NTNU

Kunnskap for en bedre verden

TMA4212 - Numerical solution of differential equations by difference methods

Project 2

Authors:
Vemund Aakre, Thorbjørn C. Djupvik, Oskar F. Jakobsen

4th April 2025

1 The Poisson Equation

The Poisson equation with homogeneous Dirichlet boundary conditions is

$$-\Delta u = f, u(0) = u(1) = 0 \tag{1}$$

We assume that the reader is familiar with the derivation of the weak formulation of the Poisson equation; find $u \in H_0^1(0,1)$ such that

$$a(u,v) := \int_0^1 u_x v_x \,\mathrm{d}x = \int_0^1 fv \,\mathrm{d}x =: F(v) \tag{2}$$

for all $v \in H_0^1(0,1)$. Restricting to the second degree Lagrange finite element space results in the linear system

$$Au_h = F \tag{3}$$

where A is the stiffness matrix, and F is the load vector. Since $u(0) = u(1) = 0$ we set the first and last entry of u_h to 0.

1.1 Second degree Lagrange finite element space

Let $\hat{K} = [0,1]$ serve as the reference element. The Lagrange interpolating polynomials (shape functions) on the nodes $(0, \frac{1}{2}, 1)$ are

$$\begin{aligned} \Psi_0(x) &= 2x^2 - 3x + 1 \\ \Psi_1(x) &= -4x^2 + 2x \\ \Psi_2(x) &= 2x^2 - x \end{aligned} \tag{4}$$

We partition the interval $K = [0,1]$ into $M+1$ points. Given a partition $0 = s_0 < s_1 < \cdots < s_M = 1$ we define the elements $K_k = [s_k, s_{k+1}]$ for $k = 0, \ldots, M-1$. Denote the size of element k by $h_k := s_{k+1} - s_k$.

In order to construct a basis on X_h^2 we need three nodes per element. Hence, a partition of K into $M+1$ points results in M segments and $2M+1$ nodes. Let x_i denote the i'th node.

In the following we use k to indicate the index of a segment and $\alpha \in \{0,1,2\}$ to indicate the node on the segment. These indices are related by the local to global map θ:

$$i = \theta(k,\alpha) := 2k + \alpha \tag{5}$$

The following bijection maps the reference element onto the k-th physical element.

$$\begin{aligned} \Phi_k : \hat{K} &\longrightarrow K \\ \xi &\longmapsto \xi s_{k+1} + (1-\xi)s_k \end{aligned} \tag{6}$$

The i'th basis function is denoted by φ_i and is defined as

$$\varphi_i(x) = \varphi_{\theta(k,\alpha)}(x) := \Psi_\alpha({\Phi_k}^{-1}(x)). \tag{7}$$

1.2 Constructing the stiffness matrix and the load vector

The elemental stiffness matrix and load vector has entries

$$[A^{K_k}]_{ij} = \int_{s_k}^{s_{k+1}} \varphi_i'(x)\varphi_j'(x) \,\mathrm{d}x \tag{8}$$

$$[F^{K_k}]_i = \int_{s_k}^{s_{k+1}} f\left(x\right) \varphi_i\left(x\right) \,\mathrm{d}x \tag{9}$$

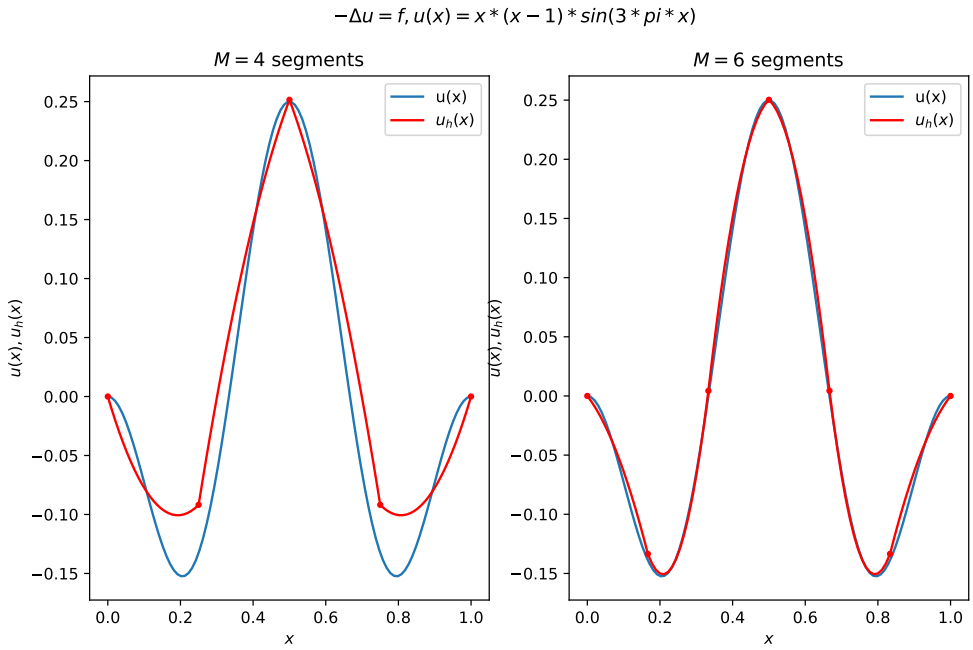

Figure 1: The blue graphs shows the exact solution. The red graphs are the numerical solutions on two different equidistant partitions of the domain. The approximated solution u_h is sampled at 100 equidistant points.

A change of variables (let $x = \Phi_k(\xi)$) and using the chain rule yields

$$[A^{K_k}]_{ij} = \frac{1}{h_k} \int_0^1 \Psi_i'(\xi)\Psi_j'(\xi)\,\mathrm{d}\xi \tag{10}$$

$$[F^{K_k}]_i = h_k \int_0^1 f\left(\Phi_k\left(\xi\right)\right)\Psi_i\left(\xi\right)\,\mathrm{d}\xi \tag{11}$$

The assembly of the stiffness matrix is just a matter of adding the elements to their respective sub-matrices (see algorithm 1). The construction of the load vector follows the same procedure. Note also that the elemental stiffness matrix can be computed exactly, for example with a Gauss-Legendre quadrature rule of appropriate degree. In contrast, the function f in the elemental load vectors forces these integrals to be computed numerically (Simpsons rule for instance).

Algorithm 1 Assemble stiffness matrix

Require: $M \geq 1$
  $A \leftarrow [0]_{(2M+1)\times(2M+1)}$
  for $k = 0, \ldots, M-1$ **do**
    $[A]_{i,j=\theta(k,0),\ldots,\theta(k,2)} \leftarrow [A]_{i,j=\theta(k,0),\ldots,\theta(k,2)} + [A^{K_k}]$ ▷ add the k-th elemental matrix
  end for

1.3 Test problem

We use the test solution $u(x) = x(x-1)\sin(3\pi x)$ with Dirichlet boundary conditions $u(0) = u(1) = 0$ (see figure 1). To test the implementation for inhomogeneous dirichlet boundary conditions we use the test solution $u(x) = x\cos(3\pi x)$, $u(0) = 0, u(1) = -1$ (see figure 2).

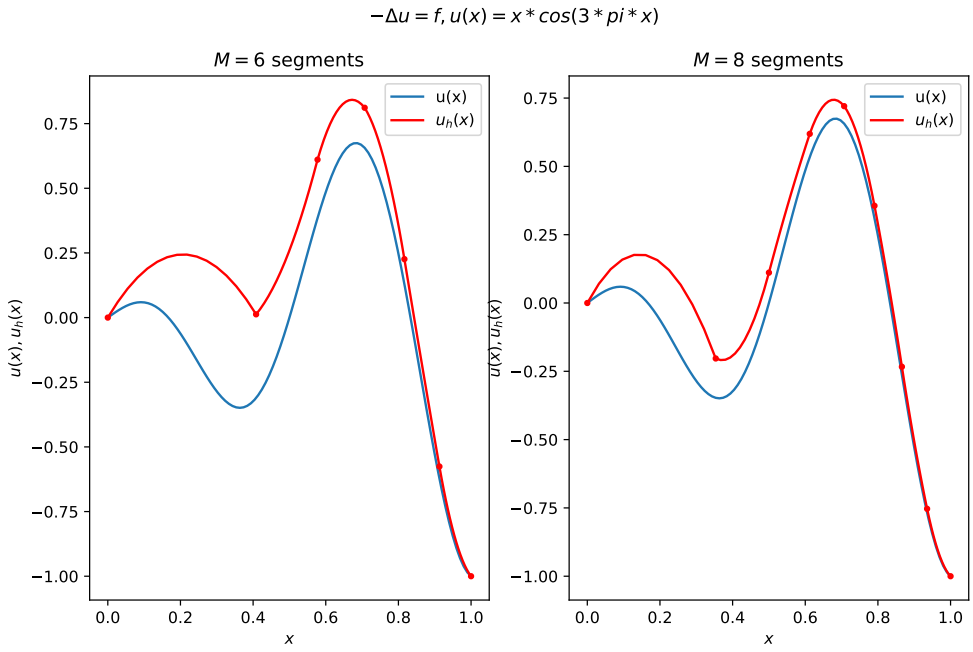

Figure 2: The blue graphs shows the exact solution. The red graphs are the numerical solutions on two different non-equidistant partitions of the domain. The approximated solution u_h is sampled at 100 equidistant points.

1.4 Convergence analysis

Theorem 1.1. *Let $u \in H_0^3((0,1))$ and $u_h \in X_h^2 \cap H_0^1$ be the solutions of the infinite and finite dimensional variational problem respectively, where h is the maximum element size in X_h^2. Assume further that F in the variational problem is bounded, and that u has a second order polynomial interpolant u_h^2 on $(0,1)$. Then*

$$||u - u_h||_{H^1} \leq Ch^2 \tag{12}$$

for some constant $C > 0$.

Proof. We first note that $a(u,v)$ is both bounded and coercive, with constants M and α respectively. Combined with the assumption that F is bounded, we then get from the Lax-Milgram theorem ([1], page 15) that the variational problem admits a unique solution.

From Cea's lemma ([1], page 18) we then obtain

$$||u - u_h||_{H^1} \leq \frac{M}{\alpha}||u - v_h||_{H^1} = \frac{M}{\alpha}(||u - v_h||_{L^2} + |u - v_h|_{H^1}) \quad \forall v_h \in X_h^2 \cap H_0^1. \tag{13}$$

Since $u(0) = u(1) = 0$, and $\{0,1\}$ are nodes of the interpolating polynomial, we obtain that $u_h^2(0) = u_h^2(1) = 0$ as well. In particular we get that $u_h^2 \in X_h^2 \cap H_0^1$.

Choosing $v_h = u_h^2$ in Cea's lemma results in

$$||u - u_h||_{H^1} \leq \frac{M}{\alpha}(||u - u_h^2||_{L^2} + |u - u_h^2|_{H^1}), \tag{14}$$

which combined with lemma 4.4 ([1], page 19) and $h \leq 1$ gives us

$$\begin{aligned} ||u - u_h||_{H^1} &\leq \frac{M}{\alpha}(C_2 h^3 |u|_{H^3} + C_1 h^2 |u|_{H^3}) \\ &= h^2 \frac{M}{\alpha}(C_2 h |u|_{H^3} + C_1 |u|_{H^3}) \\ &\leq Ch^2 \end{aligned} \tag{15}$$

□

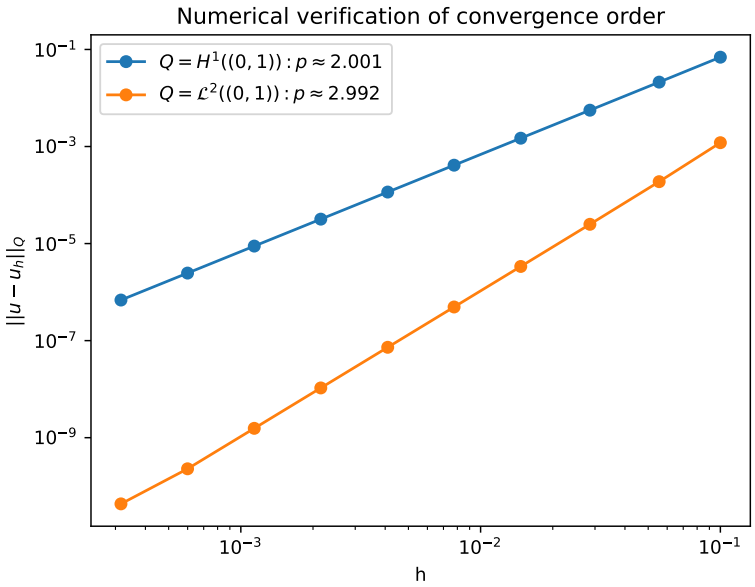

Figure 3: Convergence plot for the test solution $u(x) = x(x-1)\sin(3\pi x)$ in both $\mathcal{L}^2$ and H^1 norm. h is the element size for a equidistant partition of $(0, 1)$.

1.4.1 Numerical Verification

To numerically calculate the $\mathcal{L}^2$ and H^1 norm we write the integral as a sum over the elements and use a change of variables similar to the stiffness matrix. The inner integrals have to be calculated numerically.

$$\|u - u_h\|^2_{\mathcal{L}^2(\Omega)} = \int_\Omega (u(x) - u_h(x))^2 \, \mathrm{d}x \tag{16}$$

$$= \sum_{k=0}^{M-1} h_k \int_0^1 \left(u(\Phi_k(\xi)) - \sum_{\alpha=0}^{2} u_{\theta(k,\alpha)} \Psi_{\theta(k,\alpha)}(\xi) \right)^2 \mathrm{d}x \tag{17}$$

$$|u - u_h|^2_{H^1(\Omega)} = \int_\Omega (u'(x) - u_h'(x))^2 \, \mathrm{d}x \tag{18}$$

$$= \sum_{k=0}^{M-1} h_k \int_0^1 \left(u'(\Phi_k(\xi)) - \frac{1}{h_k} \sum_{\alpha=0}^{2} u_{\theta(k,\alpha)} \Psi'_{\theta(k,\alpha)}(\xi) \right)^2 \mathrm{d}x \tag{19}$$

We observe in figure 3 that the convergence order in the H^1 norm match the theoretical results, and that $\mathcal{L}^2$ is of one degree higher. The equation used is the same as our first test equation (see figure 1).

1.5 Generalisation

We have used second order Lagrange finite elements throughout this project. The source code is written such that any degree d of Lagrange basis functions can be used. The process here is effectively to replace every 2 in section 1.1 with d.

2 A PDE optimal control problem

We have some physical one-dimensional object, $\Omega = (0, 1)$, that we wish to give a desired temperature profile $y_d \in L^2(\Omega)$. The boundaries of the object are set to a (relative) temperature of 0. There is some heat source u that we have control over, but there is a cost $\alpha \in (0, \infty)$ associated with both cooling and heating the object.

Our problem is then to minimize the discrepancy between the desired profile and the temperature profile of the object, while simultaneously minimizing the cost of the heating/cooling. A model for the problem is then

$$\min_{y,u} \frac{1}{2}\int_0^1 |y - y_d|^2 dx + \frac{\alpha}{2}\int_0^1 u^2 dx \quad \text{s.t.} \quad \begin{cases} -\Delta y = u \\ y(0) = y(1) = 0 \end{cases} \quad \text{in the weak sense.} \tag{20}$$

We approximate this as

$$\min_{u_h, y_h \in V_h} \frac{1}{2}\|y_h - \bar{y_d}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u_h\|^2_{L^2(\Omega)} \quad \text{s.t.} \quad a(y_h, v) = \langle u_h, v\rangle_{L^2(\Omega)} \forall v \in V_h, \tag{21}$$

where $\bar{y}_d$ is the interpolation of y_d onto X_h^2, $V_h = X_h^2 \cap H_0^1(\Omega)$ and $a(u, v) = \int_0^1 u_x v_x dx$. The optimal control is then u_h, and the optimal state is y_h.

2.1 Solution

To find a solution we interpret this as a minimization problem for the unknown coefficients $\mathbf{u} = \{u_1, \ldots, u_{2N-1}\}$, $\mathbf{y} = \{y_1, \ldots, y_{2N-1}\}$ as

$$\min_{\mathbf{y},\mathbf{u} \in \mathbb{R}^{2N-1}} G(\mathbf{y}, \mathbf{u}) \quad \text{s.t.} \quad B\mathbf{y} = F\mathbf{u}. \tag{22}$$

For F and B we use each side of the constraint from (21),

$$\begin{aligned} a(y_h, v) &= \int_0^1 \left(\sum_{i=1}^{2N-1} y_i \varphi_i'\right)\left(\sum_{j=1}^{2N-1} v_j \varphi_j'\right) dx \\ &= \sum_{i=1}^{2N-1}\sum_{j=1}^{2N-1} y_i v_j \int_0^1 \varphi_i' \varphi_j' dx \\ &= \mathbf{v}^T B \mathbf{y}, \end{aligned} \tag{23}$$

where $B_{i,j} = \int_0^1 \varphi_i' \varphi_j' dx$, $i, j = 1, \ldots, 2N-1$, which is the stiffness matrix, and

$$\begin{aligned} \langle u_h, v\rangle_{L^2(\Omega)} &= \int_0^1 \left(\sum_{i=1}^{2N-1} u_i \varphi_i\right)\left(\sum_{j=1}^{2N-1} v_j \varphi_j\right) dx \\ &= \sum_{i=1}^{2N-1}\sum_{j=1}^{2N-1} u_i v_j \int_0^1 \varphi_i \varphi_j dx \\ &= \mathbf{v}^T F \mathbf{u}, \end{aligned} \tag{24}$$

where $F_{i,j} = \int_0^1 \varphi_i \varphi_j dx$, $i, j = 1, \ldots, 2N-1$. Hence $\mathbf{v}^T B\mathbf{y} = \mathbf{v}^T F\mathbf{u}$ and since the constraint holds for all $\mathbf{v} \in V_h$, we get that

$$B\mathbf{y} = F\mathbf{u}. \tag{25}$$

We find $G(\mathbf{y}, \mathbf{u})$ by rewriting the objective function of (21). Since the desired profile is an interpolation, we know that $\bar{y}_d = \sum_{i=0}^{2N} d_i \varphi_i$, where $\mathbf{d} = \begin{bmatrix} d_0 & \ldots & d_{2N} \end{bmatrix}^T$ are the coefficients, giving the

form

$$
\begin{aligned}
G(\mathbf{y},\mathbf{u}) &= \frac{1}{2}\langle y_h - \bar{y}_d, y_h - \bar{y}_d\rangle_{L^2(\Omega)} + \frac{\alpha}{2}\langle u_h, u_h\rangle_{L^2(\Omega)} \\
&= \frac{1}{2}\langle y_h, y_h\rangle_{L^2(\Omega)} + \frac{1}{2}\langle \bar{y}_d, \bar{y}_d\rangle_{L^2(\Omega)} - \langle y_h, \bar{y}_d\rangle_{L^2(\Omega)} + \frac{\alpha}{2}\langle u_h, u_h\rangle_{L^2(\Omega)} \\
&= \frac{1}{2}\int_0^1 \left(\sum_{i=1}^{2N-1} y_i\varphi_i\right)^2 dx + \frac{1}{2}\int_0^1 \left(\sum_{i=0}^{2N} d_i\varphi_i\right)^2 dx \\
&\quad - \int_0^1 \left(\sum_{i=1}^{2N-1} y_i\varphi_i\right)\left(\sum_{i=0}^{2N} d_i\varphi_i\right) dx + \frac{\alpha}{2}\int_0^1 \left(\sum_{i=1}^{2N-1} u_i\varphi_i\right)^2 dx \\
&= \frac{1}{2}\sum_{i=1}^{2N-1}\sum_{j=1}^{2N-1} y_i y_j \int_0^1 \varphi_i\varphi_j dx + \frac{1}{2}\sum_{i=0}^{2N}\sum_{j=0}^{2N} d_i d_j \int_0^1 \varphi_i\varphi_j dx \\
&\quad - \sum_{i=1}^{2N-1}\sum_{j=0}^{2N} y_i d_i \int_0^1 \varphi_i\varphi_j dx + \frac{\alpha}{2}\sum_{i=1}^{2N-1}\sum_{j=1}^{2N-1} u_i u_j \int_0^1 \varphi_i\varphi_j dx.
\end{aligned}
$$

Writing this expression in matrix-vector form,

$$
G(\mathbf{y},\mathbf{u}) = \frac{1}{2}\mathbf{y}^T F\mathbf{y} + \frac{1}{2}\mathbf{d}^T\tilde{F}\mathbf{d} - \mathbf{y}^T\hat{F}\mathbf{d} + \frac{\alpha}{2}\mathbf{u}^T F\mathbf{u}, \tag{26}
$$

where $\tilde{F} \in \mathbb{R}^{(2N+1)\times(2N+1)}$ is F, but extended to include the boundaries. $\hat{F} \in \mathbb{R}^{(2N-1)\times(2N+1)}$ is the same as $\tilde{F}$ but with the first and last row removed.

To solve equation (22) we use Lagrange multipliers and the Lagrangian

$$
\begin{aligned}
\mathcal{L}(\mathbf{y},\mathbf{u},\lambda) &= G(\mathbf{y},\mathbf{u}) - \lambda^T\left(B\mathbf{y} - F\mathbf{u}\right) \\
&= \frac{1}{2}\mathbf{y}^T F\mathbf{y} + \frac{1}{2}\mathbf{d}^T\tilde{F}\mathbf{d} - \mathbf{y}^T\hat{F}\mathbf{d} + \frac{\alpha}{2}\mathbf{u}^T F\mathbf{u} - \lambda^T\left(B\mathbf{y} - F\mathbf{u}\right),
\end{aligned} \tag{27}
$$

where $\lambda \in \mathbb{R}^{2N-1}$ is the vector of Lagrange multipliers. The solution is found by solving

$$
\nabla_{\mathbf{y}}\mathcal{L} = 0 \tag{28}
$$

$$
\nabla_{\mathbf{u}}\mathcal{L} = 0 \tag{29}
$$

$$
\nabla_{\lambda}\mathcal{L} = 0 \tag{30}
$$

for $\mathbf{y}$ and $\mathbf{u}$. Calculating the gradients, and exploiting the symmetry of F and B yields the system

$$
F\mathbf{y} - \hat{F}\mathbf{d} - B\lambda = 0 \tag{31}
$$

$$
\alpha F\mathbf{u} + F\lambda = 0 \tag{32}
$$

$$
-B\mathbf{y} + F\mathbf{y} = 0 \tag{33}
$$

Lemma 2.1. *F is positive definite.*

Proof. The basis functions φ_i are linearly independent, since they are Lagrange nodal polynomials and our nodes are unique. Since $F_{i,j} = \int_0^1 \varphi_i\varphi_j dx = \langle\varphi_i,\varphi_j\rangle_{L^2(\Omega)}$, F is a Gram-matrix corresponding to the vector space of our basis functions. Combining these results gives us that F is positive definite. □

Lemma 2.2. *B is positive definite.*

Proof. Let $\mathbf{z} \in \mathbb{R}^{2N-1}, \mathbf{z} \neq 0$. Then

$$
\mathbf{z}^T B\mathbf{z} = \sum_{i=1}^{2N-1}\sum_{j=1}^{2N-1} z_i B_{i,j} z_j = \sum_{i=1}^{2N-1}\sum_{j=1}^{2N-1} a(v_i\varphi_i, v_j\varphi_j) > 0.
$$

□

Lemma 2.3. *$BF^{-1}B$ is positive definite.*

Proof. F^{-1} is positive definite since it is the inverse of a positive definite matrix and $\ker(B) = \{0\}$ since B is positive definite, by lemma (2.2).

Let $\mathbf{z} \in \mathbb{R}^{2N-1}, \mathbf{z} \neq 0$. Then

$$\mathbf{z}^T BF^{-1}B\mathbf{z} = (B\mathbf{z})^T F^{-1}(B\mathbf{z}) > 0.$$

□

Since F is invertible, by lemma (2.1), we see that $\alpha\mathbf{u} = -\lambda$. This gives the system

$$F\mathbf{y} + \alpha B\mathbf{u} = \hat{F}\mathbf{d} \tag{34}$$
$$-B\mathbf{y} + F\mathbf{u} = \mathbf{0}. \tag{35}$$

An explicit solution can be found by seeing that $\mathbf{u} = F^{-1}B\mathbf{y}$, and getting

$$F\mathbf{y} + \alpha BF^{-1}B\mathbf{y} = \hat{F}\mathbf{d},$$

which gives the solutions

$$\mathbf{y} = \left[F + \alpha BF^{-1}B\right]^{-1} \hat{F}\mathbf{d} \tag{36}$$
$$\mathbf{u} = F^{-1}B\left[F + \alpha BF^{-1}B\right]^{-1} \hat{F}\mathbf{d}, \tag{37}$$

where $\left[F + \alpha BF^{-1}B\right]^{-1}$ exists since $\alpha > 0$ and it is therefore the inverse of the sum of positive definite matrices. The system can also be written in block matrix form as

$$\begin{bmatrix} F & \alpha B \\ -B & F \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \hat{F}\mathbf{d} \\ \mathbf{0} \end{bmatrix}. \tag{38}$$

This system has sparse sub-blocks, which is solved efficiently by a sparse solver.

2.2 Example temperature profiles

We wish to test our implementation for some different desired temperature profiles,

$$y_d = \frac{1}{2}x(1-x), \tag{39}$$
$$y_d = 1, \tag{40}$$
$$y_d = \begin{cases} 1 & \text{for} \quad x \in \left[\frac{1}{4}, \frac{3}{4}\right] \\ 0 & \text{else.} \end{cases} \tag{41}$$

We see that when the cost parameter is large ($\alpha = 10^{-1}$, figure 4) the discrepancy between the optimal state and the desired profile is quite large. However, the heating is kept relatively small (especially compared with figures 5 and 6). That is, the optimal control becomes small at the expense of the discrepancy.

In contrast, for lower costs ($\alpha = 10^{-8}$, figure 6), the discrepancy between the optimal state and the desired profile is small. Note that the distributed heat source reaches amplitudes in the interval $(-1500, 2000)$ for equations (40) and (41).

Using a typical cost parameter ($\alpha = 10^{-3}$, figure 5), we see that there is some discrepancy from the desired profile. The heating profile is reasonable.

This is to be expected given the original problem. If heating is expensive, the temperature discrepancy is large. If heating is cheap, the temperature discrepancy is small.

We have assumed that the optimal state, y, and optimal control, u, are in H_0^1, whilst the desired profile, y_d, are not restricted by H_0^1. We get interesting results if we use a desired profile not in H_0^1.

This is the case for equations (40) and (41). We see that for both high and typical costs, both the optimal control and state are smooth.

In contrast, for low costs, the optimal control oscillates rapidly and with high amplitude at the boundaries (40) and at the discontinuities (41).

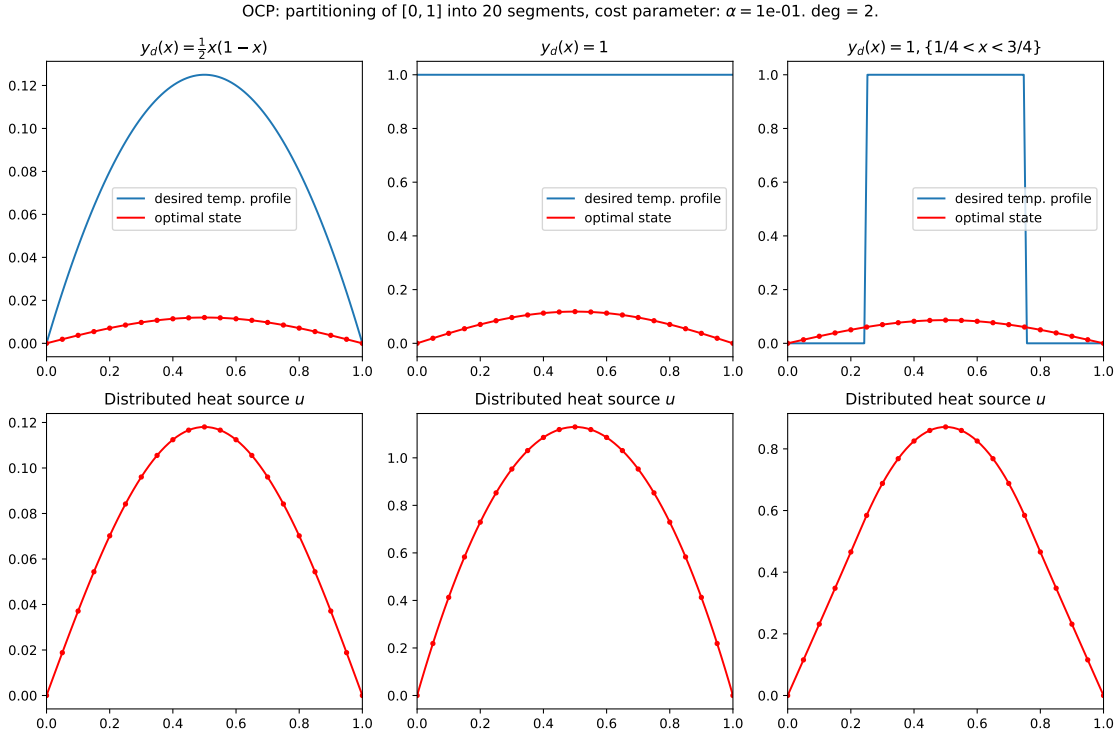

Figure 4: An optimized control problem for various desired temperature profiles. The upper row plots the desired temperature profiles and the corresponding optimal state. The lower row plots the optimal control for the corresponding problem. The cost parameter, α, is set to high value. To find the optimal state and the optimal control we have partitioned the interval into 20 elements. A second degree Lagrange finite element space has been used.

Bibliography

[1] C. Curry. *TMA4212 Part 2: Introduction to finite element methods*. 2018.

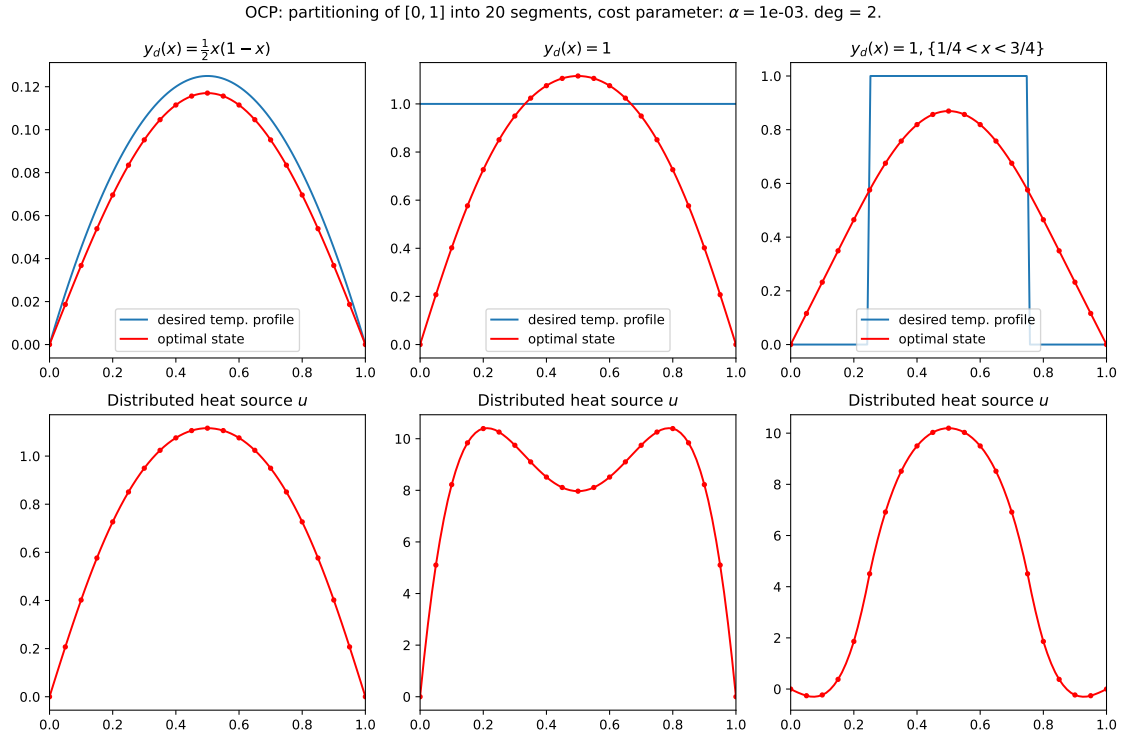

Figure 5: An optimized control problem for various desired temperature profiles. The upper row plots the desired temperature profiles and the corresponding optimal state. The lower row plots the optimal control for the corresponding problem. The cost parameter, α, is set to typical value. To find the optimal state and the optimal control we have partitioned the interval into 20 elements. A second degree Lagrange finite element space has been used.

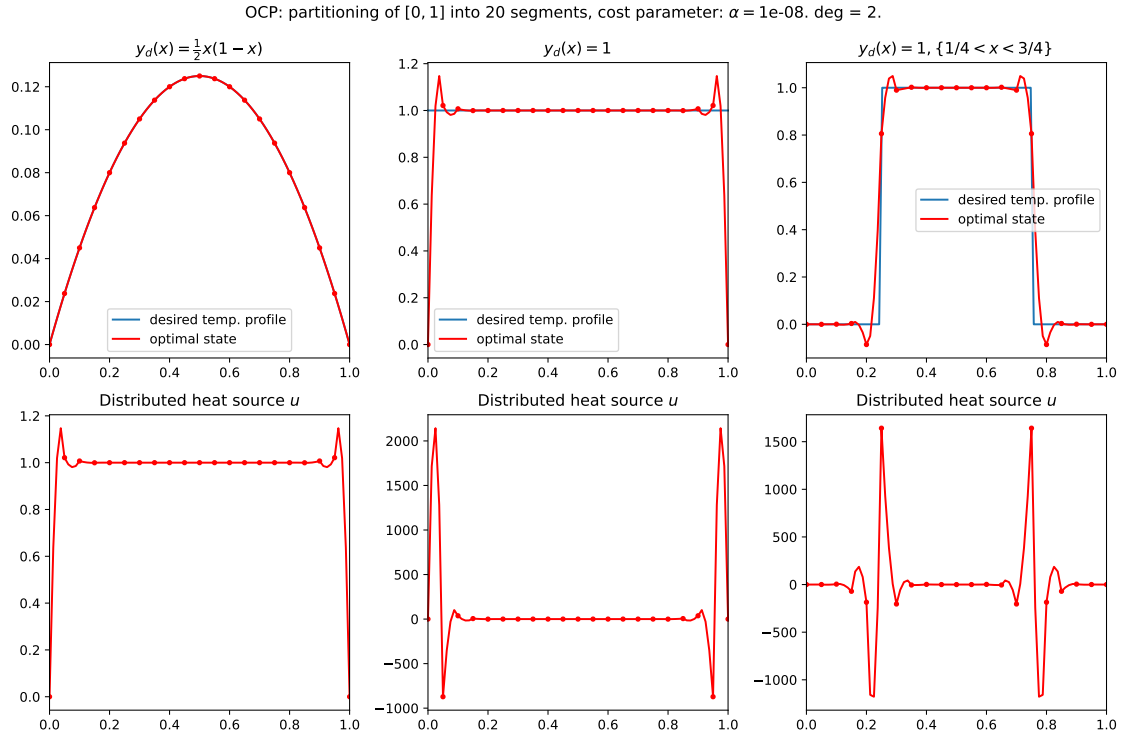

Figure 6: An optimized control problem for various desired temperature profiles. The upper row plots the desired temperature profiles and the corresponding optimal state. The lower row plots the optimal control for the corresponding problem. The cost parameter, α, is set to low value. To find the optimal state and the optimal control we have partitioned the interval into 20 elements. A second degree Lagrange finite element space has been used.