

Efficient Solution of State-Constrained Distributed Parabolic Optimal Control Problems

Richard Löscher^[0000-0002-6155-1178], Michael Reichelt^[0000-0002-9015-3419], and Olaf Steinbach^[0000-0002-2552-3022]

1 Introduction

Optimal control problems arise naturally in a wide range of applications such as cancer treatment [1] or shape optimization [2]. For a recent overview, see [4]. In this work, we consider optimal control problems of tracking type subject to the heat equation. It reads to minimize the functional \mathcal{J} , given by

$$\mathcal{J}(u_\varrho, z_\varrho) = \frac{1}{2} \int_0^T \int_\Omega [u_\varrho(x, t) - \bar{u}(x, t)]^2 dx dt + \frac{1}{2} \varrho \|z_\varrho\|_X^2, \quad (1)$$

subject to the Dirichlet boundary value problem for the heat equation,

$$\begin{aligned} \partial_t u_\varrho(x, t) - \Delta_x u_\varrho(x, t) &= z_\varrho(x, t) & \text{for } (x, t) \in Q &:= \Omega \times (0, T), \\ u_\varrho(x, t) &= 0 & \text{for } (x, t) \in \Sigma &:= \partial\Omega \times (0, T), \\ u_\varrho(x, 0) &= 0 & \text{for } x \in \Omega, \end{aligned} \quad (2)$$

where $\Omega \subset \mathbb{R}^n$, $n = 2, 3$ is a Lipschitz domain and $T > 0$ is a finite time horizon. We call Q the space-time cylinder and Σ its lateral boundary. Further \bar{u} is a given target, z_ϱ is called control, u_ϱ is called state and ϱ is a regularity parameter, which has to be properly chosen. As a mental model imagine Ω to be a room, where a desired (possibly discontinuous) temperature distribution \bar{u} should be reached and heat sources, represented by z_ϱ , can be placed at arbitrary positions in space-time. We seek to find the minimum for $u_\varrho \in X = H_{0,0}^{1,1/2}(Q) := L^2(0, T; H_0^1(\Omega)) \cap H_{0,0}^{1/2}(0, T; L^2(\Omega))$. This anisotropic Sobolev space and its relation to the heat equation has been investigated in e.g. [11]. In contrast to the wide-spread choice of measuring z_ϱ in the L^2 -norm, we use the dual norm with respect to X . This has proven to have many advantages

Richard Löscher, Michael Reichelt, Olaf Steinbach
Institute of Applied Mathematics, TU Graz, Steyrergasse 30, 8010 Graz, Austria
e-mail: {loescher, michael.reichelt, o.steinbach}@tugraz.at

from a theoretical and numeric perspective [6, 8, 9]. For brevity of notation, define the operator $B : X \rightarrow X^*$ by

$$\langle Bu_\varrho, v \rangle_Q := \langle \partial_t u_\varrho, v \rangle_Q + \langle \nabla_x u_\varrho, \nabla_x v \rangle_{L^2(Q)}, \quad (3)$$

where $\langle \cdot, \cdot \rangle_Q$ denotes the duality pairing as the extension of the L^2 -inner product with integral over the space-time cylinder Q . In [10] we have shown in detail, that

$$\|z_\varrho\|_{X^*} = \|Bu_\varrho\|_{X^*} \simeq \|u_\varrho\|_X \quad (4)$$

are equivalent norms and one may replace the original functional (1) by a state based one

$$\mathcal{J}(u_\varrho) = \frac{1}{2} \|u_\varrho - \bar{u}\|_{L^2(Q)}^2 + \frac{1}{2} \varrho \|u_\varrho\|_X^2, \quad (5)$$

as ϱ is anyway chosen by an asymptotic rule. The norm in X is realized via the operator D , defined by

$$\|u\|_{H_{0,0}^{1,1/2}(Q)}^2 := \langle \partial_t u, \mathcal{H}_T u \rangle_Q + \|\nabla_x u\|_{L^2(Q)}^2 =: \langle Du, u \rangle_Q = \|u\|_D^2, \quad (6)$$

where \mathcal{H}_T is the modified Hilbert transform [11], that only acts in the temporal direction. For given $u \in L^2(0, T)$ with Fourier series

$$u(t) = \sum_{k=0}^{\infty} u_k \sin\left(\left(\frac{\pi}{2} + k\pi\right)\frac{t}{T}\right), \quad u_k = \frac{2}{T} \int_0^T u(t) \sin\left(\left(\frac{\pi}{2} + k\pi\right)\frac{t}{T}\right) dt,$$

it is defined by

$$\mathcal{H}_T u(t) := \sum_{k=0}^{\infty} u_k \cos\left(\left(\frac{\pi}{2} + k\pi\right)\frac{t}{T}\right).$$

Albeit this is a non-local operator, we can directly compute the norm induced by D , and we will demonstrate in Section 2 an efficient implementation for function spaces, that can be written as a tensor product of a spatial finite element space with the space of piecewise linear continuous functions on a regular mesh of the time interval $(0, T)$. With the state based formulation, we can either minimize over the whole space X , which we call the unconstrained problem, or we can restrict the domain to a convex subset $\mathcal{K} \subset X$, which comprises state and control constraints as shown in [3]. However, this work only considers state constraints, where the convex set is given by

$$\mathcal{K} = \{u \in X \mid u_- \leq u \leq u_+ \text{ a.e. in } Q\} \quad (7)$$

and $u_- \leq 0, 0 \leq u_+$ are given continuous functions with $(\partial_t - \Delta_x)u_\pm \in L^2(Q)$. The constrained problem is nonlinear and can be analyzed in the framework of active set strategies. For the unconstrained problem, in [10] we derived the state based optimality system

$$(I + \varrho D) u_\varrho = \bar{u}, \quad (8)$$

derived error estimates for different classes of targets \bar{u} , as well as a quasi optimal solver on the aforementioned tensor product spaces. In Section 2 we briefly revisit the discretization, but instead of a solver, we develop a matrix free implementation of $I + \varrho D$ on the respective finite element space. We then use this matrix free operator to solve the discretized optimality system using a conjugate gradient method. This has the advantage, that we can augment this operator to yield an efficient algorithm for the active set strategy later on. The description of the active set strategy with the augmented operator is presented in Section 3. Numerical results for the state-constrained case are presented in Section 4.

2 Discretization and Efficient Implementation

This section tackles the question of discretizing the optimality system

$$(I + \varrho D) u_\varrho = \bar{u} \quad (9)$$

of the unconstrained problem. As D is non-local in time, we choose to use ansatz functions, that are tensor products of spatial and temporal functions. More precisely, let $W_{h_x} = \text{span}\{\psi_i\}_{i=1}^{M_x} \subset H_0^1(\Omega)$ be the spatial finite element space of piecewise linear basis functions ψ_i which are defined with respect to some admissible and globally quasi-uniform finite element mesh with spatial mesh size h_x . Further, let $V_{h_t} := S_{h_t}^1(0, T) \cap H_{0,\cdot}^{1/2}(0, T) = \text{span}\{\varphi_k\}_{k=1}^{N_t}$ be the space of piecewise linear functions, which are defined with respect to a uniform finite element mesh with temporal mesh size h_t . Then we construct the finite element ansatz space via

$$X_h := W_{h_x} \otimes V_{h_t} \subset X. \quad (10)$$

The Galerkin discretization of (9) leads to the equivalent system of linear equations $K_h \underline{u} = \underline{\bar{u}}$, where

$$K_h = M_{h_t} \otimes M_{h_x} + \varrho \left[A_{h_t} \otimes M_{h_x} + M_{h_t} \otimes A_{h_x} \right] \in \mathbb{R}^{N_t \cdot M_x \times N_t \cdot M_x}, \quad (11)$$

with

$$\begin{aligned} A_{h_t}[j, i] &= \langle \partial_t \varphi_i, \mathcal{H}_T \varphi_j \rangle_{L^2(0, T)}, & M_{h_t}[j, i] &= \langle \varphi_i, \varphi_j \rangle_{L^2(0, T)} & i, j &= 1, \dots, N_t, \\ A_{h_x}[\ell, k] &= \langle \nabla_x \psi_k, \nabla_x \psi_\ell \rangle_{L^2(\Omega)}, & M_{h_x}[\ell, k] &= \langle \psi_k, \psi_\ell \rangle_{L^2(\Omega)} & k, \ell &= 1, \dots, M_x. \end{aligned}$$

In [10] we have shown, that the optimal choice for the penalty parameter is $\varrho = h_x^2$ and that the matrix K_h is spectrally equivalent to the space-time mass matrix for this case. As the matrix is furthermore symmetric, this implies that a conjugate gradient method with diagonal preconditioning is applicable. Also in [10] we have solved the

generalized eigenvalue problem

$$A_{h_t} \underline{v} = \lambda M_{h_t} \underline{v} \quad (12)$$

analytically for equidistant temporal meshes. Moreover, we are able to transform any vector $\underline{v} \in V_{h_t}$ into the eigenvector basis of (A_{h_t}, M_{h_t}) with $\mathcal{O}(N_t \log N_t)$ effort, which translates to $\mathcal{O}(N_t M_x \log N_t)$ for any vector in X_h . Denote the temporal transformation matrix into the eigenvector basis by $C_{h_t}^{-1}$. The respective transformation on X_h is then given by $C_{h_t}^{-1} \otimes I_{h_x}$. We intend to use this, to compute the action

$$\underline{w} = K_h \underline{v}. \quad (13)$$

A short calculation yields

$$\underline{w} = (M_{h_t} C_{h_t} \otimes I_x) \hat{K}_h (C_{h_t}^{-1} \otimes I_x) \underline{v}, \quad (14)$$

where \hat{K}_h is the representation of K_h in the temporal eigenbasis

$$\begin{aligned} \hat{K}_h &= (C_{h_t}^{-1} M_{h_t}^{-1} \otimes I_{h_x}) K_h (C_{h_t} \otimes I_{h_x}) \\ &= I_{h_t} \otimes M_{h_x} + \varrho (\Lambda_{h_t} \otimes I_{h_x} + I_{h_t} \otimes A_{h_x}) \end{aligned} \quad (15)$$

and Λ_{h_t} is the diagonal matrix of the generalized eigenvalues λ_i of (12). The application of \hat{K}_h has effort $\mathcal{O}(N_t M_x)$, so we end up with the overall effort of $\mathcal{O}(N_t M_x \log N_t)$ for the matrix free application of K_h . This is a significant improvement over the $\mathcal{O}(N_t^2 M_x)$ effort for the direct application of K_h . Together with the feasibility of the conjugate gradient method with diagonal preconditioning, this yields a quasi optimal solver. Additionally, shared memory parallelization can be easily implemented, due to the Kronecker product structure of the arising matrices.

3 State Constraints

In the last section, we have derived a finite element discretization of the unconstrained problem. In this section, we will allow state constraints, i.e. we will consider the optimization problem on the convex set $K \subset X$, which we discretize in the following way

$$\mathcal{K}_h = \{ \underline{u} \in X_h \mid \underline{u}^- \leq \underline{u} \leq \underline{u}^+ \}, \quad (16)$$

where \underline{u}^\pm are defined via interpolation of the functions u_\pm on X_h and the inequality is element-wise. Note, that \mathcal{K}_h is still a convex set on the continuous level. For example in [3, 5] the authors show that the discrete (convex) optimization problem, to find $u_h \in K_h$ such that

$$\mathcal{J}(u_h) = \frac{1}{2} \|u_h - \bar{u}\|_{L^2(\mathcal{Q})}^2 + \frac{1}{2} \varrho \|u_h\|_X^2 \rightarrow \min \quad (17)$$

is equivalent to the nonlinear system of equations

$$\underline{F}(\underline{u}, \underline{\lambda}) = \underline{0}, \quad (18)$$

where for any $c > 0$, we can define the function \underline{F} as

$$\underline{F}(\underline{u}, \underline{\lambda}) = \begin{pmatrix} K_h \underline{u} - \underline{\lambda} - \bar{u} \\ \underline{\lambda} - \min\{0, \underline{\lambda} + c(\underline{u}^+ - \underline{u})\} - \max\{0, \underline{\lambda} + c(\underline{u}^- - \underline{u})\} \end{pmatrix}. \quad (19)$$

One step of the semismooth Newton method for this system reads as

$$\begin{pmatrix} \underline{u}^{k+1} \\ \underline{\lambda}^{k+1} \end{pmatrix} = \begin{pmatrix} \underline{u}^k \\ \underline{\lambda}^k \end{pmatrix} - [D\underline{F}(\underline{u}^k, \underline{\lambda}^k)]^{-1} \underline{F}(\underline{u}^k, \underline{\lambda}^k) \quad (20)$$

where $D\underline{F}$ is the Jacobian in the sense of slant derivatives. If one starts solving this system starting with the second line, after a lengthy calculation fully presented in [3], it turns out that \underline{u}^{k+1} and $\underline{\lambda}^{k+1}$ are complementary in the sense, that

$$\begin{cases} \underline{\lambda}^k[i] + c(\underline{u}^-[i] - \underline{u}^k[i]) > 0 \\ \underline{\lambda}^k[i] + c(\underline{u}^+[i] - \underline{u}^k[i]) < 0 \\ \text{else} \end{cases} \implies \begin{cases} \underline{u}^{k+1}[i] = \underline{u}^-[i] \\ \underline{u}^{k+1}[i] = \underline{u}^+[i] \\ \underline{\lambda}^{k+1}[i] = 0. \end{cases} \quad (21)$$

The index set, where $\underline{\lambda}^k$ is non-vanishing, is called the active set \mathcal{I}_A^k and the complementary set is called the inactive set \mathcal{I}_I^k . This consideration fully eliminates the second row of (20). For the remainder, we first introduce a split into active and inactive parts of our quantities

$$\underline{u}^{k+1} = \underline{u}_A^{k+1} \oplus \underline{u}_I^{k+1}, \quad \underline{\lambda}^{k+1} = -\underline{\lambda}_A^{k+1} \oplus \underline{\lambda}_I^{k+1}, \quad (22)$$

where $\underline{\lambda}_I^{k+1}[i] = 0$ for $i \in \mathcal{I}_I^{k+1}$ and $\underline{u}_A^{k+1}[i]$ is set according to (21) for $i \in \mathcal{I}_A^{k+1}$. Substituting this into the remaining first equation of (20) yields

$$\tilde{K}_h^{k+1} \begin{pmatrix} \underline{u}_I^{k+1} \\ \underline{\lambda}_A^{k+1} \end{pmatrix} := K_h \underline{u}_I^{k+1} + \underline{\lambda}_A^{k+1} = \bar{u} - K_h \underline{u}_A^{k+1} =: \underline{f}^{k+1}. \quad (23)$$

The left-hand side is a linear mapping $\mathbb{R}^{N_I M_x} \rightarrow \mathbb{R}^{N_I M_x}$ for every choice of active set. Due to the orthogonality of the splitting, $K_h \underline{u}_I^{k+1} + \underline{\lambda}_A^{k+1}$ is also an orthogonal sum and hence \tilde{K}_h^{k+1} inherits positive definiteness from K_h and the identity. After a correct reordering of the indices, the matrix is even block diagonal

$$\tilde{K}_h^{k+1} = \begin{pmatrix} R_I^{k+1} K_h^{k+1} P_I^{k+1} & 0 \\ 0 & R_A^{k+1} I_h P_A^{k+1} \end{pmatrix}, \quad (24)$$

where $P_{I/A}^{k+1}$ and $R_{I/A}^{k+1}$ are the canonical prolongation (by zero) and restriction with respect to their subscript index set. Now feasibility of diagonal preconditioning is evident from its feasibility for K_h . In practice the splitting is performed by setting vanishing components to zero. Then all matrices can be applied as in the unconstrained case. As \tilde{K}_h is not available as a matrix, we take the diagonal of the spectrally equivalent space-time mass for preconditioning, where we set entries corresponding to the active set to one. The semismooth Newton is converged, when the index sets are no longer changing. But remember, that this is a nonlinear problem and convergence is dependent on the initial guesses for \underline{u} and $\underline{\lambda}$! The algorithm for the overall procedure is summarized in Algorithm 1. Note, that the simple stopping criterion is no longer applicable, when e.g. underrelaxation or line-search is used. Then one needs to add additional stopping criteria. Underrelaxation is discussed in the next chapter.

Algorithm 1 Active Set Strategy for State Constraints

```

choose initial guesses for  $\underline{u}^0$  and  $\underline{\lambda}^0$ 
for  $k = 0, 1, 2, \dots$  do
  compute  $\mathcal{I}_A^{k+1}$  and  $\mathcal{I}_I^{k+1}$  according to (21)
  if  $k > 0$  and  $\mathcal{I}_A^{k+1} = \mathcal{I}_A^k$  then
    return  $(\underline{u}^k, \underline{\lambda}^k)$ 
  end if
  solve (23) for  $\underline{u}_I^{k+1}$  and  $\underline{\lambda}_A^{k+1}$  using matrix free CG with diagonal preconditioning
end for

```

4 Numerical Results

As an illustrative example, we consider the space-time cylinder $Q = \Omega \times (0, T)$, where $\Omega = (0, 1)^3$ and $T = 1$. We decompose Ω into a regular simplicial mesh Ω_h with n_x elements per spatial direction and $(0, T)$ into a regular mesh \mathcal{T}_h consisting of n_t equidistant intervals. Then we define our discrete space $X_h = W_{h_x} \otimes V_{h_t}$ with $W_{h_x} = S^1(\Omega_h) \cap H_0^1(\Omega)$ and $V_{h_t} = S^1(\mathcal{T}_h) \cap H_0^{1/2}(0, T)$, where S^1 denotes the space of piecewise linear functions on the respective mesh. We stress, that the simple mesh Ω_h is only for illustrational purposes, and that the method is applicable to any conforming, shape regular and quasi-uniform spatial mesh. The optimal choice for the regularization parameter is $\varrho = h_x^2 = n_x^{-2}$ and as target function we use

$$\bar{u}(x, t) = \sin(\pi x_1) \sin(\pi x_2) \sin(\pi x_3) \sin(\pi t) \in C^\infty(Q) \cap X. \quad (25)$$

As we have extensively validated the optimality of the strongly related solver for the unconstrained case in [10], we will only present the more interesting restricted case. We choose $u_- \equiv 0$ and $u_+ \equiv 0.8$. This implies that $\bar{u} \notin K$, even though the function is infinitely smooth. As start value for u we take the mean of the bounding functions

and we choose the Lagrange multiplier according to the first line of the root-finding problem (19), i.e.

$$\underline{u}^0 = \frac{1}{2} (\underline{u}^- + \underline{u}^+) \quad \text{and} \quad \underline{\lambda}^0 = K_h \underline{u}^0 - \underline{u}. \quad (26)$$

Further we apply underrelaxation with $\omega_N = 0.1$ in the semismooth Newton and the parameter $c = 1$. We stop the non-linear solver if the active sets do not change anymore and subsequent iterates fulfill

$$\|\underline{u}^{k+1} - \underline{u}^k\|_\infty + \|\underline{\lambda}^{k+1} - \underline{\lambda}^k\|_\infty < 10^{-3}. \quad (27)$$

The second condition is needed, due to underrelaxation, which could lead to a non-changing active set, even though the solution is still changing. For this illustrative example, simple underrelaxation is sufficient, but in general, more sophisticated strategies are advised. To gain insight on the condition number of \tilde{K}_h each nested Conjugate Gradient starts with zero initial guess and stops at a relative residual of 10^{-10} . The experiment is conducted for varying mesh sizes $n = n_x = n_t$. The number of needed Newton as well as CG iterations are recorded and listed in Table 1. Furthermore the table contains the ratio of CG iterations to Newton iterations, which stay at reasonable values. In Figure 1 we plot the solution for different discretization parameters n along the line $x_1 = x_2 = x_3 = 0.51$ and compare it to the target function. A point slightly off the center is chosen to exclude benevolent symmetry effects. It is clearly visible, that the solution approaches the target function until it reaches the upper bound. This is exactly the desired behavior.

Table 1 Numerical results for the constrained optimal control problem, with $n = n_t = n_x$.

n	DoF	Newton iter.	CG iter.	CG/Newton
2	16	36	36	1
4	256	36	612	17
8	4,096	36	1,296	36
16	65,536	38	2,173	57
32	1,048,580	64	3,814	60

Acknowledgements Part of this work has been supported by the Austrian Science Fund (FWF) under the Grant Collaborative Research Center TRR361/F90: CREATOR Computational Electric Machine Laboratory.

[7]

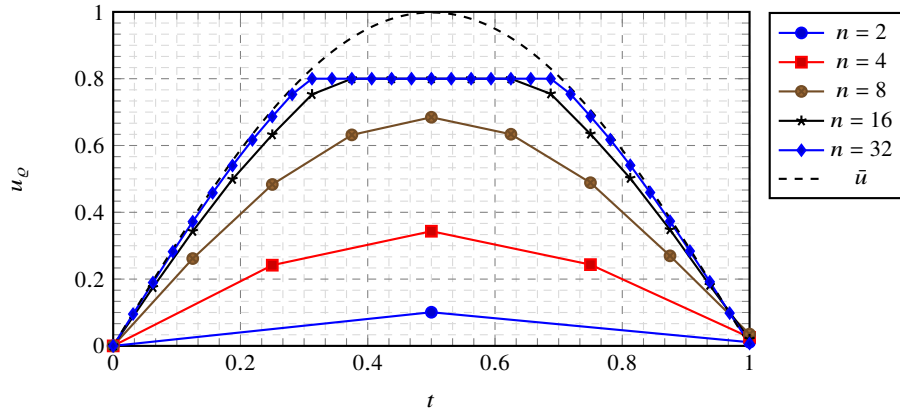


Fig. 1 Plot of the constrained solution u_Q for different discretization parameters $n = n_t = n_x$ along the line $x_1 = x_2 = x_3 = 0.51$.

References

1. P. Deuffhard, A. Schiela, and M. Weiser. Mathematical cancer therapy planning in deep regional hyperthermia. *Acta Numer.*, 21:307–378, 2012.
2. P. Gangl, S. Köthe, C. Mellak, A. Cesarano, and A. Mütze. Multi-objective free-form shape optimization of a synchronous reluctance machine. *COMPEL*, 41(5):1849–1864, 2022.
3. P. Gangl, R. Löscher, and O. Steinbach. Regularization and finite element error estimates for elliptic distributed optimal control problems with energy regularization and state or control constraints. *Comput. Math. Appl.*, 180:242–260, 2025.
4. R. Herzog, M. Heinkenschloss, D. Kalise, G. Stadler, and E. Trélat, editors. *Optimization and Control for Partial Differential Equations*. de Gruyter, Berlin, Boston, 2022.
5. M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888, 2002.
6. U. Langer, O. Steinbach, F. Tröltzsch, and H. Yang. Space-time finite element discretization of parabolic optimal control problems with energy regularization. *SIAM J. Numer. Anal.*, 59:675–695, 2021.
7. U. Langer, O. Steinbach, F. Tröltzsch, and H. Yang. Unstructured space-time finite element methods for optimal control of parabolic equations. *SIAM J. Sci. Comput.*, 43:A744–A771, 2021.
8. U. Langer, O. Steinbach, and H. Yang. Robust space-time finite element methods for parabolic distributed optimal control problems with energy regularization. *Adv. Comput. Math.*, 50:24, 2024.
9. U. Langer and M. Zank. Efficient direct space-time finite element solvers for parabolic initial-boundary value problems in anisotropic Sobolev spaces. *SIAM J. Sci. Comput.*, 43:A2714–A2736, 2021.
10. R. Löscher, M. Reichelt, and O. Steinbach. Optimal complexity solution of space-time finite element systems for state-based parabolic distributed optimal control problems. *J. Complexity*, 92:101976, 2026.
11. O. Steinbach and M. Zank. Coercive space-time finite element methods for initial boundary value problems. *Electron. Trans. Numer. Anal.*, 52:154–194, 2020.