = (1/\delta_t)(t - a_k)$ , where  $a_k = (k-1)\delta_t$ .

As a result,  $\boldsymbol{\xi}_\lambda^K = \mathbf{T}_{\boldsymbol{\xi}\boldsymbol{\gamma}}(\boldsymbol{\gamma}_a - \boldsymbol{\gamma}_b)$ , where  $\mathbf{T}_{\boldsymbol{\xi}\boldsymbol{\gamma}} = \left\{ \sum_{k=1}^{N_\lambda} \int_{\mathcal{I}_k} \mathbf{H}_K(t_f - \tau) \mathbf{h}_{kab}^\top(\tau) d\tau \right\} \mathbf{E}_k$

and

$$-\hat{\varphi}_D^K(\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b) = \boldsymbol{\gamma}_{ab}^\top (\mathbf{P}_\boldsymbol{\gamma} - \mathbf{T}_{\boldsymbol{\xi}\boldsymbol{\gamma}}^\top \rho_F (\mathbf{I} + \rho_F \mathbf{G}_K)^{-1} \mathbf{T}_{\boldsymbol{\xi}\boldsymbol{\gamma}}) \boldsymbol{\gamma}_{ab} + 2\bar{\boldsymbol{\alpha}}_K^\top \mathbf{T}_{\boldsymbol{\xi}\boldsymbol{\gamma}} \boldsymbol{\gamma}_{ab} - 2\mathbf{r}_{\boldsymbol{\gamma}_a}^\top \boldsymbol{\gamma}_a + 2\mathbf{r}_{\boldsymbol{\gamma}_b}^\top \boldsymbol{\gamma}_b,$$

where  $\boldsymbol{\gamma}_{ab} \triangleq \boldsymbol{\gamma}_a - \boldsymbol{\gamma}_b$ ,  $\mathbf{P}_\boldsymbol{\gamma} \triangleq \sum_{k=1}^{N_\lambda} \mathbf{E}_k^\top \int_{\mathcal{I}_k} \mathbf{h}_{kab}(t) \mathbf{h}_{kab}^\top(t) dt \mathbf{E}_k$ ,  $\mathbf{r}_{\boldsymbol{\gamma}_a}^\top = \sum_{k=1}^{N_\lambda} \left\{ \left[ \int_{\mathcal{I}_k} \mathbf{u}_a(t) \mathbf{h}_{kab}^\top(t) dt \right] \mathbf{E}_k \right\}$ ,

and  $\mathbf{r}_{\boldsymbol{\gamma}_b}^\top = \sum_{k=1}^{N_\lambda} \left\{ \left[ \int_{\mathcal{I}_k} \mathbf{u}_b(t) \mathbf{h}_{kab}^\top(t) dt \right] \mathbf{E}_k \right\}$ .

The problem to be numerically solved is then

$$\underline{Prob. D_\boldsymbol{\gamma}^K} : \max_{\boldsymbol{\gamma}_a, \boldsymbol{\gamma}_b \in \mathbb{R}^{N_\lambda+1}} \varphi_D^K(\boldsymbol{\lambda}_a(\boldsymbol{\gamma}_a), \boldsymbol{\lambda}_b(\boldsymbol{\gamma}_b); \rho_F). \quad (4.6)$$

## 4.2 Numerical Results for the One-dimensional Example

*Prob.  $I_K$*  and *Prob.  $D_\gamma^K$*  were numerically solved for two pairs  $(\theta_r, \beta_S)$  respectively displayed in Figures 1, 2 and Figures 7, 8, with  $\rho_F = 2000$ ,  $K = 5$ ,  $L_x = 1$  or 2, and  $N_\lambda = 30$ . For the first pair  $(\theta_r, \beta_S)$  the unconstrained problem was solved leading to the approximate solution  $\mathbf{u}_K(\cdot; \rho_F)$  which is plotted in Figure 3 (dashed blue curve, labeled  $\mathbf{u}_K$ ). Table 1 gives the  $L_2(0, t_F)$  and  $L_\infty(0, t_F)$  norms of  $\mathbf{u}_K(\cdot; \rho_F)$  and the  $L_2(\Omega)$  norm of the projection of the final-state, approximation error on  $\text{span}\{\phi_1, \dots, \phi_K\}$ .

| $\check{\mathcal{J}}_K(\mathbf{u}_K; \rho_F)$ | $\ \mathbf{u}_K\ _2$ | $\ \mathbf{u}_K\ _\infty$ | $\ \mathcal{T}_\theta^K[\mathbf{u}_K] - \theta_{ro}^K\ _2$ |
|---|----------------------|---------------------------|--|
| 160.5171                                      | 10.9233              | 43.5917                   | 0.1435   |

Table 1: Unconstrained problem for the first pair  $(\theta_r, \beta_S)$ ,  $\rho_F = 2000$ .

The constrained problem *Prob.  $I_{cK}$*  was then solved for the same pair  $(\theta_r, \beta_S)$  with the prescribed upper bound  $\mu_u$  on  $\|\mathbf{u}\|_\infty$  taken to be  $\mu_u = 30$ . Approximate solutions are then obtained for *Prob.  $D_\gamma^K$* , say  $(\gamma_a^K, \gamma_b^K)$ . The corresponding multipliers are denoted by  $\lambda_a^K$  and  $\lambda_b^K$  on the basis of which a feasible solution for *Prob.  $I_{cK}$*  is computed, namely,  $\check{\mathbf{u}}_K^R = \mathbf{u}_K^R[\lambda^K]$  where  $\lambda^K = (\lambda_a^K, \lambda_b^K)$ . Table 2 below exhibits the results to *Prob.  $I_{cK}$*  for the first pair  $(\theta_r, \beta_S)$ .

| $\check{\mathcal{J}}_K(\check{\mathbf{u}}_K^R; \rho_F)$ | $\varphi_D^K(\lambda^K)$ | $\ \check{\mathbf{u}}_K^R\ _2$ | $\ \check{\mathbf{u}}_K^R\ _\infty$ | $\ \mathcal{T}_\theta^K[\check{\mathbf{u}}_K^R] - \theta_{ro}^R\ _2$ |
|---|--------------------------|--------------------------------|-------------------------------------|--|
| 168.2210  | 167.0747                 | 10.5405                        | 30                                  | 0.1690   |

Table 2: Constrained problem for the first pair  $(\theta_r, \beta_S)$ ,  $\rho_F = 2000$ .

Recall that  $\varphi_D^K(\lambda^K)$  is a lower bound on the optimal value of *Prob.  $I_{cK}$*  and that  $\check{\mathbf{u}}_K^R$  is a feasible solution for it. Thus, as shown in Table 3,  $\check{\mathcal{J}}_K(\check{\mathbf{u}}_K^R)$  does not exceed the optimal value of *Prob.  $I_{cK}$*  (say  $\mathcal{J}_{cK}^o$ ) by more than 1.15 (or by 0.7% of  $\mathcal{J}_{cK}^o$ ) – thus,  $\check{\mathbf{u}}_K^R$  can be taken to be an "approximately - optimal" solution to *Prob.  $I_{cK}$* .

Figure 3 displays the plots of  $\check{\mathbf{u}}_K^R$  and  $\mathbf{u}_K$ . Figure 4 exhibits the plots of  $\theta_{ro}^K$  (the projection of  $\theta_{ro}$  on  $\text{span}\{\phi_1, \dots, \phi_K\}$ , in green),  $\hat{\theta}_K \triangleq \mathcal{T}_\theta^K[\mathbf{u}_K]$  (dashed blue) and  $\hat{\theta}_K^R \triangleq \mathcal{T}_\theta^K[\check{\mathbf{u}}_K^R]$  (in red).

To illustrate the role of  $\rho_F$  in getting better approximation of the desired final state, numerical results were obtained for the same pair  $(\theta_r, \beta_S)$  with  $\rho_F = 4000$ . The results are presented in Tables 3, 4 and Figures 5 and 6.

| $\check{\mathcal{J}}_K(\mathbf{u}_K; \rho_F)$ | $\ \mathbf{u}_K\ _2$ | $\ \mathbf{u}_K\ _\infty$ | $\ \mathcal{T}_\theta^K[\mathbf{u}_K] - \boldsymbol{\theta}_{ro}^K\ _2$ |
|---|----------------------|---------------------------|---|
| 187.54639                                     | 12.3752              | 46.5118                   | 0.0926  |

Table 3: Unconstrained problem for the first pair  $(\theta_r, \boldsymbol{\beta}_S)$ ,  $\rho_F = 4000$ .

| $\check{\mathcal{J}}_K(\check{\mathbf{u}}_K^R; \rho_F)$ | $\varphi_D^K(\boldsymbol{\lambda}^K)$ | $\ \check{\mathbf{u}}_K^R\ _2$ | $\ \check{\mathbf{u}}_K^R\ _\infty$ | $\ \mathcal{T}_\theta^K[\check{\mathbf{u}}_K^R] - \boldsymbol{\theta}_{ro}^R\ _2$ |
|---|---------------------------------------|--------------------------------|-------------------------------------|---|
| 211.2104  | 212.2948                              | 11.9634                        | 30                                  | 0.1305  |

Table 4: Constrained problem for the first pair  $(\theta_r, \boldsymbol{\beta}_S)$ ,  $\rho_F = 4000$ .

Comparing Tables 1 and 3, it can be noted that the increase in  $\rho_F$  from 2000 to 4000 brought about a decrease in the  $L_2(0, t_F)$ -norm of the approximation error on  $\text{span}\{\phi_1, \dots, \phi_5\}$  (from 0.1435 to 0.0926) at the expense of increases in both the  $L_2(0, t_F)$  and  $L_\infty(0, t_F)$  norms of  $\mathbf{u}_K$  (respectively, from 10.9233 to 12.3752 and from 43.5917 and 46.5118).

Similarly, in the case of constrained problems (Tables 2 and 4) it can be noted that the increase in  $\rho_F$  decreased the  $L_2(0, t_F)$ -norm of the “projected” approximation error obtained under “peak-value” constraint ( $\|\mathbf{u}\|_\infty \leq 30$ ) from 0.1690 (Table 2) to 0.1305 (Table 4). Note also that  $\mathbf{u}_K^R$  is “approximately optimal” as  $|\varphi_D^K(\boldsymbol{\lambda}^K) - \check{\mathcal{J}}_K(\mathbf{u}_K^R; 4000)|/\varphi_D^K(\boldsymbol{\lambda}^K) \approx 1.09/212.2948 \leq 0.5 \times 10^{-2}$ . In this case, due to numerical imprecision,  $\check{\mathcal{J}}_K(\cdot)$  is close to but smaller than  $\varphi_D^K(\boldsymbol{\lambda}^K)$ .

The plots of  $\mathbf{u}_K$  and  $\mathbf{u}_K^R$  and those of the corresponding approximations  $\hat{\theta}_K$  and  $\hat{\theta}_K^R$  of the desired final state are respectively displayed in Figures 5 and 6.

Numerical results were also obtained for the pair  $(\theta_r, \boldsymbol{\beta}_S)$  shown in Figures 7 and 8. First, an approximate solution  $\mathbf{u}_K$  was obtained for *Prob. I<sub>K</sub>* – see Table 5 for the values of its  $L_2(0, t_F)$  and  $L_\infty(0, t_F)$  norms and the corresponding values of the cost-functional and the  $L_2(0, 1)$  norm of the final-state error (projected on  $\text{span}\{\phi_1, \dots, \phi_K\}$ ).

| $\check{\mathcal{J}}_K(\mathbf{u}_K; \rho_F)$ | $\ \mathbf{u}_K\ _2$ | $\ \mathbf{u}_K\ _\infty$ | $\ \mathcal{T}_\theta^K[\mathbf{u}_K] - \boldsymbol{\theta}_{ro}^K\ _2$ |
|---|----------------------|---------------------------|---|
| 283.5120                                      | 13.5254              | 23.5491                   | 0.2242  |

Table 5: Unconstrained problem for the second pair  $(\theta_r, \boldsymbol{\beta}_S)$ ,  $\rho_F = 2000$ .

A numerical solution  $\check{\mathbf{u}}_K^R$  was then obtained for *Prob. I<sub>cK</sub>* with the prescribed upper limit  $\mu_{\mathbf{u}}$  on the  $L_\infty(0, t_F)$ -norm of  $\mathbf{u}$  being set at  $\mu_{\mathbf{u}} = 18$ . This was done along the same lines described above in connection with the first pair  $(\theta_r, \boldsymbol{\beta}_S)$ . Table 6 exhibits the corresponding assessment data for  $\check{\mathbf{u}}_K^R$ .

| $\check{\mathcal{J}}_K(\check{\mathbf{u}}_K^R; \rho_F)$ | $\varphi_D^K(\boldsymbol{\lambda}^K)$ | $\ \check{\mathbf{u}}_K^R\ _2$ | $\ \check{\mathbf{u}}_K^R\ _\infty$ | $\ \mathcal{T}_\theta^K[\check{\mathbf{u}}_K^R] - \boldsymbol{\theta}_{r_o}^R\ _2$ |
|---|---------------------------------------|--------------------------------|-------------------------------------|--|
| 300.2274  | 286.3859                              | 12.6191                        | 18.0000                             | 0.2655   |

Table 6: Constrained problem for the second pair  $(\theta_r, \boldsymbol{\beta}_S)$ ,  $\rho_F = 2000$ .

Note that  $\check{\mathcal{J}}_K(\check{\mathbf{u}}_K^R; \rho_F)$  may only exceed the optimal value  $\mathcal{J}_{cK}^o$  of *Prob. I<sub>cK</sub>* by less than 5% (of  $\mathcal{J}_{cK}^o$ ). Figures 9 and 10 respectively display the plots of  $\mathbf{u}_K$  (dashed blue) and  $\check{\mathbf{u}}_K^R$  and those of  $\boldsymbol{\theta}_{r_o}^K$  (the projection of  $\boldsymbol{\theta}_{r_o}$  on  $\text{span}\{\phi_1, \dots, \phi_K\}$ ),  $\check{\boldsymbol{\theta}}_K \triangleq \mathcal{T}_\theta^K[\mathbf{u}_K]$  (dashed blue) and  $\check{\boldsymbol{\theta}}_K^R \triangleq \mathcal{T}_\theta^K[\mathbf{u}_K^R]$ .

Results were also obtained for the second pair  $(\theta_r, \boldsymbol{\beta}_S)$  with  $\rho_F = 4000$ , as presented in Tables 7 and 8 and Figures 11 and 12

| $\check{\mathcal{J}}_K(\mathbf{u}_K; \rho_F)$ | $\ \mathbf{u}_K\ _2$ | $\ \mathbf{u}_K\ _\infty$ | $\ \mathcal{T}_\theta^K[\mathbf{u}_K] - \boldsymbol{\theta}_{r_o}^K\ _2$ |
|---|----------------------|---------------------------|--|
| 362.0183                                      | 15.3659              | 26.4600                   | 0.1774   |

Table 7: Unconstrained problem for the second pair  $(\theta_r, \boldsymbol{\beta}_S)$ ,  $\rho_F = 4000$ .

| $\check{\mathcal{J}}_K(\check{\mathbf{u}}_K^R; \rho_F)$ | $\varphi_D^K(\boldsymbol{\lambda}^K)$ | $\ \check{\mathbf{u}}_K^R\ _2$ | $\ \check{\mathbf{u}}_K^R\ _\infty$ | $\ \mathcal{T}_\theta^K[\check{\mathbf{u}}_K^R] - \check{\boldsymbol{\theta}}_{r_o}^R\ _2$ |
|---|---------------------------------------|--------------------------------|-------------------------------------|--|
| 387.3645  | 387.2568                              | 14.7342                        | 18                                  | 0.2063   |

Table 8: Constrained problem for the second pair  $(\theta_r, \boldsymbol{\beta}_S)$ ,  $\rho_F = 4000$ .

Again, it can be noted that increasing  $\rho_F$  brings about a better approximation to the desired final state. Note also that  $|\varphi_D^K(\boldsymbol{\lambda}^K) - \check{\mathcal{J}}_K(\mathbf{u}_K^R; 4000)|/\varphi_D^K(\boldsymbol{\lambda}^K) \approx 0.11/387.2568 \leq 0.03 \times 10^{-2}$  and hence  $\check{\mathbf{u}}_K^R$  can be regarded as ‘‘approximately optimal’’ for the constrained problem.

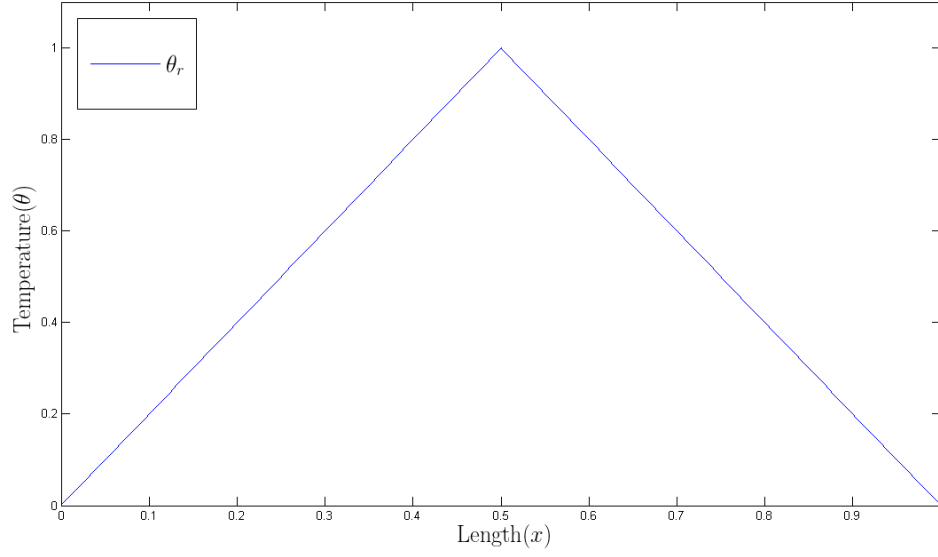


Figure 1: Example 1.  $\theta_r$ : target final state.

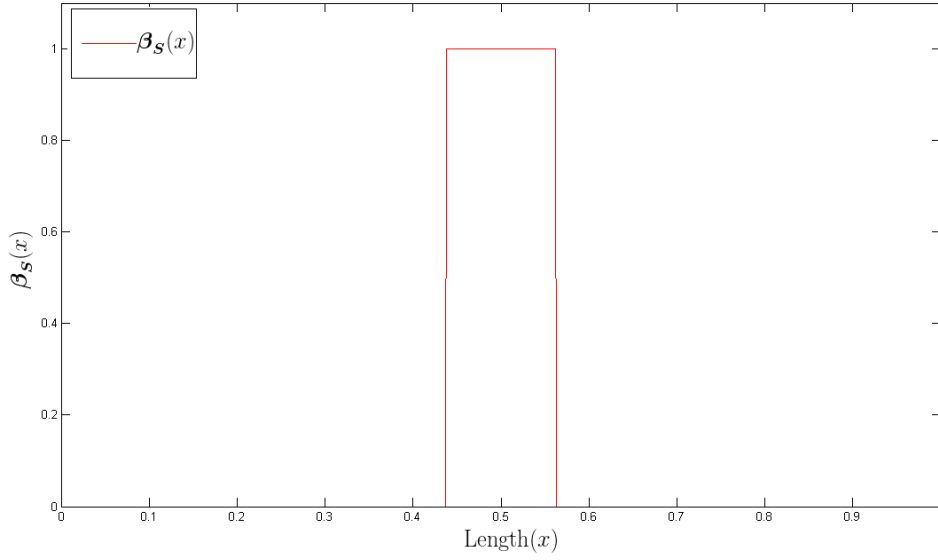


Figure 2: Example 1.  $\beta_S$ : control-to-state actuator.

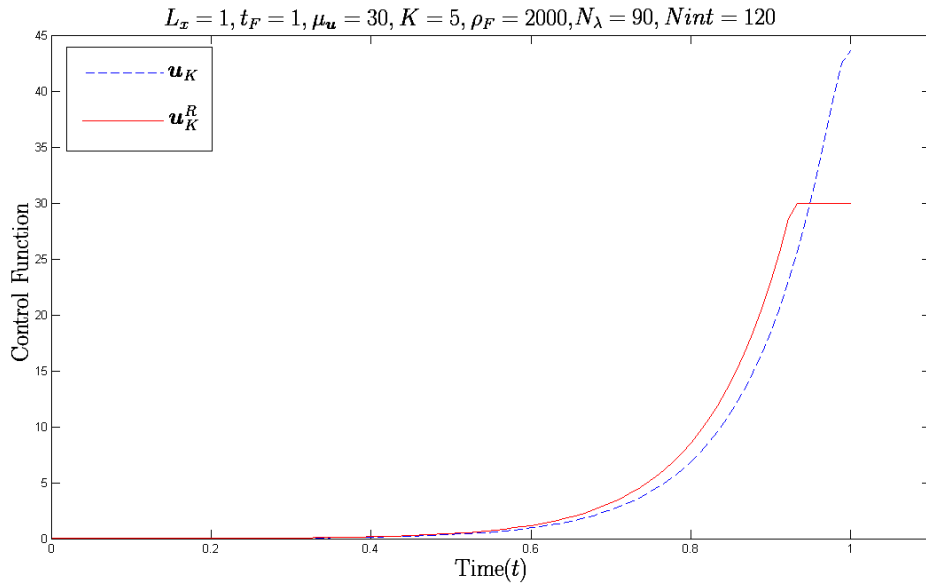


Figure 3: Example 1. Control signals  $\mathbf{u}_K$  (blue dashed),  $\mathbf{u}_K^R$  (red solid) for  $\rho_F = 2000$ .

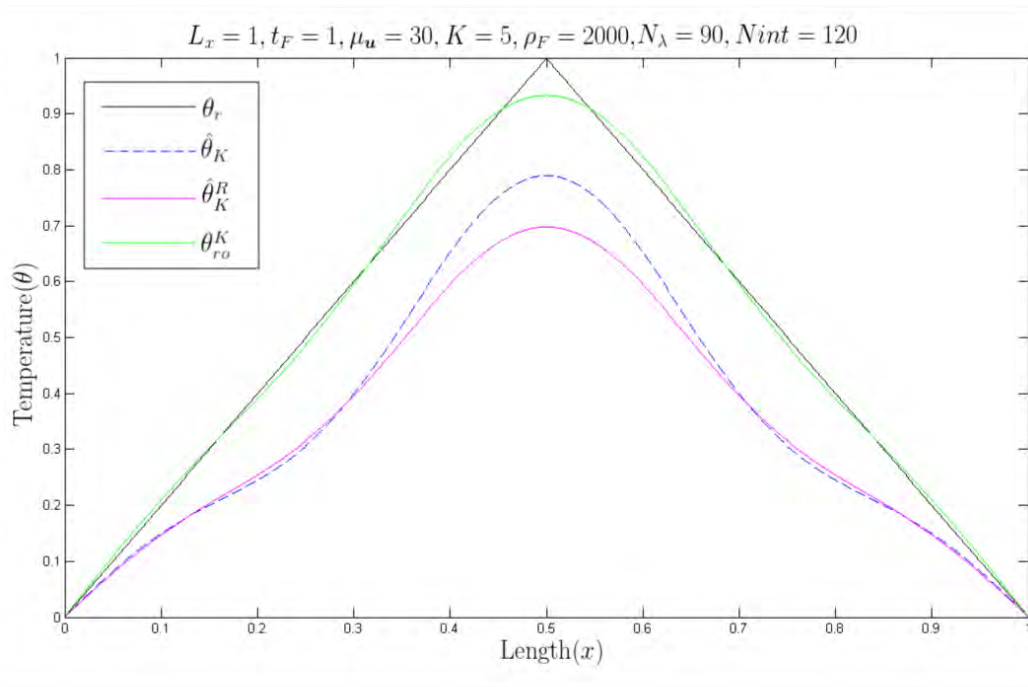


Figure 4: Example 1. Approximations to target final state for  $\rho_F = 2000$ .

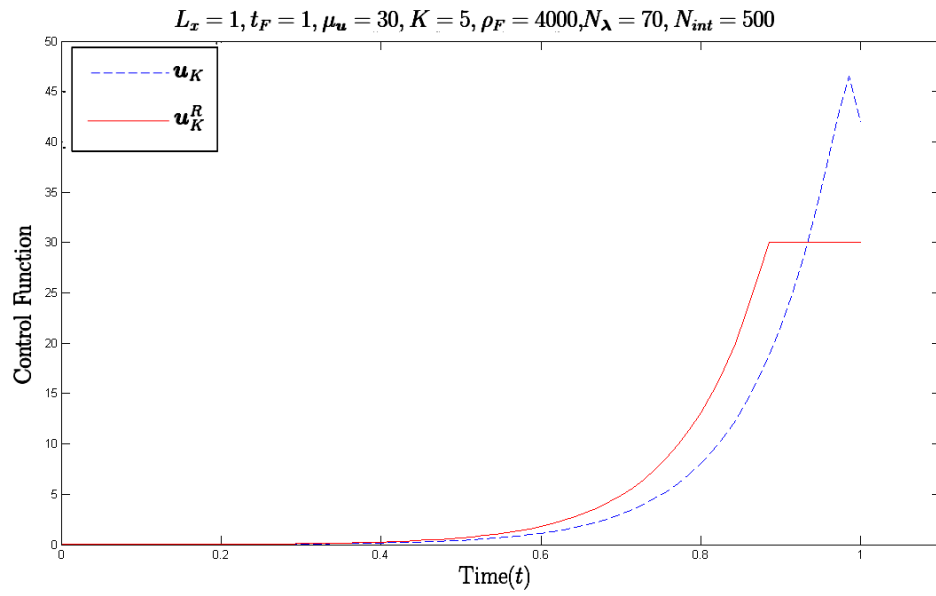


Figure 5: Example 1. Control signals  $\mathbf{u}_K$  (blue dashed),  $\mathbf{u}_K^R$  (red solid) for  $\rho_F = 4000$ .

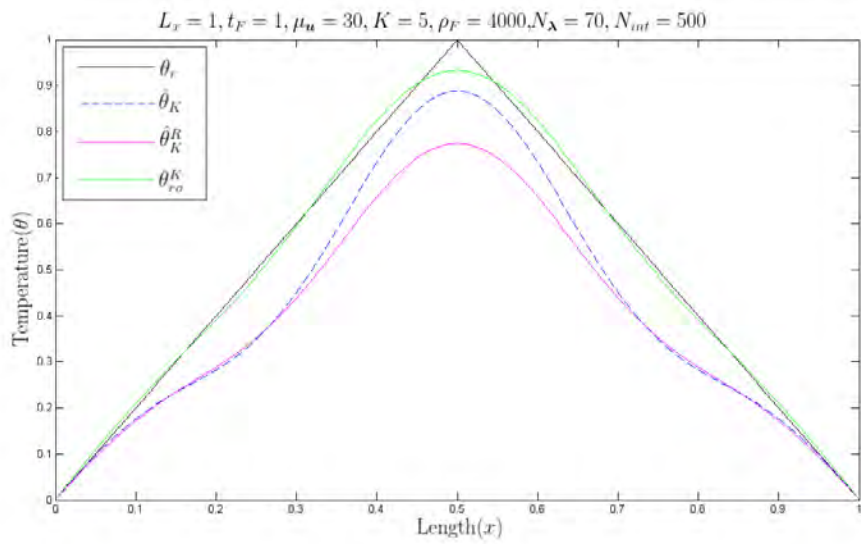


Figure 6: Example 1. Approximations to target final state for  $\rho_F = 4000$ .



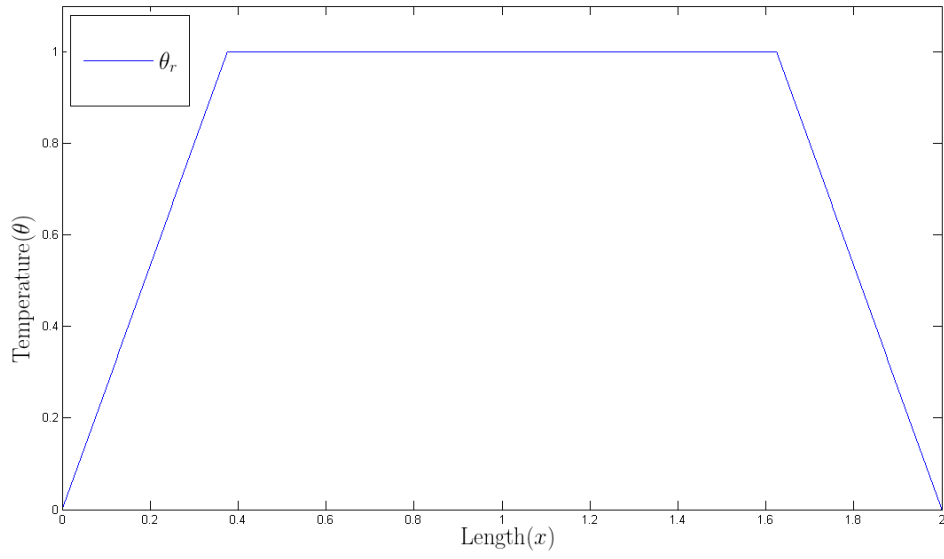


Figure 7: Example 2.  $\theta_r$ : target final state.

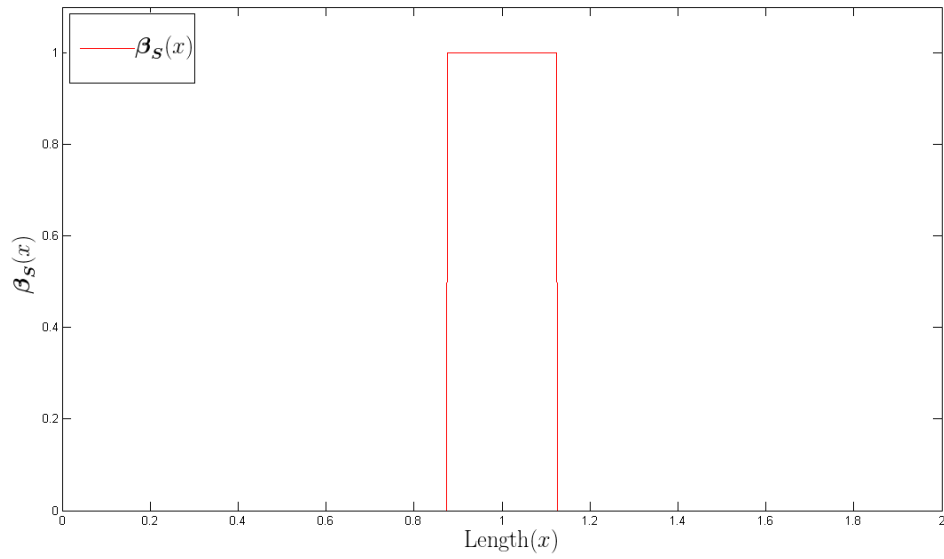


Figure 8: Example 2.  $\beta_S$ : control-to-state actuator.

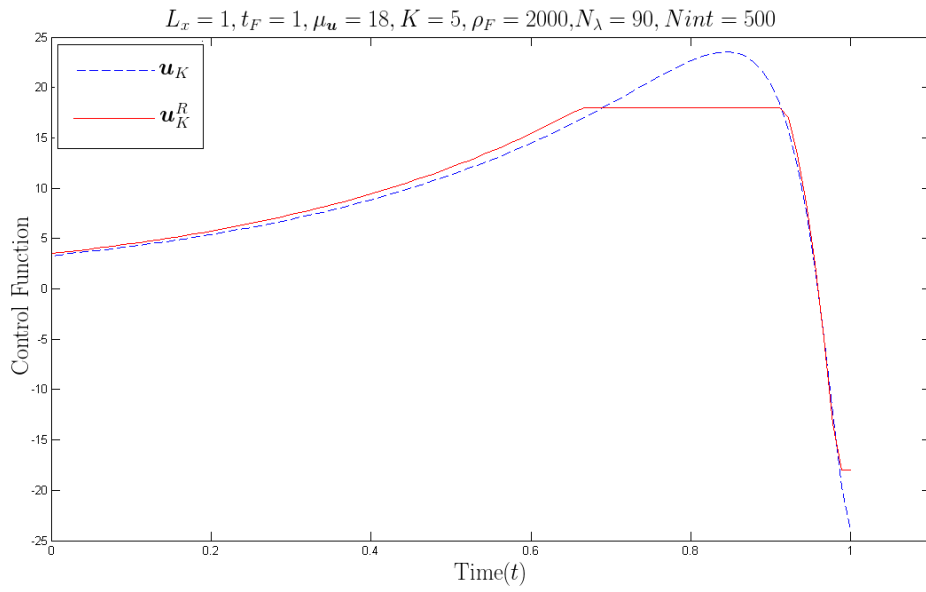


Figure 9: Example 2. Control signals  $u_K$  (blue dashed),  $u_K^R$  (red solid) for  $\rho_F = 2000$ .

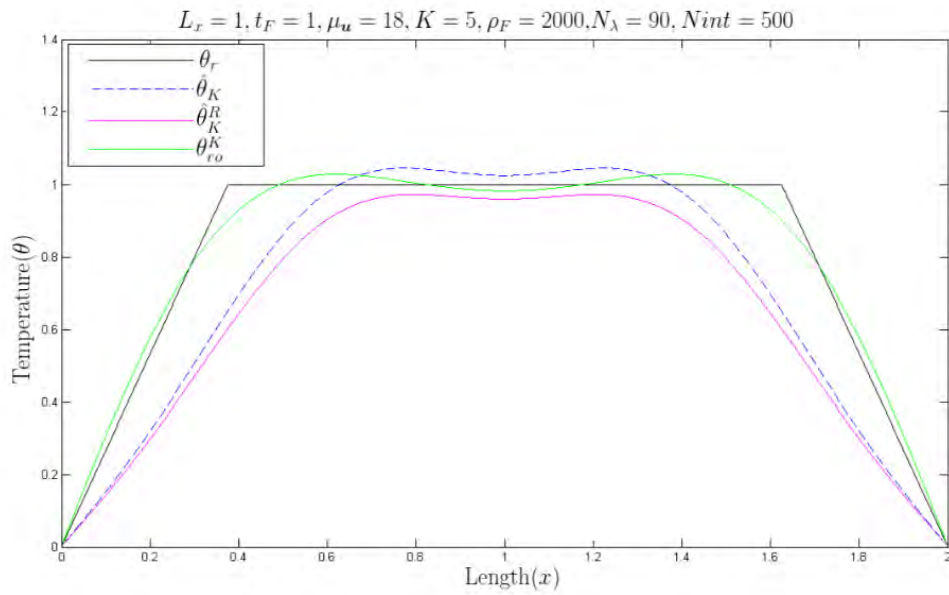


Figure 10: Example 2. Approximations to target final state for  $\rho_F = 2000$ .

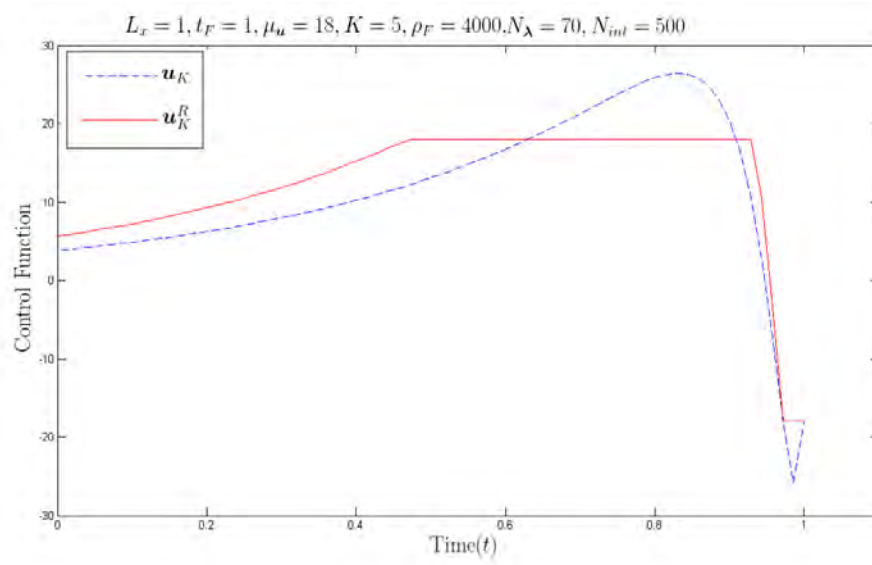


Figure 11: Example 2. Control signals  $\mathbf{u}_K$  (blue dashed),  $\mathbf{u}_K^R$  (red solid) for  $\rho_F = 4000$ .

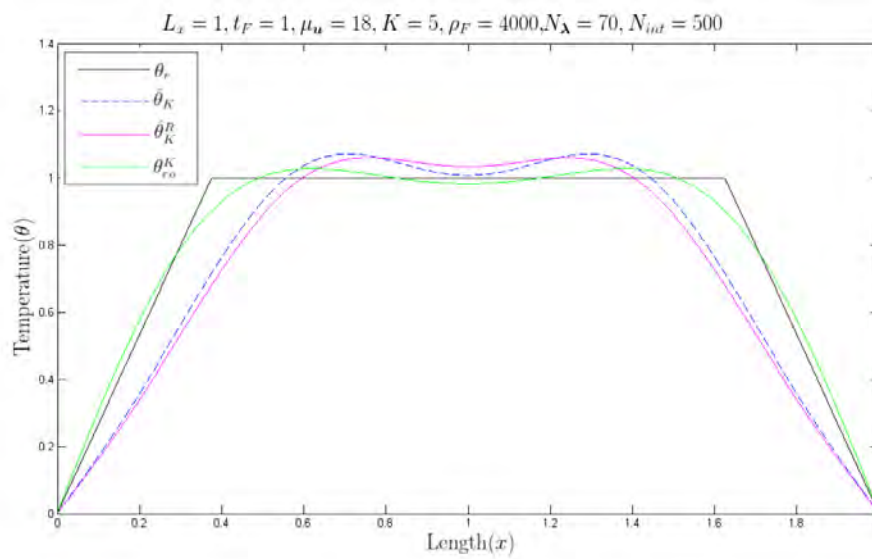


Figure 12: Example 2. Approximations to target final state for  $\rho_F = 4000$ .

Finally, the effect of the location of the “actuator”  $\beta_S$  on the final-state error  $\mathcal{T}_\theta^K[\mathbf{u}_c^K] - \theta_{ro}$  is illustrated by taking  $\beta_S$  to be centered on  $\ell_x \in (0, 2)$ , *i.e.*, by letting  $\beta_S$  to be given by  $\beta_S(x) = 1, \forall x \in (\ell_x - \delta_\beta, \ell_x + \delta_\beta), \beta_S(x) = 0$  otherwise, and computing the resulting  $\mathcal{T}_\theta^K[\mathbf{u}_c^K]$  for several values of  $\ell_x$  (with  $\delta_\beta = 0.1$ ), which are displayed in Figures 13 – 15, respectively for  $\ell_x = 3/10, \ell_x = 1$  and  $\ell_x = 2 - 3/10$ .

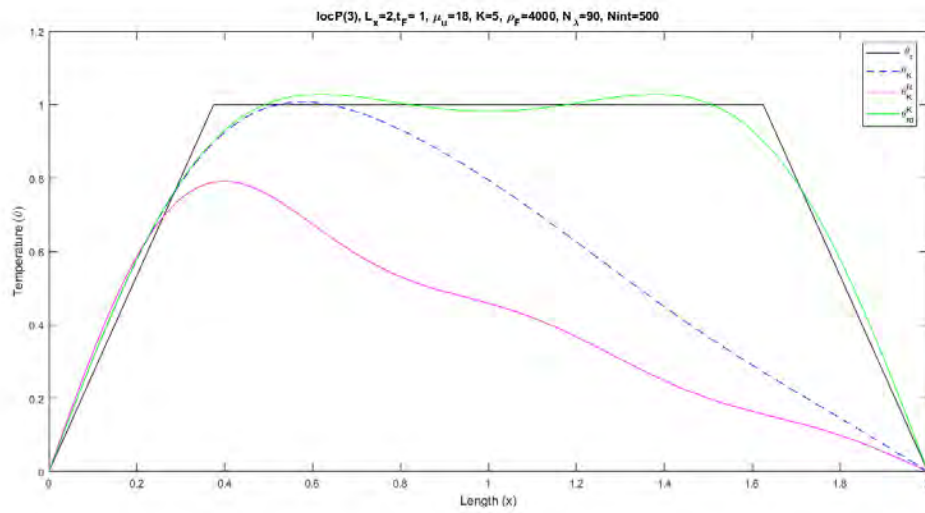


Figure 13: Example 2. Approximations to target final state for  $\rho_F = 4000, \ell_x = 3/10$ .

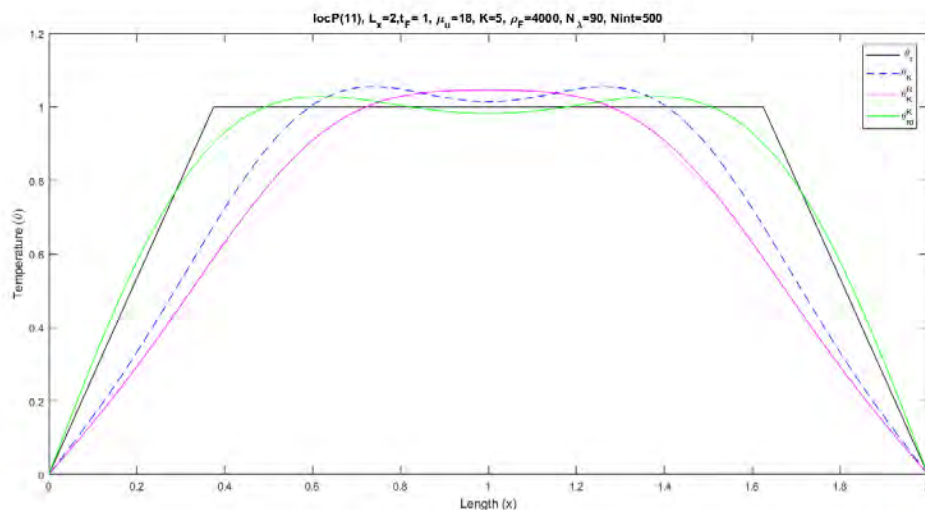


Figure 14: Example 2. Approximations to target final state for  $\rho_F = 4000, \ell_x = 1$ .

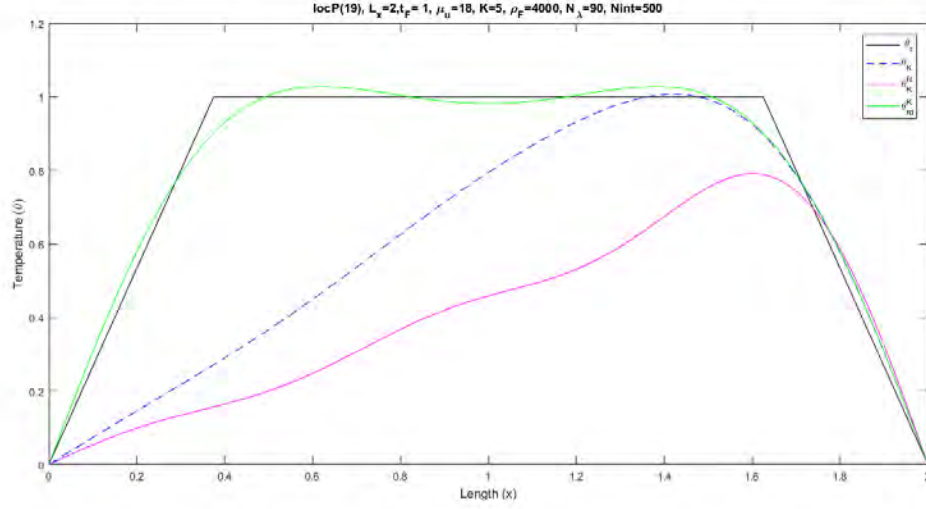


Figure 15: Example 2. Approximations to target final state for  $\rho_F = 4000$ ,  $\ell_x = 2 - 3/10$ .

### 4.3 A Two-Dimensional Example

An example is now presented of an initial/boundary-value problem defined by the heat equation on a rectangle in  $\mathbb{R}^2$ . More specifically, let  $\Omega = (0, L_x) \times (0, L_y)$ , where  $L_x, L_y \in \mathbb{R}_+$  and consider the following equation:

$$\forall (x, y) \in \Omega, \quad \frac{\partial \theta}{\partial t}(x, y, t) = k_\alpha \left\{ \frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2} \right\} (x, y, t) + \beta_{\mathcal{S}}(x, y) \mathbf{u}(t)$$

with zero initial conditions, *i.e.*,  $\forall (x, y) \in \Omega$ ,  $\theta(x, y, 0) = 0$  and homogeneous Dirichlet boundary conditions, *i.e.*,

$$\forall t \in [0, t_F], \quad \forall (x, y) \in \partial\Omega, \quad \theta(x, y, t) = 0,$$

where  $\mathbf{u} : [0, t_F] \rightarrow \mathbb{R}$  and  $\beta_{\mathcal{S}} : \Omega \rightarrow \mathbb{R}$ .

The corresponding weak, “ $K$ -th order”, Galerkin version is given by  $\forall k = 1, \dots, K$ ,

$$\left\langle \frac{\partial \theta}{\partial t}(\cdot, \cdot, t), \phi_k \right\rangle = -k_\alpha \left\{ \left\langle \frac{\partial \theta}{\partial x}(\cdot, \cdot, t), \frac{\partial \phi_k}{\partial x} \right\rangle + \left\langle \frac{\partial \theta}{\partial y}(\cdot, \cdot, t), \frac{\partial \phi_k}{\partial y} \right\rangle \right\} + \beta_{\mathcal{S}k} \mathbf{u}(t),$$

where  $i = 1, \dots, K_x$ ,  $j = 1, \dots, K_y$ ,  $k(i, j) = (i - 1)K_y + j$ ,  $K = K_x K_y$ ,  $\phi_{k(i, j)}(x, y) = \phi_i^x(x) \phi_j^y(y)$ ,  $\phi_i^x(x) = \sqrt{\frac{2}{L_x}} \sin \left[ \frac{i\pi x}{L_x} \right]$ ,  $\phi_j^y(y) = \sqrt{\frac{2}{L_y}} \sin \left[ \frac{j\pi y}{L_y} \right]$ .

As in the previous example, control signals  $\mathbf{u}_K$  and  $\mathbf{u}_c^K$  are sought by means of the problems

$$\underline{Prob. I_K} : \min_{\mathbf{u} \in L_2(0, t_F)} \check{\mathcal{J}}_K(\mathbf{u}; \rho_F) \quad \text{and} \quad \underline{Prob. I_{cK}} : \min_{\mathbf{u} \in S_{\mathbf{u}F}} \check{\mathcal{J}}_K(\mathbf{u}; \rho_F),$$

where  $\check{\mathcal{J}}_K(\mathbf{u}; \rho_F) = \|\mathbf{u}\|_{L_2(0, t_F)}^2 + \rho_F \|\mathcal{T}_\theta^K[\mathbf{u}] - \theta_{ro}\|_2^2$ ,  $\mathcal{T}_\theta^K[\mathbf{u}] = \sum_{k=1}^K c_k(t_F; \mathbf{u}) \phi_k$ ,  $\theta_r$  is the final state to be ‘‘approximately reached’’ and, as before,  $\bar{\mathbf{c}}_K(t; \mathbf{u}) = [c_1(t; \mathbf{u}) \cdots c_K(t; \mathbf{u})]^T$  is given by  $\bar{\mathbf{c}}_K(t; \mathbf{u}) = \int_0^t \mathbf{F}_K(\tau)^T \mathbf{u}(\tau) d\tau$  with  $\mathbf{F}_K$  as in (2.25). In this case,

$$\mathbf{A}_K = \text{diag}\{a_k : k = k(1, 1), \dots, k(1, K_y), k(2, 1), \dots, k(2, K_y), \dots, k(K_x, 1), \dots, k(K_x, K_y)\},$$

where  $a_{k(i,j)} = -k_\alpha \left\{ \left[ \frac{i\pi}{L_x} \right]^2 + \left[ \frac{j\pi}{L_y} \right]^2 \right\}$ ,  $\mathbf{M}_\beta^K = [\langle \beta_S, \phi_1 \rangle \cdots \langle \beta_S, \phi_k \rangle]^T$ , and  $S_{\mathbf{u}F} = \{\mathbf{u} \in L_2(0, t_F) : \text{a.e.}, |\mathbf{u}(t)| \leq \mu_{\mathbf{u}}\}$ .

Note that  $\check{\mathcal{J}}_K(\mathbf{u}; \rho_F) = \|\mathbf{u}\|_{L_2(0, t_F)}^2 + \rho_F \|\mathcal{T}_\theta^K[\mathbf{u}] - \theta_{ro}^K\|_{L_2(\Omega)}^2 + \|\theta_{ro} - \theta_{ro}^K\|_{L_2(\Omega)}^2$ , where  $\theta_{ro}^K$  is the orthogonal projection of  $\theta_{ro}$  on the span of  $\{\phi_1, \dots, \phi_K\}$ .

The numerical results shown in Tables 9 – 12 were obtained with the following problem data:  $k_\alpha = 1$ ,  $L_x = L_y = 1$ ,  $t_F = 1$ ,  $\rho_F = 8000$  and  $20000$ ,  $\mu_{\mathbf{u}} = 100$ ,  $K_x = K_y = 5$ ,  $\theta_r(x, y) = 0 \ \forall (x, y) \in \partial\Omega$ ,  $\theta_r(x, y) = 2 \ \forall (x, y) \in [L_x/10, 9L_x/10] \times [L_y/10, 9L_y/10]$ , the graph of  $\theta_r$  is the frustum of a rectangular pyramid with  $[0, L_x] \times [0, L_y]$  as basis,  $\|\theta_{ro}^K\|_2 = 1.7289$  and  $\beta_S$  is given by  $\begin{cases} \beta_S = 1 & \text{for } (x, y) \in [L_x/4, 3L_x/4] \times [L_y/4, 3L_y/4] \\ \beta_S = 0 & \text{otherwise} \end{cases}$ .

| $\check{\mathcal{J}}_K(\mathbf{u}_K; \rho_F)$ | $\ \mathbf{u}_K\ _2$ | $\ \mathbf{u}_K\ _\infty$ | $\ \mathcal{T}_\theta^K[\mathbf{u}_K] - \theta_{ro}^K\ _2^2$ |
|---|----------------------|---------------------------|--|
| 4978.00                                       | 45.6636              | 192.5735                  | 0.6037   |

Table 9: Unconstrained problem with  $\rho_F = 8000$ .

| $\check{\mathcal{J}}_K(\mathbf{u}_c^K; \rho_F)$ | $\varphi_D^K(\lambda^K)$ | $\ \mathbf{u}_c^K\ _2$ | $\ \mathbf{u}_c^K\ _\infty$ | $\ \mathcal{T}_\theta^K[\mathbf{u}_c^K] - \theta_{ro}^K\ _2^2$ |
|---|--------------------------|------------------------|-----------------------------|--|
| 5668.10   | 5485.00                  | 33.0038                | 100                         | 0.7565   |

Table 10: Constrained problem with  $\rho_F = 8000$ .

| $\check{\mathcal{J}}_K(\mathbf{u}_K; \rho_F)$ | $\ \mathbf{u}_K\ _2$ | $\ \mathbf{u}_c^K\ _\infty$ | $\ \mathcal{T}_\theta^K[\mathbf{u}_K] - \boldsymbol{\theta}_{ro}^K\ _2^2$ |
|---|----------------------|-----------------------------|---|
| 8127.40                                       | 64.4017              | 265.37                      | 0.4485  |

Table 11: Unconstrained problem with  $\rho_F = 20000$ .

| $\check{\mathcal{J}}_K(\mathbf{u}_c^K; \rho_F)$ | $\varphi_D^K(\boldsymbol{\lambda}^K)$ | $\ \mathbf{u}_c^K\ _2$ | $\ \mathbf{u}_c^K\ _\infty$ | $\ \mathcal{T}_\theta^K[\mathbf{u}_c^K] - \boldsymbol{\theta}_{ro}^K\ _2^2$ |
|---|---------------------------------------|------------------------|-----------------------------|---|
| 12281.00  | 11195.00                              | 37.8125                | 100                         | 0.7366  |

Table 12: Constrained problem with  $\rho_F = 20000$ .

Similarly to the results in the case of a one-dimensional spatial domain, Tables 9 – 11 illustrate the effect of increasing  $\rho_F$  on the decrease of the approximation errors  $\|\mathcal{T}_\theta^K[\mathbf{u}_K] - \boldsymbol{\theta}_{ro}^K\|_2$  (from 0.6037 in Table 9 to 0.4484 in Table 11) and  $\|\mathcal{T}_\theta^K[\mathbf{u}_c^K] - \boldsymbol{\theta}_{ro}^K\|_2$  (from 0.7565 in Table 10 to 0.7366 in Table 12). Note that in the latter case, increasing  $\rho_F$  from 8000 to 20000 had a small effect on the approximation error - this is due to the fact that the maximum magnitude of  $\mathbf{u}$  was kept at the same value ( $\mu_{\mathbf{u}} = 100$ ).

Again, as observed in the 1D-case, the “relatively small” difference between  $\varphi_D^K(\boldsymbol{\lambda}^K)$  and  $\check{\mathcal{J}}_K(\mathbf{u}_c^K; \rho_F)$  (3.2% for  $\rho_F = 8000$  and 8.8% for  $\rho_F = 20000$ ) indicates that  $\mathbf{u}_c^K$  is “nearly optimal” for the constrained problem - recall that  $\varphi_D^K(\boldsymbol{\lambda}^K)$  is a lower bound on  $\check{\mathbf{u}}_K(\mathbf{u}; \rho_F)$  for any  $\mathbf{u} \in S_{\mathbf{u}F}$ .

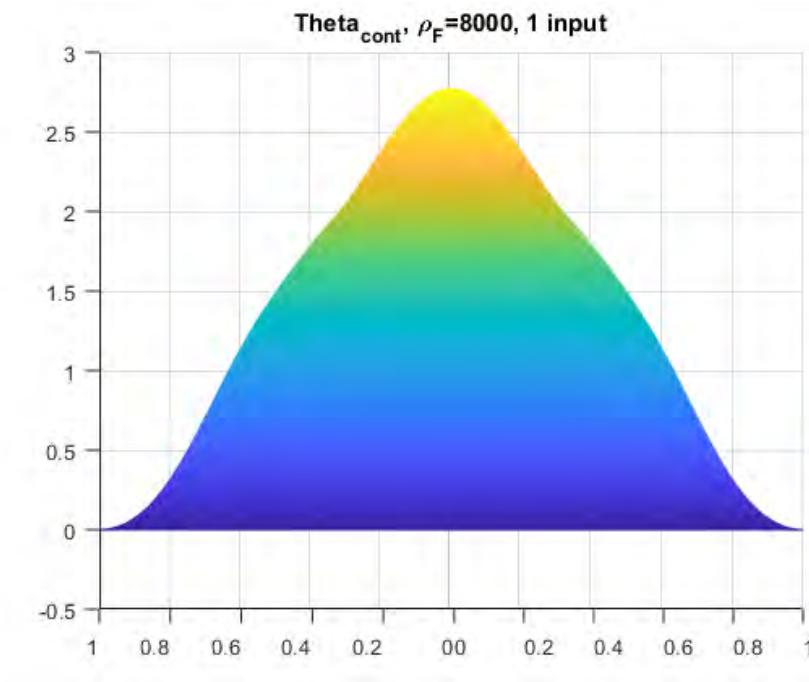


Figure 16: Transversal section of  $\mathcal{T}_\theta^K[\mathbf{u}_c^K]$  at  $\ell_x = \ell_y$  for  $\rho_F = 8000$ .

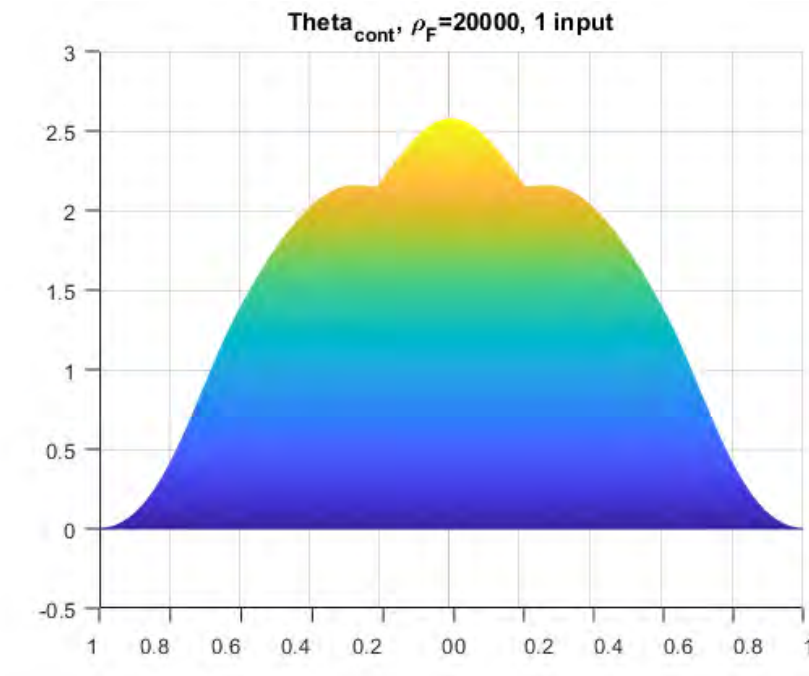


Figure 17: Transversal section of  $\mathcal{T}_\theta^K[\mathbf{u}_c^K]$  at  $\ell_x = \ell_y$  for  $\rho_F = 20000$ .



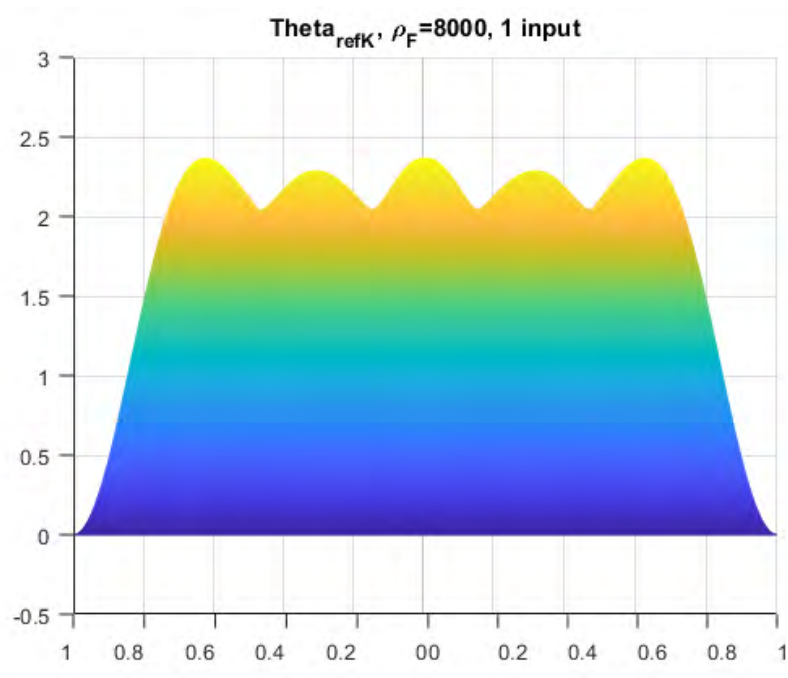


Figure 18: Transversal section of  $\theta_{ro}^K$ .

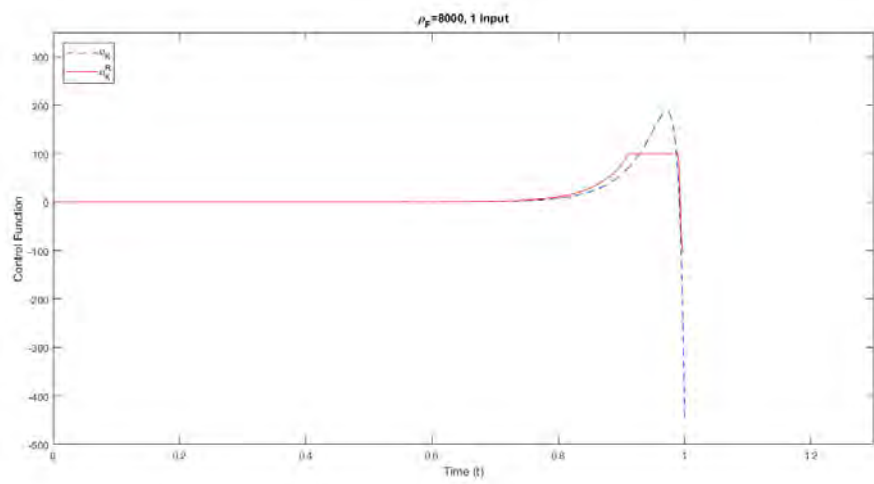


Figure 19: Graphs of  $u_K$  and  $u_c^K$  for  $\rho_F = 8000$ .

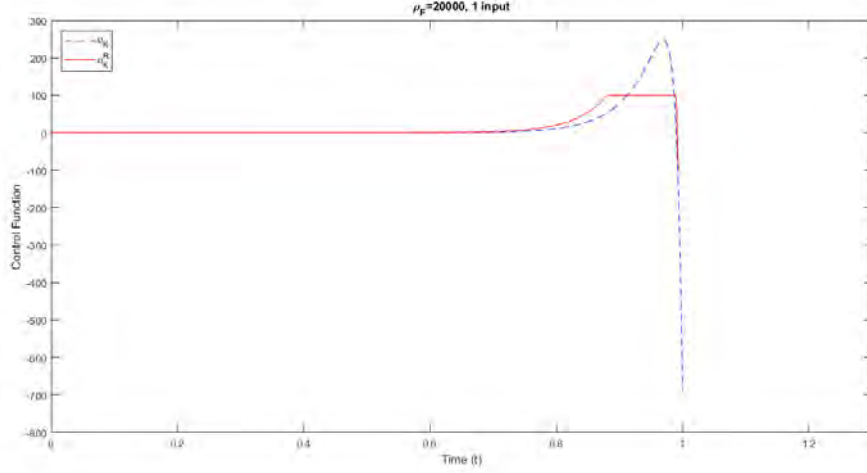


Figure 20: Graphs of  $\mathbf{u}_K$  and  $\mathbf{u}_c^K$  for  $\rho_F = 20000$ .

Figures 16 – 18 respectively display  $\mathcal{T}_\theta^K[\mathbf{u}_c^K]$ ,  $\theta_{ro}^K$ , transversal sections of the first two plot and Figures 19 – 20 display  $\mathbf{u}_K$  and  $\mathbf{u}_c^K$  for  $\rho_F = 8000$  and  $20000$ , respectively.

#### 4.4 Actuator Location

The initial/boundary value example defined by the heat equation on a rectangle  $(0, L_x) \times (0, L_y)$  in  $\mathbb{R}^2$  which was introduced above is now slightly modified to involve two scalar control signals ( $\mathbf{u}(t) \in \mathbb{R}^2$ ) and numerical results obtained searching the set of their possible “locations” will be presented.

More specifically, let the “source” term in the heat equation be given by

$$\beta_{\mathcal{S}}(x, y; \mathcal{X})\mathbf{u}(t) = \sum_{i=1}^2 \beta_{\mathcal{S}_i}(x, y; \mathcal{X}_i)u_i(t),$$

where  $\underline{\mathcal{X}} = (\mathcal{X}_1, \mathcal{X}_2)$ ,  $\mathcal{X}_i = (\mathcal{X}_i^x, \mathcal{X}_i^y) \in \mathbb{R}^2$  and  $\beta_{\mathcal{S}_i}(\cdot)$  is defined by

$$\begin{aligned} \beta_{\mathcal{S}_i}(x, y; \mathcal{X}_i) &= 1 & \forall (x, y) \in [\mathcal{X}_i^x - \delta_\beta, \mathcal{X}_i^x + \delta_\beta] \times [\mathcal{X}_i^y - \delta_\beta, \mathcal{X}_i^y + \delta_\beta] \\ \beta_{\mathcal{S}_i}(x, y; \mathcal{X}_i) &= 0 & \text{otherwise.} \end{aligned}$$

A location  $\underline{\mathcal{X}}$  will be assessed by the approximation error relative to the desired final state  $\theta_{r_o}^K$  achieved by the optimal (unconstrained) control over  $(0, t_F)$ , *i.e.*, by

$$\nu(\mathcal{X}_1, \mathcal{X}_2) = \|(\mathbf{I} + \rho_F \mathbf{G}_K(\mathbf{M}_\beta^K(\mathcal{X}_1, \mathcal{X}_2)))\bar{\theta}_{r_o}^K\|_2,$$

where  $\mathbf{M}_\beta^K$  and  $\mathbf{G}_K(\mathbf{M})$  are as in Section 3.2.

Two searches were carried with the following data:  $L_x = L_y = 1$ ,  $t_F = 1$ ,  $\rho_F = 8000$ ,  $\forall (x, y) \in (0, L_x) \times (0, L_y)$ ,  $\theta_r(x, y) = 2$ ,  $K = 5$ ,  $\delta_\beta = 0.1$ .

For the first one, a  $5 \times 5$  grid was defined by  $S_{gr} = \{0.1, 0.3, 0.5, 0.7, 0.9\}$  as  $S_{gr} \times S_{gr}$  and the set of possible locations  $S_{gr}^\mathcal{X} \triangleq \{(\mathcal{X}_1, \mathcal{X}_2) : \mathcal{X}_i \in (S_{gr} \times S_{gr}), i = 1, 2\}$  (comprised of 625 “locations”) was exhaustively searched. The minimum of  $\nu(\cdot, \cdot)$  on  $S_{gr}^\mathcal{X}$  was found to be 1.2563 and it was attained at the location  $\left(\begin{bmatrix} 0.5 \\ 0.7 \end{bmatrix}, \begin{bmatrix} 0.5 \\ 0.3 \end{bmatrix}\right)$ .

For  $N_a > N_{\alpha\varepsilon}$ , a search was also carried out on a set of  $N_a$  pseudo-random samples of a constant pdf on  $S_{\mathcal{X}_2} = \{(\mathcal{X}_1, \mathcal{X}_2) : \mathcal{X}_i \in (0, L_x) \times (0, L_y), i = 1, 2\}$ , with  $\alpha$  and  $\varepsilon$  set to  $\alpha = \varepsilon = 10^{-2}$ ,  $N_{\alpha\varepsilon} = 2/\log(1/0.99) \approx 454.5454$ , so that  $N_a$  was taken to be 500.

The minimum of  $\nu(\cdot, \cdot)$  on the 500 pseudo-random samples was found to be 1.2562 and it was attained at the location  $\left(\begin{bmatrix} 0.2854 \\ 0.4170 \end{bmatrix}, \begin{bmatrix} 0.6641 \\ 0.5468 \end{bmatrix}\right)$ .

Finally, the case of three scalar control signals was considered in the same setting. With the same values for  $\alpha$ ,  $\varepsilon$  and  $N_a$  a search on pseudo-random samples of a constant pdf on  $S_{\mathcal{X}_3} = \{(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3) : \mathcal{X}_i \in (0, L_x) \times (0, L_y), i = 1, 2\}$  was carried out leading to the minimum value 1.1431 for  $\nu(\cdot, \cdot, \cdot)$  which was attained at the location

$$\mathcal{X}_1 = \begin{bmatrix} 0.2952 \\ 0.4485 \end{bmatrix}, \quad \mathcal{X}_2 = \begin{bmatrix} 0.7628 \\ 0.2222 \end{bmatrix}, \quad \mathcal{X}_3 = \begin{bmatrix} 0.7064 \\ 0.8012 \end{bmatrix}.$$

## 5 FINITE-STATE, APPROXIMATE CONTROL OF THE NLHEQ: A HEURISTIC SCHEME BASED ON LINEARIZATION

In this chapter, a finite-state control problem will be considered in connection with the NLHEq subject to initial-value and homogeneous, Dirichlet boundary conditions. A heuristic scheme will be introduced which essentially consists of linearizing a finite-dimensional (semi-discrete) Galerkin approximation to the NLHEq, on “small” sub-intervals of  $(0, t_F)$ , and then computing the quadratically-optimal control functions for each of the corresponding linear, finite-dimensional, ordinary differential equations.

The main motivation for such a heuristic scheme rests on certain features of the non-linear problem which makes it difficult to compute approximate solutions on the basis of other possible approaches to this problem. These topics will be briefly discussed before the heuristic scheme in question is described in detail. Finally, after such a description is presented, numerical examples involving one-dimensional spatial domains will be presented at the end of this chapter.

To begin, it is recalled that the initial-value/boundary condition problem in question for the NLHEq is defined (in “weak” form) by

$$\forall \phi \in H_0^1(0, L_x), \quad \left\langle \frac{d\theta}{dt}(t), \phi \right\rangle = - \left\langle \alpha(\underline{\theta}(t)) \frac{\partial \underline{\theta}(t)}{\partial x}, \frac{\partial \phi}{\partial x} \right\rangle + \langle \underline{f}_S(t), \phi \rangle + \sum_{i=1}^m \langle \beta_{S_i}, \phi \rangle \mathbf{u}_i(t) \quad (5.1)$$

$$\text{and} \quad \langle \underline{\theta}(0), \phi \rangle = \langle g, \phi \rangle. \quad (5.2)$$

Let the solution of (5.1)–(5.2) for a given triple  $(\underline{f}_S, g, \mathbf{u})$  be denoted by  $\underline{\theta}(\cdot; \mathbf{u})$  ( $\underline{f}_S$  and  $g$  will be omitted as they remain unaltered throughout this chapter). A quadratic criterion for the choice of  $\mathbf{u}(\cdot) = [u_1(\cdot) \dots u_m(\cdot)]^T$  is then defined (as before) by:

$$\mathcal{J}_a(\mathbf{u}) \equiv \|\underline{\theta}(t_F; \mathbf{u}) - \theta_{ref}\|_2^2 + \rho_{\mathbf{u}} \|\mathbf{u}\|_{L_2(0, t_F)^m}^2,$$

where  $\theta_{ref}$  is the desired final-state (to be approximately reached).

Final-state control problems with and without “peak-value” constraints on  $\mathbf{u}$  are defined as before by:

$$\underline{\text{Prob. a:}} \quad \min_{\mathbf{u} \in L_2(0, t_F)^m} \mathcal{J}_a(\mathbf{u}) \quad \text{and} \quad \underline{\text{Prob. ac:}} \quad \min_{\mathbf{u} \in S_{\mathbf{u}F}} \mathcal{J}_a(\mathbf{u}),$$

where  $S_{\mathbf{u}F} \triangleq \{\mathbf{u} \in L_2(0, t_F)^m : \forall i = 1, \dots, m, -\mu_i \leq \mathbf{u}_i(t) \leq \mu_i, \forall t \text{ a.e. in } (0, t_F)\}$ .

Difficulties in the way of obtaining an approximate solution even to the simpler, unconstrained *Prob. a.* are the fact the mapping  $\mathbf{u} \mapsto \underline{\theta}(t_F; \mathbf{u})$  is only implicitly characterized (by (5.1) and (5.2)) and that due to (5.1) being non-linear on  $\underline{\theta}$ , convexity of the cost-functional (with respect to  $\mathbf{u}$ ) could not be established; additionally, the space of possible  $\mathbf{u}(\cdot)$  is infinite-dimensional.

A first step towards more tractable problems would be to replace (5.1)–(5.2) by a  $K$ th-order Galerkin approximation (as described in Chapter 2) which would lead to  $\underline{\theta}_K(t; \mathbf{u}) = \sum_{k=1}^{n_K} c_k(t; \mathbf{u}) \phi_k$  where  $\underline{c}_K(\cdot; \mathbf{u}) : (0, t_F) \rightarrow \mathbb{R}^{n_K}$  is the solution of the non-linear, ordinary differential equation

$$\dot{\mathbf{z}}(t) = \mathbf{R}_K(\mathbf{z}(t)) \mathbf{z}(t) + \mathbf{f}_K^S(t) + \underline{\boldsymbol{\beta}}_{SK}^T \mathbf{u}(t), \quad \forall t > 0, \quad \mathbf{z}(0) = \underline{\mathbf{g}}_K, \quad (5.3)$$

where, with  $\mathbf{G}_K^\phi$  as in Chapter 2,  $\mathbf{f}_K^S(t) = (\mathbf{G}_K^\phi)^{-1} \check{\mathbf{f}}_K^S(t)$ ,  $\underline{\boldsymbol{\beta}}_K^S = (\mathbf{G}_K^\phi)^{-1} \underline{\boldsymbol{\beta}}_K^S$ ,  $\{\check{\mathbf{f}}_K^S(t)\}_\ell = \langle \underline{f}_S(t), \phi_\ell \rangle_\ell$  and  $\{\underline{\boldsymbol{\beta}}_{SK}^T\}_{k\ell} = \langle \boldsymbol{\beta}_{S_k}, \phi_\ell \rangle$ ,

$$\mathbf{R}_K = (\mathbf{G}_K^\phi)^{-1} \check{\mathbf{R}}_K(\cdot) \text{ and } \{\check{\mathbf{R}}_K(\mathbf{z}(t))\}_{\ell k} = - \left\langle \alpha_K(\mathbf{z}(t)) \frac{\partial \phi_k}{\partial x}, \frac{\partial \phi_\ell}{\partial x} \right\rangle. \quad (5.4)$$

The corresponding, truncated cost-functional would then be given by

$$\mathcal{J}_a^K(\mathbf{u}) \equiv \|\underline{\theta}_K(t_F; \mathbf{u}) - \theta_{ref}^K\|_2^2 + \rho_u \|\mathbf{u}\|_{L_2(0, t_F)^m}^2,$$

where  $\theta_{ref}^K$  is the orthogonal projection of  $\theta_{ref}$  on  $S_K$ , i.e.,  $\boldsymbol{\theta}_r^K = [\theta_{r1}^K \dots \theta_{rn_K}^K]^T$ ,  $\boldsymbol{\theta}_r^K = (\mathbf{G}_K^\phi)^{-1} \check{\boldsymbol{\theta}}_r^K$  and  $\check{\boldsymbol{\theta}}_r^K = [\langle \theta_{ref}, \phi_1 \rangle \dots \langle \theta_{ref}, \phi_{n_K} \rangle]^T$ . Thus,  $\mathcal{J}_a^K(\mathbf{u}) = (\underline{c}_K(t_F; \mathbf{u}) - \boldsymbol{\theta}_r^K)^T \mathbf{G}_K^\phi (\underline{c}_K(t_F; \mathbf{u}) - \boldsymbol{\theta}_r^K) + \rho_u \|\mathbf{u}\|_{L_2(0, t_F)^m}^2$ .

Again, the mapping  $\mathbf{u} \mapsto \underline{c}_K(t_F; \mathbf{u})$  is only implicitly characterized by the non-linear ordinary differential equation (5.3) whose solution is  $\underline{c}_K(\cdot, \mathbf{u}) : [0, t_F] \rightarrow \mathbb{R}^{n_K}$  so that a possible approach to the problem *Prob. a<sub>K</sub>*:  $\min_{\mathbf{u} \in L_2(0, t_F)^m} \mathcal{J}_a^K(\mathbf{u})$  would be to formally replace  $\underline{c}_K$  in  $\mathcal{J}_a^K(\mathbf{u})$  by  $\mathbf{z}$  and treat  $\mathbf{z}$  and  $\mathbf{u}$  as separate decision variables linked by the non-linear equality constraint (5.3) which could, in principle, be handled by means of Lagrange multipliers. More specifically, a Lagrangian functional  $Lag_{a_K}$  would be introduced as

$$Lag_{a_K}(\mathbf{u}, \mathbf{z}, \boldsymbol{\lambda}) = \check{\mathcal{J}}_a^K(\mathbf{u}, \mathbf{z}) + \left\langle \boldsymbol{\lambda}, \dot{\mathbf{z}}(t) - \left[ \mathbf{R}_K(\mathbf{z}(t)) \mathbf{z}(t) + \mathbf{f}_K^S(t) + \underline{\boldsymbol{\beta}}_{SK}^T \mathbf{u}(t) \right] \right\rangle_{L_2(0, t_F)^m},$$

where  $\check{\mathcal{J}}_a^K(\mathbf{u}, \mathbf{z}) = (\mathbf{z}(t_F) - \boldsymbol{\theta}_r^K)^T \mathbf{G}_K^\phi (\mathbf{z}(t_F) - \boldsymbol{\theta}_r^K) + \|\mathbf{u}\|_{L_2(0, t_F)^m}^2$ ,  $\boldsymbol{\lambda} : (0, t_F) \rightarrow \mathbb{R}^m$  is a

Lagrange multiplier and the first step towards obtaining necessary optimality conditions would amount to solving (for a given  $\boldsymbol{\lambda}$ ) the problem  $\min_{\mathbf{u}, \mathbf{z}} Lag(\mathbf{u}, \mathbf{z}, \boldsymbol{\lambda})$ .

However, the difficulties associated with lack of convexity also apply to this problem. This motivates the use of global optimization techniques (NEUMAIER, 2004). To make such an approach viable, finite-dimensional sets of possible control functions and Lagrange multipliers would need to be considered (*e.g.*, piecewise linear functions on a partition of  $(0, t_F)$  in sub-intervals). Still, such methods are potentially very costly in computational terms so that random search methods (such as outlined in Chapter 4, Section 3) may also be considered as a less computationally-demanding alternative. These methods could, in principle, also be used in connection with *Prob. a<sub>K</sub>* (without introducing Lagrange multipliers) but, in any case, to evaluate  $\mathcal{J}_a^K(\mathbf{u})$  for each candidate  $\mathbf{u}$  would require that the corresponding approximate solution of (5.3) is computed by some numerical integration scheme.

Given these difficulties, which apply even to the unconstrained problem (*Prob. a<sub>K</sub>*) above, a heuristic scheme based on the linearization of the non-linear ODE in (5.3) is now introduced. The basic idea is that, as  $\mathbf{z}(\cdot)$  varies continuously with time,  $\mathbf{R}_K(\mathbf{z}(t))$  could be taken as approximately constant on small sub-intervals  $\mathcal{I}_q = [t_{q-1}, t_q]$  of  $(0, t_F)$ , *i.e.*,  $\forall t \in \mathcal{I}_q \quad \mathbf{R}_K(\mathbf{z}(t)) \approx \mathbf{R}_K(\mathbf{z}(t_{q-1}))$ . Then replacing  $\mathbf{R}_K(\mathbf{z}(t))$  by  $\mathbf{R}_K(\mathbf{z}(t_{q-1}))$  in (5.3) a control signal  $\tilde{\mathbf{u}}_q : \mathcal{I}_q \rightarrow \mathbb{R}^m$  is computed solving a linear-quadratic problem (using the results of Chapter 3). With the control signal  $\tilde{\mathbf{u}}_q$  (acting over  $\mathcal{I}_q$ ) a non-linear initial value problem is numerically solved on  $\mathcal{I}_q$  for  $\mathbf{u}_q$  and the initial condition  $\mathbf{z}(t_{q-1})$  ( $\mathbf{z}(0) = \mathbf{g}_K$ ) thereby defining  $\mathbf{z}(t_q)$ . This is repeated for all sub-intervals making up  $[0, t_F]$  and the control signal over  $[0, t_F]$  is obtained as the “concatenation” of the various  $\tilde{\mathbf{u}}_q$ s.

More precisely, this leads to the heuristic scheme described in the sequel for obtaining an “acceptable solution” to *Prob. ac<sub>K</sub>*:  $\min_{\mathbf{u} \in S_{uF}} \mathcal{J}_a^K(\mathbf{u})$ .

### Linearization scheme for *Prob. ac<sub>K</sub>*:

- (a) Divide the interval  $[0, t_F]$  into  $N_{lin}$  sub-intervals  $\mathcal{I}_q = [t_{q-1}, t_q]$ , where  $t_q = q(t_F/N_{lin})$ , for  $q = 1, \dots, N_{lin}$ .
- (b) At each step  $q = 1, \dots, N_{lin}$  with  $\mathbf{u}_q : \mathcal{I}_q \rightarrow \mathbb{R}^m$  and for a given initial value  $\mathbf{z}_q^0$ .
  - (i) the following problem is solved (using the results of Chapter 3)

$$\begin{aligned} \underline{\text{Prob. } q} : \quad & \min_{\mathbf{u}_q \in L_2(\mathcal{I}_q)^m} \|\mathbf{u}_q\|_2^2 + \rho_F \|\mathbf{z}_a^q(t_q; \mathbf{u}_q) - \mathbf{z}_{ref}^q\|_2^2 \\ & \text{subject to: } \forall t \text{ a.e. in } \mathcal{I}_q, \quad -\mu_i(t) \leq \mathbf{u}_{qi}(t) \leq \mu_i(t), \end{aligned}$$

where  $\mathbf{z}_a^q(\cdot, \mathbf{u}_q)$  is the solution of the linear initial-value problem

$$\dot{\mathbf{z}}_a^q(t) = \mathbf{R}_K(\mathbf{z}_q^0)\mathbf{z}_a^q(t) + \mathbf{f}_K^S(t) + \underline{\boldsymbol{\beta}}_{SK}\mathbf{u}_q(t) \quad \forall t \in \mathcal{I}_q, \quad \mathbf{z}_a^q(t_{q-1}) = \mathbf{z}_q^0, \quad (5.5)$$

*i. e.*,  $\forall t \in \mathcal{I}_q$ ,

$$\mathbf{z}_a^q(t; \mathbf{u}_q) = \exp[\mathbf{A}_q(t - t_{q-1})]\mathbf{z}_q^0 + \int_{t_{q-1}}^t \exp[\mathbf{A}_q(t - \tau)] \{ \mathbf{f}_K^S(\tau) + \mathbf{B}_K\mathbf{u}_q(\tau) \} d\tau,$$

where  $\mathbf{A}_q = \mathbf{R}_K(\mathbf{z}_q^0)$ ,  $\mathbf{B}_K = \underline{\boldsymbol{\beta}}_{SK}^T$ ,  $\mathbf{z}_1^0 = \mathbf{z}(0) = \underline{\mathbf{g}}_k$ ,  $\mathbf{z}_q^0 = \hat{\mathbf{z}}(t_{q-1})$ ,  $q \geq 2$ ; for  $q \geq 1$ ,  $\hat{\mathbf{z}}(t_q)$  is an estimate of the solution of (5.3) driven by  $\mathbf{u}_q$  on  $\mathcal{I}_q$  from initial condition  $\mathbf{z}_q^0$ . The optimal solution obtained in step  $k$  is denoted by  $\check{\mathbf{u}}_q$ .

- (c) The estimate  $\hat{\mathbf{z}}(t_q)$  is obtained as an approximate solution (obtained with a chosen numerical method) of

$$\dot{\mathbf{z}}(t) = \mathbf{R}_K(\mathbf{z}(t))\mathbf{z}(t) + \mathbf{f}_K^S(t) + \mathbf{B}_K\check{\mathbf{u}}_q(t), \quad \forall t \in \mathcal{I}_q, \quad \mathbf{z}(t_{q-1}) = \mathbf{z}_q^0, \quad (5.6)$$

where  $\check{\mathbf{u}}_q$  is the solution of *Prob. q*.

- (d) The control signal  $\check{\mathbf{u}} : [0, t_F] \rightarrow \mathbb{R}^m$  obtained with the iterations above is given by  $\forall t \in (t_{q-1}, t_q]$ ,  $\check{\mathbf{u}}(t) = \check{\mathbf{u}}_q(t)$ ,  $\check{\mathbf{u}}(0) = \check{\mathbf{u}}_1(0)$ .

- (e) The final-state achieved with  $\check{\mathbf{u}}$  is given by  $\hat{\mathbf{z}}(t_F)$ .

Note that for sub-intervals of equal length (*i. e.*,  $(t_q - t_{q-1}) = \delta_{lin}$ ) the fact that the linearized model has constant coefficients enables the rewriting of the problem stated in **(i)** as:

$$\begin{aligned} \underline{\text{Prob. q}} : \quad & \min_{\mathbf{u}_a: [0, \delta_{lin}] \rightarrow \mathbb{R}^m} \|\mathbf{u}_a\|_{L_2(0, \delta_{lin})}^2 + \rho_F \|\mathcal{T}_q[\mathbf{u}_a] - \mathbf{z}_q^a\|_E^2 \\ & \text{subject to: } \forall t \in [0, \delta_a], \quad \mu_i \leq \mathbf{u}_{ai}(t) \leq \mu_i, \end{aligned}$$

where  $\mathbf{z}_q^a = \mathbf{z}_{ref} - \exp[\mathbf{A}_q\delta_{lin}]\mathbf{z}_q^0 - \int_0^{\delta_{lin}} \exp[\mathbf{A}_q(\delta_{lin} - \tau)] \mathbf{f}_K^S(t_{q-1} + \tau) d\tau$ ,  $\delta_a = t_F/N_{lin}$ , and  $\mathcal{T}_q[\mathbf{u}_a] = \int_0^{\delta_{lin}} \exp[\mathbf{A}_q(\delta_{lin} - \tau)] \mathbf{B}_K\mathbf{u}_a(\tau) d\tau$ .

Thus, the linearization scheme described above amounts to solving the linear quadratic problem, *Prob. q*, above for successive  $(\mathbf{z}_q^0, \mathbf{z}_q^a, \mathbf{A}_q)$  (using the algorithm described in the next section) and numerically integrating the non-linear ODE in (5.3) over

each sub-interval from initial state  $\mathbf{z}_q^0$  ( $\mathbf{z}_1^0 = \underline{\mathbf{g}}_K$ ) to obtain  $\mathbf{z}(t_q; \check{\mathbf{u}}_q, \mathbf{z}_q^0)$  - note that (5.3) is numerically solved over  $(0, t_F)$  only once (for the obtained  $\check{\mathbf{u}}$ ).

A summary is now presented of the computational steps required to obtain a control signal to steer the NLHEq towards a desired final state. Given  $(\alpha, g, f_S; \{\boldsymbol{\beta}_S\}, \theta_{ref})$  and a family of subspaces  $\{S_K\}$  with orthonormal bases  $\{\phi_1, \dots, \phi_K\}$ .

(1) Compute

$$\mathbf{B}_K = \boldsymbol{\beta}_{S_K}^T = \begin{bmatrix} \langle \boldsymbol{\beta}_{S_1}, \phi_1 \rangle & \cdots & \langle \boldsymbol{\beta}_{S_m}, \phi_1 \rangle \\ \vdots & & \vdots \\ \langle \boldsymbol{\beta}_{S_1}, \phi_K \rangle & \cdots & \langle \boldsymbol{\beta}_{S_m}, \phi_K \rangle \end{bmatrix},$$

$$\{\mathbf{B}_K\}_{\ell i} = \langle \boldsymbol{\beta}_{S_i}, \phi_\ell \rangle, \quad \underline{\mathbf{g}}_K^T = [\langle g, \phi_1 \rangle \cdots \langle g, \phi_K \rangle], \quad \mathbf{z}(0) = \underline{\mathbf{g}}_K = \mathbf{z}_1^0 \quad \text{and} \\ \mathbf{z}_{ref} = [\langle \theta_{ref}, \phi_1 \rangle \cdots \langle \theta_{ref}, \phi_K \rangle].$$

(2) Choose  $N_{lin}$  (and  $\mathcal{I}_q$ ) as in (a) above and for each step  $q = 1, \dots, N_{lin}$  (given  $\mathbf{z}_q^0$ ).

(2.1) Compute  $\mathbf{A}_q = \mathbf{R}_K(\mathbf{z}_q^0)$ ,  $\mathbf{z}_q^a = \mathbf{z}_{ref} - \exp[\mathbf{A}_q \delta_a] \mathbf{z}_q^0$ , where  $\mathbf{R}_K$  is given by (5.4).

(2.2) Solve *Prob. q*, thus obtaining  $\check{\mathbf{u}}_q(t) = \mathbf{u}_a(t_{q-1})$ ,  $t \in \mathcal{I}_q$ , using the procedure introduced in Chapter 3, for the linear case, appropriately setting the problem data in each step.

(2.3) Obtain  $\hat{\mathbf{z}}(t_q)$  as the numerical solution of (5.6) on  $\mathcal{I}_q$  with  $\mathbf{u}_q$  set to  $\check{\mathbf{u}}_q$  obtained as in (2.2), at  $t = t_q$  (update:  $\mathbf{z}_{q+1}^0 = \hat{\mathbf{z}}(t_q)$ ).

(3) In the end for  $\check{\mathbf{u}} : [0, t_F] \rightarrow \mathbb{R}^m$  defined as in (d), solve (5.3) numerically for the interval  $[0, t_F]$  with  $\mathbf{u}$  set to  $\check{\mathbf{u}}$ .

**Remark 5.1.** *The major computational tasks required in the procedure described above are solving the  $N_{lin}$  optimal control problems on  $\mathcal{I}_q$ ,  $q = 1, \dots, N_{lin}$  (see (2.2) above) and solving  $N_{lin}$  non-linear initial value problems over  $\mathcal{I}_q$  (see (2.3) above). The former are to be solved on the basis of Chapter 4 whereas the latter have been solved in the examples below by the 4th-order Runge-Kutta method (see APPENDIX C.2).*

▽



## 5.1 Numerical Examples

In this section, numerical results are presented to illustrate the use of the linearization scheme above in the computation of control signals which purport to steer the solution of the NLHEq towards a desired, final state (over the time-interval  $[0, t_F]$ ). The diffusion coefficient  $\alpha : \mathbb{R} \rightarrow \mathbb{R}$  which is a function of the temperature  $\theta$  is taken to be by

$$\alpha(\theta) = \alpha_1 + \alpha_2 \left( \frac{\gamma_1}{\gamma_1 + \gamma_2(\theta - \theta_c)^{\gamma_3}} \right),$$

where  $\alpha_1, \alpha_2, \gamma_1, \gamma_2,$  and  $\gamma_3$  are positive coefficients and  $\theta_c$  is the “origin of the temperature scale” over which  $\alpha(\cdot)$  as above is thought to be an acceptable model for how the diffusion coefficient varies with  $\theta$  – note that for  $\theta = \theta_c$ ,  $\alpha(\theta) = \alpha_1 + \alpha_2$  and that, as  $\theta$  gets much larger than  $\theta_c$ ,  $\alpha(\theta)$  tends to  $\alpha_1$  so that the overall range of variation of  $\alpha(\theta)$  for  $\theta \geq \theta_c$  is contained in  $(\alpha_1, \alpha_1 + \alpha_2)$ , with  $\alpha(\theta)$  decreasing as  $\theta$  increases. In accordance with the homogeneous Dirichlet boundary condition, the origin of the temperature scale is taken to be the boundary temperature so that  $\theta_c$  is set to zero in the examples below.

Each of the numerical experiments reported below correspond to

- (a) A pair of values assigned to the coefficients  $\alpha_1$  and  $\alpha_2$  (with  $\alpha_1 + \alpha_2 = 1$ ) (the coefficients  $\gamma_1, \gamma_3$  and  $\theta_c$  have the following values which are held fixed throughout all experiments  $\gamma_1 = 1, \gamma_3 = 2, \theta_c = 0$ ). For the experiments in Figures 25 – 28,  $\gamma_2$  is set to  $\gamma_2 = 10$ . For the remaining experiments  $\gamma_2$  is set to  $\gamma_2 = 1$ .
- (b) One of the two desired final states given in Figure 1 and Figure 7 of Chapter 4.
- (c) One of two possible values for  $\rho_F$ .
- (d) One of two possible values for  $\mu_{\mathbf{u}}$ .

As in the examples of Chapter 4, one of the aims here is to illustrate the effect of  $\rho_F$  and  $\mu_{\mathbf{u}}$  on the final-state approximation error as well the relative difficulties posed by different functions taken to be the desired final states. In addition, by varying  $\alpha_1$ , the effect of the “amount of non-linearity” involved on the final-state error can also be illustrated (note that as  $\alpha_1 + \alpha_2 = 1$ , the smaller  $\alpha_1$  the greater the departure of  $\alpha(\cdot)$  from a constant value).

In the following experiments the values assigned to  $\rho_F$ ,  $\mu_u$  and  $\gamma_1$  are  $\gamma_1 = 1.0$ ,  $\rho_F = 10000$ ,  $\rho_F = 40000$  and  $\mu_u = 30$ ,  $\mu_u = 70$ . Roughly speaking, in the examples below  $\theta(x, t)$  varies between zero and one. Thus,  $\alpha(\theta(\cdot))$  varies between  $\alpha_1 + \alpha_2$  and  $\alpha_1 + \alpha_2 \left( \frac{1}{1 + \gamma_2} \right)$  (with  $\alpha_1 + \alpha_2 = 1$ ), *i.e.*, the range of values of  $\alpha(\cdot)$  is  $\alpha_2 \left( \frac{\gamma_2}{1 + \gamma_2} \right)$ . As a result, the “amount of non-linearity” in the NLHEq would increase with  $\alpha_2$  and  $\gamma_2$ . First, the desired final state is given in Figure 1 from Chapter 4, which for convenience will be named  $\theta_{r1}$ .

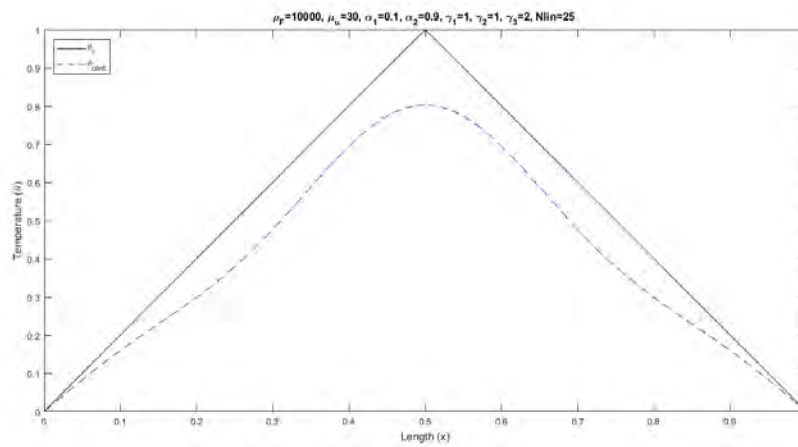


Figure 21:  $\rho_F = 10000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.1$ ,  $\alpha_2 = 0.9$ ,  $\gamma_2 = 1$ .

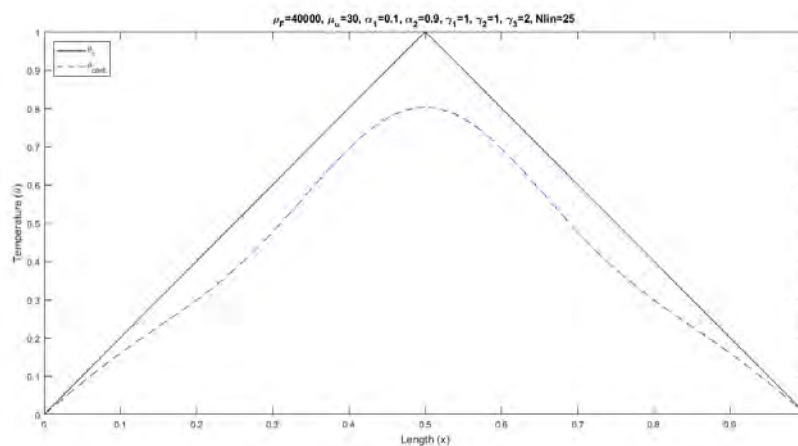


Figure 22:  $\rho_F = 40000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.1$ ,  $\alpha_2 = 0.9$ ,  $\gamma_2 = 1$ .

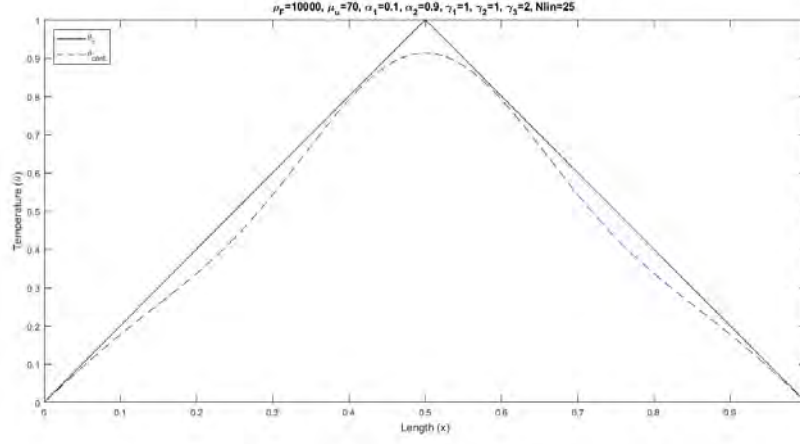


Figure 23:  $\rho_F = 10000$ ,  $\mu_u = 70$ ,  $\alpha_1 = 0.1$ ,  $\alpha_2 = 0.9$ ,  $\gamma_2 = 1$ .

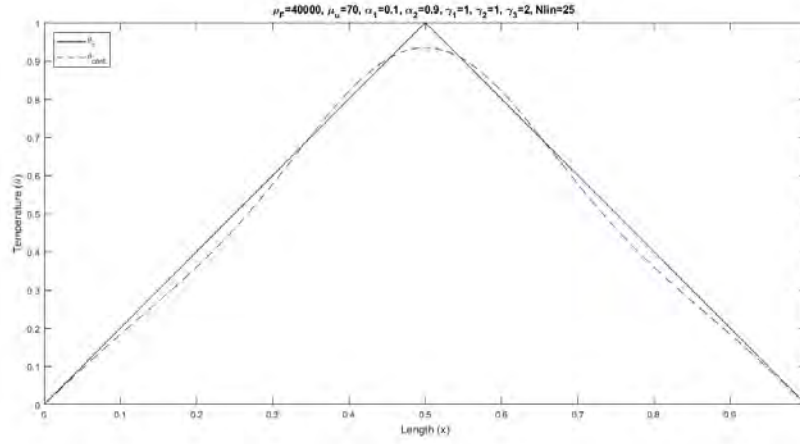


Figure 24:  $\rho_F = 40000$ ,  $\mu_u = 70$ ,  $\alpha_1 = 0.1$ ,  $\alpha_2 = 0.9$ ,  $\gamma_2 = 1$ .

| $\ e_K\ $        | $\mu_u = 30$ | $\mu_u = 70$ |
|------------------|--------------|--------------|
| $\rho_F = 10000$ | 0.1003       | 0.0398       |
| $\rho_F = 40000$ | 0.1003       | 0.0198       |

Table 13: Final state for  $\theta_{r1}$ . Approximation error norms on  $S_K$  for Figures 21 – 24 ( $\alpha_1 = 0.1$ ,  $\alpha_2 = 0.9$  and  $\gamma_2 = 1$ ).

Note that the error norms in  $S_K$ , (*i.e.*,  $\|z_{ref} - \hat{z}(t_F)\|_E$ , with  $z_{ref}$  and  $\hat{z}(t_F)$  as in (1) and (2.3) above) diminish with increasing  $\rho_F$  and  $\mu_u$  as happened in the case of the experiments with the LHEq.

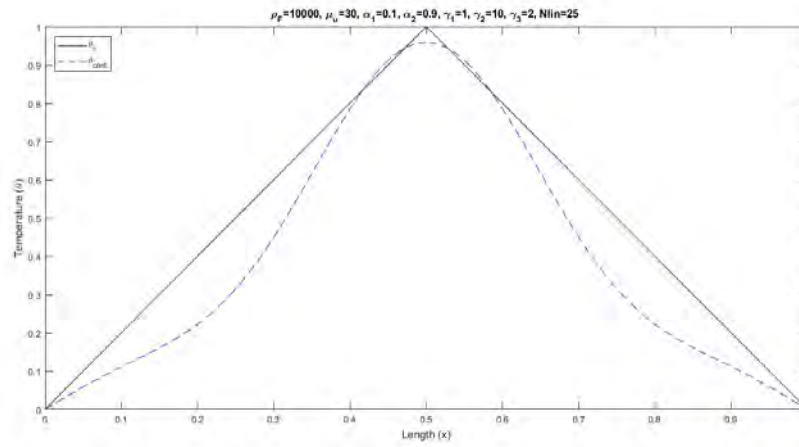


Figure 25:  $\rho_F = 10000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.1$ ,  $\alpha_2 = 0.9$ ,  $\gamma_2 = 10$ .

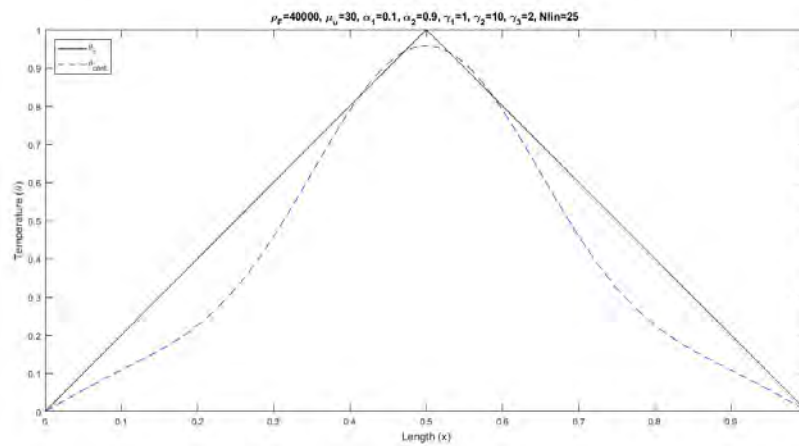


Figure 26:  $\rho_F = 40000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.1$ ,  $\alpha_2 = 0.9$ ,  $\gamma_2 = 10$ .

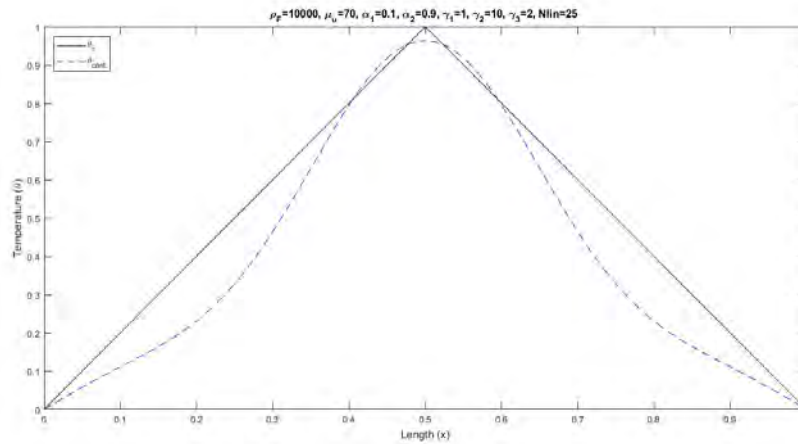


Figure 27:  $\rho_F = 10000$ ,  $\mu_u = 70$ ,  $\alpha_1 = 0.1$ ,  $\alpha_2 = 0.9$ ,  $\gamma_2 = 10$ .

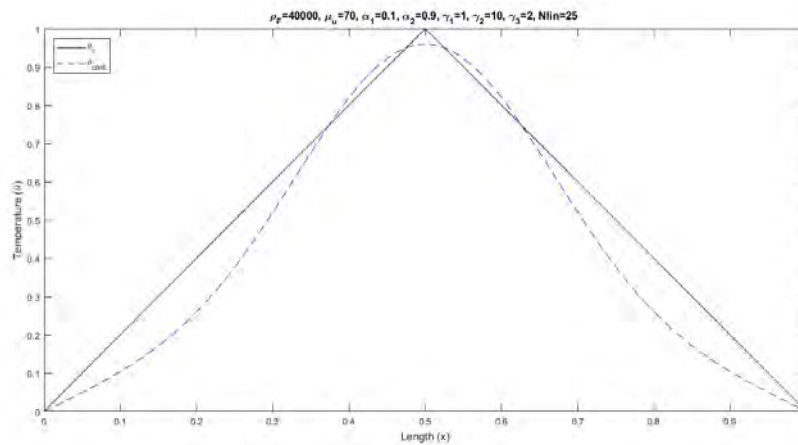


Figure 28:  $\rho_F = 40000$ ,  $\mu_u = 70$ ,  $\alpha_1 = 0.1$ ,  $\alpha_2 = 0.9$ ,  $\gamma_2 = 10$ .

| $\ e_K\ $        | $\mu_u = 30$ | $\mu_u = 70$ |
|------------------|--------------|--------------|
| $\rho_F = 10000$ | 0.1104       | 0.1035       |
| $\rho_F = 40000$ | 0.1072       | 0.0827       |

Table 14: Final state for  $\theta_{r1}$ . Approximation error norms on  $S_K$  for Figures 25 – 28 ( $\alpha_1 = 0.1$ ,  $\alpha_2 = 0.9$  and  $\gamma_2 = 10$ ).

In comparison with Table 13, the increase in  $\gamma_2$  (as remarked above, increasing the “amount of non-linearity”) brought about an increase in the approximation error for all values of  $\rho_F$  and  $\mu_u$  taken in account.

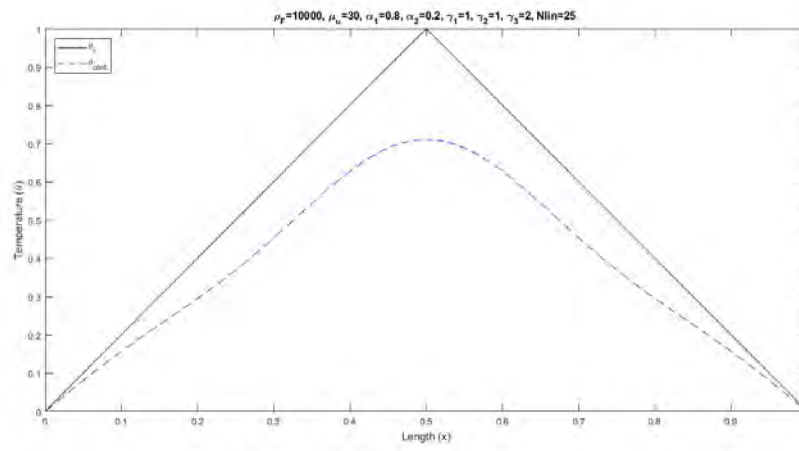


Figure 29:  $\rho_F = 10000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.8$ ,  $\alpha_2 = 0.2$ ,  $\gamma_2 = 1$ .

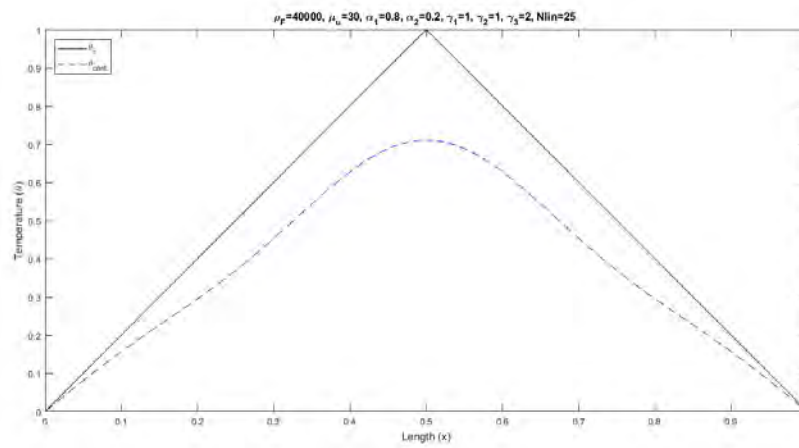


Figure 30:  $\rho_F = 40000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.8$ ,  $\alpha_2 = 0.2$ ,  $\gamma_2 = 1$ .

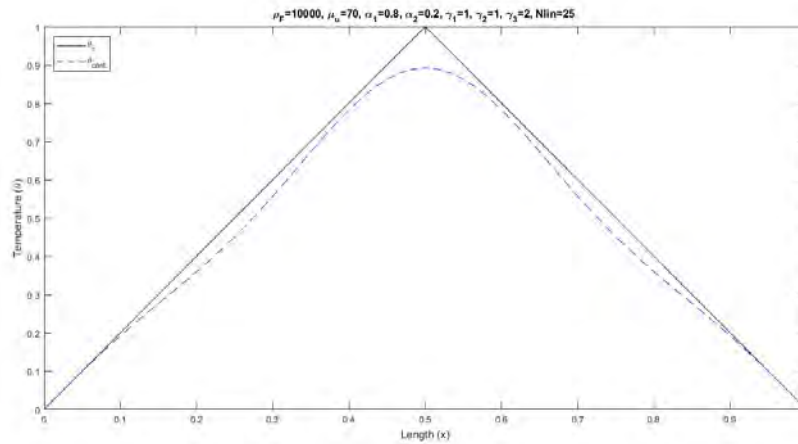


Figure 31:  $\rho_F = 10000$ ,  $\mu_{\mathbf{u}} = 70$ ,  $\alpha_1 = 0.8$ ,  $\alpha_2 = 0.2$ ,  $\gamma_2 = 1$ .

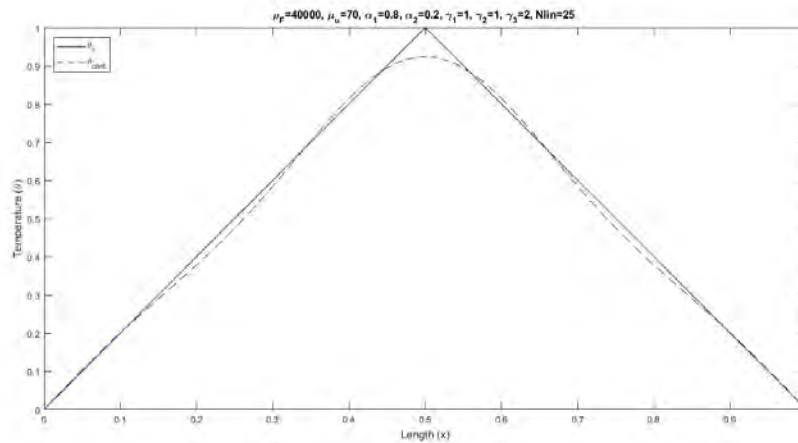


Figure 32:  $\rho_F = 40000$ ,  $\mu_{\mathbf{u}} = 70$ ,  $\alpha_1 = 0.8$ ,  $\alpha_2 = 0.2$ ,  $\gamma_2 = 1$ .

| $\ e_K\ $        | $\mu_{\mathbf{u}} = 30$ | $\mu_{\mathbf{u}} = 70$ |
|------------------|-------------------------|-------------------------|
| $\rho_F = 10000$ | 0.1381                  | 0.0324                  |
| $\rho_F = 40000$ | 0.1381                  | 0.0098                  |

Table 15: Final state for  $\theta_{r_1}$ . Approximation error norms on  $S_K$  for Figures 29 – 32 ( $\alpha_1 = 0.8$ ,  $\alpha_2 = 0.2$  and  $\gamma_2 = 1$ ).

Note that with a “small amount of non-linearity” ( $\alpha_2 = 0.2$  and  $\gamma_2 = 1$ ) the approximation errors obtained with  $\mu_{\mathbf{u}} = 70$  in Table 15 are smaller than the corresponding values in Tables 13 and 14.

Numerical experiments are now presented for the desired final state given in Figure 7 from Chapter 4, which, for convenience, will be named  $\theta_{r2}$ .

The results below (Figures 33 – 44 and Tables 16 – 18) exhibit the same behavior of approximation error with respect to different values of  $\rho_F$ ,  $\mu_u$  and the “amount of non-linearity” as observed in the examples with  $\theta_{r1}$  with performance improving with increasing  $\rho_F$ ,  $\mu_u$  and when  $\alpha(\cdot)$  is “closer” to the linear case (constant  $\alpha(\cdot)$  or “small  $\alpha_2$ ”).

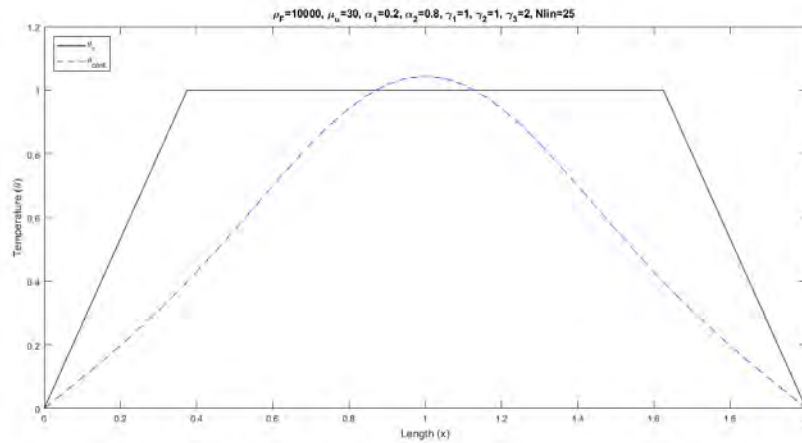


Figure 33:  $\rho_F = 10000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.2$ ,  $\alpha_2 = 0.8$ ,  $\gamma_2 = 1$ .

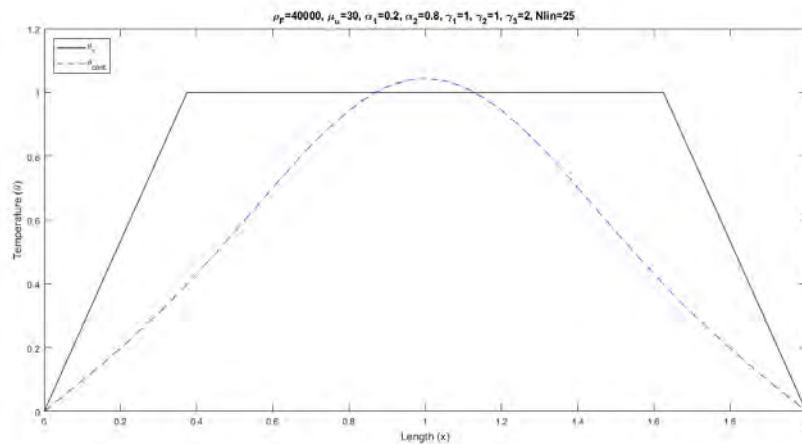


Figure 34:  $\rho_F = 40000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.2$ ,  $\alpha_2 = 0.8$ ,  $\gamma_2 = 1$ .



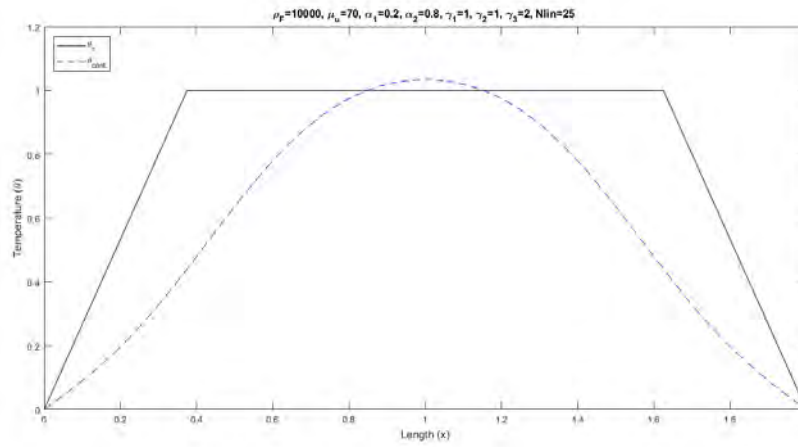


Figure 35:  $\rho_F = 10000$ ,  $\mu_u = 70$ ,  $\alpha_1 = 0.2$ ,  $\alpha_2 = 0.8$ ,  $\gamma_2 = 1$ .

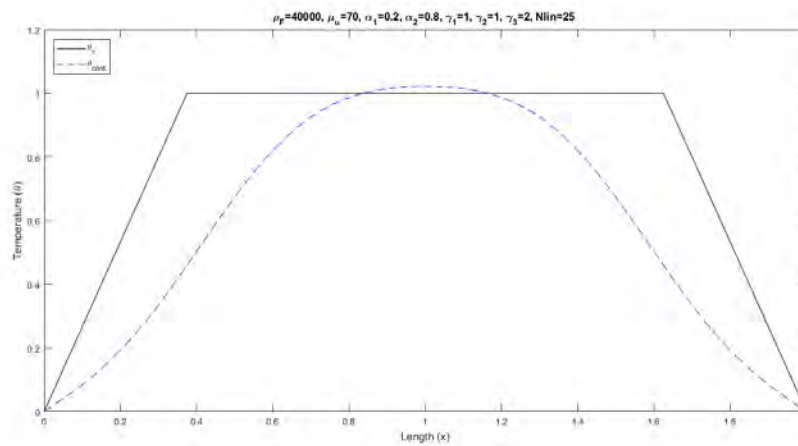


Figure 36:  $\rho_F = 40000$ ,  $\mu_u = 70$ ,  $\alpha_1 = 0.2$ ,  $\alpha_2 = 0.8$ ,  $\gamma_2 = 1$ .

| $\ e_K\ $        | $\mu_u = 30$ | $\mu_u = 70$ |
|------------------|--------------|--------------|
| $\rho_F = 10000$ | 0.4506       | 0.4061       |
| $\rho_F = 40000$ | 0.4496       | 0.3872       |

Table 16: Final state for  $\theta_{r_2}$ . Approximation error norms on  $S_K$  for Figures 33 – 36 ( $\alpha_1 = 0.2$ ,  $\alpha_2 = 0.8$  and  $\gamma_2 = 1$ ).

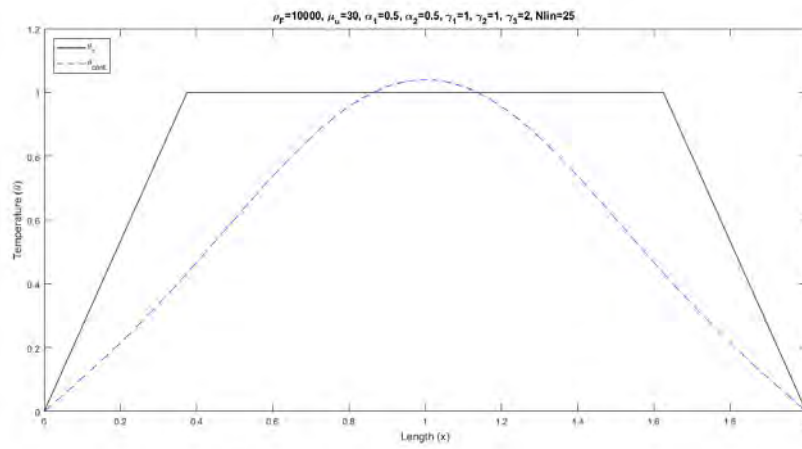


Figure 37:  $\rho_F = 10000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.5$ ,  $\alpha_2 = 0.5$ ,  $\gamma_2 = 1$ .

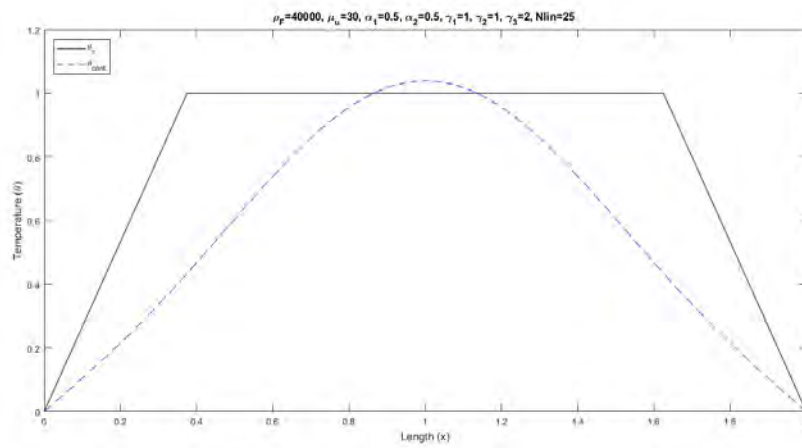


Figure 38:  $\rho_F = 40000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.5$ ,  $\alpha_2 = 0.5$ ,  $\gamma_2 = 1$ .

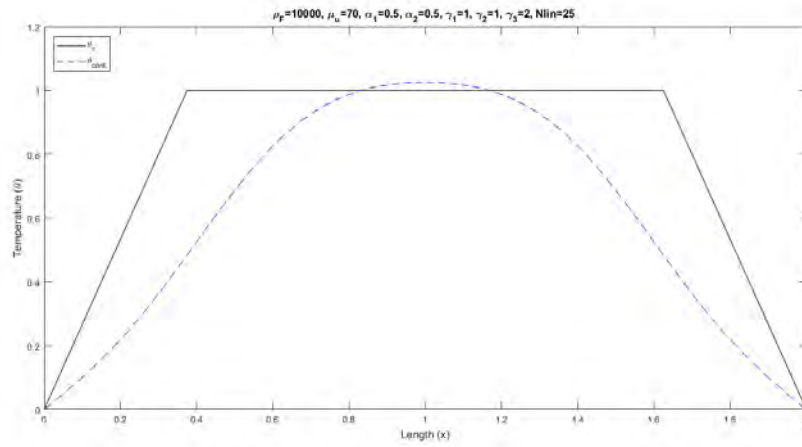


Figure 39:  $\rho_F = 10000$ ,  $\mu_u = 70$ ,  $\alpha_1 = 0.5$ ,  $\alpha_2 = 0.5$ ,  $\gamma_2 = 1$ .

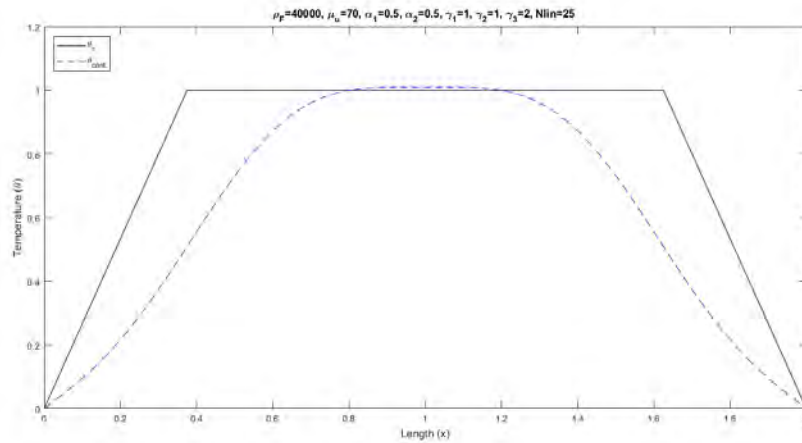


Figure 40:  $\rho_F = 40000$ ,  $\mu_u = 70$ ,  $\alpha_1 = 0.5$ ,  $\alpha_2 = 0.5$ ,  $\gamma_2 = 1$ .

| $\ e_K\ $        | $\mu_u = 30$ | $\mu_u = 70$ |
|------------------|--------------|--------------|
| $\rho_F = 10000$ | 0.4166       | 0.3660       |
| $\rho_F = 40000$ | 0.4158       | 0.3440       |

Table 17: Final state for  $\theta_{r_2}$ . Approximation error norms on  $S_K$  for Figures 37 – 40 ( $\alpha_1 = 0.5$ ,  $\alpha_2 = 0.5$  and  $\gamma_2 = 1$ ).

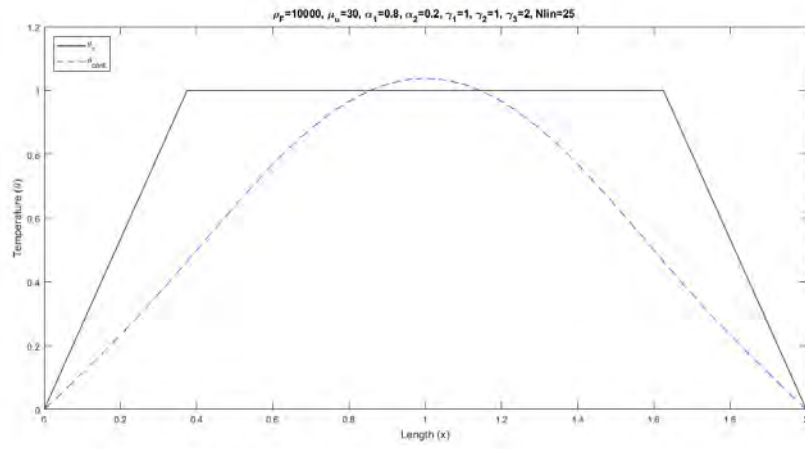


Figure 41:  $\rho_F = 10000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.8$ ,  $\alpha_2 = 0.2$ ,  $\gamma_2 = 1$ .

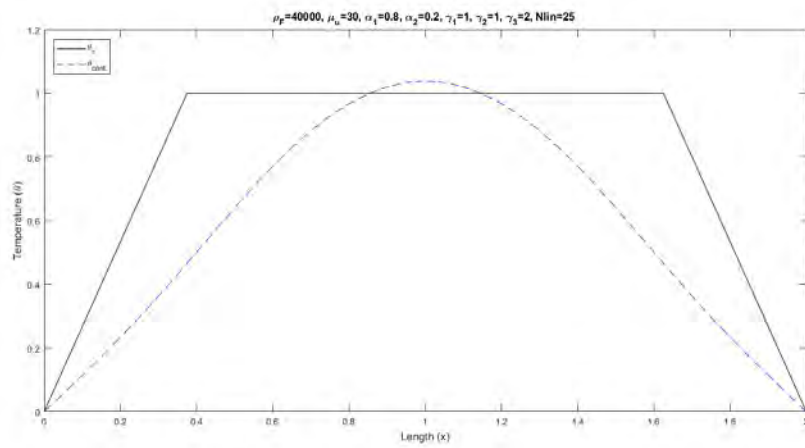


Figure 42:  $\rho_F = 40000$ ,  $\mu_u = 30$ ,  $\alpha_1 = 0.8$ ,  $\alpha_2 = 0.2$ ,  $\gamma_2 = 1$ .

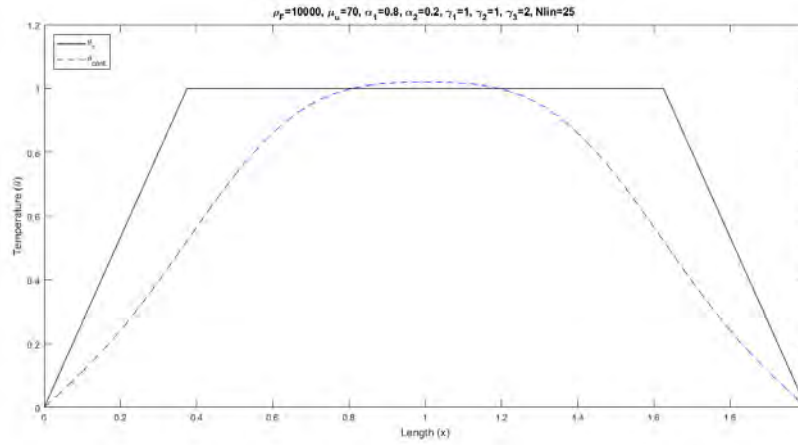


Figure 43:  $\rho_F = 10000$ ,  $\mu_u = 70$ ,  $\alpha_1 = 0.8$ ,  $\alpha_2 = 0.2$ ,  $\gamma_2 = 1$ .

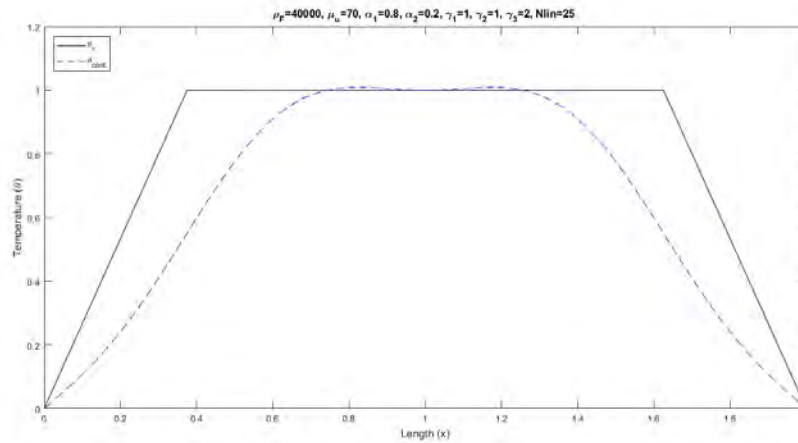


Figure 44:  $\rho_F = 40000$ ,  $\mu_u = 70$ ,  $\alpha_1 = 0.8$ ,  $\alpha_2 = 0.2$ ,  $\gamma_2 = 1$ .

| $\ e_K\ $        | $\mu_u = 30$ | $\mu_u = 70$ |
|------------------|--------------|--------------|
| $\rho_F = 10000$ | 0.3873       | 0.3318       |
| $\rho_F = 40000$ | 0.3860       | 0.3082       |

Table 18: Final state for  $\theta_{r2}$ . Approximation error norms on  $S_K$  for Figures 41 – 44 ( $\alpha_1 = 0.8$ ,  $\alpha_2 = 0.2$  and  $\gamma_2 = 1$ ).

## 6 CONCLUDING REMARKS

In this work, two types of open-loop control problems were addressed in connection with the linear heat equation in rectangular domains with Dirichlet type boundary conditions in which the control function (depending only on time) constitutes a source term. In both cases, the main objective is to impose a prescribed state (temperature distribution) at the final instant of a given time-interval. Control signals are to be selected on the basis of two optimization problems, one unconstrained and the other one involving constraints on the maximum magnitudes of the values taken by the control signals on the time-interval in question. Both problems have the same quadratic cost-functional.

Approximations for the optimal control signals are obtained on the basis of finite-dimensional Galerkin approximation for the linear heat equation. As a consequence, the resulting optimal control signals can be effectively computed. Indeed, in the unconstrained case, they are given as the output of an autonomous, finite-dimensional linear system with initial state given by the data of the original problem. Whereas, in the constrained case, using Lagrangian duality, the resulting control signals are obtained from the cascade connection of a linear system (as in the unconstrained case but with a modified initial state which depends on the “approximately-optimal” Lagrange multipliers) and a (memoryless) limiting operation. Numerical results for the 1D and 2D linear heat equations were presented to illustrate the results mentioned above.

Brief comments on the problem of choosing the location of the “point” controls were also presented together with examples to illustrate the location effects of the final-state approximation goals.

On the basis of the results obtained for the linear heat equation, a heuristic linearization scheme was introduced to address final-state control problems for the NLHEq. This scheme rests on a piecewise linearization of the finite-dimensional, non-linear ODEs corresponding to Galerkin approximations of the NLHEq. Essentially, a given time-interval is divided in contiguous sub-intervals. Starting with the “left most one”, the non-linear ODE is linearized around the given initial state. Then, an optimal control problem is solved for the resulting linear ODE using the results previously obtained. Then, a numerical integration scheme is invoked to obtain the state at the end of this sub-interval with the obtained optimal control acting on the non-linear ODE. This is iteratively repeated over the next subinterval thereby obtaining control signals over the whole interval and the state (of the non-linear ODE) reached of the end of the original interval. Some numerical results are also presented to illustrate this heuristic linearization scheme for the 1D NLHEq.

With respect to further work, it would be interesting to compare the linearization

scheme described above with other ways of choosing control signals, *e.g.*, using piecewise linear signals and random search methods or discretization of control signals obtained from state-feedback methods aiming at directing the state of a non-linear systems at each instant towards the desired final one. Additionally, the problems of choosing the location of point controls for the NLHeq could be investigated particularly in connection with the linearization scheme presented here. Another attractive topic would be the extension of the developed techniques to problems with Robin boundary conditions. The more general problem of approximately imposing a prescribed trajectory for the state variable  $\{\theta(t)\}$ ,  $t \in [0, t_F]$  would also be of interest and possibly handled by similar means. A bigger departure from the problems treated here would be to tackle control problems where the control signal is acting on the boundary of the spacial domain. Two further natural extensions of the work reported here would be to apply a similar approach to the NLHeq in 2D and 3D domains and to generalize the procedures presented here to the use of non-orthogonal basis in the sense of finite element methods, is also of considerable interest.

## REFERENCES

- AHMED, H.E. et al., *Optimization of Thermal Design of Heat Sinks: A Review*, International Journal of Heat and Mass Transfer, vol. **118**, (2018), pp. 129 – 153.
- ALVAREZ-RAMIREZ, J.; ALVAREZ, J., *Robust Temperature Control for Batch Chemical Reactors*, Chemical Engineering Science, vol. **60** (2005), no. 24, pp. 7108 – 7120.
- ATHANS, M.; FALB, P.L., *Optimal Control: An Introduction to the Theory and Its Applications*, 1st. ed, McGraw-Hill, Inc., 1966.
- BELGACEM, F.B.; BERNARDI, C.; FEKIH, H.E., *On the Dirichlet Boundary Control of the Heat Equation with Final Observation I: A space–time mixed formulation and penalization*, Asymptotic Analysis, vol. **71** (2011), no. 1 – 2, pp. 101 – 121.
- BERGAM, A.; BERNARDI, C.; MGHAZLI, Z. *A Posteriori Analysis of the Finite Element Discretization of Some Parabolic Equations*, Mathematics of Computation, vol. **74** (2004), no. 251, pp. 1117 – 1138.
- BLECK, A. et al., *Optimal Control of a Cooling Line for Production of Hot Rolled Dual Phase Steel*, Steel Research International, vol. **85** (2014), no. 9, pp. 1328 – 1333.
- BREZIS, H., *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, 1st. ed., Springer Science + Business Media, LLC, 2011.
- CLEVER, D.; LANG, J., *Optimal Control of Radiative Heat Transfer in Glass Cooling with Restrictions on the Temperature Gradient*, Optimal Control Applications Methods, vol. **33** (2012), pp. 157 – 175.
- COLLATZ, L., *The Numerical Treatment of Differential Equations*, 3rd ed., Springer-Verlag, 1966.
- CORRÊA, G.O.; LÓPEZ-FLORES, M.M.; MADUREIRA, A.L., *Posicionamento Aproximado do Estado Final para Sistemas Descritos pela Equação do Calor*, Anais do XIX Congresso Brasileiro de Automática, UFCG, Campina Grande, PB, 2012.



CURTAIN, R.F.; ZWART, H., *An Introduction to Infinite-dimensional Linear Systems Theory*, 1st. ed., Springer-Verlag New York, Inc., 1995.

DOUGLAS, J., DUPONT, T., *Galerkin Methods for Parabolic Equations*, SIAM Journal of Numerical Analysis, vol. **7** (1970), no. 4, pp. 575 – 626.

EKELAND, I.; TÉMAM, R., *Convex Analysis and Variational Problems*, 1st. ed., Society for Industrial and Applied Mathematics, 1976.

EVANS, L.C., *Partial Differential Equations*, 2nd. ed., American Mathematical Society, 2010.

GAMA, R.M.S., *Fundamentos de Mecânica dos Fluidos*, 1st. ed., EdUERJ, 2012.

\_\_\_\_\_, *Matemática Básica para Mecânica dos Meios Contínuos*, 1st. ed., EdUERJ, 2011.

HOFFMANN, K.H.; BOTKIN, N.D.; TUROVA, V.L., *Optimal Control of Ice Formation in Living Cells During Freezing*, Applied Mathematical Modelling, vol. **35** (2011), no. 8, pp. 4044–4057.

\_\_\_\_\_, *Models and Optimal Control in Freezing and Thawing of Living Cells and Tissues*, in: Leugering, G. et al. (eds), Trends in PDE Constrained Optimization. International Series of Numerical Mathematics, vol. **165**. Birkhäuser, 2014.

ITO, K.; F. KAPPEL, *The Trotter-Kato Theorem and Approximation of PDEs*, Mathematics of Computation, vol. **67** (1989), no. 221, pp. 21 – 44.

KELLEY, W.G.; PETERSON, A.C., *The Theory of Differential Equations: Classical and Qualitative*, 2nd. ed., Springer Science + Business Media, LLC, 2010.

KOGUT, P.I.; LEUGERING, G.R., *Optimal Control Problems for Partial Differential Equations on Reticulated Domains: Approximation and Asymptotic Analysis*, Birkhäuser, 2011.

KUMAR, N.; JHA, A., *Temperature Excursion Management: A Novel Approach of Quality System in Pharmaceutical Industry*, Saudi Pharmaceutical Journal, vol. **25** (2017), pp. 176 – 183.

KUNICH, K.; VEXLER, B., *Constrained Dirichlet Boundary Control in  $L^2$  for a Class of Evolution Equations*, SIAM J. Control Opt. vol. **46** (2007), no. 5, pp. 1726 – 1753.

LAPIDUS, L.; SEINFELD, J.H., *Numerical Solution of Ordinary Differential Equations*, 1st ed, Academic Press, Inc. 1971.

LAUB, A.J., *Matrix Analysis for Scientists & Engineers*, 1st. ed., Society for Industrial and Applied Mathematics, 2005.

LIONS, J.L., *Equations Différentielles Operationnelles: Et Problèmes Aux Limites.*, 1st. ed., Springer-Verlag, 1961.

\_\_\_\_\_, *Exact Controllability, Stabilization and Perturbations for Distributed Parameter Systems*, SIAM Review, vol. **30** (1988a).

\_\_\_\_\_, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes distribués*, Tome 1,2, RMA, vol. **8,9**, (1988b).

LÓPEZ-FLORES, M.M., *Posicionamento Aproximado do Estado Final para Sistemas Térmicos Descritos pela Equação do Calor*. MASTERS THESIS, Faculdade de Engenharia-UERJ, Rio de Janeiro, RJ, Brasil, 2014.

LOTOV, A.V., *Optimal Control of Cooling Process in Continuous Casting of Steel Using a Visualization-based Multi-criteria Approach*, Applied Mathematical Modelling, vol. **29**, (2005), pp. 653 – 672.

LUENBERGER, G., *Optimization by Vector Space Methods*, 1st. ed., John Wiley & Sons, 1969.

- LUO, K. et al., *A Ghost-cell Immersed Boundary Method for Simulations of Heat in Compressible Flows Under Different Boundary Conditions*, International Journal of Heat and Mass Transfer, vol. **92** (2016), pp. 708 – 717.
- MACKI, J.; STRAUSS, A., *Introduction to Optimal Control Theory*, 1st. ed., Springer-Verlag New York, Inc., 1982.
- MAHESH, C. et al., *Atmospheric-temperature-based Cooling System Control for Electronic Devices Using Internet of Things*, International Journal of Ambient Energy, <https://doi.org/10.1080/01430750.2018.1443503>, 2018.
- MARUŠIĆ-PALOKA, E., *Two Methods for Replacing Dirichlet's Boundary Condition by Robin's Boundary Condition via Penalization*, Mathematical Communications, vol. **4** (1999), pp. 27 – 33.
- MERCIER, S. et al., *Time-Temperature Management Along the Food Cold Chain: A Review of Recent Developments*, Comprehensive Reviews in Food Science and Food Safety, vol. **16** (2017), no. 4, pp. 647 – 667.
- MOROZKIN, N.D.; TKACHEV, V.I., *Control of the Process of Cooling of Ceramic Products with Allowance for the Constraints on Thermal Stresses*, Thermophysics and Aeromechanics, vol. **23** (2016), no. 3, pp. 461 – 466.
- MORRIS, K., *Design of Finite-dimensional Controllers for Infinite-dimensional Systems by Approximation*, Journal of Mathematical Systems, Estimation and Control, vol. **4** (1994), no. 2, pp. 1 – 30.
- NDAO, S.; PELES, Y.; JENSEN, M.K., *Multi-objective Thermal Design Optimization and Comparative Analysis of Electronics Cooling Technologies*, International Journal of Heat and Mass Transfer, vol. **52** (2009), pp. 4317 – 4326.
- NISE, N.S., *Control Systems Engineering*, 9th. ed, John Wiley & Sons, Inc., 2015.
- NEUMAIER, A., *Introduction to Numerical Analysis*, 1st. ed, Cambridge University Press, 2001.

\_\_\_\_\_, *Complete Search in Continuous Global Optimization and Constraint Satisfaction*, in: Arieh Iserles (ed), *Acta Numerica* (2004), vol. **13**, pp. 271 – 369.

NUGRAHA, A.S. , *The Selection of Time Step in Runge–Kutta Fourth Order for Determine Deviation in the Weapon Arm Vehicle*, *Energy Procedia*, vol. **68** (2015), pp. 363 – 369.

SHAIKH, P.H. et al., *Intelligent Multi-objective Optimization for Building Energy and Comfort Management*, *Journal of King Saud University–Engineering Sciences* (2018) vol. **30**, pp. 195 – 204.

SLATTERY, J.C., *Advanced Transport Phenomena*, 1st. ed., Cambridge University Press, 1999.

SLYADNEV, M.N. et al., *Photothermal Temperature Control of a Chemical Reaction on a Microchip Using an Infrared Diode Laser*, *Analytical Chemistry*, vol. **73** (2001), no. 16, pp. 4037 – 4044.

SÜLI, E; MAYERS, D.F., *An Introduction to Numerical Analysis*, 1st. ed., Cambridge University Press, 2003.

TAYLOR, A.E.; MANN, W.R., *Advanced Calculus*, 3rd. ed., John Wiley & Sons Inc., 1983.

TEMPO, R.; ISHII, H., *Monte Carlo and Las Vegas Randomized Algorithms for Systems and Control: An Introduction*, *European Journal of Control*, vol. **13** (2007), no. 2–3, pp. 189 – 203.

THOMÉE, V., *Galerkin Finite Element Methods for Parabolic Problems*, 2nd. ed, Springer Berlin Heidelberg, New York, 2006

TRÖLTZSCH, F., *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, American Mathematical Society, 2010.

WARTEN, R.M., *Automatic Step-Size Control for Runge-Kutta Integration*, IBM Journal of Research and Development, vol. **7**, no. 4, 1963.

WHEELER, M.F., *A Priori  $L_2$  Error Estimates For Galerkin Approximations to Parabolic Partial Differential Equations*, SIAM Journal of Numerical Analysis, vol. **10** (1973), no. 4, pp. 723 – 759.

ZHOU, K.; DOYLE, J.C.; GLOVER, K., *Robust and Optimal Control*, 1st. ed., Prentice Hall, 1996.

ZUAZUA, E., *Controllability of Partial Differential Equations and its Semi-discrete Approximations*, Discrete and Continuous Dynamical Systems, vol. **8** (2002), no. 2, pp. 469 – 513.

## APPENDIX A – ELEMENTS OF CONTINUUM MECHANICS AND THE MATHEMATICAL DESCRIPTION OF HEAT CONDUCTION IN SOLIDS

### APPENDIX A.1 – Elements from Continuum Mechanics

The following material was obtained from (GAMA, 2011) e (GAMA, 2012).

#### Divergence Theorem (for tensor fields)

Consider the arbitrary constant vector,  $\mathbf{a}$ . Then, for a tensor field  $\mathbf{S}$  sufficiently regular, we write

$$\int_{\partial\Omega} \mathbf{S}^T \mathbf{a} \cdot \mathbf{n} dS = \int_{\partial\Omega} \mathbf{S} \mathbf{n} \cdot \mathbf{a} dS = \int_{\Omega} (\text{div} \mathbf{S} dS) \cdot \mathbf{a}$$

and that

$$\int_{\partial\Omega} \mathbf{S}^T \mathbf{a} \cdot \mathbf{n} dS = \int_{\Omega} \text{div} (\mathbf{S}^T \mathbf{a}) dV.$$

Then, defining  $\text{div} \mathbf{S}$  such that

$$(\text{div} \mathbf{S}) \cdot \mathbf{a} = \text{div} (\mathbf{S}^T \mathbf{a})$$

it can be written

$$\int_{\partial\Omega} \mathbf{S} \mathbf{n} dS = \int_{\Omega} \text{div} \mathbf{S} dV.$$

#### Velocity Gradient

Let  $\mathbf{v} = v_x \mathbf{i} + v_y \mathbf{j} + v_z \mathbf{k}$  be a vector field. The *velocity gradient* of  $\mathbf{v}$  is defined by

$$\text{grad} \mathbf{v} = \begin{bmatrix} \frac{\partial v_x}{\partial x} & \frac{\partial v_x}{\partial y} & \frac{\partial v_x}{\partial z} \\ \frac{\partial v_y}{\partial x} & \frac{\partial v_y}{\partial y} & \frac{\partial v_y}{\partial z} \\ \frac{\partial v_z}{\partial x} & \frac{\partial v_z}{\partial y} & \frac{\partial v_z}{\partial z} \end{bmatrix}.$$

Then we have

$$\text{grad} \mathbf{v} = \underbrace{\mathbf{D}}_{\text{Symmetric part}} + \underbrace{\mathbf{W}}_{\text{Skew-Symmetric part}}$$

where

$$\mathbf{D} = \frac{1}{2} [\text{grad} \mathbf{v} + (\text{grad} \mathbf{v})^T] \quad \text{and} \quad \mathbf{W} = \frac{1}{2} [\text{grad} \mathbf{v} - (\text{grad} \mathbf{v})^T].$$

### Material Derivative

Let  $\boldsymbol{\omega}$  be a vector field. We define the *material derivative* as

$$\frac{D \boldsymbol{\omega}}{Dt} = \frac{\partial \boldsymbol{\omega}}{\partial t} + (\text{grad} \boldsymbol{\omega}) \mathbf{v}.$$

### Reynolds Transport Theorem

Let field  $\Psi$ , be a function of the variable  $\mathbf{x} \in \mathbb{R}^3$  (representing the position the position with spatial coordinates  $(x, y, z)$  that depend on time  $t$ , given in this case, by

$$\Psi = \hat{\Psi}(\mathbf{x}, t) = \hat{\Psi}(x, y, z, t).$$

Let a variable  $\mathbf{x} \in \mathbb{R}^3$  with coordinates  $(X, Y, Z)$  defined in such a way that, given  $\mathbf{X}$  and  $t$ ,  $\mathbf{x}$  is determined in a unique way from the regular function

$$\chi(\mathbf{X}, t) \rightarrow (x, y, z) = (\tilde{x}(X, Y, Z), \tilde{y}(X, Y, Z), \tilde{z}(X, Y, Z)).$$

Thus, we can write

$$\Psi = \hat{\Psi}(\mathbf{X}, t) = \hat{\Psi}(\tilde{x}(X, Y, Z), \tilde{y}(X, Y, Z), \tilde{z}(X, Y, Z)) = \tilde{\Psi}(\mathbf{X}, t).$$

*Reynolds Transport Theorem* establishes that, in case of regular fields, that

$$\frac{d}{dt} \int_{\Omega_t} \left[ \frac{\partial}{\partial t} (\tilde{\Psi}(\mathbf{X}, t)) + (\tilde{\Psi}(\mathbf{X}, t)) \text{tr} \left( \frac{\partial \mathbf{F}}{\partial t} \mathbf{F}^{-1} \right) \right] dV = \int_{\Omega_t} \left[ \frac{D\Psi}{Dt} + \Psi \text{div} \mathbf{v} \right] dV,$$

where  $\mathbf{F} = \tilde{\mathbf{F}}(\mathbf{X}, t)$ ,  $\frac{D}{Dt}$  represents the material derivative and  $\Omega_t$  is the body's configuration at time  $t$  (it can vary on time  $t$ ).

### **APPENDIX A.2–Green's Identities**

For given twice differentiable functions  $\theta, \phi$  in a domain  $\Omega \subset \mathbb{R}$  with boundary  $\partial\Omega$ , we have the following Green's Identities (TAYLOR; MANN, 1983, pp. 492 – 493):

### The First Identity

$$\int_{\Omega} \left( \frac{\partial^2 \theta}{\partial x^2} \phi + \frac{\partial \theta}{\partial x} \frac{\partial \phi}{\partial x} \right) d\Omega = \int_{\partial\Omega} \frac{\partial \theta}{\partial x} \phi \cdot \mathbf{n} d(\partial\Omega),$$

where  $\mathbf{n}$  is the normal (perpendicular) vector to  $\partial\Omega$ .

### The Second Identity

$$\int_{\Omega} \left( \frac{\partial^2 \theta}{\partial x^2} \phi - \theta \frac{\partial^2 \phi}{\partial x^2} \right) d\Omega = \int_{\partial\Omega} \left( \frac{\partial \theta}{\partial x} \phi - \theta \frac{\partial \phi}{\partial x} \right) d(\partial\Omega).$$

## APPENDIX A.3 – The Mathematical Description of Heat Conduction in Rigid Bodies

In this section, the deduction of the mathematical description of heat conduction in rigid bodies gives us as a result the (linear/non-linear) non-homogeneous (with a source term) heat equation. This deduction was presented in (LÓPEZ-FLORES, 2014, pp. 18 – 22) and it is reproduced here for the reader's convenience.

The *energy equation* for a continuous body, also known as *First Law of Thermodynamics*, consists on an axiom that establishes that:

*The rate of change of the quantity of energy (kinetic + internal) is equal to the rate of mechanical work done over a body (mechanical power due to the forces acting over the body) in addition to this, the rate of energy transmitted in the form of heat (heat transmitted by time unit through the boundary + the heat generation).*

This principle can be mathematically represented by

$$\frac{d}{dt} \int_{\Omega_t} \rho \left[ u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right] dV = \int_{\partial\Omega_t} (\mathbf{T}\mathbf{n}) \cdot \mathbf{v} dS + \int_{\partial\Omega_t} \rho \mathbf{g} \cdot \mathbf{v} dV + \int_{\partial\Omega_t} -\mathbf{q} \cdot \mathbf{n} dS + \int_{\Omega_t} \dot{q} dV, \quad (\text{A.3.1})$$



where:

$$\begin{aligned} \int_{\partial\Omega_t} (\mathbf{Tn}) \cdot \mathbf{v} dS & : \text{mechanical power of the surface forces (contact);} \\ \int_{\partial\Omega_t} \rho \mathbf{g} \cdot \mathbf{v} dV & : \text{mechanical power of the body's forces;} \\ \int_{\partial\Omega_t} -\mathbf{q} \cdot \mathbf{n} dS & : \text{heat flux crossing (entering) the boundary of the body;} \\ \int_{\Omega_t} \dot{q} dV & : \text{internal heat generation rate (energy).} \end{aligned}$$

The quantity  $\rho$  represents the body's density  $\Omega_t$  (the body's actual configuration), the function  $u$  represents the body's internal energy and  $\mathbf{v}$  represents the body's velocity field. The quantity  $\mathbf{q}$  represents the heat flux vector (per unit of time and area), while the quantity  $\dot{q}$  represents the rate of heat generation (per unit of time and volume). For example, when an electric current flows through a conductor,  $\dot{q}$  is positive, in average, the same as the product of the difference of potential by the current, divided by the respective volume of conductive material. The negative sign of the integral before the last one above, appears due to the fact that it represents the flux entering the body and not the flux coming out.

Rewriting (A.3.1) with the help of *Reynolds transport theorem*, in the following way,

$$\begin{aligned} \int_{\Omega_t} \left\{ \frac{D}{Dt} \left[ \rho \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) \right] + \rho \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) \operatorname{div} \mathbf{v} \right\} dV & = \tag{A.3.2} \\ = \int_{\partial\Omega_t} (\mathbf{Tn}) \cdot \mathbf{v} dS + \int_{\partial\Omega_t} \rho \mathbf{g} \cdot \mathbf{v} dV + \int_{\partial\Omega_t} -\mathbf{q} \cdot \mathbf{n} dS + \int_{\Omega_t} \dot{q} dV. \end{aligned}$$

Rewriting the internal argument of left-hand side integral of (A.3.2) as

$$\begin{aligned} & \frac{D}{Dt} \left[ \rho \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) \right] + \rho \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) \operatorname{div} \mathbf{v} = \\ & = \rho \frac{D}{Dt} \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) + \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) \frac{D\rho}{Dt} + \rho \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) \operatorname{div} \mathbf{v} \\ & = \left[ \frac{D\rho}{Dt} + \rho \operatorname{div} \mathbf{v} \right] \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) + \rho \frac{D}{Dt} \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) \end{aligned} \tag{A.3.3}$$

and considering the *continuity equation* we establish the following

$$\frac{D\rho}{Dt} + \rho \operatorname{div} \mathbf{v} = 0. \quad (\text{A.3.4})$$

Then we obtain

$$\int_{\Omega_t} \left\{ \rho \frac{D}{Dt} \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) \right\} dV = \int_{\partial\Omega_t} (\mathbf{T}\mathbf{n}) \cdot \mathbf{v} dS + \int_{\partial\Omega_t} \rho \mathbf{g} \cdot \mathbf{v} dV + \int_{\partial\Omega_t} -\mathbf{q} \cdot \mathbf{n} dS + \int_{\Omega_t} \dot{q} dV. \quad (\text{A.3.5})$$

The symmetry of the Cauchy stress tensor,  $\mathbf{T}$ , and the *divergence theorem* allows us to write

$$\begin{aligned} \int_{\partial\Omega_t} (\mathbf{T}\mathbf{n}) \cdot \mathbf{v} dS - \int_{\partial\Omega_t} \mathbf{q} \cdot \mathbf{n} dS &= \int_{\partial\Omega_t} (\mathbf{T}\mathbf{v}) \cdot \mathbf{n} dS - \int_{\partial\Omega_t} \mathbf{q} \cdot \mathbf{n} dS \\ &= \int_{\Omega_t} \operatorname{div} (\mathbf{T}\mathbf{v}) dV - \int_{\Omega_t} \operatorname{div} \mathbf{q} dV. \end{aligned} \quad (\text{A.3.6})$$

Therefore, the energy equation, (A.3.1), can be expressed as

$$\int_{\Omega_t} \left\{ \rho \frac{D}{Dt} \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) \right\} dV = \int_{\Omega_t} \operatorname{div} (\mathbf{T}\mathbf{v}) dV + \int_{\partial\Omega_t} \rho \mathbf{g} \cdot \mathbf{v} dV - \int_{\Omega_t} \operatorname{div} \mathbf{q} dV + \int_{\Omega_t} \dot{q} dV. \quad (\text{A.3.7})$$

Since the region  $\Omega_t$  is chosen arbitrarily, we can conclude that (the local form of the energy equation for a continuous body) is given by

$$\rho \frac{D}{Dt} \left( u + \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) = \operatorname{div} (\mathbf{T}\mathbf{v}) + \rho \mathbf{g} \cdot \mathbf{v} - \operatorname{div} \mathbf{q} + \dot{q}. \quad (\text{A.3.8})$$

Since

$$\operatorname{div} (\mathbf{T}\mathbf{v}) = (\operatorname{div} \mathbf{T}) \cdot \mathbf{v} + \mathbf{T} \cdot \operatorname{grad} \mathbf{v} \quad (\text{A.3.9})$$

and that

$$\frac{D}{Dt} \left( \frac{1}{2} \mathbf{v} \cdot \mathbf{v} \right) = \frac{D\mathbf{v}}{Dt} \cdot \mathbf{v} \quad (\text{A.3.10})$$

equation (A.3.8) can be written as

$$\rho \frac{Du}{Dt} + \rho \frac{D\mathbf{v}}{Dt} \cdot \mathbf{v} = (\operatorname{div} \mathbf{T}) \cdot \mathbf{v} + \mathbf{T} \cdot \operatorname{grad} \mathbf{v} + \rho \mathbf{g} \cdot \mathbf{v} - \operatorname{div} \mathbf{q} + \dot{q}. \quad (\text{A.3.11})$$

Taking in consideration that the *equation of linear motion* establishes that

$$\rho \frac{D\mathbf{v}}{Dt} = \text{div}\mathbf{T} + \rho\mathbf{g} \quad (\text{A.3.12})$$

a local form of the energy equation given by (A.3.8), it reduces to

$$\rho \frac{Du}{Dt} = \mathbf{T} \cdot \text{grad}\mathbf{v} - \text{div}\mathbf{q} + \dot{q}. \quad (\text{A.3.13})$$

The symmetry of the stress tensor and the definition of *material derivative* allows us to write

$$\rho \left( \frac{\partial u}{\partial t} + \text{grad } u \cdot \mathbf{v} \right) = -\text{div}\mathbf{q} + \mathbf{T} \cdot \mathbf{D} + \dot{q}, \quad (\text{A.3.14})$$

where  $\mathbf{D}$  is the symmetrical part of the *velocity gradient*. It is important to point out that for rigid bodies,  $\text{grad}\mathbf{v}$  is skew-symmetric and so  $\mathbf{T} \cdot \mathbf{D} = 0$ .

#### Fourier's Law

Initially we admit that the heat flux vector is solely dependent on the temperature distribution. In other words, the heat flux vector  $\mathbf{q}$  must be a function of the temperature  $\theta$  and on the spacial gradient  $\text{grad } \theta$ . To comply with the objectivity (frame invariance) principles, we have that (SLATTERY, 1999, p. 251 – 256, 273 – 275)

$$\mathbf{q} = \mathbf{K}(\theta, \text{grad } \theta) \text{grad } \theta. \quad (\text{A.3.15})$$

*Fourier's law* is a particular case of the previous equation where  $\mathbf{K}$  depends only on  $\theta$ . For isotropic materials, Fourier's law is given by

$$\mathbf{q} = -\kappa(\theta) \text{grad } \theta, \quad (\text{A.3.16})$$

where  $\kappa$  is called *thermal conductivity*.

#### Heat Conduction in an Isotropic, Rigid Solids at Rest

For a rigid opaque solid at rest, the energy equation reduces to

$$\rho \frac{\partial u}{\partial t} = -\text{div}\mathbf{q} + \dot{q}. \quad (\text{A.3.17})$$

It is known from thermodynamics that the specific heat at a constant volume for a rigid solid, denoted it here by  $c$ , to be constant and defined as  $c = (\partial u / \partial \theta)_V$ . Using this we

write

$$\rho c \frac{\partial \theta}{\partial t} = -\operatorname{div} \mathbf{q} + \dot{q}, \quad (\text{A.3.18})$$

where  $\theta$  represents the temperature field. Notice that for the given body's configuration, we can assume the density,  $\rho$ , as constant.

Considering  $\mathbf{q}$  as in (A.3.16), we then have the general equation for the heat conduction in an isotropic rigid solid at rest given by

$$\rho c \frac{\partial \theta}{\partial t} = \operatorname{div} (\kappa(\theta) \operatorname{grad} \theta) + \dot{q}. \quad (\text{A.3.19})$$

If we consider the case in which  $\dot{q}$  does not depend on the temperature  $\theta$ , it is possible to write the previous equation as

$$\frac{\partial \theta}{\partial t} - \operatorname{div} (\alpha(\theta) \operatorname{grad} \theta) = f, \quad (\text{A.3.20})$$

where  $\alpha(\theta) = \frac{\kappa(\theta)}{\rho c} > 0$  is called *thermal diffusivity* and  $f = \frac{\dot{q}}{\rho c}$ . It is important to point out that the thermal conductivity is strictly positive. "Heat flows in the opposite direction of the temperature gradient".

In engineering, in some cases, is a common practice to approximate the thermal conductivity by an average constant value. In other words, for some cases, we despise the dependency of the thermal conductivity from the temperature. In such cases we consider  $\mathbf{q} = -\kappa \operatorname{grad} \theta$ . We assume that  $\kappa$  is constant. Also, if we consider  $\dot{q}$  not depending on  $\theta$ , we write (A.3.20) as

$$\frac{\partial \theta}{\partial t} - \alpha \operatorname{div} (\operatorname{grad} \theta) = f, \quad (\text{A.3.21})$$

where  $\alpha = \frac{\kappa}{\rho c} > 0$  and  $f = \frac{\dot{q}}{\rho c}$ , as before.

### Boundary Conditions (BCs)

Now we chose initial and boundary conditions for our given partial differential equations. For this we have the *Dirichlet* (prescribed temperature) and *Neumann* boundary conditions (insulated wall), which are commonly found in the classical heat transfer literature. However, both are somewhat unrealistic.

In a more realistic sense, is not possible to prescribe a temperature and there is no perfect insulating material (the a Neumann boundary condition can be used with precision in cases when it describes symmetry). Thus, with the interest to keep a minimum resemblance with reality we should use a boundary condition that correlates the boundary temperature with the normal (perpendicular) heat flux. For example, Newton's law of

cooling, given by

$$\mathbf{q} \cdot \mathbf{n} = h(\theta - \theta_\infty) \text{ over } \partial\Omega. \quad (\text{A.3.22})$$

#### Choice of BCs for the Positioning Problem of the Final State

Nevertheless, by the relative simplicity of the resulting problems and the approximation nature (still coarse) in many cases, Dirichlet boundary conditions are commonly found in the literature, see for example (MARUŠIĆ;PALOKA, 1999), (KUNICH; VEXLER, 2007), (BELGACEM; BERNARDI; FEKIH, 2011), (LUO, K. et al, 2016).

Since the main interest is to approximately position the final state of a thermal system described by the (linear or nonlinear) heat equation, this will be formulated here as an optimal control problem subject to partial differential equations for a final time  $t_F$ , with a Dirichlet boundary condition. Even though it is not a very realistic boundary condition, it allows us to make a fast and intuitive assessment of the numerical results here obtained with previously available ones.

## APPENDIX B – MATERIAL ON SYSTEMS AND CONTROL THEORY

### APPENDIX B.1 – Open-Loop Control System

In this section, we present information related to *systems* and *control theory* that will allow the reader to follow the main ideas used throughout the text.

One of the two major configurations of a control system is an *open-loop control system*, see (NISE, 2015, pp. 6 – 7). This is a system that does not have a “feedback loop” in its configuration. This kind of system will help the reader understand, conceptually, the main ideas of our work. This system is represented in the can be presented as follows. to understand the main idea of our work.

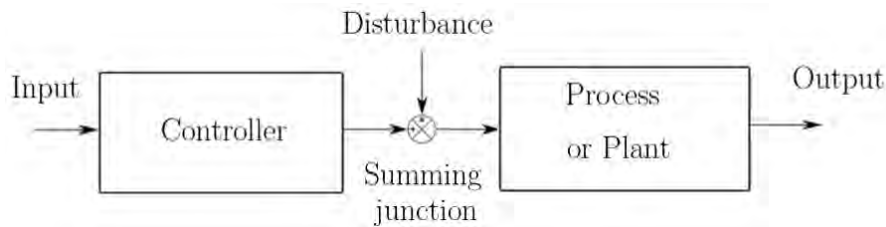


Figure 45: Block diagram of a generic open-loop control system.

As shown in Figure 45, in a generic simple open-loop control system we have two main subsystems of interest. We have a *controller* that receives an input or reference signal that “drives” the second subsystem to a specific desired state. The second subsystem is called the *process* or *plant* which gives the specified state in form of an output signal that can be called a *controlled variable*.

Any system is prone to disturbances. This other signals are shown added to the controller via summing junctions. This signals will be added algebraically to the input signal using the associated signs. As an example, the controller in a heating system consists of fuel valves and the electrical system that operates the valves. As for the plant, we can consider an air conditioning system or a furnaces, where the controlled variable is the temperature.

We can represent our problem simply by the following open-loop system.

By identifying the plant with a process, such as heat conduction in solids, a choice is to represent this system by the heat equation with its necessary initial and boundary conditions.

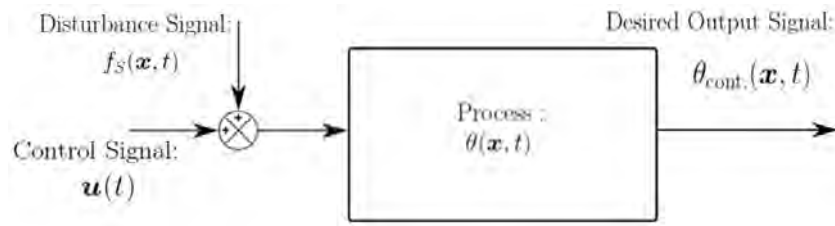


Figure 46: Open-loop system representing a process modeled by the heat equation where the control function,  $\mathbf{u}$ , depends only of time with a disturbance signal,  $f_S$ , entering the process which gives us a desired output signal  $\theta_{cont.}$ .

## APPENDIX B.2 – The Essential Elements of a Control Problem

The essential elements of a control problem are presented, as in (ATHAN; FALB, 1966, p. 3 – 4), as follows:

### 1. *A mathematical model (system) to be “controlled”.*

The mathematical model, which represents the physical system, consists of a set of relations which describe the response or output of the system for various inputs. Constraints based upon the physical situations are incorporated in this set of relations.

### 2. *A desired output of the system.*

The objective of the system is often translated into a requirement on the output. The desired output is the signal being tracked (or something close to it).

### 3. *A set of admissible inputs or “controls”.*

Since “control” signals in physical systems are usually obtained from equipment which can provide only a limited amount of force or energy, constraints are imposed upon the inputs of the systems. These constraints lead to a set of admissible inputs (or “control” signals). The desired objective can be attained by any admissible inputs, then a measure of performance or cost control is sought for which allows the choice of the “best” input.

**4. A performance or cost functional which measures the effectiveness of a given “control action”.**

When a cost functional has been decided upon, the (admissible) inputs as determined which generate the desired output and which, in so doing, minimize (optimize) the chosen performance measure. At this point, optimal-control theory is used to find a solutions to the control problem.

**APPENDIX B.3 – Mathematical Formulation of a Control Problem**

We follow the presentation in (MACKI; STRAUSS, 1982, pp. 4 – 6) of a precise mathematical formulation of the type of control problem we will be discussing. Let  $m, n$  be natural numbers, and let  $\mathbb{R}$  stand for the real numbers. If  $\mathbf{x}, \mathbf{y}$  in  $\mathbb{R}^n$ , we denote their  $i$ –th components by  $x_i, y_i$ , respectively.

Let  $\Omega_\ell$  denote the unit cube  $\mathbb{R}^m$ , *i.e.*,

$$\Omega_\ell = \{\boldsymbol{\ell} : \boldsymbol{\ell} \in \mathbb{R}^m, |\ell_i| \leq 1, i = 1, 2, \dots, m\}.$$

For  $t_1 \geq 0$ , define

$$\mathcal{U}_{ad}[0, t_1] = \{\mathbf{u}(\cdot) : \mathbf{u}(t) \in \Omega_\ell \text{ and } \mathbf{u}(\cdot) \text{ measurable on } [0, t_1]\},$$

$\mathcal{U}_{ad} = \bigcup_{t_1 > 0} \mathcal{U}_{ad}[0, t_1]$  ( $\mathcal{U}_{ad}$ –set of the admissible controls  $\mathbf{u}(\cdot)$ ). We assume that for each  $t \geq 0$  we are given a *target set*  $\mathcal{T}(t) \subset \mathbb{R}^n$  where  $\mathcal{T}(t)$  is a closed set.

We assume that the dynamics of the system, that is, the evolution of the state  $\mathbf{x}(t)$  under a given control  $\mathbf{u}(t)$ , is determined by a vector ordinary differential equation:

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0. \tag{B.1}$$

We will always assume that  $\mathbf{f}(t, \mathbf{x}, \mathbf{u}), \partial f_i / \partial x_j, \partial f_i / \partial u_k$  are all continuous ( $i, j = 1, \dots, n; k = 1, \dots, m$ ) on  $[0, \infty) \times \mathbb{R}^n \times \mathbb{R}^m$ , although most results are valid under weaker conditions. This assumption guarantees local existence and uniqueness of the solution of (B.1) for a given  $\mathbf{u}(\cdot) \in \mathcal{U}_{ad}$ . Because  $\mathbf{u}(\cdot)$  is only assumed measurable and bounded, the right side of the equation  $\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}, \mathbf{u}(t))$  is continuous in  $\mathbf{x}$  but only measurable and bounded in  $t$  for each  $\mathbf{x}$ . Therefore, solutions are understood to be absolutely continuous functions that satisfy (B.1) almost everywhere.



The solution of (B.1) for a given  $\mathbf{u}(\cdot)$  will be called the *response* to  $\mathbf{u}(\cdot)$ ; we denote it by  $\mathbf{x}[t] \equiv \mathbf{x}(t; \mathbf{x}_0, \mathbf{u}(\cdot))$ . The *control problem* is to determine those  $\mathbf{x}_0$  and  $\mathbf{u}(\cdot) \in \mathcal{U}_{ad}$  such that the associated response satisfies  $\mathbf{x}[t_1] \in \mathcal{T}(t_1)$  for some  $t_1 > 0$ ; we then say that *the control  $\mathbf{u}(\cdot)$  steers  $\mathbf{x}_0$  to the target*.

If the control  $\mathbf{u}(\cdot)$  is defined on  $[0, t_1)$  ( $t_1 \leq \infty$ ), it is not assumed that the corresponding response extends to  $[0, t_1)$ ; a given response  $\mathbf{x}(t; \mathbf{x}_0, \mathbf{u}(\cdot))$  may only exist on some subinterval of  $[0, t_1)$ .

Thus, our general control problem consists of a class of admissible *controls*  $\mathcal{U}_{ad}$ , a vector differential equation (B.1) describing the *dynamics* of our system, and a family of *target* sets  $\mathcal{T}(t)$ . One basic problem is to describe those initial states  $\mathbf{x} \in \mathbb{R}^n$  which can be steered to the target, that is, those states which are *controllable*.

#### APPENDIX B.4 – The Optimal Control Problem

The basic control problem, as in (MACKI; STRAUSS, 1982, p. 9), may have associated with it a cost functional or performance criterion. Taking the cost functional of the form

$$\mathcal{J}(\mathbf{u}(\cdot)) = \int_0^{t_1} f(\mathbf{x}[t], \mathbf{u}(t)) dt, \quad \mathbf{x}[t] \equiv \mathbf{x}(t; \mathbf{x}_0, \mathbf{u}(\cdot)),$$

where  $f$  is a given real-valued function. The *optimal control problem* is to steer  $\mathbf{x}_0$  to a state in the target, using a control  $\mathbf{u}(\cdot)$  from the appropriate class for the problem, in such a way that  $\mathcal{J}(\mathbf{u}(\cdot))$  is a minimum. More precisely, let the *successful controls* be denoted by  $\mathcal{SC}$ , *i.e.*,

$$\mathcal{SC} = \{\mathbf{u}(\cdot) \in \mathcal{U}_{ad} : \exists t_1 \geq 0 \text{ such that } \mathbf{x}(t_1; \mathbf{x}_0, \mathbf{u}(\cdot)) \in \mathcal{T}(t_1)\}.$$

Then a control  $\mathbf{u}_*(\cdot) \in \mathcal{U}_{ad}$  is *optimal* if it is successful, *i.e.*,  $\mathbf{u}_*(\cdot) \in \mathcal{SC}$ , and

$$\mathcal{J}(\mathbf{u}_*(\cdot)) \leq \mathcal{J}(\mathbf{u}(\cdot)) \quad \text{for all } \mathbf{u}(\cdot) \in \mathcal{SC}.$$

## APPENDIX C – MATERIAL ON SEMIGROUP THEORY AND THE 4TH-ORDER RUNGE–KUTTA METHOD

### APPENDIX C.1 – Semigroup Theory

The study of linear, first-order differential equations in infinite-dimensional Banach spaces provides the basic motivation for the introduction of one-parameter family of linear operators (semigroups) as defined below. Recalling that a linear ODE in  $\mathbb{R}^n$ , namely,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t), \quad t \geq 0 \quad \text{where } \mathbf{x} : [0, \infty) \rightarrow \mathbb{R}^n \quad \text{and } \mathbf{A} \in \mathbb{R}^{n \times n},$$

has solution given by  $\mathbf{x}(t) = \exp[\mathbf{A}t]\mathbf{x}_0$ , the question arises if, for  $\mathcal{X} : [0, \infty) \rightarrow \mathcal{X}$ , where  $\mathcal{X}$  is not finite-dimensional and  $\mathbf{A}$  is defined on a dense subset of  $\mathcal{X}$ , there exists a family  $\{S(t) : t \geq 0\}$  of linear mappings  $S(t) : \mathcal{X} \rightarrow \mathcal{X}$  such that the solutions of the linear ODE  $\dot{\mathbf{z}}(t) = \mathbf{A}\mathbf{z}(t)$  are likewise given by  $\mathbf{z}(t) = S(t)\mathbf{x}(0)$ . Clearly, if  $\mathbf{A}$  is bounded on  $\mathcal{X}$ , it suffices to define

$$S(t) = \exp[\mathbf{A}t] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbf{A}^k$$

(in this case, the series above converges uniformly on any finite interval  $[0, t_F]$  with respect to the induced operator norm). The definition of strongly continuous semigroup provides a generalization of  $\exp[\mathbf{A}t]$  (for unbounded  $\mathbf{A}$ ) in its role of describing the solutions of linear ODEs in infinite-dimensional spaces.

Basic definition and few fundamental results are reproduced below, see (CURTAIN; ZWART, 1995, Chapter 2, pp. 13 – 52)

**Definition C.1:** (CURTAIN; ZWART, 1995, Definition 2.1.2, p. 15). *A  $C_o$ -semigroup is an operator-valued function  $T(t)$  from  $\mathbb{R}_+$  to  $L(Z)$  ( $Z$  is a separable Hilbert space) that satisfies the following properties:*

$$T(t+s) = T(t)T(s) \quad \text{for } t, s \geq 0; \tag{C.1}$$

$$T(0) = I; \tag{C.2}$$

$$\|T(t)z_0 - z_0\| \rightarrow 0 \quad t \rightarrow 0^+ \quad \forall z_0 \in Z. \tag{C.3}$$

**Theorem C.T1** (CURTAIN;ZWART, 1995, Theorem 2.1.6 (d, e), p. 18). *A strongly continuous semigroup on a Hilbert space  $Z$   $T(t)$  has the following properties:*

(d) *If  $\omega_0 = \inf_{t>0} \left[ \frac{1}{t} \log \|T(t)\| \right]$ , then  $\omega_0 = \lim_{t \rightarrow \infty} \left[ \frac{1}{t} \log \|T(t)\| \right] < \infty$ ;*

(e)  *$\forall \omega > \omega_0$ , there exists a constant  $M_\omega$  such that  $\forall t \geq 0$ ,  $\|T(t)\| \leq M_\omega e^{\omega t}$ .*

**Definition C.2:** (CURTAIN;ZWART, 1995, Definition 2.1.8, p. 20). *The infinitesimal generator  $A$  of a  $C_0$ -semigroup on a Hilbert space  $Z$  is defined by*

$$Az = \lim_{t \rightarrow 0^+} \frac{1}{t} [T(t) - I],$$

*whenever the limit exists; the domain of  $A$ ,  $Dom(A)$ , being the set of elements in  $Z$  for which the limit exists.*

**Theorem C.T2** (CURTAIN;ZWART, 1995, Theorem 2.1.10 (b, f), p. 21). *Let  $T(t)$  be a strongly continuous semigroup on a Hilbert space  $Z$  with infinitesimal generator  $A$ . Then the following results hold:*

(b)  *$\frac{d}{dt} [T(t)z_0] = AT(t)z_0 = T(t)Az_0$  for  $z_0 \in Dom(A)$ ,  $t > 0$ ;*

(f)  *$A$  is a closed linear operator.*

**Definition C.3:** (CURTAIN;ZWART, 1995, Definition A.4.1, p. 608). *Let  $A$  be a closed linear operator on a (complex) normed linear space  $X$ . We say that  $\lambda$  is in the resolvent set  $\rho(A)$  of  $A$ , if  $(\lambda I - A)^{-1}$  exists and is a bounded linear operator on a dense domain of  $X$ . We shall call  $(\lambda I - A)^{-1}$  the resolvent operator of  $A$ .*

**Theorem C.T3** (CURTAIN;ZWART, 1995, Theorem 2.1.12 – The Hille-Yosida Theorem, p. 26). *A necessary and sufficient condition for a closed, densely defined, linear operator  $A$  on a Hilbert space  $Z$  to be the infinitesimal generator of a  $C_0$ -semigroup is that there exist real numbers  $M, \omega$ , such that for all real  $\alpha > \omega$ ,  $\alpha \in \rho(A)$ , the resolvent set of  $A$ , and*

$$\|R(\alpha, A)^r\| \leq \frac{M}{(\alpha - \omega)^r} \quad \text{for all } r \geq 1,$$

*where  $R(\alpha, A) = (\alpha I - A)^{-1}$  is the resolvent operator. In this case*

$$\|T(t)\| \leq M e^{\omega t}.$$

## APPENDIX C.2–The 4th–Order Runge–Kutta Method

In this section, short presentation of the 4th–order Runge–Kutta Method is given. Numerical methods for the initial value problem are traditionally divided into two classes: *one-step methods* and *multistep* (or multivalued) methods. One-step methods (also called *Runge-Kutta methods*) are memoryless and only make use of the most recently computed solution point to compute a new solution point, whereas multistep methods retain some memory of the history by storing, using, and updating a matrix containing old and new information, see (NEUMAIER, 2001, p. 211).

The “classical (4th–order) Runge–Kutta method” or simply the “Runge–Kutta method” it is one of the most widely used methods to solve initial value problems such as the one presented as follows.

Suppose a continuous function  $F : D \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^n$ , a point  $y_0 \in \text{int}(D)$ , and a real interval  $[x_0, x_F]$  are given. We seek a continuously differentiable function  $y : [x_0, x_F] \rightarrow \mathbb{R}^n$  with  $y'(x) = F(y(x))$ ,  $y(x_0) = y_0$ . Each such function is called a *solution* to the initial value problem

$$y' = F(y), \quad y(x_0) = y_0 \quad \text{in the interval } [x_0, x_F]. \quad (\text{C.4})$$

Here  $y$  is an unknown function (scalar or vector) of  $x$ , which we would like to approximate; The function  $F$  and the data  $(x_0, y_0)$  are given.

Choosing a step-size  $h > 0$  and define

$$\begin{aligned} y_{n+1} &= y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \\ x_{n+1} &= x_n + h \end{aligned}$$

for  $n = 0, 1, 2, 3, \dots$ , using

$$\begin{aligned} k_1 &= hF(x_n, y_n), \\ k_2 &= hF\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}\right), \\ k_3 &= hF\left(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}\right), \\ k_4 &= hF(x_n + h, y_n + k_3). \end{aligned}$$

Here  $y_{n+1}$  is the 4th–order Runge–Kutta approximation of  $y(x_{n+1})$ , and the next value ( $y_{n+1}$ ) is determined by the present value ( $y_n$ ) plus the weighted average of four increments,

where each increment is the product of the size of the interval,  $h$ , and an estimated slope specified by function  $F$  on the right-hand side of the differential equation. More specifically,  $k_2$  and  $k_3$  represent approximations to the derivative  $y'$  at points on the solution curve, intermediate between  $(x_n, y(x_n))$  and  $(x_{n+1}, y(x_{n+1}))$ , and  $F(x_n, y_n; h)$  is a weighted average of the  $k_i, i = 1, 2, 3, 4$ , the weights corresponding to those of Simpsons rule (to which the classical 4th-order Runge–Kutta method reduces when  $\frac{\partial F}{\partial y} \equiv 0$ ), see (SÜLI;MAYERS, 2003, p. 328).

In the numerical results presented in this work it was used the following quadrature form:

**Quadrature for numerical integration (to obtain  $\mathbf{A}_q$ )**

**Grid:**  $x_j = (j - 1)L_x/N_x, \quad j = 1, \dots, N_x + 1.$

$$\{\mathbf{A}_q\}_{\ell k} = \sum_{j=1}^{N_x+1} (1/2) [\mathbf{g}_q^{\mathbf{A}}(x_j) + \mathbf{g}_q^{\mathbf{A}}(x_{j+1})] (L_x/N_x),$$

where  $\mathbf{g}_q^{\mathbf{A}}(x) = \alpha_a(\mathbf{z}_q^0)(x) \frac{\partial \phi_k(x)}{\partial x} \frac{\partial \phi_\ell(x)}{\partial x}.$

To avoid some difficulties using the Runge–Kutta method presented here, a careful choice of step-size has to be taken. As an empirical rule we present the following commentary from (LAPIDUS;SEINFELD, 1971, Section 2.8, pg. 69),

Collatz Rule of Thumb

(COLLATZ, 1966) has outlined a rule-of-thumb method for indirectly measuring  $F(x, h)$  by specifying the correct magnitude of  $h$ . He suggested that in the classic Runge–Kutta calculation  $k_2$  and  $k_3$  must be approximately equal. To be more specific, it was suggested that the identity

$$\left| \frac{k_3 - k_2}{k_2 - k_1} \right| = 1.0 \tag{C.5}$$

hold when the proper value of  $h$  is used in a calculation. If the ratio is much greater than 1, the local truncation error is too large in this step and  $h$  must be decreased. If the ratio is much smaller than 1, the  $h$  should be increased. The use of (C.5) to specify  $h$  is, of course, subject to considerable interpretation and should be used only with care. As an example, if  $F(x, y) = F(x)$  then  $k_3 = k_2$  in the Runge–Kutta formulas and (C.5) breaks down. (WARTEN, 1963) has, however, used the idea with some success. He suggested that one try to estimate the local truncation error by means of a weighted linear combi-

*nation of the  $k_i$ . The weighting factors are then adjusted to fit the local truncation error for the linear differential equation  $y' = ay + b$  or its vector equivalent.*

This type of analysis continues to be of interest in the applied sciences. For example, in the recent work of (NUGRAHA, 2015), it was studied the selection of time step in Runge–Kutta fourth order for determine deviation in the weapon arm vehicle, where the simulation results showed that for stepping over a specified time of 0.01 s produced unstable simulation results, in contrast, using a time step of 0.001 s provided a more stable result. This shows how important is the choice of the time step for this method.