# OPTIMAL CONTROL OF THE THERMISTOR PROBLEM IN THREE SPATIAL DIMENSIONS, PART 2: OPTIMALITY CONDITIONS

H. MEINLSCHMIDT[†], C. MEYER[‡], J. REHBERG[§]

**Abstract.** This paper is concerned with the state-constrained optimal control of the three-dimensional thermistor problem, a fully quasilinear coupled system of a parabolic and elliptic PDE with mixed boundary conditions. This system models the heating of a conducting material by means of direct current. Local existence, uniqueness and continuity for the state system as well as existence of optimal solutions, admitting global-in-time solutions, to the optimization problem were shown in the the companion paper of this work. In this part, we address further properties of the set of controls whose associated solutions exist globally such as openness, which includes analysis of the linearized state system via maximal parabolic regularity. The adjoint system involving measures is investigated using a duality argument. These results allow to derive first-order necessary conditions for the optimal control problem in form of a qualified optimality system in which we do not need to refer to the set of controls admitting global solutions. The theoretical findings are illustrated by numerical results. This work is the second of two papers on the three-dimensional thermistor problem.

**Key words.** Partial differential equations, optimal control problems, state constraints

**AMS subject classifications.** 35K59, 35M10, 49K20

## 1. Introduction.

In this paper, we consider the state-constrained optimal control of the three-dimensional thermistor problem. In detail the optimal control problem under consideration looks as follows:

$$\left.\begin{array}{ll} \min & \dfrac{1}{2}\|\theta(T_1)-\theta_d\|^2_{L^2(E)}+\dfrac{\gamma}{s}\|\nabla\theta\|^s_{L^s(T_0,T_1;L^q(\Omega))}+\dfrac{\beta}{2}\displaystyle\int_{\Sigma_N}(\partial_t u)^2+|u|^p\,\mathrm{d}\omega\,\mathrm{d}t \\ \text{s.t.} & (1.1)\text{–}(1.6) \\ \text{and} & \qquad\theta(x,t)\le\theta_{\max}(x,t)\quad \text{a.e. in } \Omega\times(T_0,T_1), \\ & 0\le u(x,t)\le u_{\max}(x,t)\quad \text{a.e. on } \Gamma_N\times(T_0,T_1) \end{array}\right\}\quad (\mathrm{P})$$

where (1.1)–(1.6) refer to the following coupled PDE system consisting of the instationary nonlinear heat equation and the quasi-static potential equation, which is also known as *thermistor problem*:

$$\partial_t\theta-\operatorname{div}(\eta(\theta)\kappa\nabla\theta)=(\sigma(\theta)\varepsilon\nabla\varphi)\cdot\nabla\varphi \qquad \text{in } Q:=\Omega\times(T_0,T_1) \tag{1.1}$$

$$\nu\cdot\kappa\nabla\theta+\alpha\theta=\alpha\theta_l \qquad \text{on } \Sigma:=\partial\Omega\times(T_0,T_1) \tag{1.2}$$

$$\theta(T_0)=\theta_0 \qquad \text{in } \Omega \tag{1.3}$$

$$-\operatorname{div}(\sigma(\theta)\varepsilon\nabla\varphi)=0 \qquad \text{in } Q \tag{1.4}$$

$$\nu\cdot\sigma(\theta)\varepsilon\nabla\varphi=u \qquad \text{on } \Sigma_N:=\Gamma_N\times(T_0,T_1) \tag{1.5}$$

$$\varphi=0 \qquad \text{on } \Sigma_D:=\Gamma_D\times(T_0,T_1). \tag{1.6}$$

[†] Faculty of Mathematics, TU Darmstadt, Dolivostrasse 15, D-64283 Darmstadt, Germany.

[‡] Faculty of Mathematics, Technical University of Dortmund, Vogelpothsweg 87, D-44227 Dortmund, Germany.

[§] Weierstrass Institute for Applied Analysis and Stochastics Mohrenstr. 39, D-10117 Berlin, Germany.

Here $\theta$ is the temperature in a conducting material covered by the three dimensional domain $\Omega$, while $\varphi$ refers to the electric potential. The boundary of $\Omega$ is denoted by $\partial\Omega$ with the unit normal $\nu$ facing outward of $\Omega$ in almost every boundary point (w.r.t. the boundary measure $\omega$). In addition, for the boundary we have $\Gamma_D \dot{\cup} \Gamma_N = \partial\Omega$, where $\Gamma_D$ is closed within $\partial\Omega$. The functions $\eta(\cdot)\kappa$ and $\sigma(\cdot)\varepsilon$ represent heat- and electric conductivity. While $\kappa$ and $\varepsilon$ are given, prescribed functions, $\eta$ and $\sigma$ are allowed to depend on the temperature $\theta$. Moreover, $\alpha$ is the heat transfer coefficient and $\theta_l$ and $\theta_0$ are given boundary– and initial data, respectively. Finally, $u$ stands for a current which is induced via the boundary part $\Gamma_N$ and is to be controlled. The bounds in the optimization problem (P) as well as the desired temperature $\theta_d$ are given functions and $\beta$ is the usual Tikhonov regularization parameter. The precise assumptions on the data in (P) and (1.1)–(1.6) will be specified in §2. In all what follows, the system (1.1)–(1.6) is frequently also called *state system*.

The PDE system (1.1)–(1.6) models the heating of a conducting material by means of a direct current, described by $u$, induced on the part $\Gamma_N$ of the boundary, which is done for some time $T_1 - T_0$. At the grounding $\Gamma_D$, homogeneous Dirichlet boundary conditions are given, i.e., the potential is zero, inducing electron flow. Note that, usually, $u$ will be zero on a subset $\Gamma_{N_0}$ of $\Gamma_N$, which corresponds to having insulation at this part of the boundary. We emphasize that the different boundary conditions are essential for a realistic modeling of the process. The objective of (P) is to adjust the induced current $u$ to minimize the $L^2$-distance between the desired and the resulting temperature at end time $T_1$ on the set $E \subseteq \Omega$, the latter representing the area of the material in which one is interested – realized in the objective functional by the first term. The other terms are present to minimize thermal stresses (second term) and to ensure a certain smoothness of the controls (third term), whose influence to the objective functional, however, may be controlled by the weights $\gamma$ and $\beta$. The actual form of these terms and the size of the integrability orders are motivated by functional-analytic considerations, see [33, §4]. Moreover, the optimization is subject to pointwise control and state constraints. The control constraints reflect a maximum heating power, while the state constraints limit the temperature evolution to prevent possible damage, e.g. by melting of the material. Similarly to the mixed boundary conditions, the inequality constraints in (P) are essential for a realistic model as demonstrated by the numerical example within this paper. Problem (P) is relevant in various applications, such as for instance the heat treatment of steel by means of an electric current. The example considered in the numerical part of this paper deals with an application of this type.

The state system (1.1)–(1.6) exhibits some non-standard features, in particular due to the quasilinear coupling of the parabolic and the elliptic PDE, the mixed boundary conditions in (1.5)–(1.6), and the inhomogeneity in the heat equation (1.1) as well as the temperature-dependent heat conduction coefficients. Besides the quasilinear state system, the pointwise state constraints on the temperature represent another challenging feature of the optimal control problem under consideration. The Lagrange multipliers associated with constraints of this kind only provide poor regularity in general, which especially complicates the analysis of the adjoint equation.

We briefly describe the genuine aspects of our work. First of all, the discussion of the quasilinear state system alone requires sophisticated up-to-date tools from maximal elliptic and parabolic regularity theory. This concerns already local-in-time existence for solutions of (1.5)–(1.6) as carried out in the companion paper [33], but also the characterization of global-in-time solutions and continuous differentiability for the control-to-state operator. The corresponding maximal regularity results were established only recently, see e.g. [5, 21, 24] for the parabolic case and [30, Appendix], [14] for the elliptic one. Still, the analysis performed only guarantees the local-in-time existence. However, using the implicit function theorem, we show that set of control functions admitting solutions which exist globally in time form an open set, which is essential for the derivation of optimality conditions in qualified form that are useful for numerical computations. In particular, the fact the set of "global" controls is open allows to obtain in essence the same optimality system as one would obtain if one always had global-in-time solutions for every admissible control. In the derivation of first order necessary optimality conditions, we also have to consider the nonstandard second and third term in the objective functional which allowed to show existence of optimal controls in the companion paper [33]. Lastly, we also give a numerical example which underlines the necessity of both control- and state bounds in (P).

Let us put our work into perspective. Up to the authors' best knowledge, there are only few contributions dealing with the optimal control of the thermistor problem. We refer to [31, 11, 29], where two-dimensional problems are discussed. In [31], a completely parabolic problem is discussed, while [29] considers the purely elliptic counterpart to (1.1)–(1.6). In [11, 3], the authors investigate a parabolic-elliptic system similar to (1.1)–(1.6), assuming a particular structure of the controls. In contrast to [31, 29], mixed boundary conditions are considered in [11]. However, all these contributions do not consider pointwise state constraints and non-smooth data. Thus, (P) differs significantly from the problems considered in the aforementioned papers. In a previous paper [27], two of the authors investigated the two-dimensional counterpart of (P). This contribution also accounts for mixed boundary conditions, non-smooth data, and pointwise state constraints. However, the analysis in [27] substantially differs from the three dimensional case considered here. The treatment of the state system in [27] heavily rests on the classical $W^{1,p}$-results from the classical paper [19], which are no longer sufficient for the three-dimensional problem. Moreover, the heat conduction coefficient in (1.1) is assumed not to depend on the temperature in [27]. Both features allow to derive a global existence result for a suitable class of control functions. Hence, main aspects of the present work do not appear in the two-dimensional setting. Let us finally take a broader look on state-constrained optimal control problems governed by PDEs. Compared to semilinear state-constrained optimal control problems, the literature concerning optimal control problems subject to quasilinear PDEs and pointwise state constraints is rather scarce. We exemplarily refer to [10, 9], where elliptic problems are studied. The vast majority of papers in this field deals with problems that possess a well defined control-to-state operator. By contrast, as indicated above, the state-system (1.1)–(1.6) in general just admits local-in-time solutions, which requires a sophisticated treatment of the optimal control problem under consideration.

The paper is organized as follows: We collect notations, assumptions and known results needed in the sequel in §2. Then we show that the set of controls admitting global solutions to the state system is open in §3, thereby also establishing continuous differentiability of the control-to-state operator for global solutions. The optimal control problem is considered in §4 and we give first order necessary conditions for (P) in qualified form. The theoretical findings are complemented with an illustrative numerical example in §5.

**2. Notations, general assumptions and known results.** We introduce some notation and the relevant function spaces. All function spaces under our consideration are *real* ones. Let, for now, $\Omega$ be a domain in $\mathbb{R}^3$. We give precise geometric specifications for $\Omega$ in §2.1 below.

Let us fix some notations: The underlying time interval is called $J = (T_0, T_1)$ with $T_0 < T_1$. The boundary measure for the domain $\Omega$ is called $\omega$. Generally, given an integrability order $q \in (1, \infty)$, we denote the conjugated of $q$ by $q'$, i.e., it always holds $1/q + 1/q' = 1$.

Definition 2.1. *For $q \in (1,\infty)$, let $W^{1,q}(\Omega)$ denote the usual Sobolev space on $\Omega$. If $\Xi \subset \partial\Omega$ is a closed part of the boundary $\partial\Omega$, we set $W^{1,q}_\Xi(\Omega)$ to be the closure of the set $\{\psi|_\Omega :\ \psi \in C_0^\infty(\mathbb{R}^3),\ \mathrm{supp}\ \ \psi \cap \Xi = \emptyset\}$ with respect to the $W^{1,q}$-norm.*

The dual space of $W^{1,q'}_\Xi(\Omega)$ is denoted by $W^{-1,q}_\Xi(\Omega)$; in particular, we write $W^{-1,q}_\emptyset(\Omega)$ for the dual of $W^{1,q'}(\Omega)$. The Hölder spaces of order $\delta$ on $\Omega$ or order $\varrho$ on $Q$ are denoted by $C^\delta(\Omega)$ and $C^\varrho(Q)$, respectively (note here that Hölder continuous functions on $\Omega$ or $Q$, respectively, possess an unique uniformly continuous extension to the closure of the domain, such that we will mostly use $C^\delta(\overline{\Omega})$ and $C^\varrho(\overline{Q})$ to emphasize on this).

We will usually abbreviate the function spaces on $\Omega$ by leaving out the $\Omega$, e.g. we write $W^{1,q}_\Xi$ instead of $W^{1,q}_\Xi(\Omega)$ or $L^p$ instead of $L^p(\Omega)$. Lebesgue spaces on subsets of $\partial\Omega$ are always to be considered with respect to the boundary measure $\omega$, but we abbreviate $L^p(\partial\Omega, \omega)$ by $L^p(\partial\Omega)$ and do so analogously for any $\omega$-measurable subset of the boundary. The norm in a Banach space $X$ will be always indicated by $\|\cdot\|_X$. For two Banach spaces $X$ and $Y$, we denote the space of linear, bounded operators from $X$ into $Y$ by $\mathcal{L}(X;Y)$. The symbol $\mathcal{LH}(X;Y)$ stands for the set of linear homeomorphisms between $X$ and $Y$. If $X, Y$ are Banach spaces which form an interpolation couple, then we denote by $(X,Y)_{\tau,r}$ the real interpolation space, see [36]. We use $M_3$ for the set of real, symmetric $3 \times 3$-matrices. In the sequel, a linear, continuous injection from $X$ to $Y$ is called an *embedding*, abbreviated by $X \hookrightarrow Y$. For Lipschitz continuous functions $f$, we denote the Lipschitz constants by $L_f$, while for bounded functions $g$ we denote their bound by $M_g$ (both over appropriate sets, if necessary). Finally, $c$ denotes a generic positive constant.

### 2.1. Geometric setting for $\Omega$ and $\Gamma_D$.

In all what follows, the symbol $\Omega$ stands for a bounded Lipschitz domain in $\mathbb{R}^3$ in the sense of [32, Ch. 1.1.9]; cf. [23] for the boundary measure $\omega$ on such a domain.

Remark 2.2. *The thus defined notion is different from* strong Lipschitz domain*, which is more restrictive and in fact identical with* uniform cone domain*, see again [32, Ch. 1.1.9]).*

Next we define the geometric setting for the domains $\Omega$ and the Dirichlet boundary part. For this, we denote by $K$ the open unit cube in $\mathbb{R}^n$, centered at $0 \in \mathbb{R}^n$, by $K_-$ the lower half cube $K \cap \{\mathrm{x}\colon x_n < 0\}$, by $\Sigma_K = K \cap \{\mathrm{x}\colon x_n = 0\}$ the upper plate of $K_-$ and by $\Sigma_K^0$ the left half of $\Sigma$, i.e. $\Sigma_K^0 = \Sigma_K \cap \{\mathrm{x}\colon x_{n-1} \le 0\}$.

Definition 2.3. *Let $\Xi \subset \partial\Omega$ be closed within $\partial\Omega$.*

(i) *We say that $\Omega \cup \Xi$ is* regular *(in the sense of Gröger), if for any point $\mathrm{x} \in \partial\Omega$ there is an open neighborhood $U_\mathrm{x}$ of $\mathrm{x}$, a number $a_\mathrm{x} > 0$ and a bi-Lipschitz mapping $\phi_\mathrm{x}$ from $U_\mathrm{x}$ onto $a_\mathrm{x}K$ such that $\phi_\mathrm{x}(\mathrm{x}) = 0 \in \mathbb{R}^3$, and we have either $\phi_\mathrm{x}\big((\Omega \cup \Xi) \cap U_\mathrm{x}\big) = a_\mathrm{x}K_-$ or $a_\mathrm{x}(K_- \cup \Sigma_K)$ or $a_\mathrm{x}(K_- \cup \Sigma_K^0)$.*

(ii) *The regular set $\Omega \cup \Xi$ is said to satisfy the* volume-conservation condition*, if each mapping $\phi_\mathrm{x}$ in Condition (i) is volume-preserving.*

Generally, $\Xi$ is allowed to be empty in Definition 2.3. Then Definition 2.3 (i) merely describes a Lipschitz domain. Some further comments are in order:

Remark 2.4.

(i) *Condition* (i) *exactly characterizes Gröger's* regular sets*, introduced in his pioneering paper [19]. Note that the volume-conservation condition also has been required in several contexts, cf. [17] and [20].*
*Clearly, the properties $\phi_{\mathrm{x}}(U_{\mathrm{x}}) = a_{\mathrm{x}}K$ and $\phi_{\mathrm{x}}\big(\Omega \cap U_{\mathrm{x}}\big) = a_{\mathrm{x}}K_-$ are already ensured by the Lipschitz property of $\Omega$; the crucial point is the behavior of $\phi_{\mathrm{x}}(\Xi \cap U_{\mathrm{x}})$.*

(ii) *A simplifying topological characterization of Gröger's regular sets in the case of three space dimensions reads as follows (cf. [22, Ch. 5]):*

1. *$\Xi$ is the closure of its interior within $\partial\Omega$,*
2. *the boundary $\partial\Xi$ within $\partial\Omega$ is locally bi-Lipschitz diffeomorphic to the open unit interval* $(0, 1)$*.*

(iii) *It turns out that regularity together with the volume-conservation condition is not a too restrictive assumption on the mapping $\phi_x$. In particular, there are such mappings—although not easy to construct—which map the ball onto the cylinder, the ball onto the cube and the ball onto the half ball, see [18, 16]. The general message is that this class has enough flexibility to map "non-smooth" objects onto smooth ones.*

(iv) *If $\Xi$ is nonempty and $\Omega \cup \Xi$ is regular, then $\Xi$ has interior points (with respect to the boundary topology in $\partial\Omega$), and, consequently, never has boundary measure* 0*.*

The following assumption is supposed to be valid for all the remaining considerations in the paper.

Assumption 2.5. *The set $\Omega \cup \Gamma_D$ is regular with $\Gamma_D \neq \emptyset$ and satisfies the volume-conservation condition.*

### 2.2. General assumptions and known results.

We collect known results from the companion paper of this work [33] and recall some basic definitions.

Definition 2.6. *Let $\Xi \subset \partial\Omega$ be closed. Assume that $\mu$ is any bounded, measurable, $M_3$-valued function on $\Omega$ and that $\gamma \in L^\infty(\partial\Omega \setminus \Xi)$ is nonnegative. We define the operators $-\nabla \cdot \mu\nabla$ and $-\nabla \cdot \mu + \tilde{\gamma}$, each mapping $W^{1,2}_\Xi$ into $W^{-1,2}_\Xi$, by*

$$\langle -\nabla \cdot \mu\nabla\psi, \xi\rangle := \int_\Omega \mu\nabla\psi \cdot \nabla\xi \,\mathrm{d}x \quad \textit{for} \quad \psi, \xi \in W^{1,2}_\Xi$$

*and*

$$\langle (-\nabla \cdot \mu\nabla + \tilde{\gamma})\psi, \xi\rangle = \langle -\nabla \cdot \mu\nabla\psi, \xi\rangle + \int_{\partial\Omega\setminus\Xi} \gamma\,\psi\,\xi \,\mathrm{d}\omega \quad \textit{for} \quad \psi, \xi \in W^{1,2}_\Xi.$$

*In all what follows, we maintain the same notation for the corresponding maximal restrictions to $W^{-1,q}_\Xi$, where $q > 2$.*

Remark 2.7. *Let us denote the domain for the operator $-\nabla \cdot \mu\nabla$, when restricted to $W^{-1,q}_\Xi$ $(q > 2)$, by $\mathcal{D}_q(\mu)$, equipped with the graph norm. Then the estimate*

$$\| -\nabla \cdot \mu\nabla\psi\|_{W^{-1,q}_\Xi} = \sup_{\|\varphi\|_{W^{1,q'}_\Xi}=1} \left| \int_\Omega \mu\nabla\psi \cdot \nabla\varphi \,\mathrm{d}x \right| \leq \|\mu\|_{L^\infty} \|\psi\|_{W^{1,q}_\Xi} \tag{2.1}$$

*shows that $W^{1,q}_\Xi$ is embedded in $\mathcal{D}_q(\mu)$ for every bounded coefficient function $\mu$. It is also known that $\mathcal{D}_q(\mu) \hookrightarrow C^\alpha(\overline{\Omega})$ for some $\alpha > 0$ whenever $q > 3$, see [22, Thm. 3.3]. Additionally,* (2.1) *implies that the mapping*

$$L^\infty(\Omega; M_3) \ni \mu \mapsto \nabla \cdot \mu\nabla \in \mathcal{L}(W^{1,q}_\Xi; W^{-1,q}_\Xi)$$

*is a linear and continuous contraction for every $q \in (1, \infty)$.*

The fundamental result from [19] characterizes $\mathcal{D}_q(\mu)$ for $q$ close to 2 uniformly for coefficient functions $\mu$ as follows:

PROPOSITION 2.8. *Let $\mu$ and $\gamma$ be as in Definition 2.6 and suppose that either $\omega(\Xi) > 0$ or $\Xi = \emptyset$ and $\int_{\partial\Omega} \gamma \,\mathrm{d}\omega > 0$. Then there is a number $q_0 > 2$ such that*

$$-\nabla \cdot \mu \nabla + \tilde{\gamma} \colon W^{1,q}_{\Xi} \to W^{-1,q}_{\Xi}$$

*is a topological isomorphism for all $q \in [2, q_0]$. The number $q_0$ may be chosen uniformly for all coefficient functions $\mu$ with the same ellipticity constant and the same $L^\infty$-bound. Moreover, for each $q \in [2, q_0]$, the norm of the inverse of $\nabla \cdot \mu \nabla + \tilde{\gamma}$ as a mapping from $W^{-1,q}_{\Xi}$ to $W^{1,q}_{\Xi}$ may be estimated again uniformly for all coefficient functions with the same ellipticity constant and the same $L^\infty$-bound.*

We impose the following assumptions on the quantities in the state system (1.1)–(1.6) and in the optimization problem (P):

ASSUMPTION 2.9.

(i) *The functions $\sigma : \mathbb{R} \to (0,\infty)$ and $\eta \colon \mathbb{R} \to (0,\infty)$ are bounded, Lipschitzian on any bounded interval and continuously differentiable, with bounded derivatives $\eta'$ and $\sigma'$, which are also Lipschitz continuous on bounded sets,*

(ii) *the function $\varepsilon \in L^\infty(\Omega; M_3)$ takes* symmetric *matrices as values, and satisfies the usual ellipticity condition, i.e.,*

$$\operatorname*{ess\,inf}_{x\in\Omega} \sum_{i,j=1}^{3} \varepsilon_{ij}(x)_{ij}\, \xi_i\, \xi_j \geq \underline{\varepsilon}\, \|\xi\|^2_{\mathbb{R}^3} \quad \forall\, \xi \in \mathbb{R}^3$$

*with a constant $\underline{\varepsilon} > 0$,*

(iii) *the function $\kappa \in L^\infty(\Omega; M_3)$ also takes* symmetric *matrices as values, and, additionally, satisfies an ellipticity condition, that is,*

$$\operatorname*{ess\,inf}_{x\in\Omega} \sum_{i,j=1}^{3} \kappa_{ij}(x)\, \xi_i\, \xi_j \geq \underline{\kappa}\, \|\xi\|^2_{\mathbb{R}^3} \quad \forall\, \xi \in \mathbb{R}^3$$

*holds with a constant $\underline{\kappa} > 0$,*

(iv) *there is a $q \in (3,4)$ such that the mappings*

$$-\nabla \cdot \varepsilon \nabla : W^{1,q}_{\Gamma_D} \to W^{-1,q}_{\Gamma_D}$$

*and*

$$-\nabla \cdot \kappa \nabla + 1 : W^{1,q} \to W^{-1,q}_{\emptyset}$$

*each provide a topological isomorphism,*

(v) *$\theta_l \in L^\infty(J; L^\infty(\partial\Omega))$,*

(vi) *$\alpha \in L^\infty(\partial\Omega)$ with $\alpha(x) \geq 0$ a.e. on $\partial\Omega$ and $\int_{\partial\Omega} \alpha \,d\omega > 0$,*

(vii) *$u \in L^{2r}(J; W^{-1,q}_{\Gamma_D})$ for $q > 3$ as assumed in (iv) above and $r > \frac{2q}{q-3}$,*

(viii) *the integrability exponents in the objective functional satisfy $p > \frac{4}{3}q - 2$ and $s > \frac{2q}{q-3}(1 - \frac{3}{q} + \frac{3}{\varsigma})$, where $q > 3$ is as in (iv), and $\varsigma := \frac{3\mathfrak{q}}{6-\mathfrak{q}}$ with $\mathfrak{q} \in (2, \min\{q_0, 3\}]$, here $q_0$ being the number from Proposition 2.8 for $(\mu,\gamma) = (\sigma\varepsilon, 0)$,*

(ix) *$E$ is an open (not necessarily proper) subset of $\Omega$,*

(x) *$\theta_d \in L^2(E)$,*

(xi) $\theta_{\max} \in C(\overline{Q})$ *with* $\max(\max_{\overline{\Omega}} \theta_0, \operatorname{ess\,sup}_\Sigma \theta_l) \le \theta_{\max}(x,t)$ *for all* $(x,t) \in \overline{Q}$ *and* $\theta_0(x) < \theta_{\max}(T_0, x)$ *for all* $x \in \overline{\Omega}$,

(xii) $u_{\max}$ *is a given function with* $u_{\max}(x,t) \ge 0$ *a.e. on* $\Sigma_N$,

(xiii) $\beta > 0$.

Let us remark that Assumption 2.9 (iv) is somewhat different in nature from the other assumptions. In essence, it requires that the—in general unknown—domains $\mathcal{D}_q(\varepsilon)$ and $\mathcal{D}_q(\kappa)$ of the restrictions of $-\nabla\cdot\varepsilon\nabla$ and $-\nabla\cdot\kappa\nabla + 1$ to $W^{-1,q}_{\Gamma_D}$ and $W^{-1,q}_{\emptyset}$ for $q > 3$ are in fact $W^{1,q}_{\Gamma_D}$ and $W^{1,q}$, cf. also Remark 2.7. This assumption is crucial for the analysis of the state system because of $n = 3$ and is discussed in detail in [30, Appendix] and [14], see also [33, Ass. 3.4/Rem. 3.25] and the comments there. Compare also with Proposition 2.8, where it is established that Assumption 2.9 (iv) is always true for a $q > 2$, which makes the assumption superfluous if one considers only space dimension $n = 2$.

Remark 2.10. *In Assumption 2.9 (vii), we implicitly made use of the embedding* $L^{\mathfrak{p}}(\Gamma_N) \hookrightarrow W^{-1,q}_{\Gamma_D}$ *for* $\mathfrak{p} > \frac{2}{3}q$, *realized by the adjoint operator of the continuous trace operator* $\tau_{\Gamma_N} : W^{1,q'}_{\Gamma_D} \to L^{\mathfrak{p}'}(\Gamma_N)$. *In this sense, a function* $u \in L^{2r}(J; L^{\mathfrak{p}}(\Gamma_N))$ *is considered as an element of* $L^{2r}(J; W^{-1,q}_{\Gamma_D})$. *Note that the assumption* $p > \frac{4}{3}q - 2$ *in Assumption 2.9 (viii) implies* $p > \frac{2}{3}q$ *due to* $q > 3$. *See also [28, Lemma 2.7] for the required embeddings/trace operators.*

We further define

$$\mathcal{A}(\zeta) := -\nabla\cdot\eta(\zeta)\kappa\nabla + \tilde{\alpha} \tag{2.2}$$

as a mapping $\mathcal{A}\colon C(\overline{\Omega}) \to \mathcal{L}(W^{1,q}; W^{-1,q}_{\emptyset})$. Due to Assumptions 2.9 (i) and (iv), the domain $\mathcal{D}_q(\eta(\zeta)\kappa)$ of the operator $-\nabla\cdot\eta(\zeta)\kappa\nabla + \tilde{\alpha}$ as a mapping into $W^{-1,q}_{\emptyset}$ is indeed $W^{1,q}$ for *every* $\zeta \in C(\overline{\Omega})$, as shown in [33, Cor. 3.8].

The following was the main existence and uniqueness result for the state system (1.1)-(1.6) as given in [33].

Theorem 2.11. *Suppose that Assumption 2.9 is true. Let* $q \in (3,4)$ *be the number for which Assumption 2.9 (iv) is satisfied,* $r > \frac{2q}{q-3}$ *and* $\theta_0 \in (W^{1,q}, W^{-1,q}_{\emptyset})_{\frac{1}{r},r}$. *Then there exists* $T_\bullet \in (T_0, T_1]$ *and a unique pair of functions*

$$\varphi \in L^{2r}(T_0, T_\bullet; W^{1,q}_{\Gamma_D}) \quad \textit{and} \quad \theta \in W^{1,r}(T_0, T_\bullet; W^{-1,q}_{\emptyset}) \cap L^r(T_0, T_\bullet; W^{1,q}).$$

*with* $\theta(T_0) = \theta_0$ *such that the operator equations*

$$\partial_t\theta(t) + \mathcal{A}(\theta(t))\theta(t) = (\sigma(\theta(t))\varepsilon\nabla\varphi(t))\cdot\nabla\varphi(t) + \alpha\theta_l(t) \qquad \textit{in } W^{-1,q}_{\emptyset}, \tag{2.3}$$

$$-\nabla\cdot\sigma(\theta(t))\varepsilon\nabla\varphi(t) = u(t) \qquad \textit{in } W^{-1,q}_{\Gamma_D} \tag{2.4}$$

*are satisfied for almost all* $t \in J$. *This solution is in particular Hölder-continuous on* $[T_0, T_\bullet] \times \overline{\Omega}$. *If* $T_\bullet = T_1$, *the pair* $(\theta, \varphi)$ *is a* global *solution.*

Let us collect some more items from the treatment of (2.3) and (2.4) from [33] which are needed in the sequel, thereby assuming that we are in the position of Theorem 2.11. The central point in the strategy to obtain unique solutions, albeit only local-in-time, is to reduce the problem to $\theta$ only, which includes solving (2.4) uniquely for $\varphi(t)$ for a given $\theta(t) \in C(\overline{\Omega})$, cf. [33, Thm. 3.23] – recall that $\theta(t)$ is uniformly continuous for all $t \in [T_0, T_\bullet]$ by Theorem 2.11. The corresponding solution operator is given by

$$C(\overline{\Omega}) \ni \zeta \mapsto \mathcal{J}(\sigma(\zeta)) := (-\nabla\cdot\sigma(\zeta)\varepsilon\nabla)^{-1} \in \mathcal{LH}(W^{-1,q}_{\Gamma_D}, W^{1,q}_{\Gamma_D}) \tag{2.5}$$

and we have $\varphi(t) = \mathcal{J}(\sigma(\theta(t)))u(t)$ for almost every $t$ from the interval of existence $[T_0, T_\bullet]$ associated to $u$. Accordingly, we will also call $\theta$ alone a *solution* as in Theorem 2.11, since $\varphi$ is uniquely determined by $\theta$ and $u$. Further, the right-hand side of (2.3) in dependence of $\theta(t)$ and $u(t)$ is given by

$$\Psi_{u(t)}(\theta(t)) := \nabla\left[\mathcal{J}(\sigma(\theta(t)))u(t)\right]\cdot\sigma(\theta(t))\varepsilon\nabla\left[\mathcal{J}(\sigma(\theta(t)))u(t)\right] \tag{2.6}$$

which is an element of $L^{q/2}$ for almost every $t \in [T_0, T_\bullet]$.

**3. Global solutions and the set of global controls.** In this section, we consider the set of controls which admit solutions existing *globally* in time, which is natural in view of the state constraints and the end time observation in the objective of (P). We define this set as follows: Under the assumptions of Theorem 2.11, we call a control $u \in L^{2r}(J; W^{-1,q}_{\Gamma_D})$ a *global control* if the corresponding solution $\theta_u$ as given in Theorem 2.11 exists on the whole prescribed interval $J = (T_0, T_1)$ and denote the set of global controls by $\mathcal{U}_g$. From [34, Thm. 5.3], we already know that $u \equiv 0$ is a global control, cf. [33, Cor. 5.3], and thus in particular $\mathcal{U}_g \neq \emptyset$. Note that this is the (only) main point where the volume-conservation property from Assumption 2.5 becomes relevant. Moreover, we define the *control-to-state operator*

$$\mathcal{S}\colon \mathcal{U}_g \ni u \mapsto \mathcal{S}(u) = \theta_u \in W^{1,r}(J; W^{-1,q}_\emptyset) \cap L^r(J; W^{1,q}) \tag{3.1}$$

on $\mathcal{U}_g$. Let us point out that it is possible to show that optimal global controls for (P) exist, cf. [33, §4], hence $\mathcal{U}_g$ is well-suited for the optimal control problem.

The goal for this section is to show that $\mathcal{U}_g$ is "nice" enough in the sense that the control-to-state operator is well-behaved and that we are able to derive first order necessary conditions for (P) which do not contain the, in general unknown, set $\mathcal{U}_g$. To be precise, in Theorem 3.1 below we show that $\mathcal{U}_g$ is a nonempty open set and that the control-to-state operator is continuously differentiable on $\mathcal{U}_g$.

We consider the assumptions of Theorem 2.11 to be fulfilled and fixed, that means, $q > 3$ and $r > \frac{2q}{q-3}$ are given from now on, with the control-to-state operator $u \mapsto \mathcal{S}(u) = \theta_u$ given as in (3.1). In particular, there is $\varrho > 0$ such that $\theta_u \in C^\varrho(\overline{J}; C^\varrho(\overline{\Omega}))$. We consider $\varphi = \varphi_u \in L^{2r}(J; W^{1,q}_{\Gamma_D})$ to be given in dependence of $u$ and $\theta_u$ using (2.5) as explained there. The following is the main theorem for this section:

THEOREM 3.1. *The set of global controls $\mathcal{U}_g$ forms an open set in $L^{2r}(J; W^{-1,q}_{\Gamma_D})$. Moreover, the control-to-state operator $\mathcal{S}$ is continuously differentiable. For every $h \in L^{2r}(J; W^{-1,q}_{\Gamma_D})$, its derivative $\zeta_h = \mathcal{S}'(u)h \in W^{1,r}(J; W^{-1,q}_\emptyset) \cap L^r(J; W^{1,q})$ is given by the unique solution of the equation*

$$\begin{aligned}\partial_t\zeta + \mathcal{A}(\theta_u)\zeta = (\sigma'(\theta_u)\zeta\varepsilon\nabla\varphi_u)\cdot\nabla\varphi_u + \nabla\cdot\eta'(\theta_u)\zeta\kappa\nabla\theta_u \\ - 2\left(\sigma(\theta_u)\varepsilon\nabla\varphi_u\right)\cdot\nabla\left[\mathcal{J}(\sigma(\theta_u))\left(-\nabla\cdot\sigma'(\theta_u)\zeta\varepsilon\nabla\varphi_u + h\right)\right],\end{aligned} \tag{3.2}$$

*which has to hold for almost every $t \in J$ in the space $W^{-1,q}_\emptyset$, with $\zeta(T_0) = 0$.*

*Proof.* Let $\bar{u} \in L^{2r}(J; W^{-1,q}_{\Gamma_D})$ be global, i.e., the associated solution $\theta_{\bar{u}} =: \bar{\theta}$ exists on the whole time horizon $(T_0, T_1)$. We intend to apply the implicit function theorem. To this end, we show that the mapping

$$\begin{aligned}\mathcal{B}\colon \Big(W^{1,r}(J; W^{-1,q}_\emptyset) \cap L^r(J; W^{1,q})\Big) \times L^{2r}(J; W^{-1,q}_{\Gamma_D}) \\ \to L^r(J; W^{-1,q}_\emptyset) \times (W^{1,q}, W^{-1,q}_\emptyset)_{\frac{1}{r},r},\end{aligned}$$

where

$$\mathcal{B}(\theta,u) = \left(\partial_t\theta + \mathcal{A}(\theta)\theta - \Psi_u(\theta) - \alpha\theta_l, \theta(T_0) - \theta_0\right),$$

is continuously differentiable in $(\bar\theta,\bar u)$, and that the partial derivative $\partial_\theta\mathcal{B}(\bar\theta,\bar u)$ is continuously invertible (see (2.6) for the definition of $\Psi$). Note that $\mathcal{B}(\bar\theta,\bar u) = 0$. The term $\alpha\theta_l$ does not depend neither on $u$ nor on $\theta$ and is thus neglected for the rest of this proof. Let us first consider the partial derivative with respect to $u$: For each $\theta \in C(\overline{Q})$, the mapping

$$L^{2r}(J;W^{-1,q}_{\Gamma_D})^2 \ni (u,v) \mapsto (\sigma(\theta)\varepsilon\nabla\varphi_u(\theta))\cdot\nabla\varphi_v(\theta) \in L^r(J;L^{q/2})$$

gives rise to a continuous symmetric bilinear form $b_\theta(u,v)$ (cf. also (2.5) and [33, Lem. 3.22]), since for fixed $\theta \in C(\overline{Q})$ we have

$$\begin{aligned}\|b_\theta(u,v)\|_{L^r(J;L^{q/2})} \le \|\sigma(\theta)\|_{C(\overline{Q})}\|\varepsilon\|_{L^\infty}\|\mathcal{J}(\sigma(\theta))\|^2_{C(\overline{J};\mathcal{L}(W^{-1,q}_{\Gamma_D},W^{1,q}_{\Gamma_D}))}&\\ \cdot\|u\|_{L^{2r}(J;W^{-1,q}_{\Gamma_D})}\|v\|_{L^{2r}(J;W^{-1,q}_{\Gamma_D})}&.\end{aligned}$$

Accordingly, $u \mapsto \Psi_u(\bar\theta) = b_{\bar\theta}(u,u)$ is continuously differentiable, and its derivative in $\bar u$ is given by $h \mapsto 2b_{\bar\theta}(\bar u,h)$. The second component of $\mathcal{B}$ is independent of $u$. Next, we treat the derivative of $\mathcal{B}$ w.r.t. $\theta$. First, note that, due to Assumption 2.9 (i), the Nemytskii operator $\theta \mapsto \eta(\theta)$ is continuously differentiable from $C(\overline{Q})$ to $C(\overline{Q})$ and its derivative in $\bar\theta$ is given by $h \mapsto \eta'(\bar\theta)h$. With Remark 2.7, we thus find that the derivative of the function $\theta \mapsto \partial_t\theta + \mathcal{A}(\theta)\theta$ as a mapping from $W^{1,r}(J;W^{-1,q}_\emptyset) \cap L^r(J;W^{1,q})$ to $L^r(J;W^{-1,q}_\emptyset)$ in the point $\bar\theta$ is given by

$$h \mapsto \partial_t h - \nabla\cdot\eta(\bar\theta)\kappa\nabla h + \tilde\alpha h - \nabla\cdot\eta'(\bar\theta)h\kappa\nabla\bar\theta = \partial_t h + \mathcal{A}(\bar\theta)h - \nabla\cdot\eta'(\bar\theta)h\kappa\nabla\bar\theta. \quad (3.3)$$

We turn to $\theta \mapsto \Psi_{\bar u}(\theta)$. As above, due to Assumption 2.9 (i), $\theta \mapsto \sigma(\theta)$ is continuously differentiable as a mapping from $C(\overline{Q})$ to $C(\overline{Q})$ and with derivative $h \mapsto \sigma'(\bar\theta)h$ (in a point $\bar\theta$). Further, recall that the derivative of the (continuously differentiable) mapping $\mathcal{L}(X;Y) \ni A \mapsto A^{-1} \in \mathcal{L}(Y;X)$ in $A$ is given by $H \mapsto -A^{-1}HA^{-1}$. The chain rule and Remark 2.7 thus yield continuous differentiability of $\theta \mapsto \mathcal{J}(\sigma(\theta))$ as a mapping from $C(\overline{J};C(\overline{\Omega}))$ to $C(\overline{J};\mathcal{L}(W^{1,q}_{\Gamma_D};W^{-1,q}_{\Gamma_D}))$ with the derivative

$$\left[(\mathcal{J}\circ\sigma)'(\bar\theta)\right]h = -\mathcal{J}(\sigma(\bar\theta))\left[-\nabla\cdot\sigma'(\bar\theta)h\varepsilon\nabla\right]\mathcal{J}(\sigma(\bar\theta)).$$

Hence, $\theta \mapsto \varphi_{\bar u}(\theta) = \mathcal{J}(\sigma(\theta))\bar u$ is also continuously differentiable, considered as a mapping from $C(\overline{J};C(\overline{\Omega}))$ to $L^{2r}(J;W^{1,q}_{\Gamma_D})$. Continuous differentiability of the function given by $C(\overline{J};C(\overline{\Omega})) \ni \theta \mapsto \Psi_{\bar u}(\theta) \in L^r(J;L^{q/2}) \hookrightarrow L^r(J;W^{-1,q}_\emptyset)$ is now straightforward and its derivative in $\bar\theta$ is given by

$$\begin{aligned}\left[\partial_\theta\Psi_{\bar u}(\bar\theta)\right]h = -2\left(\sigma(\bar\theta)\varepsilon\nabla\left[\mathcal{J}(\sigma(\bar\theta))\bar u\right]\right)\cdot\nabla\left[\left(\left[(\mathcal{J}\circ\sigma)'(\bar\theta)\right]h\right)\bar u\right]&\\ +\left(\sigma'(\bar\theta)h\varepsilon\nabla\left[\mathcal{J}(\sigma(\bar\theta))\bar u\right]\right)\cdot\nabla\left[\mathcal{J}(\sigma(\bar\theta))\bar u\right].&\quad (3.4)\end{aligned}$$

The second component of $\mathcal{B}$, i.e., $\theta \mapsto \theta(T_0) - \theta_0$, is affine-linear and continuous from the maximal regularity space into $(W^{1,q},W^{-1,q}_\emptyset)_{\frac{1}{r},r}$ and as such has the derivative $h \mapsto h(T_0)$. It remains to show the continuous invertibility of $\partial_\theta\mathcal{B}(\bar\theta,\bar u)$. For this, we identify for almost every $t \in (T_0,T_1)$ and $h \in C(\overline{J};C(\overline{\Omega}))$ as follows:

$$B(t)h(t) = \left(\left[\partial_\theta\Psi_{\bar u}(\bar\theta)\right]h\right)(t) + \nabla\cdot\eta'(\bar\theta(t))h(t)\kappa\nabla\bar\theta(t),$$

such that $B(t)$ is from $\mathcal{L}(C(\overline{\Omega}); W^{-1,q}_{\emptyset})$ and $t \mapsto B(t) \in L^r(J; \mathcal{L}(C(\overline{\Omega}); W^{-1,q}_{\emptyset}))$. Combining (3.3) and (3.4), in order to prove that $\mathcal{B}_\theta$ is continuously invertible we need to show that the equation

$$\partial_t \xi(t) + \mathcal{A}(\bar{\theta}(t))\xi(t) = B(t)\xi(t) + f(t), \qquad \xi(T_0) = \xi_0 \tag{3.5}$$

has a unique solution $\xi \in W^{1,r}(J; W^{-1,q}_{\emptyset}) \cap L^r(J; W^{1,q})$ for every $f \in L^r(J; W^{-1,q}_{\emptyset})$ and $\xi_0 \in (W^{1,q}, W^{-1,q}_{\emptyset})_{\frac{1}{r},r}$. This, however, is exactly what is obtained by [35, Cor. 3.4], hence we have

$$\partial_\theta \mathcal{B}(\bar{\theta}, \bar{u}) \in \mathcal{LH}\big(W^{1,r}(J; W^{-1,q}_{\emptyset}) \cap L^r(J; W^{1,q}); L^r(J; W^{-1,q}_{\emptyset})\big) \times (W^{1,q}, W^{-1,q}_{\emptyset})_{\frac{1}{r},r})).$$

Thus, all requirements for the implicit function theorem are satisfied, which yields neighbourhoods $\mathfrak{V}_{\bar{u}}$ of $\bar{u}$ in $L^{2r}(J; W^{-1,q}_{\Gamma_D})$ and $\mathfrak{V}_{\bar{\theta}}$ of $\bar{\theta}$ in the maximal regularity space, such that there exists a continuously differentiable mapping $\Phi\colon \mathfrak{V}_{\bar{u}} \to \mathfrak{V}_{\bar{\theta}}$ with $\mathcal{B}(\Phi(u), u) = \mathcal{B}(\bar{\theta}, \bar{u}) = 0$ for all $u \in \mathfrak{V}_{\bar{u}}$. This shows that the set of global controls is open. Moreover, $\Phi$ locally coincides with the control-to-state operator $u \mapsto \mathcal{S}(u)$, which implies continuous differentiability for the latter.

The stated expression for $\mathcal{S}'(u)h$ is obtained by differentiating the (constant) function $u \mapsto \mathcal{B}(\mathcal{S}(u), u)$. From the second component, we then find $(\mathcal{S}'(u)h)(T_0) = 0$ in $(W^{1,q}, W^{-1,q}_{\emptyset})_{\frac{1}{r},r}$ for all $h$, and the chain rule yields

$$\mathcal{S}'(u)h = -[\partial_\theta \mathcal{B}(\mathcal{S}(u), u)]^{-1} \partial_u \mathcal{B}(\mathcal{S}(u), u)h,$$

meaning exactly that $\mathcal{S}'(u)h$ is the unique solution to the problem (3.5) with right-hand side $f = -\partial_u \mathcal{B}(\mathcal{S}(u), u)h$ and initial value 0. Inserting all formulas, we obtain the equation stated in the theorem. ∎

Remark 3.2. *One may split the equation solved by $\zeta_h = \mathcal{S}'(u)h$ in the previous Theorem 3.1 back into two equations: Introducing*

$$\Phi(\zeta) := \mathcal{J}(\sigma(\theta_u))\,(-\nabla \cdot \sigma'(\theta_u)\zeta\varepsilon\nabla\varphi_u + h) \in L^{2r}(J; W^{1,q}_{\Gamma_D}),$$

*we find that, for every $h \in L^{2r}(J; W^{-1,q}_{\Gamma_D})$, the pair $(\zeta, \pi) := (\mathcal{S}'(u)h, \Phi(\mathcal{S}'(u)h))$ is the* unique solution *of the system*

$$\begin{aligned} \partial_t \zeta + \mathcal{A}(\theta_u)\zeta &= (\sigma'(\theta_u)\zeta\varepsilon\nabla\varphi_u) \cdot \nabla\varphi_u + \nabla \cdot \eta'(\theta_u)\zeta\kappa\nabla\theta_u + 2\,(\sigma(\theta_u)\varepsilon\nabla\varphi_u) \cdot \nabla\pi \\ -\nabla \cdot \sigma(\theta_u)\varepsilon\nabla\pi &= -\nabla \cdot \sigma'(\theta_u)\zeta\varepsilon\nabla\varphi_u + h \end{aligned}$$

*with $\zeta(T_0) = 0$ (the first equation is supposed to hold in $W^{-1,q}_{\emptyset}$, the second one in $W^{-1,q}_{\Gamma_D}$, each for almost all $t \in J$). These equations are exactly the* linearized state system *for (2.3) and (2.4). This also shows, expectedly, that from a functional-analytical point of view, it makes no difference working with $\theta$ only and considering $\varphi$ as a function obtained by $\theta$, instead of considering both functions at once.*

Combining Theorem 3.1 with the fact that $u \equiv 0$ is a global control, we obtain the following

Corollary 3.3. *There is always a neighbourhood $\mathfrak{V}_0$ of 0 in $L^{2r}(J; W^{-1,q}_{\Gamma_D})$, containing only global controls, i.e., $\mathfrak{V}_0 \subseteq \mathcal{U}_g$.*

**4. Necessary optimality conditions.** This section is devoted to the derivation of necessary optimality conditions for (P) in reduced form which we will introduce first, cf. ($\mathrm{P}_u$) below. We will require some preliminary definitions for the control problem. Let us begin by introducing the admissible set

$$\mathcal{U}^{\mathrm{ad}} := \{u \in L^2(J; L^2(\Gamma_N)) \colon 0 \leq u \leq u_{\max} \text{ a.e. in } \Sigma_N\}. \tag{4.1}$$

We call a global control $u \in \mathcal{U}_g$ *feasible*, if $u \in \mathcal{U}^{\mathrm{ad}}$ and the associated state satisfies $\mathcal{S}(u)(x,t) \leq \theta_{\max}(x,t)$ for all $(x,t) \in \overline{Q}$. Due to Assumption 2.9 (xi), the control $u \equiv 0$ is a feasible one, cf. [33, Cor. 4.11].

Next, we define the actual control space for the optimal control problem, fitting the norm in the objective functional in (P), that is,

$$\mathbb{U} := W^{1,2}(J; L^2(\Gamma_N)) \cap L^p(J; L^p(\Gamma_N))$$

with the standard norm $\|u\|_{\mathbb{U}} = \|u\|_{W^{1,2}(J;L^2(\Gamma_N))} + \|u\|_{L^p(J;L^p(\Gamma_N))}$. Since $p > \frac{4}{3}q - 2$ by Assumption 2.9 (viii), this space continuously embeds into $L^s(J; W^{-1,q}_{\Gamma_D})$ for every $s \in [1,\infty]$. The precise result is as follows ([33, Prop. 4.12]):

Proposition 4.1. *Let $p > 2$. The space $\mathbb{U}$ is embedded into a Hölder space $C^{\varrho}(\overline{J}; L^{\mathfrak{p}}(\Gamma_N))$ for some $\varrho > 0$ and $2 < \mathfrak{p} < \frac{p+2}{2}$. In particular, there exists a* compact *embedding $\mathcal{E} \colon \mathbb{U} \hookrightarrow L^s(J; W^{-1,q}_{\Gamma_D})$ for every $p > \frac{4}{3}q - 2$ and $s \in [1,\infty]$.*

Note that the embedding $L^{\mathfrak{p}}(\Gamma_N) \hookrightarrow W^{-1,q}_{\Gamma_D}$ is given via $\tau^*_{\Gamma_N}$, i.e., the adjoint of the trace operator $\tau_{\Gamma_N} \colon W^{1,q'}_{\Gamma_D} \hookrightarrow L^{\mathfrak{p}'}(\Gamma_N)$, cf. Remark 2.10.

Remark 4.2. *The compact embedding for $\mathbb{U}$ into the spaces $L^s(J; W^{-1,q}_{\Gamma_D})$ plays a crucial part in establishing existence of optimal solutions for* (P)*, see [33, Thm. 4.14], which is the main reason to consider this special space as the control space instead of, say, $L^p(J; L^p(\Gamma_N))$ with $p$ sufficiently large. The other ingredients for the existence result are the second term in the objective functional which includes changing the requirements for the time integrability $s$ to $s > \frac{2q}{q-3}(1 - \frac{3}{q} + \frac{3}{\varsigma}) > \frac{2q}{q-3}$, cf. Assumption* 2.9 (viii)*. For the results in this paper, these considerations are* not *needed, i.e., one could also drop the term involving $\nabla\theta$ in the objective functional and work with $L^p(J; L^p(\Gamma_N))$ with $p > \max(\frac{4q}{q-3}, \frac{4}{3}q - 2)$ as a control space, thereby also dropping the $\partial_t u$-norm in the objective functional, and obtain the results presented below. However, in order to give a comprehensive treatment, together with [33], of the optimal control problem for the thermistor problem, we will keep the problem* (P) *for the derivation of necessary optimality conditions of first order as it stands. Modifications of the following considerations to remove the additional terms as listed above are possible and straight-forward.*

In view of the foregoing Remark 4.2, we from now on fix $s > \frac{2q}{q-3}(1 - \frac{3}{q} + \frac{3}{\varsigma})$ as in Assumption 2.9 (viii).

Definition 4.3. *Consider the embedding $\mathcal{E}$ from Proposition 4.1 with range in $L^{2s}(J; W^{-1,q}_{\Gamma_D})$, where $s > \frac{2q}{q-3}(1 - \frac{3}{q} + \frac{3}{\varsigma})$ is the integrability exponent from the objective functional. We set*

$$\mathbb{U}_g := \{u \in \mathbb{U} \colon \mathcal{E}(u) \in \mathcal{U}_g\}$$

*and define the mapping*

$$\mathcal{S}_{\mathcal{E}} := \mathcal{S} \circ \mathcal{E} \colon \mathbb{U}_g \to W^{1,s}(J; W^{1,q}) \cap L^s(J; W^{-1,q}_{\emptyset}).$$

*Moreover, we define the* reduced objective functional *$j$ obtained by reducing the objective functional in* (P) *to $u$, i.e.,*

$$j(u) = \frac{1}{2}\int_E |\mathcal{S}_{\mathcal{E}}(u)(T_1) - \theta_d|^2 \,\mathrm{d}x + \frac{\gamma}{s}\|\nabla \mathcal{S}_{\mathcal{E}}(u)\|^s_{L^s(J;L^q)} + \frac{\beta}{2}\int_{\Sigma_N} (\partial_t u)^2 + |u|^p \,\mathrm{d}\omega\,\mathrm{d}t,$$

*as a function on* $\mathbb{U}_g$. *Further, let* $\mathbb{U}^{ad} := \mathbb{U} \cap \mathcal{U}^{ad}$ *and* $\mathbb{U}_g^{ad} := \mathbb{U}_g \cap \mathcal{U}^{ad}$*, where* $\mathcal{U}^{ad}$ *is as defined in* (4.1). *Finally, the reduced optimal control problem is now given by*

$$\min_{u \in \mathbb{U}_g^{ad}} j(u) \quad \text{such that} \quad \mathcal{S}_\mathcal{E}(u)(x,t) \leq \theta_{\max}(x,t) \quad \forall (x,t) \in \overline{Q}. \tag{P$_u$}$$

One readily observes that $\mathcal{S}_\mathcal{E}$ on $\mathbb{U}_g$ is still continuously differentiable with the derivative $h \mapsto \mathcal{S}_\mathcal{E}'(u)h = \mathcal{S}'(\mathcal{E}u)\mathcal{E}h$. Let us now turn to first order necessary conditions for (P$_u$) and start with the definition of the Lagrangian function. It is well-known that the Lagrangian multipliers associated to the state constraints may, in general, only be regular Borel measures, see for instance [8]. Hence, we introduce the space $\mathcal{M}(\overline{Q})$ as the space of regular Borel measures on $\overline{Q}$ and, simultaneously, as the dual space of $C(\overline{Q})$.

Definition 4.4. *The* Lagrangian function $\mathfrak{L}\colon \mathbb{U}_g \times \mathcal{M}(\overline{Q}) \to \mathbb{R}$ *associated with* (P$_u$) *is given by*

$$\mathfrak{L}(u,\mu) = j(u) + \langle \mu, \mathcal{S}_\mathcal{E}(u) - \theta_{\max}\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})},$$

*where* $j$ *is the reduced objective functional.*

Definition 4.5. *We denote by* $\Delta_q\colon W^{1,q} \to W_\emptyset^{-1,q'}$ *the* (weak) $q$-Laplacian*, given by*

$$\langle \Delta_q \psi, \xi\rangle := \int_\Omega |\nabla \psi|^{q-2} \nabla\psi \cdot \nabla \xi \, \mathrm{d}x$$

*for each* $\psi, \xi \in W^{1,q}$.

The chain rule immediately yields the derivative of $\mathfrak{L}$ with respect to $u$:

Lemma 4.6. *The Lagrangian function* $\mathfrak{L}$ *is continuously differentiable with respect to* $u$. *Abbreviating the states by* $\theta_u := \mathcal{S}_\mathcal{E}(u)$ *and* $\theta_u' = \mathcal{S}_\mathcal{E}'(u)h$*, the partial derivative in direction* $h \in \mathbb{U}$ *is given by*

$$\begin{aligned}\partial_u \mathfrak{L}(u,\mu)h = \int_E (\theta_u(T_1) - \theta_d)\theta_u'(T_1)\,\mathrm{d}x + \gamma \int_{T_0}^{T_1} \|\nabla \theta_u(t)\|_{L^q}^{s-q} \langle \Delta_q \theta_u(t), \theta_u'(t)\rangle\,\mathrm{d}t \\ + \beta \int_{\Sigma_N} \partial_t u \partial_t h + \frac{p}{2}|u|^{p-2} uh \,\mathrm{d}\omega\,\mathrm{d}t) + \langle \mu, \theta_u'\rangle_{\mathcal{M}(\overline{Q}),C(\overline{Q})}\end{aligned} \tag{4.2}$$

*with* $\Delta_q$ *given as in Definition 4.5.*

Using the Lagrangian function and its derivative, we characterize local optima of (P$_u$). We say that that a feasible control $\bar{u}$ is *locally optimal* if there exists an $\epsilon > 0$ such that $j(\bar{u}) \leq j(u)$ for all feasible $u \in \mathbb{U}_g^{\mathrm{ad}}$ with $\|u - \bar{u}\|_\mathbb{U} < \epsilon$. As we will see in the proof of Theorem 4.8, the restriction to global controls $u \in \mathbb{U}_g$ does not influence the derivation of optimality conditions, since $\mathbb{U}_g$ is an *open* set by Theorem 3.1. This allows to derive an optimality theory without refering to the generally unknown set $\mathcal{U}_g$.

Definition 4.7. *A measure* $\bar{\mu} \in \mathcal{M}(\overline{Q})$ *is called a* Lagrangian multiplier *associated with the state constraint in* (P$_u$)*, if for a locally optimal control* $\bar{u}$ *the* KKT conditions

$$\bar{\mu} \geq 0, \tag{4.3}$$

$$\langle \bar{\mu}, \mathcal{S}_\mathcal{E}(\bar{u}) - \theta_{\max}\rangle_{C(\overline{Q})} = 0, \tag{4.4}$$

$$\langle \partial_u \mathfrak{L}(\bar{u},\bar{\mu}), u - \bar{u}\rangle_\mathbb{U} \geq 0 \quad \forall u \in \mathbb{U}^{ad} \tag{4.5}$$

*hold true. Here, (4.3) means that $\langle \bar\mu, f\rangle_{C(\overline{Q})} \geq 0$ for all $f \in C(\overline{Q})$ with $f(x,t) \geq 0$ for all $(x,t) \in Q$. Note that (4.5) has to be satisfied for* all *$u \in \mathbb{U}^{ad}$ instead of only in $\mathbb{U}^{ad}_g$, the latter being defined in Definition 4.3.*

It is well-known that, in general, a so-called regularity condition is needed in order to ensure the existence of a Lagrangian multiplier. In this case, we rely on the linearized Slater condition, which is a special form of Robinson's regularity condition.

Theorem 4.8. *Let $\bar u$ be a locally optimal control and let the following so-called* linearized Slater condition *be satisfied: There exists $\hat u \in \mathbb{U}^{ad}_g$ such that there is a $\delta > 0$ with the property*

$$\mathcal{S}_{\mathcal{E}}(\bar u)(x,t) + \mathcal{S}_{\mathcal{E}}'(\bar u)(\hat u - \bar u)(x,t) \leq \theta_{\max}(x,t) - \delta \quad \text{for all } (x,t) \in Q. \tag{4.6}$$

*Then there exists a Lagrangian multiplier $\bar\mu \in \mathcal{M}(\overline{Q})$ associated with the state constraint in ($\mathrm{P}_u$), i.e., such that (4.3)-(4.5) is satisfied.*

*Proof.* Since $\mathbb{U} \hookrightarrow L^{2s}(J; W^{-1,q}_{\Gamma_D})$ as seen in Proposition 4.1, Theorem 3.1 implies that there is an open ball $B_\delta(\bar u) \subset \mathbb{U}$ around $\bar u$ with radius $\delta > 0$ such that $B_\delta(\bar u) \cap \mathbb{U}^{\mathrm{ad}} \subset \mathbb{U}^{\mathrm{ad}}_g$. We consider the auxiliary problem

$$\left.\begin{aligned} &\min \quad j(u) \\ &\text{s.t.} \quad u \in B_\delta(\bar u) \cap \mathbb{U}^{\mathrm{ad}}, \quad \mathcal{S}_{\mathcal{E}}(u)(x,t) \leq \theta_{\max}(x,t) \quad \forall (x,t) \in \overline{Q}. \end{aligned}\right\} \tag{$\mathrm{P_{aux}}$}$$

Clearly, $\bar u$ is also a local minimizer of this problem. Moreover, in contrast to $\mathbb{U}^{\mathrm{ad}}_g$ appearing in ($\mathrm{P}_u$), the feasible set $B_\delta(\bar u) \cap \mathbb{U}^{\mathrm{ad}}$ is now *convex*. Therefore the standard Karush-Kuhn-Tucker (KKT) theory in function space can be applied to ($\mathrm{P_{aux}}$), see e.g. [37, Thm. 3.1], [8, Thm. 5.2] or [6, Thm. 3.9]. Hence, on account of the linearized Slater condition in (4.6), there exists a Lagrange multiplier $\bar\mu \in \mathcal{M}(\overline{Q})$ so that (4.3), (4.4), and

$$\langle \partial_u \mathfrak{L}(\bar u, \bar\mu), v - \bar u\rangle_{\mathbb{U}} \geq 0 \quad \forall v \in B_\delta(\bar u) \cap \mathbb{U}^{\mathrm{ad}} \tag{4.7}$$

are fulfilled. Now, let $u \in \mathbb{U}^{ad}$ be arbitrary. Then, due to convexity, $\bar u + \tau(u - \bar u) \in B_\delta(\bar u) \cap \mathbb{U}^{\mathrm{ad}}$ for $\tau > 0$ sufficiently small such that this function can be chosen as test function in (4.7), giving in turn (4.5). □

Let us now transform (4.3)-(4.5) into an optimality system involving an adjoint state. To this end, we aim to reformulate the derivative expression for $\partial_u \mathfrak{L}(\bar u, \mu)$ from Lemma 4.6 in a designated locally optimal point $\bar u$. We stress again the crucial point that we do not have to work with $\mathcal{U}_g$ or associated sets in the optimality conditions (4.3)-(4.5) by virtue of Theorem 3.1. For brevity, we define

$$\mathbb{X} = W^{1,s}(J; W^{-1,q}_{\emptyset}) \cap L^s(J; W^{1,q}) \quad \text{and} \quad X_s = (W^{1,q}, W^{-1,q}_{\emptyset})_{\frac{1}{s},s}.$$

The plan is to use the adjoint of the derivative of the control-to-state operator. We will show that $\mathcal{S}_{\mathcal{E}}'(\bar u)^*$ is associated to the solution operator (in an appropriate sense) to the *adjoint system*, which we formally introduce as follows:

Definition 4.9. *For given, fixed functions $\theta$ and $\varphi$, given terminal value $\vartheta_T$ and* *inhomogeneities $f_1, f_2, g_1, g_2$, we call the following system the* adjoint system*:*

$$\left.\begin{aligned}
-\partial_t\vartheta - \operatorname{div}(\eta(\theta)\kappa\nabla\vartheta) &= (\sigma'(\theta)\vartheta\varepsilon\nabla\varphi)\cdot\nabla\varphi - (\sigma'(\theta)\varepsilon\nabla\varphi)\cdot\nabla\psi \\
&\qquad - (\eta'(\theta)\kappa\nabla\vartheta)\cdot\nabla\theta + f_1 && \text{in } Q,\\
\nu\cdot\eta(\theta)\kappa\nabla\vartheta + \alpha\vartheta &= f_2 && \text{on } \Sigma,\\
\vartheta(T_1) &= \vartheta_T && \text{in } \Omega,\\[1ex]
-\operatorname{div}(\sigma(\theta)\varepsilon\nabla\psi) &= -2\operatorname{div}(\sigma(\theta)\vartheta\varepsilon\nabla\varphi) + g_1 && \text{in } Q,\\
\nu\cdot\nabla\sigma(\theta)\varepsilon\nabla\psi &= 2\nu\cdot\sigma(\theta)\vartheta\varepsilon\nabla\varphi + g_2 && \text{on } \Sigma_N,\\
\psi &= 0 && \text{on } \Sigma_D.
\end{aligned}\right\}\tag{4.8}$$

More specified assumptions about the inhomogeneities $f_1, f_2, g_1, g_2$ and the terminal value $\vartheta_T$ will be given in the following. Note that (4.8) is only a *formal* representation of the adjoint of the linearized system of (1.1)-(1.6). We will work with the abstract version, referring to (2.3) and (2.4) and its linearizations, cf. (3.2) or Remark 3.2.

Definition 4.10. *Let $\theta \in \mathbb{X}$ be fixed and set $\varphi = \mathcal{J}(\sigma(\theta))u$. Further, let $f \in \mathbb{X}'$, $\vartheta_T \in X_r'$, and $g \in L^{(2s)'}(J; W^{-1,q'}_{\Gamma_D})$ be given with $(2s)' = \frac{2s}{2s-1}$. The* abstract adjoint system *is given by*

$$\begin{aligned}
-\partial_t\vartheta + \partial_\theta\mathcal{A}(\theta)\vartheta = -(\eta'(\theta)\kappa\nabla\vartheta)\cdot\nabla\theta + (\sigma'(\theta)\vartheta\varepsilon\nabla\varphi)\cdot\nabla\varphi + \delta_{T_1}\otimes\vartheta_T - \delta_{T_0}\otimes\chi + f\\
- (\sigma'(\theta)\varepsilon\nabla\varphi)\cdot\nabla\left[\mathcal{J}(\sigma(\theta))^*(-2\nabla\cdot\sigma(\theta)\vartheta\varepsilon\nabla\varphi + g)\right].
\end{aligned}\tag{4.9}$$

*Here, $\delta_{T_0}$ and $\delta_{T_1}$ are Dirac measures in $T_0$ and $T_1$, obtained as the adjoints of the point evaluations in $T_0$ and $T_1$, respectively. The latter are continuous mappings from $C(\overline{J}; X_s)$ to $X_s$, such that $\delta_{T_0}\otimes\vartheta_T$ and $\delta_{T_1}\otimes\chi$ are seen as objects from $\mathcal{M}(J; X_s')$. We say that the functions $(\vartheta,\chi) \in L^{s'}(J; W^{1,q'}) \times X_s'$ are a* weak solution *of (4.9) or (4.8), if*

$$\begin{aligned}
\int_J \langle\partial_t\xi, \vartheta\rangle_{W^{1,q'}}\,\mathrm{d}t = &-\int_J\int_\Omega \langle(\eta(\theta)\kappa\nabla\vartheta)\nabla\xi\,\mathrm{d}x\mathrm{d}t - \int_J\int_\Gamma \alpha\vartheta\xi\,\mathrm{d}\omega\mathrm{d}t\\
&- \int_J\int_\Omega \left[(\eta'(\theta)\kappa\nabla\vartheta)\cdot\nabla\theta - (\sigma'(\theta)\vartheta\varepsilon\nabla\varphi)\cdot\nabla\varphi\right]\xi\,\mathrm{d}x\mathrm{d}t\\
&- \int_J\int_\Omega (\sigma'(\theta)\varepsilon\nabla\varphi)\cdot\nabla\left[\mathcal{J}(\sigma(\theta))^*(-2\nabla\cdot\sigma(\theta)\vartheta\varepsilon\nabla\varphi + g)\right]\xi\,\mathrm{d}x\mathrm{d}t\\
&+ \langle\vartheta_T, \xi(T_1)\rangle_{X_r} - \langle\chi, \xi(T_0)\rangle_{X_r} + \langle f, \xi\rangle_{\mathbb{X}}
\end{aligned}\tag{4.10}$$

*is true for all $\xi \in \mathbb{X}$. Equivalently, (4.9) holds true in $\mathbb{X}'$.*

Note that the functionals $\delta_{T_0} \times \chi$ and $\delta_{T_1} \otimes \vartheta_T$ are well-defined in $\mathbb{X}'$ due to $\mathbb{X} \hookrightarrow C(\overline{J}; X_s)$. Of course, the inhomogeneities $f_1, f_2$ and $g_1, g_2$ from (4.8) are represented by $f = f_1 + f_2$ and $g = g_1 + g_2$, respectively. Moreover, thanks to the symmetry of $\varepsilon$, one easily sees that $\mathcal{J}(\sigma(\theta))^*$ is formally selfadjoint, which is the basis of the following

Remark 4.11. *Similarly to Remark 3.2, we introduce*

$$\psi(\vartheta) := \mathcal{J}(\sigma(\theta))^*(-2\nabla\cdot\sigma(\theta)\vartheta\varepsilon\nabla\varphi + g),$$

*which allows to split* (4.9) *back into two equations, namely*

$$\begin{aligned}-\partial_t\vartheta + \partial_\theta\mathcal{A}(\theta)\vartheta &= (\sigma'(\theta)\vartheta\varepsilon\nabla\varphi)\cdot\nabla\varphi - (\eta'(\theta)\kappa\nabla\vartheta)\cdot\nabla\theta - (\sigma'(\theta)\varepsilon\nabla\varphi)\cdot\nabla\psi\\ &\qquad + \delta_{T_1}\otimes\vartheta_T - \delta_{T_0}\otimes\chi + f,\\ -\nabla\cdot\sigma(\theta)\varepsilon\nabla\psi &= -2\nabla\cdot\sigma(\theta)\vartheta\varepsilon\nabla\varphi + g,\end{aligned}$$

*to be understood as in* (4.10)*. This is exactly a very weak abstract formulation of the formal adjoint system* (4.8) *with inhomogeneities* $f = f_1 + f_2$ *and* $g = g_1 + g_2$ *and terminal value* $\vartheta_T$*. Note that the first equation is supposed to hold in* $\mathbb{X}'$*, the second one in* $L^{(2s)'}(J; W^{-1,q'}_{\Gamma_D})$*.*

We next show that the abstract adjoint (4.9) always admits a unique weak solution for $f \in \mathbb{X}'$ and $g \in L^{(2s)'}(J; W^{-1,q'}_{\Gamma_D})$. This will follow directly from Theorem 3.1 using an adjoint-approach (see e.g. [1, Ch. 7]). Since the inhomogeneity $f$ in (4.9) will in fact contain the Lagrange multiplier $\mu$ introduced in Definition 4.7, we will not investigate the adjoint system more specifically under additional regularity assumptions on $f$, since the Lagrange multipliers are in general only measures and thus limit said regularity in a crucial way anyhow. In particular, this lack of regularity is the very obstacle which permits time-derivatives for weak solutions to (4.9), cf. [1, Prop. 6.1]. Nevertheless, even in the absence of measure-valued Lagrange multipliers, the time regularity of the adjoint state is still limited by the differential operator itself, since $(\eta'(\theta)\kappa\nabla\vartheta)\cdot\nabla\theta$ is only integrable in time (as opposed to $s'$-integrable).

THEOREM 4.12. *For every terminal value* $\vartheta_T \in X_s' = (W^{1,q'}, W^{-1,q'}_\emptyset)_{\frac{1}{s'},s'}$ *and all imhomogeneities* $f \in \mathbb{X}'$ *and* $g \in L^{(2s)'}(J; W^{-1,q'}_{\Gamma_D})$*, there exists a unique weak solution* $(\vartheta, \chi) \in L^{s'}(J; W^{1,q'}) \times X_s'$ *of* (4.9) *in the sense of Definition 4.10.*

*Proof.* The equality $X_s' = (W^{1,q'}, W^{-1,q'}_\emptyset)_{\frac{1}{s'},s'}$ follows from the usual duality properties of interpolation functors, see [36, Ch. 1.11.2 and 1.3.3]. Recall the operator

$$\mathcal{B}\colon \mathbb{X}\times L^{2s}(J; W^{-1,q}_{\Gamma_D}) \to L^s(J; W^{-1,q}_\emptyset)\times (W^{1,q}, W^{-1,q}_\emptyset)_{\frac{1}{s},s},$$

from Theorem 3.1 with $r = s > \bar{r}(q,\varsigma) \geq r^*(q)$. The partial derivative w.r.t. $\theta$ of $\mathcal{B}$ was given by

$$\partial_\theta\mathcal{B}(\theta,u)\xi = (\partial_t\xi + \mathcal{A}(\theta)\xi - \nabla\cdot\eta'(\theta)\xi\kappa\nabla\theta - \partial_\theta\Psi_u(\theta)\xi, \xi(T_0))$$

with

$$\partial_\theta\Psi_u(\theta)\xi = -2(\sigma(\theta)\varepsilon\nabla\varphi)\cdot\nabla\left[\mathcal{J}(\sigma(\theta)(-\nabla\cdot\sigma'(\theta)\xi\varepsilon\nabla\varphi)\right] + (\sigma'(\theta)\xi\varepsilon\nabla\varphi)\cdot\nabla\varphi,$$

cf. (3.4), and $\varphi = \mathcal{J}(\sigma(\theta))u$. Now, let $(\vartheta,\chi)$ be from $L^{s'}(J; W^{1,q'}) \times X_s'$. We easily find

$$\langle -\nabla\cdot\eta'(\theta)\xi\kappa\nabla\theta, \vartheta\rangle_{W^{1,q'}} = \int_\Omega (\eta'(\theta)\xi\kappa\nabla\vartheta)\cdot\nabla\theta\,\mathrm{d}x = \langle(\eta'(\theta)\kappa\nabla\vartheta)\cdot\nabla\theta, \xi\rangle_{W^{1,q}} \tag{4.11}$$

and

$$\langle(\sigma'(\theta)\xi\varepsilon\nabla\varphi)\cdot\nabla\varphi, \vartheta\rangle_{W^{1,q'}} = \langle(\sigma'(\theta)\vartheta\varepsilon\nabla\varphi)\cdot\nabla\varphi, \xi\rangle_{W^{1,q}}\,. \tag{4.12}$$

Let us turn to the complicated term in $\partial_\theta \Psi_u(\theta)$. Analogously to (4.11), we find

$$\begin{aligned}\langle 2(\sigma(\theta)\varepsilon\nabla\varphi)\cdot\nabla\left[\mathcal{J}(\sigma(\theta)(-\nabla\cdot\sigma'(\theta)\xi\varepsilon\nabla\varphi)\right],\vartheta\rangle_{W^{1,q'}} &\\ = \langle \mathcal{J}(\sigma(\theta))(-\nabla\cdot\sigma'(\theta)\xi\varepsilon\nabla\varphi), -2\nabla\cdot\sigma(\theta)\vartheta\varepsilon\nabla\varphi\rangle_{W^{-1,q'}_{\Gamma_D}} &\\ = \langle -\nabla\cdot\sigma'(\theta)\xi\varepsilon\nabla\varphi, \mathcal{J}(\sigma(\theta))^*\left(-2\nabla\cdot\sigma(\theta)\vartheta\varepsilon\nabla\varphi\right)\rangle_{W^{1,q'}_{\Gamma_D}} &\\ = \langle \xi, (\sigma'(\theta)\varepsilon\nabla\varphi)\,\nabla\left[\mathcal{J}(\sigma(\theta))^*\left(-2\nabla\cdot\sigma(\theta)\vartheta\varepsilon\nabla\varphi\right)\right]\rangle_{W^{-1,q'}_{\emptyset}}. &\quad (4.13)\end{aligned}$$

Symmetry of $\kappa$ implies that $\mathcal{A}(\theta)$ is formally self-adjoint, i.e., $\mathcal{A}(\theta)^*$ maps $W^{1,q'}$ into $W^{-1,q'}_\emptyset$, but is still given as in (2.2) and Definition 2.6, respectively. Using this and equations (4.11), (4.12) and (4.13), we obtain

$$\begin{aligned}\langle \partial_\theta \mathcal{B}(\theta,u)^*(\vartheta,\chi),\xi\rangle_{\mathbb{X}} &= \langle(\vartheta,\chi),\partial_\theta\mathcal{B}(\theta,u)\xi\rangle_{L^s(J;W^{-1,q}_\emptyset)\times X_s}\\ &= \int_J \langle\partial_t\xi,\vartheta\rangle_{W^{1,q'}}\,\mathrm{d}t + \int_J\langle\mathcal{A}^*(\theta)\vartheta,\xi\rangle_{W^{1,q}}\,\mathrm{d}t\\ &\quad + \int_J \langle(\eta'(\theta)\kappa\nabla\vartheta)\cdot\nabla\theta,\xi\rangle_{W^{1,q}}\,\mathrm{d}t\\ &\quad - \int_J\langle(\sigma'(\theta)\vartheta\varepsilon\nabla\varphi)\cdot\nabla\varphi,\xi\rangle_{W^{1,q}}\,\mathrm{d}t + \langle\chi,\xi(T_0)\rangle_{X_s}\\ &\quad + \int_J\langle(\sigma'(\theta)\varepsilon\nabla\varphi)\,\nabla\left[\mathcal{J}(\sigma(\theta))^*(-2\nabla\cdot\sigma(\theta)\vartheta\varepsilon\nabla\varphi)\right],\xi\rangle_{W^{1,q}}\,\mathrm{d}t\end{aligned}$$

for all $\xi\in\mathbb{X}$. Moreover, in the proof of Theorem 3.1, $\partial_\theta\mathcal{B}(\theta,u)$ was found to be a topological isomorphism between the spaces $\mathbb{X}$ and $L^s(J;W^{-1,q}_\emptyset)\times X_s$ and consequently $\partial_\theta\mathcal{B}(\theta,u)^*$ is also a topological isomorphism between $L^{s'}(J;W^{1,q'})\times X_s'$ and $\mathbb{X}'$. In particular, for every $\mathfrak{f}\in\mathbb{X}'$ there exists a unique $p = p_{\mathfrak{f}}\in L^{s'}(J;W^{1,q'})\times X_s'$ such that $\partial_\theta\mathcal{B}(\theta,u)^*p=\mathfrak{f}$. Hence, setting

$$\bar{\mathfrak{f}} = f + \delta_{T_1}\otimes\vartheta_T - (\sigma'(\theta)\varepsilon\nabla\varphi)\,\nabla\left[\mathcal{J}(\sigma(\theta))^*g\right], \qquad (4.14)$$

the pair $(\bar\vartheta,\bar\chi) := p_{\bar{\mathfrak{f}}}$ satisfies (4.10) by the above form of $\partial_\theta\mathcal{B}(\theta,u)^*$, and is exactly the searched-for unique solution as in Definition 4.10. ◻

As hinted above, we immediately obtain the following characterization of $\mathcal{S}'(u)^*$ for given $u\in\mathbb{U}_g$:

Corollary 4.13. *Let $(\vartheta,\chi)$ be the solution of (4.10) in the sense of Definition 4.10 with inhomogeneites $f$ and $g$ and terminal value $\vartheta_T$. The adjoint linearized solution operator $\mathcal{S}'_{\mathcal{E}}(u)^*$ then assigns to $f,g$ and $\vartheta_T$ in the form $\mathfrak{f}\in\mathbb{X}'$ as in (4.14) the functional $\mathcal{E}^*\psi\in\mathbb{U}'$, where $\psi(\vartheta)\in L^{(2s)'}(J;W^{1,q'}_{\Gamma_D})$ is given by*

$$\psi(\vartheta) = \mathcal{J}(\sigma(\theta_u))^*(-\nabla\cdot\sigma(\theta_u)\vartheta\varepsilon\nabla\varphi_u),$$

*similarly to Remark 4.11.*

*Proof.* In Theorem 3.1, we found $\mathcal{S}'(u) = -[\partial_\theta\mathcal{B}(\mathcal{S}(u),u)]^{-1}\partial_u\mathcal{B}(\mathcal{S}'(u),u)$. Hence, with $\mathcal{S}'_{\mathcal{E}}(u) = \mathcal{S}'(u)\circ\mathcal{E}$, we obtain

$$\mathcal{S}'_{\mathcal{E}}(u)^*\mathfrak{f} = -\mathcal{E}^*\partial_u\mathcal{B}(\mathcal{S}_{\mathcal{E}}(u),u)^*\partial_\theta\mathcal{B}(\mathcal{S}_{\mathcal{E}}(u),u)^{-*}\mathfrak{f}.$$

In view of Theorem 4.12 and its proof, $\partial_\theta\mathcal{B}(\mathcal{S}_{\mathcal{E}}(u),u)^{-*}\mathfrak{f}$ is exactly the unique solution $(\vartheta,\chi)$ of (4.10) in the sense of Definition 4.10 with inhomogeneites $f,g$ and terminal value $\vartheta_T$. Moreover, a repetition of the first lines of (4.13) shows that

$$-\partial_u\mathcal{B}(\mathcal{S}_{\mathcal{E}}(u),u)^*(\vartheta,\chi) = \mathcal{J}(\sigma(\theta_u))^*(-\nabla\cdot\sigma(\theta_u)\vartheta\varepsilon\nabla\varphi_u) = \psi(\vartheta).$$

An application of $\mathcal{E}^*: L^{(2s)'}(J; W^{1,q'}_{\Gamma_D}) \hookrightarrow \mathbb{U}'$ yields the claim. □

Having $\mathcal{S}'_{\mathcal{E}}(u)^*$ at hand, we now proceed to establish the actual necessary optimality conditions by manipulating the variational inequality in the KKT conditions (4.5).

For a concise "strong" formulation in the following theorem, we decompose measures $\mu \in \mathcal{M}(\overline{Q})$ by restriction into $\mu = \mu_{(T_0,T_1)} + \mu_{\{T_0\}\times\{T_1\}}$, with $\mu_{(T_0,T_1)} \in \mathcal{M}((T_0,T_1)\times\overline{\Omega})$ and $\mu_{\{T_0\}\times\{T_1\}} \in \mathcal{M}((\{T_0\}\times\{T_1\})\times\overline{\Omega})$. Both measures may in turn be further decomposed into $\mu_{(T_0,T_1)} = \mu_\Omega + \mu_\Gamma$, where $\mu_\Omega \in \mathcal{M}((T_0,T_1)\times\Omega)$ and $\mu_\Gamma \in \mathcal{M}((T_0,T_1)\times\Gamma)$, and $\mu_{\{T_0\}\times\{T_1\}} = \delta_{T_0}\otimes\mu_{T_0} + \delta_{T_1}\otimes\mu_T$ with $\mu_{T_0}, \mu_T \in \mathcal{M}(\overline{\Omega})$.

THEOREM 4.14 (First Order Necessary Conditions). *Let $\bar{u} \in \mathbb{U}^{ad}_g$ be a locally optimal control such that the linearized Slater condition (4.6) is satisfied. Let $\theta_{\bar{u}} = \mathcal{S}_{\mathcal{E}}(\bar{u})$ be the state associated with $\bar{u}$ and set $\varphi_{\bar{u}} := \varphi_{\bar{u}}(\theta_{\bar{u}})$. Then there exists a Lagrangian multiplier $\bar{\mu} \in \mathcal{M}(\overline{Q})$ in the sense of Definition 4.7 and adjoint states $\vartheta \in L^{s'}(J; W^{1,q'})$ and $\psi \in L^{(2s)'}(J; W^{1,q'}_{\Gamma_D})$, such that the formal system*

$$\begin{aligned} -\partial_t\vartheta - \operatorname{div}(\eta(\theta_{\bar{u}})\kappa\nabla\vartheta) &= (\sigma'(\theta_{\bar{u}})\vartheta\varepsilon\nabla\varphi_{\bar{u}})\cdot\nabla\varphi_{\bar{u}} - (\sigma'(\theta_{\bar{u}})\varepsilon\nabla\varphi_{\bar{u}})\cdot\nabla\psi \\ &\quad - (\eta'(\theta)\kappa\nabla\vartheta)\cdot\nabla\theta + \|\nabla\theta_{\bar{u}}\|^{s-q}_{L^s(J;L^q)}\Delta_q\theta_{\bar{u}} + \bar{\mu}_\Omega && \text{in } Q, \\ \nu\cdot\eta(\theta_{\bar{u}})\kappa\nabla\vartheta + \alpha\vartheta &= \bar{\mu}_\Gamma && \text{on } \Sigma, \\ \vartheta(T_1) &= \chi_E(\theta_{\bar{u}}(T_1) - \theta_d) + \bar{\mu}_{T_1} && \text{in } \Omega, \\[1ex] -\operatorname{div}(\sigma(\theta_{\bar{u}})\varepsilon\nabla\psi) &= -2\operatorname{div}(\sigma(\theta_{\bar{u}})\vartheta\varepsilon\nabla\varphi_{\bar{u}}) && \text{in } Q, \\ \nu\cdot\sigma(\theta_{\bar{u}})\varepsilon\nabla\psi &= 2\nu\cdot\sigma(\theta_{\bar{u}})\vartheta\varepsilon\nabla\varphi_{\bar{u}} && \text{on } \Sigma_N, \\ \psi &= 0 && \text{on } \Sigma_D. \end{aligned}$$

*is satisfied in the sense of Definition 4.10 and Remark 4.11. Moreover, $\bar{u}$ is the solution of the variational inequality*

$$\begin{gathered} \int_{\Sigma_N} \partial_t\bar{u}\,\partial_t(u-\bar{u}) + \frac{p}{2}|\bar{u}|^{p-1}\bar{u}(u-\bar{u}) + \frac{1}{\beta}(\tau_{\Gamma_N}\psi)(u-\bar{u})\,\mathrm{d}\omega\,\mathrm{d}t \geq 0 \\ \textit{for all } u \in \mathbb{U}^{ad} = \{u \in \mathbb{U}\colon 0 \leq u \leq u_{\max} \textit{ a.e. in } \Sigma_N\}. \end{gathered} \tag{4.15}$$

Note that the Lagrange multiplier $\bar{\mu}$ is not active on the set $\{T_0\}\times\overline{\Omega}$ due to Assumption 2.9 (xi) and the positivity and complementary conditions (4.3) and (4.4). Hence, $\bar{\mu}_{T_0}$ is zero and does not contribute to the system of equations in Theorem 4.14. Note moreover that the variational inequality in (4.15) is just a (semilinear) variational inequality of obstacle-type in time.

*Proof of Theorem 4.14.* Let $\bar{u}$ be a locally optimal control such that the linearized Slater condition (4.6) is satisfied. Theorem 4.8 then yields the existence of a Lagrangian multiplier $\bar{\mu} \in \mathcal{M}(\overline{Q})$ such that (4.3)-(4.5) hold true. We show that these lead to the assertions.

First consider the linear continuous functional

$$\langle\chi_E(\theta_{\bar{u}}(T_1) - \theta_d), \Theta\rangle_{L^2(\Omega)} := \int_E (\theta_{\bar{u}}(T_1) - \theta_d)\Theta\,\mathrm{d}x.$$

Due to the choice of $s$, we have $X_s \hookrightarrow C(\overline{\Omega}) \hookrightarrow L^2(\Omega)$, such that the functional is also an element of $X_s'$ and $\delta_{T_1}\otimes\chi_E(\theta_{\bar{u}}(T_1) - \theta_d) \in \mathbb{X}'$. Moreover, we set $\|\nabla\theta_{\bar{u}}\|^{s-q}_{L^q}\Delta_q\theta_{\bar{u}}$ as a functional on $\mathbb{X} \hookrightarrow L^s(J; W^{1,q})$ via

$$\langle\|\nabla\theta_{\bar{u}}\|^{s-q}_{L^q}\Delta_q\theta_{\bar{u}}, \xi\rangle_{\mathbb{X}} := \int_J \|\nabla\theta_{\bar{u}}(t)\|^{s-q}_{L^q}\langle\Delta_q\theta_{\bar{u}}(t), \xi(t)\rangle_{W^{1,q}}\,\mathrm{d}t.$$

The inclusion $\mathbb{X} \hookrightarrow C(\overline{Q})$ also implies $\bar{\mu} \in \mathcal{M}(\overline{Q}) \hookrightarrow \mathbb{X}'$. Hence, inserting $\theta_{\bar{u}}' = \mathcal{S}_{\mathcal{E}}'(u)h$ in (4.2), we immediately obtain

$$\begin{aligned}\partial_{\bar{u}}\mathfrak{L}(u,\mu)h = \langle \mathcal{S}_{\mathcal{E}}'(u)^*\left[\delta_{T_1}^*\chi_E(\theta_{\bar{u}}(T_1)-\theta_d) + \gamma\|\nabla\theta_{\bar{u}}\|_{L^q}^{s-q}\Delta_q\theta_{\bar{u}} + \mu\right], h\rangle_{\mathbb{U}} \\ + \beta\int_{\Sigma_N} \partial_t u\partial_t h + \frac{p}{2}|u|^{p-2}uh \,\mathrm{d}\omega\,\mathrm{d}t\end{aligned}$$

for $h \in \mathbb{U}$. Let us introduce $(\vartheta,\chi)$ as the unique solution of (4.9) (cf. Theorem 4.12) with data $\vartheta_T = \chi_E(\theta_{\bar{u}}(T_1)-\theta_d)+\bar{\mu}_{T_1}$, $g = 0$ and $f = \gamma\|\nabla\theta_{\bar{u}}\|_{L^q}^{s-q}\Delta_q\theta_{\bar{u}}+\bar{\mu}_{(T_0,T_1)}$, which is then also the solution of the formal system (4.8) with the stated inhomogeneities $f$ and $g$ and terminal value $\vartheta_T$. Here, $\psi$ is obtained by $\psi(\vartheta) = \mathcal{J}(\sigma(\theta_{\bar{u}}))^*(-\nabla\cdot\sigma(\theta_{\bar{u}})\vartheta\varepsilon\nabla\varphi_{\bar{u}})$, cf. Remark 4.11. Corollary 4.13 now shows that

$$\partial_{\bar{u}}\mathfrak{L}(\bar{u},\bar{\mu})h = \langle\mathcal{E}^*\psi, h\rangle_{\mathbb{U}} + \beta\int_{\Sigma_N}\partial_t\bar{u}\partial_t h + \frac{p}{2}|\bar{u}|^{p-2}\bar{u}h\,\mathrm{d}\omega\,\mathrm{d}t \tag{4.16}$$

for $h \in \mathbb{U}$. It is convenient to write $\mathcal{E}$ as $\mathcal{E} = \tau_{\Gamma_N}^* \circ \mathfrak{E}$ with $\mathfrak{E}\colon \mathbb{U} \hookrightarrow L^{2s}(J;L^{\mathfrak{p}}(\Gamma_N))$ and $\tau_{\Gamma_N}^*\colon L^{2s}(J;L^{\mathfrak{p}}(\Gamma_N)) \to L^{2s}(J;W_{\Gamma_D}^{-1,q})$ with $\mathfrak{p} > \frac{2}{3}q$, see Proposition 4.1 and Remark 2.10. Then we have

$$\langle\mathcal{E}^*\psi, h\rangle_{\mathbb{U}} = \langle\tau_{\Gamma_N}\psi, \mathfrak{E}h\rangle_{L^{(2s)'}(J;L^{\mathfrak{p}'}(\Gamma_N)),L^{2s}(J;L^{\mathfrak{p}}(\Gamma_N))} = \int_{\Sigma_N}(\tau_{\Gamma_N}\psi)h\,\mathrm{d}\omega\,\mathrm{d}t, \tag{4.17}$$

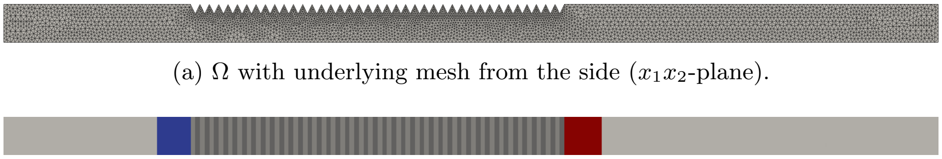
again $h \in \mathbb{U}$. Inserting (4.17) and (4.16) into (4.5), we obtain the stated variational inequality. □

Remark 4.15. *If the optimal control $\bar{u}$ in the previous theorem is an* interior *point of $\mathbb{U}^{ad}$, or if $\mathbb{U}^{ad}$ is not present at all, then one may transform the variational inequality* (4.15) *to the ordinary nonlinear differential equation of order two*

$$\partial_{tt}\bar{u} = \frac{1}{\beta}\tau_{\Gamma_N}\psi + \frac{p}{2}|\bar{u}|^{p-2}\bar{u}$$

*in the space $L^{p'}(\Gamma_N)$ as a boundary value problem with $\partial_t\bar{u}(T_0) = \partial_t\bar{u}(T_1) = 0$. In particular, $\partial_{tt}\bar{u} \in L^{(2s)'}(J;L^{p'}(\Gamma_N))$ in this case.*

**5. Application and numerical example.** As already outlined in [27] and the introduction, a typical example of an application for a problem in the form (P) is the optimal heating of a conducting material such as steel by means of an electric current. The aim of such procedures is to heat up a workpiece by electric current and to cool it down rapidly with water nozzles in order to harden it. In case of steel, this treatment indeed produces a hard martensitic outer layer, see for instance [7, Ch. 9.18] for a phase diagram and [7, Chapters 10.5/10.7 about Martensite], and is thus used for instance for rack-and-pinion actuators, to be found e.g. in steering mechanisms. The part of the workpiece to be heated up corresponds to the design area $E$ in the objective functional in (P). In order to avoid thermal stresses in the material, it is crucial to produce a homogeneous temperature distribution in the design area, which is reflected by the first term of the objective functional if we choose $\theta_d$ appropriately. The gradient term in the objective functional further enforces minimal thermal stresses. Moreover, the temperatures necesssary for the hardening process as described above are rather close to the melting point of the material, thus the state constraints are used to prevent the temperature exceeding the melting temperature



(a) $\Omega$ with underlying mesh from the side ($x_1x_2$-plane).

(b) $\Omega$ from above ($x_1x_3$-plane) with $\Gamma_N$ (left) and $\Gamma_D$ (right) emphasized.

Fig. 5.1: The computational domain $\Omega$ used in the numerical example.

$\theta_{\max}$. The control constraints in (P) represent a maximum electrical current which can be induced in the workpiece.

In the following we exhibit numerical examples for the optimal control of the three-dimensional thermistor problem in the form (P), underlining in particular the importance of the state-constraints. The considered computational domain $\Omega$ is a (simplified) three-dimensional gear-rack as seen in Figure 5.1, where the design area $E$ consists of the sawteeth. The mesh consists of about 80000 nodes, inducing 400000 cells with cell diameteres ranging from $8.8 \cdot 10^{-4}$ to $7.6 \cdot 10^{-3}$.

The heat-equation we use in the computations is as follows:

$$\varrho C_p \partial_t \theta - \operatorname{div}(\eta(\theta)\kappa\nabla\theta) = (\sigma(\theta)\nabla\varphi)\cdot\nabla\varphi.$$

It deviates from (1.1) by the factor $\varrho C_p$, the so-called the volumetric heat capacity, where $\varrho$ is the density of the material and $C_p$ is its specific heat capacity. However, since we assume $\varrho C_p$ to be constant, it certainly has no influence on the theory presented above. In [21, Remarks 6.13/15] and [25] it is laid out how to modify the analysis if one wants to incorporate a volumetric heat capacity depending on the temperature $\theta$. For a realistic modeling of the process, we use the data gathered in [12], i.e., the workpiece $\Omega$ is supposedly made of non-ferromagnetic stainless steel (#1.4301). The constants used can be found in Table 5.1 and the conductivity functions are given by

$$\sigma(\theta) := \frac{1}{a_\sigma + b_\sigma\theta + c_\sigma\theta^2 + d_\sigma\theta^3} \quad \text{for} \quad \theta \in [0, 10000] \text{ K},$$

with the constants $a_\sigma = 4.9659 \cdot 10^{-7}$, $b_\sigma = 8.4121 \cdot 10^{-10}$, $c_\sigma = -3.7246 \cdot 10^{-13}$ and $d_\sigma = 6.1960 \cdot 10^{-17}$ for the electrical conductivity (in $\Omega^{-1}\text{m}^{-1}$), and

$$\eta(\theta) := 100(a_\eta + b_\eta\theta) \quad \text{for} \quad \theta \in [0, 10000] \text{ K}$$

with $a_\eta = 0.11215$ and $b_\eta = 1.4087\cdot 10^{-4}$ for the thermal conductivity (in $\text{Wm}^{-1}\text{K}^{-1}$). Both functions are extended outside of $[0, 10000]$ in a smooth and bounded way, such that Assumption 2.9 (i) is satisfied. Note that $\varepsilon$ and $\kappa$ are each chosen as the identity matrix, as we do not account for heterogeneous materials in this numerical example. To counter-act on the different scales inherent in the problem, cf. the value for $u_{\max}$ and $\theta_0$ in Table 5.1, the model was nondimensionalized for the implementation.

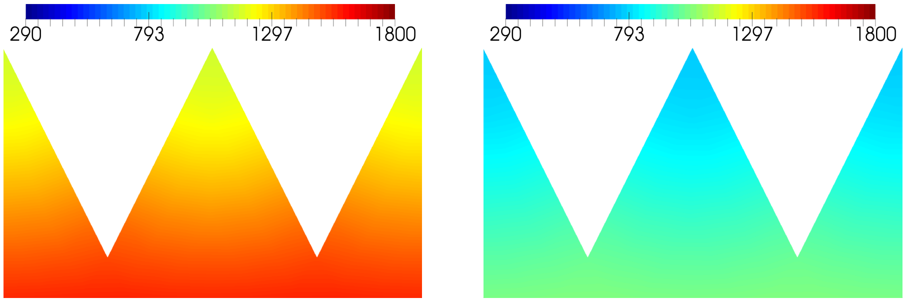
The optimization problem (P) is solved by means of a Nonlinear Conjugate-Gradients Method in the form as described in [13], modified to a projected method to account for the admissible set $\mathcal{U}_{\text{ad}}$. The method needed up to 150 iterations to meet the stopping criterion, which required the relative change in the objective functional to be smaller than $10^{-5}$. The state constraints in (P) are incorporated by a

| $\varrho$ | $C_p$ | $\alpha$ | $\theta_0$ | $\theta_l$ | $\theta_d$ | $\theta_{\max}$ | $u_{\max}$ |
|---|---|---|---|---|---|---|---|
| 7900 $\frac{\mathrm{kg}}{\mathrm{m}^3}$ | 455 $\frac{\mathrm{J}}{\mathrm{kg\,K}}$ | 20 $\frac{\mathrm{W}}{\mathrm{m}^3\,\mathrm{K}}$ | 290 K | 290 K | 1500 K | 1700 K | $10 \cdot 10^7 \frac{\mathrm{A}}{\mathrm{m}^2}$ |

Table 5.1: Material parameters used in the numerical tests

quadratic penalty approach—so-called Moreau-Yosida regularization—, cf. [26] and the references therein, where the penalty-parameter was increased up to a maximum of $10^{10}$, stopping earlier if the violation of the state constraints was smaller than $10^{-2}$ K. This resulted in a violation of $9.54 \cdot 10^{-2}$ K, which is about 0.0056% of the upper bound of 1700 K. In each step of the optimization algorithm, the nonlinear state equations (1.1)-(1.6) and the adjoint equations (4.8) have to be solved. We use an Implicit Euler Scheme for the time-discretization of these equations, whereas the spatial discretization is done via piecewise continuous linear finite elements. The nonlinear system of equations arising in each time-step is solved via Newton's method. Here, we do a semi-implicit pre-step to obtain a suitable initial guess for the discrete $\varphi$ for Newton's method. For the control, we also choose piecewise continuous linear functions in space where the values in the first and last timestep were pre-set to 0. In the calculation of the gradient of the reduced objective functional $j$, the gradient representation with respect to the $L^2(J; L^2(\Gamma_N))$ scalar product of the derivative of $u \mapsto \frac{1}{2}(\partial_t u)^2$ is needed, which one formally computes as $\partial^2_{tt} u$. We used the second order central difference quotient $\frac{u_{k+1}-2u_k-u_{k-1}}{\Delta t^2}$ to approximate $(\partial^2_{tt} u)(t_k)$ at time step $k$ with the appropriate modifications for the first and last time step, respectively. All computations were performed within the FEnICS framework [15].

For the experiment duration, we set $T_1 - T_0 = 2.0$ s with timesteps $\Delta t = 0.02$ s and $T_0 = 0.0$ s, while we use $\gamma = 10^{-8}$ and $\beta = 10^{-5}$ – this small value for $\beta$ is only possible due to the nondimensionalization performed. In the following, we elaborate on two settings: one in which we enforce the state constraint $\theta \le \theta_{\max}$ and one in which we do not.



(a) Free optimization.

(b) State constrained optimization.

Fig. 5.2: Detail of the sawteeth in $E$ at end time $t = 2.0$ s with distribution of the temperature $\theta$ in K.

Figure 5.2 shows the temperature distribution at end time $T_1 = 2.0$ s in $E$ in both cases. The desired temperature distribution close to uniformly 1500 K has been nearly achieved in the free optimization, see Figure 5.2a, at the price of very high temperature values around $\Gamma_D$ and $\Gamma_N$ already early in the heating process. We come back to this below, cf. also Figure 5.6. For the state-constrained optimization, we achieve a much worse result (note the same scales in both Figure 5.2a and 5.2b), which again corresponds to the rapid evolution to high temperatures at the critical areas, since these crucially limit the maximal amount of energy induced into the workpiece if one wants to prevent the temperature rising higher than the given bounds $\theta_{\max}$. This can also be seen in the development of the optimal controls in both cases over time, see below.

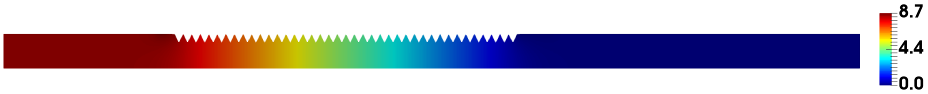


Fig. 5.3: The potential $\varphi$ (in V) associated with the optimal solution at time $t = 1.0$ s, view from the side ($x_1x_2$-plane).

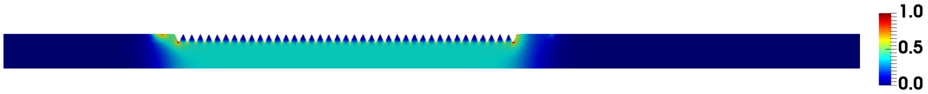
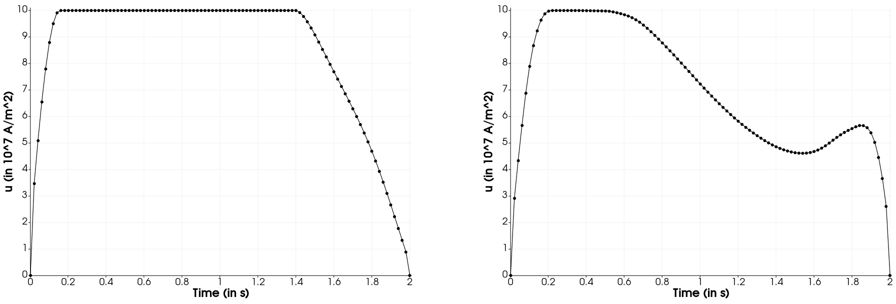


Fig. 5.4: Magnitude of the gradient $\nabla\varphi$ (in V/m) associated with the optimal solution at time $t = 1.0$ s, view from the side ($x_1x_2$-plane).

The potential $\varphi$ and its gradient $\nabla\varphi$ associated with the optimal control to the state-constrained optimization problem, at time $t = 1.0$ s are depicted in Figures 5.3 and 5.4. Here, $\nabla\varphi$ is to be understood as the projection of the potentially discontinuous gradient of $\varphi$ to the space of continuous linear finite elements. The potential $\varphi$ decreases from $\Gamma_N$ to the grounding with prescribed value $\varphi \equiv 0$ at $\Gamma_D$, cf. Figure 5.1b, thus inducing a current flow and acting as a heat source between $\Gamma_D$ and $\Gamma_N$, since the corresponding term in the heat equation $\sigma(\theta)\varepsilon\nabla\varphi \cdot \nabla\varphi$ is proportional to $|\nabla\varphi|^2$ due to the coercivity and boundedness of $\varepsilon$ and the bounds on $\sigma$. This is confirmed by the magnitude of $\nabla\varphi$ as seen in Figure 5.4. In particular one observes that $\nabla\varphi$ is very small or 0 in $E$, which means that the current flows only through the area between $\Gamma_D$ and $\Gamma_N$ and right *below* $E$, heating only this part of the workpiece.

The optimal controls are shown in Figure 5.5, taken at an arbitrary but fixed grid point in $\Gamma_N$. The high values in the control at the beginning of the process seem to be the result of the inability to heat up the tooth rack in the design-area $E$ directly as explained above, which makes heating of the teeth reliant on diffusion. This in turn requires the needed total energy to be inserted into the system as fast as possible, resulting in high control values, which also agrees with the requirement to obtain a *uniform* temperature distribution in the tooth rack. These considerations also underline the necessity of control bounds in this example. In decreasing the control values after the inital period, the opimization procedure in the free optimization is avoiding to "over-shoot", i.e., to produce a higher temperature than desired. In the case of state-constrained optimization, the presence of the state constraints forces an earlier decrease in control values in order to not violate the upper bound $\theta_{\max}$,

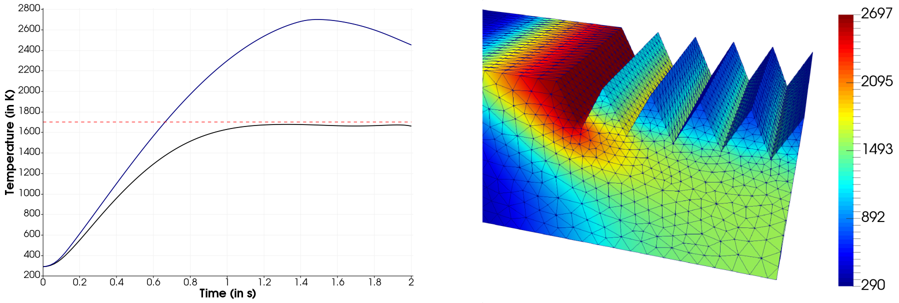


(a) For the unconstrained problem.

(b) For the state-constrained problem.

Fig. 5.5: Time plot of the optimal controls, taken at an arbitrary but fixed grid point in $\Gamma_N$.

which is then compensated by a slightly higher level of values towards the end of the simulation. This, however, is clearly not enough to make up for the earlier decrease as seen in Figure 5.2.



(a) Time plot of the temperature in a point close to $\Gamma_N$.

(b) Temperature $\theta$ in K at the critical area near $\Gamma_N$ at time $t = 1.4$ s.

Fig. 5.6: Influence and necessity of state constraints.

Figure 5.6 illustrates why state constraints are a necessary addition to an appropriate model of the industrial steel heating process. Figure 5.6a shows the temperature evolution in a point in one of the two critical regions, which are the points near $\Gamma_D$ and $\Gamma_N$, see also Figure 5.6b and the magnitude of $\nabla\varphi$ at this region in Figure 5.4. In this case, the point lies in $E$ close to $\Gamma_N$, but we emphasize that the state constraints hold in the whole $\Omega$ and are not limited to $E$. The upper line in Figure 5.6a corresponds to the temperature associated to the optimal solution of the unconstrained optimization, while the lower belongs to the state-constrained optimal solution, with the upper bound $\theta_{\max} = 1700$ K marked by the dashed line. In the free optimization case, the temperature exceeds the bounds already at about one third of the simulation time and continuous to rise to almost 1000 K above $\theta_{\max}$. On the other hand, the temperature obtained from the state-constrained case stays below the threshold, as required. Note here that the evaluated point is chosen as one of those where the temperature rises highest overall, compare the temperature distribution as seen in Figure 5.6b and the maximal temperature achieved in the free optimization case in Figure 5.6a.

Concluding from the results presented above, it becomes apparent that the prescribed time of 2.0 s is too short to heat up the workpiece in the given geometry enough to reach the required temperature for Austenite to form in the workpiece (cf. [7, Ch. 9.18]) in $E$, if melting is to be prevented.

## REFERENCES

[1] H. Amann, *Nonautonomous parabolic equations involving measures*, J. Math. Sci., 130 (2005), pp. 4780–8002.

[2] H. Amann and P. Quittner, *Optimal control problems with final observation governed by explosive parabolic equations*, SIAM J. Control Optim., 44 (2005), pp. 1215–1238.

[3] M. R. Sidi Ammi, *Optimal control for a nonlocal parabolic problem resulting from thermistor system*, Int. J. Ecological Economics & Statistics, 9 (2007), pp. 116–122.

[4] S. N. Antontsev and M. Chipot, *The thermistor problem: Existence, smoothness, uniqueness, blowup*, SIAM J. Math. Anal., 25 (1994), pp. 1128–1156.

[5] P. Auscher, N. Badr, R. Haller-Dintelmann and J. Rehberg, *The square root problem for second-order, divergence form operators with mixed boundary conditions on $L^p$*, J. Evol. Equ., 15 (2014), pp. 165–208.

[6] J. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer Science + Business Media, 2000.

[7] W.D. Callister, *Materials Science And Engineering: An Introduction*, John Wiley & Sons, 2007.

[8] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[9] E. Casas and F. Tröltzsch, *First- and second-order optimality conditions for a class of optimal control problems with quasilinear elliptic equations*, SIAM J. Control Optim., 48 (2009), pp. 688–718.

[10] E. Casas and J. Yong, *Maximum principle for state-constrained control problems governed by quasilinear elliptic equations*, Differential Integral Equations, 8 (1995), pp. 1–18.

[11] G. Cimatti, *Optimal control for the thermistor problem with a current limiting device*, IMA J. Math. Control and Information, 24 (2007), pp. 339–345.

[12] S. Clain et al., *Numerical modeling of induction heating for two-dimensional geometries*, Math. Models Methods Appl. Sci., 3 (1993), pp. 271–281.

[13] Y. H. Dai and Y. Yuan, *A Nonlinear Conjugate Gradient Method with a Strong Global Convergence Property*, SIAM J. Optim., 1 (1999), pp. 177–182.

[14] K. Disser, H.-C. Kaiser and J. Rehberg, *Optimal Sobolev regularity for linear second-order divergence elliptic operators occuring in real-world problems*, SIAM J. Math. Anal., 3 (3012), pp. 1719-1746.

[15] A. Logg, K.-A. Mardal and G. Wells (eds.), *Automated Solution of Differential Equations by the Finite Element Method*, Springer, Berlin Heidelberg, 2012.

[16] I. Fonseca and G. Parry, *Equilibrium configurations of defective crystals*, Arch. Rat. Mech. Anal., 120 (1992) pp. 245–283.

[17] J. A. Griepentrog, K. Gröger, H. C. Kaiser, and J. Rehberg, *Interpolation for function spaces related to mixed boundary value problems*, Math. Nachr., 241 (2002), pp. 110–120.

[18] J.A. Griepentrog, W. Höppner, H.-C. Kaiser, J. Rehberg, *A bi-Lipschitz, volume-preserving map form the unit ball onto the cube*, Note Mat., 1 (2008), pp. 177–193.

[19] K. Gröger, *A $W^{1,p}$–estimate for solutions to mixed boundary value problems for second order elliptic differential equations*, Math. Ann., 283 (1989), pp. 679–687.

[20] K. Gröger, *$W^{1,p}$–estimates of solutions to evolution equations corresponding to nonsmooth second order elliptic differential operators*, Nonlinear Analysis, 18 (1992), pp. 569–577.

[21] R. Haller-Dintelmann and J. Rehberg, *Maximal parabolic regularity for divergence operators including mixed boundary conditions*, J. Differ. Equations, 247 no. 5 (2009), pp. 1354–1396.

[22] R. Haller-Dintelmann, C. Meyer, J. Rehberg and A. Schiela, *Hölder continuity and optimal control for nonsmooth elliptic problems*, Appl. Math. Optim., 60 (2009), pp. 397–

428.

[23] R. Haller-Dintelmann and J. Rehberg, *Coercivity for elliptic operators and positivity of solutions on Lipschitz domains*, Arch. Math., 95 no. 5, (2010), pp. 457–468.

[24] R. Haller-Dintelmann and J. Rehberg, *Maximal parabolic regularity for divergence operators on distribution spaces*, in Parabolic Problems, The Herbert Amann Festschrift, Series: Progress in Nonlinear Differential Equations and Their Applications, Vol. 80, J. Escher, P. Guidotti, P. Mucha, J.W. Prüss, Y. Shibata, G. Simonett, C. Walker, W. Zajaczkowski, eds., Basel, 2011.

[25] M. Hieber and J. Rehberg, *Quasilinear parabolic systems with mixed boundary conditions*, SIAM J. Math. Anal., 40 (2008), pp. 292–305.

[26] M. Hintermüller and K. Kunisch, *Feasible and non-interior path-following in constrained minimization with low multiplier regularity*, SIAM J. Control Optim., 45 (2006), pp. 1198–1221.

[27] D. Hömberg, C. Meyer, J. Rehberg and W. Ring, *Optimal control for the thermistor problem*, SIAM J. Control Optim., 48 no. 5, (2010), pp. 3449–3481.

[28] D. Hömberg, K. Krumbiegel and J. Rehberg *Optimal control of a parabolic equation with dynamic boundary condition* Appl. Math. Optim., 67 (2013), pp. 3–31.

[29] V. Hrynkiv, S. Lenhart, and V. Protopopescu, *Optimal control of convective boundary condition in a thermistor problem*, SIAM J. Control Optim., 47 (2008), pp. 20–39.

[30] K. Krumbiegel and J. Rehberg, *Second order sufficient optimality conditions for parabolic optimal control problems with pointwise state constraints*, SIAM J. Control Optim, 51 (2013), pp. 304–331.

[31] H.-C. Lee and T. Shilkin, *Analysis of optimal control problems for the two-dimensional thermistor system*, SIAM J. Control Optim., 44 (2005), pp. 268–282.

[32] V.G. Maz'ya, *Sobolev Spaces*, Springer, Berlin-Heidelberg-New York-Tokyo, 1985.

[33] H. Meinlschmidt, C. Meyer and J. Rehberg, *Optimal Control of the Thermistor Problem in Three Spatial Dimensions, Part 1: Existence of Optimal Solutions*, submitted, 2016.

[34] H. Meinlschmidt and J. Rehberg, *Hölder-estimates for non-autonomous parabolic problems with rough data*, Evol. Equ. Control Theory, 6 (2016), pp. 147–184.

[35] J. Prüss, *Maximal regularity for evolution equations in $L^p$-spaces*, Conf. Semin. Mat. Univ. Bari, 285 (2002), pp. 1–39.

[36] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North Holland, Amsterdam, 1978.

[37] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.